Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

# Overcoming Challenges of Shotgun Proteomics

von

Annette Michalski

aus Hildesheim

2012

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, ........................................................................................................

Dissertation eingereicht am

1. Gutachter: Prof. Dr. Matthias Mann

2. Gutachter: Prof. Thomas Neubert, phD

Mündliche Prüfung am

# CONTENTS

## ABBREVIATIONS

| | |
|---|---|
| AGC | automatic gain control |
| CID | collision induced dissociation |
| dc | direct current |
| ESI | electrospray ionization |
| ETD | electron transfer dissociation |
| FDR | false discovery rate |
| FT | Fourier transformation |
| HCD | higher energy collisional dissociation |
| HPLC | high-pressure liquid chromatography |
| ICR | ion cyclotron resonance |
| iTRAQ | isobaric tags for relative and absolute quantitation |
| LC/MS | liquid chromatography coupled to mass spectrometry |
| MALDI | matrix-assisted laser desorption ionization |
| mRNA | messenger RNA |
| MS | mass spectrometry |
| MS/MS | tandem mass spectrometry |
| m/z | mass-to-charge ratio |
| PIF | precursor intensity fraction |
| ppm | parts per million |
| ppb | parts per billion |
| QQQ | triple quadrupole mass spectrometer |
| Q TOF | quadrupole time-of-flight mass spectrometer |
| R | mass spectrometric resolution |
| rf | radio frequency |
| RP | reversed-phase |
| SAX | strong anion exchange |
| SCX | strong cation exchange |
| SILAC | stable isotope labeling of amino acids in cell culture |
| T | Tesla (unit of magnetic field strength) |
| Th | Thompson (unit of m/z) |
| TMT | tandem mass tag |
| TOF | time-of-flight |

# 1 Introduction

## System-wide '*omics*' studies

### Proteomics, genomics, transcriptomics

For many decades, researchers have investigated the biology of cells, their diversity and their main functional building blocks, *the proteins*, in focused and small-scale experiments. Increasingly comprehensive investigations into the complement of all proteins of a cell type, *the proteome*, enable researchers to better understand their functions in the context of biological networks as a whole[1,2]. Specifically, *proteomics* as a rapidly evolving discipline deals with identification and quantification of thousands of proteins in different cell states, including interactions between individual proteins and post-translational modifications, which play a pivotal role in regulating protein activity.

Early system-wide biological studies were focused on the *genome* of an organism, the complement of all *genes*, directly or indirectly encoding for the primary structure of the proteins. The very first genome – that of bacteriophage ΦX174 - was decoded in 1977[3] by electrophoretic sequencing methods that were pioneered by Fred Sanger and became very popular during the next decades[4,5]. It took until 1995 to analyze the first genome sequences of small cellular organisms[6-8]. In 1990, the Human Genome Project was established as a joint effort of several laboratories to further advance sequencing technology and to obtain a reference sequence of man. It took 11 years for presenting a first draft[9,10] – a milestone in the era of *genomics*. The finalized sequence of 3 billion nucleotides of the human genome was published in 2007 and the number of its protein-coding genes is now estimated to be about 20,000[11].

Expression of protein coding genes requires the synthesis of *transcripts*. Therefore, the *transcriptome* - as the complement of all messenger RNA molecules in a cell type - is to a first approximation the missing link between genome sequence and the proteome. DNA microarrays are a very well established technology that is frequently applied for transcript analysis and it allows large-scale screens for a variety of applications[12,13]. However, the hybridization-based strategy is limited to the detection of known sequences that are synthesized and deposited on a chip. Quantification of transcripts is possible but not very precise[14]. Today, whole-genome sequencing as well as transcript analysis benefit greatly from *next-generation sequencing* technologies[15]. The automated sequencing-by-synthesis approaches overcome major drawbacks of microarrays and facilitate in-depth transcriptome analysis. *RNA-Seq* technology features high-throughput, improved quantification and low costs per base[16]. Due to the relatively short sequence reads, however, data analysis strategies preferably utilize a reference genome.

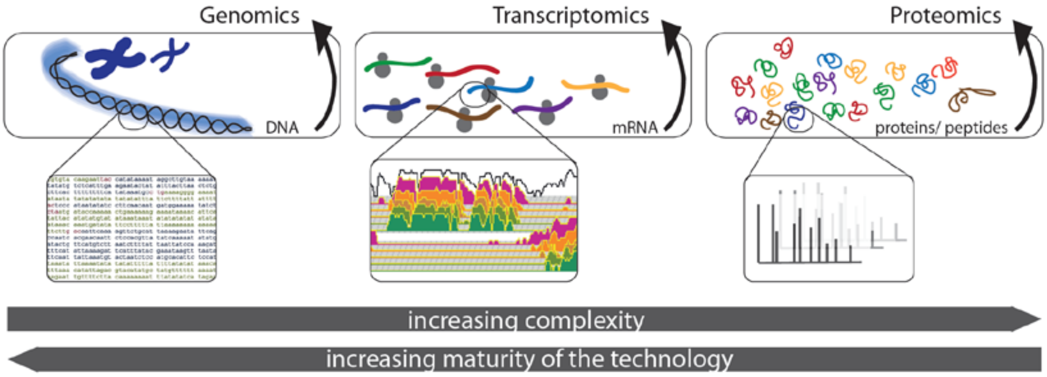


Figure 1: System-wide studies of different biopolymers: Genomics, transcriptomics and proteomics; cartoons in the lower part depict coding sequences in the genome (left), transcripts mapped to coding sequences (middle) and mass spectra of peptides (right).

## Challenges of three complementary disciplines

*'Omics'* technologies aim at unraveling biopolymer sequences, however, different functionalities result in different challenges for these disciplines. For instance, the mammalian genome is extremely complex and contains highly repetitive sections that complicate sequencing and alignment of fragments. Nonetheless, the genome of each species is a fairly stable unit and sequencing a reference genome is a onetime effort. In contrast, the transcriptome and proteome regulate biological processes in different cell types that are exposed to different environmental conditions and that are therefore constantly changing. The proteome varies considerably between mammalian cell types and tissues and the copy numbers of different proteins are spread over many orders of magnitude. Similar trends apply to transcriptome analysis and greatly increase the challenges of obtaining comprehensive data compared to genomics. For modern proteomics research, the availability of reference genomes is of fundamental importance, because data analysis almost always relies on protein databases that are derived from genome sequences[17,18]. Due to rapid technological progress, several thousand genomes of different organisms, from microorganisms to mammals, are already available today.

In contrast to genomics, completeness of transcriptomes and even more so, of proteomes, is more difficult to determine and even the number of protein coding genes is usually only an estimate. Alternative splicing and protein isoforms may or may not be counted as different proteins; therefore the size of the proteome is also a matter of definition[19]. According to a simple ‘one gene–one protein’ rule, the first *complete* proteome characterizing 4,400 different proteins of the model organism yeast was obtained in 2008. Completeness was judged by the number of proteins expected from genome-wide tagging experiments[20].

There are numerous large-scale investigations into the mammalian proteome of various cell types, tissues or body fluids. However, these studies were usually focused on specific questions rather than comprehensiveness. Mass spectrometry-based technology only recently succeeded in identifying 10,000 proteins from a human cancer cell line[21,22]. This appears to be not far from the number of functional proteins expected to be expressed in a single cell type based on RNA-Seq data[23-25]. However, using current technology, coverage of protein variants remains relatively low and in-depth analysis of proteomes remains a time-consuming and laborious effort. Thus, in order to complement whole-genome sequencing, which has become fairly routine in research[26], and transcriptome analysis that is now performed with fast and relatively inexpensive RNA-Seq technology, the established proteomics approaches need to be drastically extended as well as streamlined and automated. Successfully complementing the *‘omics’* disciplines is one of the ultimate goals of proteomics. This would enhance our understanding of system-wide processes by enabling comparative studies among genomics, transcriptomics and proteomics (Article 8). Therefore, a main objective of this thesis was to make technological contributions towards the further development of proteomics.

## Mass spectrometric instrumentation

### Gentle ionization methods

The era of *mass spectrometry-based proteomics* began with the invention of two gentle ionization methods that allow bringing proteins into the gas-phase without destroying them. Electrospray ionization (ESI) utilizes a needle to spray small droplets containing the charged sample molecules into a strong electric field[27]. The solvent evaporates during this process and the desolvated ions enter the mass spectrometer. The second approach employs a laser that ‘excavates’ and ionizes molecules from a solid matrix that is co-crystallized with the sample (matrix-assisted laser desorption ionization, MALDI)[28]. Both approaches are very suitable for ionizing large biopolymers; however, in terms of coupling to mass spectrometers they are conceptually completely different. Electrospray ionization works at atmospheric pressure and produces ions in a continuous manner, as opposed to the pulsed matrix-assisted laser desorption ionization that happens in the vacuum of the mass spectrometer. The nature of the ionization method is of great importance for the analysis strategy and the type of mass spectrometer that can be used. For example, liquid chromatography can relatively easily be combined with electrospray, whereas for matrix-assisted laser desorption ionization a spotter has to off-line collect the sample prior to mass spectrometry.

### Mass analyzers: general working principles

The mass spectrometers employed for proteomics research have undergone dramatic improvements; however, the general working principle of the mass analyzers such as quadrupoles, ion traps, time-of-flight (TOF) and ion cyclotron resonance (ICR) have generally remained the same for several decades. A remarkable exception, however, is an entirely novel electrostatic analyzer presented in 2000 by Alexander Makarov. This *Orbitrap analyzer* rapidly turned into an indispensable tool for proteomics research (see Box for *A brief history of the Orbitrap analyzer*).

## A Brief History of the Orbitrap Analyzer

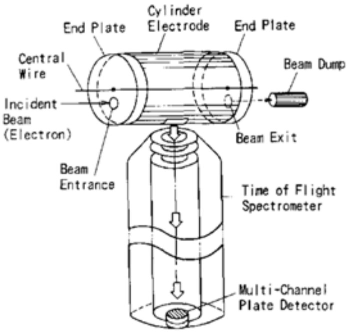


Figure 2: Kingdon trap and TOF mass spectrometer[30].

The earliest predecessor of the Orbitrap mass analyzer was the electrostatic Kingdon trap in 1923, consisting of an axial wire surrounded by cylindrical electrodes (logarithmic field)[29]. The Kingdon trap is not a mass analyzer itself, but was used in combination with a quadrupole, TOF or ICR as trapping device in the 1990s[30]. Several modifications such as to the shape of the outer electrodes by Knight *et al.*, 1981, improved the axial confinement for better trapping efficiency (addition of a quadrupole field)[31]. The trap was further developed by Makarov *et al.* introducing the spindle-like shape of the inner electrode, which induces an electrostatic field with purely harmonic potential in z-direction (quadro-logarithmic field)[32]. As a consequence, ions circulate around the central electrode on stable trajectories while also oscillating along the z-axis with a frequency that is only dependent on the mass-to-charge ratio. Image current detection and Fourier transformation of the transient successfully generated mass spectra from the *Orbitrap analyzer* [33].

External accumulation and injection of ions as confined packets turned out to be key challenges when coupling the Orbitrap analyzer to a continuous ion source. First, a linear ion trap was employed for storage and subsequent axial ejection, but this was later replaced by a more efficient curved rf-only quadrupole (*C-trap*)., this instrument was first introduced commercially only in 2005 and it has already undergone several major iterations since – and it is a main topic of this thesis.

Depending on their working principles, mass analyzers can be categorized into two major types, which are (1) beam-type and (2) trap-based analyzers (Figure 3). Quadrupole and TOF analyzers are by their nature continuously scanning devices, whereas the ICR and Orbitrap ion traps, analyzers perform sequential processes on captured ion populations to obtain a mass spectrum. Certain performance characteristics such as resolution, mass accuracy, sensitivity and in particular the duty cycle are different between these mass analyzers with important implications for liquid chromatography-based shotgun proteomics analysis.

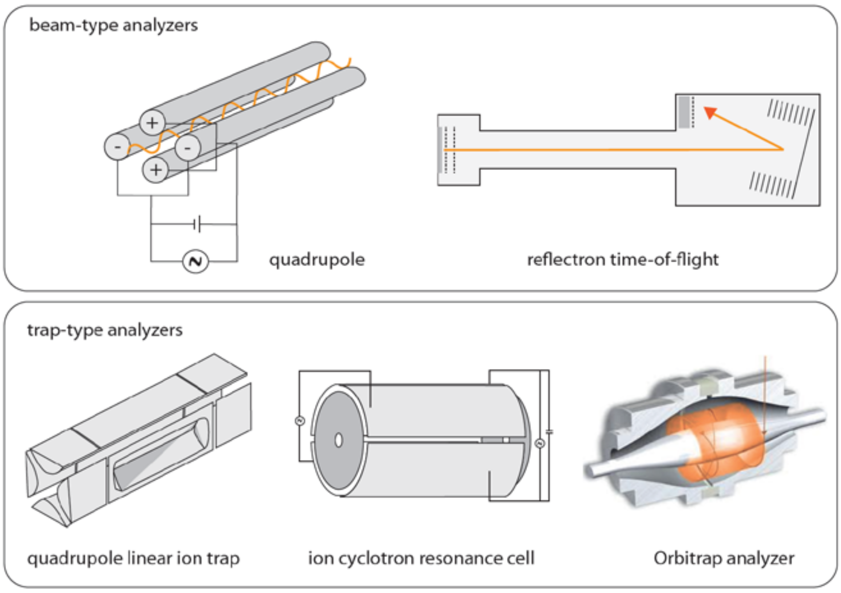

beam-type analyzers
+
-
-
+
quadrupole
reflectron time-of-flight
trap-type analyzers
quadrupole linear ion trap
ion cyclotron resonance cell
Orbitrap analyzer

Figure 3: Beam-type and trap-type mass analyzers; adapted from [2,34,35].

One of the oldest types of mass analyzer is the quadrupole, consisting of four precisely parallel metal rods. Opposite electrodes are connected and one pair receives a positive, the other pair a negative direct current (dc) potential that is superimposed by a time-dependent radio frequency (rf) potential. When ions are injected into the quadrupole in the direction of the rods, the oscillating electric field in the center of the quadrupole only allows a narrow mass-to-charge (m/z) range to pass on a stable trajectory. The remaining ions will impinge on the rods. Thus, a quadrupole rather acts as a mass filter than as a conventional mass spectrometer. Ramping the dc and rf potentials allows different narrow m/z ranges to pass the quadrupole and thereby generates a mass spectrum. The x-y-motion of ions that pass the quadrupole in z-direction, dependent on the changing potentials, is described in a second order differential equation, the *Mathieu equation*. Its solution provides two dimensionless parameters *a* and *q* that characterize the amplitude of the dc and rf current. Plotting of *a* against *q*, as shown in Figure 4, provides the so-called stability diagram of the quadrupole; the intersecting mass-scan line (*a/q* constant) defines the m/z ratios that pass the quadrupole without interruption.

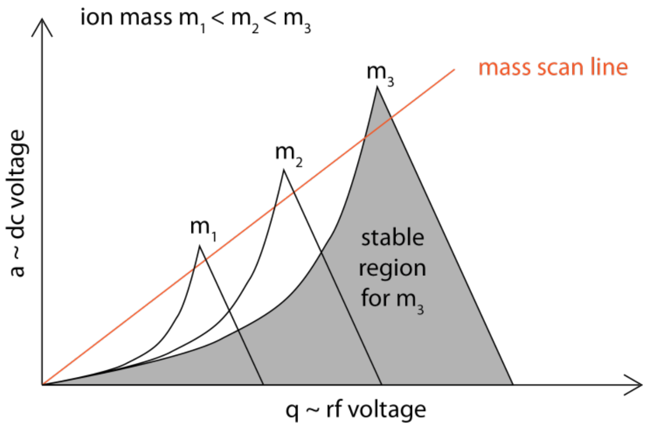


Figure 4: Plots of the dimensionless parameters *a* and *q* for different m/z values generate the stability diagram.

The mass resolution is inversely proportional to the width of the cross-section between the stable region and the mass-scan line. Quadrupoles can also be used in rf-only mode (*a = 0*), in which case they function as wide band mass filters. As such, quadrupoles are frequently employed as ion guides or as intermediate reaction region in triple-quadrupole instruments (Q2) as described at the end of this section. In general, quadrupole mass analyzers are compact, but feature rather low mass resolution. Improving the quadrupole characteristics with respect to resolution and mass range is difficult and most importantly requires substantial efforts in manufacturing. Narrower-diameter rods of hyperbolic shape together with higher rf frequency and lower acceleration potential of the ions all contribute to higher resolution. As the resolution correlates with the number of oscillations of the ions in the quadrupole, increasing the length of the rods is also beneficial, but restricted by practical limitations.

Time-of-flight instruments are beam-type analyzers that determine the m/z ratio of ions based on the different time for passing a field-free drift tube. Ions ideally enter with the same kinetic energy, therefore those with smaller m/z ratios travel faster than those with larger ones. This dispersion allows detection of the ions in the order of increasing m/z ratios. Resolution of TOF instruments is limited by the length of the field-free drift tube and the capability of the detector to identify two different masses that arrive rapidly one after the other. Reduction of the spatial and kinetic energy spread of the ions can efficiently be achieved by a reflectron, i.e. an electrostatic mirror that focuses equal m/z ratios, making high resolution (R > 10,000) available[36].

Special shapes of the flight-path can further increase the resolution; however, they may compromise sensitivity[37].

MALDI sources are very commonly coupled to time-of-flight instruments, because the pulsed ion flux is very suitable for this analyzer. Continuous ion sources such as electrospray devices can be utilized with orthogonal TOF conformations, featuring an acceleration region to push a section of the ion beam into the field-free drift tube[38] .

Ion traps are primarily devices to confine ions and store them over a period of time, but they can also be employed as mass spectrometers. They are popular because of their robustness, small footprint and because their performance characteristics are sufficient for many applications. Moreover, the trapping capability allows special applications, such as fragmentation analysis of molecules in multiple stages of mass spectrometry. There are two major designs of ion traps that were historically developed in the same order as described here. The 3D ion trap (*Paul* trap) consists of three electrodes, a central doughnut-shaped electrode accompanied by two end-cap electrodes with small apertures for inlet and ejection of ions[39]. Ion traps usually employ electron multipliers for detection, which for 3D traps are placed directly behind the exit end-cap. Inside the ion trap, a three-dimensional quadrupole field is created by superimposing dc and rf potentials similar to those in the quadrupole mass filter. Ions with suitably low energy are trapped, because their trajectories inside the ion trap allow enough collisions with the helium buffer gas, *collisional cooling*, to sufficiently reduce their kinetic energy. The mass spectrum is obtained in the next step by linear ramping of the rf amplitude, which causes sequential ejection of ions with different m/z ratios. This strategy, referred to as *mass selective instability mode*, was a major breakthrough in ion trap technology, since it allowed scanning entire mass ranges instead of only isolating single mass-to-charge ratios. It was developed by George Stafford and co-workers in 1984 [40].

Despite its popularity since the 1990s, in proteomics the 3D linear ion trap faces major limitations stemming from its performance characteristics. Very limited ion capacity inside the trap restricts the number of ions that are available for different processes such as storage, isolation, activation, and obtaining the mass or fragment mass spectrum[34]. Due to the limited physical space inside the trap, exceeding the ion capacity causes so-called *space charge effects*, i.e. repulsion of equally charged particles, which overlay the external field. This has major implications on the achievable resolution and mass accuracy. Furthermore any parameter is affected that is dependent on the number of ions, or more accurately, charges that can be accumulated, e.g. sensitivity and dynamic range. A second restriction of the 3D ion trap is its relatively low trapping efficiency. Linear ion traps address many of these limitations[41], because they store the ions in two dimensions along the z-axis of a quadrupole and are therefore less affected by space charge effects[34,42]. The design of the quadrupole linear ion trap (2D trap) by Jae Schwartz *et al.* resembles a quadrupole that is split into three sections. The central section of the three parts is the largest and is intended to store the ions, whereas the front and back sections can be used for ion manipulation and for applying an axial trapping potential. An ejection slit in one of the central rods allows ion ejection and detection by an electron multiplier as illustrated in Figure 3. In recent versions of the quadrupole linear ion trap, the sensitivity is doubled by adding a second ion ejection slit and a second electron multiplier to the opposing rod[43]. Due to the long trapping path of the linear ion trap, the tapping efficiency is higher and similarly it is capable of storing much larger ion populations than the 3D trap. Isolation is performed by resonance ion ejection, i.e. all ions with mass-to-charge ratios higher and lower than the ion of interest are ejected from the trap. In the subsequent steps, collision induced fragmentation can be employed to generate a tandem mass spectrum (MS/MS). The lowest mass-to-charge ratio that is stable in the ion trap is referred to as *low mass cut-off* and is directly proportional to the main rf amplitude applied. In particular, in tandem mass spectrometry, the low mass cut-off has important implications, because it limits the lowest fragment mass that can be retained in the trap.

In direct comparison with a 3D trap using the same detector, a linear quadrupole ion trap was found to have 15-fold higher ion capacity[34]. However, this benefit comes with the need of increased dimensions, e.g. longer rods, which need to be manufactured very precisely. As ions are spread out along the rods, different field strength at different positions may cause different time points of ejection, reducing the resolution of the ion trap mass analyzer. The performance of quadrupole linear ion traps was greatly improved by a dual cell design with two compartments separated by an aperture[43,44]. The compartments are kept at different pressures, the first one at higher pressure (approx. 7 x $10^{-3}$ mbar) to efficiently trap, isolate and fragment ions. Increased rf voltage and an isolation waveform during ion injection increase the sensitivity. The second cell is held at lower pressure (approx. 5 x $10^{-4}$ mbar) to achieve an improved resolution and higher scan rates[45].

## High resolution mass analyzers and image current detection

ICR and Orbitrap analyzer are both located at the high-performance end of the range of trap-type analyzers, but different concepts of trapping the ions are applied: ICR technology uses a static electric field for axial trapping and a strong magnetic field to force the ions onto orbital trajectories inside the ICR cell[46-48]. Due to its unique shape, the Orbitrap analyzer enables dynamic trapping in an electrostatic field[32]. Major differences to 3D or linear ion traps are detection of image currents and analysis of the resulting signal by Fourier transformation (FT). This contrasts with the counting of ions on multipliers that are sequentially ejected from the ion trap to obtain the mass spectrum. Image current detection followed by Fourier transformation allows very accurate measurement of all mass-to-charge ratios at the same time; this requires a device response in which the measured frequency of motion is only dependent on the mass-to-charge ratio of the ions (Figure 5).

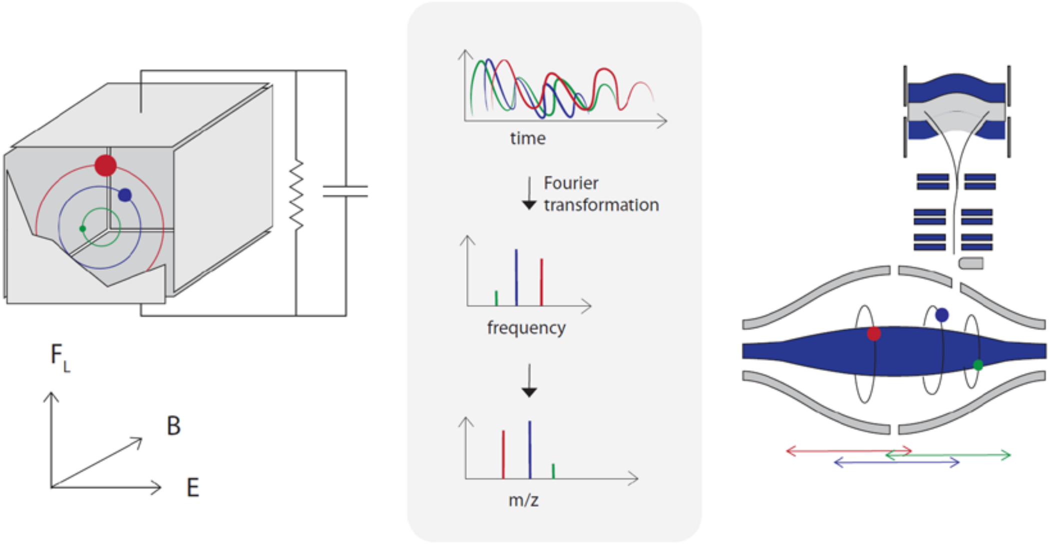


Figure 5: Working principle of ICR and Orbitrap analyzers applying image current detection and Fourier transformation.

In FT-ICR, ions are axially injected into the ICR cell (electric field E) that is present inside a strong magnetic field B so that charged particles experience the Lorenz force $F_L$. Coherence of ion motion is achieved by broadband rf excitation; ions absorb energy from the signal and start moving as phase-coherent packages on larger orbits that are specific to their mass-to-charge ratio. A set of extra electrodes detects the image current and very high resolution spectra are obtained by Fourier transformation. Recent instruments usually were equipped with 7 T magnetic fields. However, commercial instruments can contain magnets up to 18 T[49,50] and for research purposes, magnets up to 21 T have been tested[51].

For high resolution Orbitrap measurements, in contrast, the ion injection step is crucial and was a major challenge during the development of the analyzer[52]. A curved rf-only quadrupole, the *C-trap*, is employed for external collection of ions prior to injection of defined ion packages into the Orbitrap analyzer[53]. The C-trap is filled with nitrogen gas for collisional cooling of the ions, before high voltages push the ions orthogonally out of the device. They are then accelerated to high kinetic energy and enter the Orbitrap analyzer through a small aperture. Ions are captured in an increasing electrical field that contracts the radius of the ion cloud, referred to as *electrodynamical squeezing*[54], and start circulating around the spindle-shaped central electrode in rings made up of ions with specific m/z value. In addition, the orbiting ions oscillate along the central electrode as illustrated in Figure 5. The frequency of this motion is controlled by the unique shape of the trap that provides an exclusively harmonic potential along the z-direction[32]. The image current induced by the axial oscillations in detector electrodes allows precise determination of the mass-to-charge ratios, because the frequency is independent on any initial properties of the ions.

## Important parameters and instrument characteristics

In FT-based mass spectrometry, the resolution is directly related to the transient length, i.e. the duration of time for which the image current is detected. In practice, however, both ICR and Orbitrap technologies face certain limitations that restrict the resolution to some maximum value. With FT ICR technology, resolution of several millions is achieved utilizing strong magnets, but these are a major drawback of the FT ICR technology. Super-conducting magnets require liquid helium, which makes maintenance of these mass spectrometers laborious and expensive. Infinitely high resolution can theoretically be obtained with infinite transients, but in practice it is limited due to collisions of the analyte with background molecules that cause the transient to decay after a certain time when the ion packets become incoherent[55]. This effect is more prominent at higher masses - equivalent to larger diameters and collisional cross sections of the molecules - even if the mass-to-charge ratio is the same[52]. In commercial Orbitrap analyzers, the signal becomes indistinguishable from the noise after about 1 s. However, this practical limit is well matched to requirements of the chromatography time-scale, i.e. peptides eluting within several seconds, which is more important than extremely high resolution[48].

The mass resolving power of an analyzer refers to the relative width of a single peak ($m/\Delta m$), which is related to the minimum mass difference ($m_2$-$m_1$) to distinguish two peaks of equal height with a 50% valley in between them[56]. High resolution mass-spectrometry usually refers to resolution values greater 10,000, however, the resolution varies with the mass-to-charge ratio for most instrument types. Due to the linear dependence of ion motion on m/z, the resolution in FT ICR decreases as 1/(m/z). For the Orbitrap analyzer the decay for larger mass-to-charge ratios varies as $1/(m/z)^{1/2}$ because of the harmonic oscillations of ion rings along the z-axis[32]. Therefore, FT ICR resolution can be superior at lower mass-to-charge ratios, whereas this trend turns around for high mass-to-charge ratios[35,56]. Typically, the Orbitrap mass analyzers achieved a resolving power of 60,000 at m/z 400 with a 750 ms transient time; this is commonly referred to as a 1s scan[43]. Strategies to further improve the resolution of the Orbitrap mass analyzer were demonstrated by Makarov *et al.*[57] and are developed as part of this thesis (Articles 2 and 3).

