Personalized proteomic profiles enabled by advances in mass spectrometry-based proteomics

Alba Cristóbal González de Durana

dedicated to my family and partner

ISBN:978-90-393-6759-9

The research in this thesis was performed in the Biomolecular Mass Spectrometry and Proteomics Group, Utrecht University, Utrecht, The Netherlands

The research was financially supported by the Netherlands Proteomics Center (NPC)

Printed by ProefschrifMaken || www.proefschriftmaken.nl

Personalized proteomic profiles enabled by advances in mass spectrometry-based proteomics

Gepersonaliseerde proteomics profielen mogelijk gemaakt door nieuwe ontwikkelingen in de massaspectrometrische bepaling van eiwitten

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 1 mei 2017 des middags te 12.45 uur

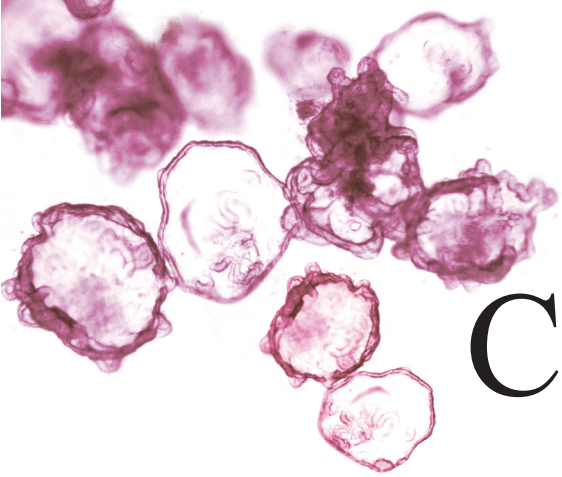
door

Alba Cristóbal González de Durana

geboren op 12 juni 1988 te Vitoria-Gasteiz, Spanje

Promotor:	Prof. dr. A. J. R. Heck
Copromotor:	Dr. S. Mohammed

Table of contents



Chapter 1

Introduction

Table of contents

Part I. Proteomics

Proteins are essential molecules that constitute the molecular entities through which genetic information is expressed.[1] Genes encoded by the DNA are first transcribed into RNA, which is then used as a template for protein synthesis by the ribosome. To study these processes globally is quite complex and therefore each of these level has traditionally their own discipline. The genetic material of an organism, the genome, is studied by genomics.[2,3] The complete set of RNA transcripts that are produced by the genome, the transcriptome, is studied by transcriptomics,[4] while the total protein complement of the genome, the proteome,[5,6] is the topic of study for proteomics.[7] However, conceptually (and in complexity) the proteome differs quite a bit from the genome (Figure 1).

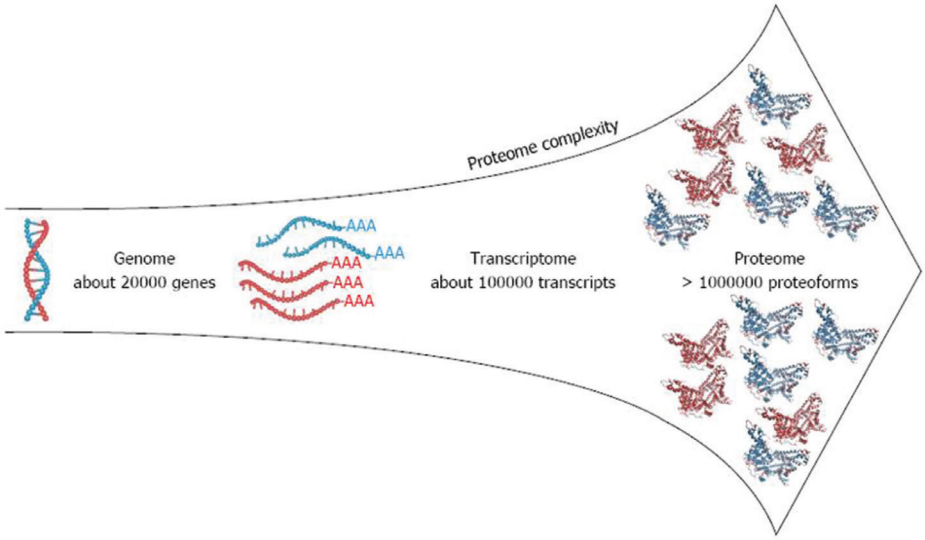


Figure 1. Overview of the increased diversity of the proteome compared to the genome and transcriptome.

Compared to the relatively static character of the genome, a much more dynamic nature is found at the proteome level.[8] The proteome is a context dependent entity and each tissue of a single organism has a unique proteome.[9] Furthermore, modifications occurring during the transcription of DNA into mRNA (alternative splicing) and during the translation of mRNA into protein (post-translational modifications (PTMs)) further increase the complexity of the proteome compared to that of the genome. PTMs are frequently occurring covalent modifications, by

cellular enzymes, transforming the side chains of amino acid residues during or after protein biosynthesis.[10] Over 300 different protein PTMs have been identified, whereby protein phosphorylation, acetylation, methylation and ubiquitination have so far been the most widely investigated.[11] PTMs can modulate molecular interactions, protein localization, activity state and turnover. Failure to control these complex molecular processes is detrimental or fatal for the survival of the cell and it is no surprise that a range of post-translationally modified proteins and their substrates are implicated in human diseases,[12] including cancers.[13,14] Another important feature to be addressed in proteomics is the broad dynamic range (difference between the most and least abundant protein) in which proteins are present in the cell, tissue or body fluid. Therefore, typically a rather large amount of starting material is needed for a 'comprehensive' proteome study.[15] Efficient sample preparation and sensitive separation and detection methods are required as amplification tools, as available for DNA, are not available for the study of proteins. Sample material levels are typically in the nanogram range and samples with limited starting material such as clinical samples,[16,17] or specifically sorted cells[18] are hard to study but often of special interest. These three major factors: the complexity of the proteome, the high dynamic range at which proteins are present in biological samples and the sensitivity of MS-based proteomics techniques for proteins of low abundance, make the study of proteins still a very challenging and intriguing task.

Currently, mass spectrometry (MS)-based studies are dominating the landscape for the study of proteins at a proteome wide level. In order to identify the proteins present, first of all, proteins need to be extracted from the biological matrix (i.e. cells, tissue and biological fluids). Disruption of the cell membrane (usually performed by sonication) and solubilization and isolation of the proteins is often the first step required in the proteomics workflow. In this manner mainly cytosolic proteins are recovered. Therefore, in order to recover proteins from a distinct location, e.g. the cell membrane, specific extraction steps are required.[19] Subsequently a protein digestion has to be performed to obtain peptides. This is typically achieved after denaturation, reduction of the protein's disulfide bonds and alkylation of the resulting reduced ends, which prevents re-folding of the proteins. The solubilized proteins are then cleaved using a sequence-specific protease. Trypsin is the most

commonly used enzyme due to its highly specific cleavage at the C-terminus of Arg and Lys amino acid residues, leading to the generation of peptides suitable for LC-MS identification. Depending on the complexity of the sample and/or the biological questions that need to be addressed, a pre-fractionation and/or an enrichment strategy can be included. Afterwards, peptides are separated by liquid chromatography and analyzed by mass spectrometry. The mass-to-charge ratio (m/z) of the peptide ions is measured first by mass spectrometry to determine the molecular mass of each precursor peptide. However, this information is often not sufficient to identify the amino acid sequence of the peptide. Therefore, peptide ions are subsequently isolated in the first mass analyzer and fragmented, a process known as tandem mass spectrometry (MS/MS). The acquired data is processed by several bioinformatics tools in order to get the peptide identified, and to find its protein of origin. Often the (relative) abundance of the proteins is also quantified based on the abundance of the peptides in the sample. These peptide centric approaches are generally classified as shotgun or bottom-up proteomics[20] (Figure 2) and represent presently the most mature and widely used approaches for protein identification, characterization and quantification.

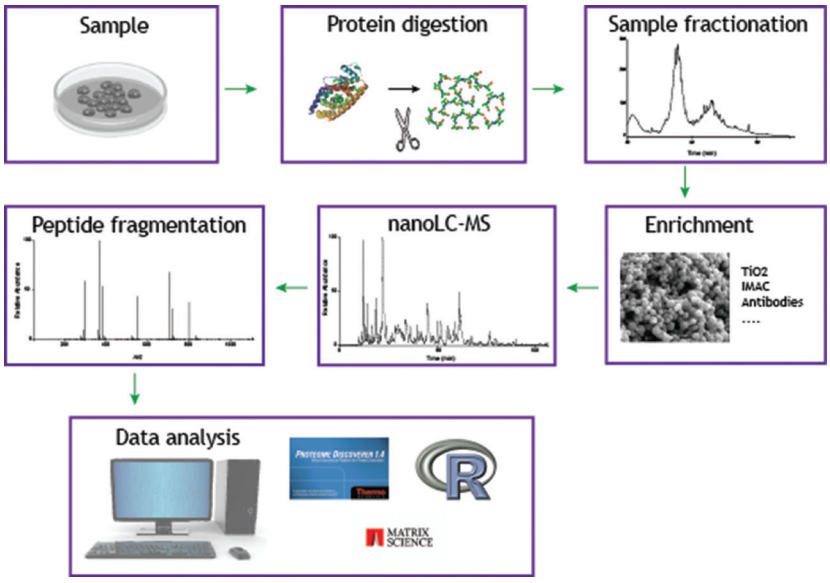


Figure 2. Overview of the generic bottom-up, or shot-gun, proteomics workflow.

Even though the bottom-up proteomics field is rapidly evolving, there are still

limitations hampering this approach. One of the major drawbacks of this approach is the increase in complexity of the sample due to the generation of peptides from an already complex mixture of proteins. Furthermore, due to the relatively short length of the generated peptides, the observation of (possibly co-occurring) neighboring PTMs and the detection of alternative splice variants is challeging.[21] The continuous developments in sample preparation, LC and MS over the last decade have enabled and facilitated large scale protein sequencing of samples. However, despite the immense progress in shot-gun proteomics, the comprehensive characterization of proteomics by bottom-up approaches still represents a major challenge. A more protein centric approach, known as top-down proteomics,[22] circumvents some of the limitations present in bottom-up approaches. Top-down proteomics faces some advantages as it provides complete molecular specificity on intact proteins, enabling the analysis of proteoforms[23] as intact proteins harbor the entire set of (co-occurring) PTMs.[24,25] The main limitations in top-down proteomics are related to the difficulty of efficiently separating proteins by chromatography or electrophoresis, and the relatively inefficient formation of fragment ions from proteins, when compared to peptides.

With the aim of circumventing the limitations of the currently used two main proteomics approaches, a compromise between both strategies; termed middle-down proteomics is gaining popularity.[26–28] This approach also uses protein digestion as in bottom-up proteomics, but aims to yield relatively larger peptides (ideally (far) above 3 kDa).[29–32] Middle-down proteomics exhibits particular advantages as the complexity of the digests decreases (as less peptides are formed), and may also allow better proteome coverage, including the identification of splice-variants and other isoforms. Furthermore, longer peptides increase the probability to detect multiple co-occurring neighboring PTMs, important to study functionally relevant PTM crosstalk.

Separation techniques

In bottom-up proteomics peptide separation is essential to reduce the sample complexity prior to MS analysis. Therefore, in order to perform as comprehensive as possible characterization of proteomes efficient liquid chromatography technologies are required prior to MS analysis. The term "chromatography", which

literally means "color writing", was introduced at the beginning of the 20th century by Twsett.[33] Twsett based the chromatographic separation on adsorption and since then several alternative separation methods had been developed.[34–36] The separation in chromatography is based on the distinct affinities of the components of the analyte towards the stationary and mobile phase. The differences in affinities arise due to the relative adsorption or partitioning between both phases. Adsorption varies due to the polarity of the components towards the stationary phase, while partition varies due to the solubility of components into different liquids. Three main factors affect the achievable resolution from a chromatographic separation: the selectivity, the retention and the efficiency (Figure 3).

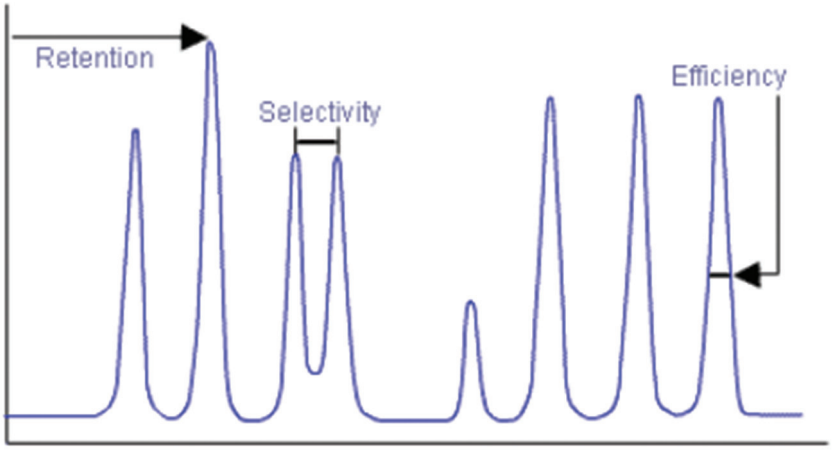


Figure 3. Schematic representation of three important parameters (retention, selectivity and efficiency) contributing to the achievable resolution in a chromatographic separation.

The selectivity is the ability of a chromatographic system to distinguish between sample components. The retention factor measures the retention of an analyte on the chromatographic column and the efficiency is a measure of the dispersion of the analyte band as it migrates through the LC system and the column. A wider analyte band leads to increased chromatographic peak width. The wider band results in a dilution effect that produces a decrease in peak height accompanied by a loss in sensitivity and resolution. Conversely, if the band broadening is minimized, narrower chromatographic bands are achieved, resulting in a higher efficiency. These sharper peaks allow higher sensitivity and resolution due to a more concentrated analyte band. It is therefore important to be aware of the

factors that influence band broadening to improve the overall chromatographic performance. Van Deemter derived an equation that includes the main factors contributing to band broadening.[37] Variations in the flow path (Eddy diffusion), dispersion of the analyte due to the concentration gradient at the outer edges of the band (longitudinal diffusion) as well as in the porous particles of the stationary phase (mass transfer) are factors present in this equation (Figure 4a).

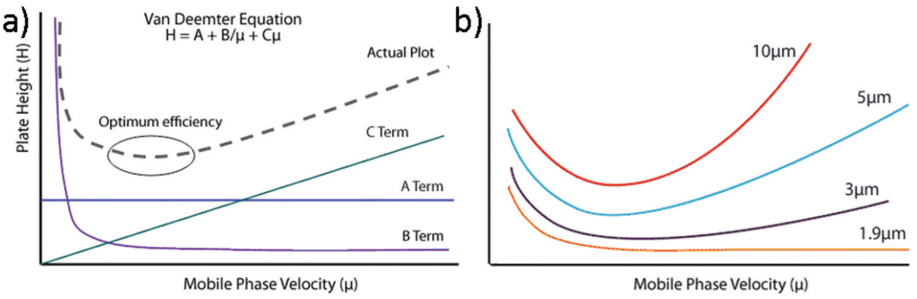


Figure 4. a) The Van Deemter equation and the representation of the parameters in the equation. The A term represents the Eddy diffusion, the B term is the longitudinal diffusion and the C term is the mass transfer. b) Smaller particle sizes yield higher overall peak efficiencies and a much wider range of usable flow rates.

- **Reversed Phase liquid chromatography**

Reversed Phase (RP) chromatography, a type of partition chromatography, is the most widely used technique for sample preparation and peptide separation in peptide-centric proteomics, pioneered by Horvath and Molnár.[38] The high separation power as well as the compatibility of the solvents used with ESI-MS made RP chromatography the method of choice for peptide separation.[39] In this type of chromatography the separation is based on the partitioning of the analytes between the hydrophobic stationary phase and the polar hydrophilic mobile phase. The mobile phase generally consist of water, an organic solvent (usually acetonitrile is used) and often small molecule additives to adjust the pH or influence interactions between the analyte and the stationary phase. Most typically C18-based materials are used for peptide separation in RP LC. Initially, peptides are dissolved in an aqueous acidic solution and are exposed to the hydrophobic stationary phase. Subsequently, the concentration of the organic solvent is gradually increased in order to elute sequentially all peptides. Subsequently, the eluted peptides are ionized

and analysed in the mass spectrometer as described further in the following section.

One development that had a great impact on RP chromatography is ultra high pressure liquid chromatography (UHPLC)[40], which is now most widely used in the proteomics field.[40–43] This system enables, due to the achievable high pressure, the use of smaller particle sizes (sub 2 um) in the column leading to a higher efficiency and a much wider range of usable flow rates (Figure 4b).

In RP-LC the pH at which the separation is carried out influences the properties of the charged residues in the peptides due to protonation or deprotonotation. Neutralization of charged residues leads to decreased hydrophilicity (increased hydrophobicity) and as a consequence changes the retention of peptides. Such features can be exploited in two-dimensional separations as discussed later.

- **Strong Cation eXchange chromatography**

Strong cation exchange (SCX) chromatography is a particular type of ion exchange chromatography (IEX), extensively employed in proteomics.[44,45] In SCX, the analytes are principally separated based on their net positive charge, as a result of the coulombic interactions between the positive charges of the molecules and the negatively charged stationary phase (the functional groups are strong acids).[46–48] As it will be discussed in the next section, SCX is widely used, especially in multi-dimensional separation approaches. Furthermore, the study of PTMs, such as phosphopeptides is an interesting application of SCX as it can be used as a selective enrichment method.[49] When the pH is kept at approximately 3, the acidic amino acids, glutamic and aspartic acid are protonated and do not contribute to the net charge of the peptide. Only basic residues and highly acidic residues like phosphorylated Ser and Thr residues contribute to the net charge of the peptide, reducing the overall positive net charge. This makes the separation of distinct groups of phosphopeptides from the non-phosphorylated peptides feasible.

- **Multidimensional separation**

Despite the high performance of RP chromatography for peptide separation, one step of separation is mostly not enough for deep proteome coverage due to the previously described complexity of the proteome. Several chromatographic techniques are currently used as a first separation step prior to RP-LC-MS/MS to decrease the complexity of the sample.[50] Fractions of the elution of the first separation/dimension are then transferred to the second dimension to further separate components that are not separated well in the first dimension. In such a two dimensional approach, a high orthogonality between the two modes of separation is desired. The separation is orthogonal if the two separation mechanisms are independent of each other, so that the distribution of component zone in one dimension is not correlated with the zone distribution in the other dimension[51]. The advantage of multidimensional separations is that its separation efficiency, measured by the peak capacity (the maximum number of resolvable peaks on a given LC column within a gradient time introduced by Giddings)[52] will be the product of each separation modes,[53] boosting further the resolving power of the system. This performance is directly dependent of the separation orthogonality.[54] If the selectivity of the two LC modes is not completely orthogonal, the achievable peak capacity will be lower. The type of chromatographic separation present in the first dimension covers different physicochemical properties of the peptides such as their polarity (hydrophilic interaction chromatography (HILIC), high pH reversed phase chromatography (HpH)) or charge (IEX). Conventionally SCX, had been commonly coupled to RP in proteomics.[44,45] Despite the clear difference in selectivity compared to the RP mode, due to the nature of tryptic peptides (mainly +2 and +3 charged) the orthogonality of these modes is also not ideal, leading to a non-optimal peak capacity of the SCX-RP combination.[55] As it was previously described, the selectivity in reversed phase chromatography is dependent on the pH. Therefore the coupling of two RP rounds is currently gaining popularity due to the high peak capacity achievable.[56–58] The orthogonality achievable operating RP at different pH values has been shown to be comparable to the SCX-RP modes.[59]

The coupling of the chromatographic separation techniques can be carried out in on-line or off-line set ups.[50] In off-line set ups fractions from the first dimension are collected and analyzed by the second dimension.[45] On the other hand, in on-

line setups the eluent from the first dimension is directly transferred onto the next one.[60] In this case the incompatibility of the mobile phases between the consecutive dimensions represents a major obstacle while in the off-line set ups this issue can be avoided. Furthermore, the design and operation are simplified making off-line set up often a more flexible configuration. However, the higher amount of sample required and longer analysis times are some of the main disadvantages of off-line set ups.

Mass spectrometry instrumentation

Mass spectrometry initially emerged in the field of physics and its beginning date back to the late 19th century.[61] The first studies were performed using gas discharge tubes[62] in which energetic ions with a wide range of velocities were produced and deflected with magnetic and electric fields. The construction of what is known as the first mass spectrometer, the 'parabola spectrograph',[63] came some years later and since then extensive work has been done in the field yielding to the generation of a broad selection of mass spectrometers. All mass spectrometers have three main components in common, an ion source, a mass analyser(s) and a detector. The ion source converts the analyte molecules into gas-phase ions, afterwards the mass analyser separates those ions and the detector records the number of ions at each m/z value. Mass analysers are the key element of mass spectrometers and can be divided according to their operation mode, which involves either trapping or non-trapping and according to the way ion separation is performed which can be either scanning or non-scanning mode.

- **Ionization techniques**

The importance of the ionization techniques lies in the fact that biomolecules can be analysed by mass spectrometry only if they are ionized and brought into the gas phase. Proteins are generally polar, non-volatile and thermally unstable compounds. Therefore, it is a great challenge to convert proteins present in the liquid or solid phase into the gas phase. The main breakthrough that made possible the study of proteins as we know it nowadays, is the development of suitable ionization techniques for these biomolecules. At the beginning of the era 'hard' ionization techniques were used which, due to the high energy

applied, were fragmenting the peptides resulting in hard to interpret spectra.[64,65] Currently, two techniques dominate the landscape, namely electrospray ionization (ESI)[66] and Matrix-Assisted Laser Desorption/Ionization (MALDI).[67] Both techniques are 'soft' ionization techniques in which a lower energy is employed avoiding the dissociation of the analyte and leading to a simplification of the spectra. Thanks to these two breakthroughs, the range of biological questions which were possible to address by mass spectrometry was substantially broadened.

- *Electrospray ionization*

The development of ESI mass spectrometry in the late 80's is attributed to Fenn.[68,69] Fenn's studies were based on the work done by Dole some years before[70] who described the idea that electrospraying a liquid containing analyte molecules might liberate these as ions and make them amenable for mass spectrometry analysis. Interestingly the principles behind electrospray were already described theoretically by Lord Rayleigh in 1882.[71] The importance of this development lead to Fenn being awarded the Nobel Prize in Chemistry in 2002.[72]

In ESI a strong electric field is applied to the analyte, which is dissolved in a solvent, as it passes through a capillary at a low flow rate and atmospheric pressure. The potential difference is applied between the conductive tip of the capillary and the orifice of the mass spectrometer, generating the electric field. The solvent containing the peptides is thereby electrostatically dispersed and this generates highly charged droplets. The high voltage produces an accumulation, via repulsion, of charged analyte ions at the tip of the capillary. This charge accumulation deforms the surface of the solvent creating the so-called Taylor cone,[73] from which charged droplets are released. Through evaporation of the volatile solvent, the charge density of the droplets increases until the so-called Rayleigh point is reached. The fission of the droplets into smaller droplets occurs as the Coulombic repulsion exceeds the surface tension of the droplet (Figure 5).[74]

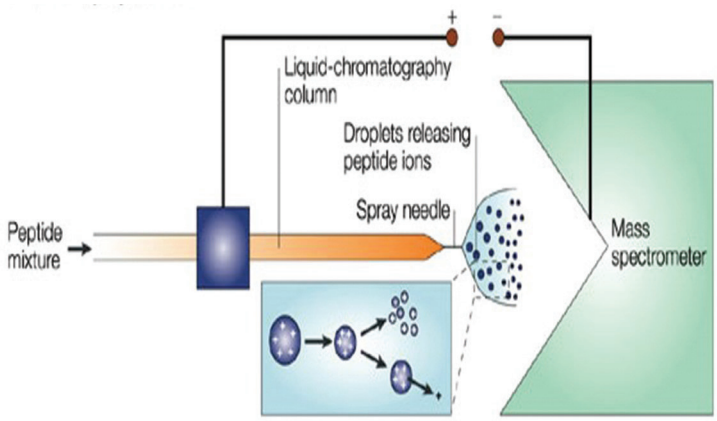


Figure 5. Schematic representation of the electrospray ionization process. Figure adapted from Steen *et al.*[75]

Regarding the ionization, two models have been proposed to explain the mechanism of how gas phase ions are generated from very small highly charged droplets: the charge residue model (CRM)[76] and the ion evaporation model (IEM).[77] The CRM model, attributed to Dole, describes the process as a series of fission events which leads to the production of small droplets each of which contains only one analyte molecule. When the last few solvent molecules evaporate, the analyte becomes a free gas phase ion by retaining some of its droplet's charge. The CRM model is the most likely mechanism for the creation of gas-phase peptide and protein ions.[78] On the other hand, the IEM model proposes that solvated ions are emitted, evaporated directly from charged droplets after the radii of the droplets decrease to a suitable size. These charged droplets are much larger and carry more charges than in the CRM model. This mechanism is thought to be favourable for small (in)organic ions.[78]

Reducing droplet size has been shown to improve ionization efficiency and this led to the introduction of microelectrospray[79,80] and nanoelectrospray ionization.[76] Nano-ESI is generally operated at flow rates ranging from 50 to 500 nL/min.[81] An essential advantage of nano-ESI is the small amount of sample required and the higher sensitivity obtained compared to ESI.[82] The creation of smaller-diameter droplets through nano-ESI also contributed to have a more efficient charging of gas phase analyte molecules. Furthermore, it leads to lower competition and suppression effects due to the presence of less analytes in smaller droplets that ultimately lead to an extension of the dynamic range.[83]

- *Matrix-Assisted Laser Desorption/Ionization*

Matrix-assisted laser desorption/ionization mass spectrometry was started by Hillenkamp and Karas with their observations made by a device called LAMMA (LAser Micropulse Mass Analyser).[84] The applicability of MALDI to larger and thermally labile biomolecules was shorty after described by the same authors.[85,86] However, it was Tanaka who, in the same year, was able to ionize large proteins by combining a laser and a matrix.[67] Tanaka, alongside with Fenn, was awarded the Nobel Prize in Chemistry in 2002 for his work in the development of these 'soft' ionization methods.[72]

MALDI provides a non-destructive vaporization and ionization of both large and small biomolecules. The analyte is first co-crystallized with a matrix compound that will strongly absorb the light from a laser. Laser radiation of this analyte-matrix mixture results in the ionization of the molecules of the matrix. The charge is then (partially) transferred from the matrix to the analyte generating peptide ions. The rapid heating, produced by the laser radiation, lead to the vaporization of the matrix, which carries the analyte with it. Thanks to the limited amount of energy transferred from the matrix to the analyte little fragmentation is observed. The process generates mostly singly charged ions (Figure 6).[75]

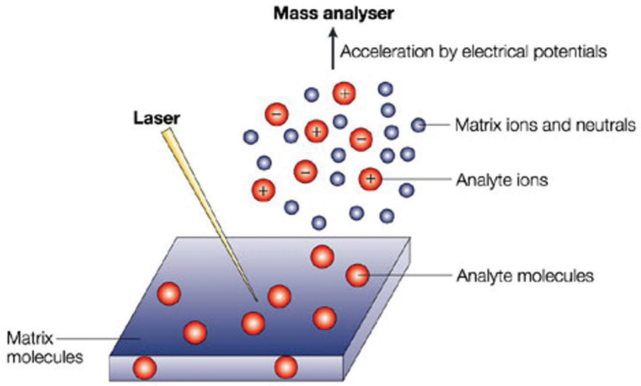


Figure 6. Schematic representation of Matrix-Assisted Laser Desorption/Ionization. Figure adapted from Steen *et al.*[75]

The desorption and the ionization processes that take place in MALDI are complex and still not fully understood. Two main theories explain the desorption of large molecules: the thermal-spike model[87] and the pressure pulse theory.[88] The first

one attributes the ejection of the molecules to poor vibrational coupling between the analyte and the matrix, while in the second theory a pressure gradient from the matrix to the surface facilitates the desorption of the molecules. Regarding the ionization process, two conflicting theories were initially proposed: the lucky survivor theory[89] and the gas phase protonation model.[90] The first one is based on the premise that all ions are already incorporated in the matrix crystals before the laser ablation. Only the "lucky" ions will be detected while the others will be neutralized or undergo extensive fragmentation due to higher charge and being more energetic. In comparison to this model, the gas phase protonation model predicts the protonation of the analyte in the gas phase by collisions with protonated or deprotonated matrix ions. It is now believed that a more complex interaction between analyte and matrix which combines both models, is responsible for the ionization.[91] Due to its pulse nature, which limits the optimal use of the sequencing speed of current mass analysers, and thanks to the ease of hyphenation of ESI to LC, the use of MALDI is currently less used in proteomics when compared to ESI.

- **Mass analyzers**
    - *Quadrupole analyzer*

Quadrupoles consist of four precisely parallel metal rods equally placed around a central axis. Opposite rods are connected and a combination of static and oscillating electric fields is applied to these rods. Gas phase ions are introduced along the axis into the middle of the radially positioned rods. Due to the direct-current (DC) and radiofrequency (RF) field applied, a turbulent environment is created in which ions may or may not possess a stable trajectory. In this way the ions are filtered and only selected m/z ions satisfying specific RF and DC settings will pass through the quadrupole. All other ions, due to their unstable trajectories, collide with the quadrupole rods and are effectively removed (Figure 7). The quadrupole mass filters (Q) were described by Paul and Steinwedel in 1953,[92] and in 1989 Paul was awarded with the Nobel Prize for his work.[93] Quadrupole mass analysers are relatively cheap, but have comparatively a somewhat lower resolution (below 10000 FWHM) and are thus nowadays often employed as mass filters in hybrid instruments using additionally higher performance mass analyzers.

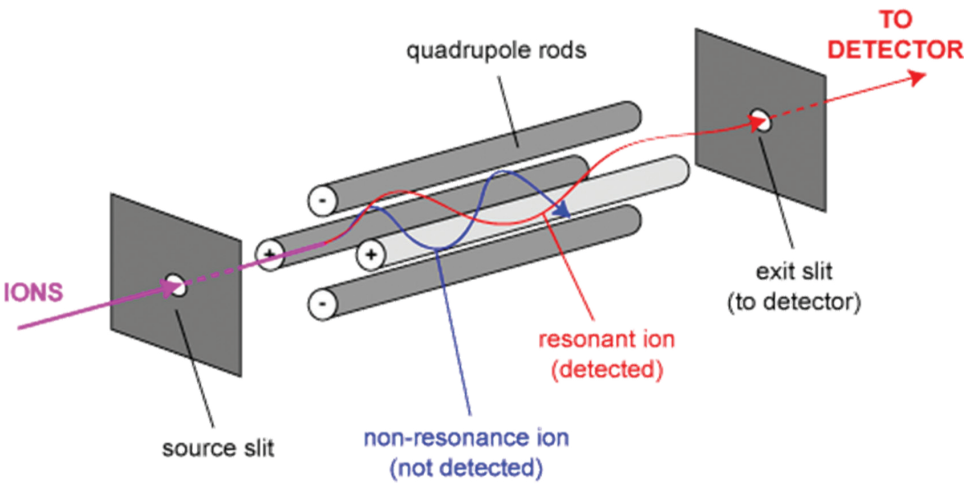


Figure 7. Schematic representation of a quadrupole mass analyser and the stable (resonant ion) and unstable trajectories of ions (non-resonance ion).

- *Quadrupolar and Linear Ion Trap analyzers*

Quadrupole ion traps, also called 3D traps or Paul ion traps consist of a circular electrode (ring electrode) and two hyperbolic caps/electrodes (end caps) (Figure 8a).[94] An oscillating electric field is applied to the ring electrode and the end-cap electrodes are kept at ground potential. Since there is no DC potential applied, the trap will allow ions of a wide m/z range to have stable trajectories in 3 dimensions, which effectively means the ions are trapped. Discrimination of m/z is performed using a changing amplitude of the RF field and resonance excitation.

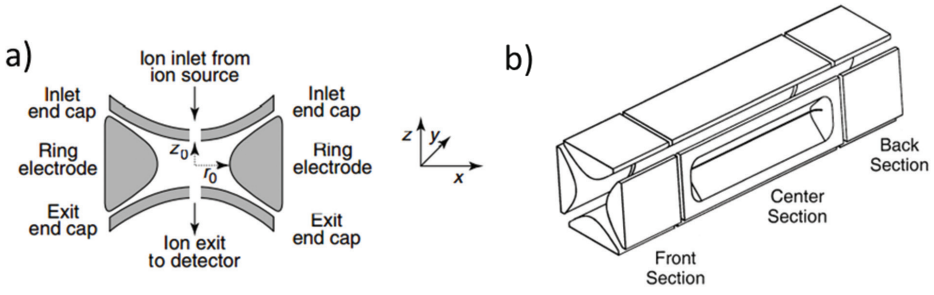


Figure 8. a) Schematic representation of a 3D ion trap. Adapted from Hoffmann *et al.*[95] b) Schematic representation of a linear ion trap. Adapted from Schwartz *et al.*[96]

A different type of an ion trap is the 2D-trap, also called linear ion trap (LIT).[96] An LIT consists of a quadrupole with two end electrodes, which can be built of two plates or short quadrupoles (Figure 8b). Ions are confined radially and a static electric potential on the end electrodes is used to confine the ions axially.[97] The design of 2D traps enables to increase the ion storage volume as compared to 3D traps. A benefit from an increased ion storage volume is a reduction of space charge effect between the trapped ions. Space charge effects limit the maximum load of an ion trap, due to repulsion of ions, which in consequence disrupts their stable trajectories.[97]

- *Time of flight analyzer*

The concept of time of flight (TOF) was first introduced by Stephens.[98] In this mass analyser ions are separated in a field free drift tube based on the different speed at which ions with distinct m/z values fly through it. The packet of ions, which are initially placed in the same location (ideally) and are at rest, is subsequently accelerated to the same kinetic energy using an electric filed. The speed of the ions, at a constant electric field, is inversely proportional to the square root of their m/z value. Since its introduction, TOF mass spectrometers have undergone a considerable evolution.[99,100] To overcome issues with different initial kinetic energies, an electrostatic mirror, called reflectron, which provided higher resolving power, was introduced.[101] The introduction of MALDI (as the laser has a pulsed behaviour rather than a continuous operation as we discussed it above),[86] and an orthogonal accelerator (this accelerator apart from allowing higher sensitivity and ion speed gave the possibility to couple the TOF to most continuous ion sources),[102] together with the revolution in digital electronics (due to the time-scale for measuring and processing signals),[103] have greatly influenced the resurgence of TOF technology in mass spectrometry.

- *Orbitrap analyzer*

Currently one of the most prominent technologies for mass spectrometry is the Orbitrap. Although Makarov was clearly the first to build a working version of an Orbitrap in 2000,[104] the principles behind Orbitrap trapping were already introduced by Kingdon in 1923.[105] The Orbitrap consists of a central spindle-like electrode surrounded by barrel-like electrode (Figure 9).[104,106] In this device neither

magnetic nor RF fields are applied, instead a combination of a static quadrupole and a logarithmic field is used. The ions combine a rotation trajectory around the central electrode with oscillations along the axial electrode in a harmonic orbit. Three frequencies, namely frequency of radial oscillations, rotation and axial, are present in this technology. However, only the axial frequency is completely independent of the energy and the position of the ions. The oscillating ions induce an image current that is detected with the help of a differential amplifier between the halves of the outer electrode. The signals are converted into frequencies using Fourier transformation. The mass spectrum can then be calculated by frequency-to-m/z conversion. High analytical performance in terms of resolution (>200,000 FMHW) and mass accuracy (< 5p.p.m. with external calibration, <2 p.p.m. with internal calibration) are achievable with this mass analyzer.

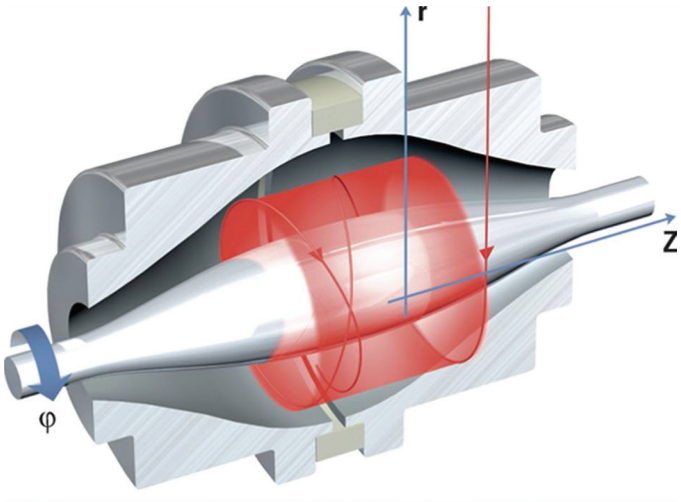


Figure 9. Schematic representation of an Orbitrap mass analyser. Adapted from Scigelova *et al.*[107]

- *Hybrid analyzers*

Nearly all modern mass spectrometers are 'hybrid' instruments in which multiple mass analysers are coupled in tandem. Hybrid instruments combining distinct mass analysers with an Orbitrap are widely used in proteomics. Among them, the Orbitrap was first coupled to a linear ion trap (LTQ).[108] In this instrument an ion trap was placed in front of the Orbitrap with an RF-only quadrupole in between.

In this way advantage of the speed and sensitivity of the ion trap and the resolution and mass accuracy of the Orbitrap was taken.[109] Further developments led to the introduction of later instruments version, such as the LTQ-Orbitrap XL,[110,111] the LTQ-Orbitrap Velos[112,113] and the Q-Exactive.[114] Two new improvements were done in the last years, which were included in the last generation of Orbitrap-based instruments. One of them is a compact, high field Orbitrap analyser which increases two-fold the resolving power.[115] The second improvement is a new method for processing transient detection signal called enhanced Fourier transform (eFT), which also increases two-fold the resolving power.[116] These two aspects were included in the LTQ-Orbitrap Elite,[115] Q Exactive HF,[117] Orbitrap Fusion and Orbitrap Fusion Lumos instruments (Figure 10).[118]

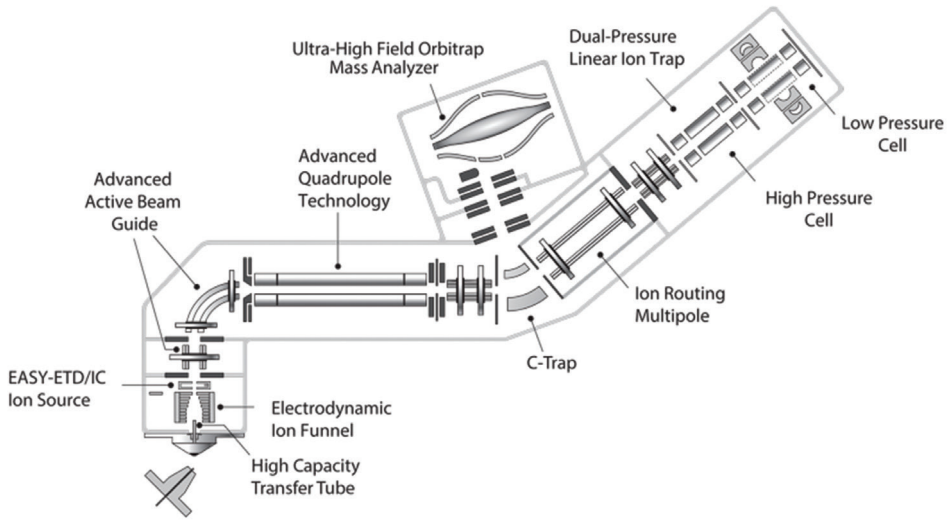


Figure 10. Schematic representation of the Orbitrap Fusion Lumos, the latest hybrid instrument developed around the Orbitrap technology.

Hybrid instruments combining other platforms are also popular in proteomics. A tandem mass spectrometer that combines quadrupole analyzers and a TOF mass analyser, the TripleTOF (Figure 11)[119], is also widely used.[120–122] In this system the quadrupoles act as mass filters, being able to selectively transmit ions to the TOF analyser. The TOF analyser enables fast detection allowing high MS/MS spectral acquisition rates (20 Hz). The speed of the instrument allows high number of identifications, increasing the depth of the research. Furthermore, a high resolving

power and mass accuracy are achievable with this hybrid platform.

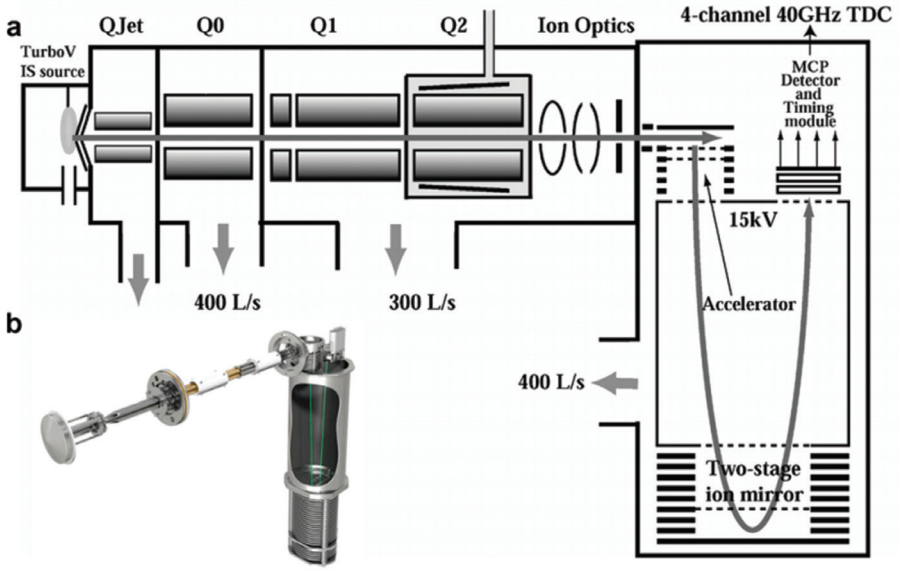


Figure 11. Schematic representation (a) and real image (b) of the TripleTOF instrument platform. Figure adapted from Andrews *et al.*[119]

Fragmentation techniques

Tandem mass spectrometry (MS/MS) is a key technique for peptide sequencing as the measurement of a peptide precursor ion m/z by MS is mostly not sufficient to identify its amino acid sequence. A subsequent analysis of the fragment ions obtained by fragmenting the peptide precursor ion is needed. Several fragmentation techniques exit and depending on the biological question and/or characteristics of the precursor ions the most suitable technique should be chosen in order to get the optimal results.

- **Collision-induced dissociation**

Collision-induced dissociation (CID), also referred to as collision-activated dissociation (CAD) is the most widely used fragmentation technique in proteomics research.[123] In CID ions are accelerated and subjected to multiple collisions with inert gas molecules (such as Helium, Nitrogen or Xenon). Subsequently the kinetic energy of the ion is transferred into internal vibrational modes, which ultimately

leads to peptide backbone fragmentation. The cleavage occurs predominantly at the peptide amide bond (CO-NH), resulting in a series of b- and y-fragment ions (Figure 12).[124]

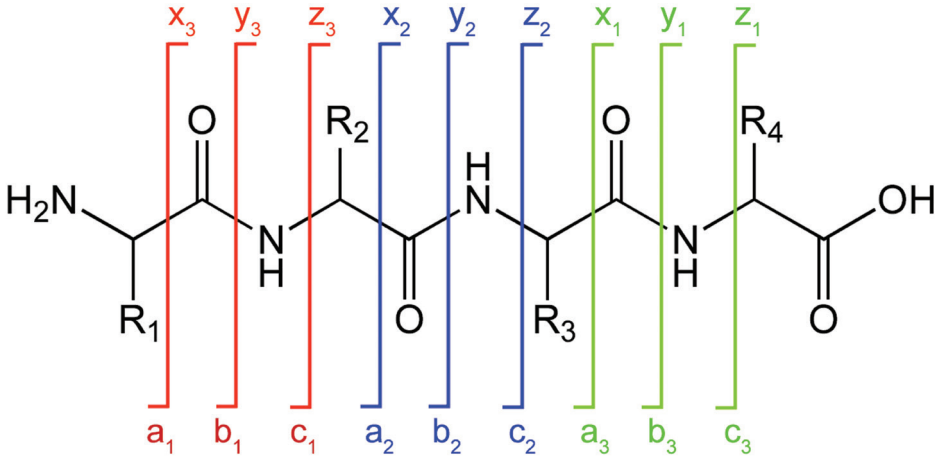


Figure 12. Nomenclature of peptide fragment ions as proposed by Roepstorff and Fohlman.[125]

CID processes can be separated primarily into three categories based on the translational energy of the precursor ion.[126] In quadrupole instruments the kinetic energy of the precursor is in the range of 1-200 eV and the fragmentation occurs mostly on the microsecond time scale. In ion trap instruments the kinetic energy of the precursor is generated using resonance and corresponds to a few eV. The timeframe of activation in ion trap CID is in the millisecond time scale. In contrast, high-energy collisions are in the keV range and are common in sector and TOF/TOF instruments. In low energy experiments (milli to microsecond timeframes) multiple collisions are employed to improve efficiency of fragmentation. The term higher-energy collisional dissociation (HCD) also refers to beam type CID and includes the generation of low molecular weight reporter ions.[110]

The observed dissociation has been rationalized by the 'mobile proton model'.[127,128] The model states that upon excitation the proton(s) migrate to different amide locations along the backbone of the peptide, as long as they are not sequestered by a basic amino acid side chain.[128] Migration to an amide nitrogen leads to weakening of the bond and combined with the energy provided by CID ultimately leads to

dissociation.

- **Electron Capture and Electron Transfer induced Dissociation**

Electron-capture dissociation (ECD) and electron-transfer induced dissociation (ETD) are relatively new fragmentation techniques in proteomics, introduced in 1998[129] and 2004[130], respectively. In ECD a heated filament source produces low-energy electrons. These electrons and the positively charged peptide ions are forced to overlap. Under such conditions an electron can be captured by the charged peptide, which induces backbone fragmentation at the N-Cα bond and generates c- and z-fragment ions. In ETD negatively charged species are generated by chemical ionization. A common chemical used as the electron transfer reagent is fluoranthene.[131] When the electron transfer reagent enters 'in contact' with the positively charged peptides an electron is transferred, which leads to peptide fragmentation in a similar manner to ECD. Therefore, the predominant ions found in ETD spectra are also c- and z-fragment ions. Both are electron-driven reactions and provide extensive fragmentation. One big difference between ECD and ETD is that ECD is primarily used in Fourier transform ion cyclotron resonance mass spectrometers while ETD enables the use of ion traps for electron-based dissociation, making ETD the most widely used electron-driven dissociation method in proteomics.

In both electron-based methods, the precursor ion charge is an important factor as the electron capture cross section is proportional to the square of the charge state of the precursor.[132] This explains the more efficient fragmentation found for higher charged peptides and the poor dissociation of double charged peptides. In order to increase the dissociation efficiency one can activate the non-dissociative electron transfer products with supplemental collisional activation.[133,131] However, In ETD the unreacted precursor remains relatively highly abundant. Recently a novel scheme combining electron-transfer and higher-energy collision dissociation (EThcD), which overcomes this limitation and increases peptide sequence coverage was introduced.[134] In EThcD after an initial ETD step, all ions are subjected to HCD fragmentation, which yields thus b/y- and c/z-type fragment ions in a single spectrum. The outstanding performance of EThcD for PTM studies, in particular for phosphosite localization has already been demonstrated.[135] The promising applicability of this new fragmentation technique for middle-down range peptides,

which involve highly charged peptides, is further described in the fourth chapter of this thesis.