Mass accuracy refers to the deviation between the actual (calculated) and the experimentally determined mass of a compound and it is dependent on the resolution of the mass analyzer. However, there is no linear correlation between these two parameters and both also depend on the peaks abundance, i.e. the signal-to-noise ratio, amongst other parameters. In general, if an instrument is correctly calibrated, high resolution (> 10,000) can provide parts per million (ppm) mass accuracy[56]. Internal calibration procedures add a reference mass, e.g. background ions that are collected and utilized as *lock masses*[58], to each spectrum for correction. This strategy is superior to external calibration before and after an analysis, because it accounts for continuous mass drifts and those on shorter time scale. External calibration, in contrast, avoids reduction of the ion capacity and thereby of the dynamic range of a trap during the actual scan.

With regard to the Orbitrap analyzer, the major overall limitation in terms of ion capacity is the C-trap[59]. Overfilling is very problematic, because it induces space charge effects that lead to non-linear mass errors for different mass-to-charge ratios within the same spectrum. Careful control of the number of ions is ensured by automatic gain control (AGC), in which injection times are matched to space charge limits in order to avoid systematic errors[60]. Efficient recalibration of shotgun proteomics runs can furthermore elegantly be achieved by post-measurement recalibration of the data (Article 5).

The scan speed of a mass analyzer is dependent on a variety of parameters, but as a rule of thumb, it is inversely correlated with resolution. Fourier transformation-based approaches are comparatively slow as the high resolution is proportional to the transient duration. 2D and 3D ion traps are rather fast scanning devices, with up to 33 000 Th/s[44]. However, enhanced resolution scans require longer scan times and ion trap technology by its nature always needs a series of consecutive processes to obtain a mass spectrum. Beam-type analyzers feature the highest scan rates, e.g. time-of-flight analyzers in the range of microseconds. But in order to collect sufficient signal, often many TOF spectra need to be merged, reducing the effective scan speed of the instrument[61].

Finally, the detection principle, Fourier transformation of an image current versus electron multipliers, implies differences in sensitivity. Electron multipliers are capable of detecting single ions, whereas FT-based detection typically requires a minimum of approximately 20 charges to detect a signal clearly distinguishable from the noise[53,62]. Latest improvements in electronics and improved thermal stability, however, make single ion detection in the Orbitrap analyzer feasible, given multiply charged ions and sufficient transient length[55], and allow covering a dynamic range of up to four orders of magnitude within a single spectrum[59].

## Hybrid instrumentation

Many of these novel developments and improvements of mass analyzers have rapidly been applied in hybrid technologies, i.e. the merging of two mass analyzers in one platform. Such instruments exploit the benefits of each component and provide a greater variety of scan modes. The most relevant performance characteristic of hybrid mass spectrometers besides resolution, mass accuracy and sensitivity is the duty cycle i.e. the speed of acquisition and the information density of the data. Merging of mass analyzers introduces a number of challenges because the ions have to travel between the devices and they have to undergo several changes of their properties to be optimally transferred, trapped, fragmented or analyzed in the corresponding part of the instrument. The major effects on the ions are changes in their kinetic energy during transfer from a continuous electrospray ion source to discontinuous trapping devices or further to beam-type analyzers. These challenges are increased in tandem mass spectrometry, which adds isolation and fragmentation of ions, i.e. changes of m/z, mass range and nature of the ions[63].

Before the era of high resolution mass spectrometers, fragmentation analysis was generally either performed in ion traps or in triple quadrupole instruments. The latter feature three consecutive quadrupoles for isolation, fragmentation and mass selection of fragments, respectively, but are not hybrids. A variety of scan modes are available that makes them versatile and popular instruments. For precursor ion scans, the first two quadrupoles are operated in an rf-only mode. Fragmentation spectra can be recorded by using the first quadrupole (Q1) for isolation and the second (Q2) as a collision chamber for fragmentation and the third quadrupole for mass selective transfer of fragment ion peaks, so-called *precursor-fragment transitions*. Triple quadrupole instruments are not generally used in shotgun proteomics but remain popular for targeted proteomics assays. Exchange of Q3 against a time-of-flight analyzer constituted the first hybrid mass spectrometer that became very popular in proteomics[64]. In contrast to early TOF instrumentation relying on matrix-assisted laser desorption ionization, Q TOF instruments that often feature an orthogonal layout between quadrupole and TOF trajectories, can readily be coupled to continuous electrospray sources and therefore to chromatography setups[65]. Furthermore, TOF/TOF instrument configurations can be employed for acquisition of fragment ion spectra in MALDI[66,67].

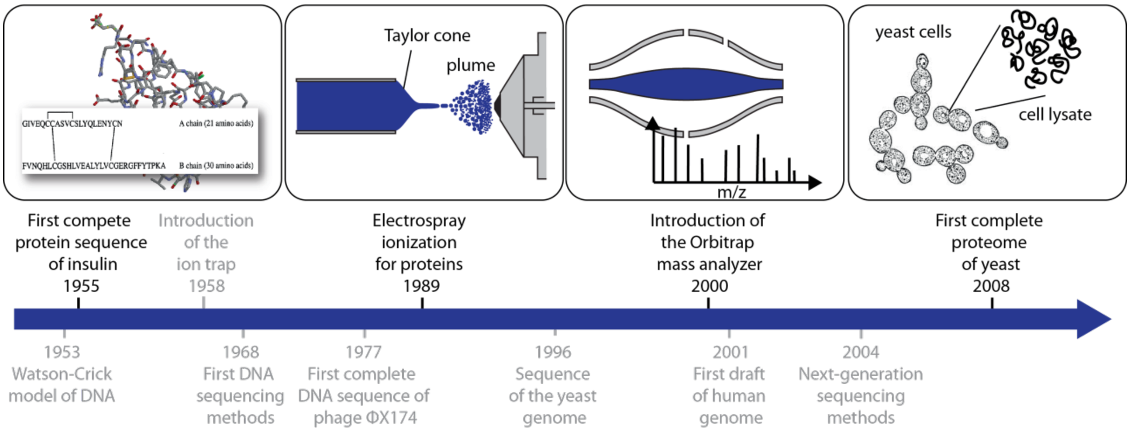
Due to its high resolution and duty cycle of tandem mass analysis, coupling of a linear ion trap with an FT ICR analyzer created a powerful platform for proteomics research[68]. This hybrid allows parallel acquisition of very high resolution full scans (MS) and very fast tandem mass spectra acquisition (MS/MS) with high sensitivity in the ion trap analyzer. The analogue instrument configuration featuring the Orbitrap mass analyzer, the *LTQ Orbitrap*, was introduced only two years later[53].

Collision induced fragmentation in triple quadrupole or Q TOF instruments is relatively efficient and does not suffer from the low mass cut-off inherent to ion trap mass spectrometry. As different physical regions of the instruments are used for tandem mass spectrometry, the process is referred to as *tandem in space* analysis, as opposed to *tandem in time* analysis that is performed in trap based instruments[69]. The latter apply lower energy to the trapped ions, which frequently results in cleavage only of the weakest bonds of the molecules, which then do not undergo any further fragmentation reactions. This often occurs with serine or threonine phosphorylated peptides, for instance, and it can lead to poor fragmentation spectra. However, this problem can be partially overcome by multistage activation, which provides supplemental fragmentation for the main peptide fragments to obtain sequence information[70]. The second generation linear ion trap Orbitrap hybrid mass spectrometer is equipped with a collision cell at the far end of the C-trap that allows beam-type fragmentation at higher collisional energies than ion traps[71]. The fragment ions can then be analyzed with high resolution and high mass accuracy in the Orbitrap analyzer. Higher energy collisional dissociation (HCD) of peptides resembles triple quadrupole fragmentation and is as efficient. This HCD-based *high-high* strategy, referring to high resolution for MS and MS/MS spectra, is routinely usable in shotgun proteomics runs and enables several novel approaches for data analysis (Articles 4, 6, 7).

## Concepts of mass spectrometry-based proteomics

### Historical perspective

The fundamental concept of protein sequencing was pioneered by Fred Sanger in 1955, who presented the first complete protein sequence - that of insulin - based on electrophoretic methods[72-76]. Mass spectrometry was only introduced into protein analysis much later and numerous further developments were required for mass spectrometry-based proteomics to become a mature technology. These included many breakthroughs in the MS instrumentation itself but importantly also in the upstream and downstream workflow. Indeed, first investigations into the complement of proteins in a cell type were performed by biochemical separation such as 2D-PAGE[1] and frequently used antibody-based detection methods. Mass spectrometry was already established for the analysis of small molecules and the ion trap analyzer (*Paul* trap), for which Wolfgang Paul[39] received the Nobel Prize in Physics in 1989, was a major contribution. Finally, the development of soft ionization methods made protein analysis feasible. John Fenn was awarded a share of the Nobel Prize in Chemistry in 2002 for the invention of electrospray ionization (ESI)[27]. This ionization method  had great impact in the field of proteomics together with the novel method matrix-assisted laser desorption ionization (MALDI)[28], which was established by Franz Hillenkamp and Michael Karas around the same time. A protocol to extract proteins from 2D-PAGE[77] after separation and fractionation according to the isoelectric point and size of proteins, and the introduction of nanoelectrospray with reduced flow rate for the first time allowed sequencing of single protein species at high sensitivity[78]. As electrospray ionization of molecules is performed at atmospheric pressure, this technology immediately led to the introduction of on-line coupling of liquid chromatography devices to mass spectrometers, allowing the on-line separation of complex mixtures just before mass analysis (LC/MS), which was a major step towards large-scale analyses[2,79]. Together with the ion trap, the triple quadrupole mass spectrometer (QQQ) and the quadrupole time-of-flight hybrid instrument (Q TOF) were already popular in the early days of mass spectrometry-based proteomics. Today, the Orbitrap analyzer has proven to be especially suitable for proteomics analyses and was also employed for the first *complete proteome* analysis (Figure 6).



GIVEQCCASVCSLYQLENYCN
A chain (21 amino acids)
FVNQHLCGSHLVEALYLVCGERGFFYTPKA
B chain (30 amino acids)
Taylor cone
plume
m/z
yeast cells
cell lysate
First compete protein sequence of insulin 1955
Introduction of the ion trap 1958
Electrospray ionization for proteins 1989
Introduction of the Orbitrap mass analyzer 2000
First complete proteome of yeast 2008
1953 Watson-Crick model of DNA
1968 First DNA sequencing methods
1977 First complete DNA sequence of phage ΦX174
1996 Sequence of the yeast genome
2001 First draft of human genome
2004 Next-generation sequencing methods

Figure 6: Key steps from the first protein sequence towards modern proteomics research.

Mass spectrometry-based proteomics research can be classified into two major approaches: *top-down* versus *bottom-up*. In the first and very intuitive approach intact proteins are ionized and enter the mass spectrometer for determination of their intact and fragment mass-to-charge ratios. The success of this strategy is limited by the complexity of protein mixtures, the fact that each protein has many forms with different masses and that proteins can be difficult to solubilize and separate from each other. Furthermore, electrospray ionization transfers multiple charges to large molecules, which complicates spectra interpretation. For purified proteins the molecular weight is a limiting factor depending on the instrument resolving power. However, at least theoretically, top-down proteomics strategies offer the potential of complete sequence coverage from tandem mass spectra (protein MS/MS) and localization capabilities of post-translational modifications on the entire protein sequence.

## The shotgun proteomics workflow

The vast majority of mass spectrometry-based proteomics studies apply the bottom-up approach, often referred to as *shotgun proteomics*. Protein mixtures are analyzed in an indirect way by cleaving the long amino acid chains in proteins with specific proteases into shorter peptides. The most commonly used enzyme for this purpose is trypsin, due to its high specificity for arginine and lysine[80]. The resulting identified peptides have an average length of about 10 AA and are suitable for analysis in positive ion mode, because the basic amino acids are at the C-terminus of each peptide. When using electrospray ionization, the majority of tryptic peptides carry two to four net positive charges depending on their length. In principle, any other protease can also be selected for protein digestion. LysC has many of the advantages of trypsin, including the desired peptide length as well as sequence specificity C-terminally of lysine and is therefore also frequently used. GluC, AspN or LysN mainly occur in proteomics contexts in order to increase protein sequence coverage[21] or for special applications, e.g. *de novo* sequencing based on electron transfer dissociation (ETD) derived fragments[81]. Protein digestion can be performed in solution, but for removal of detergents such as SDS, it is often desirable to perform in-gel protocols or to employ the filter-aided sample preparation protocol[82]. The resulting peptide mixtures are extremely complex; therefore, liquid-chromatography is almost always applied to separate the peptides. At low pH, achieved by adding formic or acidic acid to the samples and solvents, the protonated peptides bind to the widely used $C_{18}$ material of a chromatography column. During elution with increasing percentage of organic solvent, peptides detach according to their polarity and are on-line sprayed into the mass spectrometer. In shotgun proteomics, optimum gradient separation is important to increase the depth of mass spectrometric analysis and column dimensions are chosen to fit the analytical problem. Very complex mixtures greatly benefit from longer columns and smaller bead size. These provide better separation, due to their increased number of theoretical plates as a consequence of longer interaction with the stationary material[83]. Longer columns and smaller bead size greatly increase the backpressure of the column, unless the flowrate is significantly reduced, which may have the unfortunate consequence of peak-broadening[84]. Increasing the temperature and working with higher pressure overcomes these drawbacks. Recently, the standard high pressure liquid chromatography systems (HPLC) with a pressure limit of several hundred bar have been complemented by UHPLC systems that allow up to 1000 bar. These allow flowrates up to 250 nl/min in 50 cm columns with 1.8 µl beads[85]. Together with the latest generation of Orbitrap-based mass spectrometer, such as the Q Exactive (Article 2), these optimized chromatography set-ups have a high potential for in depth single-run analysis of complex mixtures avoiding any pre-fractionation steps.

In routine practice, efficient reduction of the sample complexity is usually carried out off-line by gel-based methods, size-exclusion or ion-exchange chromatography. Figure 7 illustrates how these optional steps can be performed at the protein or at the peptide level. A very straightforward strategy employed in our laboratory first applies StageTips[86] for strong anion-exchange chromatography and uses different pH buffers for step-wise elution of peptides onto $C_{18}$ StageTips[87]. We also employ OFFGEL technology to provide 12-14 peptide fractions based on isoelectric focusing[88]. These fractions are consecutively analyzed by LC/MS. Alternatively, 2D-chromatography approaches, e.g. multidimensional protein identification technology (MudPIT), facilitate automated fractionation in a two column setup[79]. The peptide mixture is loaded onto a strong cation-exchange column (SCX) and eluted in discrete portions using a stepped salt gradient. Each fraction is collected on a reverse-phase column (RP); the peptides are retained, while the salt buffers divert to waste, and afterwards each fraction is eluted by an organic solvent gradient directly into the mass spectrometer.

Peptide elution is monitored in the full scans covering an m/z range of 300 – 1650 Th and additional information about the peptides is collected in tandem mass spectra. Therefore, the top N – usually 5 to 20 – most abundant precursor ions in each full scan are selected in an automated manner by the instrument software. They are separately accumulated, fragmented and analyzed. A dynamic exclusion list ensures that each precursor is selected only once during typical peptide elution times. Due to its largely unbiased nature, this data-dependent top N strategy is applied for *discovery proteomics* studies investigating various aspects of cell biology[2]. In contrast, *hypothesis-driven approaches* target pre-selected peptides using inclusion lists or multi reaction monitoring (MRM) assays and aim to detect and quantify proteins of prior interest. This strategy verifies the presence of even very low abundant candidates and can be less time consuming than shotgun proteomics[89].

Peptides fragment in a very specific way and cleavage of the peptide backbone is usually most dominant. The amino acid sequence is determined from the tandem mass spectra in conjunction with the accurate precursor mass by an automated database search. Software tools are then applied to re-assemble proteins from the identified peptides; for details see section *Structure and analysis of shotgun proteomics data*.

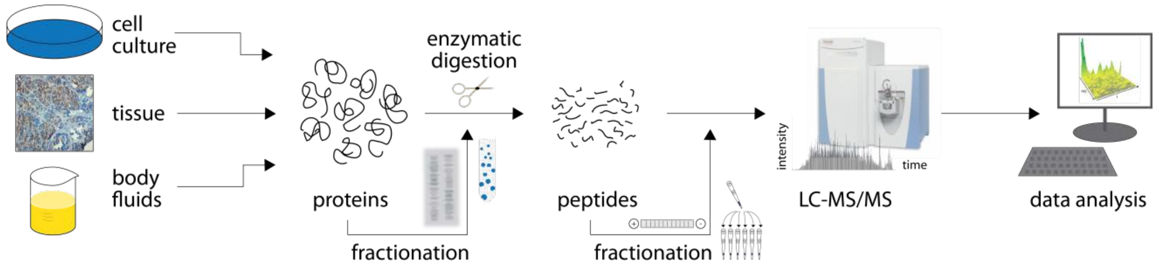


Figure 7: Typical shotgun proteomics workflow.

## Quantitative proteomics

Observing as many different proteins as possible in a single cell type or tissue is a first qualitative step to obtain insight into the biological system under study. Many questions, however, need to be addressed by quantitative comparisons between different cell states, e.g. healthy *versus* disease, or by evaluating the influence of systematic perturbations on the proteome, because this can directly reflect biological function. As mass spectrometry by its nature is not quantitative, different isotope labeling strategies have been developed for pursuing *quantitative proteomics* studies[41,90,91] (Figure 8). Replacing isotopes introduces a difference in the mass of a compound, while minimally affecting chemical properties or biology. For example SILAC (Stable Isotope Labeling of Amino acids in Cell culture), which is a *metabolic labeling* strategy, uses amino acids with specific heavy atoms that are incorporated into proteins and therefore allows labeling of the entire proteome of cultured cells[92]. The SILAC approach has been extended to whole animals, for instance with mice that are fed with a special diet[93]. In a spike-in format[94] SILAC can even be used to quantify human tissues[95]. *Chemical labeling* is a broadly applicable strategy, because it can be used for proteins independent of their origin. Heavy or light isotope tags are covalently bound to peptides - most commonly dimethyl groups[96,97], TMT[98] or iTRAQ[99]. The tags in the latter methods are designed to lead to the exact same precursor mass for the conditions to be compared, thus they are indistinguishable in an MS spectrum. Fragmentation of the labeled peptides, however, reveals different reporter ions in the low mass region of the tandem mass spectrum. These strategies allow relative quantification of two or more conditions in a single LC/MS run. The multiplexing capability of chemical labeling is often higher than that of metabolic labeling approaches; however, the MS/MS-based quantification of reporter ions suffers from contaminating precursor ions that may distort the ratios of the peaks and degrade accuracy.

Recent strategies try to overcome this drawback, but add complexity to the experiment[100,101]. In contrast, the abundance information contained in full scans not only allows precise quantification using several scans, but can also be used in label-free approaches that are completely software-based and increasingly accurately provide ratios of up- or down-regulated proteins when comparing between two or more different LC/MS runs. Label-free quantification requires very high reproducibility and minimal retention time shifts, and it is particularly successful for very high ratios[102].

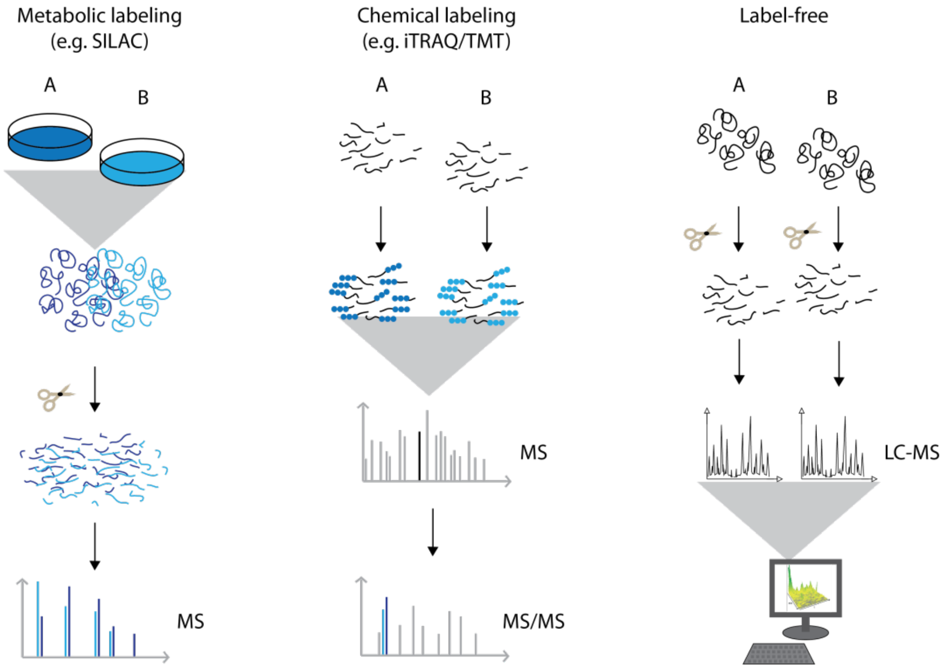


Figure 8: Quantitative proteomics strategies; light and dark blue indicate the presence of light and heavy isotopes in the samples and mass spectra, respectively.

Finally, the analysis of qualitative and quantitative large-scale proteomics datasets represents a challenge in itself, because the amount of data is far beyond what human experts can manually cope with. Modern studies can contain hundreds of thousands of mass spectra and several gigabytes of data. The discipline of *computational proteomics* evolved to handle the data in statistically proper ways and to extract the maximum amount of information

## Structure and analysis of shotgun proteomics data

### Information content in the LC-MS map

Shotgun proteomics data feature three dimensions at the MS level: (1) the mass range and (2) the signal intensity inherent to the mass spectra, and (3) the retention time axis reflecting the order of the peptides eluting from the chromatographic column. The three dimensional elution profiles of the peptide features can be modeled from consecutive full scans. In high resolution mass spectrometry, the isotope patterns are resolved and represented in the LC-MS map. Figure 9 illustrates the extremely high complexity of proteomics data acquired from a complete mammalian cell lysate separated over a two hour gradient.

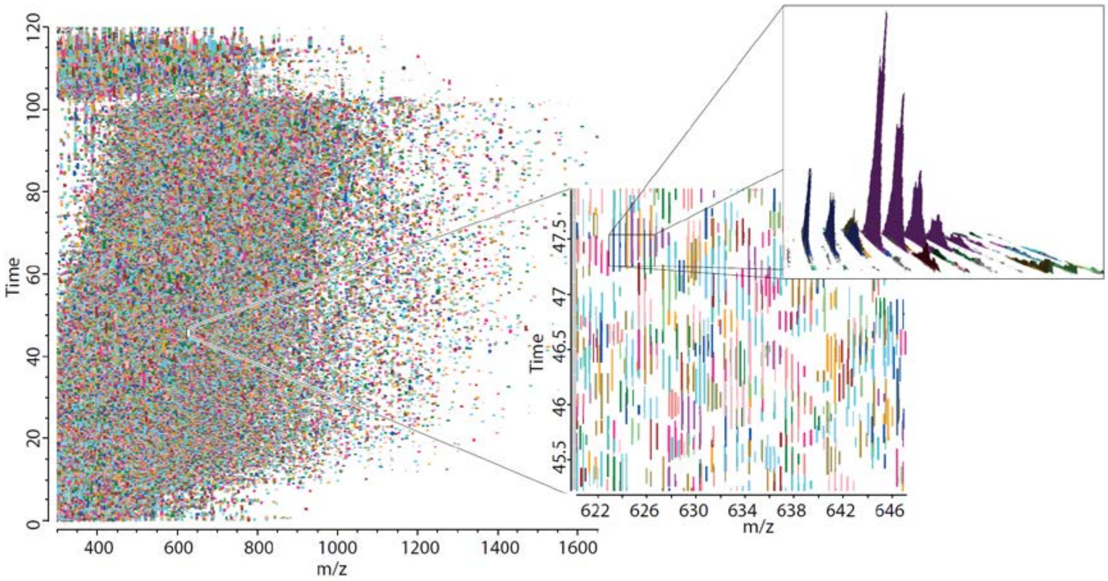


Figure 9: LC-MS map of a mammalian cell lysate and zoom in 2D; selected isotope clusters are depicted as 3D elution profiles.

As high resolution mass analysis allows distinction between the isotope peaks of a cluster, the monoisotopic mass of each peptide can be determined with high accuracy. The MaxQuant software is tailored to such high resolution data and its post-processing algorithms achieve mass accuracies below 100 parts per billion (ppb) for most peaks[103]. A Gaussian curve is fitted to the three central data points of each MS peaks and an intensity-weighted mass deviation is calculated for the assembled three-dimensional peak profiles of all features detected in the LC-MS map. Non-linear mass recalibration based on peptide charge pairs allows calculating an individual mass accuracy for each peptide precursor taking the abundance of the peak and the number of measurements into consideration[104]. This very high mass accuracy is a potent filter option for database search as it significantly reduces the search space of possible peptide precursors. Along the same lines, confident assignment of the isotope clusters functions as noise filter for the data. One of the primary motivations for creating the MaxQuant software environment was SILAC-based quantification; recently this was extended by sophisticated label-free quantification algorithms[102].

## Peptide fragmentation spectra

A second layer of information about the peptides, which is not displayed in the LC-MS map, are the fragmentation spectra that are acquired either in parallel or alternately with the full scans. Depending on the capabilities of the instrument, up to 20 tandem mass spectra are acquired in each duty cycle, i.e. between two consecutive full scans, usually in a data-dependent fashion (Figure 10).

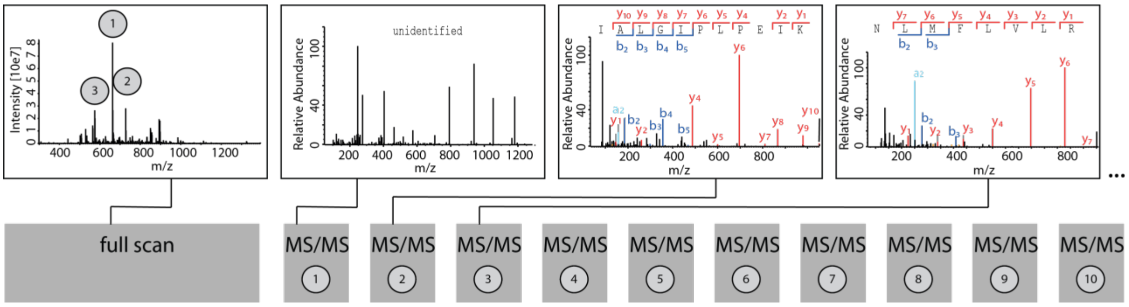


Figure 10: Schematic of a data-dependent MS and MS/MS cycle (top10); peptide candidates targeted in the MS/MS scans (1-10) are selected based on their peak intensity in the previous full scan (left). The depicted examples show that the first fragment ion spectrum (1) remained unidentified, whereas (2) and (3) could be assigned a peptide sequence during data analysis.

For a tandem mass spectrum, the precursor ion of interest is isolated in a selection window of a few Th and typically accumulated for several tens of milliseconds, until a desired number of ions is collected. Fragmentation of the precursor ions is subsequently carried out in the appropriate section of the mass spectrometer. MS/MS spectra used to be recorded at nominal resolution in the ion trap part of hybrid Orbitrap analyzers, but recently it has become feasible to acquire both full scans and fragment ion scans at high resolution[43,105].

The composition of tandem mass spectra is dependent on the peptide fragmentation chemistry induced by different dissociation principles. Collisions of protonated peptides with neutral gas atoms in CID and HCD result in predominant cleavage of amid bonds in the peptide backbone. Thus, they generate sequence ladders of b- and y-type ions from the N- and C-terminus, respectively. Complementary fragmentation methods, termed electron capture dissociation or electron transfer dissociation, induce peptide fragmentation by a single electron or an electron transferred from negatively charged fluoranthene molecules, respectively[106,107]. The peptide radical ions generated in this way fragment into c- and z•-type ions by cleavage of the N-$C_\alpha$ bonds. In each case, the regular backbone ions provide information on the amino acid sequence. Tandem mass spectra also feature a great number of additional peaks that are derived from various other fragmentation pathways, but these are not generally used in automated peptide identification (Articles 4, 6 and 7).

## Data analysis strategies

Software tools usually follow one of two major approaches for peptide identification in large-scale datasets[18]. One is solely based on the data in the tandem mass spectrum and is referred to as *de novo* sequencing. It frequently utilizes graph-based approaches[108,109]. Fragment ion peaks represent the nodes of the graph, while the edges are the mass differences between two peaks. The amino acid sequence is determined by finding the longest and most likely path through the graph. *De novo* sequencing has the advantage of being completely independent of any previous knowledge such as protein or gene sequences stored in databases. This allows, in principle, to identify novel proteins and genes or to work with unsequenced organisms. High resolution tandem mass spectra are particularly advantageous for *de novo* sequencing approaches, because the high mass accuracy allows distinguishing amino acids with the same nominal mass such as Q from K and amino acid pairs with very similar mass. However, very high spectrum quality and sufficiently long runs of successive fragment ions are required.

Routine shotgun proteomics studies aim to generate as many peptide identifications as possible. Therefore database supported identification strategies are more suitable, because these tolerate lower data quality. Tandem mass spectra are matched against theoretical spectra created from protein sequences of the organism being studied. Protein databases usually refer to the sequence of the decoded reference genome; however, there are various approaches to incorporate protein variant data from other sources such as transcriptome data generated by RNA-seq approaches (Article 8). Depending on the experimental conditions, several parameters in a peptide database search are defined by the user. These include enzyme specificity, fixed and variable modifications and the mass tolerance, which depends on the MS system that was used. The peptide search engine then compares a list of peptide sequences to the spectrum and assigns peptide spectral matches with a score reflecting the confidence of identification. The commercial search engine *Mascot*[110] as well as the recently released Andromeda search engine (Article 4) apply probability-based scoring models. This means that the score is calculated based on the probability that matches between observed and calculated peaks in the tandem mass spectrum occurred randomly. The peptide identification score can be extended by an expectation value providing a statistical confidence measure and taking the accurate precursor mass into consideration. The latter is often sufficient to independently determine the molecular composition of short peptide fragments[104].

The number of incorrect peptide spectrum matches in the entire dataset is controlled by restricting the false discovery rate (FDR) among all identified sequences. According to the target decoy approach, the database is extended by reversed sequences and the resulting number of matches that correspond to such reverse hit directly yields the FDR[111]. In general, 1% false positives identifications are accepted at the peptide level; however, the protein FDR needs to be restricted in a similar manner to provide the same level of confidence in protein hits. When reassembling the proteins from the identified peptides, correct identifications mostly feature several unique peptides, whereas incorrect proteins identifications are often identified by just a single peptide sequence. In shotgun proteomics, overlapping or similar protein sequences, e.g. isoforms, are only distinguishable if differentiating peptides were identified; otherwise compatible protein sequences are combined into protein groups.

# 2 Challenges and Limitations in Shotgun Proteomics: A Baseline Study

**Article 1 More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS**

Annette Michalski, Jürgen Cox, Matthias Mann,
*J Proteome Res*, 2011, 10(4),1785.

## Prologue

With the advent of high resolution hybrid mass spectrometers in shotgun proteomics research, the accurate precursor mass derived from MS spectra complements the tandem mass spectrometric information for peptide identification. The most common strategy is data-dependent selection of the N most abundant precursor ions from the previous full scan. Advantages of this strategy are that (1) the most abundant peptide features are fragmented, which guarantees the highest available quality of tandem mass spectra and that (2) the duty cycle of the mass spectrometer is as high as possible. Conversely, the approach is limited by the sequencing speed of the instrument and - for low abundant ions - by the sensitivity of the mass analyzer, which is related to the brightness of the ion source and transmission of the ion beam.

In 2009, when this thesis was started, a novel linear ion trap Orbitrap mass spectrometer with 5-10 fold improved sensitivity, the LTQ Orbitrap Velos, was introduced[43]. For the first time, very high resolution and high mass accuracy full scans and tandem mass spectra could be routinely acquired in large-scale datasets. Higher energy collisional fragmentation (HCD) rapidly enabled a novel standard for shotgun proteomics experiments, even for the analysis of phosphorylated peptides[112,113]. In our experience, the identification rate in peptide mixtures resulting from mammalian cell lysates is usually higher than 50%, which is facilitated by high MS and MS/MS spectra quality. This translates to 10,000 unique peptide sequences and 1,800 proteins during 90 minutes gradient time. This was significantly higher than the identification rates obtained with low resolution ion trap measurements.

Motivated by the overwhelming complexity present in the LC-MS map of standard shotgun proteomics analyses, we were interested in estimating the capabilities and the potential of the latest mass spectrometric instrumentation with respect to mammalian cell lysate. Therefore, we carefully investigated and quantified the peptide features, represented by isotope clusters in the LC-MS map, over the entire elution time. This analysis revealed that more than 100,000 features, likely representing peptides, are detectable in the full scans. Obviously, this number is dependent on the full scan resolution as well as on the cycle time, i.e. the number of full scans available to detect the isotope clusters in the first place. Experiments with higher resolution and increased scan speed revealed that the number of features can even be significantly higher than our rather conservative estimate of 100,000. The accessibility of these peptide features would appear to only be limited by the sequencing speed of the mass spectrometer. However, high sequencing speed is only beneficial if the targeted precursor ions are sufficiently abundant and the instrument is sufficiently sensitive to collect the requested number of ions for the fragment spectrum in an appropriate time interval. Interestingly, our study revealed that complex peptide mixtures contain many candidates that fulfill these criteria and could therefore potentially have been identified, if sequencing speed had been fast enough to target them.

We also analyzed the detrimental influence of co-eluting precursor ions in the same isolation window on the *identifiability* of the tandem mass spectra. This effect is particularly severe for low abundance peptides. Strikingly the median precursor intensity fraction (PIF), a key measure to estimate the identifiability, was only 0.14 for all detected HeLa peptides. Incorporating a fractionation step only achieved slight improvements. Due to decreasing transmission efficiency and overlapping of isotope clusters, reduction of the size of the isolation window before peptide fragmentation is not an appropriate solution, either. Thus, insufficient purity of precursor isolation sets a principal limitation to data-dependent analysis of complex mixtures. At the same time, our article reveals that the potential of the shotgun proteomics strategy is not yet fully exploited and challenges manufacturers to further improve instrument capabilities.

In the long run, findings such as ours may encourage mass spectrometrists to develop novel strategies for data acquisition. Independently, first steps have been taken towards data-independent approaches and scanning of broader mass ranges, which deliberately generate very complex fragmentation spectra[114,115]. This, however, provides huge challenges to informatics analysis tools and so far, these attempts still suffer from limited dynamic range and insufficient sensitivity.

Journal of proteome research

ARTICLE

pubs.acs.org/jpr

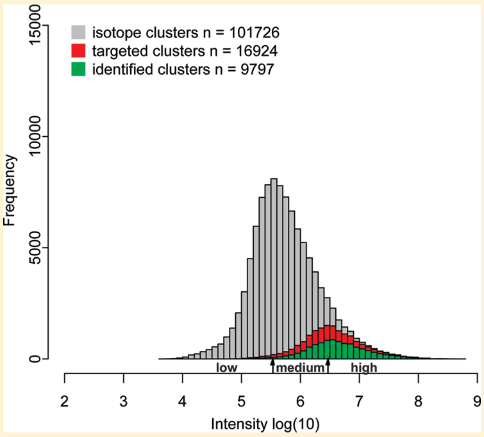
# More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC−MS/MS

Annette Michalski, Juergen Cox, and Matthias Mann*

Department for Proteomics and Signal Transduction at the Max-Planck Institute of Biochemistry, Martinsried, Germany

S *Supporting Information*

**ABSTRACT:** Shotgun proteomics entails the identification of as many peptides as possible from complex mixtures. Here we investigate how many peptides are detectable by high resolution MS in standard LC runs of cell lysate and how many of them are accessible to data-dependent MS/MS. Isotope clusters were determined by MaxQuant and stringently filtered for charge states and retention times typical of peptides. This resulted in more than 100 000 likely peptide features, of which only about 16% had been targeted for MS/MS. Three instrumental attributes determine the proportion of additional peptides that can be identified: sequencing speed, sensitivity, and precursor ion isolation. In our data, an MS/MS scan rate of 25/s would be necessary to target all peptide features, but this drops to less than 17/s for reasonably abundant peptides. Sensitivity is a greater challenge, with many peptide features requiring long MS/MS injection times (>250 ms). The greatest limitation, however, is the generally low proportion of the target peptide ion intensity in the MS/MS selection window (the "precursor ion fraction" or PIF). Median PIF is only 0.14, making the peptides difficult to identify by standard MS/MS methods. Our results aid in developing strategies to further increase coverage in shotgun proteomics.



**KEYWORDS:** tandem mass spectrometry, shotgun proteomics, peptide identification, proteome identification, dynamic range

## INTRODUCTION

In bottom-up, MS-based proteomics, protein mixtures are digested to peptides with proteases.[1,2] Most typically, peptides are separated with liquid chromatography, online electrosprayed[3] and analyzed by MS scans. The most abundant signals in the MS scans are isolated, fragmented and analyzed by an MS/MS scan. The MS and MS/MS information is combined and used to identify the peptides and proteins in sequence databases. This basic scheme, supplemented with peptide quantification at the MS or MS/MS level, is successfully used in a wide variety of biological applications.[4−6] A longstanding challenge of the field is to identify as many peptides as possible in LC−MS/MS runs. Previously, using low resolution instruments, peptide identification rates were as low as a few percent of all fragmentation events.[7] High resolution MS, for example in the form of hybrid instruments such as the linear ion trap Orbitrap,[8] combined with advanced computational proteomics algorithms, now routinely enabled identification of more than half of all MS/MS spectra.[9] The Orbitrap analyzer combines very high resolution (60 000 at $m/z$ 400) as well as high sensitivity and for complex mixtures a dynamic range of at least $10^3$. In particular, the high mass resolution, which is a precondition for distinguishing coeluting peptides with similar masses, contributes to the success of shotgun proteomics. Highly confident assignment of the charge state allows definite distinction of isotope clusters of multiply charged peptides from noise peaks. In routine biological applications, when coupled to reversed phase chromatography, this allows the identification of several thousand peptides and proteins with gradients of a few hours.[10]

Even more identifications, and even complete proteomes, can be obtained by more sophisticated sample preparation such as prefractionation or by prolonged gradients.[11] However, it would be attractive to increase the information that can be drawn from single LC−MS/MS runs, because this implies less sample consumption and measuring time. A recent technological improvement in the Orbitrap platform has been the introduction of the LTQ Orbitrap Velos, which provides sufficient speed and sensitivity for analyzing fragment ions with high resolution and ppm mass accuracy in a standard "top10" method.[12] However, despite ppm or subppm level accuracy in the MS and MS/MS scans, the precursor isolation in the linear ion tap still requires a window of a few Th to ensure sufficient sensitivity. This low resolution precursor isolation is inherent in the technology of ion trap and quadrupole devices. In combination with the very high sample complexity typical of proteomics, it inevitably leads to the cofragmentation of precursor ions that happen to be present in the same isolation window. Figure 1 illustrates the density of

**Received:** October 21, 2010
**Published:** February 10, 2011

ACS Publications © 2011 American Chemical Society

dx.doi.org/10.1021/pr101060v

Figure 1:

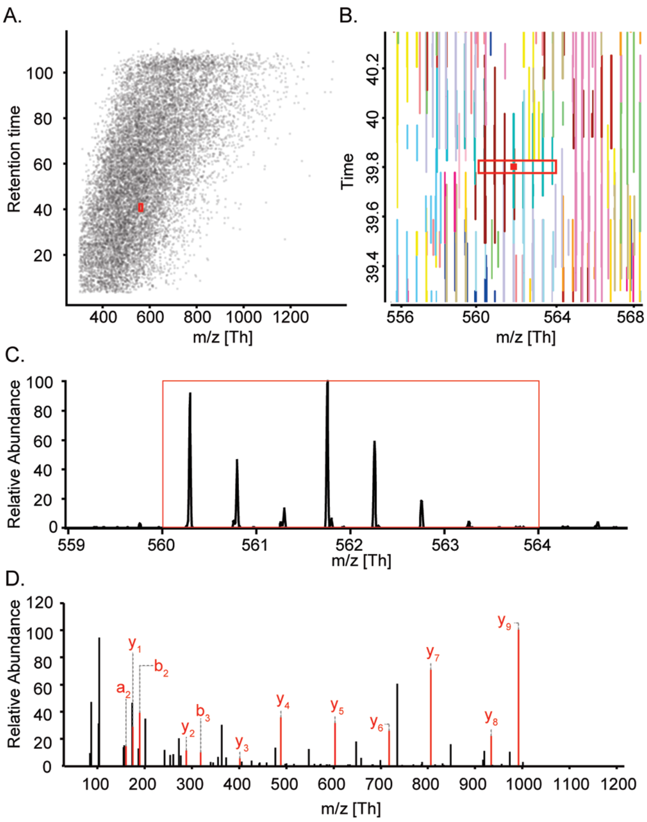


**Figure 1.** Visualization of peptides measured by LC−MS/MS in complex mixtures. (A) LC−MS map showing a black square for each detected peptide. (B) Zoom of the red rectangle in (A), with eluting isotope clusters and a selection window for the MS/MS marked in red. (C) Precursor selection window in the MS spectrum. (D) Mixture MS/MS spectrum resulting from fragmentation of the ions in (C).

peptides eluting in a typical LC−MS/MS run as well as an example of a cofragmentation event in an LC−MS map and in an MS view. Co-fragmentation of peptides raises several challenges. In quantification methods that are based on low mass reporter ions, both peptides contribute to the quantified intensities. These spectra should be discarded, or their acquisition avoided in the first place.[13] More generally, cofragmented, mixture or "chimera spectra" may reduce peptide identification rates[14] but they also present the opportunity to identify more than one peptide from an MS/MS spectrum. Several algorithms for second peptide identification have been published with this goal in mind, generally with low resolution MS/MS data.[15−18]

As shown above, researches have so far mainly concentrated on improving identification rates of targeted peptides. However, given the fact that most MS/MS spectra can now be identified, it would be interesting to investigate the peptides that are present in LC−MS runs but that are not targeted for MS/MS. To our knowledge, there is no in-depth study of the extent and the attributes of this population. In particular, it is not clear if these peptides are in principle accessible to shotgun proteomics. Here we determine how many peptide species are detectable with state-of-the-art LC−MS and LC−MS/MS using the feature detection algorithms of MaxQuant combined with a SILAC analysis. This establishes a conservative estimate of the number of peptides present in standard MS analysis of complex cellular lysates. We then examine the mass spectrometric sequencing speed and sensitivity necessary to target the peptide population not currently fragmented. We also investigate the target peptide ion intensity in the MS/MS selection window, which we term the "precursor ion fraction" or PIF, as a function of the standard parameters employed in LC−MS/MS. Most detectable peptides have a very low PIF and we discuss implications of this finding for further development towards comprehensive proteomics analysis.

## ■ EXPERIMENTAL PROCEDURES

### SILAC Labeling

HeLa S3 cervix carcinoma cells were grown in RPMI 1640 medium supplemented with bovine serum and Penicillin/Streptomycin (1:1000). One out of two populations of HeLa S3 cervix carcinoma cells were grown in RPMI 1640 medium supplemented with dialyzed bovine serum and two isotopic variants of lysine ($^{13}C_6$,$^{15}N_2$-L-Lys) and arginine ($^{13}C_6$,$^{15}N_4$-L-Arg). The cell populations were harvested by spinning for 5 min at 400 g. The cells were washed with phosphate-buffered saline and mixed (heavy-to-light 3:1). Lysis was carried out by douncing (30−40 strokes) in 5 volumes of Hepes KOH (10 mM), pH 7.9, $MgCl_2$ (1.5 mM) and KCl (10 mM) in a homogenizer on ice. The nuclear and cellular constituents were separated by centrifugation for 15 min at 3900 rpm, before the crude cytoplasmic supernatant was ultracentrifuged for 1 h at 60 000× *g*. Dilution with glycerol (10% final concentration) and NaCl (150 mM final concentration) yielded the cytoplasmic extract that was snap-frozen in liquid nitrogen.

### Protein Digestion

Total HeLa cell lysate (unlabeled and SILAC-labeled) or *E.coli* lysate were treated with a urea (6 M) and thiourea (2M) solution. The proteins were incubated with Dithiotreitol (DTT) (1 mM) for 30 min and iodoacetamide (IAA) (55 mM) for 20 min at room temperature. Lys-C (1 $\mu$g/50 $\mu$g protein) (Wako) was added and the mixture was incubated for 3 h at room temperature. After 1:4 dilution with water, trypsin (1 $\mu$g/50 $\mu$g protein) (Promega) was added and the sample was incubated for 12 h at room temperature. The digestion was stopped by addition of formic acid (3%).

Prefractionation of HeLa lysate was carried out by 1D-SDS-PAGE (4−12% Novex mini-gel) (Invitrogen) in three separate lanes. Colloidal Coomassie (Invitrogen) was used for staining of the proteins before each lane was cut into 8 slices. All gel slices were subjected to reduction of the proteins with DTT and subsequently alkylated with IAA. In-gel digestion with trypsin was carried out at 37 °C for 12 h followed by extraction of the tryptic peptides with 3% TFA in 30% ACN. Organic solvent was removed and the peptide mixture was concentrated and desalted on reversed phase $C_{18}$ StageTips.[19] The peptides were eluted twice shortly before high resolution LC−MS/MS analysis with 20 $\mu$L buffer B (80% ACN in 0.5% acetic acid) into a 96 sample well plate (Abgene). Organic solvents were removed in a SpeedVac concentrator and the final concentration was adjusted with buffer A* (2% ACN in 0.1% TFA).

### LC−MS/MS Analysis

The peptide mixture was separated by a nanoflow HPLC (Proxeon Biosystems, now Thermo Fisher Scientific) and coupled online to an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) with a nanoelectrospray ion source (Proxeon Biosystems). Loading of the routinely used amount of peptides (2 $\mu$g) onto a $C_{18}$-reversed phase column (15 cm long, 75 μm inner diameter) was carried out with a maximal flow rate of 500 nL/min controlled by IntelliFlow technology. The chromatography columns were packed in-house with ReproSil-Pur $C_{18}$-AQ 3 μm resin (Dr. Maisch) in buffer A (0.5% acetic acid). Peptides were eluted with a linear gradient of 5−60% buffer B (80% ACN and 0.5% acetic acid) at a flow rate of 250 nL/min over 60, 90, or 200 min depending on the experiment. Total LC−MS/MS time was about 40−50 min longer due to loading and washing. Data were acquired using a data-dependent "top 10" method, dynamically choosing the most abundant precursor ions from the survey scan (mass range 300−1650 Th) in order to isolate them in the LTQ and fragment them by HCD.[20] Dynamic exclusion was defined by a list size of 500 features and exclusion duration of 90 s. Early expiration was disabled to decrease the resequencing of isotope clusters. The isolation window for the precursor selection was varied in different runs between 1, 2, 4, 8, 16, and 32 Th. For the survey scan a target value of 1 000 000 and a resolution of 30 000 at $m/z$ 400 were set, whereas the target value for the fragment ion spectra was set to 40 000 ions and the resolution to 7500 at $m/z$ 400. The lower threshold for targeting precursor ions in the MS scans was 5000 counts.

### Data Analysis

The mass spectrometric raw data were analyzed with the MaxQuant software[9] (version 1.1.1.17). A false discovery rate (FDR) of 0.01 for proteins and peptides and a minimum peptide length of 6 amino acids were required. The mass accuracy of the precursor ions was improved by the time-dependent recalibration algorithm of MaxQuant. The Andromeda search engine[21] was used to search the MS/MS spectra against the IPI human database (containing 87 061 entries) combined with 262 common contaminants and concatenated with the reversed versions of all sequences. Enzyme specificity was set to trypsin specificity, allowing cleavage N-terminal to proline. Further modifications were cysteine carbamidomethylation (fixed) as well as protein N-terminal acetylation and methionine oxidation (variable). MaxQuant was used for scoring of the peptides for identification. A maximum of two missed cleavages were allowed. Peptide identification was based on a search with an initial mass deviation of the precursor ion of up to 7 ppm. The fragment mass tolerance was set to 20 ppm on the $m/z$ scale. Analysis of the data provided by MaxQuant was performed in the R scripting and statistical environment[22] supplemented by the ggplot2 package.[23]

The data sets used for analysis have been deposited at TRANCHE (www.proteomecommons.org).

## ■ RESULTS AND DISCUSSION

### Determination of the Number of Detectable Peptides

To obtain a conservative estimate of the number of peptides detectable with high resolution MS, we prepared lysate from a mammalian cell line and analyzed it on the LTQ-Orbitrap Velos platform. We used a standard "high−high" strategy, meaning that an MS scan in the Orbitrap analyzer at 30,000 resolution was followed by up to 10 MS/MS events obtained by HCD with fragment spectrum analysis at 7,500 resolution (see Experimental Procedures).

We chose a "high−high" strategy over the commonly applied "high−low" strategy with CID and LTQ detection of fragments used on Orbitrap XL instruments, even though the "high-low" strategy in principle would need less precursor ions, making it more sensitive. MS/MS scan speeds are similar but the "high−low" strategy allows for parallel operation, giving it an advantage in total targeted precursors. Nevertheless, in previous investigations we have found that the "high−high" strategy identifies more peptides overall,[12,24] and we therefore used it in these investigations. In any case, the results obtained here are mainly concerned with the number and nature of precursor ions, which is independent of the MS/MS detection. Those results that depend on MS/MS acquisition can easily be converted to instruments with different sensitivity or scan speeds.

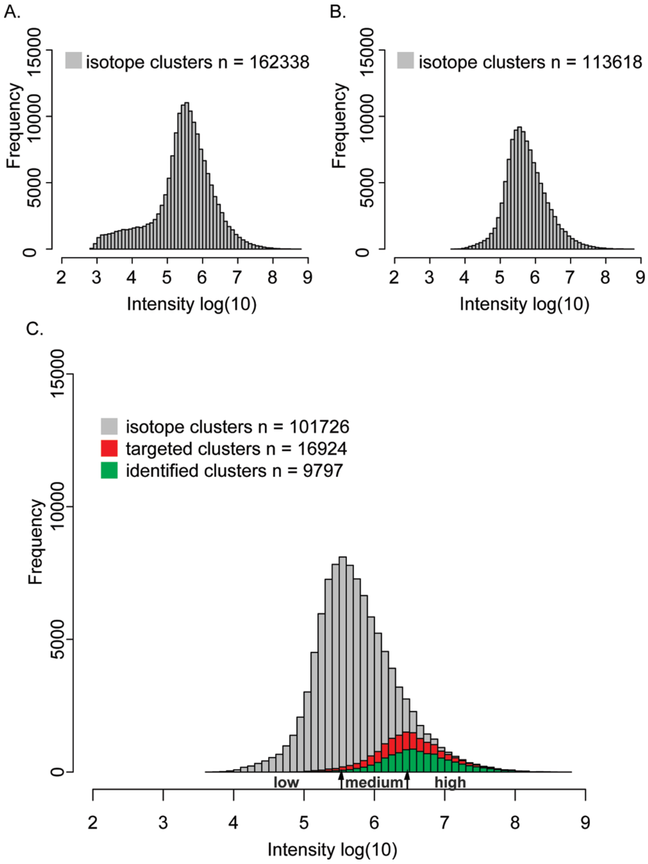


**Figure 2.** Histogram of intensity of the detected features in an LC−MS/MS run. (A) Isotope clusters determined by MaxQuant. (B) Same as (A) but filtered between 5 and 30% organic phase. (C) Same as (B) but filtered for charge state >1. Gray, all peptides; red, targeted peptides; green, targeted and identified peptides.

We determined peaks in the LC−MS data in the MaxQuant environment using standard criteria.[9] Briefly, MaxQuant analyzed the 3805 high resolution MS scans in the MS data (RAW file) and assigned isotope clusters whenever two isotope states (usually $^{12}C$ and one $^{13}C$) occurred with a mass tolerance of 5 ppm and in at least two consecutive (or next to consecutive) MS scans. These criteria are reliable noise filters.[9] Figure 2A shows a histogram of the 162 338 isotope clusters resulting from this analysis as a function of the peptide signal. The distribution is generally log-normal but there is clearly an additional population at very low peptide intensities. We suspected that the second distribution was partly due to nonpeptide compounds. Indeed when plotting only peaks eluting between the start and the end of the main population in the Total Ion Current (TIC) graph (10 and 100 min; Suppl. Figure 1, Supporting Information), the very low intensity population disappeared (Figure 2B). This reduced the number of detected peaks by 30% to 113 618. Since chemical noise is almost always singly charged, whereas tryptic peptides are likely to be at least doubly charged, we next filtered the data by charge state. This is a strongly conservative step, because we routinely also observe singly charged identified peptides. The filtering resulted in a final number of 101 726 likely peptide features (Figure 2C). We performed an analysis to estimate the proportion of peptides with multiple charge states among all detectable peptides. Peptide features that carry the same mass (±0.005 Da) and elute in a retention time window of ±7 s, which is equivalent to two full scans before or after the intensity maximum of the reference peak, are considered to be the same peptide. According to these criteria we find that 13 550 peptide features (13%) result from multiple charge states. This number is in the same order of magnitude as the proportion of multiple charge state identifications that we find among the identified peptides (17%). The repeated fragmentation of different charge states of the same peptide could be reduced by more sophisticated data dependent acquisition tools. Overall, accounting for the fragmentation of different charge states as well as resequencing of previously targeted peptides would not appreciably change the statistics.

Data-dependent approaches are applied in a majority of proteomics experiments. Our data clearly show that even state-of-the-art mass spectrometers such as the LTQ Orbitrap Velos employed here do not yet come close to targeting the enormous number of peptides in complex proteomics samples. Out of all apparent peptide features that are eluting between the start of peptide elution to the end of the gradient 16 924 (16%) were targeted for MS/MS, 9797 of which led to identification (success rate of 58% at 1% FDR). The overall number of MS/MS scans was 21 906. Similar to the distribution of all (gray) peptides, the distribution of targeted (red) peptides and identified (green) peptides are well behaved on a log scale. Interestingly, red and green peptides have the same intensity distribution, whereas both are wedged into the high intensity tail of the gray peptide distribution. This indicates that within the targeted range, identification success depends little on peptide intensity. The data dependent acquisition mode ensures that the most abundant peptides are fragmented and since the instrument is working at maximum fragmentation speed, there is no capacity for fragmenting the less abundant precursor ions. The low intensity limit of the gray distribution does not even indicate absence of further peptides but rather that the dynamic range limit of the mass analyzer for MS is reached.

We next investigated the physicochemical properties of the three populations (gray, green, red) to ensure that the detected peptide-like features indeed represent peptides. A comparison of their mass, $m/z$, retention time and charge distributions (Suppl. Figure 2, Supporting Information) provided strong evidence for this assumption.