Database search strategies and statistics

Mass spectrometry based proteomics, as other high-throughput-omics technologies, needs a reliable and fast analysis of the large amount of experimental data obtained. Once the MS/MS spectra are acquired the identification of the peptides is computationally achieved. Nowadays there are several strategies for peptide identification,[136] however a database search algorithm is the most commonly used. Experimentally acquired fragmentation spectra are compared against theoretically constructed spectra based on the genome derived protein databases.[137] For each MS/MS spectrum, software is used to determine which peptide sequence in a database of protein sequences generates the best match. Besides the MS/MS spectrum additional parameters are required by the software for faster and more reliable peptide identification. Among those parameters are the sequence database, the taxonomy, the enzyme used, the fixed and variable modifications, the mass tolerance and type of mass spectrometer used. Each gene/protein entry in the database is digested in silico, using the given parameters. If the calculated mass of a peptide matches that of an observed peptide, the masses of the expected fragment ions are calculated and compared with the experimental values.[138] Many scoring algorithms have been developed to decide which peptide sequence best matches a given spectrum and can be classified in four basic approaches: descriptive, interpretative, stochastic and probability-based.[139] Among all the available algorithms the most widely used, and the ones applied in the different chapters of this thesis, are Mascot[140] and Sequest.[141] Mascot is a probability-based algorithm and calculates the probability whether an observed match between experimental data and peptide sequences found in a reference database has occurred by chance (Figure 13). The match with the lowest probability of occurring by chance is returned as the most significant match. The significance of the match depends on the size of the database that is being queried. Sequest uses a descriptive model for peptide fragmentation and correlative matching to a tandem mass spectrum. It uses a two-tiered scoring scheme to assess the quality of the match: the preliminary score (an empirically derived score that restricts the number of sequences analyzed

in the correlation analysis) and the XCorr (a cross-correlation score between the experimental and the theoretical spectra).

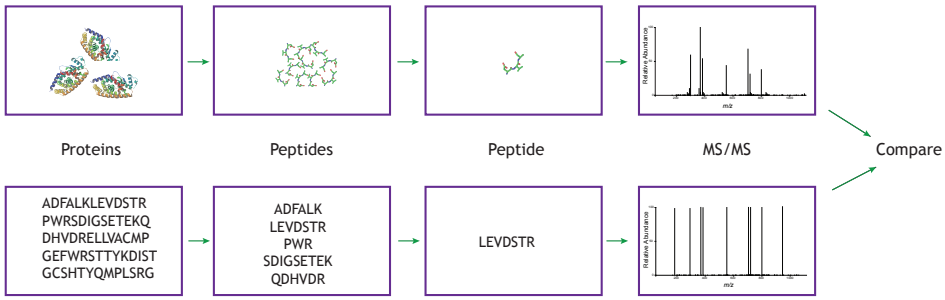


Figure 13. Schematic concept of MS/MS database searching.

In addition to these calculated peptide scores either by Mascot or Sequest, the estimation of a False Discovery Rate (FDR) is the usual way to measure the quality of a set of search results.[142] The FDR converts the scores into probabilities by searching them against a decoy database (which contains reversed, randomized or shuffled sequences) and is estimated as the ratio of decoy database matches to target database matches.

Alternatively, to database search approaches, de novo sequencing allows identification of proteins through direct interpretation of the MS/MS spectra in a database-independent manner.[143–145] In this approach the reconstruction of the original peptide sequence is done without knowledge of the genomic sequences or even the organism from which the sample was taken. A main advantage of this approach is the possibility of identifying novel peptides as there is no match of spectra to already exited sequences in a database as in the previously described algorithms.

Quantitative proteomics

Quantitative information is required to study the highly dynamic nature of the proteome; however, MS-based proteomics is as such not a quantitative method. The main reason is that protein and peptide ionization is a rather complex process in which analytes compete to take the available charges. Therefore, differences in

the physical and chemical properties of the generated peptides (such as charge state, length, amino acid composition, PTMs) lead to differences in ion intensities for the peptides, even when they belong to the same protein. Thus, quantitative methods require comparative experiment designs. These can be broadly classified in two forms, absolute quantification (measuring the absolute amount of the protein in the sample) or relative quantification (measuring the relative change in protein amount between two states).[146] Relative quantification can be achieved either by employing differential stable isotope labeling or by label-free approaches (Figure 14). Labeled peptides are chemically identical to their native counterpart, only differing in mass, and therefore the two peptides also behave identically during chromatographic and mass spectrometric analysis. Given that a mass spectrometer can recognize the mass difference between the labeled and unlabeled forms of a peptide, quantification is achieved by comparing their respective signal intensities.[147] The label can be introduced at various steps during sample preparation: it can be introduced into proteins or peptides metabolically, chemically or enzymatically, or provided by spiked synthetic peptide standards. The limitation of protein quantification in complex samples by stable isotope based methods mainly lies in signal interferences caused by co-eluting components of similar mass. Therefore, the most efficient way to optimize the quantitative analysis is to decrease sample complexity by increasing chromatographic gradient times or by including a pre-fractionation step prior to LC-MS/MS analysis (as discussed it earlier).

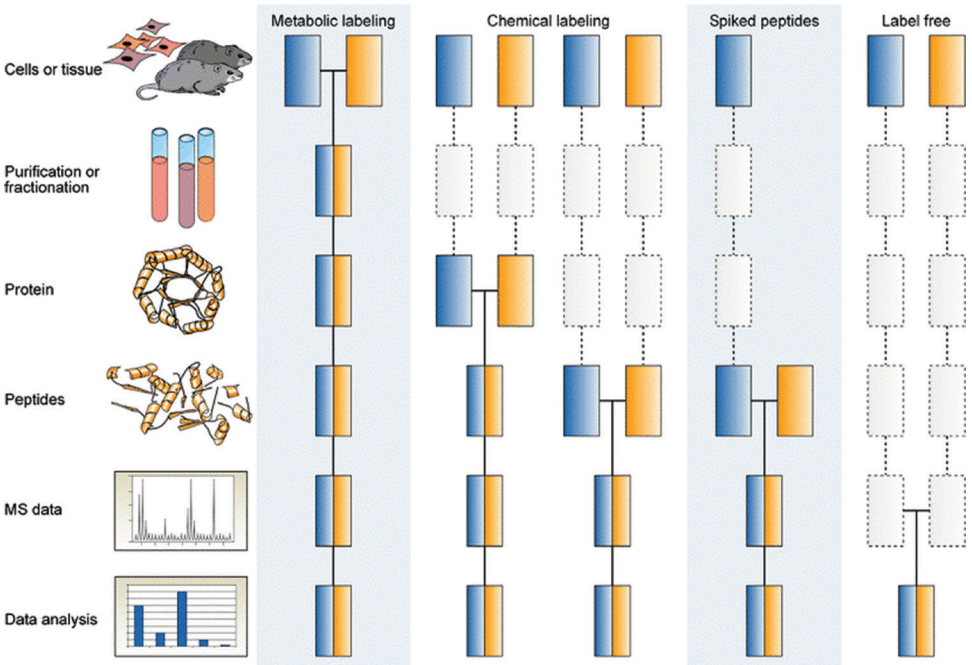


Figure 14. Common schemes for quantitative mass spectrometry. Boxes in blue and yellow represent two to be compared experimental conditions. Dashed lines indicate presence of experimental variability. Adapted from Bantscheff *et al.*[147]

- **Metabolic labeling**

The earliest possible point for introducing a stable isotope label is by metabolic labeling during cell growth and division through the growth medium, which is consequently incorporated into proteins and peptides. This can be applied not only to cultured cells but also to complex organisms, such as plants or animals. Metabolic labeling was first achieved with ^{15}N-enriched cell culture medium,[148] but it was established as a high-throughput approach by using stable-isotope-labeled essential amino acids in mammalian cell cultures (SILAC).[149,150] The cell culture medium contains $^{13}C_6$-arginine and $^{13}C_6$-lysine, which ensures that all tryptic cleavage products of a protein (except for the very C-terminal peptide) carry at least one labeled amino acid resulting in a constant mass increment over the non-labeled counterpart. Protein identification is based on fragmentation spectra of at least one of the co-eluting 'heavy' and 'light' peptides and relative quantitation is performed by comparing the intensities of isotope clusters of the intact peptide in

the survey spectrum. An accurate quantification method is obtained thanks to the early combination of the samples, which minimizes experimental error or bias in the workflow.

- **Chemical labeling**

In the case of chemical labeling, the label is introduced to proteins or peptides through a chemical reaction, which increases the applicability of this approach compared to metabolic labeling. Chemical labeling methods include isotope-coded affinity tag (ICAT),[151,152] isobaric tag for relative and absolute quantification (iTRAQ),[153] tandem mass tags (TMT),[154] dimethyl labeling[155,156] and ^{18}O labeling.[157] Dimethyl labeling is a widely applied approach for protein quantification in proteomics. In this labeling approach all primary amines (the N terminus and the side chain of lysine residues) in a peptide mixture are converted to dimethylamines. By using combinations of several isotopomers of formaldehyde and cyanoborohydride (^2D and ^{13}C), peptide triplets can be obtained that differ in mass by a minimum of 4 Da between samples (Figure 15a). These reagents have the advantage of being inexpensive and the reaction can be applied to any type and amount of sample in a simple and fast way. The labeled samples are mixed and simultaneously analyzed by LC-MS/MS and the mass difference of the dimethyl labels is used to compare the peptide (Figure 15b). The labeling can be performed in-solution, online with LC-MS and on-column using solid phase extraction columns.[158] A drawback of this approach, and chemical labeling in general, is that different samples can be mixed only after enzymatic digestion and labeling have been performed separately, ultimately leading to higher variability compared to metabolic labeling.

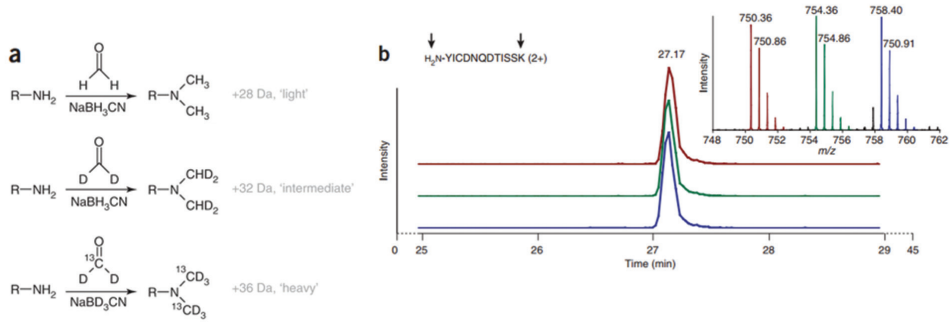


Figure 15. Dimethyl labeling approach for protein quantification. a) Labeling scheme of triplex stable isotope dimethyl labeling. b) Extracted ion chromatograms and mass spectra of a BSA peptide. Adapted from Boersema *et al.*[158]

- **Spiked synthetic peptide standards**

The use of synthetic peptides was described in the early 80s,[159] and years later the use of reference peptides that contain heavy isotope labels and are further chemically identical to the original peptides (the so-called AQUA peptides) [160] became broadly applied. The main advantage of the use of spiked peptides is that peptide concentrations are accurately determined. Peptides are synthesized with incorporated stable isotopes as ideal internal standards to mimic native peptides formed by proteolysis, which can also be prepared with covalent modifications (i.e. phosphorylation, methylation, acetylation) that are chemically identical to naturally occurring PTMs. Such AQUA internal standard peptides are added to the extract after enzymatic digest. Subsequent comparison of the mass spectrometric response to the endogenous peptide in the sample by multiple reaction monitoring (MRM) allows quantitatively measurement of the absolute levels of proteins and post-translationally modified proteins.

- **Label free quantification**

The two most widely applied label-free quantitative methods are spectral counting and peptide peak intensity measurements. Spectral counting is based on the observation that more abundant proteins will produce more MS/MS spectra than less abundant proteins, and abundant peptides are sampled more often in fragment ion scans than low abundance peptides. Relative protein quantitation by spectral counting involves the comparison of the number of fragment spectra identifying peptides of a given protein to estimate their relative protein abundance.[161] On the other hand, relative quantitation using peptide peak intensity measurements involves comparing the MS peptide ion intensities belonging to a given protein across multiple LC-MS runs where each run represents an experiment.[162] The ion chromatograms for every peptide are extracted from an LC-MS run, and their peak areas are integrated over the chromatographic time scale. This approach requires correlation of retention time with m/z ion features and charge state to avoid discrepancy in matching common ions detected in each run.

These strategies are easy to perform and cost-effective. Furthermore, an unlimited numbers of samples can be compared and they correlate well with protein abundances in complex biological samples.[163] However, the application

of these methods requires, apart from a robust LC-MS system, robust computing power and advanced algorithms. The algorithms need to be capable of handling chromatographic peak alignment, spectral counting and peptide ion intensity measurements in order to calculate changes in protein abundances in complex biological samples.[164]

Part II. Intestinal stem cell biology and disease

Adult stem cells

1

The human body consist of many different organs that are each build up by numerous different specialized cells. At the origin of nearly every organ of the human body are specialized cells, from where the (re)generation of the organ is orchestrated. These cells have been termed Adult Stem Cells. Adult stem cells, regardless of their origin, have two unique properties:[165] their capacity to self-renew to produce more stem cells and to differentiate into diverse specialized cell types. When an adult stem cell divides, the new cell can become a different cell with a more specific function or it can remain a stem cell. Due to their stage of development, adult stem cells, have more limited differentiation potential compared to embryonic stem cells, they are multi-potent. Multipotency describes progenitor cells, which are able to give rise to other types of cells albeit within a limited range. Most adult stem cells are lineage-restricted and play important roles on local tissue repair and regeneration,[166] which makes them an important research subject.

▪ Anatomy of the adult intestine

The primary function of the intestinal tract is the digestion and absorption of nutrients and fluids.[167] The intestinal tract is essentially a tube lined with a specialized simple and truly multitasking epithelium, a single-cell layer termed the mucosa. This tissue, apart from efficiently performing the digestion and absorption of nutrients, must maintain an effective barrier against potentially harmful organisms and substances present in the intestinal lumen.[168] Several differentiated cell types mediate the functions of the intestinal epithelium:[169] Enterocytes, Enteroendocrine cells, Goblet cells, Tuft cells and Paneth cells. Each cell type has a specific function. Enterocytes are highly polarized cells responsible for absorbing and transporting nutrients across the epithelium.[167] Enteroendocrine cells coordinate gut functioning through specific peptide hormone secretion. Goblet cells secrete protective mucins and trefoil proteins that are required for the movement and effective expulsion of gut contents, and in addition they provide protection against shear stress and chemical damage. Tuft cells have recently been described to be essential in the initiation of immune responses.[170] Paneth cells contribute to innate

immunity by producing antimicrobial peptides that regulate the gut microbiota.[171] Furthermore, the epithelium contains at least two other cells types with poorly defined functions: cup cells[172] and Peyer's patch-associated microfold cells.[173] The relative abundance of the differentiated cell types varies markedly within distinct segments of the intestine in order to adapt to the specific function of each segment. Indeed, the intestinal tract is functionally and anatomically divided into two well-defined segments: the small intestine and the large intestine, or colon. The small intestine can be subdivided into duodenum, jejunum and ileum. The epithelium of the small intestine is organized into crypts and villi.[168] Crypts, also known as crypts of Lieberkühn (as Lieberkühn was the first to describe the structure and function of these crypts in the 18th century), are invaginations into the underlying connective tissue. In contrast, villi are finger-like protrusions pointing toward the lumen, which dramatically increase the absorptive surface area of the small intestine (Figure 16a). The situation in the colon resembles the small intestine as the large intestine is also folded into crypts. However, the colon has a flat surface epithelium characterized by a high density of goblet cells in order to facilitate the passage of stool and the absence of Paneth cells. Deep secretory cells are suggested to supply a niche function in the colon[174] (Figure 16b).

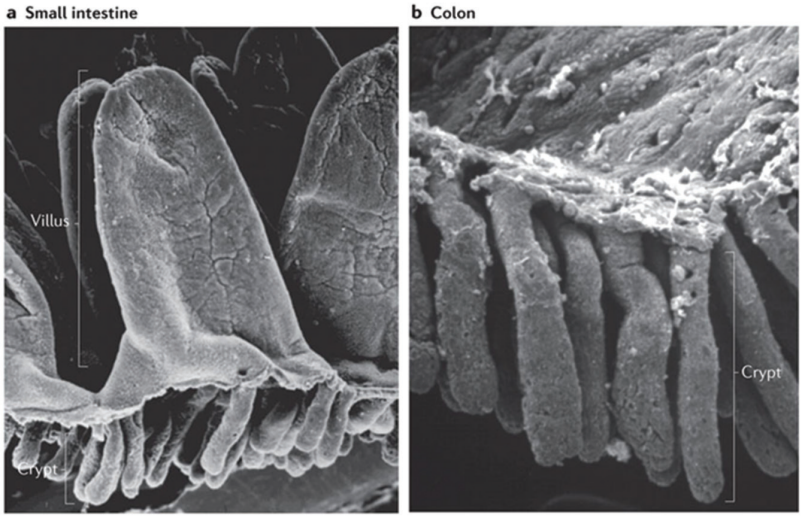


Figure 16. Intestinal epithelium. Scanning electron micrographs of the small intestine (a) and colon (b) in which the structural organization of these organs can be observed. The small intestine is organized into crypts and villus whereas the colon is also folded into crypts but carries no villus. Adapted from Magney *et al.*[175]

- **The intestinal stem cells**

The persistent aggression from potentially lethal organisms and carcinogens present in the intestinal lumen induces a remarkably high rate of cell death,[176,177] which requires a vigorous self-renewing process to maintain optimal function. In fact, the intestine is the most rapidly self-renewing tissue in mammals and thus it has long been assumed that the crypts must harbor functional stem cells.[178] To definitively identify stem cells, one needs to be able to identify and/or mark those cells. Consequently, intestinal stem cells had long remained elusive hampered by the lack of unique molecular markers. Two schools of thoughts existed regarding the exact identity and location of intestinal stem cells (Figure 17). On the one hand, the "stem cell zone" model defined by Cheng and Leblond[176] suggested the crypt-base columnar cells (CBC) as the precursor cells that generate the differentiated cell types of the epithelium of the small intestine. These CBC cells are present in the bottom of the crypts, wedged between the Paneth cells (Figure 17a). On the other hand, the "+4 position" model described by Potten and colleagues[179] proposed that stem cells reside at position +4 relative to the crypt bottom, directly above the Paneth cells. These +4 cells were identified as DNA label-retaining cells (Figure 17b).

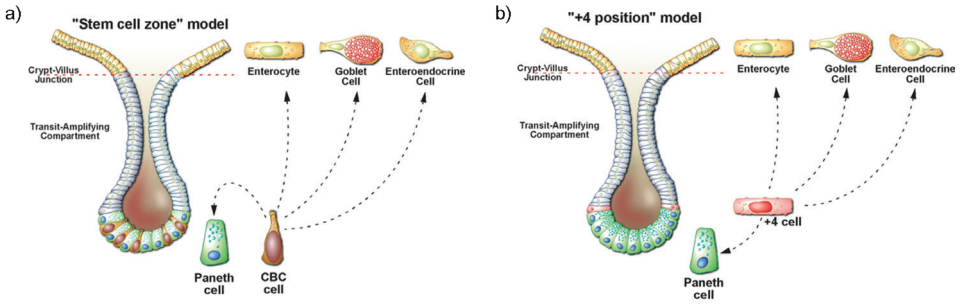


Figure 17. Two opposing models defining the exact identity of the intestinal stem cells. a) The "stem cell zone" model was proposed in the early 1970s and states that small, undifferentiated, cycling cells intermingled with the Paneth cells are likely to be the true intestinal stem cells. b) In the"+4 position" model proposed in the late 1950s, it was assumed that the crypt base is exclusively populated by terminally differentiated Paneth cells and the stem cells must therefore be located just above the Paneth cells at the +4 position. Figure adapted from Barker *et al.*[166]

The breakthrough for the definitive identification of intestinal stem cells was the discovery of a specific marker for the CBC cells.[180] Wnt signals are the driving forces behind the proliferative activity of the intestinal epithelium both in normal physiology and in colorectal cancer (CRC).[181–183] It was hypothesized that certain Wnt target genes may be specifically expressed in stem cells.[184,185] Indeed a subset of these target genes was uniquely expressed in a limited number of cells in the bottom of the crypts, more specifically by the CBC cells (as opposed to Paneth cells which are also localized to crypt bottoms). The gene Lgr5 was one of these uniquely expressed in CBC cells. Lgr5 stands for leucine-rich-repeat-containing G-protein-coupled receptor 5, also known as Grp49.[186] Lgr5 expression is restricted to the CBC cells, defined by Cheng and Leblond,[176] that are squeezed in between the Paneth cells. Using genetic lineage tracing it was demonstrated that $Lgr5^{+ve}$ cells generate all the differentiated cell types of the epithelium of the small intestine.[187] By morphology, the $Lgr5^{+ve}$ cells are readily distinguishable from the adjacent Paneth cells,[166] and fulfill the definition of an adult stem cell in displaying longevity and multipotency in the small intestine and in the colon. Interestingly, further studies have demonstrated that when $Lgr5^{+ve}$ stem cells are damaged, the +4 cells suggested by Potten *et al.*[179] to be the stem cells of the intestine, revert into cycling $Lgr5^{+ve}$ stem cells.[188,189] These findings manifest the plasticity of the stem and progenitor cells and unify earlier theories about the identity of crypt stem cells.[190]

Finding an intestinal stem cell marker greatly improved the understanding of epithelial stem cells biology[191] and the unique epithelial anatomy and homeostasis of the small intestine and colon (Figure 18).[168] The stem cells of the intestine and their progeny, the transit-amplifying (TA) cells reside at the crypts base. TA cells spend approximately 2 days in the crypts, in which they divide 4-5 times of unusually short duration, 12-16h, generating some 300 cells per crypt every day.[192] When the TA cells reach the crypt-villi junction, they rapidly and irreversibly differentiate into the specialized intestinal epithelial cell types. These differentiated cells continue to migrate upwards towards the tip of the villi, where after 2-3 more days they undergo apoptosis and are shed into the gut lumen. Thus, the epithelium is in a continuous upward movement and only Paneth cells escape from this flow. Paneth cells also derive from stem cells but persist for 3-6 weeks at the crypt base, located in between the stem cells.[193]

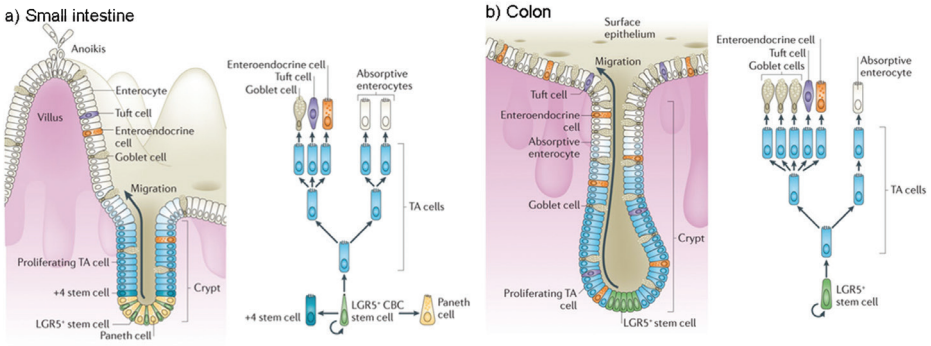


Figure 18. Epithelial self-renewal in the intestinal epithelium. In the intestine the stem cells are present in the bottom of the crypts. They continuously generate transit-amplifying cells, which differentiate into the different cells present in the intestine. The difference in the abundance of the differentiated cells is also presented in this figure. In the small intestine (a) the stem cells are intercalated with Paneth cells while in the colon (b) no Paneth cells are present and deep secretory cells are suggested to supply the niche function. Figure adapted from Barker.[168]

- **Impact of the intestinal stem cell identification**

The ability to identify populations of Lgr5^{+ve} stem cells from the adult intestine has opened up exciting opportunities to exploit the clinical potential of these adult stem cells. Adult stem cells have the potential to revolutionize regenerative medicine with their unique capabilities to self-renew and differentiate into various phenotypes. Furthermore, with the Lgr5-based genetic tools in hand it is also possible to isolate and genetically modify adult intestine stem cells on demand and study their role in tissue homeostasis and disease. Intestinal stem cells renew the epithelium that lines the intestine and it is known that alterations in the biology of these stem cells can lead to various disorders, including cancer.[194] For instance, some studies revealed that deletion of APC (a negative regulator of the Wnt signaling pathway)[195] in intestinal stem cells transforms the cells and leads to adenoma formation.[196] Current clinical models, such as cell lines and animal models, have been used for many years to investigate the mechanisms of diseases and design novel treatment strategies. However, they fail to reflect important features of tumor cells (such as loss of tumor heterogeneity in cell lines and drift in the stromal components from human to mouse in animal models),[197] making the interpretation of the genomic changes of cancer in terms of prognosis, drug response or patient outcome difficult to extract from such studies.

Organoids

Long-term ex vivo culture methods for adult stem cells provide possibly a more suitable model to understand cancer biology; however, faithful recapitulation of the complex stem cell niche in a culture dish is a challenging task. Several ex vivo culture strategies had been developed without successful establishment of a long-term culture that can recapitulate the morphology of the adult epithelium.[198–200] A better understanding of the conditions to maintain these adult stem cells,[201–203] in combination with the work in the field of three-dimensional cultures,[204,205] has allowed the establishment of a new preclinical model, termed organoids by Clevers *et al.*[206] Organoids are ever-expanding three-dimensional epithelial structures with all the hallmarks of in vivo epithelial tissue. Single crypts isolated from the small intestine or single sorted Lgr5^{+ve} stem cells were shown to be capable of generating villus-like epithelial domains. These organoids contain all the differentiated cell types and have a remarkably similar architecture and composition comparable to normal gut epithelium. Importantly, the homeostasis of these organoids faithfully recapitulates the in vivo situation: cells are continuously generated at the bottom of the crypt-like compartments, proliferate, differentiate and are shed into the central cyst lumen approximately 5 days later (Figure 19).

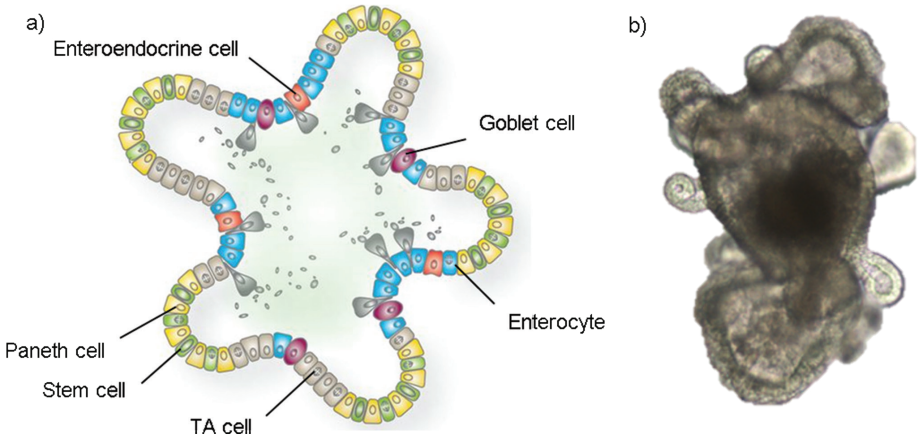


Figure 19. Intestinal organoids. a) Schematic representation of a small intestinal organoid, consisting of a central lumen lined by villus-like epithelium and associated crypt compartments. Adapted from Leushacke *et al.*[207] b) Successfully growing organoid derived from a single Lgr5^{+ve} cell after 13 days of culture. Adapted from Sato *et al.*[206]

Specific growth requirements are needed to derive small intestine organoids from mouse intestinal epithelium: Wnt signaling is an essential driving force of intestinal self-renewal. Inhibition of Wnt signaling in the intestine by either knocking out TCF4[181] or over expressing Dickopf (a secreted inhibitor of Wnt signaling),[208,209] blocks self-renewal. R-spondin is an essential factor to activate Wnt signaling,[210,211] and systemic expression of R-spondin induces strong hyperproliferation of the intestinal epithelium.[208] EGF signaling is also associated with intestinal proliferation[212] whereas transgenic expression of Noggin induces ectopic crypt formation along the crypt villus axis. Furthermore, these pathways are also found to be key players in malignant transformation of the intestinal epithelium. Over 90% of CRC carry Wnt pathway mutations, 50% RAS mutations (EGF signaling) and SMAD4 mutations (TGF-β). In other words, signaling pathways that are essential for normal development and renewal are also key pathways in cancer development. Finally, as isolated intestinal cells undergo anoikis outside the normal tissue context[213] and laminin is enriched at the crypt base,[214] Matrigel (a laminin and collagen rich matrix)[215] is used to support intestinal epithelial growth. Organoids generated with these essential cocktails of media can be propagated for a minimum of 1.5 years.[216] They can be passaged, expanded and frozen and can be genetically manipulated via transfection of DNA or siRNA, or infection with an adeno- or lentivirus.[217–220]

Thanks to the faithful recapitulation of the homeostasis and architecture of the functional intestinal epithelium multiple applications of the organoid technology, such as developmental and stem cell biology[221,222] as well as disease modelling[223–225], are feasible. To accurately model human intestinal diseases, such as CRC, extensive efforts have been made to adapt the organoid culture technology for the ex vivo growth of human large intestinal tissue as it requires some crucial modifications.[194,226] Recently, a detailed investigation into genotype-to-phenotype correlations has been reported using several patient-derived organoids.[225] The feasibility of organoids for high-throughput drug screening and the possibility of using patient-derived organoids for personalized therapy design was demonstrate. These organoids provide a potentially unlimited supply of well-characterized patient material, circumventing some of the limitations of current models (i.e. lack of genetically stable cell lines and need of extensive colonies of animals) (Figure 20).

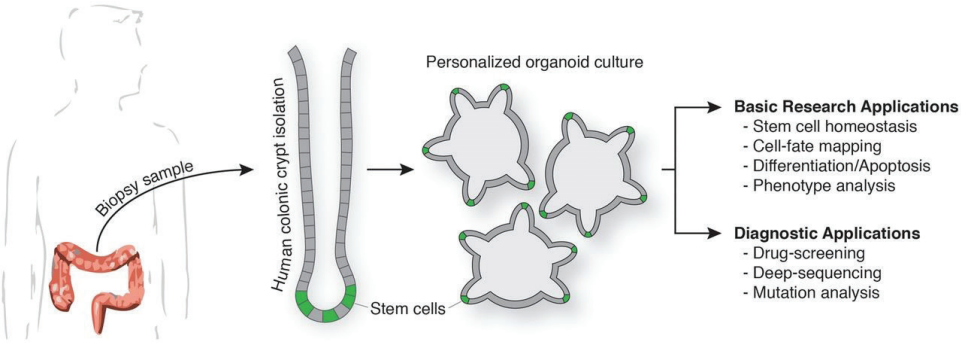


Figure 20. Schematic representation of organoid applications in basic and applied science. Human intestinal organoids have recently been generated from patient biopsies. Personalised organoid cultures of the human intestinal epithelium open up several basic research opportunities and translational applications. Adapted from Leushacke *et al.*[207]

Colorectal Cancer

CRC is one of the leading causes of death for adults, being the third most common cancer type in women and the second in men within Europe. Since the 1980s the incidence has decreased largely due to changes in risk factors and the introduction of colorectal cancer screening.[227,228]

Although Wnt signaling is very important in the normal physiology of the intestine, it was first recognized for its association with CRC. CRC usually develops from a benign precursor lesion, called the adenoma, which is visible on the mucosal surface of the colon. Through a combination of mutational activation of oncogenes and mutational inactivation of tumor suppressor genes, the adenoma gradually progresses into an advanced adenoma and subsequently into an invasive cancer. The activation of the Wnt signaling pathway,[229] by means of mutations in the APC gene[230] or β-catenin,[231] is regarded as the initiating event in colorectal cancer. As a common result, β-catenin accumulates in the nucleus and constitutively binds to the TCF/LEF transcription factors. TCF4 is physiologically expressed in all epithelial cells of the intestine.[232] The inappropriate formation of β-catenin/TCF4 complexes results in transcriptional activation of WNT/TCF4 target genes, initiating the transformation of intestinal epithelial cells into cancer cells.[231] The following steps include mutations of the KRAS[233] and PI3K oncogenes.[234] These events are continued by the mutational inactivation of the TGF-β response[235] and

the inactivation of the p53 pathway by mutation of the TP53 suppressor gene.[236,237] Although the alterations often occur in this order, the accumulation of the changes is likely more important than the order of occurrence. This sequence is known as the adenoma-adenocarcinoma sequence and it takes approximately 10 years for an adenoma to develop into a malignant tumor (Figure 21).[238]

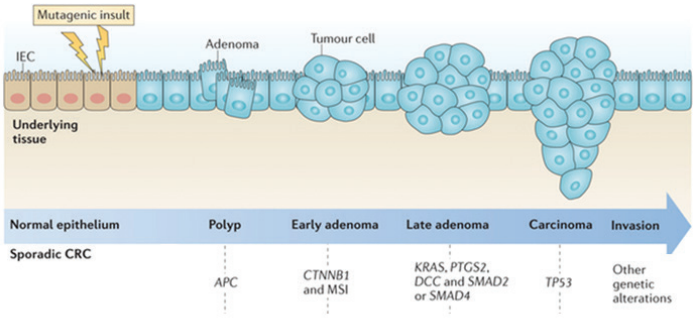


Figure 21. Adenoma-adenocarcinoma sequence first proposed in the 1980s. The transformation of normal colorectal epithelium into an adenoma and ultimately into an invasive and metastatic tumor is described. Adapted from West *et al.*[239]

There are several factors that contribute to the development of CRC, which can be divided into either dietary and lifestyle factors or inherited mutations. Around 20-25% of the cases correspond to the later, while most of the cases, 75-80%, are due to dietary and lifestyle factors.[240,241] However, the mechanisms underlying CRC development appear to be complex and heterogeneous and may entail quite a few patient-specific features.[242,243] To reach new and better treatments, the molecular mechanisms behind colorectal cancer and more generally the intestine need to be thoroughly investigated. Intestinal organoid culture enables to expand normal and tumor epithelial cells in vitro retaining their stem cell self-renewal and multiple differentiation and allows studying the genetic events involved in CRC.

References

1. Nelson, D. L. & Cox, M. M. Lehninger. Principles of Biochemistry. (2012).
2. Hutchison, C. A. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. 35, 6227–6237 (2007).
3. Metzker, M. L. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46 (2010).
4. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12, 87–98 (2011).
5. Wasinger, V. C. *et al.* Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis 16, 1090–4 (1995).
6. Wilkins, M. R. *et al.* From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. Bio/Technology 14, 61–65 (1996).
7. James, P. Protein identification in the post-genome era: the rapid rise of proteomics. Q. Rev. Biophys. 30, 279–331 (1997).
8. Godovac-Zimmermann, J. & Brown, L. R. Perspectives for mass spectrometry and functional proteomics. Mass Spectrom. Rev. 20, 1–57 (2001).
9. Wilkins, M. R. *et al.* Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It. Biotechnol. Genet. Eng. Rev. 13, 19–50 (1996).
10. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. Nat. Biotechnol. 21, 255–261 (2003).
11. Walsh, C. T. Posttranslational Modification of Proteins: Expanding Nature's Inventory. (Roberts and Company, 2006).
12. Nørregaard Jensen, O. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr. Opin. Chem. Biol. 8, 33–41 (2004).
13. Cohen, P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. Eur. J. Biochem. 268, 5001–10 (2001).
14. Stowell, S. R., Ju, T. & Cummings, R. D. Protein glycosylation in cancer. Annu. Rev. Pathol. 10, 473–510 (2015).
15. Altelaar, A. F. M. & Heck, A. J. R. Trends in ultrasensitive proteomics. Curr. Opin. Chem. Biol. 16, 206–13 (2012).
16. Wiśniewski, J. R. *et al.* Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. Mol. Syst. Biol. 8, 611 (2012).
17. Waanders, L. F. *et al.* Quantitative proteomic analysis of single pancreatic islets. Proc. Natl. Acad. Sci. U. S. A. 106, 18902–7 (2009).
18. Di Palma, S. *et al.* Highly sensitive proteome analysis of FACS-sorted adult colon stem cells. J. Proteome Res. 10, 3814–9 (2011).
19. Mayne, J. *et al.* Bottom-Up Proteomics (2013–2015): Keeping up in the Era of Systems Biology. Anal. Chem. 88, 95–121 (2016).
20. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. 113, 2343–2394 (2013).
21. Chait, B. T. Mass Spectrometry: Bottom-Up or Top-Down? Science (80-.). 314, (2006).
22. Reid, G. E. & McLuckey, S. A. 'Top down' protein characterization via tandem mass spectrometry. J. Mass Spectrom. 37, 663–675 (2002).

23. Smith, L. M., Kelleher, N. L. & Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. Nat. Methods 10, 186–7 (2013).

24. Siuti, N. & Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. Nat. Methods 4, 817–21 (2007).

25. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature 480, 254–8 (2011).

26. Forbes, A. J., Mazur, M. T., Patel, H. M., Walsh, C. T. & Kelleher, N. L. Toward efficient analysis of >70 kDa proteins with 100% sequence coverage. Proteomics 1, 927–33 (2001).

27. Wu, S., Kim, J., Hancock, W. S. & Karger, B. Extended Range Proteomic Analysis (ERPA): a new and sensitive LC-MS platform for high sequence coverage of complex proteins with extensive post-translational modifications-comprehensive analysis of beta-casein and epidermal growth factor receptor (EGFR). J. Proteome Res. 4, 1155–1170 (2005).

28. Garcia, B. A. What does the future hold for Top Down mass spectrometry? J. Am. Soc. Mass Spectrom. 21, 193–202 (2010).

29. Wu, S.-L. *et al.* Dynamic profiling of the post-translational modifications and interaction partners of epidermal growth factor receptor signaling after stimulation by epidermal growth factor using Extended Range Proteomic Analysis (ERPA). Mol. Cell. Proteomics 5, 1610–27 (2006).

30. Zhang, J., Wu, S.-L., Kim, J. & Karger, B. L. Ultratrace liquid chromatography/mass spectrometry analysis of large peptides with post-translational modifications using narrow-bore poly(styrene-divinylbenzene) monolithic columns and extended range proteomic analysis. J. Chromatogr. A 1154, 295–307 (2007).

31. Wu, C. *et al.* A protease for 'middle-down' proteomics. Nat. Methods 9, 822–4 (2012).

32. Laskay, Ü. A., Lobas, A. A., Srzentić, 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. 12, 5558–69 (2013).

33. Tswett, M. Physikalish-chemische studien uber das chlorophyll. die adsorptionen. Ber. Dtsch. Bot. Ges. 24, 316–323 (1906).

34. Chen, X. & Ge, Y. Ultrahigh pressure fast size exclusion chromatography for top-down proteomics. Proteomics 13, 2563–6 (2013).

35. Bouvier, E. S. P. & Koza, S. M. Advances in size-exclusion separations of proteins and polymers by UHPLC. TrAC Trends Anal. Chem. 63, 85–94 (2014).

36. Martin, A. J. & Synge, R. L. A new form of chromatogram employing two liquid phases: A theory of chromatography. 2. Application to the micro-determination of the higher monoamino-acids in proteins. Biochem. J. 35, 1358–68 (1941).

37. van Deemter, J. J., Zuiderweg, F. J. & Klinkenberg, A. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. Chem. Eng. Sci. 5, 271–289 (1956).

38. Molnár, I. & Horváth, C. Separation of amino acids and peptides on non-polar stationary phases by high-performance liquid chromatography. J. Chromatogr. 142, 623–40 (1977).

39. Shen, Y. & Smith, R. D. Proteomics based on high-efficiency capillary separations. Electrophoresis 23, 3106–24 (2002).

40. MacNair, J. E., Lewis, K. C. & Jorgenson, J. W. Ultrahigh-Pressure Reversed-Phase Liquid Chromatography in Packed Capillary Columns. Anal. Chem. 69, 983–989 (1997).

41. MacNair, J. E., Patel, K. D. & Jorgenson, J. W. Ultrahigh-pressure reversed-phase capillary liquid chromatography: isocratic and gradient elution using columns packed with 1.0-micron particles. Anal. Chem. 71, 700–8 (1999).

42. Shen, Y. *et al.* Automated 20 kpsi RPLC-MS and MS/MS with Chromatographic Peak Capacities of 1000–1500 and Capabilities in Proteomics and Metabolomics. Anal. Chem. 77, 3090–3100 (2005).

43. Jorgenson, J. W. Capillary Liquid Chromatography at Ultrahigh Pressures. Annu. Rev. Anal. Chem. 3, 129–150 (2010).

44. Yufeng Shen *et al.* Ultra-High-Efficiency Strong Cation Exchange LC/RPLC/MS/MS for High Dynamic Range Characterization of the Human Plasma Proteome. (2004). doi:10.1021/AC034869M

45. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res. 2, 43–50 (2003).

46. Isobe, T., Takayasu, T., Takai, N. & Okuyama, T. High-performance liquid chromatography of peptides on a macroreticular cation-exchange resin: Application to peptide mapping of Bence-Jones proteins. Anal. Biochem. 122, 417–425 (1982).

47. Alpert, A. J. & Andrews, P. C. Cation-exchange chromatography of peptides on poly(2-sulfoethyl aspartamide)-silica. J. Chromatogr. A 443, 85–96 (1988).

48. Cachia, P. J., van Eyk, J., Chong, P. C. S., Taneja, A. & Hodges, R. S. Separation of basic peptides by cation-exchange high-performance liquid chromatography. J. Chromatogr. A 266, 651–659 (1983).

49. Mohammed, S. & Heck, A. J. Strong cation exchange (SCX) based analytical methods for the targeted analysis of protein post-translational modifications. Curr. Opin. Biotechnol. 22, 9–16 (2011).

50. Di Palma, S., Hennrich, M. L., Heck, A. J. R. & Mohammed, S. Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. J. Proteomics 75, 3791–813 (2012).

51. Gilar, M., Petra Olivova, Amy E. Daly, A. & Gebler, J. C. Orthogonality of Separation in Two-Dimensional Liquid Chromatography. (2005). doi:10.1021/AC050923I

52. Giddings, J. C. Maximum number of components resolvable by gel filtration and other elution chromatographic methods. Anal. Chem. 39, 1027–1028 (1967).

53. Giddings, J. C. Two-dimensional separations: concept and promise. Anal. Chem. 56, 1258A–1270A (1984).

54. Liu, Z., Patterson, D. G. & Lee, M. L. Geometric Approach to Factor Analysis for the Estimation of Orthogonality and Practical Peak Capacity in Comprehensive Two-Dimensional Separations. Anal. Chem. 67, 3840–3845 (1995).

55. Gilar, M., Daly, A. E., Kele, M., Neue, U. D. & Gebler, J. C. Implications of column peak capacity on the separation of complex peptide mixtures in single- and two-dimensional high-performance liquid chromatography. J. Chromatogr. A 1061, 183–92 (2004).

56. Yang, F., Shen, Y., Camp Ii, D. G. & Smith, R. D. High pH reversed-phase chromatography with fraction concatenation as an alternative to strong-cation exchange chromatography for two-dimensional proteomic analysis. Expert Rev. Proteomics 9, 129–134 (2012).

57. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-Line High-pH Reversed-Phase Fractionation for In-Depth Phosphoproteomics. doi:10.1021/pr500893m

58. Gilar, M., Olivova, P., Daly, A. E. & Gebler, J. C. Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. J. Sep. Sci. 28, 1694–1703 (2005).

59. Wang, Y. *et al.* Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. Proteomics 11, 2019–26 (2011).

60. Wolters, D. A., Washburn, M. P. & Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. Anal. Chem. 73, 5683–90 (2001).

61. Wien, W. Untersuchungen über die elektrische Entladung in verdünnten Gasen. Ann. Phys. 313, 244–266 (1902).

62. Goldstein, E. Vorläufige Mittheilungen über elektrische Entladungen in verdünnten Gasen, Monatsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin. (The Academy, 1876).

63. J.J., T. Rays of positive electricity. Proc. R. Soc. A 89, 1–20 (1913).

64. Nier, A. O. A Mass Spectrometer for Isotope and Gas Analysis. Rev. Sci. Instrum. 18, 398 (1947).

65. Inghram, M. G. & Gomer, R. Mass Spectrometric Analysis of Ions from the Field Microscope. J. Chem. Phys. 22, 1279 (1954).

66. Fenn, J., Mann, M., Meng, C., Wong, S. & Whitehouse, C. Electrospray ionization for mass spectrometry of large biomolecules. Science (80-.). 246, 64–71 (1989).

67. Tanaka, K. *et al.* Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 2, 151–153 (1988).

68. Yamashita, M. & Fenn, J. B. Electrospray ion source. Another variation on the free-jet theme. J. Phys. Chem. 88, 4451–4459 (1984).

69. Whitehouse, C. M., Dreyer, R. N., Yamashita, M. & Fenn, J. B. Electrospray interface for liquid chromatographs and mass spectrometers. Anal. Chem. 57, 675–679 (1985).

70. Malcolm Dole, L. L. Mack, R. L. Hines, R. C. Mobley, L. D. F. and M. B. A. Molecular Beams of Macroions. J. Chem. Phys. 49, 2240 (1968).

71. Rayleigh, Lord. XX. On the equilibrium of liquid conducting masses charged with electricity. Philos. Mag. Ser. 5 14, 184–186 (2009).

72. John B. Fenn, Koichi Tanaka, K. W. Press release: 'The Nobel Prize in Chemistry 2002'. Nobel Found. 1–13 (2002).

73. Taylor, G. Disintegration of Water Drops in an Electric Field. Proc. R. Soc. A Math. Phys. Eng. Sci. 280, 383–397 (1964).

74. Gomez, A. & Tang, K. Charge and fission of droplets in electrostatic sprays. Phys. Fluids 6, 404–414 (1994).

75. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 5, 699–711 (2004).

76. Wilm, M. S. & Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? Int. J. Mass Spectrom. Ion Process. 136, 167–180 (1994).

77. Iribarne, J. V. On the evaporation of small ions from charged droplets. J. Chem. Phys. 64, 2287 (1976).

78. Kebarle, P. & Verkerk, U. H. Electrospray: from ions in solution to ions in the gas phase, what we know now. Mass Spectrom. Rev. 28, 898–917

79. Emmett, M. R. & Caprioli, R. M. Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins. J. Am. Soc. Mass Spectrom. 5, 605–613 (1994).

80. Andren, P. E., Emmett, M. R. & Caprioli, R. M. Micro-Electrospray: Zeptomole/attomole per microliter sensitivity for peptides. J. Am. Soc. Mass Spectrom. 5, 867–9 (1994).

81. Schmidt, A., Karas, M. & Dülcks, T. Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: when does ESI turn into nano-ESI? J. Am. Soc. Mass Spectrom. 14, 492–500 (2003).

82. Karas, M., Bahr, U. & Dülcks, T. Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. Fresenius. J. Anal. Chem. 366, 669–676 (2000).