To independently verify that the gray distribution mainly represents peptides we performed a SILAC experiment by mixing equal amounts of heavy and light labeled cell populations. MaxQuant detected SILAC pairs using the isotope criteria described above in addition to the requirement for pairs of coeluting isotope clusters with the precise mass difference of the SILAC label. This allows detecting peptide pairs even if they were not fragmented and identified, and distinguishes them from chemical noise with near certainty. Note that due to the more stringent criteria for SILAC pair detection we found that a minimum intensity of $10^5$ is needed for SILAC pair detection and considerably fewer peptide features are expected in this experiment (Suppl. Figure 3, Supporting Information). While the overall number of isotope clusters in the SILAC sample (167 811) is comparable to the number of isotope clusters in the unlabeled HeLa sample (162 338), we only detected 33 281 SILAC pairs in total (representing 66 562 isotope clusters), at least one of which was fragmented in 36% of the cases (Suppl. Table 1, Supporting Information). For 69% of these we unambiguously identified a peptide sequence in the database (Suppl. Figure 4, Supporting Information). Examination of the mass, $m/z$ and retention time distributions of the SILAC population and comparing these to the distributions of the unlabeled peptides described above (Suppl. Figure 5, Supporting Information), revealed that there was again excellent agreement for all of these properties, supporting that the features detected above were indeed peptides.

To ascertain that the above numbers of the unlabeled HeLa cell lysate were typical, we repeated the analysis with two independent experiments, which yielded very similar results (Suppl. Table 2, Supporting Information). Furthermore, we analyzed *E. coli* cell lysate, with similar results except that overall numbers were reduced by about one-third (Suppl. Table 2). This is not surprising because bacterial proteomes are somewhat less complex than mammalian proteomes (the genome is only about 20% as large and there is no alternative splicing). The proportions between detected, targeted and identified peptides were the same as in the mammalian proteome, suggesting that they reflect characteristics of complex proteomes in general (Suppl. Table 2).

Together, our results demonstrate that at least 100 000 detectable peptide peaks elute under standard LC−MS/MS conditions and this number is even larger when using faster scan rates. Although many of these peaks would redundantly identify the same peptide sequences or peptide sequences of already identified proteins, our results suggest that a single LC−MS run contains sufficient peptides to identify a large part of the cellular proteome.

### Requirements for MS/MS Sequencing Speed

Having established that at least 100 000 distinct peptide peaks elute over a standard gradient, we next investigated the MS/MS sequencing speed that would be necessary to target them. First we plotted the actual frequency of MS/MS events over the gradient (Figure 3). From the start of peptide elution to the end of the gradient, the instrument was sequencing at its maximum rate, indicating that there were sufficient precursor peaks above threshold value at any elution time. Next, we divided the peptide features into high abundance (>3 × $10^6$ cps; the maximum of the identified peptide distribution), medium abundance (>4 × $10^5$ cps; down to the lowest identified peptide intensities) and low abundance (the rest of the gray peptide distribution in Figure 2C). The peptide features of these three abundance classes are plotted as a function of retention time in Figure 3. Even among the high abundance peptides, there are several occasions where more than three peptides per second elute (dark gray line exceeds the red line in Figure 3). This is possibly the reason that not all high abundance peptides were targeted for MS/MS. The medium abundance peptides, which is the entire population that exceeds the threshold for MS/MS picking, already requires a sequencing speed of 17 MS/MS events per second. If all detectable peptides should be targeted, then a sequencing speed of up to 25 MS/MS events per second would be necessary (Figure 3). Clearly, further improvement of the sequencing speed over the 3.3 MS/MS spectra per second achieved here, would be desirable and is certainly within reach. When keeping the same threshold for picking of peaks, this would have the effect of targeting the entire median abundance peptide distribution (the red population would encompass the right half of the peptide population). However, precursor ion injection times for MS/MS are increasingly limited at high sequencing speed. For this and other reasons that we investigate below, high MS/MS frequency by itself would not necessarily lead to a corresponding increase in identified peptides (green population).

## Requirements for Sensitivity in MS/MS

Next we investigated the MS/MS sensitivity necessary for successful fragmentation of the detectable peptide features. In figure 4A the intensity at the peak apex versus the ion injection time necessary to reach the target value of 40 000 ions appears as a linear relationship in a log–log plot. The ion injection times can be determined for the high abundance peptide population defined above and the medium abundant population below $3 \times 10^6$, as well as the border to the low abundance peptide population at $4 \times 10^5$ (Figure 2C). To cover the medium abundant peptide features, ion injection times up to 100 ms are necessary. The low intensity proportion of the gray peptides would need to be accumulated for 2 s to reach their target value.

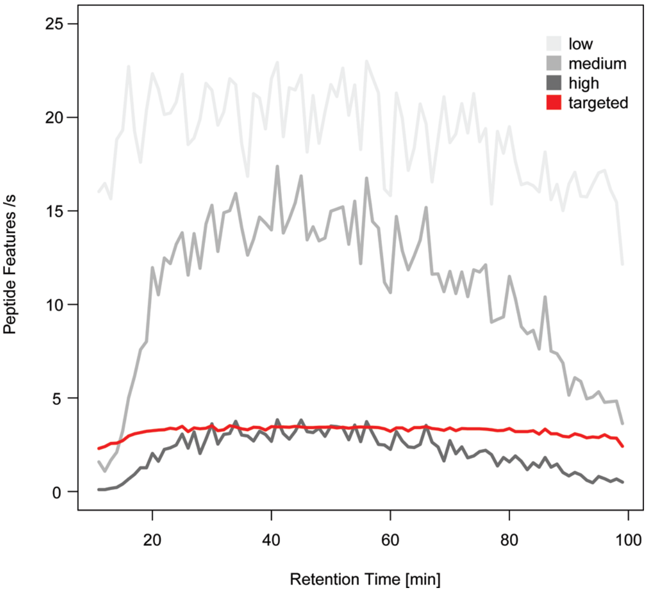


**Figure 3.** Peptide features eluting as a function of retention time. Gray scale values for high, medium and low refer to the peptide intensity regions marked in Figure 2. The red line marks the actual MS/MS frequency.

Furthermore, in complex samples the targeted peptide is usually accompanied by other candidates in the isolation window that reduce the ion injection time needed for the requested target value (i.e., the PIF is not 1). Figure 4B shows that this effect significantly reduces the ion injection times in particular for the low abundant peptides if a fixed overall number of ions are accumulated by the automatic gain control of the instrument. This is unwelcome because it reduces the number of desired precursor ions contributing to the MS/MS spectrum. Finally, isolation and fragmentation of the precursor ion does not always happen at the peak apex during chromatographic elution, thus injection times would tend to be longer than the ones we calculate.

In conclusion, sensitivity in the MS/MS mode would be limiting for at least half of the detectable peptide population, even on an extremely high sensitivity instrument such as the LTQ Orbitrap Velos.

## Required Precursor Intensity Fraction

Co-elution of peptides with similar mass is a general challenge in shot-gun proteomics, because precursor ion selection is low resolution and these peptides are often cofragmented. This is illustrated in Figure 1, where the isolation window spans two peptide precursors, giving rise to a mixed MS/MS spectrum. To quantify the effects of this interference, we introduce a parameter called the "precursor ion fraction" (PIF). The PIF is defined as the fraction of ion current in the isolation window that is due to the targeted precursor ion and therefore ranges between 0 and 1. The PIF is determined for each targeted precursor ion and each

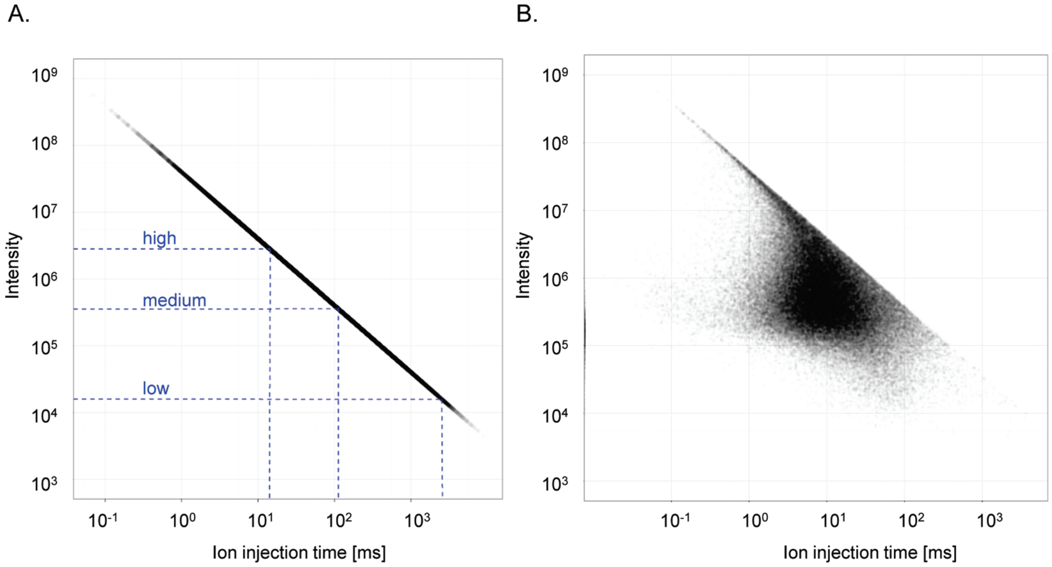


**Figure 4.** Calculated MS/MS ion injection time for the desired target value of 40 000 based on the apex intensity of the precursor ions. (A) Relationship between ion injection time and apex intensity assuming a PIF equal to one. The dashed lines indicate the high, medium and low abundance cases from Figure 2C. (B) Corrected injection times of the target precursor after considering the actual PIF.

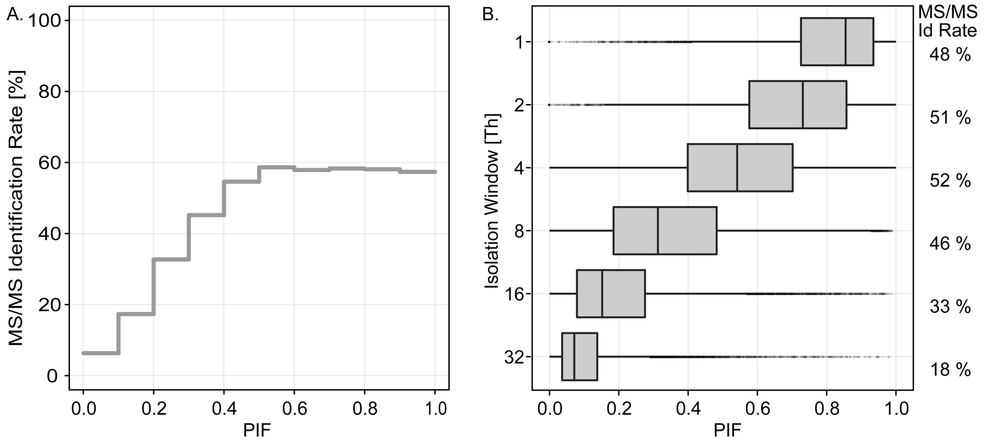


**Figure 5.** Dependence of peptide identification rates on (A) precursor ion fraction (PIF) and (B) isolation window as a box plot summary of runs with different isolation window widths.

tandem mass spectrum based on the closest full scan, which could be the previous or the consecutive one. For peptide features that are not targeted for fragmentation, the PIF is determined from the full scan closest to the peak apex. As a consequence, the PIF is *per se* unrelated to any MS/MS settings, such as the peak picking threshold for fragmentation. We first plotted the dependence of the MS/MS identification rate on the PIF (Figure 5A). At high PIF, the identification rate was nearly 60% and it stayed at this high level until a PIF value of 0.5. As expected the identification rate drops drastically at lower PIFs. This is in agreement with a recent study that determined the decrease in identification score after *in silico* merging of low resolution MS/MS spectra of different peptides.[14] Those authors defined the percentage of the intensity of the highest contaminating peak versus the intensity of the targeted peak (set at 100%) as "percent chimera intensity" (PCI), and noted a decrease in identification score at PCI values higher than 20%. In contrast, the PIF also takes into account any contributions from multiple peptide precursors or other coeluting species in the entire isolation window. Note that very few peptides (<150) with very low PIF (<0.2) were targeted in our experiment, and these may represent favorable cases for identification. Therefore the deleterious effect of the low PIF on identification success rate may be even larger. In any case, our data demonstrates that PIFs below 0.5 are generally detrimental for identification success in data driven shotgun proteomics, even for high accuracy MS/MS data.

The above experiment was performed with an isolation window of 4 Th. The extent of coelution of additional precursor ions is strongly influenced by the width of the isolation window of the targeted precursor ions and we therefore systematically varied the isolation window and determined the resulting distribution of PIFs. The median PIF of targeted peptides decreased from 0.85 at an isolation window of 1 Th to 0.07 at an isolation window of 32 Da (Figure 5B). At 2 Th or 4 Th, the median PIF was 0.73 and 0.54, respectively, both above the value where the identification rates start to decline. The best success might be expected with a particularly narrow isolation window. However, we observed that an isolation window of 4 Th resulted in the highest number MS/MS identifications (Figure 5B). This is related to the fact that narrow isolation windows in quadrupole devices generally limit transmission. Furthermore, the isolation window is not perfectly rectangular but rather has a somewhat rounded shape, limiting the retention of ions at the outer edges. Therefore, the typically used values of 2 or 4 Th are optimal in practice, despite the beneficial influence of narrow selection windows on the PIF.

It might appear obvious that reducing the sample complexity would help in increasing the PIF. Indeed, the analysis of HeLa cell lysate fractionated by one-dimensional gel electrophoresis into eight gel slices improved the median PIF, but this effect was only minor (Suppl. Figure 4A, Supporting Information). This is likely due to the very high complexity of the HeLa cell lysate itself, which allows more peptides to be detected as the dynamic range is increased by fractionation. It is possible that the median PIF would increase substantially upon very extensive fractionation. However, due to the measurement times involved, such strategies are not feasible for most proteomics experiments. We also checked, if the PIF would improve upon more extensive chromatography based separation of the peptide mixture. However, we found that the PIF decreases somewhat, when the length of the gradient was increased from 30 to 90 and then to 200 min (from 0.65 to 0.55 and to 0.45; Suppl. Figure 4B, Supporting Information). As the elution widths of the peaks increased, presumably those low abundant peptide features that were not detected in a shorter gradient became detectable by MaxQuant, because their elution profiles were sampled more often. These low abundance peaks tend to have a low PIF and therefore they lower the median PIF of the experiment. Together, these experiments show that neither sample fractionation nor longer gradients improve the median PIFs substantially. We speculate, however, that increased chromatographic resolution would contribute proportionally to improving the PIF of the already detected peaks, but would again lead to the detection of additional, low abundance peaks in the LC−MS map.

In Figure 6, we have plotted the identified peptides as a function of *m*/*z* scale and retention time in the LC−MS plane. The graph has the typical shape with smaller peptides, which are generally more hydrophilic, eluting first and larger peptides tending to elute later. Thus, only a specific part of the plane is occupied by eluting peptides. Color coding the peptides according to their PIF clearly shows that the center of the distribution has most peptides with low PIF (red). Peptides with very high PIF (green) tend to be outside the typical elution envelope. The part of the LC−MS/MS run that is richest in peptides that can be

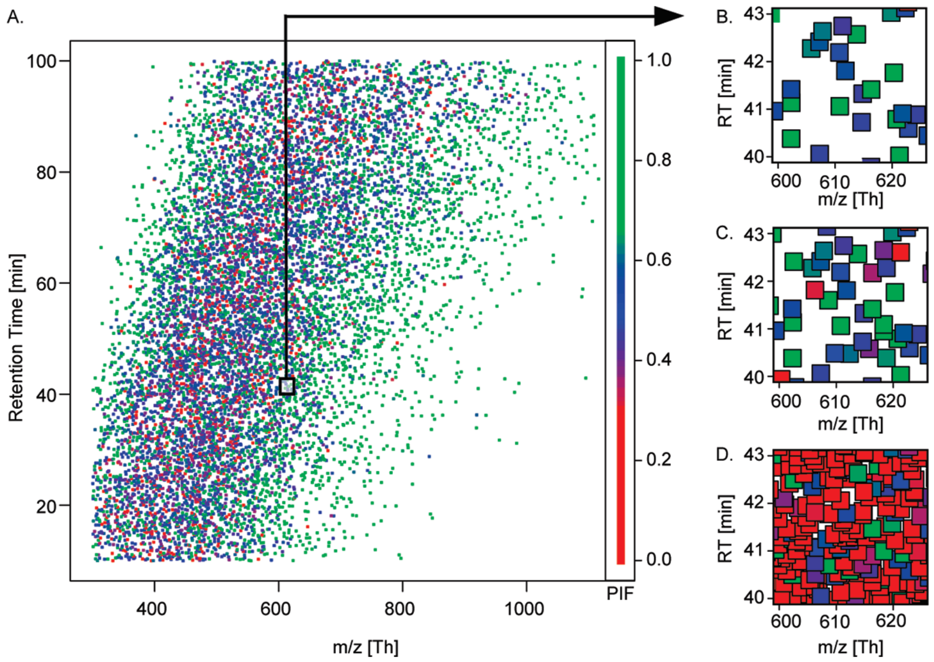


**Figure 6.** (A) LC−MS map of the complex peptide mixture colored by the PIF. Zoom of the indicated region for (B) all identified peptides, (C) all targeted peptides, and (D) all detected peptide precursor features.

fragmented and identified is also the one that tends to have the worst PIFs.

A zoom of a typical peptide-rich region is shown in Figure 6B. Each fragmented and identified peptide is overlaid with a potential MS/MS selection window assuming a median elution length of 40 s and an isolation window of 4 Th. Reflecting the median PIF of 0.54 for this experiment, there is substantial overlap of the potential selection windows. Figure 6C repeats the same plot but for all targeted peptides. The full extent of the PIF effect is revealed when plotting potential selection windows for all eluting peptide features in this region, whether targeted for fragmentation or not. Clearly the vast majority of peptide peaks are overlapping with one or more potential selection windows, demonstrating that cofragmentation would be the norm rather than the exception, if all peptide features were to be target in data dependent LC−MS/MS. This can also be confirmed on theoretical grounds: even if 100 000 peptide features were to elute equally spaced in the LC−MS map, there would only be room for 23 625 peptides with a PIF of 1, given a gradient length of 90 min and an $m/z$ range of 300−1000 Th for at least doubly charged precursors in a 4 Th isolation window.

As shown above for a particular example, considering lower abundant peptide features decreases the median PIF. To quantify this relationship for the entire data set, we plotted the PIF as a function of the maximum abundance of the isotope cluster as determined by MaxQuant (Figure 7A). Not surprisingly, there is a wide range of measured PIFs for each peptide intensity. However, the median PIF follows a clear trend, with PIF close to 1 occurring regularly for very high intensity peptide signals ($>10^8$ cps) whereas they are largely absent for very low abundance peptide signals ($<10^5$ cps). Strikingly, the median PIF of all peptide features detected is only 0.14 (Figure 7B). This means that for half of the peptide features, the ion current provided by the precursor of interest to the MS/MS spectrum would only be 14% or less, resulting in low identification success of these mixture MS/MS spectra.

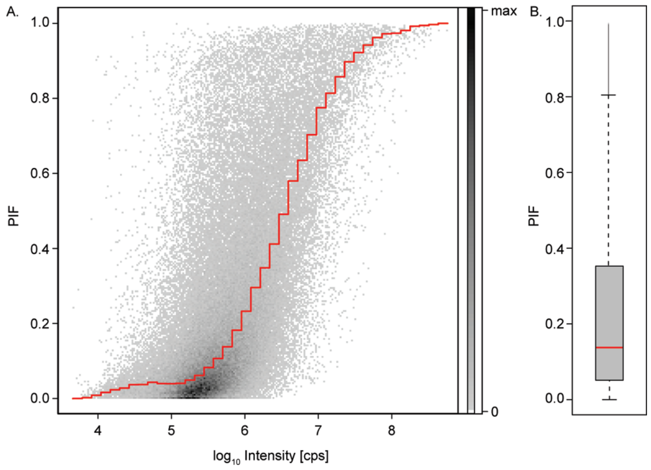


**Figure 7.** (A) PIF as a function of peptide feature intensity. The red line indicates the median signal intensity bins of all peptide features. (B) Box plot of (A). The median of all peptide features is 0.14 and is indicated by the red line.

## ■ CONCLUSION AND OUTLOOK

The number of identifiable peptides and their reproducibility between runs have been of great interest in shotgun proteomics. Here we have instead focused on the total number of detectable peptides, regardless of whether they were targeted for fragmentation or not. Using high resolution MS and the MaxQuant computational environment, we found that more than 100 000 isotope features, likely representing peptides, elute in a standard LC−MS/MS run. Thus, many more peptides can be detected at the LC level than are currently targeted and identified by data dependent LC−MS/MS. We systematically investigated the factors required to enable their identification. This analysis showed that a 10-fold higher sequencing speed would allow targeting all of the currently detectable peptide features.

Increased ion current by a factor of 10−100 would bring most of the peptides within the sensitivity range required for successful MS/MS. Most strikingly, we found that the vast majority of peptides have close neighbors in the LC−MS map and that in the median case only 14% of the peptide current in the isolation window is due to the precursor ion. Therefore cofragmentation of peptides would occur for almost all of the more than 100 000 peptide features, if they were targeted for fragmentation. The theoretically most effective approach for solving this problem would be high resolution precursor selection. However, this is not possible given existing technology. Even the reduction of the selection window to 2 Th results only in an increase of the PIF of 0.29 but at the expense of precursor intensity and identification success. A practical approach to deal with this effect is a demultiplexing algorithm (see for example ref 21) in conjunction with the appropriate overfilling of ions in the precursor isolation window. All our results are based on the dynamic range of state of the art instruments. The above trends will be intensified as the dynamic range and the resolution of mass spectrometers is enhanced and will further increase the number of detected precursor ions.

Our findings have interesting implications for the further development of shotgun proteomics. For example, an increase in sequencing speed is welcome and necessary, but not sufficient. Likewise, ongoing improvements in the sensitivity of instruments will yield benefits for the foreseeable future, but will not by themselves make all detectable features identifiable. Instead, our data suggest that a classical data driven MS and MS/MS strategy will eventually have limitations because virtually all MS/MS spectra will be mixture spectra. Therefore, not only advanced inclusion and exclusion features for peak selection, but also some form of multiplexing of MS/MS in a high resolution format will likely need to be a component of future shotgun proteomics strategies. Alternatively, if the peptide identity can be established in separate experiments, it may be transferable for many of the peptides that were not targeted for sequencing. However, this will need strict reference standards and sophisticated algorithms. Finally, we note that for many proteomic samples of low complexity, the depth of coverage achievable today is already sufficient and even for very complex proteomes, depending on the question asked, it may not be necessary to identify all detectable peptide features.

## ASSOCIATED CONTENT

### S Supporting Information

Supplementary figures and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*Matthias Mann, Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18 D-82152, Martinsried, Germany. E-mail: mmann@biochem.mpg.de. Fax: +49 89 8578 2219.

## ACKNOWLEDGMENT

We thank other members of our group for fruitful discussions and help. Drs. Boris Macek and Stevan Horning gave helpful advice. This work was partially supported by the European Union 7th Framework Program (HEALTH-F4-2008-201648/PROSPECTS).

## ABBREVIATIONS:

AGC, automatic gain control; cps, counts per second; FDR, false discovery rate; LC, liquid chromatography; MS, mass spectrometry; MS/MS, tandem mass spectrometry; PIF, precursor ion fraction; TIC, total ion current.

## REFERENCES

(1) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–82.

(2) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–7.

(3) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246* (4926), 64–71.

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

(5) Yates, J. R., 3rd; Gilchrist, A.; Howell, K. E.; Bergeron, J. J. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6* (9), 702–14.

(6) Choudhary, C.; Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell. Biol.* **2010**, *11* (6), 427–39.

(7) Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6* (7), 577–83.

(8) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **2006**, *78* (7), 2113–20.

(9) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–72.

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

(11) de Godoy, L. M.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008**, *455* (7217), 1251–4.

(12) 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. *Mol. Cell. Proteomics* **2009**, *8* (12), 2759–69.

(13) Wenger, C.; Phanstiel, D.; Coon, J., A Real-Time Data Acquisition Method for Improved Protein Quantitation on Hybrid Mass Spectrometers. Proceedings of the 58th ASMS Conference on Mass Spectrometry and Allied Topics May 23−27, 2010; Salt Lake City, UT.

(14) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152–60.

(15) Masselon, C.; Pasa-Tolic, L.; Lee, S. W.; Li, L.; Anderson, G. A.; Harkewicz, R.; Smith, R. D. Identification of tryptic peptides from large databases using multiplexed tandem mass spectrometry: simulations and experimental results. *Proteomics* **2003**, *3* (7), 1279–86.

(16) Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005**, *5* (16), 4096–106.

(17) Bern, M.; Finney, G.; Hoopmann, M. R.; Merrihew, G.; Toth, M. J.; MacCoss, M. J. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **2010**, *82* (3), 833–41.

(18) Wang, J.; Perez-Santiago, J.; Katz, J. E.; Mallick, P.; Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* **2010**, *9* (7), 1476–85.

(19) Rappsilber, J.; Ishihama, Y.; Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **2003**, *75* (3), 663–70.

(20) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709–12.

(21) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R.; Olsen, J. V.; Mann, M. Andromeda — a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**10.1021/pr101065j.

(22) Ihaka, R.; Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph Stat.* **1996**, *5* (3), 16.

(23) Wickham, H. *ggplot2: elegant graphics for data analysis*; Springer: New York, 2009.

(24) Nagaraj, N.; D'Souza, R. C.; Cox, J.; Olsen, J. V.; Mann, M. Feasibility of large scale phosphoproteomics with HCD fragmentation. *J. Proteome Res.* **2010**, *9* (12), 6786–94.

# 3 Novel Orbitrap Instrumentation with Improved Speed and Resolution

**Article 2** **Mass Spectrometry-Based Proteomics Using Q Exactive, a High-Performance Benchtop Quadrupole Orbitrap Mass Spectrometer**

Annette Michalski, Eugen Damoc, Jan-Peter Hauschild, Oliver Lange, Andreas Wieghaus,
Alexander Makarov, Nagarjuna Nagaraj, Jürgen Cox, Matthias Mann, Stevan Horning
*Mol Cell Proteomics*, 2011, 10(9): M111.011015.

**Article 3** **Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes**

Annette Michalski*, Eugen Damoc*, Oliver Lange, Eduard Denisov, Dirk Nolting,
Mathias Müller, Rosa Viner, Jae Schwartz, Philip Remes, Michael Belford, Jean-Jacques
Dunyach, Jürgen Cox, Stevan Horning, Matthias Mann, and Alexander Makarov
*Mol Cell Proteomics*, 2012, 11: O111.013698.

(*) equal contribution

## Prologue

Mass spectrometry is the analytical technique of choice for diverse applications including but not limited to proteomics and biological questions. Analyzing a wide variety of starting materials requires different technologies and specifications, but also a high degree of flexibility of the instruments.

The Orbitrap analyzer has proven to be a very versatile tool and it is well accepted by researchers from many different areas, due to its high performance in conjunction with relatively low maintenance efforts. Interestingly, the Orbitrap hardware itself did not undergo any changes since it was introduced to the market in 2005[53], but the associated instrument platforms were modified and improved in several iterations. The original format of a hybrid instrument with an up-front linear ion trap analyzer (LTQ Orbitrap), was extended by an extra collision cell attached to the C-trap, enabling higher energy collisional dissociation (HCD)[71] in the LTQ Orbitrap XL. Furthermore, optional ETD capability was added to the instrument[107]. The linear ion trap was replaced by a dual pressure double linear trap that provides more efficient trapping, fragmentation and faster scanning (LTQ Orbitrap Velos)[43]. An important device permitting greatly improved sensitivity was the stacked ring ion guide located behind the inlet, and a shorter version of the heated capillary that has a larger diameter in later instrument versions; together this increased the ion current by up to a factor 10. Beyond the hybrids, a very robust and simple instrument without mass selection device was developed especially for small molecule applications in a benchtop format (Exactive)[116]. Thus, Orbitrap instrumentation was improved considerably and the platforms became more and more powerful over several years - without touching the Orbitrap analyzer.

Proteomics was a major field of application for Orbitrap instrumentation from the beginning and a very demanding area as outlined in Article 1. In this thesis, we therefore co-developed and evaluated several hardware and software features of two novel Orbitrap instrument types prior to their release in 2011. In close collaboration with the research and development team from Thermo Scientific, Bremen, our efforts were primarily focused on shotgun proteomics measurements, quality control and intuitive usage, as well as the development of novel acquisition strategies. The projects were largely carried out on prototype instruments featuring only selected tools or preliminary versions of the final user-interface.

The quadrupole Orbitrap combination, Q Exactive, is particularly successful in shotgun proteomics applications because of its very high sequencing speed. One cycle in a top 10 method with resolution of 50,000 and 12,500 at m/z 400 for MS and MS/MS scans, respectively, takes only slightly above one second (Article 2). This is achieved by collecting ions in the C-trap or fragmenting them in the HCD cell simultaneously to the previous scan in the Orbitrap. Furthermore, an enhanced Fourier transformation algorithm doubles the resolution at a given transient length, which can also be used to increase sequencing speed. The shortest transient that can be selected is 64 ms (corresponding to the lowest resolution 12,500 at m/400). Theoretically this time can fully be used for accumulating ions without impacting the fastest cycle time given by this minimum scan time. In practice, the time span is slightly shorter due to some overhead: about 50 ms ensure entirely parallel operation. In shotgun proteomics experiments this functionality permits another mode of operation that applies a fixed ion injection time. Instead of selecting a specific target value that is controlled by predictive AGC, ions for the tandem mass spectrum are accumulated for a defined time period. That strategy makes optimal use of instrument capabilities and ensures very well defined, regular cycles. Finally, multiplexing scan modes at the MS and MS/MS level provide a variety of options to further improve sensitivity or sequencing speed of the instrument. In principle, it provides the Q Exactive with the capability to implement targeted proteomics approaches based on the ability of the quadrupole to rapidly switch between m/z ranges to be isolated. These strategies can be applied in conjunction with the data-dependent strategy.

In 2011, the latest generation hybrid instrument, the Orbitrap Elite, was introduced. This marked the first time that the dimensions of the Orbitrap analyzer were changed in a commercial instrument. Importantly, the compact Orbitrap is not only smaller than the previous version, but the ratio between the diameters of the outer and the central electrodes was modified. This provides a stronger electrostatic field and consequently about a factor two higher resolving power at the same transient length. Together with the enhanced Fourier transformation described above, resolution is improved by a factor four. The instrument also underwent improvements with regard to the ion optics for better robustness. Furthermore, a novel scan mode in the ion trap, *rapid CID*, was introduced. For shotgun applications, it is most beneficial to reduce the resolution for the sake of sequencing speed, unless very low abundant samples are analyzed. The ultra-high resolution can also be applied advantageously in top-down proteomics as demonstrated in Article 3.

⌘ Author's Choice

© 2011 by The American Society for Biochemistry and Molecular Biology, Inc.
**This paper is available on line at http://www.mcponline.org**

# Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer* Ⓢ

**Annette Michalski‡, Eugen Damoc§, Jan-Peter Hauschild§, Oliver Lange§, Andreas Wieghaus§, Alexander Makarov§, Nagarjuna Nagaraj‡, Juergen Cox‡, Matthias Mann‡¶, and Stevan Horning§¶**

**Mass spectrometry-based proteomics has greatly benefitted from enormous advances in high resolution instrumentation in recent years. In particular, the combination of a linear ion trap with the Orbitrap analyzer has proven to be a popular instrument configuration. Complementing this hybrid trap-trap instrument, as well as the standalone Orbitrap analyzer termed Exactive, we here present coupling of a quadrupole mass filter to an Orbitrap analyzer. This "Q Exactive" instrument features high ion currents because of an S-lens, and fast high-energy collision-induced dissociation peptide fragmentation because of parallel filling and detection modes. The image current from the detector is processed by an "enhanced Fourier Transformation" algorithm, doubling mass spectrometric resolution. Together with almost instantaneous isolation and fragmentation, the instrument achieves overall cycle times of 1 s for a top10 higher energy collisional dissociation method. More than 2500 proteins can be identified in standard 90-min gradients of tryptic digests of mammalian cell lysate— a significant improvement over previous Orbitrap mass spectrometers. Furthermore, the quadrupole Orbitrap analyzer combination enables multiplexed operation at the MS and tandem MS levels. This is demonstrated in a multiplexed single ion monitoring mode, in which the quadrupole rapidly switches among different narrow mass ranges that are analyzed in a single composite MS spectrum. Similarly, the quadrupole allows fragmentation of different precursor masses in rapid succession, followed by joint analysis of the higher energy collisional dissociation fragment ions in the Orbitrap analyzer. High performance in a robust benchtop format together with the ability to perform complex multiplexed scan modes make the Q Exactive an exciting new instrument for the proteomics and general analytical communities.** ***Molecular & Cellular Proteomics 10: 10.1074/mcp.M111.011015, 1–11, 2011.***

Mass spectrometry-based proteomics often involves the analysis of complex mixtures of proteins derived from cell or tissue lysates or from body fluids, posing tremendous analytical challenges (1–3). After proteolytic digestion, the resulting peptide mixtures are separated by liquid chromatography and online electrosprayed for mass spectrometric (MS) and tandem mass spectrometric (MS/MS) analysis. Because tens of thousands of peptides elute over a relatively short time and with ion signals different by many orders of magnitude (4, 5), mass spectrometers have been pushed to even higher sensitivity, sequencing speed, and resolution (6, 7). In current shotgun proteomics there are mainly four mass spectrometric separation principles: quadrupole mass filters, time of flight (TOF)[1] mass analyzers, linear ion traps, and Orbitrap™ analyzers. These are typically combined in hybrid configurations. Quadrupole TOF instruments use a quadrupole mass filter to either transmit the entire mass range produced by the ion source (for analysis of all ions in MS mode) or to transmit only a defined mass window around a precursor ion of choice (MS/MS mode). In the latter case ions are activated in a collision cell and resulting fragments are analyzed in the TOF part of the instrument with very high repetition rate. This TOF part of quadrupole TOF instruments replaces the final quadrupole section of triple quadrupole instruments, which are today mainly used for targeted proteomics (8–10).

The quadrupole TOF instruments achieve peptide separation "in space", meaning the ions are separated nearly instantaneously by passing through either the quadrupole section, in which only a chosen small mass range has stable trajectories, or by traversing the TOF section. In contrast, trapping instruments such as linear ion traps separate ions "in time" by applying external RF-DC fields to a stationary ion population that allow only a certain ion population to stably

From the ‡Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany; §Thermo Fisher Scientific (Bremen) GmbH, Hanna-Kunath-Strasse 11, 28199 Bremen, Germany

Received May 17, 2011, and in revised form, June 3, 2011
⌘ *Author's Choice*—Final version full access.
Published, MCP Papers in Press, June 3, 2011, DOI 10.1074/mcp.M111.011015

[1] The abbreviations used are: TOF, time-of-flight; AIF, all ion fragmentation; CID, collision induced dissociation; ETD, electron transfer dissociation; FDR, false discovery rate; FT, Fourier transform; HCD, higher energy collisional dissociation; HPLC, high performance liquid chromatography; LTQ, linear trap quadrupole; MS/MS, tandem mass spectrometry; pAGC, predictive automatic gain control; RF, radio frequency; SIM, selected ion monitoring.

remain in the trap (see ref (11)) for the concept of separation and fragmentation in time *versus* in space).

The Orbitrap mass analyzer was developed about ten years ago by Makarov. It consists of a small electrostatic device into which ion packets are injected at high energies to orbit around a central, spindle-shaped electrode (12–14). The image current of the axial motion of the ions is picked up by the detector and this signal is Fourier transformed (FT) to yield high resolution mass spectra. Commercially, the Orbitrap analyzer was first introduced in 2005 in a hybrid instrument (15). In proteomics and related fields, this combination of a low resolution linear ion trap with the high resolution Orbitrap analyzer—termed "LTQ Orbitrap"—has now become widespread (16, 17). The LTQ Orbitrap instruments represent a multistage trap combination (Fig. 1). In MS mode the linear trap performs the function of collecting the ion population, passing them on to an intermediate C-trap for injection and analysis in the Orbitrap analyzer at high resolution. In MS/MS mode the linear ion trap only retains a chosen mass window, which is activated by a supplemental RF field leading to fragmentation of the trapped precursor ions, and records the signal of a mass dependent scan at low resolution. Note that the high resolution MS scan can be performed at the same time as the low resolution MS/MS scans in the linear ion trap. Recently, an improved linear ion trap Orbitrap analyzer combination termed "LTQ Orbitrap Velos" has been introduced (18). It features an S-lens with up to 10-fold improved ion transmission from the atmosphere, a dual linear ion trap, and a more efficient Higher energy Collisional Dissociation (HCD) cell interfaced directly to the C-trap (18). HCD fragmentation is similar to the fragmentation in triple quadrupole or quadrupole TOF instruments and its products are analyzed with high mass accuracy in the Orbitrap analyzer (19). Thus, the LTQ Orbitrap or LTQ Orbitrap Velos instruments offer versatile fragmentation modes depending on the analytical problem (20–22).

Taking advantage of the small size of the Orbitrap analyzer a standalone benchtop instrument termed "Exactive" has been introduced mainly for small molecule applications. However, because of the absence of mass selection, its use in proteomics is limited to non-mass selective fragmentation of the entire mass range (called "All Ion Fragmentation" (AIF) on this instrument (23)).

The combination of a quadrupole mass filter with an Orbitrap analyzer has not yet been reported. We reasoned that such a quadrupole trap combination might offer unique and complementary advantages to the hybrid mass spectrometers described above. In particular, a quadrupole Exactive instrument or "Q Exactive" would be able to select ions virtually instantaneously because of the fast switching times of quadrupoles and it would be able to fragment peptides in HCD mode on a similarly fast time scale. Furthermore, because of the small size and mature technology used in current quadrupole mass filters, this analyzer combination should have a small footprint and be particularly robust. Finally, the ability to separate "in space" and analyze MS and MS/MS ranges at high resolution in the Orbitrap analyzer offers the promise of enabling efficient multiplexed

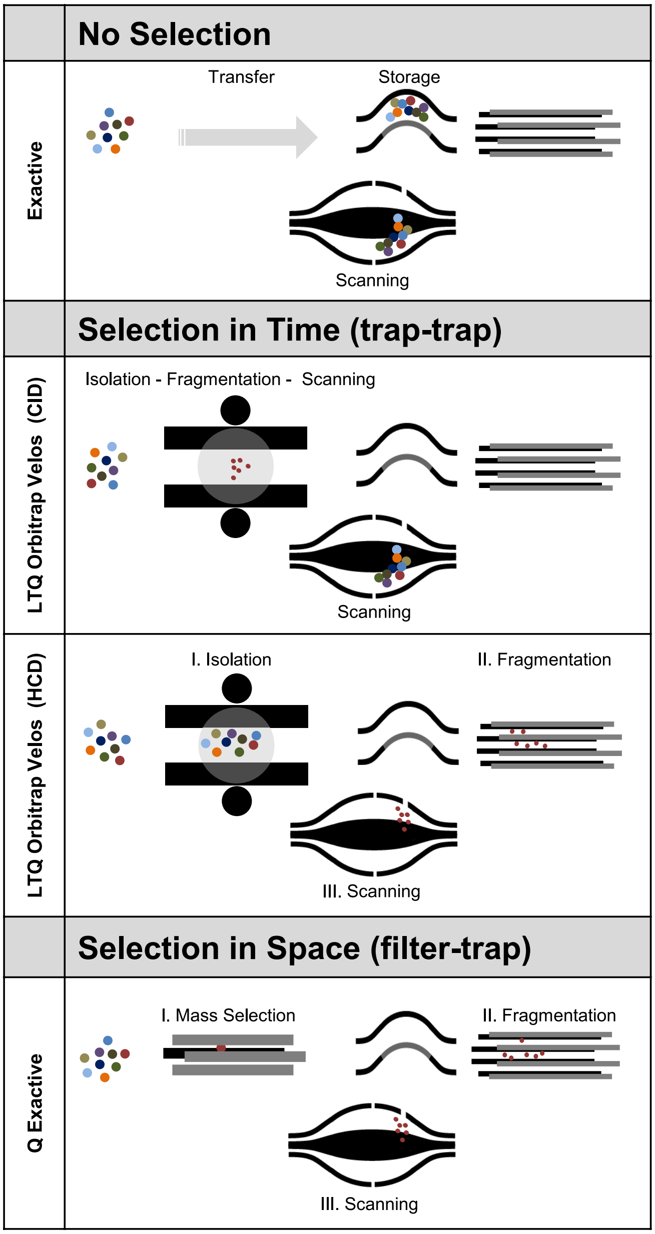


FIG. 1. **Mass spectrometers incorporating an Orbitrap analyzer.** The Exactive is a standalone instrument without mass selection. The total ion population is collected in the C-trap and injected into the Orbitrap analyzer (see text and Fig. 2 for details on detector components). In the LTQ Orbitrap Velos combination, ions can be selected "in time" by mass selective scans in the linear ion trap. In CID mode, the LTQ and Orbitrap operate as separate mass spectrometers. In HCD mode its function is to isolate a particular precursor, which is then fragmented in the HCD cell. In contrast, in the Q Exactive mass selection is "in space" as ions of only a specified *m/z* range have stable trajectories and are transferred to the storage or fragmentation devices before Orbitrap analysis.

scan modes not currently applied in proteomics research using trapping instruments.

## EXPERIMENTAL PROCEDURES

*Construction of a Quadrupole Orbitrap Instrument*—The Q Exactive instrument includes an atmospheric pressure ion source (API), a stacked-ring ion guide (S-lens) in the source region, a quadrupole mass filter, a C-trap, an HCD cell, and an Orbitrap mass analyzer as shown in Fig. 2. Ions are formed at atmospheric pressure (in this work in a nanoelectrospray ion source), pass through a transfer tube to an S-lens described in (18) and then via an injection multipole into a bent flatapole. The bent flatapole has 2-mm gaps between its rods, oriented in such a way that the line of sight from the S-lens is open for clusters and droplets to fly unimpeded out of the flatapole.

After collisional cooling in the bent flatapole, ions are transmitted via a lens into a hyperbolic quadrupole ($r_0 = 4$ mm), capable of isolating ions down to an isolation width of 0.4 Th at m/z 400. The quadrupole is followed by its exit lens combined with a split lens used to gate the incoming ion beam. A short octapole then brings ions into the C-trap interfaced to an HCD cell with axial field (18). The gas-filled HCD cell is separated from the C-trap only by a single diaphragm, allowing easy HCD tuning. Fragmentation of ions in the HCD cell is achieved by adjusting the offset of the RF rods and the axial field to provide the required collision energy. As long as this offset remains negative relative to the C-trap and the HCD exit lenses, all fragments remain trapped inside the HCD cell, even if the offset of the RF rods is varied. This allows to introduce multiple precursor ions and to fragment them at their optimum collision energy without compromising the storage of preceding injections. The summed ion population can then be transferred back into the C-trap, ejected into the Orbitrap analyzer and analyzed in a single Orbitrap detection cycle. This opens the possibility of fundamentally new, "multiplexing" modes of operation. In practice, the useful number of ion injections for a single Orbitrap detection is limited by the sum of the individual inject times being lower than the time for the Orbitrap scan.

A new challenge posed by interfacing the Orbitrap analyzer to a quadrupole is the automatic gain control (AGC) of weak ion signals. This problem was addressed by using an AGC pre-scan for a full MS spectrum with subsequent prediction of the ion currents for the weak signals on the basis of their share in total ion current (predictive AGC).

The mass range covered by the instrument is *m/z* 50–4000, with the range of mass selection reaching *m/z* 2500. Acquisition speed ranges from 12 Hz for resolving power 17,500 at *m/z* 200 (corresponding to 12,500 at *m/z* 400) to 1.5 Hz for resolving power 140,000 at *m/z* 200 (corresponding to 100,000 at *m/z* 400). Vacuum in the Orbitrap compartment is typically below $7 \times 10^{-10}$ mBar, which makes the analyzer adequate for high resolution analysis of most analytes, including large peptides and small proteins.

The ability to fill the HCD cell or the C-trap with ions while a previous Orbitrap detection cycle is still ongoing is another important innovation that allows to significantly reduce the influence of low ion currents on acquisition speed and quality of spectra.

*Processing of Transients Using Magnitude and Phase Information (eFT)*—Transients detected in the Orbitrap mass analyzer are processed using an enhanced version of Fourier Transformation (eFT™) for conversion of transients into frequency and then m/z. Details of the technique can be found in (24). Both eFT and conventional FT make use of complex numbers, which can be represented by magnitude and phase, or by real and imaginary components. As the initial phase of the ion package appears to be dependent on initial parameters of the ions in a very complex way (25), FT spectra have to be presented in the so-called magnitude mode, which amounts to disregarding the phase information. However, in Orbitrap mass spectrometers the built-in excitation-by-injection mechanism (26) provides an initial phase of ion oscillations that is almost m/z independent. This synchronization allows converting spectra in such a way that the real component of data can be utilized, which results in narrower peaks. In practice, eFT uses a combination of the magnitude and the real component of the signal to improve mass accuracy and peak shape.

Better accuracy of synchronization is achieved, if detection starts as early as possible after ion injection. For this reason, modifications of preamplifier and Orbitrap analyzer were introduced to reduce the delay between ion injection and start of transient detection from almost 10 ms to a fraction of a millisecond.

Practical implementation of eFT achieves between 1.8- and 2-fold increase of resolving power for the same transient (except for rapidly decaying signals, for example from proteins, where the gain is reduced to about 1.4-fold because of "hard sphere" collisions with background gas). The dual-spectrum online processing is computationally demanding but still fast enough to be completed in the LC MS time scale. Thus cycle time is still determined by transient acquisition and ion injection times and not by processing of the data. The eFT method is sensitive to precise synchronization of the instrument electronics and remaining shot-to-shot jitter, so that final mass accuracy is comparable to that of traditional magnitude mode FT spectra. Side-lobes in eFT spectra are comparable to those in conventional FT spectra.

*Preparation of HeLa Lysates*—HeLa cells were lysed and the pellet was dissolved in a urea (6 M) and thiourea (2 M) solution. Proteins were reduced with dithiotreitol (1 mM) for 30 min at room temperature followed by alkylation with iodoacetamide (55 mM) for 20 min in the dark. The mixture was incubated with LysC (1 $\mu$g/50 $\mu$g protein) (Wako, Richmond, VA) at room temperature for 3 h before 1:4 dilution with water. Incubation with trypsin (1 $\mu$g/50 $\mu$g protein) (Promega, Madison, WI) was carried out for 12 h at room temperature. The digestion was stopped by addition of formic acid (3%). Organic solvent was removed in a SpeedVac concentrator. The peptide mixture was desalted on reversed phase $C_{18}$ StageTips (27). Directly before analysis, peptides were eluted into 8 well autosampler vials with 60 $\mu$l buffer B (80% acetonitrile in 0.5% acetic acid). Organic solvent was removed in a SpeedVac concentrator and the final sample volume was adjusted with buffer $A^*$ (2% acetonitrile in 0.1% trifluoroacetic acid) to 12 $\mu$l.

*LC MS/MS Analysis for Q Exactive and LTQ Orbitrap Velos*—A nanoflow HPLC instrument (Easy nLC, Proxeon Biosystems, now Thermo Fisher Scientific) was coupled on-line to a Q Exactive or an LTQ Orbitrap Velos mass spectrometer (both from Thermo Fisher Scientific) with a nanoelectrospray ion source (Proxeon). Chromatography columns were packed in-house with ReproSil-Pur $C_{18}$-AQ 3 $\mu$m resin (Dr. Maisch GmbH) in buffer A (0.5% acetic acid). The peptide mixture (5 $\mu$g) was loaded onto a $C_{18}$-reversed phase column (15 cm long, 75 $\mu$m inner diameter) and separated with a linear gradient of 5–60% buffer B (80% acetonitrile and 0.5% acetic acid) at a flow rate of 250 nL/min controlled by IntelliFlow technology over 90 min. Because of loading and washing steps, the total time for an LC MS/MS run was about 40–50 min longer.

MS data was acquired using a data-dependent top10 method dynamically choosing the most abundant precursor ions from the survey scan (300–1650 Th) for HCD fragmentation. Target values on Q Exactive were similar to those typically used on an LTQ Orbitrap Velos. Determination of the target value is based on predictive Automatic Gain Control (pAGC) in both instruments. However, the LTQ Orbitrap Velos is equipped with electron multipliers, which allows scaling of the number of ions in a direct manner. In contrast, scaling of the number of ions is more indirect on the Q Exactive accounting for the difference in target values for the same S/N. Dynamic exclusion duration was 60 s with early expiration disabled on the LTQ Orbitrap Velos. Isolation of precursors was performed with a 4-Th window and MS/MS scans were acquired with a starting mass of 100 Th. Survey scans were acquired at a resolution of 70,000 at *m/z* 200 on the Q Exactive and 30,000 at *m/z* 400 on the LTQ Orbitrap Velos (see Results and Discussion and Table I for conversion of resolution values to different m/z values). Resolution for HCD spectra was set to 17,500 at *m/z* 200 on the Q Exactive and 7500 at *m/z* 400 on the LTQ Orbitrap Velos. Normalized collision energy was 30 eV for the Q Exactive and 35 eV for the LTQ Orbitrap Velos—they are not identical because of different scaling functions in the instrument software. The underfill ratio, which specifies the minimum percentage of the target value likely to be reached at maximum fill time, was defined as 0.1% on the Q Exactive. For the LTQ Orbitrap Velos the lower threshold for targeting a precursor ion in the MS scans was 5,000 counts. Both instruments were run with peptide recognition mode enabled, but exclusion of singly charged and unassigned precursor ions was only enabled on the LTQ Orbitrap Velos. This was because of the higher sequencing speed of the Q Exactive and a slightly different precursor selection algorithm for the data-dependent scans. However, in practice there was not much difference between the settings with regard to the number of identified unique peptides and proteins.

To demonstrate multiplexing of selected ion monitoring (SIM) scans, a method alternating full scans and SIM scans over the entire gradient was set up on the Q Exactive. The 92 min range in which peptides eluted was divided into 23 segments of 4 min duration. For each of these segments, three SIM windows of 2 Th width were defined, centered around 69 randomly chosen, low abundance precursor ions observed in these elution time windows in a previous top10 run. Pre-selection of these low abundance peptides was carried out manually based on the msms.txt file resulting from MaxQuant analysis. The method for multiplexed SIM scans was specified using the "Targeted SIM" template in the Q Exactive method editor. Resolution was set to 140,000 at *m/z* 200 and a target value of 1e6 ions for both scan types was chosen. The maximum ion injection time was set to 10 ms for the full scan and to 100 ms for each of the multiplexed SIMs. The inclusion list was saved in the global list features and in the data-dependent settings page "inclusion" was set to "on." Multiplexing of MS/MS spectra was done in exactly the same format as the standard top10 method, except that "msx" in the method setup of the data-dependent scans was set to 2 for multiplexing the fragment ions of two consecutively selected precursors.

*Analysis of Proteomic Data*—The mass spectrometric raw data from top10 methods were analyzed with the MaxQuant software (developmental version 1.1.1.32) (28). The false discovery rate (FDR) was set to 0.01 for proteins and peptides, which had to have a minimum length of 6 amino acids. MaxQuant was used to score peptides for identification based on a search with an initial allowed mass deviation of the precursor ion of up to 7 ppm. The allowed fragment mass deviation was 20 ppm. Search of the MS/MS spectra against the International Protein Index human data base (version 3.68, 87,061 entries) combined with 262 common contaminants was performed using the Andromeda search engine (29). Enzyme specificity was set as C-terminal to Arg and Lys, also allowing cleavage at proline bonds and a maximum of two missed cleavages. Carbamidomethylation of cysteine was set as fixed modification and N-terminal protein acetylation and methionine oxidation as variable modifications. MaxQuant applied time-dependent recalibration to the precursor masses for improved mass accuracy. Further analysis of the data provided by MaxQuant was performed in the R scripting and statistical environment (30). The data sets used for analysis are deposited at Tranche (www.proteomecommons.org).

## RESULTS AND DISCUSSION

Our goal was to construct a high performance quadrupole Orbitrap mass spectrometer in a compact format. Details of the hardware are in Experimental Procedures but here we give a brief overview. We started by building on the Exactive platform. The Exactive does not have mass selection capability and was developed mainly for small molecule applications (31). However, it can be equipped with a higher energy collisional dissociation cell (HCD) at the far side of the C-trap. Thus the detection system of the Exactive already allows HCD fragmentation (19) albeit without mass selection. This mode is called "AIF" for All Ion Fragmentation on this instrument and can also be used in proteomics (23). To support mass selective MS/MS scans in the Q Exactive, the transmission from electrospray source to vacuum was increased up to 10-fold, for which we used the S-lens employed in the LTQ Orbitrap Velos (18). New, rapidly switching electronics systems controlling the instrument were incorporated. Apart from some inlet ion optics changes the Orbitrap analyzer is the same as in previous Orbitrap analyzers. The Orbitrap voltage is 5 kV as it is on the Exactive and therefore higher than the 3.5 kV on LTQ Orbitrap instruments. The Q Exactive also employs a 90° bent ion path from the source toward the mass analyzer in common with the Exactive and in contrast to the LTQ Orbitrap instruments. The defining difference of the Q Exactive compared with the Exactive is the presence of a mass selective quadrupole analyzer between the ion source and the C-trap (Fig. 2). This quadrupole is the same as that used in triple quadrupole Access instruments, however, it features a modified RF-generator capable of driving selection of wide mass selection windows.

From a practical point of view, maintenance of the Q Exactive is similar to that of the Exactive. The quadrupole mass filter has very few tunable parameters and the instrument is automatically calibrated in a few minutes.

*Mass Spectrometric Resolution*—In the analysis of complex mixtures, peptides of similar mass often co-elute and therefore resolution is a key parameter of a mass spectrometer in these applications (7). Shotgun proteomics on the LTQ Orbitrap instruments is usually performed with 30,000 or 60,000 resolution at *m/z* 400. (Note that resolution decreases with the square root of the *m/z* value in Orbitrap analyzers.) High intrinsic resolution of an instrument allows short transients and hence short cycle times in topN methods—facilitating deep coverage of the proteome.

Because of the higher voltage of the Q Exactive, resolution at the same transient length is 20% higher. More importantly, we here employ eFT of the transients, which boosts resolution by a factor 1.8 to 2.0 (for further explanation see Experimental Methods). A similar principle has recently been described by Marshall and coworkers for FT ICR (32).