83. Cech, N. B. & Enke, C. G. Practical implications of some recent studies in electrospray ionization fundamentals. Mass Spectrom. Rev. 20, 362–387 (2001).

84. Karas, M., Bachmann, D. & Hillenkamp, F. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Anal. Chem. 57, 2935–2939 (1985).

85. Karas, M., Bachmann, D., Bahr, U. & Hillenkamp, F. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. Int. J. Mass Spectrom. Ion Process. 78, 53–68 (1987).

86. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal. Chem. 60, 2299–301 (1988).

87. Vertes, A. & Levine, R. D. Sublimation versus fragmentation in matrix-assisted laser desorption. Chem. Phys. Lett. 171, 284–290 (1990).

88. Johnson, R. E., Sundqvist, B. U. R. & Ens, W. Laser-pulse ejection of organic molecules from a matrix: Lessons from fast-ion-induced ejection. Rapid Commun. Mass Spectrom. 5, 574–578 (1991).

89. Karas, M., Glückmann, M. & Schäfer, J. Ionization in matrix-assisted laser desorption/ionization: singly charged molecular ions are the lucky survivors. J. Mass Spectrom. 35, 1–12 (2000).

90. Zenobi, R. & Knochenmuss, R. Ion formation in MALDI mass spectrometry. Mass Spectrom. Rev. 17, 337–366 (1998).

91. Jaskolla, T. W. & Karas, M. Compelling evidence for Lucky Survivor and gas phase protonation: the unified MALDI analyte protonation mechanism. J. Am. Soc. Mass Spectrom. 22, 976–88 (2011).

92. Paul, W. & Steinwedel, H. Notizen: Ein neues Massenspektrometer ohne Magnetfeld. Zeitschrift für Naturforsch. A 8, (1953).

93. Norman F. Ramsey, Hans G. Dehmelt, W. P. Press release: 'The Nobel Prize in Physics 1989'. Nobel Found. (1989).

94. March, R. E. An Introduction to Quadrupole Ion Trap Mass Spectrometry. J. Mass Spectrom. 32, 351–369 (1997).

95. Hoffmann, E. de & Stroobant, V. Mass Spectrometry: Principles and Applications. (Wiley, 2007).

96. Schwartz, J. C., Senko, M. W. & Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. J. Am. Soc. Mass Spectrom. 13, 659–69 (2002).

97. John E. P. Syka, William J. Fies, J. Fourtier transform quadrupole mass spectrometer and method. 4,755,670 (1988).

98. Stephens, W. E. A pulsed mass spectrometer with time dispersion. Phys. Rev 69, 674–792 (1946).

99. Guilhaus, M., Mlynski, V. & Selby, D. Perfect Timing: Time-of-flight Mass Spectrometry†. Rapid Commun. Mass Spectrom. 11, 951–962 (1997).

100. Cotter, R. J. Peer Reviewed: The New Time-of-Flight Mass Spectrometry. Anal. Chem. 71, 445A–51A (1999).

101. Mamyrin, B. A., Karataev, V. I., Shmikk, D. V. & Zagulin, V. A. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. Sov. Phys. JETP 37, (1973).

102. Dawson, J. H. J. & Guilhaus, M. Orthogonal-acceleration time-of-flight mass spectrometer.

Rapid Commun. Mass Spectrom. 3, 155–159 (1989).

103. Guilhaus, M. Principles and Instrumentation in Time-of-flight Mass Spectrometry. J. Mass Spectrom. 30, 1519–1532 (1995).

104. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. Anal. Chem. 72, 1156–1162 (2000).

105. Kingdon, K. H. A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. Phys. Rev. 21, 408–418 (1923).

106. Makarov, A. Mass spectrometer. U.S. Patent 5,886,346 (1999).

107. Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. Fourier transform mass spectrometry. Mol. Cell. Proteomics 10, M111.009431 (2011).

108. Makarov, A. *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal. Chem. 78, 2113–20 (2006).

109. Makarov, A., Denisov, E., Lange, O. & Horning, S. Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer. J. Am. Soc. Mass Spectrom. 17, 977–982 (2006).

110. Olsen, J. V *et al.* Higher-energy C-trap dissociation for peptide modification analysis. Nat. Methods 4, 709–12 (2007).

111. McAlister, G., Phanstiel, D., Good, D., Berggren, T. & Coon, J. Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. Analytical Chemistry 79, 3525–3534 (2007).

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

113. Second, T. P. *et al.* Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. Anal. Chem. 81, 7757–65 (2009).

114. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol. Cell. Proteomics 10, M111.011015 (2011).

115. Michalski, A. *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 11, O111.013698-O111.013698 (2011).

116. Lange, O., Damoc, E., Wieghaus, A. & Makarov, A. Enhanced Fourier transform for Orbitrap mass spectrometry. Int. J. Mass Spectrom. 369, 16–22 (2014).

117. Scheltema, R. A. *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 13, 3698–708 (2014).

118. Senko, M. W. *et al.* Novel parallelized quadrupole/linear ion trap/Orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. Anal. Chem. 85, 11710–4 (2013).

119. Andrews, G. L., Simons, B. L., Young, J. B., Hawkridge, A. M. & Muddiman, D. C. Performance Characteristics of a New Hybrid Quadrupole Time-of-Flight Tandem Mass Spectrometer (TripleTOF 5600). Anal. Chem. 83, 5442–5446 (2011).

120. Selevsek, N. *et al.* Reproducible and Consistent Quantification of the Saccharomyces cerevisiae Proteome by SWATH-mass spectrometry. Mol. Cell. Proteomics 14, 739–749 (2015).

121. Zhang, Y. *et al.* Data for a comprehensive map and functional annotation of the human cerebrospinal fluid proteome. Data Br. 3, 103–107 (2015).

122. Jones, K. A. *et al.* Immunodepletion Plasma Proteomics by TripleTOF 5600 and Orbitrap

Elite/LTQ-Orbitrap Velos/Q Exactive Mass Spectrometers. J. Proteome Res. 12, 4351–4365 (2013).

123. Han, X., Aslanian, A. & Yates Iii, J. R. Mass Spectrometry for Proteomics. Curr Opin Chem Biol 12, 483–490 (2008).

124. Shukla, A. K. & Futrell, J. H. Tandem mass spectrometry: dissociation of ions by collisional activation. J. Mass Spectrom. 35, 1069–1090 (2000).

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

126. Sleno, L. & Volmer, D. A. Ion activation methods for tandem mass spectrometry. J. Mass Spectrom. 39, 1091–112 (2004).

127. Dongré, A. R., Jones, J. L., And, Á. S. & Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. J. Am. Chem. Soc. 118, 8365–8374 (1996).

128. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. J. Mass Spectrom. 35, 1399–406 (2000).

129. Roman A. Zubarev, Neil L. Kelleher, A. & McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. J. Am. Chem. Soc. 120, 3265–3266 (1998).

130. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 101, 9528–33 (2004).

131. Swaney, D. L. *et al.* Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. Anal. Chem. 79, 477–85 (2007).

132. Sharon J. Pitteri, Paul A. Chrisman, A. & McLuckey, S. A. Electron-transfer ion/ion reactions of doubly protonated peptides: Effect of elevated bath gas temperature. Anal. Biochem. 77, 5662–5669 (2005).

133. Horn, D. M., Ge, Y. & McLafferty, F. W. Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins. Anal. Chem. 72, 4778–84 (2000).

134. Frese, C. K. *et al.* Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. Anal. Chem. 84, 9668–9673 (2012).

135. Frese, C. K. *et al.* Unambiguous Phosphosite Localization using Electron-Transfer/Higher-Energy Collision Dissociation (EThcD). J. Proteome Res. 12, 1520–1525 (2013).

136. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics 73, 2092–2123 (2010).

137. Nesvizhskii, A. I. Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol. Biol. 367, 87–119 (2007).

138. Cottrell, J. S. Protein identification using MS/MS data. J. Proteomics 74, 1842–1851 (2011).

139. Sadygov, R. G., Cociorva, D. & Yates, J. R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat. Methods 1, 195–202 (2004).

140. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–67 (1999).

141. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom.

5, 976–989 (1994).

142. Benjamini, Y. & Hochberg, Y. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300 (1995).

143. Sakurai, T., Matsuo, T., Matsuda, H. & Katakuse, I. PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. Biol. Mass Spectrom. 11, 396–399 (1984).

144. Bartels, C. Fast algorithm for peptide sequencing by mass spectroscopy. Biol. Mass Spectrom. 19, 363–368 (1990).

145. Seidler, J., Zinn, N., Boehm, M. E. & Lehmann, W. D. De novo sequencing of peptides by MS/MS. Proteomics 10, 634–49 (2010).

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

147. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. 389, 1017–1031 (2007).

148. Oda, Y., Huang, K., Cross, F. R., Cowburn, D. & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. Proc. Natl. Acad. Sci. U. S. A. 96, 6591–6 (1999).

149. Ljiljana Paša-Tolić *et al.* High throughput proteome-wide precision measurements of protein expression using mass spectrometry. J. Am. Chem. Soc. 121, 7949–7950 (1999).

150. Ong, S.-E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics 1, 376–86 (2002).

151. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–9 (1999).

152. Han, D. K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat. Biotechnol. 19, 946–51 (2001).

153. Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol. Cell. Proteomics 3, 1154–69 (2004).

154. Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. 75, 1895–1904 (2003).

155. Hsu, J.-L., Huang, S.-Y., Chow, N.-H. & Chen, S.-H. Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. Anal. Chem. 75, 6843–6852 (2003).

156. Boersema, P. J., Aye, T. T., van Veen, T. A. B., Heck, A. J. R. & Mohammed, S. Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates. Proteomics 8, 4624–4632 (2008).

157. Miyagi, M. & Rao, K. C. S. Proteolytic18O-labeling strategies for quantitative proteomics. Mass Spectrom. Rev. 26, 121–136 (2007).

158. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat. Protoc. 4, 484–94 (2009).

159. Desiderio, D. M. & Kai, M. Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. Biol. Mass Spectrom. 10, 471–479 (1983).

160. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc. Natl. Acad. Sci. 100, 6940–6945 (2003).

161. Liu, H., Rovshan G. Sadygov, A. & John R. Yates, I. A model for random sampling and

estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201 (2004).

162. Zhu, W. *et al.* Mass spectrometry-based label-free quantitative proteomics. J. Biomed. Biotechnol. 2010, 840518 (2010).

163. Wang, G., Wu, W. W., Zeng, W., And, C.-L. C. & Shen, R.-F. Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. J. Proteome Res. 5, 1214–1223 (2006).

164. Nahnsen, S., Bielow, C., Reinert, K. & Kohlbacher, O. Tools for Label-free Peptide Quantification. Mol. Cell. Proteomics 12, 549–556 (2013).

165. Weissman, I. L. Stem cells: units of development, units of regeneration, and units in evolution. Cell 100, 157–68 (2000).

166. Barker, N., van de Wetering, M. & Clevers, H. The intestinal stem cell. Genes Dev. 22, 1856–64 (2008).

167. van der Flier, L. G. & Clevers, H. Stem Cells, Self-Renewal, and Differentiation in the Intestinal Epithelium. Annu. Rev. Physiol. 71, 241–260 (2009).

168. Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. Nat. Rev. Mol. Cell Biol. 15, 19–33 (2013).

169. Sancho, E., Batlle, E. & Clevers, H. Live and let die in the intestinal epithelium. Curr. Opin. Cell Biol. 15, 763–70 (2003).

170. Gerbe, F. *et al.* Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. Nature 529, 226–230 (2016).

171. Ayabe, T. *et al.* Secretion of microbicidal |[alpha]|-defensins by intestinal Paneth cellsin response to bacteria. Nat. Immunol. 1, 113–118 (2000).

172. Madara, J. L. Cup cells: structure and distribution of a unique class of epithelial cells in guinea pig, rabbit, and monkey small intestine. Gastroenterology 83, 981–94 (1982).

173. Neutra, M. R. Current concepts in mucosal immunity. V Role of M cells in transepithelial transport of antigens and pathogens to the mucosal immune system. Am. J. Physiol. 274, G785-91 (1998).

174. Altmann, G. G. Morphological observations on mucus-secreting nongoblet cells in the deep crypts of the rat ascending colon. Am. J. Anat. 167, 95–117 (1983).

175. Magney, J. E., Erlandsen, S. L., Bjerknes, M. L. & Cheng, H. Scanning electron microscopy of isolated epithelium of the murine gastrointestinal tract: Morphology of the basal surface and evidence for paracrinelike cells. Am. J. Anat. 177, 43–53 (1986).

176. Cheng, H. & Leblond, C. P. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine. V. Unitarian Theory of the origin of the four epithelial cell types. Am. J. Anat. 141, 537–61 (1974).

177. Barker, N., Bartfeld, S. & Clevers, H. Tissue-resident adult stem cell populations of rapidly self-renewing organs. Cell Stem Cell 7, 656–70 (2010).

178. Barker, N. *et al.* Identifying the stem cell of the intestinal crypt: strategies and pitfalls. Cell Stem Cell 11, 452–60 (2012).

179. Potten, C. S. Extreme sensitivity of some intestinal crypt cells to X and gamma irradiation. Nature 269, 518–21 (1977).

180. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 449, 1003–1007 (2007).

181. Korinek, V. *et al.* Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. Nat. Genet. 19, 379–83 (1998).

182. Nusse, R. Wnt signaling in disease and in development. Cell Res. 15, 28–32 (2005).

183. Sancho, E., Batlle, E. & Clevers, H. Signaling pathways in intestinal development and cancer. Annu. Rev. Cell Dev. Biol. 20, 695–723 (2004).

184. van de Wetering, M. *et al.* The β-Catenin/TCF-4 Complex Imposes a Crypt Progenitor Phenotype on Colorectal Cancer Cells. Cell 111, 241–250 (2002).

185. Van der Flier, L. G. *et al.* The Intestinal Wnt/TCF Signature. Gastroenterology 132, 628–632 (2007).

186. Hsu, S. Y., Liang, S.-G. & Hsueh, A. J. W. Characterization of Two LGR Genes Homologous to Gonadotropin and Thyrotropin Receptors with Extracellular Leucine-Rich Repeats and a G Protein-Coupled, Seven-Transmembrane Region. Mol. Endocrinol. 12, 1830–1845 (1998).

187. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 449, 1003–7 (2007).

188. Buczacki, S. J. A. *et al.* Intestinal label-retaining cells are secretory precursors expressing Lgr5. Nature 495, 65–69 (2013).

189. Tian, H. *et al.* A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. Nature 478, 255–259 (2011).

190. Clevers, H. Stem Cells: A unifying theory for the crypt. Nature 495, 53–54 (2013).

191. Barker, N., Tan, S. & Clevers, H. Lgr proteins in epithelial stem cell biology. Development 140, 2484–2494 (2013).

192. Marshman, E., Booth, C. & Potten, C. S. The intestinal epithelial stem cell. Bioessays 24, 91–8 (2002).

193. Ireland, H., Houghton, C., Howard, L. & Winton, D. J. Cellular inheritance of a Cre-activated reporter gene to determine paneth cell longevity in the murine small intestine. Dev. Dyn. 233, 1332–1336 (2005).

194. Jung, P. *et al.* Isolation and in vitro expansion of human colonic stem cells. Nat. Med. 17, 1225–1227 (2011).

195. Kinzler, K. W. & Vogelstein, B. Lessons from Hereditary Colorectal Cancer. Cell 87, 159–170 (1996).

196. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. Nature 457, 608–611 (2009).

197. Sachs, N. & Clevers, H. Organoid cultures for the analysis of cancer phenotypes. Curr. Opin. Genet. Dev. 24, 68–73 (2014).

198. Evans, G. S., Flint, N., Somers, A. S., Eyden, B. & Potten, C. S. The development of a method for the preparation of rat intestinal epithelial cell primary cultures. J. Cell Sci. 219–31 (1992).

199. Whitehead, R. H., Demmler, K., Rockman, S. P. & Watson, N. K. Clonogenic growth of epithelial cells from normal colonic mucosa from both mice and humans. Gastroenterology 117, 858–65 (1999).

200. Ootani, A. *et al.* Sustained in vitro intestinal epithelial culture within a Wnt-dependent stem cell niche. Nat. Med. 15, 701–706 (2009).

201. Barker, N., van de Wetering, M. & Clevers, H. The intestinal stem cell. Genes Dev. 22, 1856–64 (2008).

202. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. Nature 457, 608–11 (2009).

203. Muñoz, J. *et al.* The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. EMBO J. 31, 3079–91 (2012).

204. Abbott, A. Cell culture: biology's new dimension. Nature 424, 870–2 (2003).

205. Tanner, K. & Gottesman, M. M. Beyond 3D culture models of cancer. Sci. Transl. Med. 7, 283ps9 (2015).

206. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 459, 262–5 (2009).

207. Leushacke, M. & Barker, N. Ex vivo culture of the intestinal epithelium: strategies and applications. Gut 63, 1345–54 (2014).

208. Pinto, D., Gregorieff, A., Begthel, H. & Clevers, H. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. Genes Dev. 17, 1709–13 (2003).

209. Kuhnert, F. *et al.* Essential requirement for Wnt signaling in proliferation of adult small intestine and colon revealed by adenoviral expression of Dickkopf-1. Proc. Natl. Acad. Sci. 101, 266–271 (2004).

210. Kim, K.-A. *et al.* Mitogenic influence of human R-spondin1 on the intestinal epithelium. Science 309, 1256–9 (2005).

211. de Lau, W. *et al.* Lgr5 homologues associate with Wnt receptors and mediate R-spondin signalling. Nature 476, 293–297 (2011).

212. Dignass, A. U. & Sturm, A. Peptide growth factors in the intestine. Eur. J. Gastroenterol. Hepatol. 13, 763–70 (2001).

213. Hofmann, C. *et al.* Cell-cell contacts prevent anoikis in primary human colonic epithelial cells. Gastroenterology 132, 587–600 (2007).

214. Sasaki, T., Giltay, R., Talts, U., Timpl, R. & Talts, J. F. Expression and distribution of laminin alpha1 and alpha2 chains in embryonic and adult mouse tissues: an immunochemical approach. Exp. Cell Res. 275, 185–99 (2002).

215. Hughes, C. S., Postovit, L. M. & Lajoie, G. A. Matrigel: a complex protein mixture required for optimal growth of cell culture. Proteomics 10, 1886–90 (2010).

216. Sato, T. & Clevers, H. Growing self-organizing mini-guts from a single intestinal stem cell: mechanism and applications. Science 340, 1190–4 (2013).

217. Koo, B.-K. *et al.* Controlled gene expression in primary Lgr5 organoid cultures. Nat. Methods 9, 81–83 (2011).

218. Andersson-Rolf, A., Fink, J., Mustata, R. C. & Koo, B.-K. A video protocol of retroviral infection in primary intestinal organoid culture. J. Vis. Exp. e51765 (2014). doi:10.3791/51765

219. Fujii, M., Matano, M., Nanki, K. & Sato, T. Efficient genetic engineering of human intestinal organoids using electroporation. Nat. Protoc. 10, (2015).

220. Broutier, L. *et al.* Culture and establishment of self-renewing human and mouse adult liver and pancreas 3D organoids and their genetic manipulation. Nat. Protoc. 11, 1724–1743 (2016).

221. Farin, H. F. *et al.* Paneth cell extrusion and release of antimicrobial products is directly controlled by immune cell–derived IFN-γ. J. Exp. Med. 211, (2014).

222. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. Nature 469, 415–418 (2011).

223. Grabinger, T. *et al.* Ex vivo culture of intestinal crypt organoids as a model system for assessing cell death induction in intestinal epithelial cells and enteropathy. Cell Death Dis. 5, e1228 (2014).

224. Dekkers, J. F. *et al.* A functional CFTR assay using primary cystic fibrosis intestinal organoids. Nat. Med. 19, 939–945 (2013).

225. van de Wetering, M. *et al.* Prospective Derivation of a Living Organoid Biobank of Colorectal

Cancer Patients. Cell 161, 933–945 (2015).

226. Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology 141, 1762–72 (2011).

227. Malvezzi, M. et al. European cancer mortality predictions for the year 2015: does lung cancer have the highest death rate in EU women? Ann. Oncol. 26, 779–86 (2015).

228. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2015. CA. Cancer J. Clin. 65, 5–29 (2015).

229. Fodde, R. & Brabletz, T. Wnt/β-catenin signaling in cancer stemness and malignant behavior. Curr. Opin. Cell Biol. 19, 150–158 (2007).

230. Powell, S. M. et al. APC mutations occur early during colorectal tumorigenesis. Nature 359, 235–7 (1992).

231. Morin, P. J. et al. Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. Science 275, 1787–90 (1997).

232. Barker, N., Huls, G., Korinek, V. & Clevers, H. Restricted high level expression of Tcf-4 protein in intestinal and mammary gland epithelium. Am. J. Pathol. 154, 29–35 (1999).

233. Bos, J. L. et al. Prevalence of ras gene mutations in human colorectal cancers. Nature 327, 293–7 (1987).

234. Bardelli, A. Mutational Analysis of the Tyrosine Kinome in Colorectal Cancers. Science (80-.). 300, 949–949 (2003).

235. Markowitz, S. et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 268, 1336–8 (1995).

236. Baker, S. J. et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science 244, 217–21 (1989).

237. Rodrigues, N. R. et al. p53 mutations in colorectal cancer. Proc. Natl. Acad. Sci. U. S. A. 87, 7555–9 (1990).

238. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–67 (1990).

239. West, N. R., McCuaig, S., Franchini, F. & Powrie, F. Emerging cytokine networks in colorectal cancer. Nat. Rev. Immunol. 15, 615–629 (2015).

240. de la Chapelle, A. Genetic predisposition to colorectal cancer. Nat. Rev. Cancer 4, 769–80 (2004).

241. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264 (2016).

242. Fearon, E. R. Molecular genetics of colorectal cancer. Annu. Rev. Pathol. 6, 479–507 (2011).

243. Markowitz, S. D. & Bertagnolli, M. M. Molecular origins of cancer: Molecular basis of colorectal cancer. N. Engl. J. Med. 361, 2449–60 (2009).

Chapter 2

In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis

In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis.

Alba Cristobal,[1,2,*] Marco L. Hennrich,[1,2,*] Piero Giansanti,[1,2] Soenita S. Goerdayal,[1,2] Albert J.R. Heck,[1,2,§] Shabaz Mohammed[1,2,§]

Analyst, 18 May 2012

[1]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

[2]Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

[*]These authors contribute equally to this work

Abstract

Here, we describe an in-house built ultra-high pressure liquid chromatography (UHPLC) system, with little complexity in design and high separation power combined with convenience in operation. This system enables the use of long columns of 40 cm packed with 1.8 μm particles generating pressures below 1000 bar. Furthermore, the system could be operated at flow rates between 50 to 200 nL/min while maintaining its separation power. Several gradients were optimized ranging from 23 to 458 minutes. With the longest gradient we identified over 4500 protein groups and more than 26000 unique peptides from 1 μg of a human cancer cell lysate in a single run using an Orbitrap Velos; a level of performance often seen solely using multidimensional separation strategies. Further experiments using a mass spectrometer with faster sequencing speeds, like the TripleTOF 5600, enabled us to identify over 1400 protein groups in a 23 minute gradient. The TripleTOF 5600 performed especially well, compared to the Orbitrap Velos, for the shorter gradients used. Our data demonstrate that the combination of UHPLC with high resolution mass spectrometry at increased sequencing speeds enables comprehensive proteome analysis in single runs.

Introduction

In recent years mass spectrometry (MS) has taken a central place in proteomics due to its sensitivity, increasingly broad availability and ideal applicability for protein and peptide analysis.[1–3] Mass spectrometers are continuously improving, powered by the need to analyze more complex samples with higher speed and sensitivity. A combination of developments in new instrumentation[4–6] and fragmentation methods[7,8] has improved the analytical power of MS by orders of magnitude. However, the overwhelming complexity of the proteome still demands that proteome samples are pre-fractionated and separated at either the protein and/or peptide level prior to delivery to the mass analyzer.

It is by now generally considered that at least two dimensions of separations are needed to cover a proteome in a reasonable depth. Such developments were initiated by Giddings[9] who first described the principles of two-dimensional separation.

The concept of multidimensional protein identification technology (MudPIT) was brought to the attention of the proteomics community by Yates and co-workers.[10] Since then, several groups used the combination of strong cation exchange (SCX) with reverse phase (RP) chromatography for reducing the complexity and dynamic range of the samples. Gygi and co-workers combined SCX-RP chromatography in an off-line approach for the analysis of the yeast proteome.[11] Chen *et al.* built an online system coupling capillary isoelectric focusing (CIEF) with capillary RP and showed its excellent performance.[12] Additionally, most separations are being further developed to possess ever higher efficiencies. The final step in most multiple dimensions strategies is reversed phase chromatography.[13,14] A number of different configurations have been developed to optimize RP nanoLC. The most basic design consists simply of the analytical column operating at nanoliters flow rates where a few microliter of sample are directly loaded onto the column.[15,16] Column dimensions most commonly used are 75 μm i.d. with a length of 15 cm operating at a flow rate of 250 nL/min.[17] However, LC designs incorporating a trap now dominate the landscape. Within the world of trap column systems, two main philosophies share the limelight;[18] the first is based on a two pump system with a trap over a valve,[19] the second possesses a single pump and uses vent lines and passive splitting.[20,21] Both configurations also led to the development of convenient commercial chips.[22,23]

A number of different avenues are being explored to further improve the power of reversed phase separation. To achieve better separation several parameters regarding the column can be changed. Longer columns will generate more theoretical plates, leading to a higher efficiency. However, the analysis time will be longer and the column back pressure rises. Another factor that influences the efficiency of the column is its inner diameter. Jorgenson showed that the performance improves as the column diameter decreases.[24] Furthermore, column efficiency is dramatically influenced by the particle size of the stationary phase. By decreasing the particle size, the analyte diffusion is minimized and the resolution increases. The major obstacle is the high back pressure created by such small particles,[24] which is inversely proportional to the square of the particle size.[25] When dealing with sub-2 μm particles, a new term has been coined, namely Ultra-High Pressure Liquid Chromatography (UHPLC). Jorgenson and co-workers were the first to demonstrate the power of UHPLC,[25,26] which led to a commercially available high

pressure instrument being introduced by Swartz *et al.* soon after.[27] More and more groups are starting to use UHPLC coupled to MS for proteomic studies. For instance, Mechtler and co-workers showed a linear relation between the peak capacity and the number of identified peptides using a UHPLC system hyphenated to an LTQ-Orbitrap Velos.[28] In contrast, Ishihama and co-workers mentioned that the peak capacity is not correlated with the number of peptide identifications.[29] Shen *et al.* reached peak capacities of 1000-1500 by using particles of 1.4-3 μm on a system that could operate at pressures of approximately 1400 bar.[30] Mann and co-workers have also demonstrated a system containing an analytical column packed with 1.8 μm particles; however, they were only able to use the column on a nanoHPLC system and thus used an oven and reduced flow rates to maintain an acceptable back pressure. Nevertheless, the new column design was capable of allowing the coverage of a large part of the yeast proteome without any prefractionation step.[16] Also monolithic columns have been introduced as an alternative to packed capillary columns. Ishihama and co-workers obtained a 5-fold larger peak response using a monolithic column compared with that obtained by a conventional particle-packed capillary column.[29] Smith and co-workers described how to prepare monolithic columns and also tested their performance.[31]

Here, we try to combine convenience with the high separation power of UHPLC. A design inspired by Licklider[20] and Ishihama[21] was constructed and evaluated with respect to peak capacity, sensitivity and general performance for peptide and protein identification. This home-built system, using 1.8 μm particles, was coupled to two different mass spectrometers, an Orbitrap Velos and a TripleTOF, with the aim of obtaining the best results for different gradient times. A benchmark quality of the constructed system is that it enabled the identification of over 4500 protein groups in a single RP-LC analysis using 1 μg of human cancer cell lysate.

Material and Methods

Chemicals and Materials

Bovine Serum Albumin and iodoacetamide were supplied by Sigma-Aldrich (Steinheim, DE). Ammonium bicarbonate and dithiothreitol were purchased from Fluka (Buchs, CH), and urea from Merck (Darmstadt, DE). PhosSTOP

Phosphatase Inhibitor Cocktail tablets, Complete Mini EDTA-free Cocktail tablets and Lys-C were obtained from Roche Diagnostics (Mannheim, DE), and trypsin from Promega Corporation (Madison, WI, USA). The Bradford solution for the Bradford protein assay was supplied by Bio-Rad Laboratories (Hercules, CA, USA).

For the fabrication of the trap and analytical columns the following chemicals were used: acetone and 2-propanol were supplied from Merck (Darmstadt, DE), and methanol HPLC grade from Biosolve B.V. (Valkenswaard, NL). The packing material used was a kind gift from Dr. Bill Barber of Agilent Technologies, Zorbax SB-C18, 1.8 µm (Agilent, Santa Clara, CA, USA).

The water used in all experiments was obtained from a Milli-Q purification system (Millipore, Bedford, MA, USA).

Sample preparation

BSA, HeLa and E. coli samples were prepared as described previously:[32] Prior to digestion proteins were reduced (with dithiothreitol) and carbamidomethylated (with iodoacetamide). For the digestion Lys-C and trypsin were used. First, proteins were digested with Lys-C at 37 °C for 4 hours and subsequently digested with trypsin overnight at 37 °C. In the case of the HeLa and E. coli lysates, the protein concentration was previously estimated by a Bradford assay.

Preparation of packed capillary columns

Fused-silica capillary analytical and trap columns were prepared by using slurries.[33] Frits were prepared as described by Cortes *et al.*[34] The frit solution consisted of 300 µL potassium silicate and 100 µL formamide. Polymerization was performed for 1 hour at 100°C. The packing of the column was done by back flushing the material from a capillary of wider inner diameter (75 µm). For preparing the slurry half of a mini spatula of the material was mixed with methanol:2-propanol in a 3:1 ratio. The slurry was placed in a glass vial containing a magnetic stirrer. An initial 75 µm capillary (with a frit) was filled and then the material with the solvent was flushed back to the desired capillary of 50 µm which was connected with a teflon tubing. Compared with the analytical column, the trap column is a double frit capillary. In

this case, after filling the desired length with the packing material, the second frit has to be placed.

UHPLC and mass spectrometers

Nano-UHPLC-MS/MS was performed on an Agilent 1290 Infinity System (Agilent Technologies, Waldbronn, DE) connected to different mass spectrometers. An ultra-high pressure valve was used to handle the split and vented columns (VICI, C72-1C96). T-pieces (IDEX, MicroTee Assy SS 360 µm) that can handle 1200 bar were used to connect trap and analytical columns. The UHPLC was equipped with a double frit trapping column (Agilent Zorbax SB-C18, 1.8 µm material, 0.5 cm x 100 µm) and a single frit analytical column (Agilent Zorbax SB-C18, 1.8 µm material, 40 cm x 50 µm). The buffers used were: buffer A 0.1 M acetic acid in water and buffer B 0.1 M acetic acid in 80% acetonitrile.

Trapping was performed at 3 µL/min with 100% buffer A for 10 minutes. For the elution the flow rate was passively split to 100 nL/min. This column flow rate can be measured via a 0.3 mm ID Teflon sleeve butt-connected to the end of the analytical column. Checking the distance that the flow needs to travel through the Teflon sleeve in five minutes, it is possible to calculate the column flow. If the distance is between 7-8 mm, the flow through the column is ~100 nL/min. The tested gradients varied in length between 23 and 458 minutes, which correspond to analysis times of 45 to 480 minutes, due to the sample loading and washing step. Exact gradients can be found in Supplementary Table S1.

Conventional HPLC was also performed on the Agilent 1290 but with a few modifications. A double frit trapping column (Reprosil-pur C18 3 µm, Dr Maisch, Ammerbuch-Entringen, DE, 2 cm x 100 µm) and a single frit analytical column (Reprosil-pur C18 3 µm, Dr Maisch, Ammerbuch-Entringen, DE, 40 cm x 50 µm) were used.

The column, in both cases, was directly connected to an in-house pulled and gold-coated fused silica needle (with an 5 µm o.d. tip).

For the Orbitrap "Classic" (Thermo Fisher Scientific, Bremen, DE), a voltage of 1.7 kV was applied to the needle. The survey scan was from 350 to 1500 m/z at a

resolution of 60000. The 10 most intense precursors were selected for subsequent fragmentation using a data-dependent acquisition.

For the LTQ-Orbitrap Velos (Thermo Fisher Scientific, Bremen, DE), a voltage of 1.7 kV was applied to the needle. The survey scan was from 350 to 1500 m/z at a resolution of 30000 and for the MS2 the resolution was set to 7500. The 10 most intense precursors were selected for subsequent fragmentation (HCD) using a data-dependent acquisition.

For the TripleTOF 5600 (AB Sciex, Concord, ON, CA), a voltage of 2.7 kV was applied to the needle. The survey scan was from 350 to 1250 m/z and the high resolution mode was utilized, reaching a resolution of up to 40000. Tandem mass spectra were acquired in high sensitivity mode with a resolution of 20000. The 20 most intense precursors were selected for subsequent fragmentation using an information dependent acquisition, with a minimum acquisition time of 50 ms.

Data analysis

The raw files obtained from the Velos and the Classic were first converted to Mascot generic format (mgf) files using Proteome Discoverer (version 1.3). The mgf files derived from the Velos were deisotoped and charge deconvoluted with the H-Score script, described elsewhere.[35] The raw files collected from the TripleTOF were first recalibrated based on two background ions with m/z values of 371.1012 and 445.1200. The calibrated raw files were converted to mgf by the AB Sciex MS Data Converter (version 1.1 beta) program.

All data was analyzed with Mascot using Proteome Discoverer for submission. The spectra were searched against SwissProt database (version 56.2), and Homo Sapiens was the selected taxonomy. Trypsin was set as the enzyme and up to two missed cleavages were allowed. Cysteine carbamidomethylation was set as fixed modification and oxidation of methionine as variable modification. Peptide tolerance was initially set to 50 ppm for both instruments. However the MS/MS tolerance was 0.05 Da for the Velos, 0.15 Da for the TripleTOF and 0.6 Da for the Orbitrap Classic. Filtered peptides assigned with a Mascot score lower than 20 were discarded. The results were also filtered using Percolator[36,37] to an FDR below 1% using the target-decoy strategy described elsewhere.[38] The peptides with less than 7

amino acid residues were also discarded.

The obtained raw and msf files are available on Tranche at https://proteomecommons.org using the following hash:

wLY+buLAHzNqpY1/Ryvz2/wmzF52a+GX+csgRkk+YvgUFVgiLl84fxxiqAy/q++ZF1NtP5hYP9Jy9Uff6VtMO3q5tn0AAAAAAAAkNA==. The passphrase is: 5rawandmsffiles5.

For testing the resolving power of the system the peak capacity was calculated.[39] To this end, the same thirteen random peaks were chosen in all the chromatograms. Their peak with at 4σ was calculated. These peaks were chosen across the gradient and represented peaks from the observed range of intensities.

Results and Discussion

A simple and easy to use nano-UHPLC system was developed for proteomics analysis. We opted to use a passive split and vented column system, a configuration described by Licklider *et al.*,[20] and Ishihama *et al.*.[21] A slightly refined version was presented in the same year by Meiring *et al.*.[40] Figure 1a contains a schematic of the design.

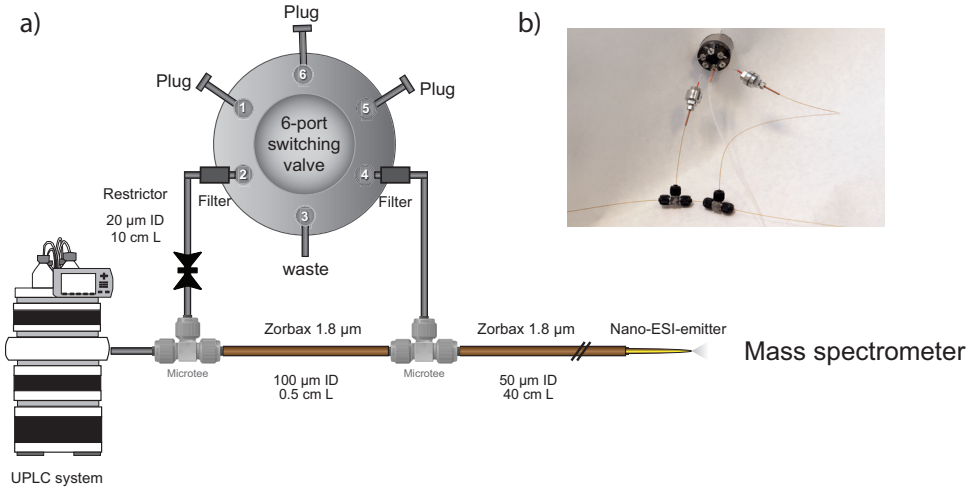


Figure 1. Design of the new UHPLC-MS system. (a) Schematic representation of the design. The trap and analytical columns are made of C18 1.8 mm material. The dimensions of the

trap are 100 µm x 0.5 cm and for the analytical column 50 µm x 40 cm. They are connected by a metal T-piece. During analytical separation, the flow rate is split by a restrictor (a 20 µm i.d. capillary of 10 cm length). A needle, which is connected to the column by a sleeve, directs the peptides to the mass spectrometer. (b) Photo of the system.

Control of the passive split and the vent line between the trap and analytical column is performed by a switching valve. We opted for the use of a 1290 HPLC (Agilent) which allows accurate isocratic flow down to ~5 µL/min and gradient generation down to 50 µL/min. All the methods start with a loading step followed by a gradient. The loading of the sample is performed at an unsplit flow of 3 µL/min during 10 minutes, which is obtained by blocking the passive split and opening the vent line to waste. In the separation mode, the passive split line is opened while the vent line is closed. The advantage of such a system is the small dead volume between the trap and the column, which is essentially the minute gap in the second T-piece. In the separation mode the flow is set to 100-300 µL/min in the UHPLC, depending on the length of the column and the desired flow rate. This primary flow is passively split to 50 to 200 nL/min. Naturally, the end flow rate is controlled by a combination of the pump, the resistance of the passive split and the back pressure generated by the column. Such a design has a number of advantages. Fast loading and washing of the sample is possible and thus the injection volume is not an issue. Accurate gradient generation is possible due to the use of microliter flow rates and a dedicated mixing chamber. The dimensions of the trap are 0.5 cm x 100 µm, which has a capacity of at least 5 µg of peptide digest although we found it was not necessary to go above 2 µg (see below). The analytical column was varied in length from 10 to 40 cm x 50 µm. The diameter of the particles used for both columns is 1.8 µm. We envisage that such dimensions will lead to pressures of close to 1000 bar at a flow rate of 100 nL/min for the longest column. Consideration is paid to all the connections due to the expected high pressure including the choice of T-pieces.

An initial characterisation was necessary to establish the relationship between column dimensions, flow rate and back pressure for the reversed phase material used in this work. As the particle size is smaller, higher pressure was, of course, expected. Jorgenson has summarized trends between the packing material and the pressure required.[24] Initial experiments were performed with a column of 10 cm in length. Different methods from 30 to 180 minutes were run with a passive flow

rate of 100 nL/min. We found the pressure was approximately ~200 bar, suggesting a 40 cm long analytical column was feasible within our upper pressure limit. Subsequently, we attempted to discover the optimum parameters for this design. These include the flow rate with regard to separation power and dead time with regard to the chromatography. Attention was also placed on the ionization and, more specifically, sensitivity and stability. Analysis of a tryptic digest of BSA was performed at flow rates from 50 to 200 nL/min using a standard short gradient. We chose a gradient from 10 to 44% solvent B over 23 minutes using the 10 cm analytical column. The width of several peaks in these analyses at different flow rates was calculated. We found that the peak width does not vary with the flow rate, obtaining an average width of 0.13 minutes at full width half maximum (FWHM). Figure 2a shows representative chromatograms of some of the flow rates and Figure 2b contains information of the calculated average peak widths.

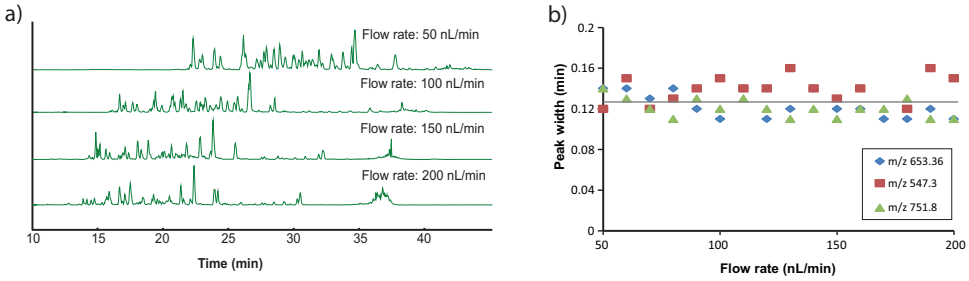


Figure 2. Independence of the efficiency of separation with respect to flow rate. In figure a) the base peak chromatograms of 23 minutes gradient time of a tryptic digest of BSA running at flow rates of 50, 100, 150 and 200 nL/min are shown. It can be clearly seen how the retention time decreases as faster flow rates are applied. However, the efficiency does not decrease with the flow rate as it is represented in figure b). It shows a scatter plot of the width of three BSA peaks at FWHM (with m/z values of 653.36, 547.3 and 751.8) against the flow rates tested (from 50 to 200 nL/min).

Supporting Figure S1 shows the chromatograms of all 16 of the tested flow rates. As consistent with the van Deemter equation,[41] the plate height of the column becomes less dependent on the linear velocity of the sample as it passes through the column when dealing with sub-2 μm particles. Several groups have already reported this trend.[21,24] As expected, the sensitivity was highest at the lowest flow rate and signal

diminished with increasing flow rate. Conversely, the dead time was longest at the lowest flow rate and decreased with increasing flow. As a compromise between efficient use of time and sensitivity a flow rate of 100 nL/min was chosen although the system can operate in a more sensitive manner.

Once the flow rate was established we then focussed on obtaining an optimal separation. The flow rate/back pressure ratio suggested that a column of 40 cm was feasible. Such a system was successfully constructed and the exact schematics can be found in Figure 1a. We found that we can load samples at up to 5 µL/min without issues. In the end we chose 10 minutes for loading the sample and the total 'overhead' to the gradient time was 22 minutes. The vented column design allows fast loading and equilibrium times which are simply not possible with LC designs where one uses a nanolitre flow rate pump. Our design is appropriate for short as well as long gradients. A number of gradients were tested. These 8 gradients had analysis times of 45, 60, 90, 120, 180, 240, 360 and 480 minutes, which correspond to 23, 38, 68, 98, 158, 218, 338 and 458 minutes gradient time. The final methods are summarized in Supporting Table S1. Note, the starting percentage of organic in the shortest gradients, 23 and 38 minutes, was lower than in the longer ones. On a similar theme, the amount of organic we chose for the end point decreased with increasing gradient time. These trends are in line with what is expected from a chromatographic system operating under a partitioning model. The gradients were performed by increasing the percentage of organic solvent. At the beginning of the analysis, when the mobile phase strength is low, the peptides will be partitioned wholly into the stationary phase at the head of the column. As the mobile phase strength increases, the peptides will begin to partition into the mobile phase and move along the column. At some point during elution, the peptides are wholly partitioned into the mobile phase, and will be moving with the same linear velocity as the mobile phase. Performance of these gradients with respect to peak width and peak capacity, was evaluated with an E. coli digest. In this case E. coli was used instead of BSA, as it is a more complex sample and it provides several peptides at all points in the gradient. The amount of sample injected was 250 ng for each method. Gradients from 68 to 458 minutes were performed. In Figure 3 the obtained chromatogram for the longest method is shown. All raw files are available in Tranche, see supplementary information.

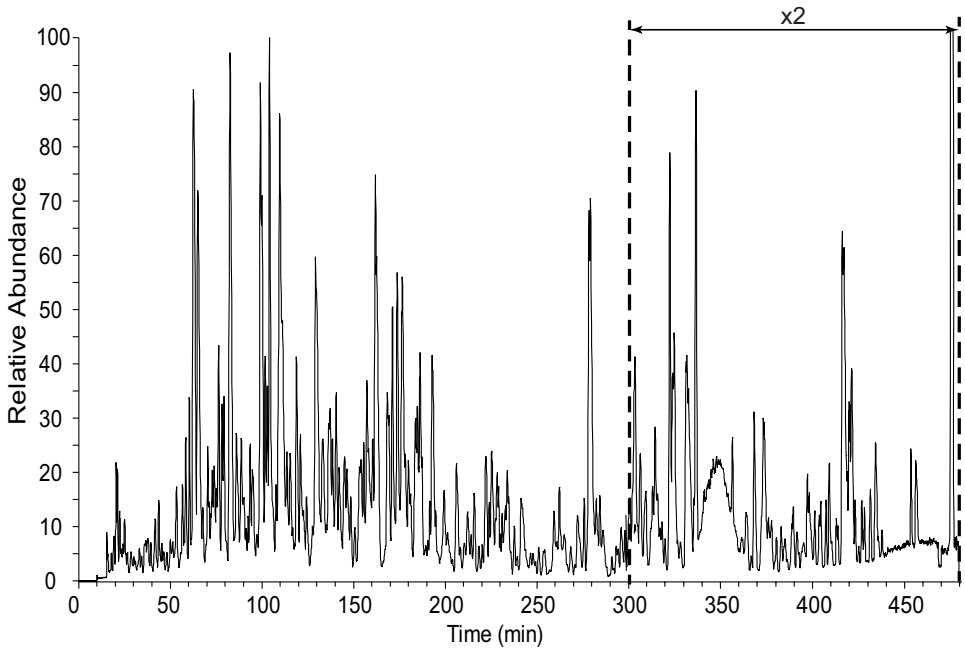


Figure 3. Base peak chromatogram for the analysis of E. coli digest with a gradient time of 458 minutes. The chromatogram was generated with the UHPLC (with a 40 cm analytical column) coupled to an Orbitrap Velos and injecting 250 ng of sample.

From these analyses, the width of a number of peaks across the gradient was calculated. We observed that peak widths increased with increasing retention. We noticed, unsurprisingly, the intensity of the peaks also influences its width. Hence, peaks of different intensities were also chosen. The width was calculated at 4σ which is defined as the peak width at base (13.4% of the peak height). The average peak widths obtained for 68, 158 and 458 minutes gradient times are 0.35, 0.66 and 1.10, respectively. The values for the peak width at 4σ, retention times and intensities of these peaks are summarized in the Supplementary Table S2.