In Fig. 3, the resolution of the Q Exactive is demonstrated for the tetra peptide MRFA. As apparent from the widths of the isotope peaks and by their spacing, a resolution of more than 90,000 was achieved at *m/z* 524. The figure also illustrates the effect of turning the eFT algorithm on and off (although in normal operation of the Q Exactive, eFT is always

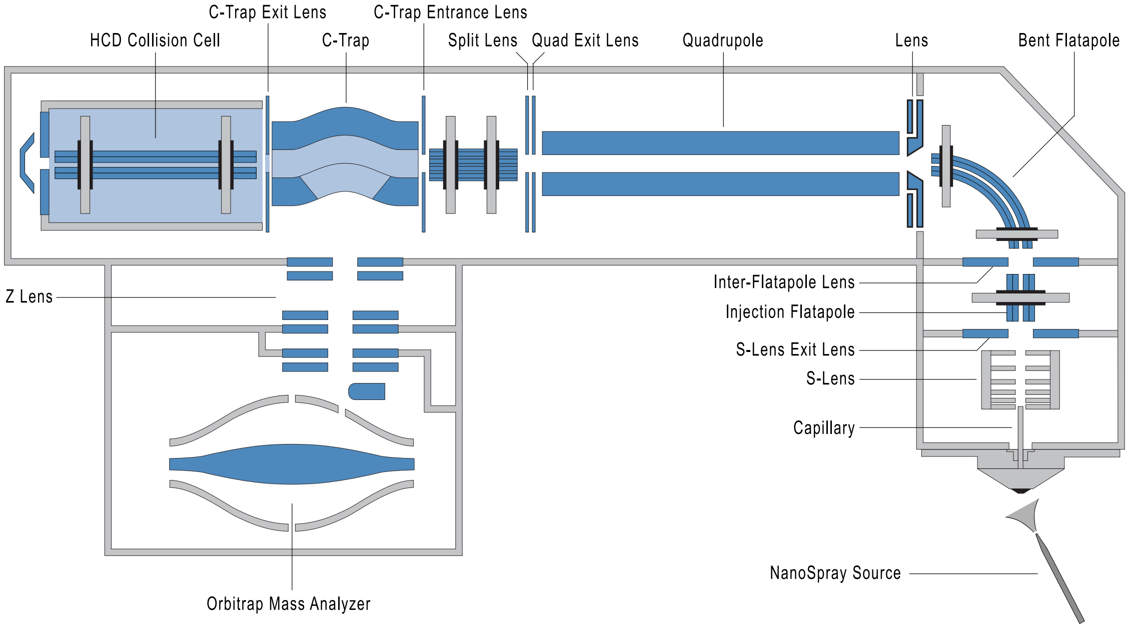


FIG. 2. **Construction details of the Q Exactive.** This instrument is based on the Exactive platform but incorporates an S-lens, a mass selective quadrupole, and an HCD collision cell directly interfaced to the C-trap. Note that the drawing is not to scale.

on). With eFT enabled, the instrument clearly resolves the two isotopes of the same nominal mass that are because of two $^{13}$C carbon atoms or the sulfur atom contained in methionine ($^{13}C_2$ *versus* $^{34}$S).

In the Exactive instrument, resolution is specified at *m/z* 200 because of its small molecule applications and this convention is kept in the Q Exactive. We provide a table to aid comparison between resolution values at *m/z* 200 and *m/z* 400 used with the LTQ Orbitrap instruments for the four possible transient lengths on the Q Exactive (Table I). The standard 60,000 resolution scan on the LTQ Orbitrap instruments uses a 768 ms transient. On the Q Exactive, a resolution of 100,000 (at *m/z* 400) is reached with a 512 ms transient. We use a resolution of 50,000 (at *m/z* 400), corresponding to a 256 ms transient length as a standard in proteomics experiments. This resolution is only slightly lower than that normally used on LTQ Orbitrap instruments, but takes less than half of the time. For MS/MS experiments we employ a resolution of 17,500 at *m/z* 200 (12,500 at *m/z* 400), which is achieved with a transient length of 64 ms. This value is substantially higher than the 7,500 resolution (at *m/z* 400) typical for HCD experiments on the LTQ Orbitrap Velos. The higher resolution in MS/MS spectra helps in assigning fragments of large precursors, however, the 64 ms transient was mainly chosen because even shorter transients would decrease the signal to noise in the MS/MS spectra.

*Cycle Times for MS and MS/MS Analysis*—As the quadrupole only serves as a selection device, the Q Exactive cannot perform MS and MS/MS operations in parallel (Fig. 1). On the other hand, the Q Exactive—unlike the LTQ Orbitrap Velos—fills ions in parallel to Orbitrap transient acquisition. Therefore, we next tested its overall cycle times and compared them to other Orbitrap instruments.

For the analysis of complex peptide mixtures, topN experiments consisting of a survey scan followed by N MS/MS scans are typically performed. Depending on the complexity of the mixture, N is usually between three and 20, and top10 is a widely used standard method. As explained above, a transient of 256 ms results in a resolution of 50,000 (at *m/z* 400), which is appropriate for proteomic applications. We combined this survey scan with ten 64 ms MS/MS scans (resolution 12,500 at *m/z* 400). If accumulation of the ions to the desired target count happened entirely in parallel with transient detection and if there was no overhead, this method would take 896 ms. The actually measured time for this sequence was 1.06 s, indicating that all overhead times together amounted only to about 160 ms (Fig. 4). This figure even included the automatic gain control scan performed before each full scan (whereas fill times for MS/MS scans are determined by "predictive AGC").

Completion of a full top10 method in about 1 s is exceedingly fast and compares favorably with top10 HCD methods on a Velos instrument. For example, the number of HCD spectra in such a Velos based method in a recent study was 3.3 MS/MS/s over the entire gradient (4). Even compared with top10 CID methods on the Velos instrument, which have the advantage of parallel acquisition, the Q Exactive proved to be faster in our hands (18). The reasons for the very fast cycle times are fivefold: (1) eFT allows using short transient times (2) ion filling is done in parallel with detection (3) overhead times of electronics components have been minimized (4) precursor

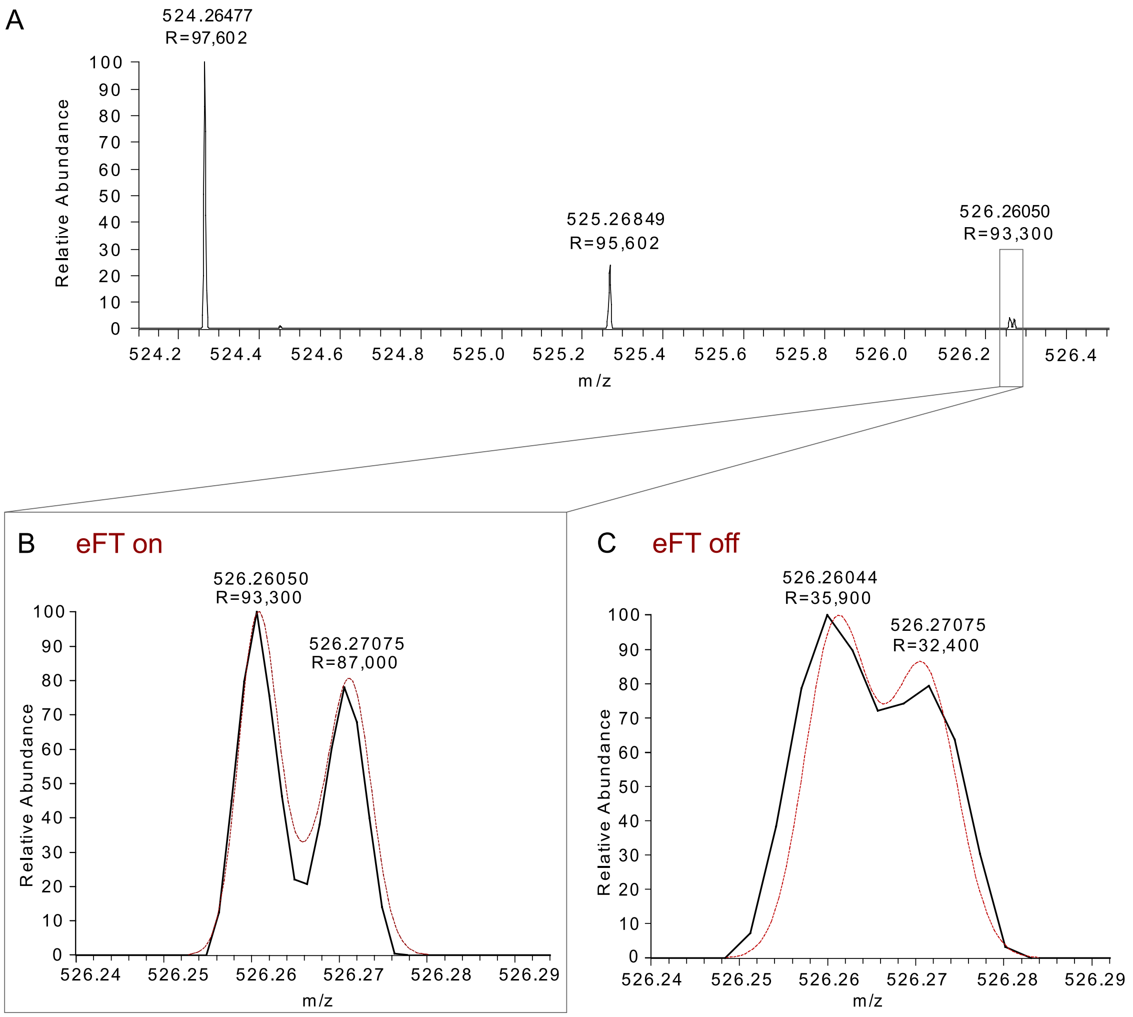


FIG. 3. **Resolution of the Q Exactive using eFT.** *A*, Isotope cluster of the MRFA peptide from a mass scan with a 512 ms transient employing eFT. *B*, Zoom into *A* demonstrating resolution of the $^{13}C_2$ isotope from the $^{34}S$ isotope. The *red* curve is the simulated signal for MRFA. *C*, The same isotopes as in *B* measured with the same transient but without enabling eFT.

TABLE I
*Four Q Exactive resolution settings and transient times*

| Resolution @ *m/z* = 200 Th | Resolution @ *m/z* = 400 Th | Transient length |
|---|---|---|
| 17,500 | 12,500 | 64 ms |
| 35,000 | 25,000 | 128 ms |
| 70,000 | 50,000 | 256 ms |
| 140,000 | 100,000 | 512 ms |

selection is done "in space" in a few ms and (5) HCD peptide fragmentation is nearly instantaneous.

Because of the parallel ion accumulation in the Q Exactive, fill times shorter than the transient length do not affect the overall cycle time. In our experiments with complex mixtures (see below), fill times for full scans were in the range of 1 to 10 ms, and for MS/MS scans they were generally between 5 and 50 ms leading to completely parallel acquisition and detection in almost all cases. The fill times observed here are similar to those of the LTQ Orbitrap Velos (18), indicating comparable sensitivity of both instrument types in full scan and HCD MS/MS mode.

*Q Exactive Performance for Proteome Analysis*—To characterize the performance of the Q Exactive for shotgun proteomics, we prepared a digest of a mammalian cell line (Experimental Methods). The peptide mixture was separated by on-line HPLC in a 90 min gradient by standard methods used in our laboratories. The entire analysis was done in triplicate and for comparison it was also performed on an LTQ Orbitrap Velos. Fig. 5*A* shows a heat map of the MS signals generated by peptides eluting from the column over the 90 min. The inset in the heat map is a zoom of a typical region (Fig. 5*B*), showing the complexity of eluting isotope patterns in this peptide mixture derived from whole cell lysate. As can be seen on the left hand scale, MS scans occurred every second and consequently eluting peptide peaks were well sampled. The MS spectrum in Fig. 5*C* depicts a single MS scan intersecting the zoomed region and indicates a triply charged precursor that

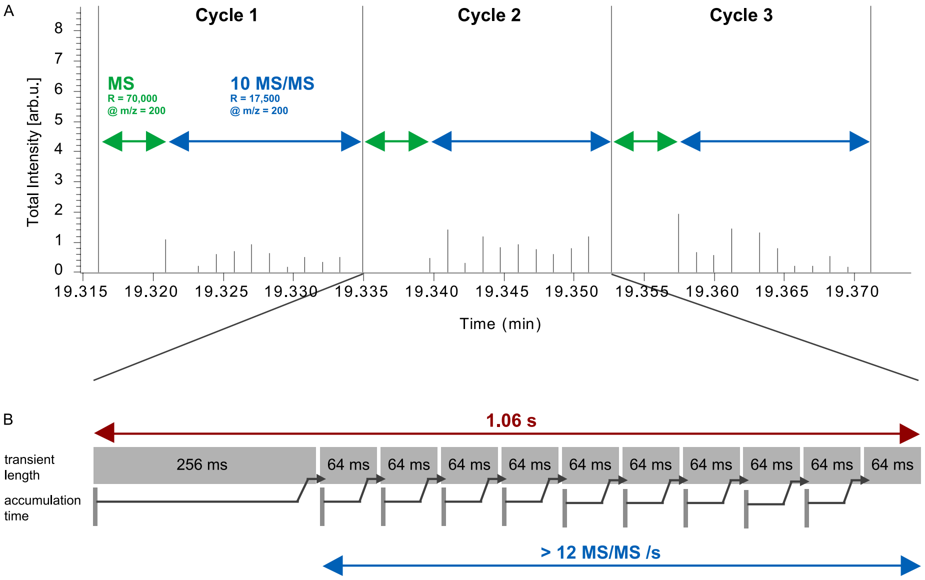


FIG. 4. **Cycle times for a top10 method on the Q Exactive.** *A*, Large ticks represent the total cycle consisting of MS and MS/MS scans. Duration for the MS survey scans is indicated by the green arrows (resolution 70,000 at *m/z* 200 or 50,000 at *m/z* 400) and for the MS/MS scans by the *blue* arrows (resolution 17,500 at *m/z* 200 or 12,500 at *m/z* 200). The *x* axis indicates chromatographic elution time and the *y* axis the total spectral intensity. *B*, Total cycle time for a top10 method is about 1 s and fragmentation frequency is more than 12 Hz. The lower trace indicates parallel ion accumulation for the following scan. Note that peptide ion accumulation times for typical LC column loads are generally shorter than transient times (as indicated in this example) and that they therefore do not add to cycle times.

was selected for fragmentation. Note that on the Q Exactive all fragmentation is performed by HCD and MS/MS spectra are always acquired with high resolution. This enables unambiguous recognition of charge states as illustrated in Fig. 5*D* and high fragment mass accuracy.

The data were analyzed in MaxQuant with the integrated Andromeda search engine (28, 29). Table IIA lists the results of the database search of the Q Exactive data. The total number of MS scans was in excess of 5,000 and the total number of MS/MS scans in excess of 35,000. (Note that top10 sequencing is only performed when there are sufficient peptide candidates in the MS scan that meet selection criteria for fragmentation.) The number of isotope patterns detected was close to 150,000, a very high number considering that the gradient was not particularly long (4, 5), presumably because of the short MS and MS/MS cycle time of 1 s. On average 12,563 unique peptides were identified in each run, for a total of 16,255 peptides in the triplicate analysis. These peptides mapped to an average of 2,557 proteins per run, and a total of 2,864 proteins of the HeLa proteome with the three 90 min gradients (supplemental Tables S1–S5).

For comparison, we performed the same analysis on an LTQ Orbitrap Velos. As expected, significantly more unique peptides were identified by the Q Exactive in the single LC runs (12,253 *versus* 10,207 on average; increase of 23%). Thus the Q Exactive, despite its compact format, represents an advance in the analysis of complex peptide mixtures as typically analyzed in shotgun proteomics. Note, however, that the above comparison only considers relatively short analysis of very complex mixtures and compares HCD fragmentation on both instruments. A detailed comparison of the two instrument types would require additional experiments and should also take into account that the Velos instrument can perform high or low resolution CID, in addition to HCD. Furthermore, the LTQ Orbitrap Velos, unlike the Q Exactive, is available with an ETD unit thereby providing a complementary fragmentation approach.

*Multiplexing at the MS and MS/MS Levels*—The linear ion trap Orbitrap analyzer combination is very versatile because it is comprised of two fully functional mass spectrometers. This allows isolation and fragmentation in different parts of the instruments and offers considerable flexibility in combining isolation and fragmentation events. Although many of these operation modes are not possible on the Q Exactive, it turns out that this novel combination of a mass filter and an Orbitrap analyzer also enables unique scan events. The principle feature making these scan modes possible is the fact that mass selection occurs "in space" which is extremely fast (Fig. 1). This should allow almost arbitrarily complex "mixing and matching" of MS and MS/MS mass ranges followed by high resolution analysis in the Orbitrap analyzer.

Fig. 6 illustrates two such multiplexed scan modes, one at the MS level and one at the MS/MS level. In selected ion monitoring (SIM) scans, a narrow mass range is accumulated providing increased signal to noise for particular ions of interest. SIM scans are useful in many applications but they are not often performed on Orbitrap instruments. This is because (1) the isolation of the SIM mass range in the linear ion trap is relatively time consuming, (2) there is a space charge limit on the number of ions that can be cleanly isolated, and (3) the analysis of even a single SIM scan takes considerable time. With few exceptions such as the lock mass injection to correct the mass scale (33), multiple mass range filling of the Orbitrap has not been implemented. The Q Exactive does not have the above limitations and, for example, allows selecting several SIM mass ranges of interest (Fig. 6*A*). In this mode, the C-trap is used as a storage device, which is filled with the desired number of ions from up to ten different SIM windows. These ions are together injected into the Orbitrap analyzer and measured in the same way as full mass ranges. Because fill

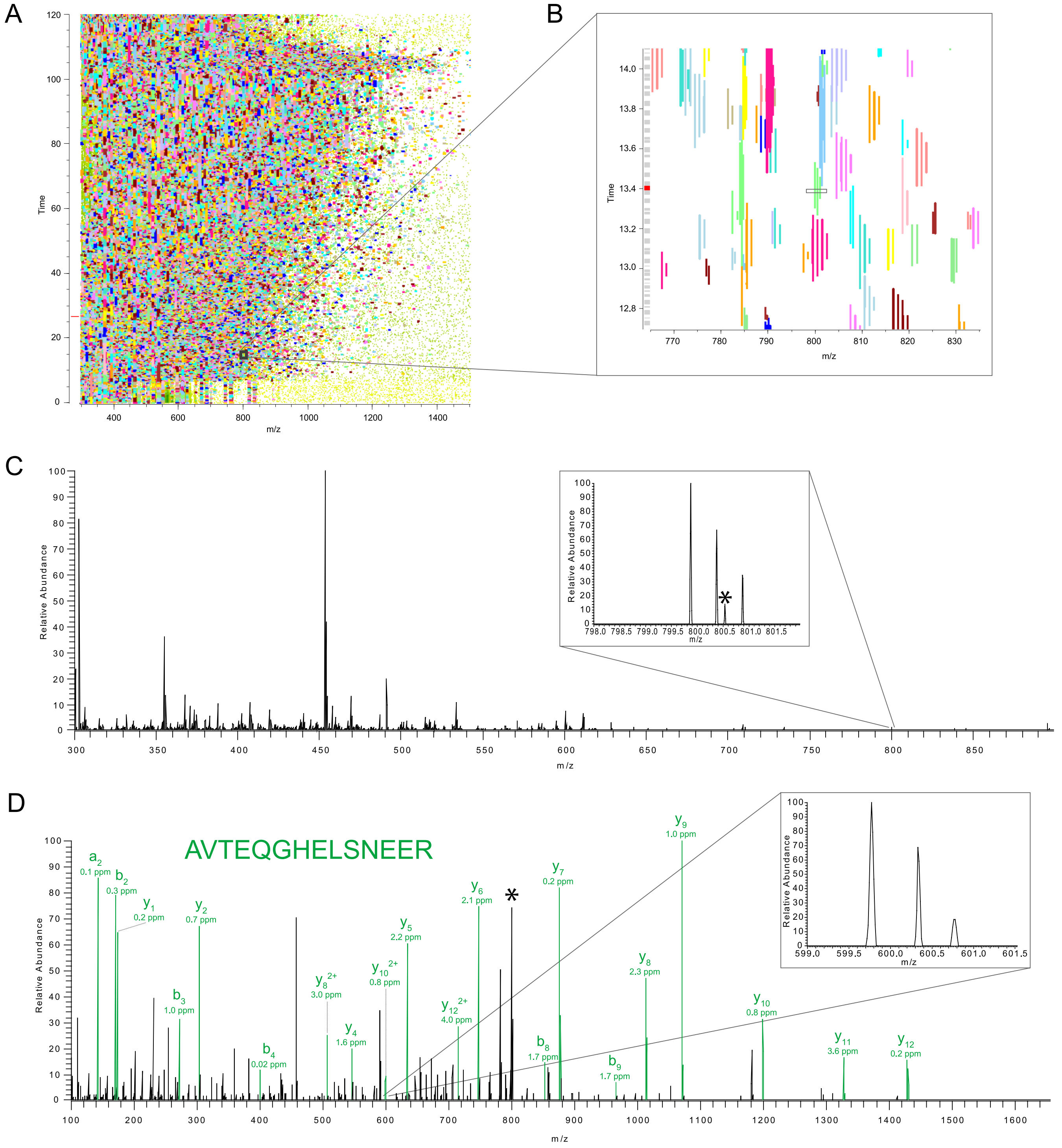


FIG. 5. **Proteome analysis with the Q Exactive.** *A*, Heat map of an LC MS/MS run of a peptide mixture resulting from proteolytic digestion of a HeLa lysate. *B*, Zoom of a typical part of the heat map. Marks on the left hand side represent the MS survey scans of each MS and MS/MS cycle and are separated by 1 s. *C*, Survey spectrum showing 50,000 resolution (at *m/z* 400) and the isotope pattern of a triply charged precursor in green; the asterisk indicates a co-eluting precursor ion. *D*, MS/MS spectrum of the precursor shown in *C* with 12,500 resolution (at *m/z* 400) and zoom of a doubly charged fragment ion.

times are typically much shorter than MS transient times, multiplexed SIM scans use Orbitrap instrument time much more efficiently. We demonstrate this concept in Fig. 6*B*–6*D* where a complete HeLa cell lysate was run in a 90 min gradient. (Note that such a gain does not occur on TOF instruments because they are not limited by scan times.)

TABLE II

*A, Peptide identification from HeLa lysate triplicate analysis on a Q Exactive (90 min gradient)*

| | MS spectra | MSMS spectra | Identifications [%] | Unique peptides | Proteins | Isotope clusters |
|---|---|---|---|---|---|---|
| HeLa (1) | 5427 | 35203 | 37.23 | 12298 | 2513 | 146138 |
| HeLa (2) | 5098 | 35911 | 38.35 | 12830 | 2601 | 143556 |
| HeLa (3) | 5274 | 35348 | 38.23 | 12560 | 2557 | 144336 |
| **Σ Triplicates** | | | **37.94** | **16255** | **2864** | |

*B, Peptide identification from HeLa lysate triplicate analysis on an LTQ Orbitrap Velos (90 min gradient)*

| | MS spectra | MSMS spectra | Identifications [%] | Unique peptides | Proteins | Isotope clusters |
|---|---|---|---|---|---|---|
| HeLa (1) | 2012 | 19818 | 55.64 | 10420 | 1895 | 125738 |
| HeLa (2) | 2102 | 19103 | 56.64 | 9855 | 1843 | 120553 |
| HeLa (3) | 2005 | 19634 | 56.37 | 10347 | 1906 | 126717 |
| **Σ Triplicates** | | | **56.21** | **14401** | **2242** | |

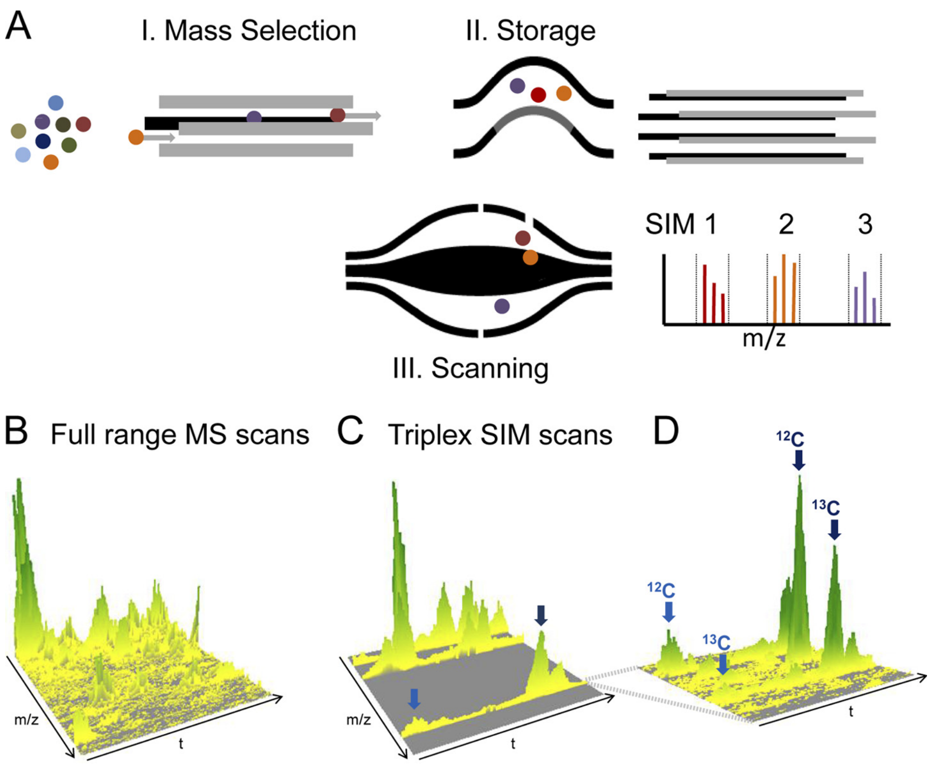


FIG. 6. **Multiplexing at the MS level.** *A*, The quadrupole mass filter is set to transmit a specific SIM mass range. After accumulation of the desired number of ions, the quadrupole is rapidly switched to the next SIM window up to the total number of SIM windows to be monitored. The combined SIM ranges are analyzed together in one high-resolution scan as shown in cartoon form in the inset. *B*, Three-dimensional representation of a 4-min segment of full range MS scans from a 90 min LC run of HeLa peptides. *C*, Visualization of the data from triplex scans acquired directly after each full scan in the same segment. Signals of the targeted low abundant peptides are clearly visible in the three SIM scans in *C* but virtually absent in the full scans in *B*. *D*, Zoom of one of the SIM ranges in *C* show that the SIM range contains peptides of even lower abundance than the targeted one marked with the dark blue arrow.

Three peptides known to be of low abundance from a previous top10 run were selected to define mass ranges (LTGMAFRVPTANVSVVDLTCR, DMIILPEMVGSMVGVYNGK, DAATIMQPYFTSNGLVTK). Fig. 6*B* represents the full range MS scans. In Fig. 6*C* and 6*D*, depicting a zoom into one of the SIM windows, the peptides are clearly visible with very good signal to noise. The ion injection time of most full scan was less than 1 ms, whereas ions for each of the SIM scans were accumulated for 100 ms accounting for the drastically improved signal-to-noise. At this high sensitivity, other peptides emerged from the background but were clearly resolved from the targeted peptide. These three multiplexed SIM windows were analyzed together in 140,000 resolution scans (0.5 s), adding little to the overall cycle time. Switching time to position the quadrupole at each mass window was 6 ms. Clearly such multiplexed SIMs could play an important role in targeted peptide analysis and peptide quantification.