The peak capacities were also calculated in order to evaluate the resolving power of our UHPLC configuration. The peak capacity was calculated as described in the Material and Methods section. As expected, the peak capacity increases as the analysis time increases. A peak capacity higher than 400 was reached for the longest

analysis time. Figure 4 contains a plot of peak capacity versus gradient time. It can be concluded that there is a non-linear increase, increasing the analysis time will not proportionally increase the peak capacity. Importantly, our design produces minor returns on separation once the gradient is over 4 hours. Several studies reported a similar trend for peak capacity in proteomics analysis.[28–30,42]

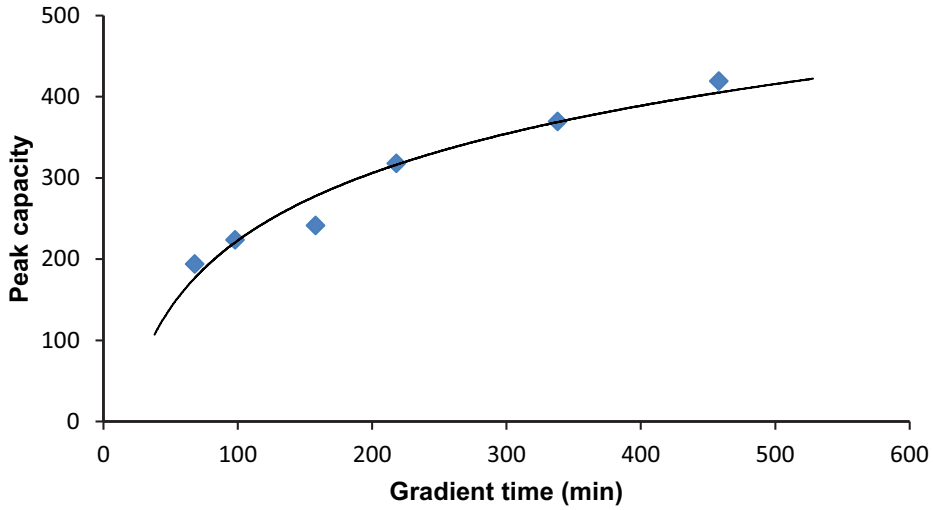


Figure 4. Trend of the peak capacity with respect to gradient times. A scatter plot of peak capacity against gradient time utilised for the 40 cm analytical column. A non-linear increase is observed.

Once that the system was optimized, we wanted to check if the UHPLC system improves LCMS performance when compared to our conventional 3 µm particle based HPLC system. The same methods, from 60 to 180 minutes analysis time were run using a tryptic lysate digestion of a human cancer cell line (HeLa cells) with the LC system coupled to an Orbitrap "Classic". Between 10 and 30% more was identified by the UHPLC system compared to the HPLC system (Supporting Figure S2). The difference in performance between the two designs increased with longer analysis times (Supplementary Tables S3-S4).

We then switched the LC system to be coupled to an Orbitrap Velos mass spectrometer to evaluate what is the maximum performance possible. The same HeLa tryptic lysate digestion was applied for the evaluation. We chose to increase the amount of material injected for the longest gradients applied in order to

maintain approximately the same peak signal intensity across analyses. From 23 up to 98 minutes gradient time, 250 ng of HeLa lysate was used and for the longer gradients 1 µg was injected. We evaluated the performance of the UHPLC system using a number of metrics: peptide spectrum matches (PSMs), unique peptides and protein groups (Figure 5). Similar to what was observed for the peak capacity, the number of PSMs identified increased with gradient time, but not linearly. We noticed that the number of candidates chosen per survey scan started to drop at the longer gradients i.e. the mass spectrometer did not 'observe' 10 new peaks in each survey scan. The number of unique peptides and protein groups also increased with longer gradient times (Figure 5 and Supplementary Tables 5-12). Nevertheless, over 4500 protein groups could be identified using 1 µg of cellular lysate at the longest gradient making our design competitive with dedicated commercial offerings.[28]

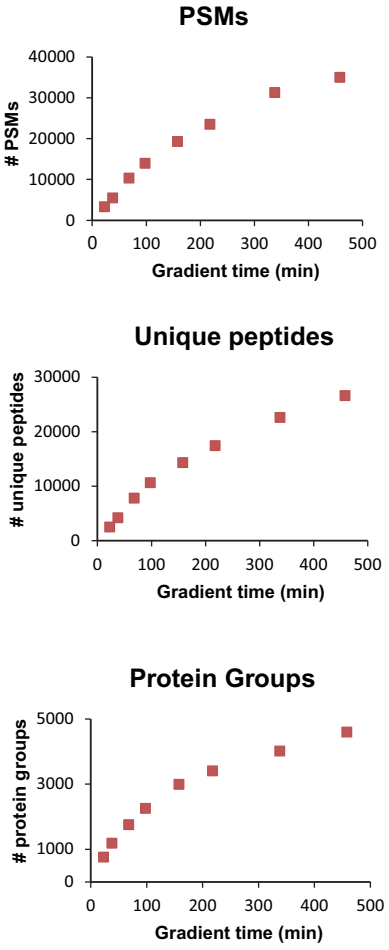


Figure 5. Identifications with the UHPLC coupled to the LTQ-Orbitrap Velos. Scatter plots of the number of PSMs, unique peptides and protein groups identified at the different gradient times. An increasing trend is shown although the values start plateauing.

Some recent mass spectrometers have sequencing speeds above 10 Hz.[6,43,44] We hypothesized that such fast instruments would not improve upon the Velos at long gradients (the Velos had already shown that it possessed sufficient speed for longer analyses), but may still be advantageous at shorter gradients. In order to test this hypothesis, the UHPLC was coupled to the AB Sciex TripleTOF (TT) mass spectrometer that can be operated at a sequencing speed of 20 Hz.[6] Using this system we tested the same gradients, from 23 to 458 minutes, with the same HeLa lysate digest in triplicate. The number of PSMs, unique peptides and protein groups were indeed found to be higher when compared to the Orbitrap Velos at the shortest gradients time (Figure 6 and Supplementary Tables 13-20). Most clearly, in a 23 minutes gradient, an average of 1435 protein groups were identified by the TripleTOF, almost doubling the amount identified by the Orbitrap Velos, demonstrating the power of UHPLC and fast MS sequencing speeds. When longer gradients are used the differences between both instruments decreased. In the longest gradients, 345 and 458 minutes, the Velos was practically the same as the TripleTOF.

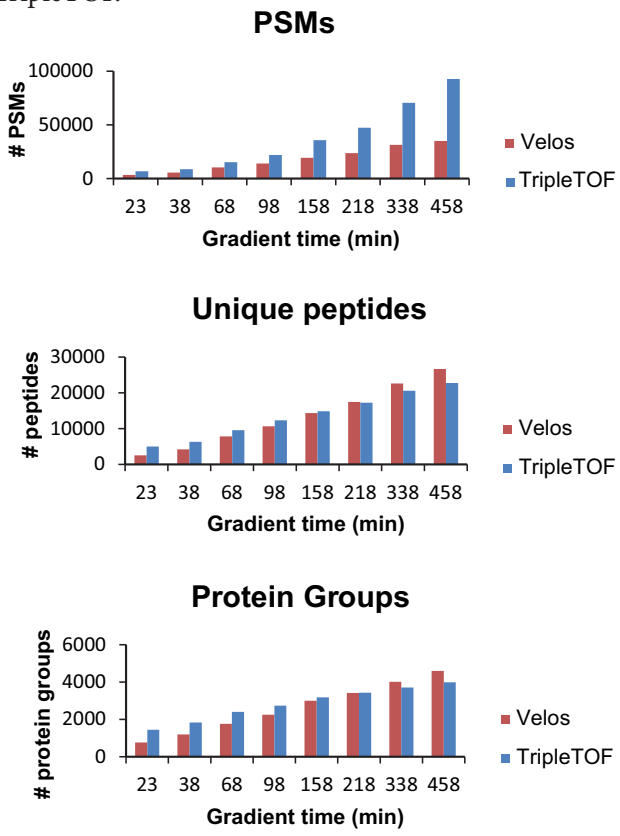


Figure 6. Identifications with the UHPLC coupled to the TripleTOF. Histograms of the number of PSMs, unique peptides and protein groups identified with the TripleTOF compared with the Orbitrap Velos at the different gradient times.

This indicates that when sequencing speed is not an issue both systems perform to a similar level. We did observe a little more from the Velos which, if real, could be due to automatic gain control (AGC) of the Velos. Such gain control will allow the Velos to accumulate sufficient ions for a successful sequencing event. It is generally accepted that short columns are appropriate for short gradients and long columns for long gradients. Our intention with our design was to allow long gradients to be performed while maintaining the possibility to perform quick analyses. It was pleasing to see that our short analyses were similar to those reported by Ishihama *et al.*[45] in which the authors used a 15 cm column for a 65 minutes gradient on the TripleTOF. Furthermore, the performance of our long gradient was similarly comparable to an equivalent gradient performed by the same authors where they used a 4 m monolithic column and the TripleTof for MS detection.

We compared the peptide and protein populations observed by the two mass spectrometers at the longest gradient time where both instruments were no longer troubled by complexity and never maxed out their sequencing speed (Figure 7). As can be seen in Figure 7a the overlap between two replicas performed on the TripleTOF appeared quite similar to the overlap of the analyses performed on the Velos and TripleTOF. This suggests what was observed by the two systems was essentially the same peptide populations which are dictated by the UHPLC system and the sample.

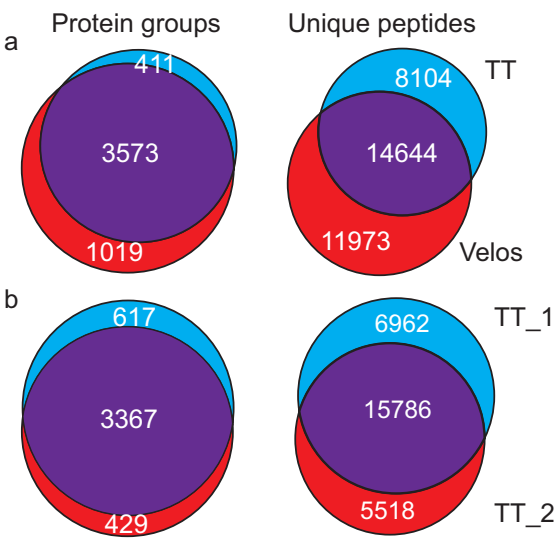


Figure 7. Comparison between analytical replicates on the same instrument and between two mass spectrometers. Figure 7a is a Venn diagram displaying the overlap between the protein groups (left) and peptides (right) identified by the Velos and the TripleTOF using the same digest and a 458 gradient. Figure 7b is a Venn diagram displaying the overlap between the protein groups (left) and peptides (right) identified by 2 TripleTOF analytical replicates using the same digest and a 458 gradient.

Conclusions

In this work we describe an in-house built UHPLC system designed to contain a passive split and a vented column. This configuration allows accurate gradient generation and fast loading of samples. Such a design generated peak capacities of over 400. In the longest gradient time, of 458 minutes, a total of 4592 protein groups and 26617 unique peptides could be identified from a human cancer cell lysate using only 1 µg of material. In the shortest gradient of 23 minutes, 1435 protein groups and 4836 unique peptides could be obtained. The short gradient benefited greatly from the use of the latest generation of mass spectrometers which possess sequencing speeds above 10 Hz. Taking into account all these results, we can conclude that the separation power of our system competitive to commercial offerings; however, our design is simple to implement and easy to use.

Acknowledgements

The authors acknowledge all members of the Heck-group. This work was in parts supported by the PRIME-XS project, Grant Agreement Number 262067, funded by the European Union Seventh Framework Program; The Netherlands Proteomics Centre (NPC), embedded in The Netherlands Genomics Initiative, and the Centre for Biomedical Genetics; and The Netherlands Organization for Scientific Research (NWO) with the VIDI Grant 700.10.429 for S.M. Dr. Bill Barber of Agilent Technologies for kindly providing the 1.8 µm Zorbax SB-C18 particles, and Dr. Javier Muñoz for helping with graphical design.

Supporting Information

Supplementary data set containing Supplementary Figures S1-S3 and Supplementary Tables S1-S20. Additional information is available as .raw and .msf file on Tranche at https://proteomecommons.org using the following hash:

wLY+buLAHzNqpY1/Ryvz2/wmzF52a+GX+csgRkk+YvgUFVgiLl84fxxiqAy/q++ZF1NtP5hYP9Jy9Uff6VtMO3q5tn0AAAAAAAAkNA==. The passphrase is: 5rawandmsffiles5.

References

1. Yates JR 3rd. Mass spectrometry as an emerging tool for systems biology. BioTechniques. 2004;36(6):917–919.
2. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.
3. Bisson, N., James, D.A., Ivosev, G., Tate, S.A., Bonner, R., Taylor, L. & Pawson, T. Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor. Nat. Biotechnol. 2011;29(7):653–658.
4. Olsen, J.V., Schwartz, J.C., Griep-Raming, J., Nielsen, M.L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E.R., Senko, M., Makarov, A., Mann, M. & Horning, S. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Molecular & Cellular Proteomics. 2009;8(12):2759–2769.
5. Second, T.P., Blethrow J.D., Schwartz J.C., Merrihew G.E., MacCoss M.J., Swaney D.L., Russell J.D., Coon J.J. & Zabrouskov V. Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. Anal. Chem. 2009;81(18):7757–7765.
6. Andrews GL, Simons BL, Young JB, Hawkridge AM, Muddiman DC. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). Anal. Chem. 2011;83(13):5442–5446.
7. Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S. & Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. Nat Meth. 2007;4(9):709–712.
8. Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(26):9528 –9533.
9. Giddings J.C. Two-dimensional separations: concept and promise. Anal. Chem. 1984;56(12):1258A–1270A.
10. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotech. 2001;19(3):242–247.
11. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res. 2003;2(1):43–50.
12. Chen J, Lee CS, Shen Y, Smith RD, Baehrecke EH. Integration of capillary isoelectric focusing with capillary reversed-phase liquid chromatography for two-dimensional proteomics separation. Electrophoresis. 2002;23(18):3143–3148.
13. Emmett MR, Caprioli RM. Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins. Journal of the American Society for Mass Spectrometry. 1994;5(7):605–613.
14. Chervet JP, Ursem M, Salzmann JP. Instrumental requirements for nanoscale liquid chromatography. Anal. Chem. 1996;68(9):1507–1512.
15. Masuda T, Sugiyama N, Tomita M, Ishihama Y. Microscale phosphoproteome analysis of 10,000 cells from human cancer cell lines. Anal. Chem. 2011;83(20):7698–7703.
16. Thakur SS, Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J. & Mann, M. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Molecular & Cellular Proteomics. 2011;10(8):M110.003699.
17. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J. Proteome Res. 2011;10(4):1785–1793.

18. Andrews GL, Shuford CM, Burnett JC Jr, Hawkridge AM, Muddiman DC. Coupling of a vented column with splitless nanoRPLC-ESI-MS for the improved separation and detection of brain natriuretic peptide-32 and its proteolytic peptides. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 2009;877(10):948–954.

19. Devreese B, Vanrobaeys F, Van Beeumen J. Automated nanoflow liquid chromatography/tandem mass spectrometric identification of proteins from Shewanella putrefaciens separated by two-dimensional polyacrylamide gel electrophoresis. Rapid Commun. Mass Spectrom. 2001;15(1):50–56.

20. Licklider LJ, Thoreen CC, Peng J, Gygi SP. Automation of Nanoscale Microcapillary Liquid Chromatography–Tandem Mass Spectrometry with a Vented Column. Anal. Chem. 2002;74(13):3076–3083.

21. Y. Ishihama; J. Rappsilber; J.S. Andersen; M. Mann. In: Proceedings of the 50th Annual Conference on Mass Spectrometry and Allied Topics, Orlando, FL,USA, June, 2002.

22. Yin H, Killeen, K., Brennen, R., Sobek, D., Werlich, M. & van de Goor, T. Microfluidic Chip for Peptide Analysis with an Integrated HPLC Column, Sample Enrichment Column, and Nanoelectrospray Tip. Anal. Chem. 2004;77(2):527–533.

23. Richard Gallagher, Raimund Peter, Kathryn Pickup, Kristin Samuelsson, Ian Wilson, Leonard Dillon & David Douce. In: Proceedings of the 59th Annual Conference on Mass Spectrometry and Allied Topics, Denver, CO,USA, June, 2011.

24. Jorgenson JW. Capillary liquid chromatography at ultrahigh pressures. Annu Rev Anal Chem (Palo Alto Calif). 2010;3:129–150.

25. MacNair JE, Patel KD, Jorgenson JW. Ultrahigh-Pressure Reversed-Phase Capillary Liquid Chromatography: Isocratic and Gradient Elution Using Columns Packed with 1.0-μm Particles. Anal. Chem. 1999;71(3):700–708.

26. MacNair JE, Lewis KC, Jorgenson JW. Ultrahigh-Pressure Reversed-Phase Liquid Chromatography in Packed Capillary Columns. Anal. Chem. 1997;69(6):983–989.

27. Swartz ME. UPLC™: An Introduction and Review. Journal of Liquid Chromatography & Related Technologies. 2005;28(7-8):1253–1263.

28. Köcher T, Swart R, Mechtler K. Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. Anal. Chem. 2011;83(7):2699–2704.

29. Iwasaki, M., Miwa, S., Ikegami, T., Tomita, M., Tanaka, N. & Ishihama, Y. One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the Escherichia coli proteome on a microarray scale. Anal. Chem. 2010;82(7):2616–2620.

30. Shen, Y., Zhang, R., Moore, R.J., Kim, J., Metz, T.O., Hixson, K.K., Zhao, R., Livesay, E.A., Udseth, H.R. & Smith R.D. Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000-1500 and capabilities in proteomics and metabolomics. Anal. Chem. 2005;77(10):3090–3100.

31. Luo Q, Shen, Y., Hixson, K.K., Zhao, R., Yang, F., Moore, R.J., Mottaz, H.M. & Smith, R.D. Preparation of 20-μm-i.d. Silica-Based Monolithic Columns and Their Performance for Proteomics Analyses. Anal. Chem. 2005;77(15):5028–5035.

32. Helbig AO, Gauci, S., Raijmajers, R., van Breukelen, B., Slijper, M., Mohammed, S. & Heck A.J.R. Profiling of N-Acetylated Protein Termini Provides In-depth Insights into the N-terminal Nature of the Proteome. Molecular & Cellular Proteomics. 2010;9(5):928–939.

33. Kirkland JJ, Destefano JJ. The art and science of forming packed analytical high-performance liquid chromatography columns. J Chromatogr A. 2006;1126(1-2):50–57.

34. Cortes HJ, Pfeiffer CD, Richter BE, Stevens TS. Porous ceramic bed supports for fused

silica packed capillary columns used in liquid chromatography. Journal of High Resolution Chromatography. 1987;10(8):446–448.

35. Savitski MM, Mathieson T, Becher I, Bantscheff M. H-Score, a Mass Accuracy Driven Rescoring Approach for Improved Peptide Identification in Modification Rich Samples. J. Proteome Res. 2010;9(11):5511–5516.

36. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Meth. 2007;4(11):923–925.

37. van den Toorn Henk WP, Muñoz J., Mohammed S., Raijmakers R., Heck AJ., van Breukelen B. RockerBox: analysis and filtering of massive proteomics search results. J. Proteome Res. 2011;10(3):1420–1424.

38. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods. 2007;4(3):207–214.

39. Gilar M, Daly AE, Kele M, Neue UD, Gebler JC. Implications of column peak capacity on the separation of complex peptide mixtures in single- and two-dimensional high-performance liquid chromatography. J Chromatogr A. 2004;1061(2):183–192.

40. Meiring HD, Van Der Heeft E, Ten Hove GJ, De Jong APJM. Nanoscale LC-MS(n): technical design and applications to peptide and protein analysis. Journal of Separation Science. 2002;25(9):557–568.

41. van Deemter JJ, Zuiderweg FJ, Klinkenberg A. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. Chemical Engineering Science. 1956;5(6):271–289.

42. Hennrich ML, Groenewold V, Kops GJPL, Heck AJR, Mohammed S. Improving depth in phosphoproteomics by using a strong cation exchange-weak anion exchange-reversed phase multidimensional separation approach. Anal. Chem. 2011;83(18):7137–7143.

43. Michalski A, Damoc, E., Hauschild, J.P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M. & Horning, S. Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. Molecular & Cellular Proteomics. 2011;10(9). Available at: http://www.mcponline.org/content/10/9/M111.011015.abstract. Accessed January 24, 2012.

44. Michalski A, Damoc, E., Lange, O., Denisov, E., Nolting, D., Müller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J.J., Cox, J., Horning, S., Mann, M. & Makarov, A. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Molecular & Cellular Proteomics. 2011. Available at: http://www.mcponline.org/content/early/2011/12/09/mcp.O111.013698.abstract. Accessed January 24, 2012.

45. Iwasaki M, Sugiyama N, Tanaka N, Ishihama Y. Human proteome analysis by using reversed phase monolithic silica capillary columns with enhanced sensitivity. Journal of Chromatography A. (0). Available at: http://www.sciencedirect.com/science/article/pii/S0021967311016013. Accessed December 20, 2011.

Supplementary Figures

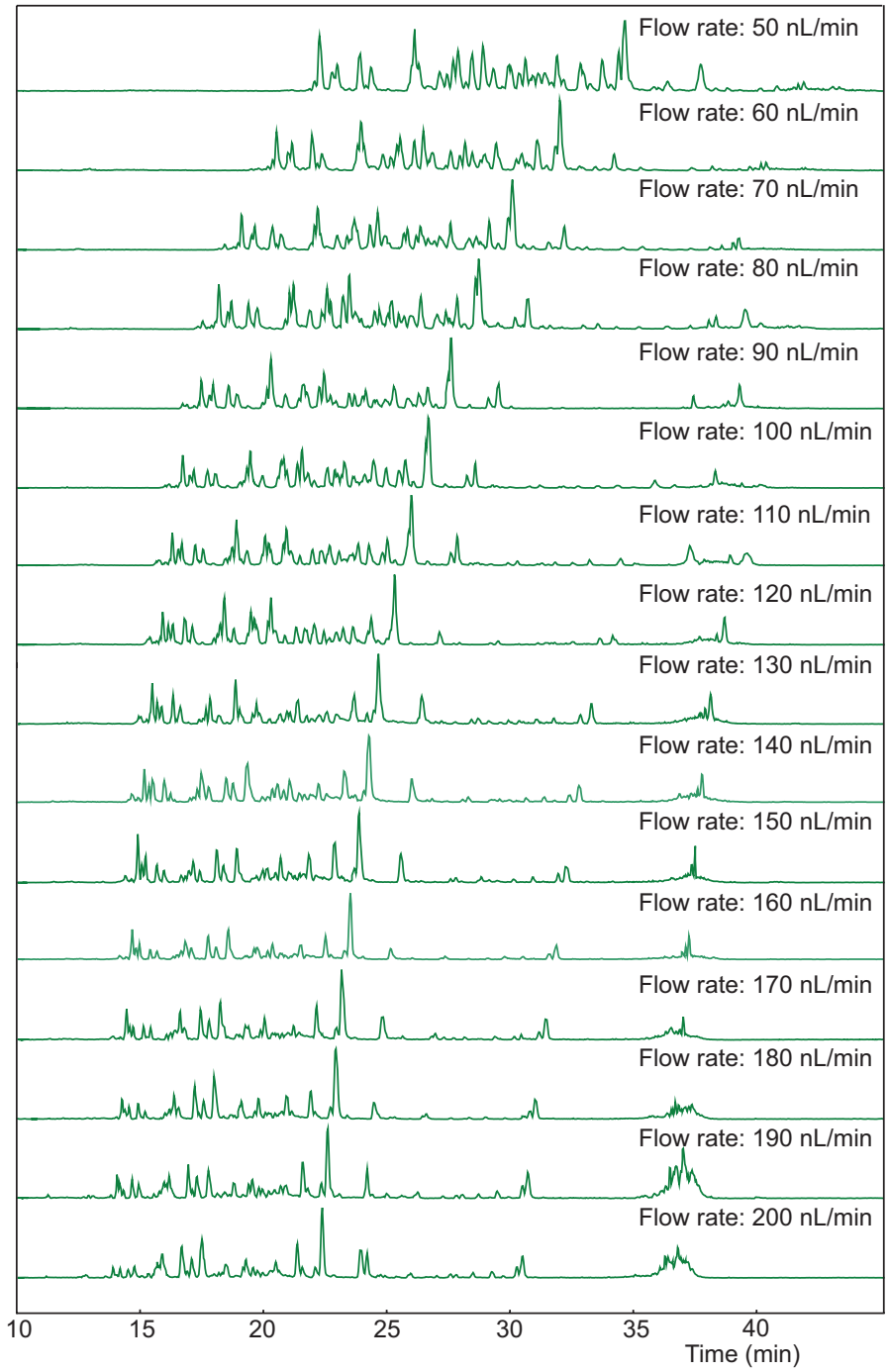


Figure S1: Base peak chromatograms at all the flow rates tested, from 50 to 200 nL/min. The first 10 minutes, which correspond to the loading of the sample, are not shown.

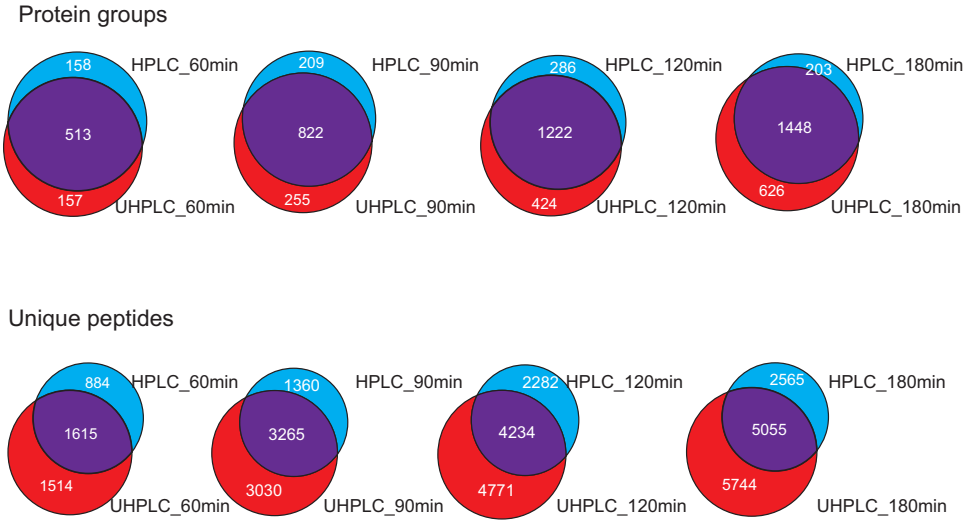


Figure S2: Comparison between the performance of the HPLC and the UHPLC. Venn diagrams displaying the overlap between the protein groups and peptides identified by the HPLC and UHPLC systems using the same digest from 60 to 180 minutes analysis times.

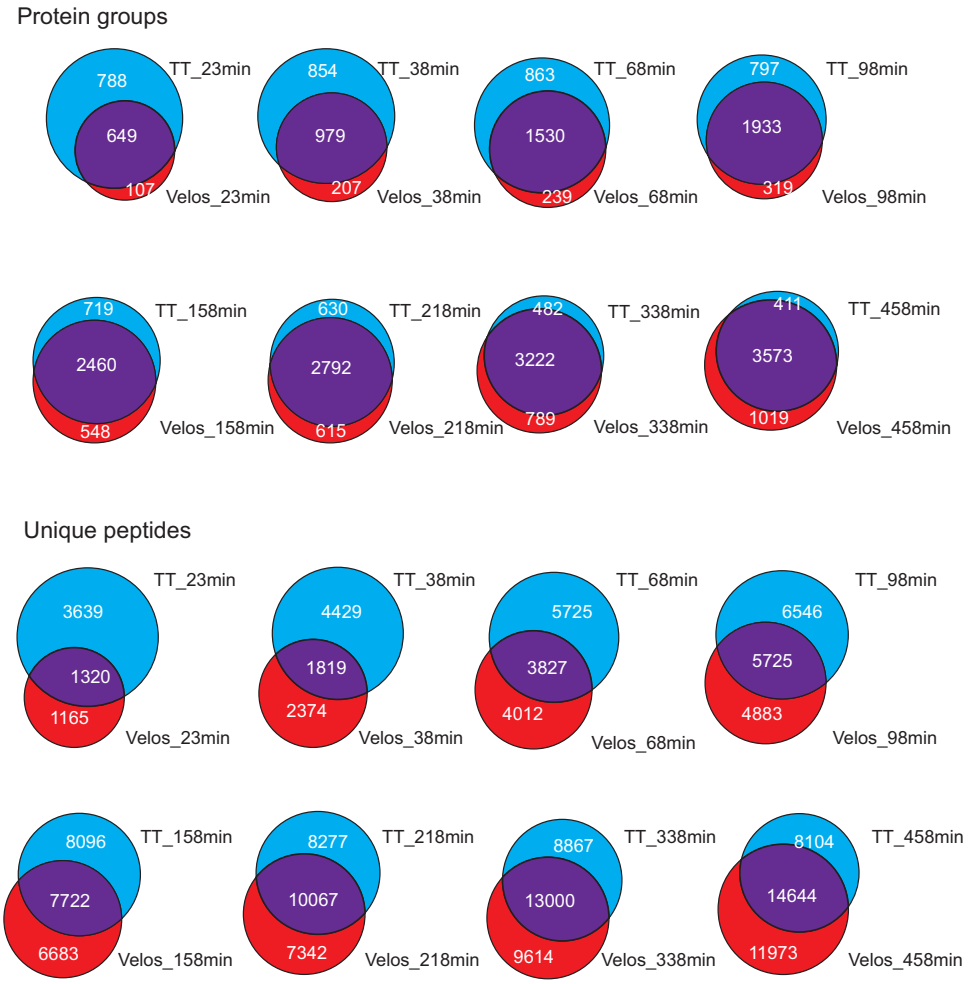
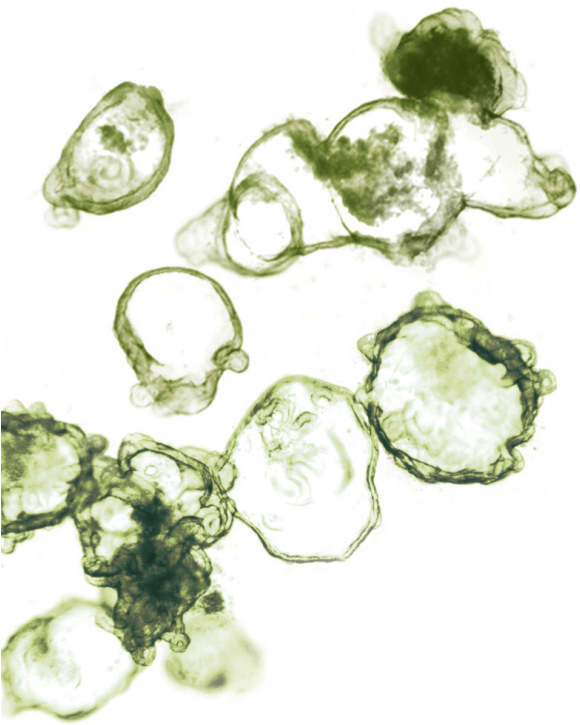


Figure S3: Comparison between the performance of the Orbitrap Velos and the TripleTOF. Venn diagrams displaying the overlap between the proteins groups and peptides identified by the UHPLC system coupled to the Orbitrap Velos and the TripleTOF for all the gradient times (from 23 to 458 minutes).

Chapter 3

Personalized proteome profiles of healthy and tumor human colon organoids reveal both individual diversity and basic features of colorectal cancer



Personalized proteome profiles of healthy and tumor human colon organoids reveal both individual diversity and basic features of colorectal cancer

Alba Cristobal[1,2], Henk W.P. van den Toorn[1,2], Marc van de Wetering[3], Hans Clevers[3,4], Albert J. R. Heck [1,2] and Shabaz Mohammed[1,2,5,6]

Cell reports, 3 January 2017

[1] Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

[2] Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

[3] Princess Maxima Center for pediatric oncology, Uppsalalaan 8, 3584 CT Utrecht, Netherlands

[4] Hubrecht Institute, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, Netherlands

[5] Department of Biochemistry, University of Oxford, New Biochemistry building, South Parks Road, Oxford, OX1 3QU Oxford, UK

[6] Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Mansfield Road, Oxford OX1 3TA, UK

Abstract

Diseases at the molecular level are complex and patient-dependent necessitating development of strategies that enable precision treatment to optimize clinical outcomes. Organoid technology has recently been shown to have the potential to recapitulate the in vivo characteristics of the original individual's tissue in a three dimensional in vitro culture system. Here, we present a quantitative mass spectrometry based proteomic analysis and a comparative transcriptomic analysis of human colorectal tumor and healthy organoids derived, in parallel, from seven patients. Although gene and protein signatures can be derived to distinguish the tumor organoid population from healthy organoids, our data clearly reveals that each patient possesses a distinct organoid signature at the proteomic level. We demonstrate that a personalized patient specific organoid proteome profile can be related to the diagnosis of a patient and with future development contribute to the generation of personalized therapies.

Introduction

Colorectal cancer (CRC) is one of the leading causes of death for adults. CRC usually develops from a benign precursor lesion, an adenoma, which is visible on the mucosal surface of the colon. The combination of mutational activation of oncogenes and mutational inactivation of tumor suppressor genes gradually leads to advanced adenoma and subsequently into an invasive cancer. Activation of the Wnt signaling pathway, through mutations in the APC gene[1] or β-catenin,[2] is regarded as the initiating event in colorectal cancer. The following steps include mutations of the KRAS[3] and PI3K oncogenes.[4] These events are continued by accumulation of additional mutations that inactivate the TGF-β response[5] and the p53 pathway.[6,7] Although the alterations often occur in this order, the accumulation of the changes is likely more important than the order of occurrence[8] The mechanisms underlying CRC development appear to be complex and heterogeneous and may entail patient-specific features.[9,10] To develop superior treatments, the molecular mechanisms underlying intestinal biology and colorectal cancer need to be thoroughly investigated. Previous work using model cell lines and organisms has

helped to provide an understanding of both intestinal biology and cancer.[11,12] The Wnt pathway is the dominant force behind this high proliferative activity and by studying the target genes of this signaling pathway it was found that some genes appeared to be restricted to a limited number of cells in the crypts. One of these genes, Lgr5, was identified as a stem cell marker in both the small intestine and the colon.[13,14] A better understanding of the conditions maintaining these adult stem cells[15–17] in combination with analysis of three-dimensional cultures[18,19] has allowed the establishment of organoids, a promising preclinical model. Organoids are ever-expanding three-dimensional epithelial structures with all the hallmarks of in vivo epithelial tissue.[20] Matrigel, which resembles the complex extracellular environment found in many tissues, is required for the culture of these organoids as well as a cocktail of growth factors (WNT, R-spondin1, Noggin and EGF).[21] The phenotype and karyotype of these organoids remain unchanged over time, thus representing a valuable model to study the biology of stem cells and for evaluating their contribution to tissue homeostasis and disease. Organoids have numerous valuable applications, and may be used to advance our knowledge of disease mechanisms, regenerative medicine and personalized precision medicine.[22,23] A detailed investigation, by us, into genotype-to-phenotype correlations was reported using several patient-derived organoids.[24] Furthermore, a comprehensive overview of the range of available organoids as well as all potential applications was recently published and emphasizes how rapidly the organoid field is developing.[25] The closer the used preclinical model is to the affected patients, the more it will be effective for translation to the patient.

Mass spectrometry (MS)-based proteomics has become the preferred choice for system-wide protein characterization thanks to advances in every stage of the proteomics workflow, although improvements at the MS instrumentation level deserves praise.[26,27] Proteomics has developed to the point at which it is an indispensable tool for molecular and cellular biology as well as for systems biology.[28–31] In the context of colorectal cancer several noteworthy proteomics studies have recently been published. Most relevant to the work described here, Mann and coworkers performed a comparative proteomics analysis of micro-dissected tissue from normal colon and adenocarcinoma.[32] Slebos and coworkers characterized APC-driven proteomic differences using specific stable cell lines.[33] Jimenez and coworkers have done extensive studies on colorectal cancer proteomics, such as in

a genomic context,[34] as well as studying the tissue secretome in order to identify CRC candidate biomarkers.[35] Aebersold and coworkers have recently proposed a set of non-invasive prognostic biomarkers by targeted proteomics experiments[36] based on circulating plasma proteins.[37] Roche and coworkers investigated the phosphoproteome of mouse xenografts in order to elucidate the contribution of the non-receptor tyrosine kinase SRC to colorectal cancer.[38]

Notwithstanding the fact that such studies provided insight into some of the mechanisms underlying CRC, they did not address the bio-variability that may be present in individuals. Given the huge prospect of patient derived organoids and given that differences in phenotype are closely related to changes in the proteome, we set out to evaluate personalized proteomes (adding transcriptome data for reference). Matched healthy and colorectal tumor-derived organoids of 7 individual patients with well-characterized, albeit distinct cancer-related mutational backgrounds, were analyzed. Next to a proof of concept on whether personalized proteome profiling can be achieved using patient derived organoids, an important aim of our study is to monitor how mutational differences in individual patients are reflected in each personalized proteome.

Experimental Procedures

Patient material background and sample preparation

Organoids were prepared as previously described[24] and the samples were named based on the previous publication in order to facilitate the correlation of the results. Organoids were lysed and digested with Lys-C and trypsin. The resulting peptides were chemically labelled using stable isotope dimethyl labelling as described before.[39] Prior to the MS analysis, samples were fractionated to reduce the complexity using a strong cation exchange (SCX) system. A detailed description of the sample preparation can be found in the Supplemental Experimental Procedures.

nanoUHPLC and mass spectrometry

The SCX fractions containing doubly and triply charged peptides (25 fractions from each SCX) were reconstituted in 10% formic acid and analyzed using a nano-

UHPLC Proxeon system (Easy-nLC 1000, Thermo Scientific, Odense, DK) coupled to different mass spectrometers. Different amount of sample was injected based on the SCX UV trace. The injected samples were first trapped on an in-house packed trap column (ReproSil-Pur C18-AQ, 3 μm (Dr. Maisch GmbH, Ammerbuch, Germany) 2 cm × 100 μm) before being separated in an in-house made analytical column (Zorbax SB-C18, 1.8 μm (Agilent Technologies, Baltimore, MD, USA) 50 cm × 50 μm) at a constant temperature of 40 degrees. The buffers used were: solvent A containing 0.1 M acetic acid in water and solvent B containing 0.1 M acetic acid in 80% acetonitrile.

Peptides were loaded into the trap column at a constant pressure of 800 bars with 30 μL of solvent A and chromatographically separated in the analytical column at a flow rate of 100 nL/min. LC methods of a duration of 120 min or 180 min were used: 7-30% solvent B within 91 or 151 min, 30-100% solvent B within 3 min, 100% solvent B for 2 min and 13 min 100% solvent A. The column effluent was directly connected to an in-house pulled and gold-coated fused silica needle (with a 5 μm o.d. tip).

For the LTQ-Orbitrap Elite (Thermo Scientific, Bremen, DE), a voltage of 1.7 kV was applied to the needle. The survey scan was recorded with a 350 to 1500 m/z scan range at a resolution of 30000 and for the MS2 the resolution was set to 7500. The 10 most intense precursors were selected for subsequent fragmentation in a data-dependent acquisition mode as described before[40] using CID and ETD-IT activation techniques.

For the Q-Exactive (Thermo Scientific, Bremen, DE), a voltage of 1.7 kV was applied to the needle. The survey scan was recorded with the same scan range as for the LTQ-Orbitrap Elite but at a resolution of 35,000 and for the MS2 the resolution was set to 17,500. The 20 most intense precursors were selected for subsequent fragmentation using HCD as the activation technique.

Data analysis

The raw files obtained from both instruments (Elite and Q-Exactive) were processed with Proteome Discoverer (PD, version 1.4, Thermo Scientific, Bremen, DE). The spectra were searched against a modified SwissProt human database (version 56.2) to which Matrigel proteins were added that were determined separately using a triplicate proteomics experiment. Searching was done with Mascot (version 2.5.1, Matrix Science, London, UK), with the following parameters: trypsin digestion

with up to two allowed missed cleavages, cysteine carbamidomethylation as fixed modification, oxidation of methionine, dimethyl labeling (light and intermediate) of lysine residues and the peptide N termini as dynamic modifications. Peptide tolerance was set to 50 ppm for both instruments, MS/MS tolerances were set to 0.6 Da for ETD-IT and 0.05 Da for ETD-FT and HCD. The quantification protocol for PD was Double dimethyl labeling, with a mass precision of 2 ppm for consecutive precursor mass scans.

The results were filtered using Percolator[41,42] to an FDR below 1%. We further only accepted peptides with at least 6 amino acid residues, a Mascot ion score of at least 20, and search engine rank 1.

All further analyses were performed in RStudio 0.98.1103, with R version 3.1.2.

To obtain a global picture of protein levels for the tumor versus the healthy organoids, we performed t-tests on the log2 transformed protein ratios, considering the different tissue donors as biological replicate measurements. Only proteins identified by more than two peptides were considered. Tests were performed whenever at least three measurements (i.e. patient ratios) were available for a protein. A detailed description of the complete analyses can be found in the Supplemental Experimental Procedures.

Results and Discussion

Background on the human colon organoids

As organoids may constitute an accurate representation of the patients from which they are derived, we set out to analyze their proteomes in order to decipher the differences between healthy and tumor colon at the protein level between a number of patients. Among the seven patients, three were female and four male, with an age range between 59 and 81 years. Biopsies were taken from the ascending colon or the sigmoid colon, and adjacent healthy tissue was also obtained in order to grow matched "healthy" organoids. Information about the patients can be found in the Table S1. Earlier deep sequencing of these samples had revealed that each of the patients had a distinct mutational profile as summarized in Figure S1.[24] The samples selected for the proteomic study were based on the mutation profile of

the most common tumor driver genes in CRC (APC, TP53, KRAS and PIK3CA). A representative population of the different CRC mutational profiles found in all the patients is present in the chosen organoids. These differences in mutational state may be reflected in differences in disease state and development, and possibly also differences in the proteomes. The samples were named based on the previous publication[24] in order to facilitate the correlation of the results.

Personalized proteomics profiling of human colon organoids

We applied a quantitative mass spectrometry based proteomics workflow to evaluate the protein expression in the organoids to derive so-called personalized proteomes. A schematic representation of this workflow is given in Figure 1. Following lysis, proteins from seven matched healthy and tumor-derived patient organoids were extracted. After enzymatic digestion the resulting peptides were 'dimethyl' labeled to allow accurate quantification,[39] in which peptides from the healthy organoids were labeled with light isotopes and peptides belonging to the tumor organoids with heavy isotopes. These organoids contain all the cell types of the tissue, constituting a somewhat complex (in terms of peptide population) sample. Therefore, after mixing the light and medium labeled peptides, a pre-fractionation step, by strong cation exchange (SCX), was included to reduce the sample complexity. Furthermore, in order to gain an in depth analysis of the proteome, two complementary mass spectrometers were utilized allowing two peptide identification strategies, namely an ETD enabled Elite and a Q-Exactive with HCD. SCX allows fractionation based on charge and so fractions could be a priori chosen that would be optimal for ETD or HCD based sequencing.

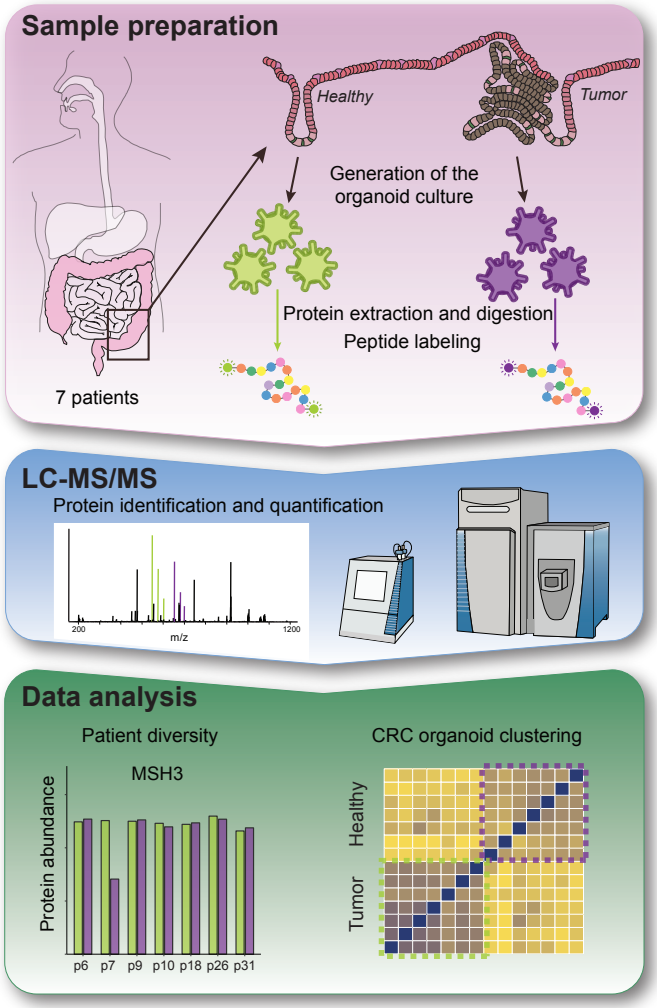


Figure 1. Experimental design. Surgically resected tissue was obtained from previously untreated colorectal cancer patients. Organoids from tumor and normal tissue were cultured by isolating the crypts and resuspended in Matrigel. The tumor and normal organoids were lysed and the proteins were extracted and subsequently digested into peptides. The resulting peptides were differentially labeled using the dimethyl labeling strategy, mixed and pre-fractionated by strong cation exchange chromatography and analyzed by LC-MS/MS. The MS results were analyzed using Proteome Discoverer and the statistical analysis was performed in R. From the data analysis a combination of patient diversity and general features were extracted as can be seen in the real data represented in the bottom part of the figure. See also Figure S1.

Some experimental challenges needed to be addressed in the proteome analysis,

such as those caused by missing values in one of the two labeling channels. In order to include proteins that are expressed only in either tumor or healthy organoids in our analysis ("on-off" proteins), we used missing value substitution, as described in detail in the Experimental Procedures section. Initial filtering provided approximately 7600 identified proteins for each matched (tumor - healthy) organoid pair and an average of 6340 proteins could be quantified per patient. Hierarchical clustering of the identified peptides across the patients was performed, revealing a patient centric pattern, as can be seen in Figure S2A. By applying an additional filter of considering only proteins with at least 2 peptides identified generated an average number of 5323 quantified proteins per patient. A "Significance B" analysis[43] was performed to discern the differentially expressed proteins in each patient. On average, approximately 400 proteins were assigned as differentially expressed in each individual patient. A representation of these data can be found in Figure S2B and a detailed description of the proteins can be found in Table S2. Additionally, we decided to check how consistently these proteins were changing across all patients. From Figure 2, in which the differentially expressed proteins of each patient are highlighted in the other patients, the high diversity present in these CRC patients can be clearly observed. There are substantial differences in the protein populations that are significantly different between healthy and tumor organoids of each patient.