In complex mixture analysis, sequencing speed can be a limiting factor. In principle, fragmentation of several precursors with simultaneous recording of the fragments can further boost the number of analyzable MS/MS events per unit time. Although in principle possible with a linear ion trap, in practice the necessary multiple rounds of peptide isolation and fragmentation preclude such an option on the LTQ Orbitrap. The Q Exactive, however, can successively isolate different precursor ions and fragment them in the HCD cell each at an individual normalized collision energy suitable for its properties. As each population of precursor ions is only

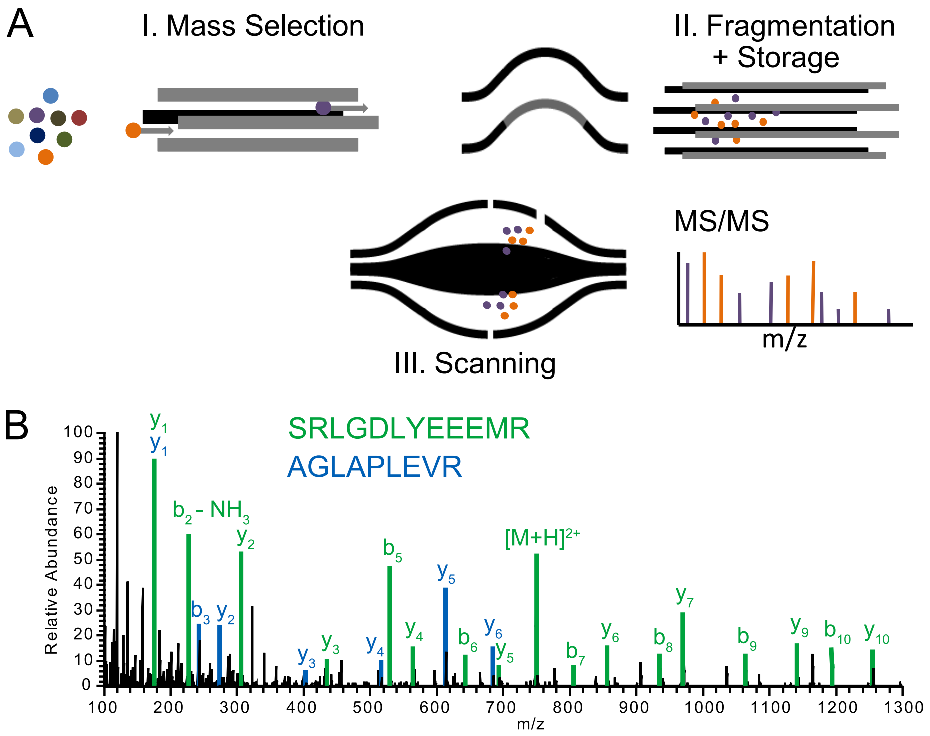


FIG. 7. **Multiplexing at the MS/MS level.** *A*, Different precursor ions are mass selected in the quadrupole, fragmented in turn by HCD and stored in the HCD cell. The combined fragment populations are measured together in the Orbitrap analyzer (depicted in cartoon form in the inset) *B*, Duplexed MS/MS spectrum.

fragmented once during injection into the HCD cell, all fragment ions are stored successively in the HCD cell and then they undergo joint analysis in the Orbitrap analyzer (see Fig. 7*A*). To demonstrate this, we specified a top10 method with multiplex degree of two. Fig. 7*B* shows a representative example of a multiplexed MS/MS spectrum analyzed at the normal MS/MS resolution setting of 12,500 (at *m/z* 400). Visual inspection of the spectrum clearly reveals extensive sequence information from both sequences. Interpretation of these deliberately multiplexed spectra, as opposed to cofragmented precursors, is aided by the fact that the same number of ions can be fragmented for each targeted precursor. Although it is clear that multiplexed MS/MS scans are easily possible on the Q Exactive platform, further investigations will be necessary to determine any resulting gain in peptide identifications. However, increased sequencing speed is not the only application of multiplexed MS/MS. As an example, the fragmentation of different charge states of multiply charged proteins could yield interesting structural information.

*Conclusions and Outlook*—We have described the construction and initial performance evaluation of a new type of mass spectrometer, the combination of a quadrupole mass filter with the Orbitrap analyzer. The quadrupole is one of the most robust and mature mass filters. Its combination with the relatively recently introduced Orbitrap mass spectrometer allowed realization of a high performance instrument with a small footprint and straightforward operation. These characteristics make the instrument an interesting addition to the proteomics toolbox, especially as proteomics is performed more and more by nonspecialist groups with biological or biomedical background.

Performance of the Q Exactive for complex peptide mixtures compares well with current LTQ Orbitrap instruments such as the LTQ Orbitrap Velos. Although the Q Exactive only offers the HCD fragmentation mode, we have shown here that HCD speed and sensitivity are not limiting. In fact, parallel filling of the ions combined with nearly instantaneous ion selection and fragmentation allowed implementation of a top10 method with 1-s cycle times. In comparison, the MS/MS scan rate of quadrupole TOF instrumentation could reach nominal speed up to 50 MS/MS per second, but, because of the lower transmission of TOF, signal to noise in each scan will be severely compromised unless high sample loads are used. An interesting novel feature of the Q Exactive is its ability to multiplex MS and MS/MS mass ranges, almost without limitations, which we have demonstrated here with two examples: multiplexed SIM mass ranges and multiplexed MS/MS spectra. In contrast to quadrupole TOF instrumentation, where scan speed in SIM mode is limited mainly by the time needed to reach acceptable signal to noise ratio, multiplexed SIM scans allow decoupling spectral acquisition speed from the speed of acquiring SIM scans and thus to utilize the full high transmission to the Orbitrap analyzer. We anticipate that the Q Exactive will enable additional interesting multiplexing capabilities in the future.

*Acknowledgments*—We thank our colleagues at Thermo Fisher Scientific, especially Ulf Froehlich, Andreas Kuehn, and Florian Grosse-Coosmann, and at the Max Planck Institute, especially Nadin Neuhauser, Tami Geiger, and Richard Scheltema, for help and fruitful discussions.

* This project was supported by the European Commission's 7th Framework Program PROteomics SPECificat ion in Time and Space (PROSPECTS, HEALTH-F4-2008-021,648).

S This article contains supplemental Tables S1 to S4.

¶ To whom correspondence should be addressed: M. M. at the Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. E-mail: mmann@biochem.mpg.de. or S. H. at Thermo Fisher Scientific, Hanna-Kunath-Straβe 11, D-28199 Bremen, Germany. E-mail: stevan.horning@thermofisher.com

Data availability: Supplementary data is available with this publication at the MCP web site. Raw MS files are uploaded to Tranche (www.proteomecommons.org) as "Michalski *et al.* Q Exactive" Hash code to access 8 RAW files: Nbh0v8NbSxDGuN/qunMhs Cz2z+rNP6YtKM1/uW2r2a2FEt9fUWESQH5XE1mGzU1BsPxp VWSlHtfTeeufl2hkW54eL54AAAAAAAAIDQ ==.

## REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Yates, J. R., 3rd, Gilchrist, A., Howell, K. E., and Bergeron, J. J. (2005) Proteomics of organelles and large cellular structures. *Nat. Rev.* **6,** 702–714
3. Walther, T. C., and Mann, M. (2010) Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* **190,** 491–500
4. Michalski, A., Cox, J., and Mann, M. (2011) More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793
5. Köcher, T., Swart, R., and Mechtler, K. (2011) 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
6. Domon, B., and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* **312,** 212–217
7. Mann, M., and Kelleher, N. L. (2008) Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 18132–18138
8. Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D. A., and White, F. M. (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 5860–5865
9. Addona, T. A., Abbatiello, S. E., Schilling, B., Skates, S. J., Mani, D. R., Bunk, D. M., Spiegelman, C. H., Zimmerman, L. J., Ham, A. J., Keshishian, H., Hall, S. C., Allen, S., Blackman, R. K., Borchers, C. H., Buck, C., Cardasis, H. L., Cusack, M. P., Dodder, N. G., Gibson, B. W., Held, J. M., Hiltke, T., Jackson, A., Johansen, E. B., Kinsinger, C. R., Li, J., Mesri, M., Neubert, T. A., Niles, R. K., Pulsipher, T. C., Ransohoff, D., Rodriguez, H., Rudnick, P. A., Smith, D., Tabb, D. L., Tegeler, T. J., Variyath, A. M., Vega-Montoto, L. J., Wahlander, A., Waldemarson, S., Wang, M., Whiteaker, J. R., Zhao, L., Anderson, N. L., Fisher, S. J., Liebler, D. C., Paulovich, A. G., Regnier, F. E., Tempst, P., and Carr, S. A. (2009) Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **27,** 633–641
10. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., and Aebersold, R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138,** 795–806
11. Louris, J. N., Cooks, R. G., Syka, J. E. P., Kelley, P. E., Stafford, G. C., and Todd, J. F. J. (1987) Instrumentation, applications, and energy deposition in quadrupole ion-trap tandem mass-spectrometry. *Anal. Chem.* **59,** 1677–1685
12. Hardman, M., and Makarov, A. A. (2003) Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal. Chem.* **75,** 1699–1705
13. Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72,** 1156–1162
14. Scigelova, M., and Makarov, A. (2006) Orbitrap mass analyzer–overview and applications in proteomics. *Proteomics* **6,** 16–21
15. Syka, J. E., Marto, J. A., Bai, D. L., Horning, S., Senko, M. W., Schwartz, J. C., Ueberheide, B., Garcia, B., Busby, S., Muratore, T., Shabanowitz, J., and Hunt, D. F. (2004) Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J. Proteome Res.* **3,** 621–626
16. Makarov, A., Denisov, E., Lange, O., and Horning, S. (2006) Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *J. Am. Soc. Mass Spectrom.* **17,** 977–982
17. Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., and Horning, S. (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **78,** 2113–2120
18. 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., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell Proteomics* **8,** 2759–2769
19. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–712
20. Macek, B., Waanders, L. F., Olsen, J. V., and Mann, M. (2006) Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Mol. Cell Proteomics* **5,** 949–958
21. McAlister, G. C., Berggren, W. T., Griep-Raming, J., Horning, S., Makarov, A., Phanstiel, D., Stafford, G., Swaney, D. L., Syka, J. E., Zabrouskov, V., and Coon, J. J. (2008) A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer. *J. Proteome Res.* **7,** 3127–3136
22. McAlister, G. C., Phanstiel, D. H., Westphall, M. S., and Coon, J. J. (2011) Higher-energy collision-activated dissociation without a dedicated collision cell. *Mol. Cellular Proteomics*, 10(5):O111.009456
23. Geiger, T., Cox, J., and Mann, M. (2010) Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell Proteomics* **9,** 2252–2261
24. Lange, O., Makarov, A., Denisov, E., and Balschun, W. (2010) Accelerating spectral acquisition rate of Orbitrap mass spectrometry. *Proc. 58th Conf. Amer. Soc. Mass Spectrom.*
25. Vining, B. A., Bossio, R. E., and Marshall, A. G. (1999) Phase correction for collision model analysis and enhanced resolving power of fourier transform ion cyclotron resonance mass spectra. *Anal. Chem.* **71,** 460–467
26. Makarov, A. (2009) In: March, R. E., and Todd, J. F. J., eds. *Practical Aspects of Trapped Ion Mass Spectrometry.* Vol 4*: Theory and Instrumentation*, CRC Press (Taylor & Francis)
27. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
28. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
29. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **11,** 1794–1805
30. Ihaka, R., and Gentleman, R. (1996) R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5,** 299–314
31. Bateman, K. P., Kellmann, M., Muenster, H., Papp, R., and Taylor, L. (2009) Quantitative-qualitative data acquisition using a benchtop Orbitrap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **20,** 1441–1450
32. Beu, S. C., Blakney, G. T., Quinn, J. P., Hendrickson, C. L., and Marshall, A. G. (2004) Broadband phase correction of FT-ICR mass spectra via simultaneous excitation and detection. *Anal. Chem.* **76,** 5756–5761
33. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell Proteomics* **4,** 2010–2021

⌘ Author's Choice

© 2012 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

# Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes*[S]

**Annette Michalski‡**, Eugen Damoc§**, Oliver Lange§, Eduard Denisov§, Dirk Nolting§, Mathias Müller§, Rosa Viner¶, Jae Schwartz¶, Philip Remes¶, Michael Belford¶, Jean-Jacques Dunyach¶, Juergen Cox‡, Stevan Horning§, Matthias Mann‡‖, and Alexander Makarov§**

**Although only a few years old, the combination of a linear ion trap with an Orbitrap analyzer has become one of the standard mass spectrometers to characterize proteins and proteomes. Here we describe a novel version of this instrument family, the Orbitrap Elite, which is improved in three main areas. The ion transfer optics has an ion path that blocks the line of sight to achieve more robust operation. The tandem MS acquisition speed of the dual cell linear ion trap now exceeds 12 Hz. Most importantly, the resolving power of the Orbitrap analyzer has been increased twofold for the same transient length by employing a compact, high-field Orbitrap analyzer that almost doubles the observed frequencies. An enhanced Fourier Transform algorithm—incorporating phase information—further doubles the resolving power to 240,000 at *m/z* 400 for a 768 ms transient. For top-down experiments, we combine a survey scan with a selected ion monitoring scan of the charge state of the protein to be fragmented and with several HCD microscans. Despite the 120,000 resolving power for SIM and HCD scans, the total cycle time is within several seconds and therefore suitable for liquid chromatography tandem MS. For bottom-up proteomics, we combined survey scans at 240,000 resolving power with data-dependent collision-induced dissociation of the 20 most abundant precursors in a total cycle time of 2.5 s—increasing protein identifications in complex mixtures by about 30%. The speed of the Orbitrap Elite furthermore allows scan modes in which complementary dissociation mechanisms are routinely obtained of all fragmented peptides.** ***Molecular & Cellular Proteomics 11: 10.1074/mcp.O111.013698, 1–11, 2012.***

In many mass spectrometric applications, the resolving power of the instrument is of pivotal importance. Ultimate resolution has so far been obtained by Fourier Transform Mass Spectrometry (1) and in a recent example, Marshall and co-workers detected more than 26,000 components in a single spectrum of a crude oil mixture (2). In ion cyclotron resonance (ICR)[1] Fourier transform mass spectrometry, resolution is determined by the length of the transient and by the strength of the magnetic field. Increasingly larger magnets have allowed resolution in excess of one million for small molecules. The relatively recently introduced Orbitrap™ analyzer utilizes a different physical principle to obtain high resolution (3–6). The signal is recorded from the image current produced by ion packets which oscillate around and along the spindle-shaped inner electrode of the trap: the higher the electric field, the larger the number of oscillations per unit time and the higher the resolving power. To increase field strength, several design options can be pursued, including increasing the radius of the inner electrode of the device (7). Here we describe an Orbitrap analyzer that achieves higher resolving power through reduced trap dimensions. Resolution is further increased by making use of the phase information during Fourier Transformation (8–11). This ultra high resolution Orbitrap analyzer was combined with other instrumental improvements to construct a novel linear ion trap Orbitrap hybrid mass spectrometer termed the Orbitrap Elite.

We describe principles of this instrument and characterize its operation for both intact protein analysis and for bottom up peptide mixture analysis. Top down protein analysis has previously mainly been performed with Fourier transform (FT) ICR instruments because of their very high resolving power (12–

From the ‡Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany; §Thermo Fisher Scientific (Bremen) GmbH, Hanna-Kunath-Strasse 11, 28199 Bremen, Germany; ¶Thermo Fisher Scientific, 355 River Oaks Parkway, San Jose, California 95134

Received August 19, 2011, and in revised form, December 1, 2011
⌘ *Author's Choice*—Final version full access.
Published, MCP Papers in Press, December 12, 2011, DOI 10.1074/mcp.O111.013698

[1] The abbreviations used are: AGC, automatic gain control; CID, collision induced dissociation; ETD, electron transfer dissociation; FDR, false discovery rate; FT, Fourier transform; HCD, higher energy collisional dissociation; HPLC, high performance liquid chromatography; ICR, ion cyclotron resonance; IPI, International Protein Index; MS/MS, tandem mass spectrometry; SIM, selected ion monitoring.

14). One of the challenges in using top down approaches in proteomics has been to obtain cycle times commensurate with liquid chromatography tandem MS (LC MS/MS) time scales (15). The linear ion trap Orbitrap has also been employed for top down proteomics (16–19). Here we take advantage of the ultra high resolution of the Orbitrap Elite to enable fast LC MS/MS compatible top-down scan methods.

In bottom-up proteomics typically very complex peptide mixtures are analyzed (20–22). Online LC MS runs contain evidence for tens of thousands of peptides (23, 24) and this places a premium on the resolution of the survey (MS) scans. A popular shotgun proteomics method on the linear ion trap Orbitrap (LTQ Orbitrap or LTQ Orbitrap Velos) is a "1 s" survey scan with 60,000 resolution at *m/z* 400 (768 ms transient), and ion trap collision-induced dissociation (CID) scans of the ten or twenty most abundant ions ("high resolution" "low resolution" or "high–low" top10 method). Here we explore topN methods with much higher resolution survey scans as well as an increased number of fragmentation events per cycle enabled by "rapid CID" scans. A "high–high" strategy (high resolution MS as well as MS/MS (25)) has been routinely made possible on Orbitrap instruments by higher energy collisional dissociation (HCD) with the advent of the LTQ-Orbitrap Velos (26). We show that this strategy benefits from the shorter transients and higher resolving power possible on the Orbitrap Elite.

It has been demonstrated that a combination of two fragmentation methods can greatly augment sequence related information in peptide MS/MS (27–29) and we explore this dual approach with CID and HCD fragmentation of the same precursor ions.

## EXPERIMENTAL PROCEDURES

The Orbitrap Elite is a further development of the LTQ Orbitrap Velos (26). This hybrid instrument combines a Velos PRO dual cell differential pressure linear ion trap mass spectrometer with a high field Orbitrap mass analyzer (Fig. 1*A*). The Velos PRO builds on the LTQ Velos (30) and its extensions include (1) a new generation of ion optics consisting of a 45° rotated bent quadrupole Q0, a neutral beam blocker, and an octopole ion transfer device, (2) faster ion trap mass analysis scan speed of 66,000 amu/s, and (3) a higher dynamic range detection system for improved quantitation performance, and (4) the addition of beam-type collisional dissociation capabilities for the stand-alone ion trap system.

*New Generation Ion Optics*—The S-lens consists of a set of stainless steel apertures to which an RF voltage is applied, alternate lenses having opposite (180°) phase. This device is used in a high pressure regime (low millibar) to efficiently focus the ion beam emerging from a transfer tube through a final exit lens (31). Droplets and solvent clusters exiting the transfer tube are kept from passing into the downstream ion optics by a curved quadrupole ion guide. The ion guide has been rotated 45° with respect to the orientation in the LTQ Velos so that noncharged droplets and solvent clusters can pass through the gap between the quadrupole rods, rather than impinge on the rod surface itself. This significantly reduces the potential for contamination of the quadrupole ion guide. The new geometry allows for a stainless steel rod to be positioned in the region of curvature of the quadrupole ion guide where it serves as a neutral beam blocker. The combination of the 45° rotated quadrupole ion guide and the neutral beam blocker reduces the rate of contamination and improves the longevity of the ion optics system. A short octopole, Q00 ($r_0$ 5.56 mm, rod diameter 2 mm, length 28.58 mm operated at 3 MHz, 800 Vpp), located between the exit lens and curved quadrupole, has also been added and replaces the quadrupole device in this region in the LTQ Velos. This octopole tends to be more robust to contamination and it is also used as a dissociation device in the stand-alone Velos PRO (32).

*Faster Scan Speed*—The dual cell differential pressure linear ion trap allows for substantially accelerated scan rates and higher resolution owing to the lower pressure in the mass analyzing cell of the ion trap (30). The normal scan rate on the LTQ Velos is 33,000 amu/s, which typically achieves peak widths of 0.34 amu at *m/z* 1822 and is sufficient to separate isotopes of triply charged ions. By optimizing operating conditions such as the resonance ejection amplitude and the phase relationship between the resonance ejection and trapping RF signals, a small sacrifice in resolution can give a large improvement in scan rate. In the Velos PRO, the scan rate has been doubled to 66,000 amu/s while still maintaining better than unit resolution, achieving an average peak width at half height of 0.47 amu at *m/z* 1822. With this rapid scan rate up to 12.5 MS/MS scans can be performed per second (rapid CID or rCID).

*Higher Dynamic Range Detection System*—Because of the faster scan rates the ion currents generated when performing mass analysis are also increased. Therefore, a higher dynamic range detection system is required. Discrete dynode electron multipliers have been developed (SGE Analytical Science Pty Ltd, Australia) that replace the continuous channel electron multipliers in the LTQ Velos. These new electron multipliers have linear outputs up to 160 $\mu$A yielding six orders of magnitude dynamic range. A 24 bit analog-to-digital converter is employed in the electrometer circuitry which matches the performance of the discrete dynode multipliers. The wider linear dynamic range of this detection system increases the precision and accuracy for doing quantitative analysis while also offering enhanced limits of detection.

*High-field Orbitrap Analyzer*—The Orbitrap mass analyzer generally consists of an outer barrel-like electrode of maximum radius R2 and a central spindle-like electrode along the axis of maximum radius R1, with the outer electrode maintained at the virtual ground of the preamplifier, while the central electrode is at a voltage -Ur ($Ur>0$ for positive ions) (3). In a standard Orbitrap analyzer, $R1 = 6$ mm and $R2 = 15$ mm (5), whereas the high-field analyzer described here is more compact with $R1 = 5$ mm and $R2 = 10$ mm (Fig. 1*C*), *i.e.* the outer electrode is scaled down by a factor of 1.5. Similarly to an analyzer described recently (7), a decrease of the R2/R1 ratio from 2.5 to 2 allows an increase in the frequency in addition to the above scaling factor, thus bringing the total gain of frequency to ~1.8-fold. This is accompanied by an increase of the injection ion energy of ~1.4-fold. Despite the increase in space charge density in the analyzer by a factor of $(1.5)^3 \approx 3.4$, the additional shielding provided by the relatively thicker central electrode keeps space-charge induced frequency shifts even slightly below those in the standard analyzer.

As the injection slot was scaled down by the same factor as the outer electrode to avoid compromising the quality of the field inside the analyzer, an additional focusing of the incoming ion beam became necessary. This was achieved by adding a miniature Einzel lens in the form of a 1-mm plate with a 2 mm ID orifice at a voltage of up to 1000 V, which was sandwiched between two similar plates at 0 V and separated by gaps of 1 mm. This assembly was mounted on the same block as the deflector at the entrance to the Orbitrap analyzer.

Scaling down of the outer electrodes appeared also to slightly reduce their capacitance to ground and to each other, which in turn allowed the use of lower-capacitance transistors in the image current preamplifier. The resulting sensitivity increase of about 30% resulted in the same signal-to-noise ratio for the same number of ions as in the standard analyzer even for twice shorter transients.

We found that reduction of the gap between the outer and the central electrodes requires an almost proportional improvement of machining accuracy of the electrodes. This was achieved by rigorous refinement of the existing manufacturing and measurement techniques.

*Transient Processing with eFT*—FT of a digitized transient is a fast processing method but it requires relatively long detection times to achieve high resolving powers. It is thus desirable to further increase the resolving power for a given acquisition time. We applied a newly developed enhanced version of the Fourier Transformation (eFT™), which is also employed in another novel instrument, the Q Exactive (11). Details of the technique can be found in ref (10). Briefly, both eFT and conventional FT make use of complex numbers, which can be represented by magnitude and phase. As the initial phase of the ion package typically depends on initial parameters of the ions in a very complicated way (8), FT spectra normally have to be presented in the so-called magnitude mode, which amounts to disregarding the phase information. However, in Orbitrap mass spectrometers the built-in excitation-by-injection mechanism (33) provides an initial phase of ion oscillations that is almost independent of *m/z*. This synchronization allows converting spectra in such a way that they correspond to zero initial phase for all *m/z* values (so-called absorption spectra) and exhibit narrower peaks. In practice, eFT uses a combination of magnitude and absorption spectra along with Hanning apodization, triple zero-filling, and additional filtering to improve mass accuracy and peak shape.

Better accuracy of spectra conversion is achieved if detection starts as early as possible after ion injection. Therefore, the following modifications of the preamplifier and Orbitrap analyzer were introduced: (1) High-speed diode bridges have replaced mechanical relays that were previously used for protection of the preamplifier during pulsing of the Orbitrap central electrode, such that the preamplifier is always ready for detection, (2) the capacitance between deflector at the entrance to the Orbitrap analyzer and each of detection electrodes was balanced by modifying the deflector geometry, (3) the capacitance between each of detection electrodes and ground was reduced and also balanced by replacing ceramic isolators with quartz ones as well as by changing the geometry of the Orbitrap holder. (4) The capacitance between Einzel lens elements at the entrance to the Orbitrap analyzer and each of these detection electrodes was minimized by implementing this lens as a miniature ceramic printed-circuit board.

Together, these measures allowed reducing the delay between ion injection and start of transient detection from almost 10 ms to a fraction of a millisecond. In addition to improved eFT, this reduction of delay allows to capture the entire first beat of the transient (see *e.g.* (34)) even for large proteins like intact antibodies and therefore significantly improves sensitivity of Orbitrap detection for top down analysis.

The practical implementation of the eFT achieves up to twofold increase of resolving power for the same transient. For rapidly decaying signals, for example from proteins, this gain is reduced to about 1.4-fold because of background collisions (8). The dual-spectrum online processing is computationally demanding but still fast enough to be completed in the LC MS time scale. Thus cycle time is still determined by transient acquisition and ion injection times and not by processing of the data. The eFT method is sensitive to precise synchronization of the instrument electronics and remaining shot-to-shot jitter, so that final mass accuracy is comparable to that of traditional magnitude mode FT spectra. Side-lobes in eFT spectra are comparable to those in conventional FT spectra.

*Sample Preparation*—Intact proteins (all from Sigma Aldrich) were dissolved in buffer A (98% water, 2% acetonitrile (ACN), 0.1% formic acid) prior to LC MS analysis. For direct infusion experiments, protein stock solutions were diluted in 50% ACN, 50% water, 0.1% formic acid.

HeLa cells were lysed in urea (6 M) and thiourea (2 M) solution. The protein mixture was reduced with dithiotreitol (1 mM) for 30 min at room temperature and alkylated with iodoacetamide (55 mM) for 20 min. The proteins were first digested with LysC (1 $\mu$g/50 $\mu$g protein) (Wako, Richmond, VA) for 3 h at room temperature. The sample was diluted (1:4) with water before 12 h incubation with trypsin (1 $\mu$g/50 $\mu$g protein) (Promega, Charbonnières, France) at room temperature. Formic acid (3%) was added to the mixture to quench enzyme activity. The peptide mixture was desalted on reversed phase $C_{18}$ StageTips (35) and eluted into 8 well autosampler vials with 60 $\mu$l buffer B (80% ACN in 0.5% acetic acid). ACN was removed in a SpeedVac concentrator. The sample volume was adjusted with buffer $A^*$ (2% ACN in 0.1% TFA) to 12 $\mu$l.

*LC MS/MS Analysis*—Intact proteins were separated on a nanobore analytical column (75 $\mu$m ID × 10 cm) with an integral fritted nanospray emitter (PicoFrit®, New Objective, Inc., Woburn, MA) containing 5 $\mu$m polymeric reversed-phase media (1000 Å pore size) using an EASY-nLC system (Thermo Scientific), which was operated at a flow rate of 300 nL/min. A linear gradient of each 20 min 10–50% buffer B and 50–80% buffer B (80% ACN in 0.1% formic acid) was applied. This setup was extended with a trap column (150 $\mu$m ID × 2 cm) containing identical chromatographic material. MS data were acquired by a relatively low resolution survey scan (10 microscans, resolution 15,000 at *m/z* 400), followed by a data dependent selected ion monitoring (SIM) scan of the most abundant (5 microscans, resolution 120,000 at *m/z* 400, isolation width 10 Th) and a data dependent HCD scan of the most abundant ion of the SIM scan (5 microscans, resolution 120,000 at *m/z* 400, isolation width 10 Th, normalized collision energy 18%). Dynamic exclusion duration of 10 s was enabled. A cycle time of 5.8 s was achieved. A modified method in which the data dependent SIM scan was replaced by a data dependent HCD scan at a different normalized collision energy, was used to increase the number of identified fragment ions.