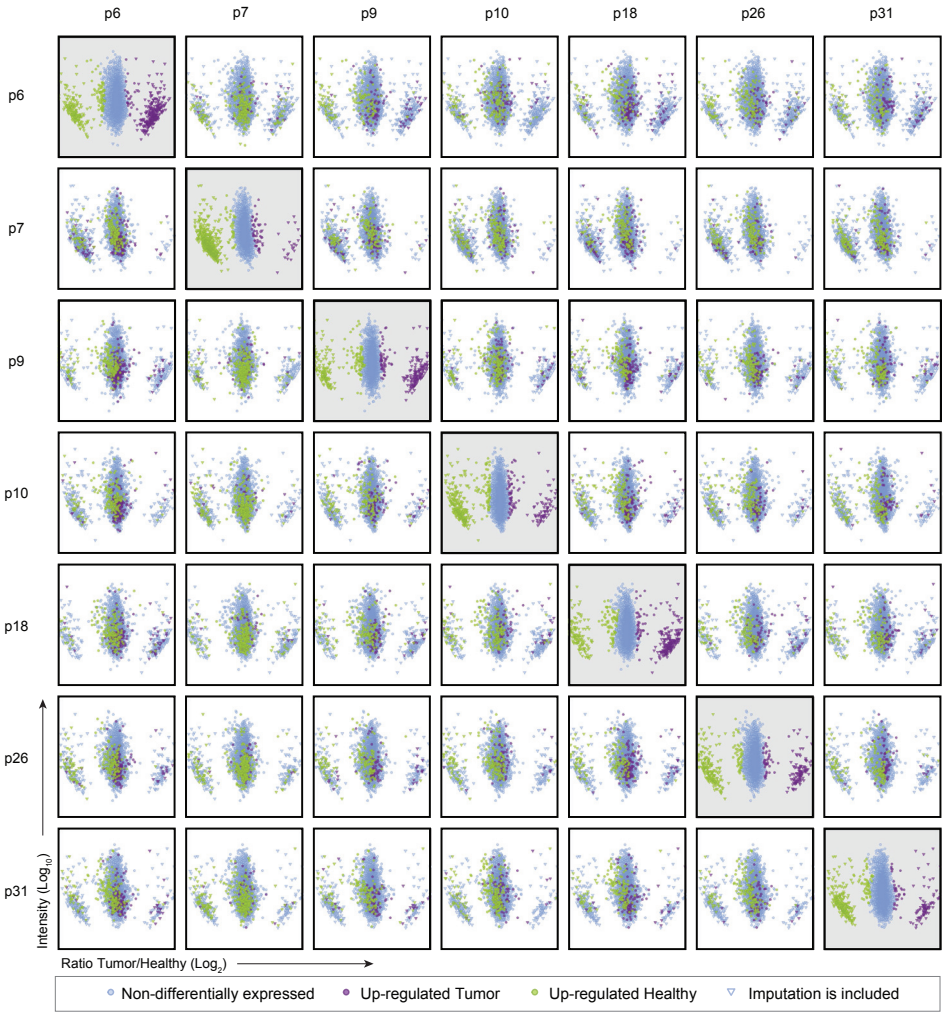


Figure 2. Personalized proteomics profiles. MA plots of the seven patients, in which the protein ratio (log_2) is plotted against its intensity (log_{10}). To visualize the heterogeneity of the samples, "Significance B" analysis was performed on the ratios of each of the patients. See also Figure S2.

Additionally, we re-analyzed a quantitative microarray experiment performed on organoids obtained from the same patients. A total of 21681 gene mRNA transcripts were quantified across the 7 patients. These 21681 transcript genes translate to 16861 proteins. Following a similar statistical treatment as used for the proteomics data,

an average of 130 transcripts were significantly changing at the transcriptome level across the 7 patients (Figure S2C). From the, on average, 400 significantly changing proteins at the proteome level, less than 80 were recapitulated in the respective transcriptomes, demonstrating the additional perspective provided by the study of the proteome (Table S2). This limited overlap is not unusual considering the abundance of post-transcriptional and -translational regulatory mechanisms and has been observed in several studies.[44,45]

Organoids from individuals exhibit distinct personalized proteomics profiles

- *Microsatellite instability*

The DNA mismatch repair protein MSH3 showed a strong down-regulation in the tumor organoid of patient 7 (female, 81 years old, whereby the biopsy was taken from the ascending colon) while remaining unchanged in all other patients (Figure 3A). MSH3 is a component of the post-replicate DNA mismatch repair system (MMR) which maintains genomic stability. The deficiency of this protein has already been related to colorectal cancer and a loss of its expression has been shown to be frequent in MLH1 deficient colorectal cancers.[46,47] MLH1, one of the four mismatch repair genes,[48,49] was also found down-regulated exclusively for this patient in our dataset and furthermore, a strong down-regulation was present at the transcriptome level as highlighted by van de Wetering *et al.*[24] Moreover, we were pleased to find a frame shift deletion on MSH3 for the patient 7 in the mutational study, which confirmed our findings from the proteomics dataset (as summarized in the Figure S1).

Genomic instability is a characteristic of all intestinal malignancies. As we described vide supra the majority of the colorectal cancers follow the classical adenoma-adenocarcinoma sequence and display chromosomal instability. Another form of genomic instability contributing to colorectal cancer is microsatellite instability (MSI). Approximately 15% of the colorectal cancer cases present high-frequency MSI, a hypermutatable phenotype caused by defects in the DNA mismatch repair (MMR) system. Apart from MSH3 and MLH1, there are other proteins showing

a strong down-regulation specifically in this patient, such as PMS2, RPL22 and MARCKS that all have been linked to microsatellite instability. Expression data representing these proteins can be found in Figure S3A. For instance, PMS2, together with MLH1 is one of the four mismatch repair genes, and it has been shown to be mutated in hereditary non-polyposis colorectal cancer, leading to the production of an abnormally short or inactive PMS2 protein.[50,51] In our dataset, this protein was quantified in 4 patients, and it only shows a down-regulation in patient 7. Another example is RPL22, for which a high mutation frequency of this gene has been reported in microsatellite instable colorectal cancer.[52] We were able to quantify RPL22 in all of the patients, finding exclusively a strong down-regulation in patient 7. This is in agreement with a frame shift deletion found in the sequencing data only for patient 7. MARCKS is another protein only down-regulated in the tumor organoid of patient 7, while in the other patients no change between healthy and tumor organoid was observed. MARCKS has already been identified in colorectal cancer as a major target of inactivation through coding microsatellite instability.[53,54] The valuable insight of the proteomics data can be clearly seen in this case, in which these 4 proteins were found significantly changing at the proteome level only for this patient, but only MARCKS appears to be significantly changing at the transcriptome level.

Due to the difference in the mutational pathway between the two genomic instability types, distinct and diverging clinical outcomes can be obtained. Patients with MSI are associated with a longer survival and improved clinical outcome than those with microsatellite stability (MSS).[55] Distinct responses to cancer drugs is also observed between MSI and MSS which makes recognition of molecular subtypes of colorectal cancer essential for future personalized treatment.[56]

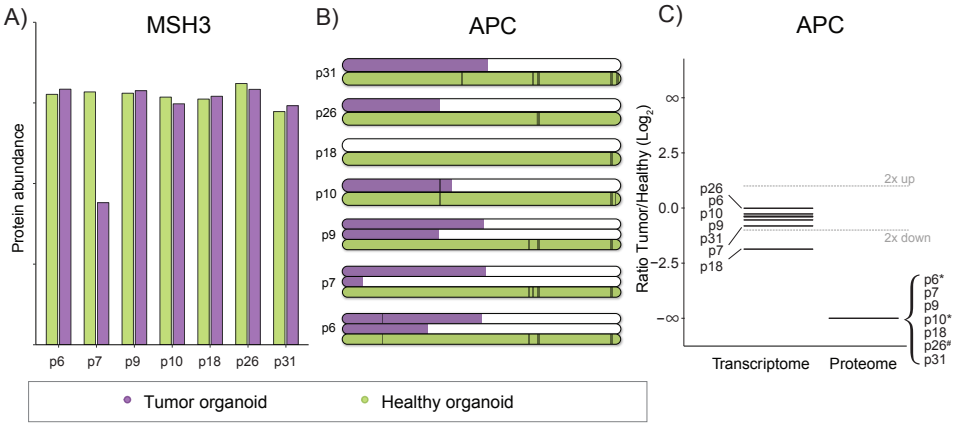


Figure 3. Overview of the patient diversity. (A) Plot representing the protein abundance for MSH3 in the healthy and tumor organoids across the different patients. Green is used for the healthy organoid and purple is used for the tumor one. A significant down-regulation is observed exclusively in the tumor organoid from p7. (B) Patient expression values of APC. The mutation pattern of the protein APC based on previously published exome sequencing data[24] for the different patients is represented with the quantified peptides highlighted with black bars. The upper purple line(s) represents the protein(s) found in the tumor organoid and the green line the protein found in the healthy organoid. The length of the purple line depends on in frame shift mutations or non-sense mutations. (C) The ratio between the tumor and healthy organoids found at the transcriptome and proteome level for the different patients is shown. *To be in line with the rest of our analyses a ratio is determined for "on-off" situations. #used to include an expression value based only in one peptide. See also Figure S3.

- *The Wnt signaling pathway in patients' healthy and tumor organoids*

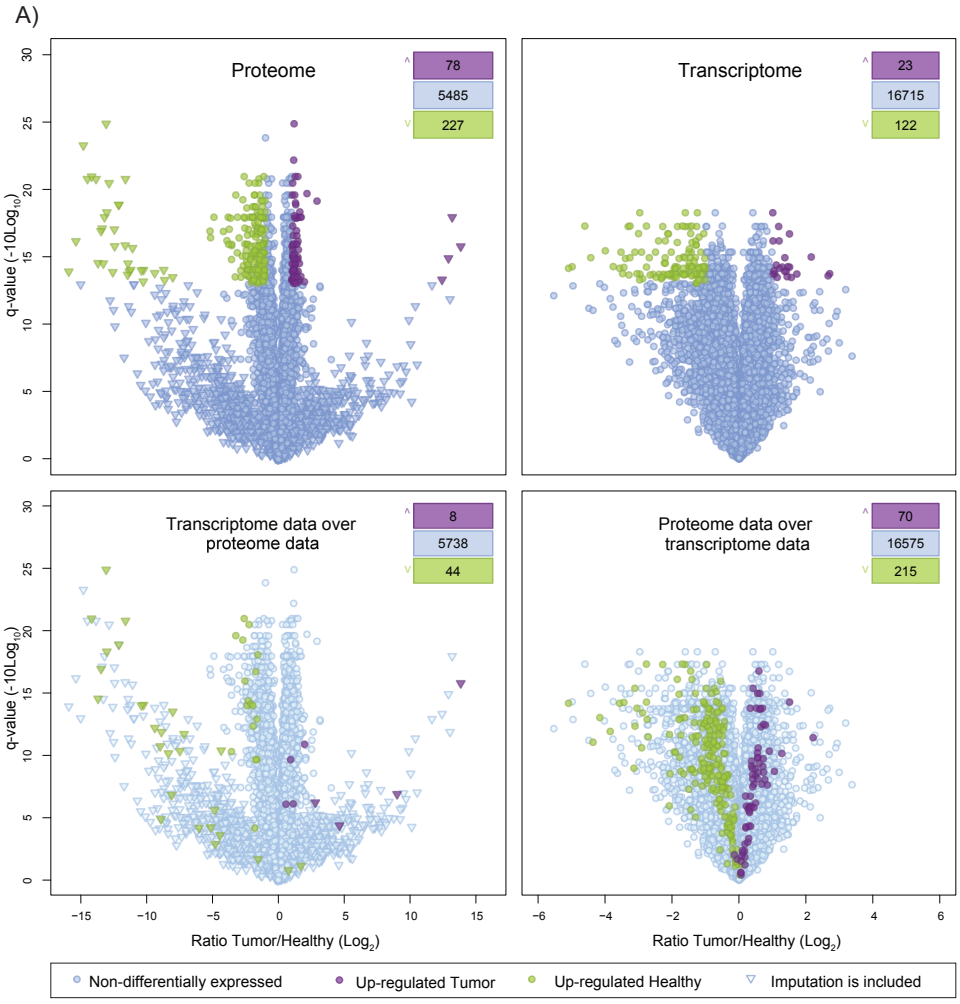
Activation of the Wnt signaling pathway is a principal event in the development of colorectal cancer. Constitutively activated β-catenin signaling, due to APC deficiency or β-catenin mutations that prevent its degradation, leads to unbalanced stem cell generation, maintaining the cells in a progenitor state [13]. We detected APC in all 7 patients, with peptides associated (almost exclusively) with the healthy organoids leading to extreme "on-off" ratios for this protein. The vast majority of detectable peptides originated from the C-terminus part which is truncated in the tumor organoids (Figure 3B). In 5 of the 7 patients APC was only detected in the healthy organoids due to C-terminal peptides being exclusively observed which would be missing in the truncated APC present in the tumor organoids. In one of the

patient's a single peptide was detected in a part of the APC that should be present in both healthy and tumor organoids yet only signal was detected in the healthy. Since we mix the samples before MS detection we generate no bias between the healthy and tumor organoid with respect to protein detection. Nevertheless, the result requires both the transcript and protein information for an accurate interpretation of the data and highlights the need for a proteogenomic approach. A representation of the expression values found at the proteome and transcriptome level for this protein are displayed in Figure 3C. Mutational inactivation of APC leads to the inappropriate stabilization of β-catenin.[57] β-catenin (CTNNB1) was detected in all 7 patients (Figure S3B); however, it only showed a light up-regulation in one of the patients (patient 10), while in the rest of the patients there was no difference between the healthy and tumor organoids. Stabilized β-catenin translocates to the cell nucleus and interacts with transcription factors of the TCF/LEF family, leading to the transcription of Wnt target genes. Groucho/TLE proteins are transcriptional co-repressors that interact with proteins from the LEF/TCF family in the absence of a Wnt signal (and therefore β-catenin), keeping the pathway in an inactive state.[58,59] However, in the presence of Wnt signal and if TLE1 is absent, we hypothesize that when β-catenin does not have to compete with TLE1, the signal might be stronger. TLE1 was quantified in our dataset in 6 out of 7 patients with a strong down-regulation in most of the tumor organoids, which is in concordance with the transcriptomics experiment. A representation of the expression values found at the proteome and transcriptome level can be found in Figure S3B.

Above, we largely focused on the patient specific differences observed in the proteome of healthy and tumor organoids, describing patient-specific features. These relevant patient-specific differences were not always present at the transcriptome level. To demonstrate the additional insight provided by proteomics we performed immunohistochemistry (IHC) for the proteins MUC5A, MGMT, AMACR and PHLDB2 on the organoids for patients p9, p18 and p26. We chose these proteins and patients as they only show expression changes at the protein level. The IHC confirmed our findings (Figure S3C), demonstrating the value of proteomics for any biological material.

Establishing a human CRC organoid expression profile by quantitative proteomics and transcriptomics

Across all seven CRC patients we were able to determine expression levels of 8510 proteins. For the subsequent analysis, proteins with at least 2 peptides identified and quantified in at least 3 out of 7 patients were used, which diminishes the total number to 5790 proteins (Figure S4A). Certain proteins may present an "on-off" situation between tumor and healthy organoids. Since these proteins may be (most) important in the tumor state, we used imputed values from Proteome Discoverer, albeit only when necessary to obtain a ratio. We found a limited number of changes, constituting 5.3% of the detected proteins, which correspond to 78 proteins up-regulated in the tumor organoids and 227 down-regulated. A representation of these data can be seen in the upper left panel of Figure 4A. On closer inspection of the quantified proteins it is pleasing to note that several proteins previously described to be specific for colorectal cancer, such as PALM3 and GPR56[32] are also observed in our data. The two proteins showed an up-regulation in both datasets, namely in the micro-dissected tumor tissue, and in our tumor-organoids. Moreover, when comparing our whole dataset obtained from the colon organoids with a recent proteomics study on colorectal cancer from The Cancer Genome Atlas (TCGA) using human tumor tissue from 90 patients,[44] a high overlap is obtained (see Figure S4B). The TCGA data allowed the authors to determine 5 CRC subtypes based on the expression level of approx. 1600 proteins. Although our choice of patients was based on the mutation profile of the most common tumor driver genes in CRC (APC, TP53, KRAS and PIK3CA) we organized our data according to these subtypes and found a broad agreement (Table S3). Taking into consideration the overlap obtained between our study and previous studies about the characterization of the organoids,[24] we can consider organoids as a promising proteomic model with multiple applications.


A)
Proteome
78
5485
227
Transcriptome
23
16715
122
Transcriptome data over proteome data
8
5738
44
Proteome data over transcriptome data
70
16575
215
q-value (-10Log10)
Ratio Tumor/Healthy (Log2)
Non-differentially expressed
Up-regulated Tumor
Up-regulated Healthy
Imputation is included

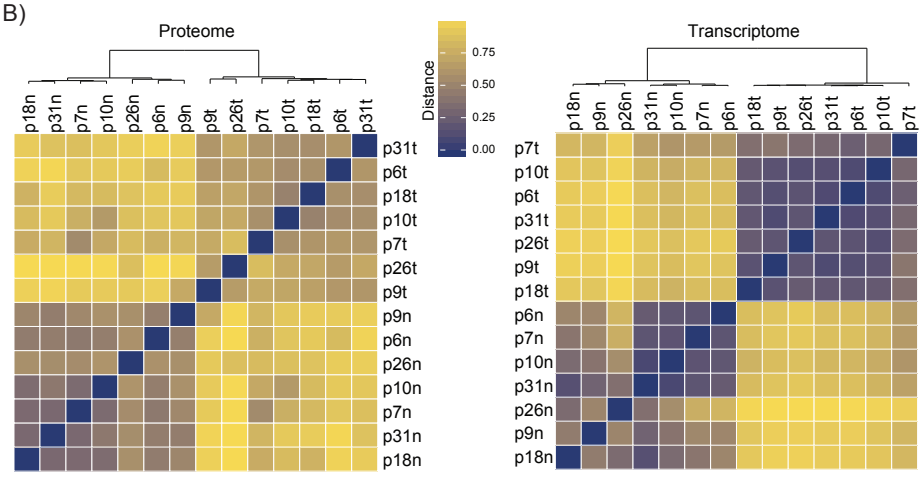

B)
Proteome
Transcriptome
Distance
0.75
0.50
0.25
0.00
p18n p31n p7n p10n p26n p6n p9n p9t p26t p7t p10t p18t p6t p31t
p31t p6t p18t p10t p7t p26t p9t p9n p6n p26n p10n p7n p31n p18n
p18n p9n p26n p31n p10n p7n p6n p18t p9t p26t p31t p6t p10t p7t
p7t p10t p6t p31t p26t p9t p18t p6n p7n p10n p31n p26n p9n p18n

Figure 4. CRC organoids overview. (A) Volcano plots in which the average protein ratio (log_2) against the q-value ($-10log_{10}$) are plotted. In the upper panel the data obtained from the proteomics experiment is represented on the left and data from the transcriptomics experiment on the right. From the proteomics data a total of 78 up-regulated and 227 down-regulated proteins are found. Regarding the transcriptome data a total of 23 up-regulated and 122 down-regulated transcripts are found. In the lower panel a direct representation of the differentially expressed proteins from the transcriptomic analysis in the proteomic dataset are presented on the left and vice versa on the right panel. Although q-values do not correlate at all, ratios do correspond well between the two technologies. (B) Correlation heat maps of the differentially expressed proteins from the proteomics (left) and transcriptomics (right) datasets. Tumor and healthy organoids are separated by hierarchical clustering. See also Figure S4.

Among the 305 significantly changing proteins found in our dataset, several proteins have already been related to CRC. MYO1C was found down-regulated in the tumor organoids and it has been recently suggested as a tumor suppressor gene candidate related to Tp53.[60] Desmocollin-2 has been previously reported as being decreased or absent in CRC studies[61,62] and is also down-regulated in our dataset. UBE2C, also known as UBCH10 is essential for cell progression and it has been related to several cancers[63] and is observed with an overexpression in the tumor organoids. Using CRC cell lines, depletion of this protein resulted in suppression of cellular growth, whereas overexpression promoted cell proliferation and oncogenic cellular growth.[64] The possibilities of inhibiting UBE2C for the treatment of CRC have also been studied.[65] HspBP1 is a co-chaperone that binds to and inhibits the activity of Hsp70 and is also found to be overexpressed in the tumor organoids. High expression values of Hsp70 had been previously reported for several tumors and the level of HspBP1 has also been reported as elevated in some tumors.[66]

Following a similar statistical treatment as that used for the proteomics data, 23 transcripts were found to be up-regulated and 122 were down-regulated in the tumor versus the healthy organoids consistently over the 7 patients. A representation of these data is summarized in the upper right panel of Figure 4A and a detailed description of the proteins can be found in Table S3.

Next we used hierarchical clustering for the differentially expressed proteins of both techniques, to classify all organoids. Figure 4B shows the correlation heat maps at the proteome and the transcriptome level, respectively and in both cases the samples

cluster by tumor versus healthy. For 5320 proteins quantified in our proteome data we obtained corresponding data at the transcriptome level. Focusing first on the earlier defined differentially expressed proteins, only 22 proteins showed similar behavior at the mRNA level, as can be seen in Figure 5.

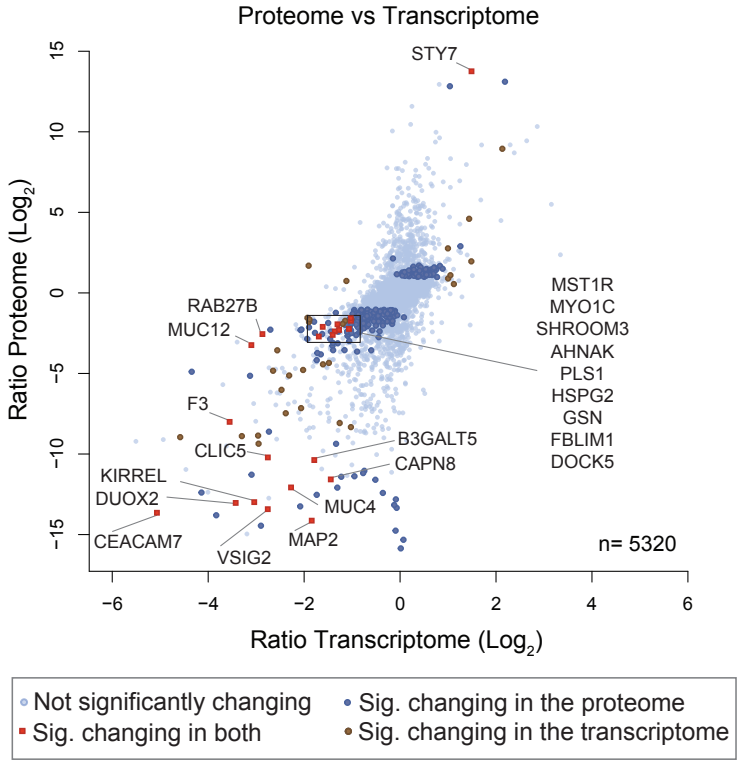


Figure 5. Overlap between the proteome and transcriptome. Comparison of expression ratios between the proteins and their corresponding transcripts (5320 protein-transcript pairs). Focusing on the differential proteins, 22 proteins showed alike behavior at the transcript level and only one of them is up-regulated while the rest are down-regulated in both datasets. See also Figure S5.

Out of these 22 proteins only one shows an up-regulation in both datasets, Synaptotagmin-7 (SYT7). SYT7 is also known as Prostate cancer-associated protein 7;[67] however, a direct relation with CRC has not yet been shown. Among the other 21 down-regulated proteins were CEAM7, Mucin12 and Gelsolin. Carcinoembryonic antigen-related cell adhesion molecule 7, CEAM7, is also known as Carcinoembryonic antigen CGM2 and is a member of the CEA family.

CEA is a classical tumor marker for several types of cancer.[68] CEAM7 is known to be expressed in normal colon and rectum and down-regulated in colorectal adenocarcinomas,[69,70] which is in agreement with our results. Mucin12 is a transmembrane mucin which we found down-regulated in our dataset which is in concordance with previous publications.[71–73] Gelsolin (GSN), a protein with structural molecule activity, is a multifunctional actin binding protein that is found down-regulated in CRC. It has been suggested that the loss of this protein is a general mechanism associated with the transition from noninvasive stage to the late invasive stage of the neoplastic process.[74] A detailed description of the overlapping proteins can be found in Table S4.

We suspected the limited overlap between both techniques (Figure S5A) maybe related to our filtering criteria being too stringent. To test our hypothesis, we decided to color-annotate all differentially expressed proteins on the measured transcript data and vice versa. A consistent trend in expression among the proteome and mRNA datasets is observed; however, at the defined stringency level the correlation is not high, as displayed in the lower panels of Figure 4A. In order to broaden the stringency, without being overly tolerant, we decided to include the proteins found differentially expressed in one of the techniques that could be observed changing in the same direction using the other technique. A Venn diagram representing this profile with the overlapping proteins highlighted, constituting the CRC organoid profile, in red is shown in Figure S5B. A total of 308 proteins are included in this dataset. The main molecular function of these proteins can be divided into proteins with catalytic activity (35%), binding proteins (28%) and proteins with structural molecule activity (12%). A pie chart with the representation of the molecular function obtained in Panther can be found in Figure S5C. Among the binding proteins, Gastrotropin (FABP6) is found with an overexpression in the tumor organoids. Gastrotropin, also called Fatty Acid-Binding Protein 6, has already been reported as being related to colon carcinogenesis, presenting a higher expression value in cancer tissue compared to normal tissue.[75] Proteins related to other cancers are also found in this dataset. For example, BCAS1 which has been described as a strong candidate oncogene for breast cancer.[76] Surprisingly, compared to breast cancer in which up-regulation is observed for this protein, an opposite trend is found in colorectal cancer.[77] The same trend is found in our study, once again organoids recapitulating observations made with biopsies.

Conclusion

Clinical models, such as cultured cancer cell lines and animal models, have been used for many years to investigate the mechanisms of diseases and design promising treatment strategies. It is well known that these model systems all have their limitations and, ideally, studies should be performed, in vivo, on patient material; however, this is currently not always realistic. Recently, a promising research technology has been developed, so-called organoids, that allows the continuous culture of a variety of healthy and diseased human tissues. These organoids provide a potentially unlimited supply of well characterized patient material, circumventing some of the limitations of current models (i.e. lack of genetically stable cell lines and need of extensive colonies of animals). Several studies have already highlighted the potential for applications of organoid technology in the study of human cancer.[24,78–82] The amenability of this technology for genomic and functional analysis as well as for high-throughput drug screening is beginning to be explored. Here, we set out to make a proteomic characterization of healthy and tumor colon organoids. Although sample amount is a genuine limiting factor in proteomics, we demonstrate here the feasibility of generating personalized human proteome profiles of human healthy and tumor colon derived organoids. Fourteen organoid samples (7 tumor/7 healthy) from seven patients were quantitatively analyzed. System-wide data analysis revealed strong patient specific features. A single patient with a hypermutated phenotype in colorectal cancer characterized by microsatellite instability stood clearly out evidenced by several patient specific proteins, which have already been linked to microsatellite instability. Patients with microsatellite instability have a different clinical outcome compared to those exhibiting chromosomal instability (the most frequently occurring genomic instability in colorectal cancer), illustrating why personalized treatment is essential for the survival of individual patients.

Although we observed that patient proteomes exhibit a clear personalized profile, we were still able to extract some tumor-specific characteristics across all patients from the proteomics data by focusing, for instance, on one of the most frequently mutated signaling pathways in colorectal cancer, Wnt signaling. Expression levels of several proteins belonging to this pathway changed significantly and consistently, when tumor- and healthy organoid proteomes were compared, also consistent with

several of these proteins previously being reported as colorectal cancer biomarker proteins.

The present study further highlights the potential of organoids as model systems for personalized cancer research, demonstrating that advanced proteomics analysis is nowadays feasible and meaningful. A better categorization of CRC would be possible by further extending the work presented in this paper by investigating larger number of patients, and or the proteome profiles of organoids when treated by anti-cancer drugs. With proteomics becoming ever more sensitive and faster, these goals may be achieved already in the near future, expanding the technical toolbox for optimizing personalized cancer treatment.

Author contributions

A.C., M. vd W., H.C., A.J.R.H. and S.M. designed the study. A.C. performed all proteomics experiments. M. vd W. prepared all organoid samples, and carried out the transcriptome analysis. H. vd T. carried out and supervised the computational analysis. All authors helped interpret the results and wrote together the manuscript.

Accession numbers

The raw data and result files have been deposited to the ProteomeXchange Consortium via the PRIDE[83] partner repository with the dataset identifier PXD004149. Username: reviewer54519@ebi.ac.uk and password: 6Gl8skWk

Acknowledgements

The authors acknowledge Harm Post for technical support on the SCX. This work was supported by the Roadmap Initiative Proteins@Work (project number 184.032.201), funded by The Netherlands Organisation for Scientific Research (NWO) and with the VIDI grant 700.10.429 for S. M. M. vd W. is supported by Stichting Virtutis Opus and Stichting Vrienden van het Hubrecht. We thank Harry Begthel for his invaluable help with the immunohistochemistry.

References

1. Powell, S. M. *et al.* APC mutations occur early during colorectal tumorigenesis. Nature 359, 235–7 (1992).
2. Morin, P. J. *et al.* Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. Science 275, 1787–90 (1997).
3. Bos, J. L. *et al.* Prevalence of ras gene mutations in human colorectal cancers. Nature 327, 293–7 (1987).
4. Bardelli, A. Mutational Analysis of the Tyrosine Kinome in Colorectal Cancers. Science (80-.). 300, 949–949 (2003).
5. Markowitz, S. *et al.* Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 268, 1336–8 (1995).
6. Baker, S. J. *et al.* Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science 244, 217–21 (1989).
7. Rodrigues, N. R. *et al.* p53 mutations in colorectal cancer. Proc. Natl. Acad. Sci. U. S. A. 87, 7555–9 (1990).
8. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–67 (1990).
9. Fearon, E. R. Molecular genetics of colorectal cancer. Annu. Rev. Pathol. 6, 479–507 (2011).
10. Markowitz, S. D. & Bertagnolli, M. M. Molecular origins of cancer: Molecular basis of colorectal cancer. N. Engl. J. Med. 361, 2449–60 (2009).
11. Marshman, E., Booth, C. & Potten, C. S. The intestinal epithelial stem cell. Bioessays 24, 91–8 (2002).
12. Barker, N. *et al.* Identifying the stem cell of the intestinal crypt: strategies and pitfalls. Cell Stem Cell 11, 452–60 (2012).
13. van de Wetering, M. *et al.* The β-Catenin/TCF-4 Complex Imposes a Crypt Progenitor Phenotype on Colorectal Cancer Cells. Cell 111, 241–250 (2002).
14. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 449, 1003–7 (2007).
15. Barker, N., van de Wetering, M. & Clevers, H. The intestinal stem cell. Genes Dev. 22, 1856–64 (2008).
16. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. Nature 457, 608–11 (2009).
17. Muñoz, J. *et al.* The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. EMBO J. 31, 3079–91 (2012).
18. Abbott, A. Cell culture: biology's new dimension. Nature 424, 870–2 (2003).
19. Tanner, K. & Gottesman, M. M. Beyond 3D culture models of cancer. Sci. Transl. Med. 7, 283ps9 (2015).
20. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 459, 262–5 (2009).
21. Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology 141, 1762–72 (2011).
22. Ohta, Y. & Sato, T. Intestinal tumor in a dish. Front. Med. 1, 14 (2014).
23. Leushacke, M. & Barker, N. Ex vivo culture of the intestinal epithelium: strategies and applications. Gut 63, 1345–54 (2014).

24. van de Wetering, M. *et al.* Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients. Cell 161, 933–945 (2015).

25. Clevers, H. Modeling Development and Disease with Organoids. Cell 165, 1586–97 (2016).

26. Yates, J. R. The revolution and evolution of shotgun proteomics for large-scale proteome analysis. J. Am. Chem. Soc. 135, 1629–40 (2013).

27. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. Curr. Opin. Chem. Biol. 24, 11–7 (2015).

28. Cravatt, B. F., Simon, G. M. & Yates, J. R. The biological impact of mass-spectrometry-based proteomics. Nature 450, 991–1000 (2007).

29. Larance, M. & Lamond, A. I. Multidimensional proteomics for cell biology. Nat. Rev. Mol. Cell Biol. 16, 269–280 (2015).

30. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat. Rev. Genet. 14, 35–48 (2013).

31. Cox, J. & Mann, M. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. Annu. Rev. Biochem. 80, 273–299 (2011).

32. Wiśniewski, J. R. *et al.* Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. Mol. Syst. Biol. 8, 611 (2012).

33. Halvey, P. J., Zhang, B., Coffey, R. J., Liebler, D. C. & Slebos, R. J. C. Proteomic consequences of a single gene mutation in a colorectal cancer model. J. Proteome Res. 11, 1184–95 (2012).

34. Jimenez, C. R. & Fijneman, R. J. A. Proteomics of Colorectal Cancer in a Genomic Context: First Large-scale Mass Spectrometry-Based Analysis from the Cancer Genome Atlas. Clin. Chem. 61, 1126–8 (2015).

35. de Wit, M. *et al.* Colorectal cancer candidate biomarkers identified by tissue secretome proteome profiling. J. Proteomics 99, 26–39 (2014).

36. Surinova, S. *et al.* Non-invasive prognostic protein biomarker signatures associated with colorectal cancer. EMBO Mol. Med. 7, 1153–65 (2015).

37. Surinova, S. *et al.* Prediction of colorectal cancer diagnosis based on circulating plasma proteins. EMBO Mol. Med. 7, 1166–78 (2015).

38. Sirvent, A., Vigy, O., Orsetti, B., Urbach, S. & Roche, S. Analysis of SRC oncogenic signaling in colorectal cancer by stable isotope labeling with heavy amino acids in mouse xenografts. Mol. Cell. Proteomics 11, 1937–50 (2012).

39. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat. Protoc. 4, 484–94 (2009).

40. Frese, C. K. *et al.* Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. J. Proteome Res. 10, 2377–88 (2011).

41. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods 4, 923–5 (2007).

42. Spivak, M., Weston, J., Bottou, L., Käll, L. & Noble, W. S. Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. J. Proteome Res. 8, 3737–45 (2009).

43. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–72 (2008).

44. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. Nature 1–21 (2014). doi:10.1038/nature13438

45. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control.

Nature 473, 337–42 (2011).

46. Haugen, A. C. *et al.* Genetic instability caused by loss of MutS homologue 3 in human colorectal cancer. Cancer Res. 68, 8465–8472 (2008).

47. Plaschke, J. *et al.* Loss of MSH3 protein expression is frequent in MLH1-deficient colorectal cancer and is associated with disease progression. Cancer Res. 64, 864–70 (2004).

48. Wheeler, J. M., Bodmer, W. F. & Mortensen, N. J. DNA mismatch repair genes and colorectal cancer. Gut 47, 148–53 (2000).

49. Hemminki, A. *et al.* Loss of the wild type MLH1 gene is a feature of hereditary nonpolyposis colorectal cancer. Nat. Genet. 8, 405–10 (1994).

50. Hampel, H. *et al.* Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). N. Engl. J. Med. 352, 1851–60 (2005).

51. Hendriks, Y. M. C. *et al.* Heterozygous mutations in PMS2 cause hereditary nonpolyposis colorectal carcinoma (Lynch syndrome). Gastroenterology 130, 312–22 (2006).

52. Ferreira, A. M. *et al.* High Frequency of RPL22 Mutations in Microsatellite-Unstable Colorectal and Endometrial Tumors. Hum. Mutat. 35, 1442–1445 (2014).

53. Woerner, S. M. *et al.* Microsatellite instability of selective target genes in HNPCC-associated colon adenomas. Oncogene 24, 2525–2535 (2005).

54. Kim, N.-G. *et al.* Identification of MARCKS, FLJ11383 and TAF1B as putative novel target genes in colorectal carcinomas with microsatellite instability. Oncogene 21, 5081–5087 (2002).

55. Gryfe, R. *et al.* Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. N. Engl. J. Med. 342, 69–77 (2000).

56. Sinicrope, F. A. & Sargent, D. J. Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. Clin. Cancer Res. 18, 1506–12 (2012).

57. Rubinfeld, B. *et al.* Association of the APC gene product with beta-catenin. Science 262, 1731–4 (1993).

58. Roose, J. *et al.* The Xenopus Wnt effector XTcf-3 interacts with Groucho-related transcriptional repressors. Nature 395, 608–12 (1998).

59. Daniels, D. L. & Weis, W. I. Beta-catenin directly displaces Groucho/TLE repressors from Tcf/Lef in Wnt-mediated transcription activation. Nat. Struct. Mol. Biol. 12, 364–371 (2005).

60. Hedberg Oldfors, C. *et al.* Analysis of an independent tumor suppressor locus telomeric to Tp53 suggested Inpp5k and Myo1c as novel tumor suppressor gene candidates in this region. BMC Genet. 16, 80 (2015).

61. Funakoshi, S., Ezaki, T., Kong, J., Guo, R. J. & Lynch, J. P. Repression of the Desmocollin 2 Gene Expression in Human Colon Cancer Cells Is Relieved by the Homeodomain Transcription Factors Cdx1 and Cdx2. Mol. Cancer Res. 6, 1478–1490 (2008).

62. Khan, K. *et al.* Desmocollin switching in colorectal cancer. Br. J. Cancer 95, 1367–70 (2006).

63. Xie, C., Powell, C., Yao, M., Wu, J. & Dong, Q. Ubiquitin-conjugating enzyme E2C: A potential cancer biomarker. Int. J. Biochem. Cell Biol. 47, 113–117 (2014).

64. Fujita, T. *et al.* Overexpression of UbcH10 alternates the cell cycle profile and accelerate the tumor proliferation in colon cancer. BMC Cancer 9, 87 (2009).

65. Chen, S. *et al.* Association of clinicopathological features with UbcH10 expression in colorectal cancer. J. Cancer Res. Clin. Oncol. 136, 419–426 (2009).

66. Raynes, D. A., Graner, M. W., Bagatell, R., McLellan, C. & Guerriero, V. Increased expression of the Hsp70 cochaperone HspBP1 in tumors. Tumour Biol. 24, 281–5 (2004).

67. Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D. & Klingler, T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res. 9, 1198–203 (1999).

68. Ballesta, A. M., Molina, R., Filella, X., Jo, J. & Giménez, N. Carcinoembryonic antigen in staging and follow-up of patients with solid tumors. Tumour Biol. 16, 32–41 (1995).

69. Thompson, J. *et al.* Down-regulation of carcinoembryonic antigen family member 2 expression is an early event in colorectal tumorigenesis. Cancer Res. 57, 1776–84 (1997).

70. Thompson, J. *et al.* CGM2, a member of the carcinoembryonic antigen gene family is down-regulated in colorectal carcinomas. J. Biol. Chem. 269, 32924–31 (1994).

71. Williams, S. J. *et al.* Two novel mucin genes down-regulated in colorectal cancer identified by differential display. Cancer Res. 59, 4083–9 (1999).

72. Matsuyama, T. *et al.* MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. Int. J. Cancer 127, 2292–9 (2010).

73. Packer, L. M., Williams, S. J., Callaghan, S., Gotley, D. C. & McGuckin, M. A. Expression of the cell surface mucin gene family in adenocarcinomas. Int. J. Oncol. 25, 1119–26 (2004).

74. Gay, F. *et al.* In colon carcinogenesis, the cytoskeletal protein gelsolin is down-regulated during the transition from adenoma to carcinoma. Hum. Pathol. 39, 1420–30 (2008).

75. Ohmachi, T. *et al.* Fatty acid binding protein 6 is overexpressed in colorectal cancer. Clin. Cancer Res. 12, 5090–5 (2006).

76. Collins, C. *et al.* Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma. Proc. Natl. Acad. Sci. U. S. A. 95, 8703–8 (1998).

77. Correa, R. G., de Carvalho, A. F., Pinheiro, N. A., Simpson, A. J. & de Souza, S. J. NABC1 (BCAS1): alternative splicing and downregulation in colorectal tumors. Genomics 65, 299–302 (2000).

78. Weeber, F. *et al.* Preserved genetic diversity in organoids cultured from biopsies of human colorectal cancer metastases. Proc. Natl. Acad. Sci. U. S. A. (2015). doi:10.1073/pnas.1516689112

79. Boj, S. F. *et al.* Organoid models of human and mouse ductal pancreatic cancer. Cell 160, 324–38 (2015).

80. Gao, D. *et al.* Organoid cultures derived from patients with advanced prostate cancer. Cell 159, 176–87 (2014).

81. Karthaus, W. R. *et al.* Identification of multipotent luminal progenitor cells in human prostate organoid cultures. Cell 159, 163–75 (2014).

82. Drost, J. *et al.* Sequential cancer mutations in cultured human intestinal stem cells. Nature 521, 43–7 (2015).

83. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 44, D447-56 (2016).

Supplementary Figures

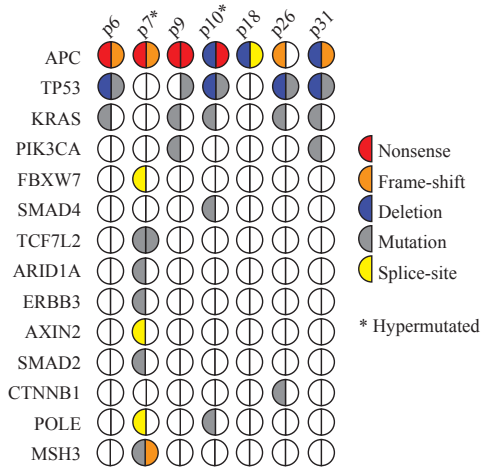


Figure S1. Related to Figure 1. Overview of the mutations found in the tumor organoids.

Whole-exome sequencing experiment revealed the mutations, insertions and deletions present in the tumor organoids. The most commonly altered genes in CRC were well represented (APC, TP53, KRAS, PIK3CA) and of the 7 tumor samples analyzed 2 displayed hypermutation (>10 mutations/Mb). Figure adapted from van de Wetering *et al*., 2015.

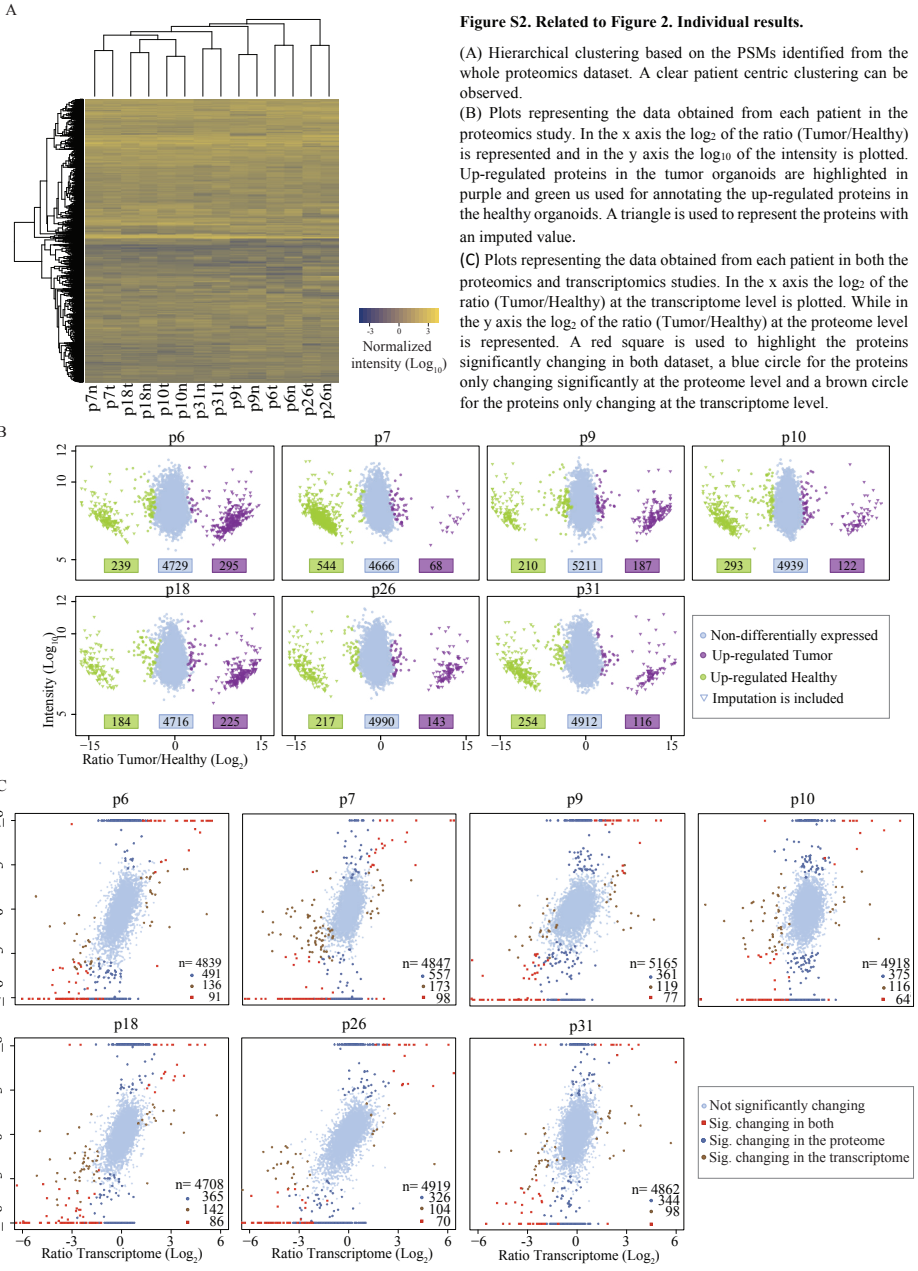


Figure S2. Related to Figure 2. Individual results.

(A) Hierarchical clustering based on the PSMs identified from the whole proteomics dataset. A clear patient centric clustering can be observed.

(B) Plots representing the data obtained from each patient in the proteomics study. In the x axis the $\log_2$ of the ratio (Tumor/Healthy) is represented and in the y axis the $\log_{10}$ of the intensity is plotted. Up-regulated proteins in the tumor organoids are highlighted in purple and green us used for annotating the up-regulated proteins in the healthy organoids. A triangle is used to represent the proteins with an imputed value.

(C) Plots representing the data obtained from each patient in both the proteomics and transcriptomics studies. In the x axis the $\log_2$ of the ratio (Tumor/Healthy) at the transcriptome level is plotted. While in the y axis the $\log_2$ of the ratio (Tumor/Healthy) at the proteome level is represented. A red square is used to highlight the proteins significantly changing in both dataset, a blue circle for the proteins only changing significantly at the proteome level and a brown circle for the proteins only changing at the transcriptome level.

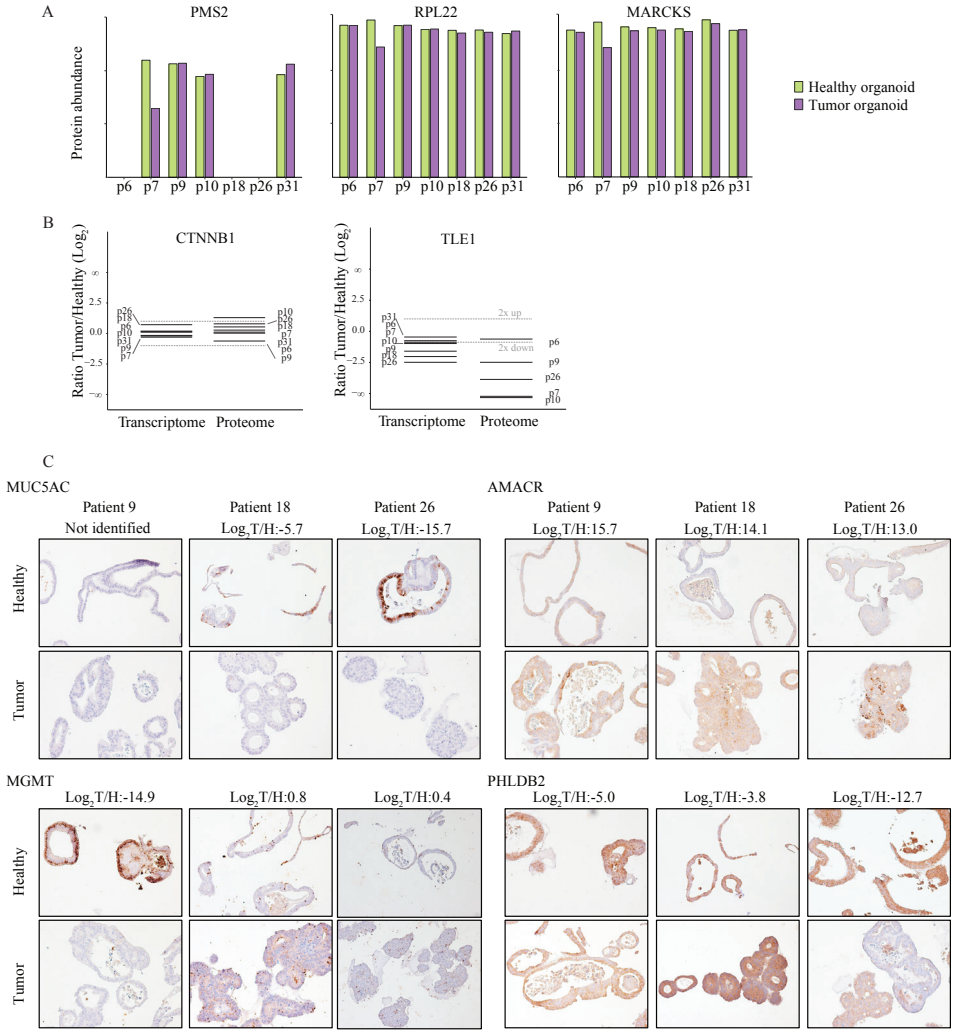


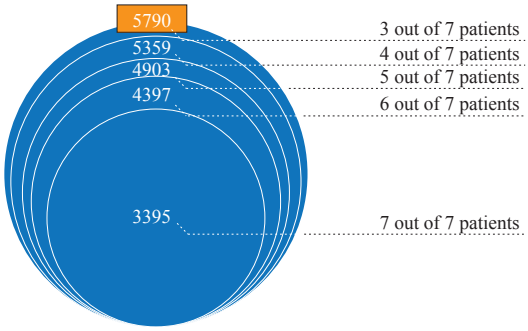
Figure S3. Related to Figure 3. Personalized proteomics profiles.

(A) Plot representing the protein abundance found in the healthy and tumor organoids for PMS2, RPL22 and MARCKS, proteins related to microsatellite instability. In all cases, a down-regulation is exclusively found in patient 7.

(B) Representations of the expression values for CTNNB1 and TLE1, proteins related to the Wnt signaling pathway, both at the proteome and transcriptome level.

(C) Immunohistochemical staining of organoids derived from p9, p18 and p26 with antibodies against MUC5AC, MGMT, AMACR and PHLDB2 confirming the differential protein abundance found in our mass spectrometry experiment (magnification=200x). These proteins display a significant difference across the patients at the proteome level (ratio Tumor/Healthy $\log_2$ values are shown), while either no significant difference is observed at the transcriptome level (MUC5AC and MGMT) or not identified at all at the transcriptome level (AMACR and PHLDB2).

A



B

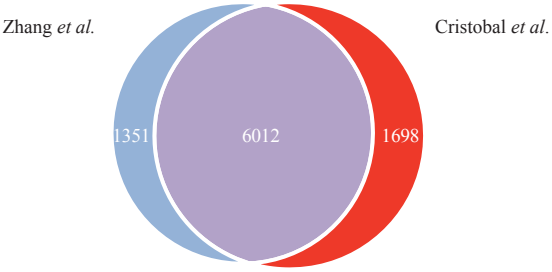


Figure S4. Related to Figure 4. General features

(A) Circle diagram representing the overlapping quantified proteins across the 7 patients analyzed in the proteomic experiment. In order to generate this plot the proteins quantified, with at least 2 peptides identified per protein are used. For the data analysis in which the patients are treated as a group, the proteins quantified in at least 3 out of the 7 patients are used, which corresponds to 5790 proteins.
(B) Venn diagram representing the overlap between the identified protein in the TCGA study (Zhang *et al.*, 2014) and our proteomics results. From the 7363 proteins that were identified in the study of the tissue obtained from 90 patients, 6012 proteins were also identified in our organoid study.

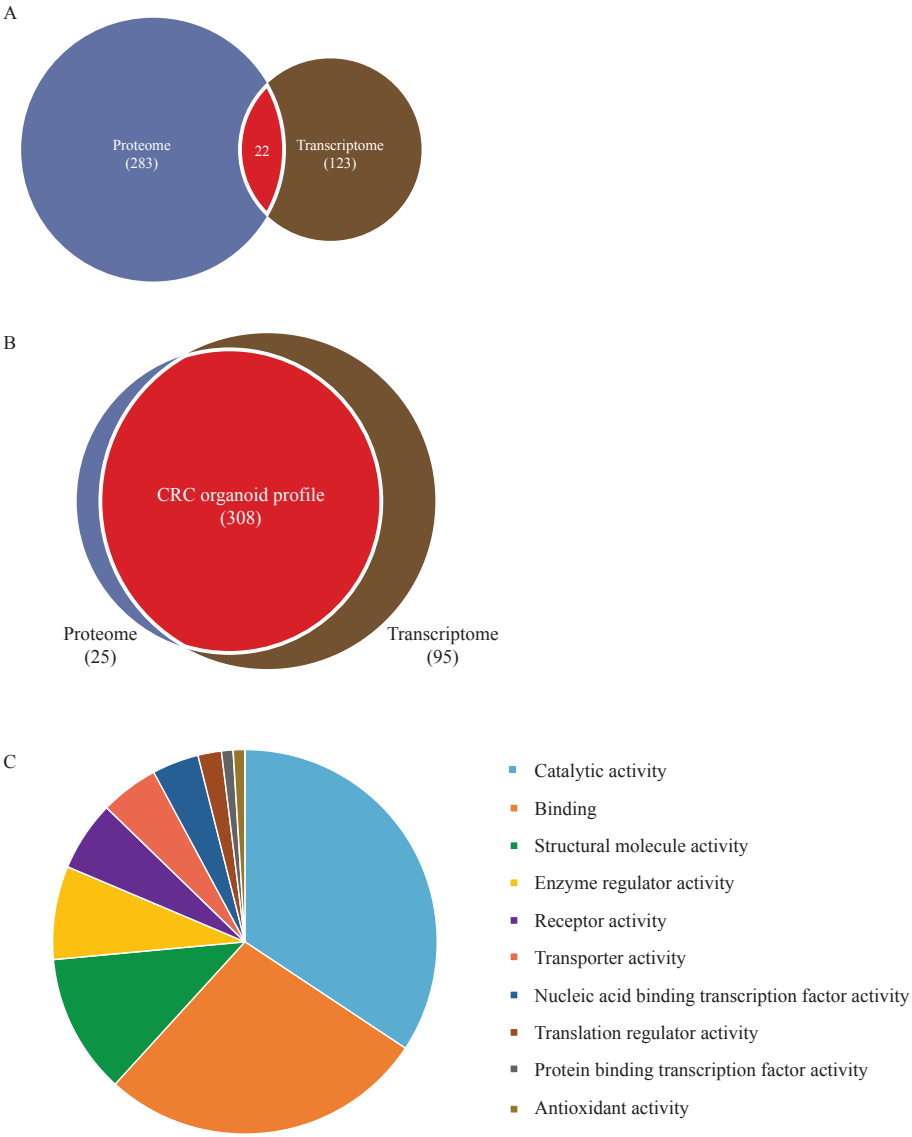


Figure S5. Related to Figure 5. CRC organoid profile.

(A) Venn diagram representing the overlap between the significantly changing proteins from the proteome (blue) and transcriptome (brown) data sets.
(B) Venn diagram representing the overlap between the significantly changing proteins in one of the two data sets that are also changing in the same direction on the other data set, without the need of being significant. The overlapping 308 proteins (red) correspond to the CRC organoid profile.
(C) Pie chart representing the molecular function of the CRC organoid profile. The main molecular function of these proteins can be divided into proteins with catalytic activity (35%), binding proteins (28%) and proteins with structural molecule activity (12%).

Supplemental experimental procedures

Chemicals and Materials

Iodoacetamide (IAA), sodium deoxycholate (SDC), formaldehyde, D-formaldehyde and sodium cyanoborohydride were supplied by Sigma-Aldrich (Steinheim, DE). Ammonium bicarbonate and dithiothreitol (DTT) were purchased from Fluka (Buchs, CH), and ethyl acetate from Merck (Darmstadt, DE). PhosSTOP Phosphatase Inhibitor Cocktail tablets and Complete Mini EDTA-free Cocktail tablets were obtained from Roche Diagnostics (Mannheim, DE), trypsin from Promega Corporation (Madison, WI, USA) and lysyl endopeptidase (Lys-C) from Wako (Richmond, VA, USA). Pierce BCA (bicinchoninic acid assay) protein assay kit was supplied by Thermo Scientific (Rockford, IL, USA). Matrigel was purchased from BD Biosciences (Franklin Lakes, NJ, USA) and the components of the Human Intestinal Stem Cell medium were supplied by Vivogen (Beograd, RS). The water used in all experiments was obtained from a Milli-Q purification system (Millipore, Bedford, MA, USA).

Patient Material Background and organoid culture

Colonic tissues (normal as well as tumor tissue) were obtained from The Diakonessen Hospital Utrecht with informed consent and the study was approved by the ethical committee. After the isolation of healthy crypts and tumor epithelium, organoids were cultured in Human Intestinal Stem Cell medium (HISC) resuspended in Matrigel. The composition of HISC is: Basal culture medium with 50% Wnt conditioned medium, 20% R-Spondin conditioned medium, 10% Noggin conditioned medium, 1 x B27, 1,25 mM n-Acetyl Cysteine, 10 mM Nicotinamide, 50 ng/ml human EGF, 10 nM Gastrin, 500 nM A83-01, 3 uM SB202190, 10 nM Prostaglandine E2 and 1 x Primocin.

Sample preparation for proteomics analysis

Organoids were lysed in a buffer containing 2.5% w/v SDC in a solution of 50 mM ammonium bicarbonate pH 8.2 with protease and phosphatase inhibitors. The lysis was performed at 95 degrees for 5 minutes. Protein concentration was estimated by a BCA protein assay. Prior to the digestion the proteins (200 µg per sample) were reduced with 2 mM DTT for 25 min at 56 degrees and alkylated with 4 mM IAA for 30 min at room temperature in the dark. In order to avoid over alkylation an extra step of DTT was added. For the digestion of the proteins Lys-C and trypsin were used. First, proteins were digested with Lys-C at 37 degrees for 4 hours (enzyme/substrate ratio 1:75) and after a three-fold dilution, they were subsequently digested

with trypsin (enzyme/substrate ratio 1:100) at 37 degrees for 4 hours more. After digestion the samples were acidified with 0.5%TFA and a liquid-liquid extraction with ethyl acetate was performed in order to remove the remaining SDC.

The resulting peptides were chemically labelled using stable isotope dimethyl labelling as described before (Boersema *et al.* 2009). In short, peptides belonging to the healthy samples from the different patients were labelled with a mixture of formaldehyde and sodium cyanoborohydride ('light' reagent). For the peptides corresponding to the tumor organoids, D-formaldehyde with cyanoborohydride ('medium' reagent) was used. The 'light' and 'medium' dimethyl labelled samples were mixed in a 1:1 ratio based on total peptide intensities.

Prior to the MS analysis, samples were fractionated to reduce the complexity using a strong cation exchange (SCX) system. Briefly, seven SCXs were performed on an Agilent 1100 HPLC system (Agilent Technologies, Waldbronn, Germany) using a C18 Opti-Lynx (Optimized Technologies, Oregon, OR) trapping cartridge and a Zorbax BioSCX-Series II column (0.8 mm i.d. × 50 mm length, 3.5 μm). The labelled peptides were dissolved in 10% FA and loaded onto the trap column at 100 μl/min and subsequently eluted onto the SCX column with 80% acetonitrile and 0.05% FA. SCX solvent A consisted of 0.05% formic acid in 20% acetonitrile, while solvent B was 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. The following gradient was used: 0–0.01 min (0–2% B); 0.01–8.01 min (2–3% B); 8.01–14.01 min (3–8% B); 14.01–28 min (8–20% B); 28–38 min (20–40% B); 38–48 min (40–90% B); 48–54 min (90% B); 54–60 min (0% B). After injection of ~100 ug of sample a total of 50 SCX fractions were collected and dried in a vacuum centrifuge.

Transcriptomics analysis

RNA from the organoid tumor samples and the paired normal samples was hybridized on Affymetrix Human Gene 2.0 ST arrays. The raw CEL files were processed with Affymetrix Power Tools (ATP) using the Hg19 genome build and NetAffx annotation dating from 09-30-2012. Between-array normalisation was performed using rma-sketch, within APT. This resulted in an intensity matrix of 21681 genes.

Histology procedures

Organoids were collected, fixed in 10% neutral buffered formalin and embedded in paraffin. Sections were subjected to H&E and Alcian blue staining as well as immunohistochemical staining. The following primary antibodies were used for immunohistochemical staining: MUC5AC (NCL-HGM-45M1, Novus Biologicals),

1:200, PHLDB2 (NBP2-38238, Novus Biologicals), 1:50, MGMT (#MAB16200, Millipore), 1:500, AMACR (ab12498, Abcam), 1:125). Antigen retrieval was performed according to the manufacturer's recommendations.

Data analysis - Global differential expression analysis

Protein ratio analysis - Missing quantification data values

Proteome Discoverer has the option to replace missing quantification values for peptides with a minimum intensity value. In our hands, these "imputed" values strongly influenced the protein ratios compared to the same protein ratios without the imputed peptide quantities. On the other hand, we tried to retain the proteins that would not be detected without this imputation since they could indicate a presence/absence situation between our conditions. We therefore decided to use the following compromise in our analysis: We only used the imputed peptides if they were necessary to obtain a protein ratio at all. The imputation state of these proteins were marked, to show that the calculated ratios were completely determined by the imputation and therefore indicating an absence of the protein in either condition. To achieve this separation, we obtained the ratio values from the exported tabular data from PD, both with and without imputation. If a protein ratio was present without imputed values, it was used, otherwise the value for the protein of the file containing imputed values was used, and the ratio was marked as "imputed". At the same, we circumvented the limitation of having only three decimal places in the reported ratio values from PD. We obtained all reports with reversed channel ratio as well, using their reciprocal values to replace the original ratios below 1.

To obtain a global picture of protein levels for the tumor versus the healthy organoids, we performed t-tests on the $\log_2$ transformed protein ratios, considering the different tissue donors as biological replicate measurements. Only proteins identified by more than two peptides were considered. Tests were performed whenever at least three measurements (i.e. patient ratios) were available for a protein, regardless of imputation. Although the t-test is not strictly valid for these imputed ratios, it is used here to discern the low-expressed proteins from the more highly expressed ones in an objective way. False discovery rates were estimated over all t-test results using the Benjamini and Hochberg (Benjamini & Hochberg 1995).

Intensity based semi-quantification

As a means of representing the individual protein levels in each tissue type, ungrouped peptide lists (i.e. PSM lists) were exported, with the "raw quant values" from the dimethyl quantification performed by Proteome Discoverer. The values

$\log_{10}$ transformed and scaled using the R scale {base} function, with both centering and scaling enabled. Hierarchical clustering was performed on proteins with measurements in all patients, using Euclidean distance and Ward's method for clustering (Ward JH 1963). The protein intensities used in Figure 2B, Figure 3B and Figure S2 were obtained in a more simple way: the Area from the PD report was divided according to the ratio and represented that way.

Transcriptomics ratios were calculated based on the Affymetrix array intensities and analyzed the same way as the proteomics data, with an important distinction. By lack of material, the healthy sample of patient 31 was no longer available. Since all arrays cluster together based on their cancer state, we constructed an average of RNA levels between the remaining patients and used it in further analyses.

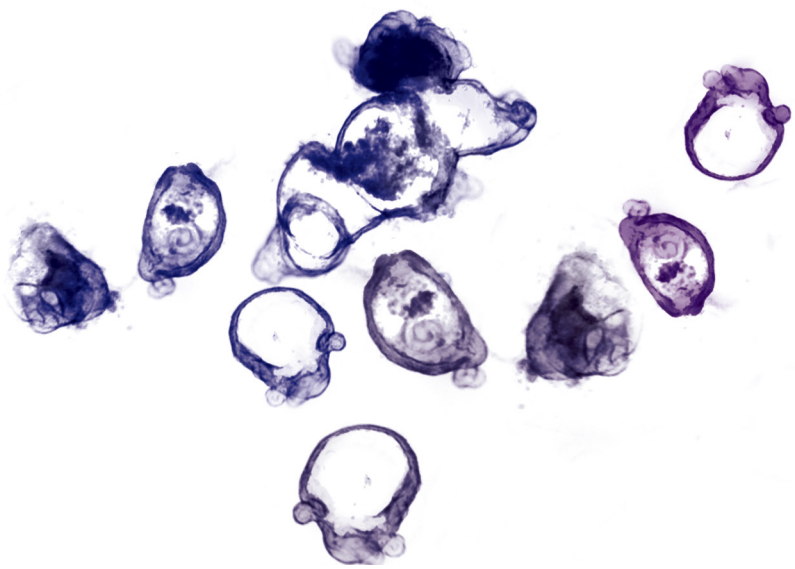
Supplemental references

Benjamini, Y. & Hochberg, Y., 1995. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. Journal of the Royal Statistical Society. Series B: Methodological, 57, pp.289–300.

Ward JH, J., 1963. Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association. Journal of the American Statistical Association, 58, 236–244.

Chapter 4

Toward an optimized workflow for middle-down proteomics



Toward an optimized workflow for middle-down proteomics

Alba Cristobal[1,2,*], Fabio Marino[1,2,*], Harm Post[1,2], Henk W.P. van den Toorn[1,2], Shabaz Mohammed[1,2,3] and Albert J. R. Heck [1,2]

Analytical Chemistry, 10 March 2017

[1]Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

[2] Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

[3]Departments of Chemistry and Biochemistry, University of Oxford, New Biochemistry building, South Parks Road, Oxford, OX1 3QU Oxford, UK

*The authors contributed equally to this work

Abstract

Mass spectrometry (MS)-based proteomics workflows can crudely be classified into two distinct regimes, targeting either relatively small peptides (i.e. 0.7 kDa < Mw < 3.0 kDa) or small to medium sized intact proteins (i.e. 10 kDa < Mw < 30 kDa), respectively termed bottom-up and top-down proteomics. Recently, a niche has started to be explored covering the analysis of middle-range peptides (i.e. 3.0 kDa < Mw < 10 kDa), aptly termed middle-down proteomics. Although middle-down proteomics can follow, in principle, a modular workflow similar to that of bottom-up proteomics, we hypothesized that each of these modules would benefit from targeted optimization to improve its overall performance in the analysis of middle-range sized peptides. Hence, to generate middle-range sized peptides from cellular lysates we explored the use of the proteases Asp-N and Glu-C, and a non-enzymatic acid induced cleavage. To increase the depth of the proteome, a strong cation exchange (SCX) separation, carefully tuned to improve the separation of longer peptides, combined with reversed phase-liquid chromatography (RP-LC) using columns packed with material possessing a larger pore size, was used. Finally, after evaluating the combination of potentially beneficial MS settings, we also assessed the peptide fragmentation techniques, including higher-energy collision dissociation (HCD), electron-transfer dissociation (ETD) and electron-transfer combined with higher-energy collision dissociation (EThcD) for characterization of middle-range sized peptides. These combined improvements clearly improve the detection and sequence coverage of middle-range peptides and should guide researchers to explore further how middle-down proteomics may lead to an improved proteome coverage, beneficial for, among other things, the enhanced analysis of (co-occurring) post-translational modifications.

Introduction

At the present time, two quite distinct approaches co-exist in mass spectrometry-based proteomics studies; bottom-up, peptide centric approaches and top-down, protein centric approaches. In bottom-up proteomics, proteins are initially enzymatically digested, after which the resulting peptides are typically separated by liquid chromatography (LC) and analyzed by tandem mass spectrometry (MS/MS).[1] In proteomics, trypsin represents the 'golden standard' for proteases

as it generates peptides that are relatively facile to separate by LC and analyze by MS. Therefore, the bottom-up approach is robust and enables high-throughput analysis, allowing the identification and quantification of thousands of proteins from complex lysates. However, sample complexity still imposes a heavy burden on separations prior to MS detection and, especially when using trypsin, a large part of the generated peptides are too small for successful binding to the stationary phases typically used in proteomics. Moreover, peptides smaller than 5 amino acids are typically not analyzed by MS/MS since they provide too little of an analytical value for unambiguous identification. Therefore, quite a portion of tryptic peptides may not be detected or are intentionally ignored.

In top-down proteomics, intact proteins are analyzed,[2] which circumvent issues related to peptide-centric proteomics such as the protein inference problem.[3] Top-down proteomics provides complete molecular specificity on intact proteins, enabling the analysis of proteoforms[4] as intact proteins harbor the entire set of (co-occurring) PTMs.[5,6] The main limitations in top-down proteomics are related to the difficulty of efficiently separating proteins and the relatively inefficient formation of fragment ions from larger proteins. Nevertheless, top-down is now becoming feasible, also in a relatively high-throughput manner, albeit mainly limited to proteins with a Mw up to ~30 kDa. Sequence information on these proteins can be obtained by employing collision-induced dissociation (CID) which selectively fragments the most labile bonds in a protein, which typically provides only limited protein coverage.[7] Use of electron capture/transfer dissociation (ECD/ETD)[8,9] was shown to improve the sequence coverage. Additionally, ETD preserves most common labile PTMs during fragmentation, facilitating confident PTM site localization. More recent fragmentation methods, such as EThcD and UV photodissociation, are emerging as efficient fragmentation alternatives for top-down proteomics.[10–12]

As top-down proteomics typically covers relatively small proteins (10 < Mw < 30 kDa), and trypsin based bottom-up proteomics focuses generally on peptides (0.7 < Mw < 3 kDa), evidently a gray zone exists between these proteomics approaches. This gap has been filled by a third approach, nowadays, called middle-down proteomics.[13–15] This approach also uses protein digestion, but aims to yield relatively larger peptides (ideally (far) above 3 kDa).[16–19] Middle-down proteomics exhibits particular advantages as the complexity of the digests decreases (as fewer

peptides are formed), and may also allow better proteome coverage, including the identification of splice-variants and other isoforms. Longer peptides also increase the probability of detecting multiple co-occurring neighboring PTMs, important to study functionally relevant PTM crosstalk. Most of the pioneering middle-down proteomics studies have been limited to specific applications, for instance, on recombinant monoclonal antibodies,[20,21] ubiquitin chains,[22] and N-terminal histone chains.[23–25] Thus, middle-down proteomics is not yet routinely used for the analysis of full proteomes. We argue that an important bottleneck has been that neither the pipeline developed for bottom-up nor that for top-down analyses is directly applicable and optimal for middle-down proteomics. Therefore, here, we aimed to describe and critically evaluate a workflow optimized for the detection of middle-range sized peptides.

The first experimental barrier in the high-throughput middle-down proteomics workflow is the lack of a proteolytic enzyme that can produce controlled populations of middle-range sized peptides (3-10 kDa). A protease cutting at a well-defined peptide length or size would obviously be ideal; however, such an idealase remains to be discovered.[26] Some known single-residue specific proteases such as Lys-C,[13,27,28] Glu-C,[29–31] Asp-N[32] and Lys-N[33,34] are thought to produce, when compared to trypsin, a higher proportion of middle-range sized peptides and may thus provide second-best alternatives for middle-down proteomics. Also some non-enzymatic approaches may be exploited to generate longer peptides. Microwave-accelerated acid hydrolysis, which produces Asp-selective chemical cleavage, had been explored by using a number of acid modifiers.[35,36] We evaluated whether the proteases Asp-N and Glu-C or a non-enzymatic acid induced digestion protocol provide middle-range sized peptides. Secondly, additional analytical issues remain to be optimized for middle-down proteomics, such as the chromatography applied to separate the longer peptides. For this reason, we optimized conditions for sample clean-up and separation of these longer peptides by multidimensional LC. Finally, MS transmission and detection as well as MS/MS fragmentation conditions should be adjusted.[37] We fine-tuned the critical MS parameters for improving the detection of middle-range sized peptides. Moreover, we evaluated the performances of ETD, HCD and EThcD as fragmentation schemes on all the peptide-digests obtained to define the best sequencing method for distinct populations of middle-range sized peptides.

Experimental section

Sample preparation

HeLa digests were prepared as described previously.[38] For preparing the Asp-N and Glu-C digests, the protocols reported by Giansanti *et al.* were used.[39] For the acid induced non-enzymatic digestion, the lysate was diluted to a final concentration of 0.1 µg/µl using a solution of formic acid (FA) (final concentration of 2% FA) and incubated at 100 °C for 1 hour. A detailed description of the used protocols can be found in the supporting information.

Sample clean-up and pre-fractionation

Following digestion, sample clean-up was performed by using solid-phase extraction (SPE) columns; C18 with a 300 Å pore size. Prior to the MS analysis, samples were fractionated using strong cation exchange (SCX) chromatography. A detailed description of the used methods can be found in the supporting information.

LC-MS and LC-MS/MS set up

Nano-UHPLC-MS/MS was performed on an Agilent 1290 Infinity System connected to an Orbitrap Fusion. Fused-silica capillary analytical and trap columns were prepared as previously described.[40] The UHPLC was equipped with a double frit trapping column and a single frit analytical column. ReprosilPur C18 (3 µm particles, 120 Å pore size 2 cm x 100 µm) was used as a trap column, and Zorbax SB-C18 (1.8 µm particles 80 Å 40 cm x 50 µm) for the analytical column. For the 300 Å pore size set up, the used materials were as follows: Zorbax SB-C18 (3.5 µm particles, 300 Å pore size, 2 cm x 100 µm) for the trap column and Zorbax SB-C18 (1.8 µm particles 300 Å 40 cm x 50 µm) for the analytical column. The column, in both cases, was directly connected to an in-house pulled and gold-coated fused silica needle (with a 5 µm o.d. tip), and a voltage of 2.0 kV was applied. The survey scan range was from 350 to 1500 m/z at a resolution of 60000 (200 m/z) with an AGC target of 4e5. The most intense precursor ions were selected for subsequent fragmentation at Top Speed within a 3 seconds duty cycle. A resolution of 30000 (200 m/z) and a maximum injection time of 125 ms were found to be ideal for MS/MS. The AGC target for the MS/MS was set to 1e5. When HCD was used, 35% collision energy (CE) was applied; in the case of EThcD, 40% supplemental activation (SA)

was selected and when ETD was used 10% SA was applied. Additionally, charge triggered MS/MS, instead of intensity triggered, was tested for the EThcD charge method.

Data analysis

The RAW files were processed by using Proteome Discoverer, and the spectra were searched against the UniProt human database. Database searching was performed with Sequest HT and Mascot, and the results were filtered using Percolator[41,42] to a peptide and protein FDR < 1%. A detailed description of the data analysis can be found in the supporting information.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[43] partner repository with the dataset identifier PXD004910, with the following account details:

Username: reviewer09480@ebi.ac.uk

Password: xWhaEOKh

Results and Discussion

Evaluation of enzymatic digestion approaches for the generation of middle-range sized peptides

This study was aimed at providing an objective and comprehensive evaluation of the optimization steps required for the generation and analysis by LC-MS/MS of middle-range sized peptides (Figure S-1). We first set out to generate peptide digests of cell lysates, which in principle could yield higher numbers of middle-range sized peptides. Proteases that target single residues such as Asp-N, Glu-C and Lys-C, are thought to produce peptides with distinct peptide length distributions compared to trypsin.[27,44–46] In our initial experiments, we tested Asp-N and Glu-C and digested a HeLa lysate, while a previously published trypsin data set was used as a reference.[47] Noteworthy, the median Mw of all identified peptides was found to be around 1.9 kDa for Glu-C and Asp-N, an increase of about 25% compared to

1.5 kDa median Mw for trypsin (Figure S-2, Table S-1). Although this increase in median Mw is significant, ideally we would generate for middle-down proteomics peptides with a median Mw > 3 kDa, and thus, the results obtained were found to be sub-optimal. Therefore, we also set out to explore other means to generate peptides from proteins, using a non-enzymatic digestion.

Testing a non-enzymatic digestion for the generation of middle-range sized peptides

Chemical cleavage methods have been used as an alternative to enzymatic proteolysis to generate peptides from proteins.[35] In acid-induced chemical digestion, proteins are incubated at high temperatures in formic acid (FA) diluted solutions.[36,49] Diluted FA was reported to cleave proteins mainly at the C-termini of aspartic acid (D).[36] We hypothesized that due to this specificity, FA-induced digestion would potentially yield peptides of middle-sized length.[35] To our knowledge, acid-induced chemical digestion has not been generally explored in full complex proteome analysis.[50] Interestingly, the median Mw of all identified peptides in the FA induced digestion was similar to Asp-N and Glu-C. In addition, the frequency of occurrence of missed cleaved peptides was on par with that for the Glu-C digest datasets (Figure S-2).

Adapting clean-up and chromatographic conditions to maximize the retention and separation of middle-range sized peptides

Once a considerable population of middle-range sized peptides is generated, it is crucial to avoid their loss in the subsequent steps of the proteomics workflow. The in silico Mw distributions generated based on the Asp-N and Glu-C specificities are 2.39 and 2.37 kDa, respectively. We analyzed experimental data obtained from previous in-depth proteomics studies generated by Asp-N and Glu-C and a much lower average Mw value was found.[52,53] We thus hypothesized that this discrepancy can be caused by losses of middle-range sized peptides which can be due to sub-optimal separation chemistry being used for sample clean-up, pre-fractionation and/or for the choice of unfavorable LC conditions. We thus optimized a range of steps in the proteomics workflow (Figure S-1), prior to MS detection, which

were aimed to minimize the losses of middle-range sized peptides. When dealing with these peptides one of the main aspects to consider is their ability of being adsorbed (mass transfer) and efficiently separated by the porous particles used in the employed chromatographic techniques. A poor choice of pore size will thus result in inadequate retention and loss of resolution. Conventionally particles with 80-120 Å pore sizes are used for the separation of tryptic peptides. However, a more appropriate pore size for the separation of middle-range sized peptides is 300 Å, as has been previously reported.[54] We therefore decided to test the latter pore size material for peptide desalting with solid-phase extraction (SPE) columns, pre-fractionation and reversed phase chromatography.

Strong Cation Exchange as a selective pre-fractionation mode for middle-range sized peptides

For in depth proteomics studies, typically one or more fractionation steps are included to decrease sample complexity and boost the number of identifications.[53,55–57] When evaluating options for multidimensional separation schemes used in proteomics approaches, we prioritized binding and separation of longer peptides with high frequency of basic residues in order to match our digested peptide populations. We, thus, hypothesized that SCX at low pH (typically ~3) could be appropriate as a fractionation technique. As initially hypothesized, we found a good correlation between the charge and the Mw of the peptides generated by the alternative digestions (Figure S-3a). Interestingly, for middle-down proteomics approaches, the above-described correlation can be exploited by selecting the appropriate fractions containing peptides of higher Mw.

Optimizing the pore size for the separation of middle-range sized peptides

We first set up a side-by-side comparison between the conventional RP C18 material and the 300 Å pore size material, using the same stationary phase chemistry in order to test if the difference in pore size generates a bias against higher Mw peptides. For the comparison, the same amount of Asp-N SCX fractions

was injected and identical MS/MS settings were used. We monitored the achieved peak capacity, retention of middle-range sized peptides, the attainable back pressure (see Supporting information), the overall retention times and the total number of identifications (Figure S-4a). Only a modest decrease in the overall peak capacity (about 8%) was observed using the 300 Å pore size column. However, when focusing on peptides with Mw > 4 kDa, improved peak widths and increased area under the curves (AUCs) were generally observed (Figure S-4b), ratifying the choice of the 300 Å pore size column for these peptides. The 300 Å pore size column marginally outperformed the conventional RP C18 column in terms of number of identified proteins (4895 vs 4722) and peptides (34307 vs 32907). Focusing on peptides identified exclusively with each of the columns, the better performance of the larger pore size material for middle-sized peptides became clearly visible (Figure S-3b). The 300 Å pore size column retained and eluted more efficiently peptides with Mw > 1.5 kDa, confirming our hypothesis that conventional pore size materials typically used for the separation of tryptic peptides negatively affects the retention and separation of middle-range sized peptides.

Tuning MS settings for improving the detection of middle-range sized peptides

Next, we evaluated several combinations of MS settings, which we thought could be optimized for the detection of middle-range sized peptides. We used the high-field Orbitrap analyzer, which enables higher resolution compared to the previous Orbitrap version at the same transient length which is advantageous for the analysis of larger peptides.[58] We evaluated the effects of resolution settings, for both MS and MS/MS scans, on the total number of peptide identifications, particularly the identification success rate of middle-range sized peptides and instrument duty cycle. For this comparison, several Asp-N SCX fractions, containing pools of peptides with distinct charge states, were chosen. We evaluated three methods varying the resolution at both the MS and MS/MS level (Table S-3). By increasing the resolution for MS/MS scans, we obtained a higher number of unique peptides, higher spectral quality (higher XCorr medians), and more middle-range sized peptides for any of the analyzed peptide population (Table S-3, top). Applying higher resolution for MS scans only resulted in a lower number of identified

unique peptides, largely due to the lower duty cycle (Table S-3, bottom). Thus, our experiments showed that the chosen higher resolution setting in the MS/MS mode is essential to correctly measure the isotope spacing associated with each ion fragment, improving the identification rates of especially highly charged, longer peptides. Furthermore, we opted to evaluate the effect of longer MS/MS injection times to increase the number of detectable ion fragments and therefore improve peptide sequence coverage. We tested 3 maximum injection times ranging from 75 (typical value for shotgun experiments) to 125 ms with the latter allowing us to increase the sequence fragmentation coverage from 88% to 91% and almost no consequences on the number of unique peptides identified (Table S-4). Optimized method parameters can be found in Table S-5.

Exploring the benefits of multiple fragmentation methods for the detection of middle-range sized peptides

We evaluated the performance of a number of fragmentation methods in identifying middle-range sized peptides from the Asp-N, Glu-C and FA HeLa digests (Figure 1). In each of these analyses, HCD fragmentation gave the highest number of unique peptides identified, which is an inherent effect due to its shorter duty cycle (twice as many MS/MS scans) compared to ETD and EThcD (Table S-1). However, EThcD showed superior success rates, defined as the number of PSMs divided by the total number of MS/MS acquired. The EThcD success rate was above 55% for both Asp-N and Glu-C, while ETD and HCD success rates were considerably lower, about 36%. Higher score distributions were also obtained for EThcD across the entire Mw range (Figure 1).

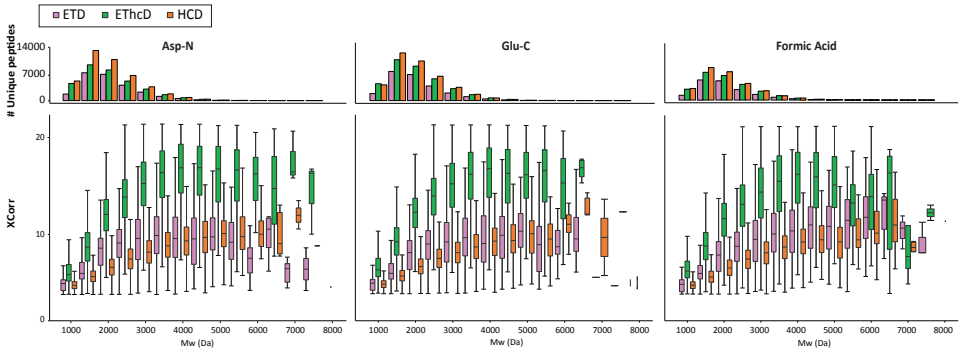


Figure 1. Performance of the fragmentation techniques (ETD, EThcD and HCD) with respect to peptide Mw in each of the 3 applied digestion protocols. The comparison is based on the number of identified unique peptides as well as on the quality of spectra represented by their XCorr distribution (box plots).

We further calculated the peptide sequence coverages for each digestion method achieved with all the fragmentation methods (Figure 2). The median peptide sequence fragmentation coverage of HCD was 65%, followed by ETD with 82% of coverage. EThcD displayed medians of sequence fragmentation coverage of 95%. This is partly due to the contribution of multiple ion series (c/z and b/y),[10] advantageous especially for middle-range sized peptides because they enable unambiguous sequence determination and multiple PTMs identification and site localization.[59,60] We focused again on distinct Mw sub-classes, peptides displaying a Mw < 2.5 kDa, peptides with 2.5 < Mw < 4 kDa, and the latter classification contains peptides with Mw > 4 kDa (Figure S-5b, c and d, respectively). Strikingly, for peptides identified at Mw > 4 kDa, the median peptide sequence coverage obtained by EThcD was constantly around 90%, confirming its unique potential for middle-down proteomics.

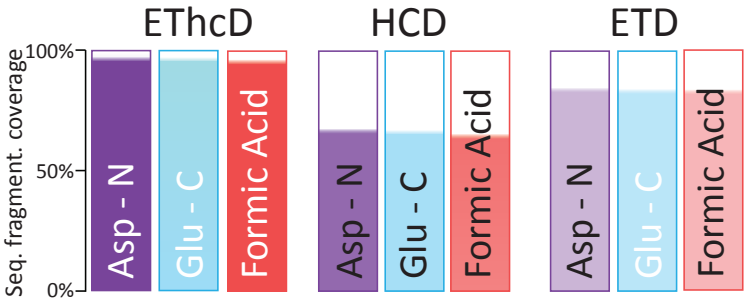


Figure 2. Peptide sequence fragmentation coverage obtained by each fragmentation method in the 3 applied digestion schemes. The median peptide sequence fragmentation coverage was calculated and represented taking into consideration all the SCX fractions for each digestion scheme.

Next to the peptide sequence coverage obtained by a single fragmentation method, we specifically looked at the performance of each sequencing technique in terms of number of unique peptides identified and XCorr distribution with respect to charge (z) and Mw ranges. (Figure 3 and Figure S-6-7). The score distributions and therefore the quality of the EThcD spectra are superior in all the studied ranges, while they yielded more identified peptides for the higher charge states (z 5 to 7). Interestingly, EThcD outperformed the other techniques also for lower charged peptides (z 2 to 4). The potential of ETD for the identification of peptides with Mw < 3 kDa, especially for highly charged peptides ($z \geq +5$), is clear even though the quality of the spectra is lower than with EThcD fragmentation. Additionally, as expected, the power of HCD for highly charged peptides is diminished compared to +2 and +3 charge states.

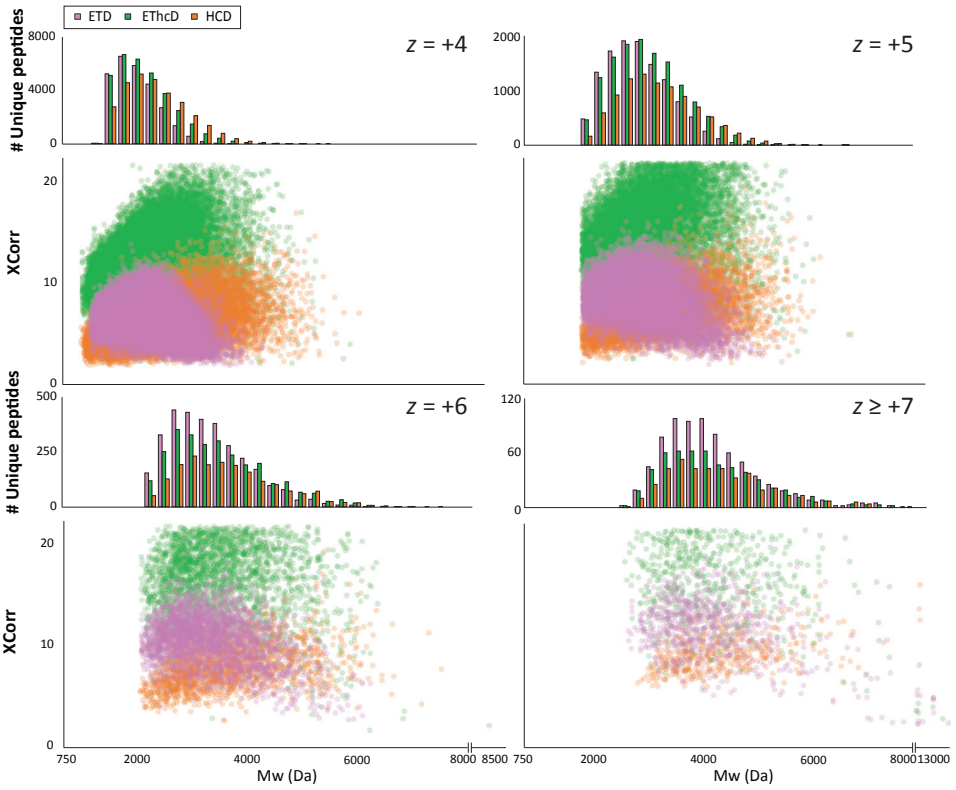


Figure 3. Performance of the peptide fragmentation techniques ETD, EThcD and HCD for peptides with respect to their Mw and charge states (z). Combined data from Asp-N, Glu-C and FA HeLa digests whereby the fragmentation parameters for ETD, EThcD and HCD were optimized. The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their z and Mw ranges.

During data dependent MS/MS triggering, the abundance of low mass peptides in the mixture can obscure the precursor selection and thus fragmentation of middle-range sized peptides.[62] We showed the correlation between charge state of peptides and their Mw (Figure S-3a) and reasoned that triggering MS/MS based on the highest charge state peptides present in the MS scan would further favor the detection of middle-range sized peptides. We evaluated the performance of this method compared to the usual intensity-based method with Asp-N SCX fractions. Indeed, not only did the peptide Mw median increase by 0.13 kDa by using the charge dependent MS/MS but also we could increase the number of total identified peptides (Figure S-8).

We also evaluated the success rate for each fragmentation technique with respect to Mw of peptide (Figure 4). The analysis confirmed the greater performance of EThcD for middle-down range peptides and of ETD especially for Mw > 6 kDa. However, despite these improved performances for longer peptides, the identification rates seem to drastically decrease at Mw > 4 kDa. One major cause of the decline in efficiency of identification for larger peptides has been related to the need of an increasing ion population required to maintain signal to noise for the generated fragment ion species.[61] The data suggests our workflow is allowing the transmission of a considerable number of middle range peptides but MS/MS efficiency is found to limit identification rates.

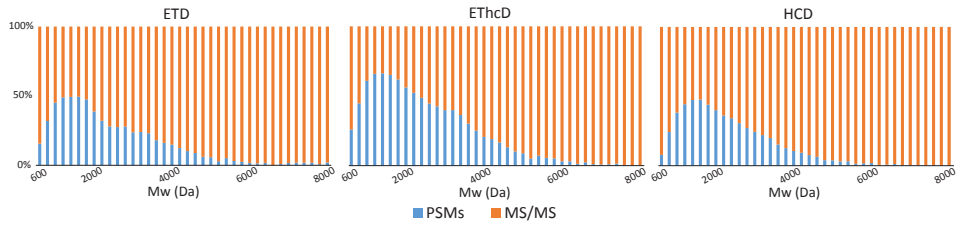


Figure 4. Ions selected for MS/MS fragmentation matched to PSMs binned by peptide Mw. The % (in blue) of precursors matched to PSMs is calculated for each Mw bin for all the different fragmentations after merging all datasets from the different digestions. The data clearly indicate the superior performance of EThcD at the higher Mw bins.

Effect of isotope deconvolution on Mascot and Sequest for the identification of middle-range peptides

Triggered by our analysis of success rate with respected to Mw, showing that a great part of the precursors with Mw > 4 kDa is not readily identified, we questioned if we could improve the identification of the high Mw precursors (thus highly charged too) by isotope deconvolution of the MS/MS spectra and use of alternative search engines. Fragmentation spectra of highly charged and high Mw precursors potentially contain multiply charged fragment ions, which are often weighted differently/weaker than singly charged ions by certain search engines.[63] We applied the H-Score script to our data,[64] searched with the digestion specificity that was dictated by our observation (summarized above, in Figure 5 for Asp-N

and similarly in S-9 for the other digestions). We compared the number of unique peptides identified, XCorr distribution and Mw median for increasing Mw ranges (Figure 5 for Asp-N, and S-9 for all data, Table S-6). Our analysis showed that deconvolution of the MS/MS leads to a decrease in XCorr distribution and number of unique peptides identified, while the median of Mw did not significantly increase. Most likely the decrease in performance of deconvoluted data is due to the cross-correlation algorithm used by Sequest which rewards data rich spectra irrespective of redundancy with respect to fragments present with a range of charge states.[65] We thus performed the same comparison, with the well-known, probability-based search engine Mascot. In this case, we obtained an increase of unique peptides identified, Mascot score distributions and Mw medians of identified peptides for all the digestions and fragmentation techniques (Figure 5 for Asp-N, and S-9 for all data, Table S-6). Interestingly, the score distribution was found to be very similar for all 3 fragmentation schemes, perhaps a product of the probability-based scoring. This clearly indicates that Mascot benefits of the isotope deconvolution of multiply charged ions as has been reported previously.[63] Nevertheless, the results were comparable to those obtained with Sequest and no deconvolution.

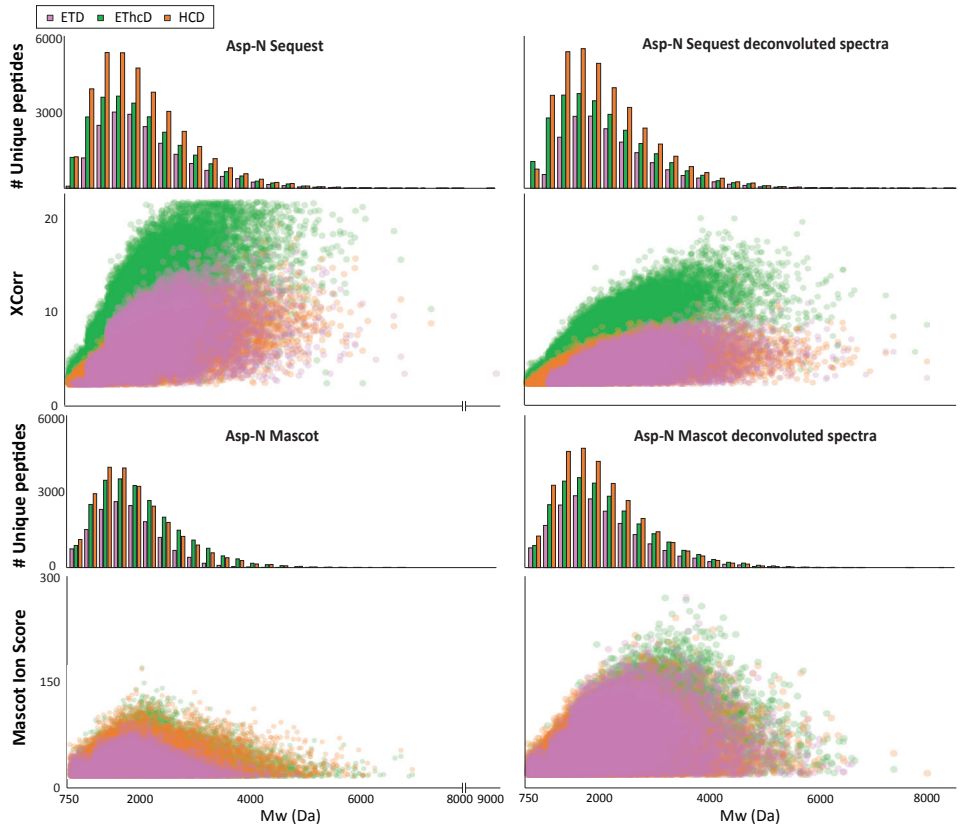


Figure 5. Effect of deconvolution on Sequest and Mascot searches for the identification of Asp-N middle-range peptides. Specific searches were performed by Sequest HT and Mascot on the Asp-N HeLa digestions data sets for each of the optimized fragmentation methods (ETD, EThcD and HCD). The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their Mw ranges for deconvoluted and non deconvoluted spectra searched by Sequest HT and Mascot.

Conclusions

Here, we present a step-by-step tuning of the "standard" bottom-up proteomics workflow with the aim to improve the generation, separation, detection and identification of middle-down range peptides. We first show that one of the main barriers for the development of middle-down proteomics is the generation of the desired peptides (3.5 kDa < Mw < 10 kDa). A substantial gain in median Mw of the peptides can be made by using Glu-C and Asp-N instead of trypsin, although the median Mw then still only reaches ~ 2 kDa. Alternatively, we generated peptides by formic acid-induced digestion. This approach, which had not yet been explored at a proteome-wide scale, provides an alternative for creating peptides, on par in number of achievable peptide and protein identifications with enzyme based approaches, although it suffers from efficiency/sensitivity issues. Interestingly, the average Mw of the peptides generated in the FA digest was found to be similar to that of the Asp-N and Glu-C digests. We next showed that the detection of these longer peptides could be improved by using an optimized SCX separation method and columns packed with larger pore size materials. In testing fragmentation methods, we demonstrate the excellent performance of EThcD for the analysis of middle-range sized peptides, not only presenting the highest identification rate (up to 57% success rate) but also providing higher peptide sequence coverage (up to 95% peptide sequence fragmentation coverage) compared to HCD and ETD fragmentations methods. Furthermore, the Mascot search engine benefited from deconvolution while Sequest showed a minor negative effect. Nevertheless, we also found that the number of effectively identified precursors seems to decrease at higher Mw > 4 kDa across all fragmentation techniques suggesting that the MS/MS sequencing may be the major limiting factor in our workflow. Cumulatively, the optimizations made here in peptide generation, separation, detection, fragmentation and identification, expand the number of unambiguous identifications of especially the middle-size range peptides and therefore, are a major step forward toward an optimized workflow for middle-down proteomics.

Corresponding author

A.J.R.H. (a.j.r.heck@uu.nl) and S.M. (shabaz.mohammed@chem.ox.ac.uk)

Notes

The authors declare no competing financial interest

Acknowledgements

The authors acknowledge all members of the Heck-lab and especially Soenita Goerdayal and Teck Yew Low for technical support and advice. This work was supported by The Netherlands Organisation for Scientific Research (NWO) via the Roadmap Initiative Proteins@Work (project number 184.032.201), the VIDI grant 700.10.429 for S.M. and the TOP-Punt Grant 718.015.003 for A.J.R.H. A.J.R.H. acknowledges additional support through the European Union Horizon 2020 programme FET-OPEN project MSmed, project number 686547.

SUPPORTING INFORMATION

Available: additional figures and tables containing detailed information of the optimization steps performed in the study. This material is available free of charge via the Internet at: http://pubs.acs.org.

References

(1) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R. Chem. Rev. 2013, 113 (4), 2343–2394.

(2) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Biochem. Biophys. Res. Commun. 2014, 445 (4), 683–693.

(3) Nesvizhskii, A. I.; Aebersold, R. Mol. Cell. Proteomics 2005, 4 (10), 1419–1440.

(4) Smith, L. M.; Kelleher, N. L.; Consortium for Top Down Proteomics. Nat. Methods 2013, 10 (3), 186–187.

(5) Siuti, N.; Kelleher, N. L. Nat. Methods 2007, 4 (10), 817–821.

(6) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. Nature 2011, 480 (7376), 254–258.

(7) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. J. Mass Spectrom. 2000, 35 (12), 1399–1406.

(8) Zubarev, R. A. Curr. Opin. Biotechnol. 2004, 15 (1), 12–16.

(9) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Proc. Natl. Acad. Sci. U. S. A. 2004, 101 (26), 9528–9533.

(10) Frese, C. K.; Altelaar, A. F. M.; van den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. Anal. Chem. 2012, 84 (22), 9668–9673.

(11) Morrison, L. J.; Brodbelt, J. S. J. Am. Chem. Soc. 2016, 138 (34), 10849–10859.

(12) Tamara, S.; Dyachenko, A.; Fort, K. L.; Makarov, A. A.; Scheltema, R. A.; Heck, A. J. R. J. Am. Chem. Soc. 2016, 138 (34), 10860–10868.

(13) Forbes, A. J.; Mazur, M. T.; Patel, H. M.; Walsh, C. T.; Kelleher, N. L. Proteomics 2001, 1 (8), 927–933.

(14) Wu, S.; Kim, J.; Hancock, W. S.; Karger, B. J. Proteome Res. 2005, 4 (4), 1155–1170.

(15) Garcia, B. A. J. Am. Soc. Mass Spectrom. 2010, 21 (2), 193–202.

(16) Wu, S.-L.; Kim, J.; Bandle, R. W.; Liotta, L.; Petricoin, E.; Karger, B. L. Mol. Cell. Proteomics 2006, 5 (9), 1610–1627.

(17) Zhang, J.; Wu, S.-L.; Kim, J.; Karger, B. L. J. Chromatogr. A 2007, 1154 (1–2), 295–307.

(18) Wu, C.; Tran, J. C.; Zamdborg, L.; Durbin, K. R.; Li, M.; Ahlf, D. R.; Early, B. P.; Thomas, P. M.; Sweedler, J. V; Kelleher, N. L. Nat. Methods 2012, 9 (8), 822–824.

(19) Laskay, Ü. A.; Lobas, A. A.; Srzentić, K.; Gorshkov, M. V; Tsybin, Y. O. J. Proteome Res. 2013, 12 (12), 5558–5569.

(20) Fornelli, L.; Ayoub, D.; Aizikov, K.; Beck, A.; Tsybin, Y. O. Anal. Chem. 2014, 86 (6), 3005–3012.

(21) Srzentić, K.; Fornelli, L.; Laskay, Ü. A.; Monod, M.; Beck, A.; Ayoub, D.; Tsybin, Y. O. Anal. Chem. 2014, 86 (19), 9945–9953.

(22) Valkevich, E. M.; Sanchez, N. A.; Ge, Y.; Strieter, E. R. Biochemistry 2014, 53 (30), 4979–4989.

(23) Garcia, B. A.; Siuti, N.; Thomas, C. E.; Mizzen, C. A.; Kelleher, N. L. Int. J. Mass Spectrom. 2007, 259 (1–3), 184–196.

(24) Moradian, A.; Kalli, A.; Sweredoski, M. J.; Hess, S. Proteomics 2014, 14 (4–5), 489–497.

(25) Shvartsburg, A. A.; Zheng, Y.; Smith, R. D.; Kelleher, N. L. Anal. Chem. 2012, 84 (10), 4271–4276.

(26) Tsiatsiani, L.; Heck, A. J. R. FEBS J. 2015, 282 (14), 2612–2626.

(27) Wu, S.-L.; Kim, J.; Hancock, W. S.; Karger, B. J. Proteome Res. 2005, 4 (4), 1155–1170.

(28) Boyne, M. T.; Garcia, B. A.; Li, M.; Zamdborg, L.; Wenger, C. D.; Babai, S.; Kelleher, N. L. J. Proteome Res. 2009, 8 (1), 374–379.

(29) Sidoli, S.; Schwämmle, V.; Ruminowicz, C.; Hansen, T. A.; Wu, X.; Helin, K.; Jensen, O. N. Proteomics 2014, 14 (19), 2200–2211.

(30) Sidoli, S.; Lin, S.; Karch, K. R.; Garcia, B. A. Anal. Chem. 2015, 87 (6), 3129–3133.

(31) Taverna, S. D.; Ueberheide, B. M.; Liu, Y.; Tackett, A. J.; Diaz, R. L.; Shabanowitz, J.; Chait, B. T.; Hunt, D. F.; Allis, C. D. Proc. Natl. Acad. Sci. U. S. A. 2007, 104 (7), 2086–2091.

(32) Swaney, D. L.; Wenger, C. D.; Coon, J. J. J. Proteome Res. 2010, 9 (3), 1323–1329.

(33) Taouatas, N.; Drugan, M. M.; Heck, A. J. R.; Mohammed, S. Nat. Methods 2008, 5 (5), 405–407.

(34) Hohmann, L.; Sherwood, C.; Eastham, A.; Peterson, A.; Eng, J. K.; Eddes, J. S.; Shteynberg, D.; Martin, D. B. J. Proteome Res. 2009, 8 (3), 1415–1422.

(35) Cannon, J.; Lohnes, K.; Wynne, C.; Wang, Y.; Edwards, N.; Fenselau, C. J. Proteome Res. 2010, 9 (8), 3886–3890.

(36) Li, A.; Sowder, R. C.; Henderson, L. E.; Moore, S. P.; Garfinkel, D. J.; Fisher, R. J. Anal. Chem. 2001, 73 (22), 5395–5402.

(37) Di Palma, S.; Hennrich, M. L.; Heck, A. J. R.; Mohammed, S. J. Proteomics 2012, 75 (13), 3791–3813.

(38) Helbig, A. O.; Gauci, S.; Raijmakers, R.; Van Breukelen, B.; Slijper, M.; Mohammed, S.; Heck, A. J. R. Mol Cell Proteomics 2010, 9 (5), 928–939.

(39) Giansanti, P.; Tsiatsiani, L.; Yew Low, T.; R Heck, A. J. Nat. Protoc. 2016, 11.

(40) Cristobal, A.; Hennrich, M. L.; Giansanti, P.; Goerdayal, S. S.; Heck, A. J. R.; Mohammed, S. Analyst 2012, 137 (15), 3541–3548.

(41) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Nat. Methods 2007, 4 (11), 923–925.

(42) Spivak, M.; Weston, J.; Bottou, L.; Käll, L.; Noble, W. S. J. Proteome Res. 2009, 8 (7), 3737–3745.

(43) Vizcaíno, J. A.; Csordas, A.; Del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H. Nucleic Acids Res. 2016, 44 (D1), D447-56.

(44) Molina, H.; Horn, D. M.; Tang, N.; Mathivanan, S.; Pandey, A. PNAS 2007, 104 (7), 2199–2204.

(45) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. Proc. Natl. Acad. Sci. U. S. A. 2002, 99 (12), 7900–7905.

(46) Chi, A.; Huttenhower, C.; Geer, L. Y.; Coon, J. J.; Syka, J. E. P.; Bai, D. L.; Shabanowitz, J.; Burke, D. J.; Troyanskaya, O. G.; Hunt, D. F. Proc. Natl. Acad. Sci. U. S. A. 2007, 104 (7), 2193–2198.

(47) Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Pääbo, S.; Mann, M. Mol. Syst. Biol. 2011, 7, 548.

(48) Drapeau, G. R. Methods Enzymol. 1977, 47, 189–191.

(49) Inglis, A. S. Methods Enzymol. 1983, 91, 324–332.

(50) Cannon, J. R.; Edwards, N. J.; Fenselau, C. J. Mass Spectrom. 2013, 48 (3), 340–343.

(51) Alam, A.; Mataj, A.; Yang, Y.; Boysen, R. I.; Bowden, D. K.; Hearn, M. T. W. Anal. Chem. 2010, 82 (21), 8922–8930.

(52) Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H. Mol. Cell. Proteomics 2014, 13 (6), 1573–1584.

(53) Wiśniewski, J. R.; Zougman, A.; Mann, M. J. Proteome Res. 2009, 8 (12), 5674–5678.

(54) Pearson, J. D.; Lin, N. T.; Regnier, F. E. Anal. Biochem. 1982, 124 (1), 217–230.

(55) Batth, T. S.; Olsen, J. V. Methods Mol. Biol. 2016, 1355, 179–192.

(56) Mohammed, S.; Heck, A. J. Curr. Opin. Biotechnol. 2011, 22 (1), 9–16.

(57) Di Palma, S.; Mohammed, S.; Heck, A. J. R. Nat. Protoc. 2012, 7 (11), 2041–2055.

(58) Scheltema, R. A.; Hauschild, J.-P.; Lange, O.; Hornburg, D.; Denisov, E.; Kuehn, A.; Makarov, A.; Mann, M.; Performance, H.; Chromatography, L.; Damoc, E. Mol. Cell. Proteomics 2014, 13 (12), 3698–3708.

(59) Frese, C. K.; Zhou, H.; Taus, T.; Altelaar, A. F.; Mechtler, K.; Heck, A. J.; Mohammed, S. J Proteome Res 2013, 12 (3), 1520–1525.

(60) Frese, C. K.; Maarten Altelaar, A. F.; Van Den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. Anal. Chem. 2012, 84 (22), 9668–9673.

(61) Riley, N. M.; Mullen, C.; Weisbrod, C. R.; Sharma, S.; Senko, M. W.; Zabrouskov, V.; Westphall, M. S.; Syka, J. E. P.; Coon, J. J. J. Am. Soc. Mass Spectrom. 2016, 27 (3), 520–531.

(62) Cannon, J. R.; Edwards, N. J.; Fenselau, C. J. Mass Spectrom. 2013, 48 (3), 340–343.

(63) Frese, C. K.; Altelaar, A. F. M.; Hennrich, M. L.; Nolting, D.; Zeller, M.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. J. Proteome Res. 2011, 10 (5), 2377–2388.

(64) Savitski, M. M.; Mathieson, T.; Becher, I.; Bantscheff, M. J. Proteome Res. 2010, 9 (11), 5511–5516.

(65) Eng, J. K.; McCormack, A. L.; Yates, J. R. J. Am. Soc. Mass Spectrom. 1994, 5 (11), 976–989.

Supporting information:

Contents:

SUPPLEMENTARY EXPERIMENTAL PROCEDURE

Chemicals and materials

Iodoacetamide was supplied by Sigma-Aldrich (Steinheim, DE). Ammonium bicarbonate (AMBIC) and dithiothreitol were purchased from Fluka (Buchs, CH), while urea and formic acid from Merck (Darmstadt, DE). Endoproteinase Asp-N and Glu-C, PhosSTOP phosphatase inhibitor cocktail tablets and complete mini EDTA-free cocktail tablets were obtained from Roche Diagnostics (Mannheim, DE). Bradford protein assay was supplied by Bio-Rad Laboratories (Hercules, CA, USA). For the fabrication of the trap and analytical columns the following chemicals were used: acetone and 2-propanol were supplied from Merck (Darmstadt, DE), and methanol HPLC grade from Biosolve B.V. (Valkenswaard, NL). The packing materials used were Zorbax SB-C18, 1.8 µm 80 Å and 300 Å pore size and 3.5 µm 300 Å from Agilent (Santa Clara, CA, USA) and ReprosilPur 120 Å C18, 3 µm from Dr. Maisch GmbH (Ammerbuch, DE). The reversed phase C18 300 Å solid phase extraction (SPE) columns were purchased from Grace Vydac (Columbia, MD, USA). The water used in all experiments was obtained from a Milli-Q purification system (Millipore, Bedford, MA, USA).

Sample preparation

HeLa digests were prepared as described previously:[38] prior to digestion the proteins were reduced (with dithiothreitol) and carbamidomethylated (with iodoacetamide). The protein concentration was estimated by a Bradford assay and subsequently the cell lysate was split into three samples of 200 µg for the digestion by the three applied digestion protocols. For preparing the Asp-N and Glu-C digests, the protocols reported by Giansanti *et al.* were used.[39]

Formic Acid non-enzymatic digestion

For the acid induced non-enzymatic digestion, the HeLa lysate was diluted to a final concentration of 0.1 µg/µl using a solution of formic acid (final concentration of 2% FA) and incubated at 100 °C. We tested incubation times from 30 min to 4 h, and studied parameters including cleavage specificity, total peptide identifications, relative yield of middle-range sized peptides and the unwanted occurrence of known side reactions (Table S-2).[36] The digests were analyzed in single runs and

the same amount of sample (based on starting material) was injected. As a result of this evaluation we selected a 1h incubation time due to the generation of a high number of middle-sized peptides while retaining a high cleavage specificity and keeping the occurrence of side reactions low (Table S-2). Of note, compared to enzymatic methods, the acid hydrolysis at any of the tried conditions provided lower sensitivity. According to literature, this phenomenon has been attributed to the distribution of peptide products which leads to a decrease in the absolute quantity of any single peptide due to the lower specificity of the method.[35,51] We cannot rule out this hypothesis, but we also believe that the overall digestion efficiency may also play an important role for this observed phenomenon, as we noticed a substantial increase in the signal of the UV trace at the end of the SCX gradient (see Figure s-3a), likely representing partially digested proteins.

Sample clean-up and pre-fractionation

Following digestion sample clean-up was performed using solid-phase extraction (SPE) columns; C18 with a 300 Å pore size. Prior to the MS analysis, samples were fractionated to reduce the complexity by using strong cation exchange (SCX) chromatography. Briefly, SCX was performed on an Agilent 1100 HPLC system (Agilent Technologies, Waldbronn, Germany) using a Zorbax BioSCX-Series II column (50 mm × 0.8 mm, 250 Å 3.5 µm). SCX solvent A consisted of 0.05% formic acid in 20% acetonitrile, while solvent B was 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. ~200 µg peptides were dissolved in 10% FA and loaded onto the SCX column with buffer A. Special attention was given not only to the pore size of the SCX material, but we also applied a modulated gradient which favors the separation of highly charged peptides. In this way, the elution window of peptides with $z < +4$ is reduced, favoring the separation and collection of higher charged peptides (Figure s-3a). In more detail the following gradient was used: 0–5 min (0% B); 5-7 min (0-2% B); 7-15 min (2-3% B); 15-25 min (3-8% B); 25-35 min (8-20% B); 35-45 min (20-40% B); 45-51 min (40-90% B); 51-55 min (90-90% B); 55-56 min (90-0% B) and 56-100 min (0% B). A total of 50 SCX fractions were collected, pooled into 11 fractions and dried in a vacuum centrifuge.

LC-MS and LC-MS/MS set up

Nano-UHPLC-MS/MS was performed on an Agilent 1290 Infinity System (Agilent

Technologies, Waldbronn, DE) connected to an Orbitrap Fusion (Thermo Fisher Scientific, San Jose, CA). Fused-silica capillary analytical and trap columns were prepared as previously described.[40] The UHPLC was equipped with a double frit trapping column and a single frit analytical column. A ReprosilPur C18 (3 µm particles, 120 Å pore size 2 cm x 100 µm) was used as a trap column, and Zorbax SB-C18 (1.8 µm particles 80 Å 40 cm x 50 µm) for the analytical column. For the 300 Å pore size set up the used materials were: Zorbax SB-C18 (3.5 µm particles, 300 Å pore size, 2 cm x 100 µm) for the trap, and Zorbax SB-C18 (1.8 µm particles 300 Å 40 cm x 50 µm) for the analytical column. The column, in both cases, was directly connected to an in-house pulled and gold-coated fused silica needle (with a 5 µm o.d. tip). In both systems the generated back pressure was comparable, although the conventional RP C18 material blocked several times when middle-sized peptides were analyzed, likely due to poor mass transfer and precipitation. A voltage of 2.0 kV was applied to the needle and the ion transfer tube temperature was increased to 275 degrees. The survey scan range was from 350 to 1500 m/z at a resolution of 60000 (200 m/z) with an AGC target of 4e5. The most intense precursor ions were selected for subsequent fragmentation at Top Speed within a 3 seconds duty cycle. A resolution of 30000 (200 m/z) and a maximum injection time of 125 ms were found to be ideal for MS/MS. The AGC target for the MS/MS was set to 1e5. When HCD was used 35% collision energy (CE) was applied, in the case of EThcD 40% supplemental activation (SA) was selected and when ETD was used 10% SA was applied. Additionally charge triggered MS/MS, instead of intensity triggered, was tested for the EThcD charge method.

Data analysis

The RAW files were processed using Proteome Discoverer (PD, version 2.1, Thermo Scientific, Bremen, DE) and the spectra were searched against the UniProt human database (version 2015_04). Searching was was performed using Sequest HT and the following parameters were used: unspecific searches with cysteine carbamidomethylation as fixed modification and oxidation of methionine as dynamic modifications. In the case of the FA induced digestion two additional dynamic modifications were included: formylation of the N-terminus and the conversion from Glutamate to pyro-Glutamate. Specific searches were performed by Sequest HT and Mascot (version 2.5.1, Matrix Science, London, UK) using the

same modifications. Peptide tolerance was set to 10 ppm and MS/MS tolerance was set to 0.05 Da. The results were filtered using Percolator[41,42] to a peptide and protein FDR < 1%. We further only accepted peptides with an Xcorr of at least 2. We performed an in-silico digest for the overall population of observed peptides for each enzyme, taking missed cleavages into account. The median of the peptide masses was then calculated in R.[65] The parameters used to perform the in-silico digestion were: Glu-C cleavage C-terminal of E with maximum 2 missed cleavages. Asp-N cleaves N-terminally of DE and maximum 4 missed cleavages were allowed. Peptide sequence fragment coverage was calculated using in-house developed scripts. Theoretical ion series were calculated for each fragmentation method (b and y for CID and HCD, c and z for ETD and b, y, c and z for EThcD). Matching was performed with a tolerance of 0.05 Da, for peaks with intensities higher than 5% of the base peak. The global fragmentation coverage was calculated based on all possible fragments, disregarding the exact breakage positions. The H-score script was used to deconvolute the mgf files exported from PD.[63]

Supplementary Figures

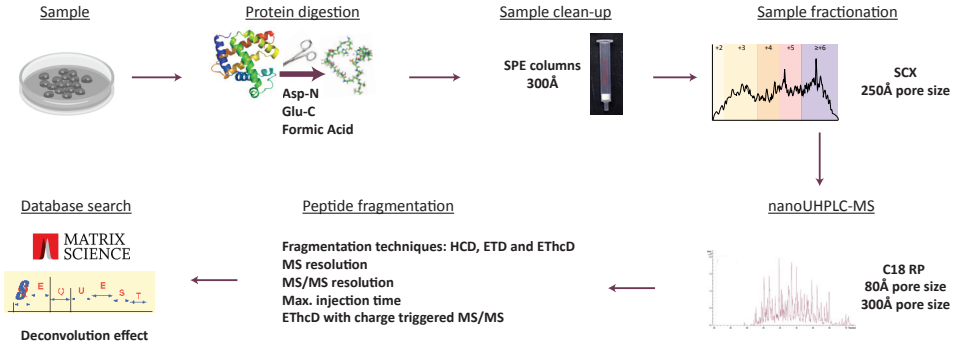


Figure S-1. Detailed experimental flow chart of the tested parameters in order to create an optimized workflow for middle-down proteomics.

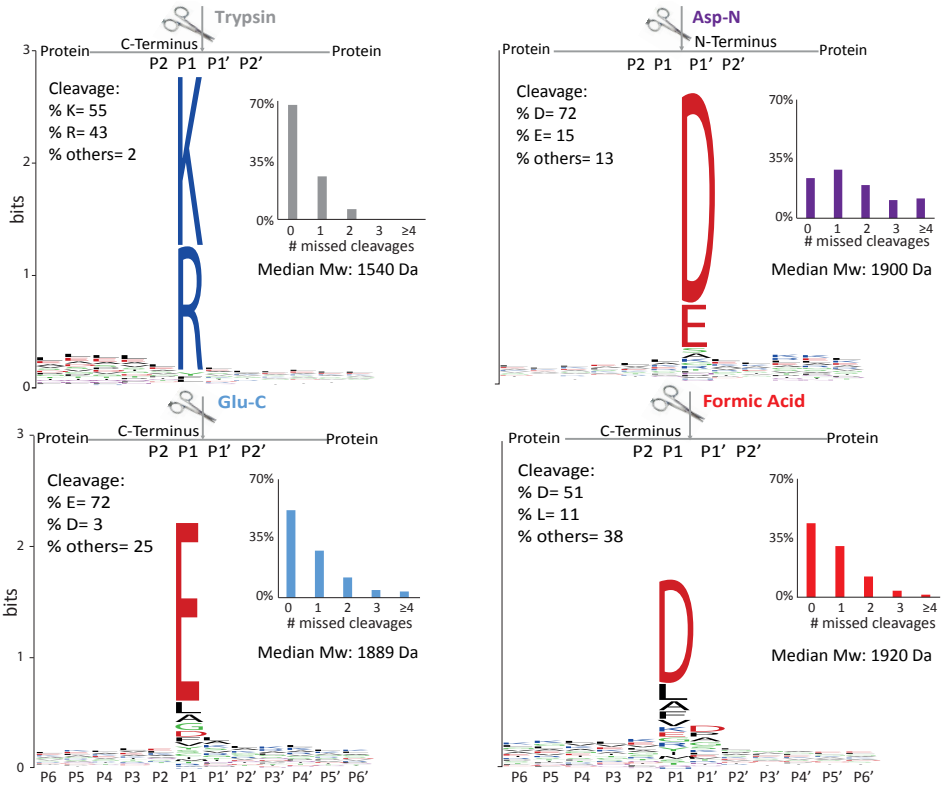


Figure S-2. Cleavage specificity and number of missed cleavages observed in trypsin, Asp-N, Glu-C and formic acid HeLa lysate digests. The cleavage specificity, displayed by relative

frequency of occurrence, is depicted at the C-terminal (P6 to P1) and N-terminal (P1' to P6') end of the cleavage site. Asp-N showed a high cleavage specificity for the N-termini side of Aspartate (D) residues (72%) and lower cleavage frequency at the N-termini side of Glutamate (E) (15 %). At a pH of approx. 8 and in ammonium bicarbonate buffer Glu-C mainly cleaves at the C-terminal side of Glutamate (E) residues (72%). The inset in each panel displays the proportion of missed cleavages and the median Mw of all identified peptides.

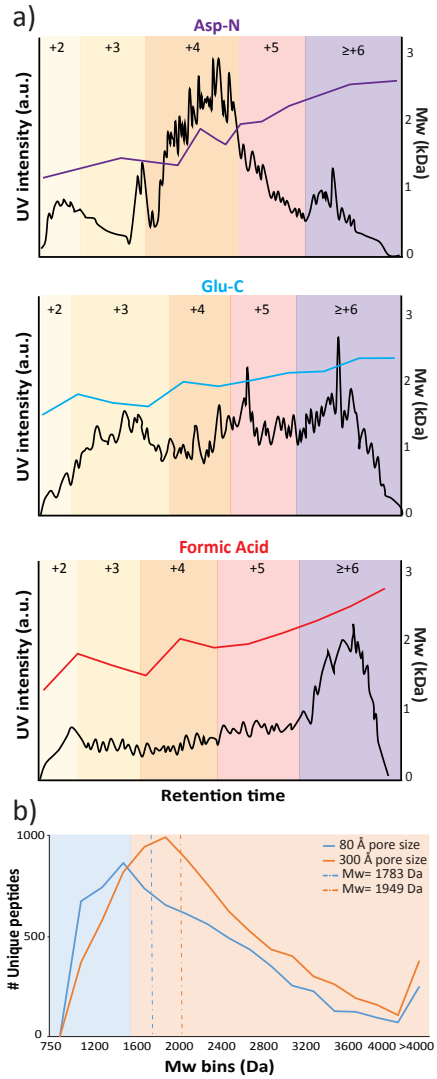


Figure S-3. Performance of the optimized SCX separation and the effect of the material pore size for the analysis of middle-range sized peptides.

The UV traces (arbitrary unit) detected in the SCX separations of the, from top to bottom, Asp-N, Glu-C and FA digestions, together with the in solution charge state distribution of peptides identified in the SCX fractions. The correlation between the charge state and the median of the molecular weight for the different SCX fractions is also represented, using the secondary y-axis of the graphs.

Peptide Mw distribution observed by using the conventional (80 Å) and the larger pore size (300 Å) materials for the columns in the UHPLC system. Eleven SCX Asp-N fractions were analyzed and the median Mw in these fractions was calculated using the uniquely identified peptides.

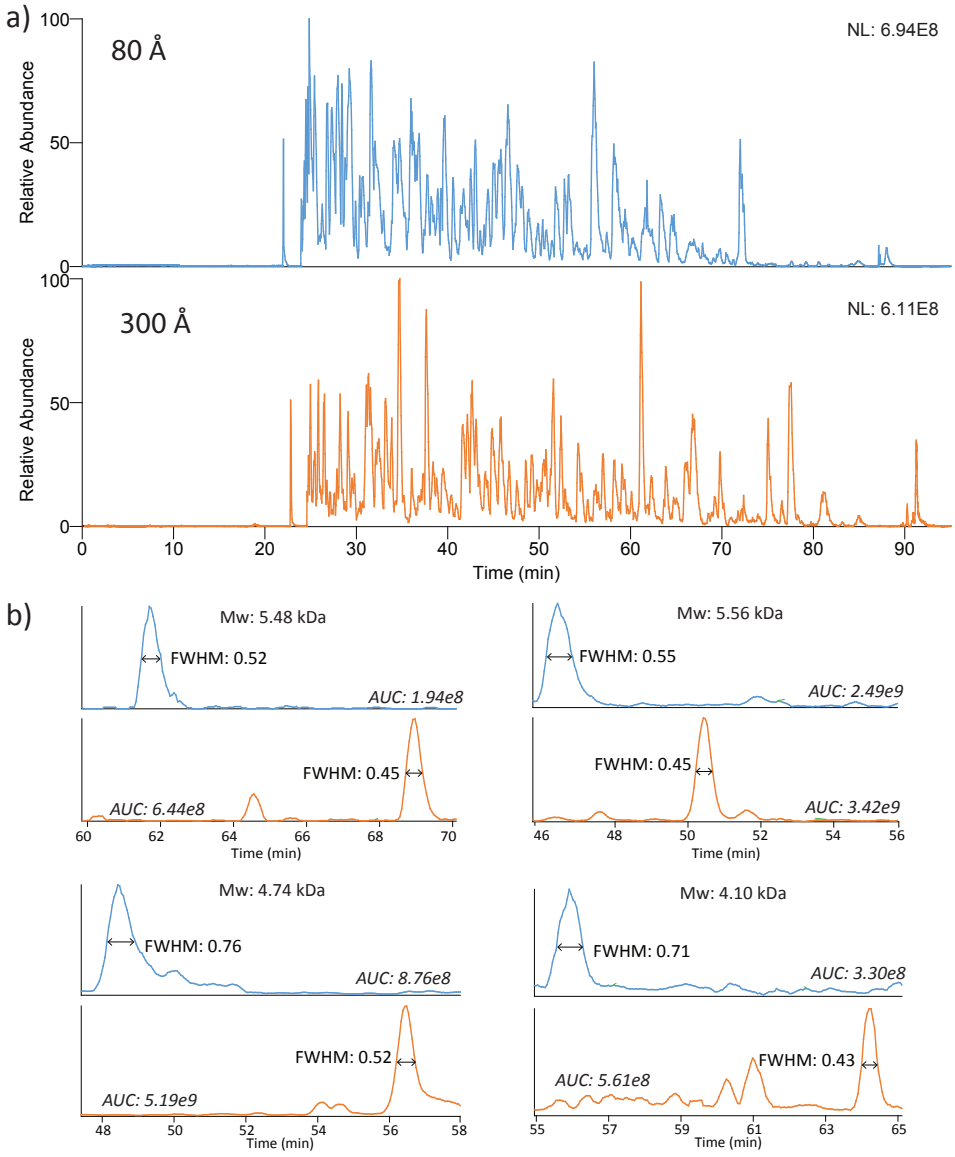


Figure S-4. Performance of the 80 and 300 Å pore size material in RP chromatography.

Chromatograms obtained by injecting the same amount of starting material (late SCX fraction of the Asp-N digest) analyzed on the 80 Å (upper panel, blue line) and the 300 Å (bottom panel, orange line) pore size columns.

Illustrative extracted ion chromatograms of four peptides exhibiting a Mw > 4 kDa analyzed by using either the 80 Å (blue line) or 300 Å (orange line) pore size columns. The area under the curve (AUC) and the full weight at half maximum (FWHM) are depicted in the panels.

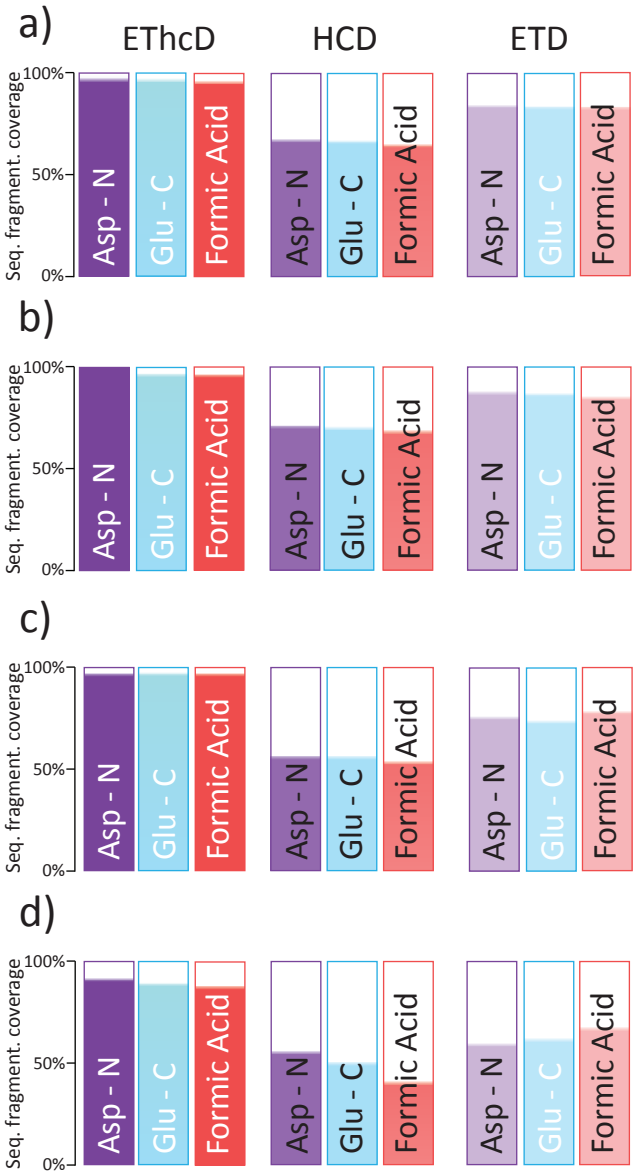


Figure S-5. Peptide sequence fragmentation coverage obtained by each fragmentation method in the three applied digestion schemes. The median peptide sequence fragmentation coverage was calculated and represented taking into consideration a) the whole dataset, b) peptides with 0 < Mw < 2.5 kDa, c) peptides with 2.5 < Mw < 4 kDa and d) peptides with Mw > 4 kDa.

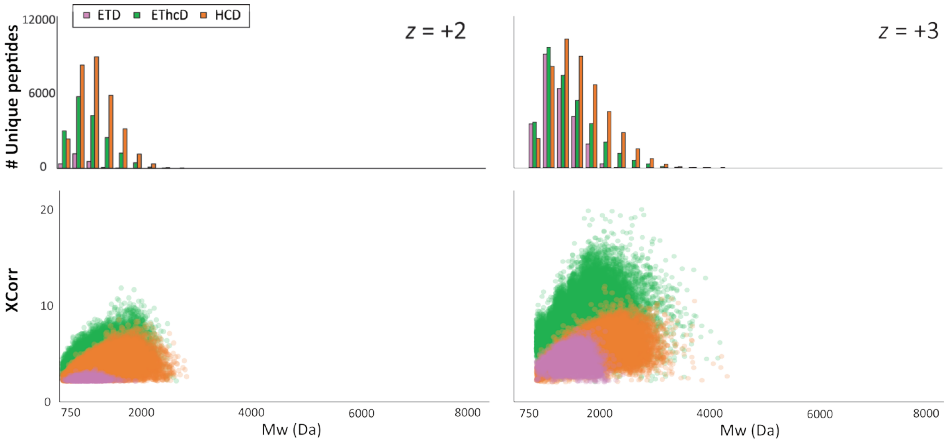
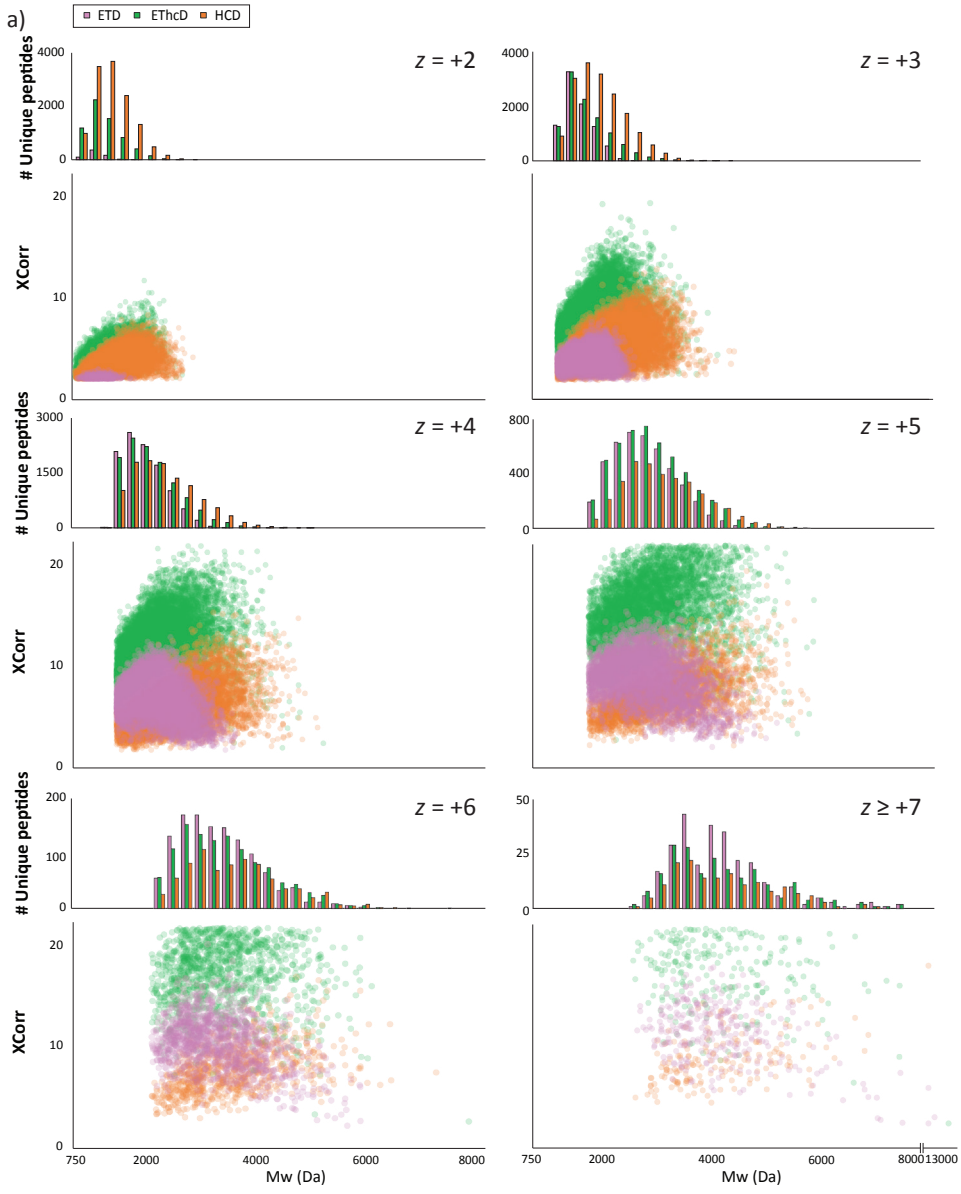
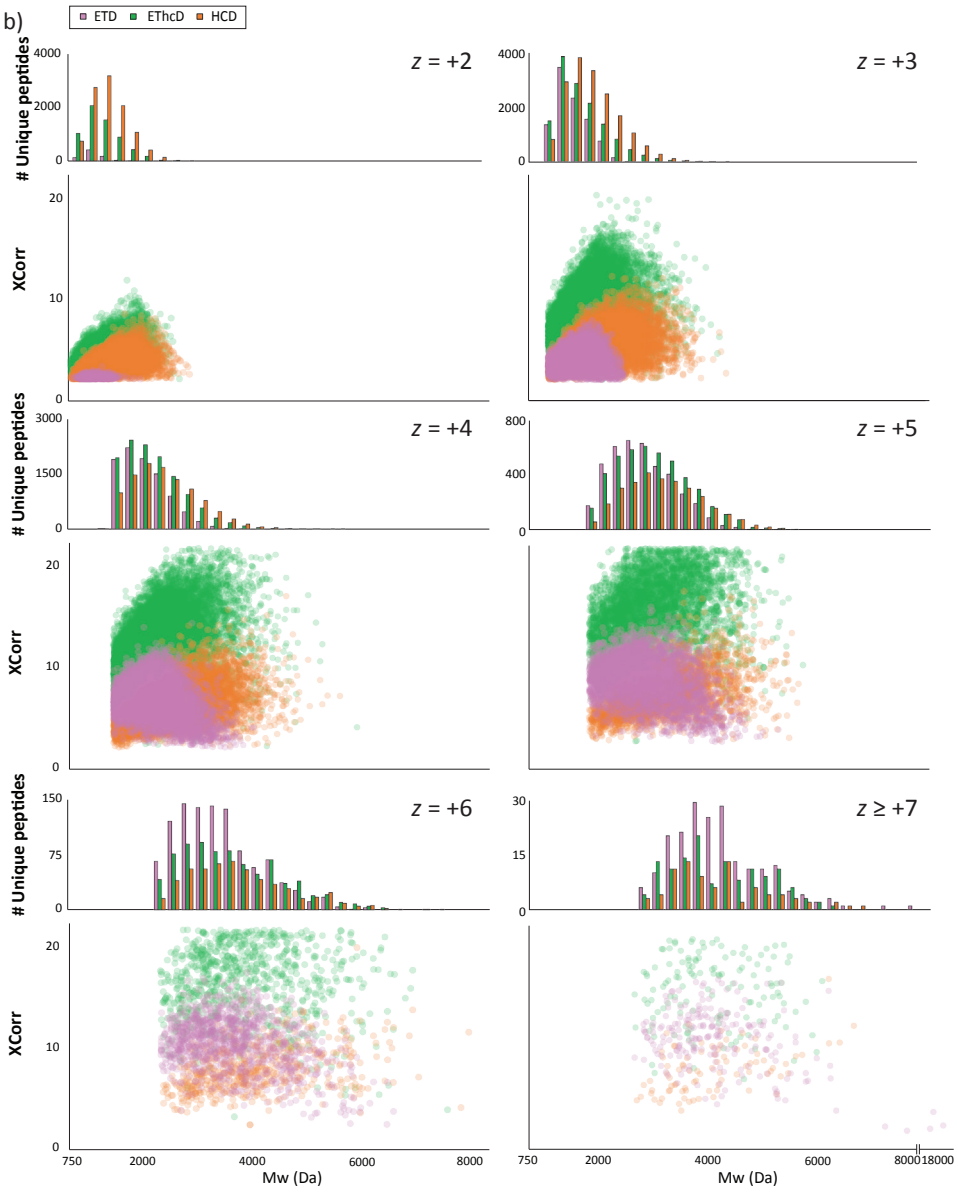


Figure S-6. Performance of the peptide fragmentation techniques ETD, EThcD and HCD for z = +2 and z = +3 peptides binned by their different Mw values. Combined data from Asp-N, Glu-C and FA induced HeLa digestions analyzed under the same LC settings, with optimized fragmentation parameters for ETD, EThcD and HCD. The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their z and Mw ranges.


a)
ETD
EThcD
HCD
Unique peptides
XCorr
z = +2
z = +3
z = +4
z = +5
z = +6
z ≥ +7
Mw (Da)
Mw (Da)


b)
ETD
EThcD
HCD
z = +2
z = +3
z = +4
z = +5
z = +6
z ≥ +7
Unique peptides
XCorr
Mw (Da)
Mw (Da)

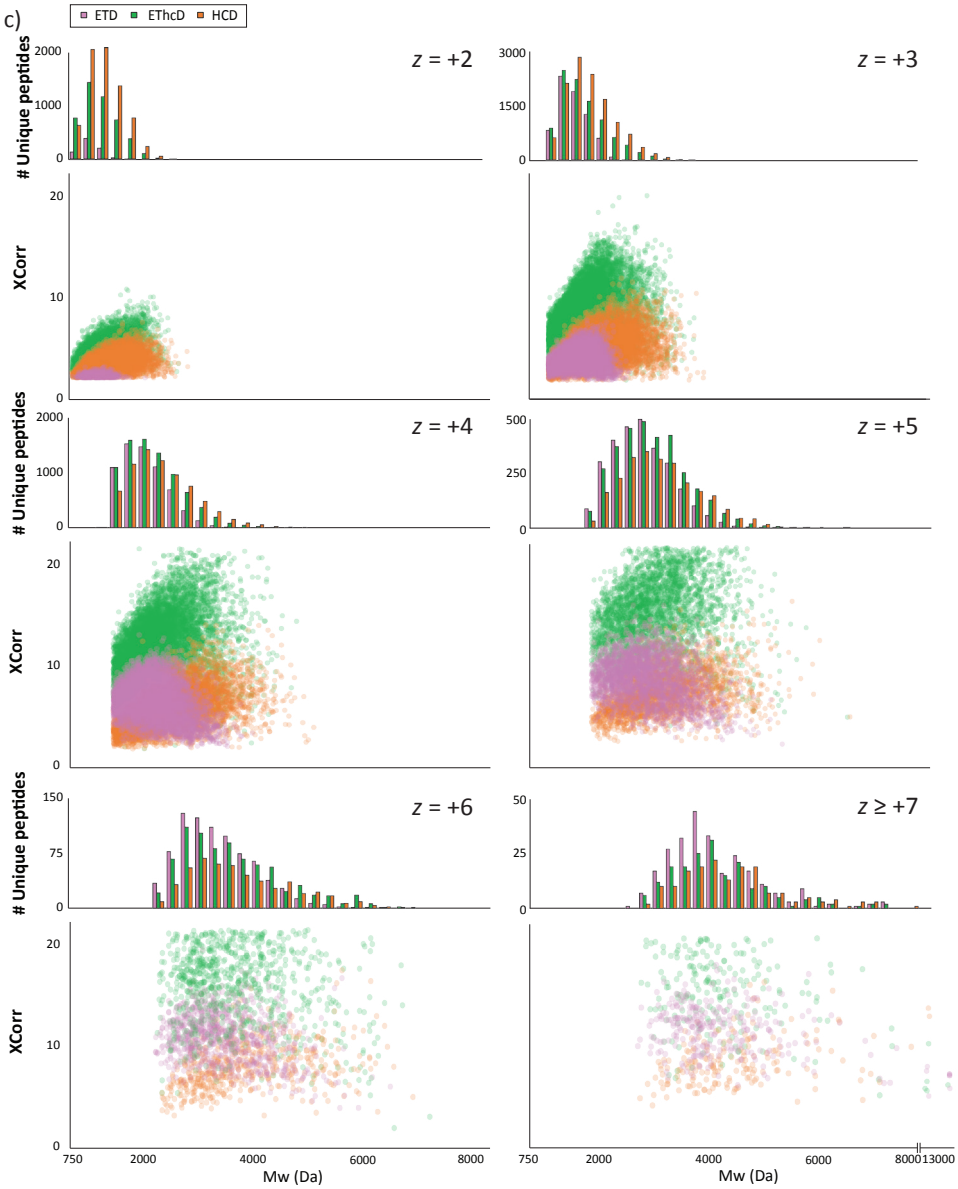


Figure S-7. Performance of the peptide fragmentation techniques ETD, EThcD and HCD for each digestion data set (Asp-N, Glu-C and FA) binned by their different Mw values. Data from Asp-N (a), Glu-C (b) and FA (c) induced HeLa digestions analyzed under the same LC settings, with optimized fragmentation parameters for ETD, EThcD and HCD. The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their z and Mw ranges.

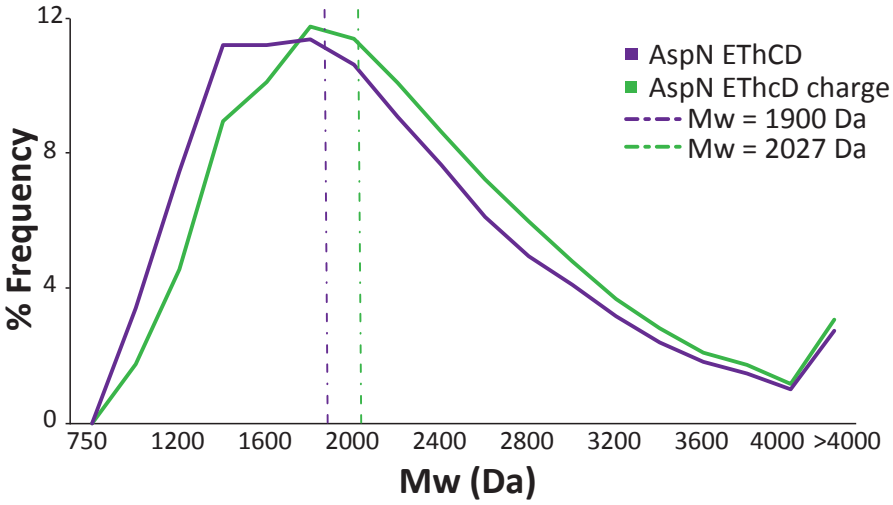
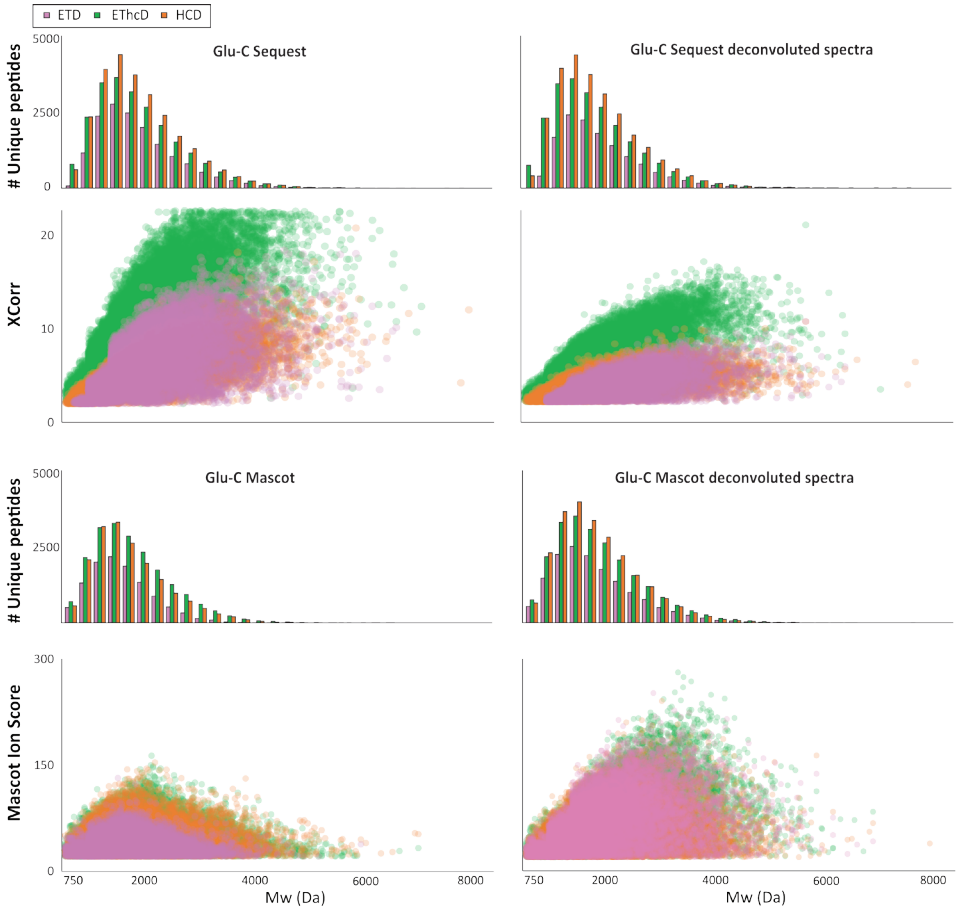


Figure S-8. Comparison of the Mw distribution of the identified peptides in conventional EThcD and EThcD with charge triggered MS/MS. Frequency distribution of the Mw of the identified peptides by EThcD (purple) and EThcD with charge triggered MS/MS (green) when the eleven Asp-N SCX fractions are analyzed using identical experimental conditions. The median of the Mw is represented by dashed lines.


ETD
EThcD
HCD
Glu-C Sequest
Glu-C Sequest deconvoluted spectra
Unique peptides
5000
2500
0
XCorr
20
10
0
Glu-C Mascot
Glu-C Mascot deconvoluted spectra
Unique peptides
5000
2500
Mascot Ion Score
300
150
0
750
2000
4000
6000
8000
Mw (Da)
750
2000
4000
6000
8000
Mw (Da)

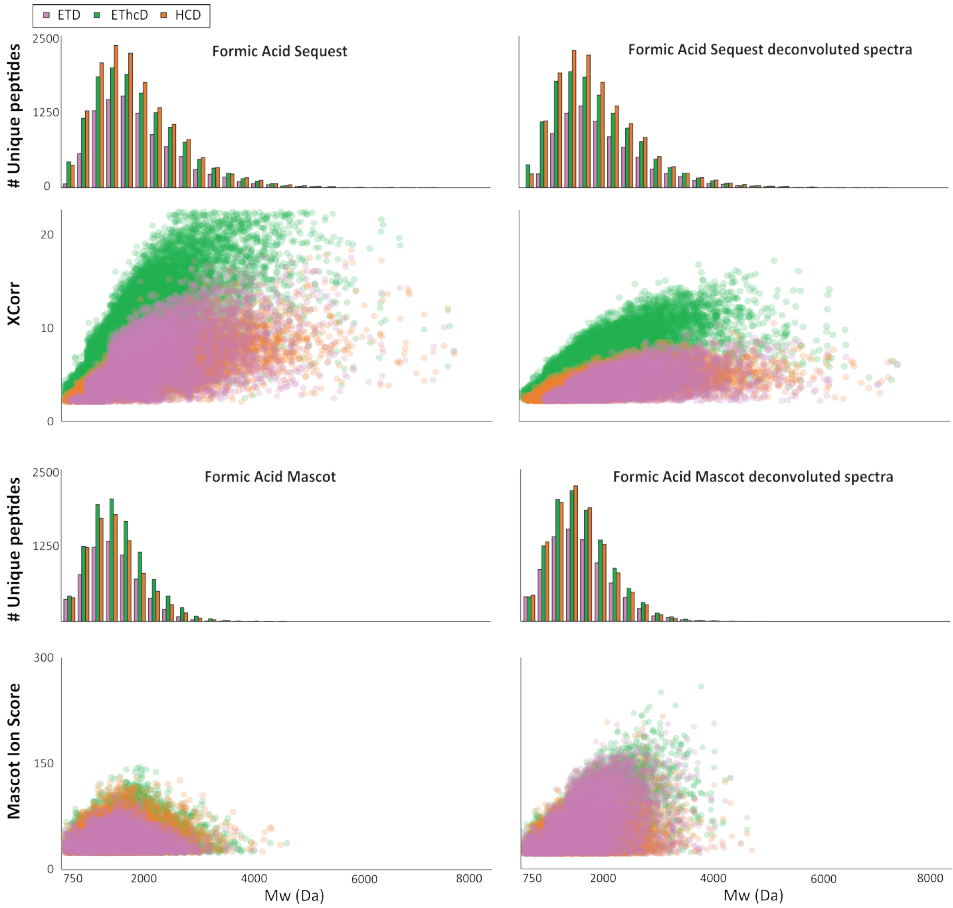


Figure S-9. Effect of deconvolution on Sequest and Mascot searches for the identification of middle-range peptides. Specific searches were performed by Sequest HT and Mascot on Glu-C (a) and FA (b) induced HeLa digestions data sets for each of the optimized fragmentation methods (ETD, EThcD and HCD). The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their Mw ranges for deconvoluted and non deconvoluted spectra search by Sequest HT and Mascot.

Table legend

Table S-1. Summary of the protein and peptide identifications obtained for the Asp-N, Glu-C and FA initiated digest using the three different fragmentation modes (EThcD, HCD and ETD). The identifications rate (Id rate) and the median Mw are also depicted.

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)
Asp-N	EThcD	4895	34307	100695	178059	56.6	1900.0
Asp-N	HCD	5099	44239	133823	349702	38.3	1870.9
Asp-N	ETD	4681	25228	66643	185089	36	1952.1
Glu-C	EThcD	4354	37141	121197	211235	57.4	1888.0
Glu-C	HCD	4374	41195	134990	372620	36.2	1888.8
Glu-C	ETD	3935	24833	67447	192643	35.0	1913.0
Formic Acid	EThcD	3054	26520	63791	186812	34.2	1920.0
Formic Acid	HCD	2942	29515	72166	343772	21.0	1891.0
Formic Acid	ETD	2693	18107	36147	170125	21.3	1930.9

Table S-2. Performance of the FA induced digestion at different incubation times. The table depicts also the percentage of observed cleavages at the C-terminus of D, the number of identified unique peptides, the median Mw of all detected peptides and the percentage of the unwanted side reactions observed.

Digestion	Incubation Time	% C-Term D	Unique peptides	Median Mw (Da)	% Side reactions
Formic Acid	30 min	41	5755	2009	1.0
Formic Acid	1 h	51	6000	1950	1.3
Formic Acid	2 h	52	6523	1919	2.1
Formic Acid	3h	48	6485	1920	3.0
Formic Acid	4h	43	6351	1916	4.3

Table S-3. Summary of the optimization of the resolution both at the MS and MSMS level, including the number of identifications, quality of the spectra and median Mw of SCX fractions containing peptides of different charge states.

	MS 120000 MS/MS 30000				MS 60000 MS/MS 15000			
Charge state	Xcorr	Unique Peptides	Mw (Da)	MS/MS	Xcorr	Unique Peptides	Mw (Da)	MS/MS
+3	6	1486	1439	11299	6	1463	1423	11623
+4	8	2162	2389	12802	8	2023	2349	13331
+5	10	2547	2459	13472	9	2368	2425	14302
+6	9	1722	2648	11420	9	1697	2620	12463

	MS 120000 MS/MS 30000				MS 60000 MS/MS 30000			
Charge state	Xcorr	Unique Peptides	Mw (Da)	MS/MS	Xcorr	Unique Peptides	Mw (Da)	MS/MS
> +5	13	3049	2607	17895	13	3190	2604	19095
> +5	13	1598	2893	16239	13	1623	2893	16943
> +5	12	860	3073	15516	11	934	3051	16206

Table S-4. Summary of the optimization for the injection times, including the number of identifications, quality of the spectra and the median peptide sequence coverage.

Max. injection time (ms)	Xcorr	Unique Peptides	# MS/MS	Peptide sequence coverage (%)
75	6	5429	20977	88
100	7	5433	19103	89
125	7	5389	17884	91

Table S-5. Summary of the parameters optimized for the different fragmentation techniques.

Fragmentation	SA/CE	MS resolution	MSMS resolution	Max. injection time (ms)
ETD	10	60000	30000	125
EThcD	40	60000	30000	125
HCD	35	60000	30000	125

Table S-6. Summary of the effect of deconvolution on Mascot and Sequest for the identification of middle-range peptides. Specific searches were performed by Sequest on non deconvoluted (a) and on deconvoluted spectra (b) as well as by Mascot on non deconvoluted (c) and on deconvoluted spectra (d). The number of peptides and proteins and summarized as well as the identifications rate (Id rate) and the median Mw and Score (XCorr and Ion Score, respectively).

a)

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5554	26688	82663	178059	46.4	1867	9.6
Asp-N	HCD	6722	36307	119994	349702	34.3	1831	4.5
Asp-N	ETD	5069	19147	54451	185089	29.4	1958	6.2
Glu-C	EThcD	4723	22892	80802	211235	38.3	1852	10.1
Glu-C	HCD	5272	25639	95203	372620	25.5	1860	4.7
Glu-C	ETD	4188	15486	44267	192643	23.0	1894	5.6
Formic Acid	EThcD	3265	12922	33314	186812	17.8	1915	9.5
Formic Acid	HCD	3751	14364	38332	343772	11.2	1894	4.7
Formic Acid	ETD	2885	8935	18812	170125	11.1	1954	5.5

b)

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5445	26421	81341	178059	45.7	1884	6.2
Asp-N	HCD	6810	35728	116324	349702	33.3	1868	3.4
Asp-N	ETD	5105	17312	49108	185089	26.5	2034	3.6
Glu-C	EThcD	4674	22905	79831	211235	37.8	1858	6.4
Glu-C	HCD	5246	25902	93664	372620	25.1	1877	3.5
Glu-C	ETD	4128	13205	37272	192643	19.3	1991	3.5
Formic Acid	EThcD	3294	12968	33325	186812	17.8	1932	6.3
Formic Acid	HCD	3755	14273	37904	343772	11.0	1936	3.5
Formic Acid	ETD	2850	7876	16397	170125	9.6	2050	3.5

c)

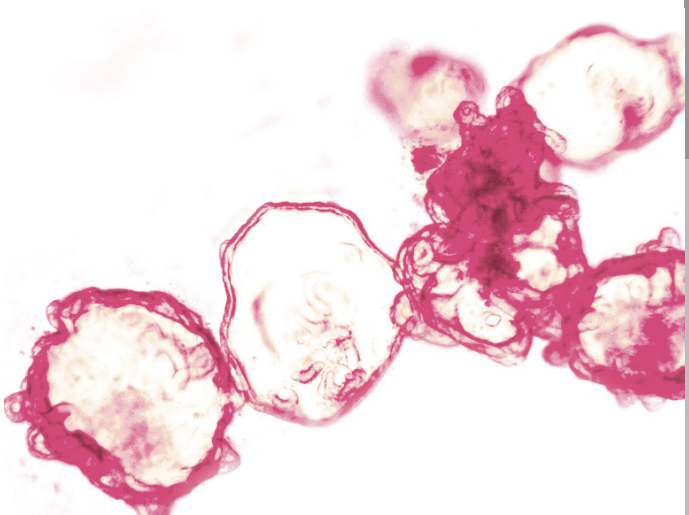
Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5218	23228	83580	178059	46.9	1833	43.1
Asp-N	HCD	6306	23580	117487	349702	33.6	1728	38.2
Asp-N	ETD	4781	14278	60471	185089	32.7	1738	35.7
Glu-C	EThcD	4578	20724	82673	211235	39.1	1809	44.3
Glu-C	HCD	4978	18771	96786	372620	26.0	1741	39.3
Glu-C	ETD	3952	11655	49602	192643	25.7	1703	36.0
Formic Acid	EThcD	2908	9922	29198	186812	15.6	1661	42.5
Formic Acid	HCD	3222	8244	32646	343772	9.5	1605	35.7
Formic Acid	ETD	2524	6184	18839	170125	11.1	1631	36.2

d)

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5292	25631	86196	178059	48.4	1913	66.1
Asp-N	HCD	6492	31033	130986	349702	37.5	1827	49.3
Asp-N	ETD	4985	19314	69766	185089	37.7	1903	62.2
Glu-C	EThcD	4553	22700	85306	211235	40.4	1878	67.4
Glu-C	HCD	5104	23870	106082	372620	28.5	1846	49.7
Glu-C	ETD	4089	15354	56311	192643	29.2	1846	62.3
Formic Acid	EThcD	2928	10721	29705	186812	15.9	1712	58.1
Formic Acid	HCD	3388	10544	34862	343772	10.1	1684	43.6
Formic Acid	ETD	2610	7726	20546	170125	12.1	1715	56.8

Chapter 5

Summary, samenvatting, future outlook, curriculum vitae, publications, acknowledgements



Summary

The major aim of the work presented in this thesis was to generate personalized proteomics profiles by improving the chromatographic aspects of the proteomic experiment.

In the first chapter an overview of proteomics is given and several practical aspects of a proteomic workflow are highlighted. An important aspect is the high complexity and the intrinsic dynamic range of the samples analyzed in proteomics. In order to reduce this complexity and perform a comprehensive study a separation strategy (involving one or more separation steps) is required. A detailed description of reversed phase chromatography is given in this chapter as is the most widely used chromatographic methodology in proteomics. Furthermore, an overview of mass spectrometers is given which includes details of both ionization sources and mass analyzers. Protein quantification is also described as in chapter 3 a quantitative proteomics workflow was applied. A separate section for intestinal stem cell biology describes the characteristics of adult stem cells, in general, and of intestinal stem cells, in particular. The impact of the identification of intestinal stem cells is also addressed and how this led to the development of organoids. Moreover, applications of organoids to investigate disease, colorectal cancer in this case, is further explored.

In the second chapter we aimed to build an in-house ultra-high pressure liquid chromatography system with a high separation efficiency and resolution. An easy to implement design was constructed. Particles with small diameter (1.8 μm) were used in the stationary phase with the aim of obtaining an increased efficiency compared to conventionally used 3 μm particles. Thanks to advancements in HPLC system design, particularly with respect to generating and maintaining pressures, we were able to implement columns containing smaller particles (1.8 μm) in dimensions of 50 μm x 40 cm. Several flow rates were studied and gradients ranging from 23 to 458 minutes were optimized. The level of performance of this system was tested by coupling it to an Orbitrap Velos achieving over 4500 proteins with the longest analysis time. Furthermore, a mass spectrometer with faster sequencing speed was

used (TripleTOF) which enables us to identify over 1400 proteins in a 23-minute gradient.

In the third chapter we performed personalized proteome profiles of healthy and tumor human colon organoids. Peptides were extracted from the organoid samples and dimethyl labeled in order to obtain quantitative information. The optimized reversed phase chromatography described in chapter 2 was used here after SCX prefractionation, to perform a comprehensive proteomics study. Furthermore, two state-of-the art mass spectrometers were used which allowed us to perform optimal peptide sequencing for each SCX fraction. Proteins extracted from organoids grown from seven patients were analyzed and the data was combined with transcriptomics data. Although generic features of colorectal cancer could be obtained across all the patients, our data clearly revealed that each patient possesses a distinct organoid signature at the proteomic level.

Key to a comprehensive proteomics study is knowing how PTMs are interconnected, it is important to get an insight into the specific proteoforms present. This is a problem in bottom-up shotgun proteomics approaches since this interconnectedness is removed by the proteolysis step and the generation of peptides. Sadly, it is currently suboptimal to perform full protein sequencing (top down) in a high-throughput manner. A compromise is found in middle-down proteomics, which tries to sequence larger peptides (ideally above 3 kDa) in a shotgun-type experiment. In the fourth chapter we set out to optimize a workflow for the emerging technology of middle-down proteomics. Starting from the sample preparation to the state-of-the-art chromatographic separations and alternative MS fragmentation techniques for the identification of larger peptides are covered. The improvements made to the workflow clearly boost the detection and sequence coverage of middle-range peptides making a step forward towards the identification of specific proteoforms.

Samenvatting

Het hoofddoel van het hier gepresenteerde werk is het genereren van gepersonaliseerde profielen op basis van eiwitonderzoek, in het bijzonder door het verbeteren van de chromatografische aspecten van de experimentele methoden.

In het eerste hoofdstuk wordt een overzicht gegeven van het onderzoek naar de eiwitinhoud in (biologische) specimen (verder hieronder proteomics genoemd), en een aantal praktische aspecten van de proteomics methode worden nader belicht. Een belangrijk probleem is de hoge complexiteit en het de kwantitatieve variatie van eiwit die worden aangetroffen in biologische monsters. Om deze complexiteit te reduceren en een zo compleet mogelijke studie uit te voeren, worden chromatografische methodes gebruikt voor de scheiding van het materiaal. In dit hoofdstuk wordt omgekeerde-fase chromatografie beschreven, als een van de meest voorkomende methodes in proteomics. Daarnaast wordt een overzicht gegeven van de massaspectrometers, met de ionen-bronnen en massa-analysatoren. De kwantificatie van eiwitten wordt ook beschreven als voorbereiding op hoofdstuk 3 waar deze techniek wordt gebruikt. Het biologische aspect van stamcellen van de darm wordt besproken, met het ontdekken van de stamcellen, en hoe dit leidde tot het ontwikkelen van de micro-organen in het laboratorium, de zogenaamde organoïden. De toepassing van deze organoïden in het onderzoek naar ziekte, in dit geval colorectale kanker, wordt beschreven.

In het tweede hoofdstuk wordt een door ons ontwikkeld ultrahogedruk vloeibare-chromatografiesysteem met een hoge scheidingsefficiëntie en een hoog oplossend vermogen beschreven. Niettegenstaande deze positieve eigenschappen is het ontwerp eenvoudig en is de methode makkelijk uit te voeren. Een hogere efficiëntie wordt bereikt door het gebruik van korrels met een kleine diameter (1.8 µm) in de stationaire fase, vergeleken bij de conventionele korreltjes met een diameter van 3 µm. Dankzij innovaties op het gebied van op het gebied van het maken en stabiliseren van de druk in ultrahogedruk vloeibare-chromatografiesystemen (UHPLC), waren we in staat 1.8 µm korrels in scheidingskolommen van 50 µm x 40 cm te gebruiken. Verschillende stroomsnelheden werden getest en oplosmiddelgradiënten tussen de 23 en 458 minuten werden geoptimaliseerd. De prestaties werden getest door de kolom te koppelen aan een Orbitrap Velos massaspectrometer, waarbij meer dan 4500 eiwitten konden worden geïdentificeerd bij de langste analysetijd. Daarnaast

werd een massaspectrometer met een hogere identificatiefrequentie (TripleTOF) getest, waarbij bleek dat met een gradiënt van slechts 23 minuten alsnog meer dan 1400 eiwitten konden worden geïdentificeerd.

In het derde hoofdstuk keken we naar de persoonlijke eiwitprofielen op basis van verscheidene organoïden, die waren gekweekt uit gezond en tumorweefsel van dezelfde patiënt. Eiwitfragmenten werden geëxtraheerd uit de organoïden en gelabeld met behulp van op basis van stabiele isotopen verzwaarde dimethyl groepen, om relatieve kwantificatie te kunnen toepassen. De geoptimaliseerde omgekeerde-fase chromatografie, zoals beschreven in hoofdstuk 2, wordt hier gebruikt na fractionering met een sterke kationenwisselaar (SCX). Daarenboven werd selectief gebruik gemaakt van twee zeer moderne massaspectrometers, waardoor we een optimale methode voor het bepalen van de aminozuurvolgorde konden toepassen voor elke SCX-fractie. Eiwitten afkomstig van organoïden van verschillende patiënten werden geanalyseerd, en de resultaten werden gecombineerd met expressiewaarden van het RNA uit deze organoïden. Naast het globale beeld van kanker-gerelateerde eiwitten in alle patiënten, kwamen er ook duidelijke patiënt-specifieke eiwitprofielen aan het licht.

De sleutel tot een alomvattende eiwitstudie is de onderlinge relatie van de verschillende post-translationele modificaties onderling op verschillende vormen van het eiwit, de zogenaamde proteovormen. Door de enzymatische klieving van eiwitten tot peptiden, wat de basis is van peptide-gebaseerde niet-gerichte proteomics (bottom-up shotgun proteomics) wordt deze relatie verbroken en kan niet meer worden herleid. Helaas is het momenteel niet goed mogelijk om van hele eiwitten de sequentie te bepalen (top-down proteomics) op alle eiwitten in het monster. Daarvoor is een compromis ontwikkeld, dat probeert de lengte van de geïdentificeerde peptiden zo groot mogelijk te houden, bij voorkeur groter dan 3 kDa, in een niet-gericht experiment. In het vierde hoofdstuk beschrijven we een geoptimaliseerde methode voor deze opkomende technologie van “middle down proteomics”. Het hele proces, vanaf het prepareren van het monster, de moderne chromatografische scheiding, en alternatieve fragmentatiemethodes voor de identificatie van de sequentie worden behandeld. De verbeteringen van de methodologie geven een duidelijke verbetering van detectiegraad en sequentiedekkingsgraad van de middengrote peptides. Deze technieken zijn een belangrijke stap voorwaarts naar de identificatie van de specifieke proteovormen.

Future outlook

Mass spectrometry-based proteomics has progressed tremendously over the years.[1,2] Two systems have turned out to be crucial for in-depth proteome characterization: the chromatography setup preceding online peptide analysis and the mass spectrometers themselves. Regarding the separation step, new developments in HPLC pumps, chromatographic supports and separation strategies have been explored. Pumps capable of generating high pressure and LC systems capable of handling those pressures have been developed.[3,4] These systems significanlty increased the performance of liquid chromatography as the use of smaller particles became feasible, which directly increases the efficiency of the separation.[5,6] Stationary phases with distinct pore sizes and chemistries have opened up a broad diversity of possibilities for peptide separation.[7,8] The use of multidimensinal separations to address the complexity of proteomic samples has also been explored and improved during the last years.[9,10] Even though these developments improved the ability to separate complex mixtures of peptides, improvements at the MS instrumentation level deserves special praise.[1] There has been a rapid advance in resolving power, mass accuracy, sensitivity and scan rate of mass spectrometers.[11] Thanks to the achievable high mass resolution coeluting peptides of similar mass are readily distinguished, which is a precondition for their accurate quantification.[12,13] In addition, hybrid mass analyzers have been introduced which have significantly improved proteomic analysis.[14,15]

Therefore, LC and MS have matured immensely through cumulative technological advances in instrumentation which have positioned mass spectrometry-based proteomics in an outstanding position facilitating the routine characterization of proteomes[16] and making proteomics an indispensable tool for systems biology.[17–20] Systems biology has largely been explored by genomic methods such as microarrays and next generation sequencing (NGS).[21] Recent advances in NGS techniques, including whole genome sequencing and total RNA sequencing allow for the generation of near-complete inventories of genetic variation in a system and its transcribed repertoire.[22] These techniques are extremely powerful and allow routine large-scale studies. However, they are not sufficient to understand the complexity of biological systems and by their nature cannot directly interrogate the proteome. Several studies have shown the poor correlation between mRNA and protein levels,

making quantitative mRNA insufficient to predict protein expression levels.[23–26] This poor correlation can be attributed to regulatory processes such as post-translational modifications and protein degradation regulation.[27,28] Subsequently, proteomic studies ideally complement genomic and transcriptomic analysis, and allow a more comprehensive understanding of biological systems.

Currently studies integrating multi-omics approaches are becoming popular,[26,29,30] particularly in the study of disease processes.[31–33] The immense complexity of molecular and cellular processes hampers the identification and interpretation of the molecular causes of disease. Therefore, in this perspective bringing together information from various biological scales should provide added insight into the fundamental mechanisms underlying physiology, development and the emergence of disease. One of the major challenges in multi-omics approaches is the integration of data from different origins. A huge amount of data is generated by the different techniques and the integrations is a complex procedure that is still a challenge even though great effort has been done during the last decade.[34–36]

Proteomics research propels basic research as discussed above; however, currently proteomics holds great potential for personalized medicine. Personalized medicine promises to offer the right treatment for the right patient at the right time.[37] Personalized medicine is rooted in the hy pothesis that diseases are heterogeneous, from their causes to rates of progression to their response to drugs. Each person's disease might be unique and therefore that person needs to be treated as an individual.[38] The link between an individual's molecular and clinical profiles is provided by personalized medicine, allowing physicians to make the right patient-care decisions. Safer and more effective treatment of disease, reduced detrimental side effects together with increased efficiency and productivity of pharmaceutical industries are some of the benefits of this way of understanding medicine. However, many challenges need to be addressed for the realization of personalized medicine, such as well-characterized patient populations, a detailed understanding of the biological pathways of disease and sophisticated computational methodologies among others. A well-characterized patient population might now be feasible thanks to the introduction of a novel and promising preclinical model, the organoids.[39] Organoids are ever-expanding three-dimensional epithelial structures with all the hallmarks of in vivo epithelial tissue described in detail in this thesis and already

applied in several studies.[40–45] The feasibility of organoids for high-throughput drug screening and the possibility of using patient-derived organoids for personalized therapy design has been demonstrated.[42] These organoids provide a potentially unlimited supply of well-characterized patient material, circumventing some of the limitations of current models (i.e. lack of genetically stable cell lines and need of extensive colonies of animals). Proteomics have already contributed to personalized medicine,[46–48] and multi-omics approaches, previously discussed, are also of special interested in personalized medicine.[49,50]

In conclusion, future technical developments in chromatography and mass spectrometry together with appropriate preclinical models will facilitate the translation of bench research into the clinic and become an excellent approach for personalized proteomics studies. Organoid technology will play an important role in personalized proteomics and the integration with multi-omic approaches will be particularly relevant in cancer research. New proteomics workflows including the use of alternative digestion protocols and multiple fragmentation techniques will allow a deeper proteome coverage. This deeper coverage will facilitate the identification of the specific proteoforms present in the samples, which is of special interest for personalized proteomics.

References

1. Yates, J. R. The revolution and evolution of shotgun proteomics for large-scale proteome analysis. J. Am. Chem. Soc. 135, 1629–40 (2013).
2. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. Curr. Opin. Chem. Biol. 24, 11–7 (2015).
3. Jorgenson, J. W. Capillary Liquid Chromatography at Ultrahigh Pressures. Annu. Rev. Anal. Chem. 3, 129–150 (2010).
4. Swartz, M. E. UPLC ™ : An Introduction and Review. J. Liq. Chromatogr. Relat. Technol. 28, 1253–1263 (2005).
5. Köcher, T., Swart, R. & Mechtler, K. Ultra-High-Pressure RPLC Hyphenated to an LTQ-Orbitrap Velos Reveals a Linear Relation between Peak Capacity and Number of Identified Peptides. Anal. Chem. 83, 2699–2704 (2011).
6. Cristobal, A. *et al.* In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis. Analyst 137, 3541 (2012).
7. Boersema, P. J., Mohammed, S. & Heck, A. J. R. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. Anal. Bioanal. Chem. 391, 151–9 (2008).
8. Mohammed, S. & Heck, A. J. Strong cation exchange (SCX) based analytical methods for the targeted analysis of protein post-translational modifications. Curr. Opin. Biotechnol. 22, 9–16 (2011).

9. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. 113, 2343–94 (2013).

10. Di Palma, S., Hennrich, M. L., Heck, A. J. R. & Mohammed, S. Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. J. Proteomics 75, 3791–813 (2012).

11. Yates, J. R. The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis. Perspective (2013).

12. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. 389, 1017–1031 (2007).

13. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal. Bioanal. Chem. 404, 939–965 (2012).

14. Andrews, G. L., Simons, B. L., Young, J. B., Hawkridge, A. M. & Muddiman, D. C. Performance Characteristics of a New Hybrid Quadrupole Time-of-Flight Tandem Mass Spectrometer (TripleTOF 5600). Anal. Chem. 83, 5442–5446 (2011).

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

16. Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. Mol. Cell 49, 583–590 (2013).

17. Cravatt, B. F., Simon, G. M. & Yates, J. R. The biological impact of mass-spectrometry-based proteomics. Nature 450, 991–1000 (2007).

18. Larance, M. & Lamond, A. I. Multidimensional proteomics for cell biology. Nat. Rev. Mol. Cell Biol. 16, 269–280 (2015).

19. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat. Rev. Genet. 14, 35–48 (2013).

20. Cox, J. & Mann, M. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. Annu. Rev. Biochem. 80, 273–299 (2011).

21. Schuster, S. C. Next-generation sequencing transforms today's biology. Nat. Methods 5, 16–18 (2007).

22. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12, 87–98 (2011).

23. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19, 1720–30 (1999).

24. Ghazalpour, A. *et al.* Comparative Analysis of Proteome and Transcriptome Variation in Mouse. PLoS Genet. 7, e1001393 (2011).

25. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. Nature 473, 337–42 (2011).

26. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. Mol. Biosyst. 5, 1512–26 (2009).

27. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13, 227 (2012).

28. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. FEBS Lett. 583, 3966–3973 (2009).

29. Low, T. Y. *et al.* Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 5, 1469–78 (2013).

30. Mootha, V. K. *et al.* Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell 115, 629–40 (2003).

31. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. Nature 1–21 (2014). doi:10.1038/nature13438

32. Celis, J. E. *et al.* Integrating proteomic and functional genomic technologies in discovery-driven translational breast cancer research. Mol. Cell. Proteomics 2, 369–77 (2003).

33. Chen, G. *et al.* Discordant protein and mRNA expression in lung adenocarcinomas. Mol. Cell. Proteomics 1, 304–13 (2002).

34. Cox, B., Kislinger, T. & Emili, A. Integrating gene and protein expression data: pattern analysis and profile mining. Methods 35, 303–314 (2005).

35. Nie, L., Wu, G., Culley, D. E., Scholten, J. C. M. & Zhang, W. Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications. Crit. Rev. Biotechnol. 27, 63–75 (2007).

36. Rogers, S. in Bioinformatics for Omics data(Clifton, N.J.) 719, 133–151 (2011).

37. Ginsburg, G. S. & McCarthy, J. J. Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol. 19, 491–6 (2001).

38. Forler, S., Klein, O. & Klose, J. Individualized proteomics. J. Proteomics 107, 56–61 (2014).

39. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 459, 262–5 (2009).

40. Grabinger, T. *et al.* Ex vivo culture of intestinal crypt organoids as a model system for assessing cell death induction in intestinal epithelial cells and enteropathy. Cell Death Dis. 5, e1228 (2014).

41. Dekkers, J. F. *et al.* A functional CFTR assay using primary cystic fibrosis intestinal organoids. Nat. Med. 19, 939–945 (2013).

42. van de Wetering, M. *et al.* Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients. Cell 161, 933–945 (2015).

43. Clevers, H. Modeling Development and Disease with Organoids. Cell 165, 1586–97 (2016).

44. Gao, D. *et al.* Organoid cultures derived from patients with advanced prostate cancer. Cell 159, 176–87 (2014).

45. Boj, S. F. *et al.* Organoid models of human and mouse ductal pancreatic cancer. Cell 160, 324–38 (2015).

46. Koomen, J. M. *et al.* Proteomic Contributions to Personalized Cancer Care*. Mol. Cel-lular Proteomics 7, 1780–1794 (2008).

47. Geho, D. H. *et al.* Role of proteomics in personalized medicine. Per. Med. 3, 223–226 (2006).

48. Jain, K. K. in Advances in protein chemistry and structural biology 102, 41–52 (2016).

49. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148, 1293–1307 (2012).

50. Cristobal, A. *et al.* Personalized Proteome Profiles of Healthy and Tumor Human Colon Organoids Reveal Both Individual Diversity and Basic Features of Colorectal Cancer. Cell Rep. 18, 263–274 (2017).

Curriculum Vitae

I was born on the 12th of June of 1988 in Vitoria-Gasteiz, Spain. I entered the University of the Basque Country in 2006 and after five years of studies I received my Chemistry degree in July 2011. During my last year of university I participated in an Erasmus exchange program. I performed an analytical research project in the University of Utrecht in the Biomolecular Mass Spectrometry and Proteomics group. During this exchange of 10 months I started to get familiar with proteomics and gained knowledge in the methodology and instrumentation. In October 2011 I started my PhD at the same group under the supervision of Prof. Dr. Albert J.R. Heck and Dr. Shabaz Mohammed. During my PhD I had the opportunity to combine method development and optimization projects with the application of those methods to the study of cancer. Moreover, in order to study this disease, instead of using the widely studied cell lines and animal models, I got the opportunity to work with a novel and promising preclinical model, the organoids.

Publications

Toward an optimized workflow for middle-down proteomics.
Cristobal A, Marino F, Post H, van den Toorn HWP, Mohammed S, Heck AJR.
Analytical Chemistry, 2017 Mar 10. doi: 10.1021/acs.analchem.6b03756

Personalized proteome profiles of healthy and tumor human colon organoids reveal both individual diversity and basic features of colorectal cancer.
Cristobal A, van den Toorn HWP, van de Wetering M, Clevers H, Heck AJR and Mohammed S.
Cell Reports, 2017 Jan 3;18(1):263-274. doi: 10.1016/j.celrep.2016.12.016.

Combining deep sequencing, proteomics, phosphoproteomics, and functional screens to discover novel regulators of sphingolipid homeostasis.
Lebesgue N, Megyeri M, Cristobal A, Scholten A, Chuartzman SG, Voichek Y, Scheltema RA, Mohammed S, Futerman AH, Schuldiner M, Heck AJR,Lemeer S.
J. Proteome Res., 2016 Nov 14. doi: 10.1021/acs.jproteome.6b00691

Rb and FZR1/Cdh1 determine CDK4/6-cyclin D requirement in C. elegans and human cancer cells.
The I, Ruijtenberg S, Bouchet BP, Cristobal A, Prinsen MB, van Mourik T, Koreth J, Xu H, Heck AJ, Akhmanova A, Cuppen E, Boxem M, Muñoz J, van den Heuvel S.
Nat Commun., 2015 Jan 6;6:5906. doi: 10.1038/ncomms6906.

Characterization and usage of the EASY-spray technology as part of an online 2D SCX-RP ultra-high pressure system.
Marino F, Cristobal A, Binai NA, Bache N, Heck AJ, Mohammed S.
Analyst, 2014 Dec 21;139(24):6520-8. doi: 10.1039/c4an01568a.

Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis.
Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, Schafer S, Hübner N, van Breukelen B, Mohammed S, Cuppen E, Heck AJ, Guryev V.
Cell Reports, 2013 Dec 12;5(5):1469-78. doi: 10.1016/j.celrep.2013.10.041.

Robust phosphoproteome enrichment using monodisperse microsphere-based immobilized titanium (IV) ion affinity chromatography.
Zhou H, Ye M, Dong J, Corradini E, Cristobal A, Heck AJ, Zou H, Mohammed S.
Nat Protoc., 2013 Mar;8(3):461-80. doi: 10.1038/nprot.2013.010.

In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis.
Cristobal A, Hennrich ML, Giansanti P, Goerdayal SS, Heck AJ, Mohammed S.
Analyst, 2012 Aug 7;137(15):3541-8. doi: 10.1039/c2an35445d.

Acknowledgments

After these 5 years of my life as a PhD student is nice to look back and realize all the help and support I have received in order to arrive to the end of the PhD. First, I would like to start with my beloved family. Aita, Ama y mi queridísimo tatito, ¿cómo haber llegado hasta aquí sin vosotros? ¡Imposible! Con vuestro apoyo todo ha sido mucho más llevadero, incluso en los momentos más difíciles. Muchísimas gracias por estar siempre ahí, por las conversaciones por Skype, por las visitas, los envíos de revistas y jamón de pato que tanta ilusión me hace recibir, por intentar hacer que la distancia que nos separa no sea tan grande y por animarme y mandarme la energía necesaria para llegar hasta dónde estamos hoy. ¡Un besote muy muy grande para los tres!

Everything started when I came to The Netherlands as an Erasmus exchange student and I became a member of the Heck Lab. Thanks to Shabaz's and Albert's belief on me, I started my PhD. Thanks to both of you for offering me the possibility to do my PhD with you! Albert I always found it amazing how you managed to be involved with my projects and give so many advices despite the fact that you are always very busy. Thanks a lot for you time and support! Shabaz, all the help, the support and the knowledge you transferred to me are invaluable. Even when you moved to Oxford and I thought everything was going to be a chaos in my PhD without you, you managed to keep on supervising me and motivating me from the distance, which I really appreciate. Thanks a lot! Maarten, thanks for the support you gave me during these years, mainly since Shabaz left. You were always there when I was facing problems with my projects to help me, thanks! Simone, I´m very glad to have met you and thank you very much for being the bridge between this PhD and my Postdoc!

An essential collaboration of this PhD involved Marc van de Wetering and Hans Clevers. Thanks to both of you for the nice work and collaboration together. Marc, a special thanks to you for your time and patience!

An extremely special THANKS goes to my dear paranymphs: Anita and Matina! If I think back on the good memories I collected over these years you are present in most of them. I have travelled with both of you (Liege, Madrid, Strasbourg, New York, Washington…) and greatly enjoyed the time we spent together. Also in Utrecht we had great times going out, dancing, going to yoga, having dinners… But very importantly I thank both of you for supporting and motivating me to finish this stage of my career. I really enjoyed the time with you girls, the apparent clam of Anita and the spontaneity of Matina, which make of both of you a great combination! I'm really really happy to have met you!

During these years I had the chance to meet many people and I would like to thank all the members of the Heck lab, previous and present, as in different ways you all contributed to my daily life as a PhD student. Unfortunately some of you left too early, however, your impact on me was still very strong. Marco Hennrich, I'm finally done! I learned many many things from you when I arrived to the lab and enjoyed talking to you about both scientific and non-scientific topics. It was extremely sad to see you leaving! Javierrrr, ¡menos mal que estabas tú cuando llegue

al laboratorio! Poder hablar en castellano por un rato, que me explicases tantas cosas, el gusto que tienes para las figuras y las presentaciones… ya estoy esperando las críticas del formato de la tesis ;) Menos mal que tanto melocotón, como calamar y el hola hola caracola siguen en tus emails, ¡siempre me sacas una sonrisa con ellos! Al igual que con el 'zenkiu':D Y bueno bueno… ¡todo un honor tenerte en mi comité de defensa! ¡Eskerrik asko y seguiremos en contacto!

And… how could I have survived these years without you, Henk?! I'm sorry for all the hundreds of thousands of times that I went to your office to get something clear out of my data… But at the end we managed to have a nice publication as well as nice conversations together, thanks!

Special thanks to: Harmianini, since the beginning we understood each other pretty well even though your skill as a human calculator was complicating my days in the lab. After the Summer School in Brixen, where we had a great time together, we got closer and having you all these years here was great. Your help was and will always be precious. I wish you all the best in this exciting new chapter of your life. Jamilaki! I miss you so much! I really enjoyed having you as an office mate and all our conversations. Charlottina! Our overlapping time in the lab was short but you quickly became one more of us. David, usted también es parte de ese grupo de amigos que con el tiempo hemos ido creando. No estaba segura si dirigirme a usted en castellano o en inglés por los problemas que tenemos para entendernos en nuestra lengua maternal… :D Ha sido un placer coincidir con usted en esta etapa de mi vida y muchas gracias por todos los consejos, el hecho de que piense todo tanto a veces tiene su lado bueno ;) Y de un tico a otro tico, JuanMa! Un gusto haberte conocido aunque en tu caso haya sido al final de esta etapa, pero bueno, todavía nos queda un tiempo en este país. Anna Ressa, it was really nice meeting you and discussing with you the frustration that was sometimes appearing during these years and how to overcome it (shall we dance?)!

Nadine, I also spent most of my PhD having you in the lab, it was always so easy to go and talk to you, and I also had a great time going horseback riding together! Thanks for the nice and funny moments together. Renske, we also shared the passion for horses and had nice time together during these last years. I wish you all the best in this new exciting chapter of your life! Teck, I really learned a lot talking to you about all that biology knowledge that I'm lacking, thanks for your patience! Gianluquino, so sad not to hear your 'a verrrr' when you see me. Kyle, thanks for all the nice movie nights! Liana, I really enjoyed our conversations and I'm happy for the lovely daughter Wim and you have! Lucrèce, it was great getting to know you, spending time together, having nice dinners, nice macarons, many conversations… I really miss you! But I know your new life in Paris is going more than great, which makes me really happy :D Violette and Luc, the lovely French couple, we also shared many great and funny moments together and I'm also aware of how good is everything going also back in France. And together with these French people, I can´t forget Nicolas and Clement. Nicolas, we started at a similar time and spent many hours and days working together in the lab until late and even though it was tiring we had a nice time together and I'm happy to know that your new life chapter has started so promisingly. Clementttt, so nice all the time we spent together! We started getting to know each other in the great Triple TOF team together with Anita

and from there we spent a lot of time together during the last years. I really enjoy being your friend both inside and outside the lab. Merci beaucoup! ;)

A special THANKS also goes to my work-husband, Fabiolon! Such a nice experience meeting you and going together through this period, which have had a lot of funny moments as well as hard moments, when you were always there trying to motivate me. Thanks a lot! I also miss you a lot, but I know you are enjoying a lot in that beautiful country, cycling every now and then.

Many thanks to Harm, Arjan, Mirjam, Soenita, Jamila, Luc and Anja, who kept and keep the lab running so smoothly. I will also take the opportunity to thank my office mates: Donna, Pepijn, Ele, Matina, Charlotte and GanYin. I would also like to thank Corine and Geert, who really helped me during these years: Corine, you are always there willing to help with all the paper work, thanks a lot for your time. And Geert, we are so lucky to have you! You are always there to fix all the computer related issues and even if you are not in the office you are still reachable and willing to help, thanks a lot!

I'm really sorry I didn't mention all the group members, but as I said at the beginning I really would like to thank you all for all the time and experiences we shared over these years!

También me gustaría agradecer a mis amigas de siempre (Maialen, Leire, Niso) que aunque estemos en la distancia siempre encontramos un rato para un Skype y ponernos al día, una escapada o una visita,… ¡muchas gracias chicas, sois un gran apoyo! Y Nagore!! ¡Muchísimas gracias por todo el tiempo que me has dedicado con el formateo de la tesis! Y muchas gracias también a todos mis amigos y conocidos que se han preocupado por saber cómo me iba el doctorado durante estos años y me han animado a llegar hasta aquí.

Y por último, pero siendo una de las personas más importantes durante este doctorado, ¡Aitor! :D Juntos decidimos empezar esta etapa y nuestra vida juntos. En estos 5 años nos han pasado muchas cosas y sé que aunque dejar Vitoria no fue nada fácil para ti, ¡juntos lo logramos y estoy muy muy orgullosa de lo que has conseguido aquí! Muchísimas gracias por aguantarme tanto en los momentos de super emoción como en los de decepción profunda que he tenido estos años, y por esa calma que siempre me transmites cuando me estoy agobiando. Eskerrik asko mi bonito! Y ahora, ¡a por el siguiente capítulo juntos!

Alba