The peptide mixture from a tryptic HeLa digest was separated with a linear gradient of 5–60% buffer B (80% ACN and 0.5% acetic acid) at a flow rate of 300 nL/min on a $C_{18}$-reversed phase column (75 $\mu$m ID × 15 cm) packed in-house with ReproSil-Pur $C_{18}$-AQ 3 $\mu$m resin (Dr. Maisch GmbH) in buffer A (0.5% acetic acid). An Easy-nLC chromatography system (Thermo Scientific) was on-line coupled to the Orbitrap Elite instrument (Thermo Scientific) via a Nanospray Flex Ion Source (Thermo Scientific). MS data were acquired in a data-dependent strategy selecting the fragmentation events based on the precursor abundance in the survey scan (300–1650 Th). The resolution of the survey scan was varied between 60,000, 120,000, and 240,000 at *m/z* 400 Th with a target value of 1e6 ions and 1 or 2 microscans. Low resolution CID MS/MS spectra were acquired with a target value of 5000 ions in normal and rapid CID scan mode. MS/MS acquisition in the linear ion trap was partially carried out in parallel to the survey scan in the Orbitrap analyzer by using the preview mode (first 192 ms of the MS transient). The maximum injection time for MS/MS was varied between 25 ms and 200 ms. HCD MS/MS spectra were acquired with a resolution of 15,000 and a target value of 40,000, setting the first mass to 120 Th. Dynamic exclusion was 60 s and early expiration was disabled. The isolation window for MS/MS fragmentation was set to 2 Th.

*Data Analysis*—High-resolution mass spectra of the intact proteins were deconvoluted using the Xtract software (Thermo Scientific) and further processed with ProSightPC (36, 37) (Thermo Scientific). The

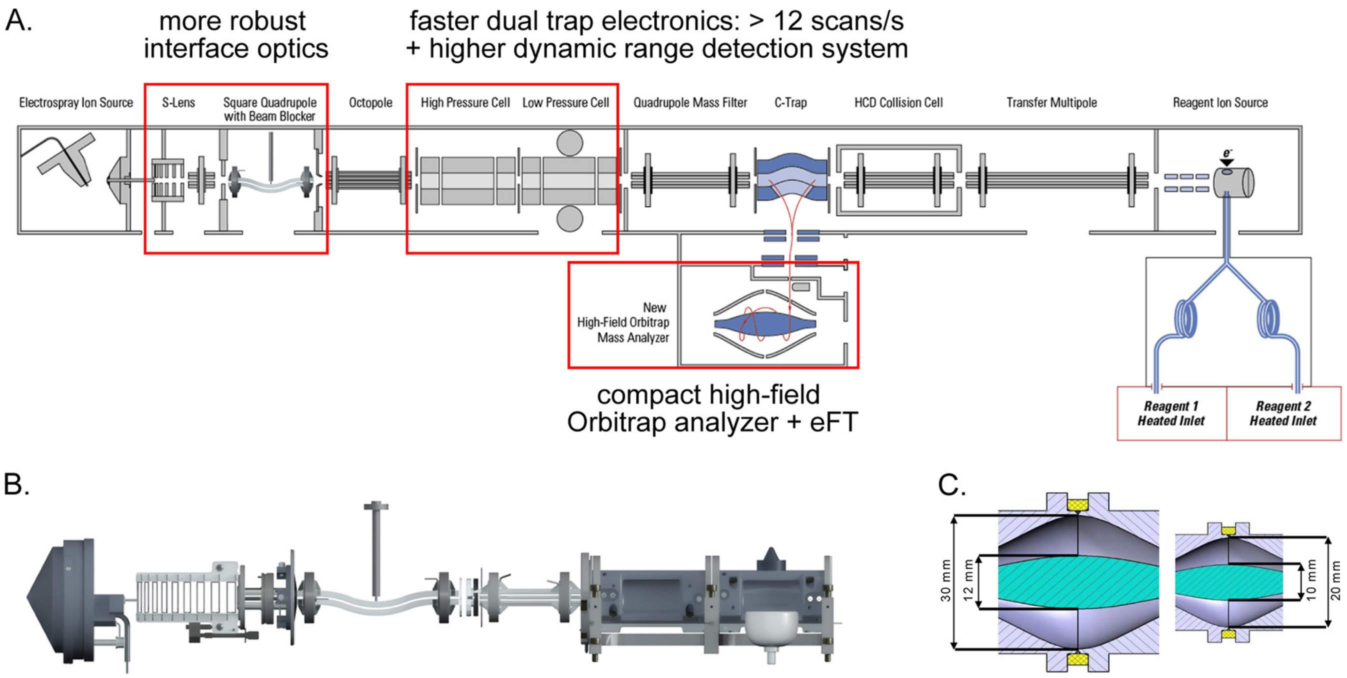

A.
more robust interface optics
faster dual trap electronics: > 12 scans/s + higher dynamic range detection system
Electrospray Ion Source
S-Lens
Square Quadrupole with Beam Blocker
Octopole
High Pressure Cell
Low Pressure Cell
Quadrupole Mass Filter
C-Trap
HCD Collision Cell
Transfer Multipole
Reagent Ion Source
New High-Field Orbitrap Mass Analyzer
compact high-field Orbitrap analyzer + eFT
Reagent 1 Heated Inlet
Reagent 2 Heated Inlet
B.
C.
30 mm
12 mm
10 mm
20 mm

FIG. 1. **The Orbitrap Elite mass spectrometer.** *A*, Novel elements compared with the LTQ Orbitrap Velos are highlighted and encompass the source region, the dual linear ion trap and the Orbitrap analyzer. ETD fragmentation is optional. *B*, Computer model of the inlet ion optics, showing the S-lens on the left and the dual ion trap on the right. The bent, square transfer quadrupole allows neutrals to leave the ion optics and impinge on the depicted beam blocker. *C*, Comparison of dimensions of the standard (*left*) to the compact, high-field Orbitrap analyzer (*right*).

analysis of the mass spectrometric RAW data was carried out using the MaxQuant software environment (developmental version 1.2.0.23) applying standard settings unless otherwise noted. Peptide scoring for identification is based on a search with an initial allowed mass deviation of the precursor ion of up to 7 ppm. To further improve the precursor mass accuracy a time-dependent recalibration algorithm was applied. Fragment mass tolerance was 0.5 Da for low resolution and 20 ppm for high resolution spectra. Enzyme specificity was defined as C-terminal to Arg and Lys including proline bond cleavage and a maximum of two missed cleavages. Carbamidomethylation of cysteine was set as fixed modification and N-terminal protein acetylation and methionine oxidation as variable modifications. MS/MS spectra were searched against the IPI human data base (version 3.68, 87,061 entries) combined with 262 common contaminants by the Andromeda search engine, with 2$^{nd}$ peptide identification enabled (38). The FDR for proteins and peptides (which had to have at least 6 amino acids) was set to 0.01. Results provided by MaxQuant were further analyzed using the R scripting and statistical environment. The data sets are provided at Tranche (www.proteomecommons.org) using the following hash: ZtXTc5NwSIxEwDzZKlD3b14XYnUCo7nKbWAnIGRZXORy3eoXWSodhh/w7SZBxTAZbqoDPDs8FGkVcBBpe8N1fC1+M1EAAAAAAAAANYw==.

## RESULTS AND DISCUSSION

*Overview of the Orbitrap Elite*—Like its predecessors, the LTQ Orbitrap and LTQ Orbitrap Velos, the instrument is a combination of two mass analyzers, a linear ion trap and an Orbitrap mass analyzer (Fig. 1). It offers up to three fragmentation modes, CID or ETD in the dual linear ion trap, or HCD in its dedicated collision cell. The fragment ions produced by CID or ETD are normally analyzed in the ion trap at low resolution and in parallel with the acquisition of the MS transient. However, they can also be transferred to the C-Trap and recorded in the Orbitrap analyzer. HCD fragments are always analyzed in the Orbitrap analyzer.

In the Orbitrap Elite instrument, as discussed in detail in EXPERIMENTAL PROCEDURES, the interface optics in the front part of the instrument now incorporates a rotated square transfer quadrupole with a beam blocker preventing neutral or low charged material from passing. The dual ion trap was equipped with new electron multipliers that tolerate up to 10-fold higher ion currents, which improves the dynamic range of the device. This in turn allowed further speed-up of the low resolution CID fragmentation mode. This scan mode was named rCID for rapid CID and it allows acquisition of up to 12.5 MS/MS spectra per second (EXPERIMENTAL PROCEDURES). The changes in interface optics together with the improved dual ion trap constitute the Orbitrap Velos Pro instrument, whereas the central feature of the Orbitrap Elite instrument described in this publication is a novel Orbitrap analyzer with drastically improved resolving power.

A detailed description of the principles and construction of the novel Orbitrap analyzer is given in EXPERIMENTAL PROCEDURES. Briefly, the major hardware change was a reduction in the inner diameter of the outer electrode from 30 to 20 mm whereas the size of the spindle-shaped inner electrode was only reduced from 12 to 10 mm (Fig. 1*C*). These smaller dimensions, with an increased ratio of inner to outer electrode diameters, lead to a higher field strength, resulting in almost doubling of the resolving power at the same scan time. An improved signal processing algorithm (eFT), which takes phase information into account (EXPERIMENTAL PROCEDURES), provides a further boost by a factor up to 2. Together, the Orbitrap Elite instrument achieves about fourfold higher resolution at the same transient length. This translates into 240,000 resolution at *m/z* 400 with the standard 768 ms transient—a radical improvement over the 60,000 resolving power achieved by the predecessor instrument. This very high resolution per unit time can alternatively be used to shorten cycle times at the same resolving power. An overview of the different resolution settings as a function of transient length can be found in Table I. Even the shortest transient (48 ms), results in a resolving power of 15,000, twice that of the 96 ms scan typically used for MS/MS on the Orbitrap Velos. In principle, even shorter transients would produce sufficient resolution for MS/MS. However, we decided against such scan modes because the ratio of useful transient time compared with overhead times and ion filling times would become unfavorable.

TABLE I

*Resolution and transient length of Orbitrap hybrid instruments. The highest resolution on Orbitrap Elite (indicated by an asterisk) can only be activated when using the developer's kit*

| Resolution | | Transient |
|---|---|---|
| LTQ Orbitrap | Orbitrap Elite | |
| – | 15,000 | 48 ms |
| 7500 | 30,000 | 96 ms |
| 15,000 | 60,000 | 192 ms |
| 30,000 | 120,000 | 384 ms |
| 60,000 | 240,000 | 768 ms |
| 120,000 | 480,000* | 1536 ms |

*Top Down at an LC Time Scale*—High resolution is an important requirement for resolving the different charge states of intact proteins investigated in top-down proteomics. We selected carbonic anhydrase II, a frequently used standard in top-down proteomics, to investigate the advantages of the increased resolving power and sequencing speed of the instrument. We devised a method that alternates between a survey scan with 15,000 resolution followed by a high resolution SIM scan of a particular charge state acquired in a data dependent manner. This same precursor m/z is subsequently

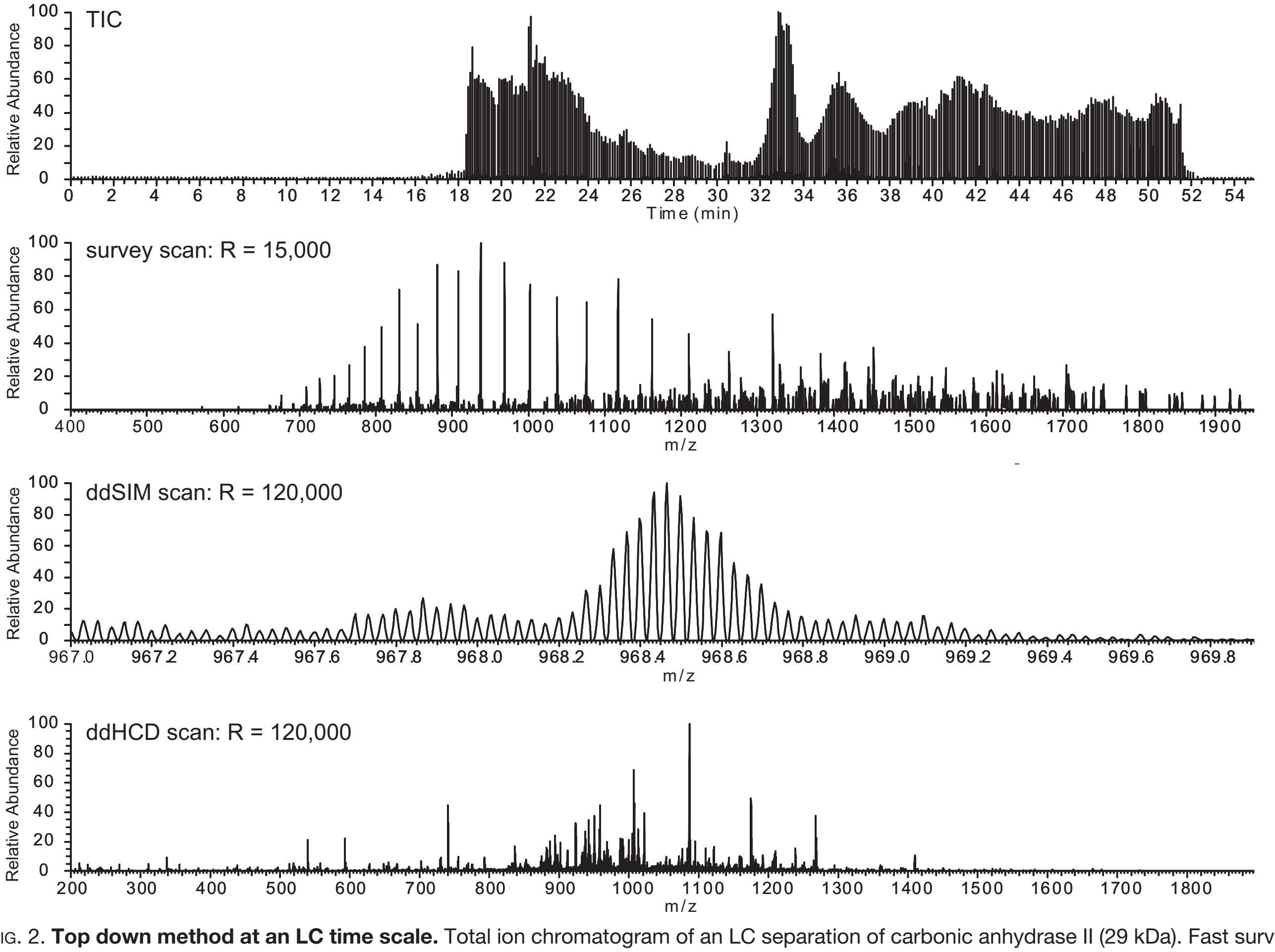


FIG. 2. **Top down method at an LC time scale.** Total ion chromatogram of an LC separation of carbonic anhydrase II (29 kDa). Fast survey scans reveal the charge envelope and are followed by high resolution SIM and HCD scans. Overlapping fragment isotope distributions are clearly resolved from each other.

fragmented by HCD and analyzed at a high resolving power of 120,000 (Fig. 2). We found that the Orbitrap Elite is capable of isotopically resolving and measuring this 29 kDa protein with a root mean square mass accuracy below 2 ppm under these conditions. Averaging times for five to six microscans of the HCD fragmentation spectra were much shorter than in the LTQ Orbitrap Velos, without reducing signal-to-noise. Deconvolution with Xtract of a single, six microscan HCD spectrum, in which charge state 34+ carbonic anhydrase was fragmented, revealed 30 b-type and 31 y-type ions (supplemental Fig. S1). For enolase (46.6 kDa), 15 b-type and 20 y-type ions were identified following fragmentation of the 58+ precursor charge state by HCD. For both of these model proteins optimizing the MS/MS parameters such as increasing the collision energy but mainly averaging of up to 24 microscans increased the number of identified fragment ions to 49 b-type and 48 y-type ions for carbonic anhydrase II and 30 b-type and 42 y-type ions for enolase (supplemental Fig. S1).

*Ultrahigh Orbitrap Resolution for Top Down Experiments*—The implementation of the compact Orbitrap analyzer and eFT signal processing allows the Orbitrap Elite to reach a comparably high resolution as a 17 Tesla FT ICR instruments at *m/z* 400 Th and with standard signal processing methods. In contrast to the inversely linear dependence of FT ICR resolving power on *m/z*, however, the Orbitrap resolving power is inversely proportional to the square root of *m/z* (1). Therefore, for subsecond acquisition on an LC-MS time scale its resolving power at *m/z* 1000 already corresponds to a 25 Tesla FT-ICR instrument, and it should be particularly suitable for top down experiments of larger proteins such as BSA (66.4 kDa) and enolase (46.64 kDa). Using static electrospray conditions we found that charge state 47+ of intact yeast enolase could readily be baseline resolved at a resolution setting of 240,000 corresponding to 768 ms transients (Fig. 3*A*). To reach this resolution, a vacuum better than $10^{-10}$ Torr was necessary. On the predecessor instrument, partial resolution of intact yeast enolase could occasionally be observed (Fig. 3*B*).

*Parallel topN CID Method for Bottom Up Proteomics*—For the analysis of complex peptide mixtures, a standard mode of operation is the acquisition of a survey spectrum in the Orbitrap analyzer while the linear ion trap isolates, dissociates, and scans the fragments. For this "high-low" mode, we evaluated the influence of the higher resolution in the survey spectra and of the faster MS/MS scans. LTQ Orbitrap instruments select precursors for fragmentation on the basis of a snapshot or preview spectrum—the first 192 ms of the MS transient—after which time the CID scans are initiated in the LTQ. However the resolution is increased fourfold, allowing isotopic resolution for all charge states and the entire *m/z* range (300–1650) at higher signal-to-noise. We found that this increased the quality of precursor ion selection.

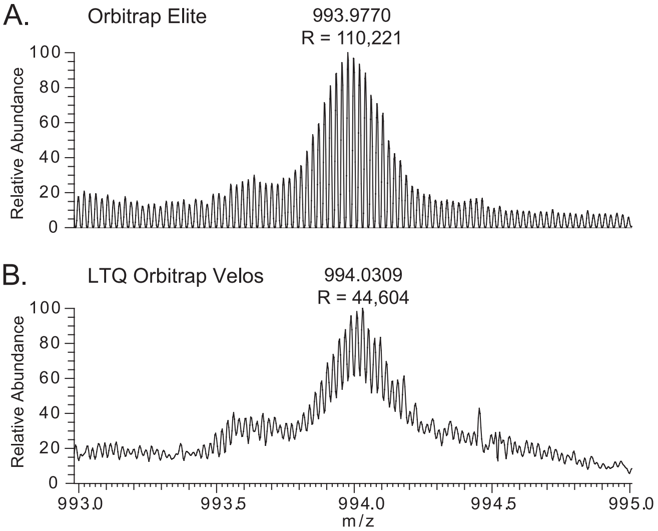


FIG. 3. **Isotope resolved spectrum of enolase with Orbitrap Elite and LTQ Orbitrap Velos instruments.** Spectra were acquired with transients of *A*, 768 ms on the Orbitrap Elite and *B*, 1536 ms on the LTQ Orbitrap Velos. In these conditions, the superior resolution of the Orbitrap Elite instrument (>2 × higher in 2 × shorter transient time) helps to baseline resolve the 47+ charge state of the intact yeast enolase.

To investigate the instrument capability for the analysis of very complex peptide mixtures such as HeLa cell lysate, we started with a digested standard of 400 ng that was analyzed using different methods. Comparison of the normal CID and rapid CID scan modes revealed that rCID produced significantly more fragmentation events and therefore this mode was chosen for all subsequent experiments. The scheme in Fig. 4*A* shows the timing sequence of MS and MS/MS scans at different MS transient lengths. For a 192 ms survey scan, resolution is 60,000 and there is no parallel operation between MS and MS/MS scans. Due to rCID an entire top20 method takes 2.7 s, easily compatible with peptide LC elution profiles. At 384 ms resolution is 120,000 and a few MS/MS scans are performed in parallel with the survey scan. At the full 768 ms transient (240,000 resolution), about six CID spectra are performed in parallel, while total cycle time is still unchanged (Fig. 4*A*). Therefore, the longest transients appear to be advantageous because the increased resolution comes "for free" as it does not cost extra measurement time.

We make use of the fixed cycle time to explore the benefits of high resolving power on complex peptide mixture analysis. Three top20 methods were established as outlined above using 60,000, 120,000, and 240,000 resolution for the survey scan (Fig. 4*A*). When analyzing the HeLa peptide mixture with these methods, we found that the number of potential peptide features (isotope clusters) detected by MaxQuant, nearly doubled from 93,000 at 60,000 resolution to 148,000 at 240,000 resolution under otherwise identical conditions (Table II). Fig. 4*B* shows a zoom into the LC MS map of the 240,000 reso-

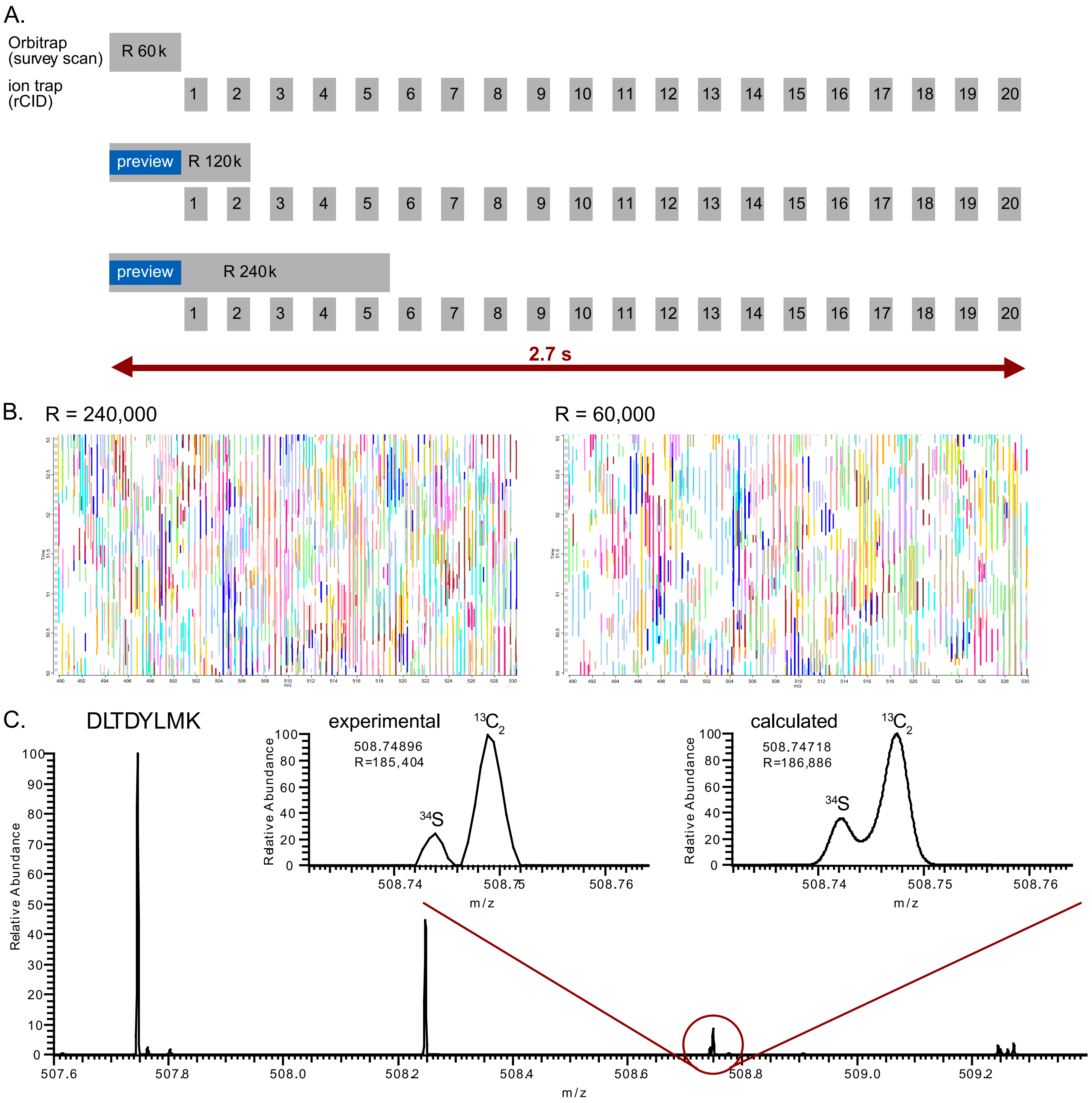


FIG. 4. **Parallel CID top20 method and ultra high resolution survey scans.** *A*, High resolution MS scan at three different transient lengths followed by 20 CID MS/MS scans in the linear ion trap. Note that cycle time is unaffected by the resolution of the full scan. Preview refers to the portion of the survey scan that is used to select precursor ions for fragmentation. *B*, LC MS heat map of peptides eluting over a 3 min elution time interval in a 40 Th range. More detail is visible in the ultra high resolution setting (*left* panel) compared with the normal resolution setting (*right* panel). *C*, Separation of isobaric species in standard LC MS/MS analysis. The $^{34}S$ isotope containing peak is clearly resolved from the $^{13}C_2$ isotope.

lution *versus* the 60,000 resolution run and demonstrates the rich feature set in complex peptide mixtures with 540 *versus* 320 isotope clusters, respectively.

ICR Fourier transform mass spectrometry measurements of small molecules can resolve the fine structure in the isotope patterns of small molecules (isobaric species with the same nominal mass) (39), although to our knowledge this has not yet been reported in proteomics LC MS/MS experiments. We inspected the methionine and cysteine containing peptides and found that the $^{34}S$ isotope and $^{13}C_2$ peaks ($\Delta M = 0.011$

Da) were clearly resolved from each other (Fig. 4*C*). Thus, the high resolution immediately indicates the presence of a sulfur atom in the peptide.

Based on the above observations, we selected a 240,000 resolution survey scan and top20 CID scan as the standard high-low method. Analysis of the 2h gradient of the 400 ng HeLa peptide sample identified 11,543 unique peptides and 2268 proteins (supplemental Tables. S1–S4), an increase of 25–30% over the analysis of the same sample on the LTQ Orbitrap Velos. Although not demonstrated here, quantification accuracy was also observed to increase due to the several-fold higher resolution.

TABLE II
*Number of isotope clusters depending on survey scan resolution at equal cycle time*

| Resolution | Cycle time | MS scans | Isotope clusters |
|---|---|---|---|
| 60,000 | 2.76 | 1902 | 93,624 |
| 120,000 | 2.87 | 1821 | 124,808 |
| 240,000 | 2.85 | 1843 | 148,085 |

*High Resolution MS/MS Methods*—The Orbitrap Elite offers CID, HCD and optional ETD fragmentation modes. Analysis of the fragments is either in parallel mode in low resolution in the linear ion trap (for CID and ETD) or in sequential mode in the Orbitrap analyzer (for CID, ETD, and HCD). In principle, all these modes can be mixed and matched according to the analytical question under investigation. Fig. 5 depicts three prototypical combinations of scan modes. In a pure HCD mode, a survey scan is acquired and is followed immediately by N HCD spectra that are also recorded in the Orbitrap analyzer. A 384 ms survey scan already provides resolution of 120,000, which is adequate for most applications and which helps to limit the overall cycle time in this mode. HCD spectra are acquired at the lowest possible resolution (48 ms transients; 15,000 resolution at *m/z* 400). A complete top15 sequence takes 3.3 s, about the same time required for a top10 method with half the resolution in MS and in MS/MS mode in the predecessor instrument. We measured the 400 ng HeLa cell lysate sample with this method and obtained significantly improved numbers of peptide identifications compared with

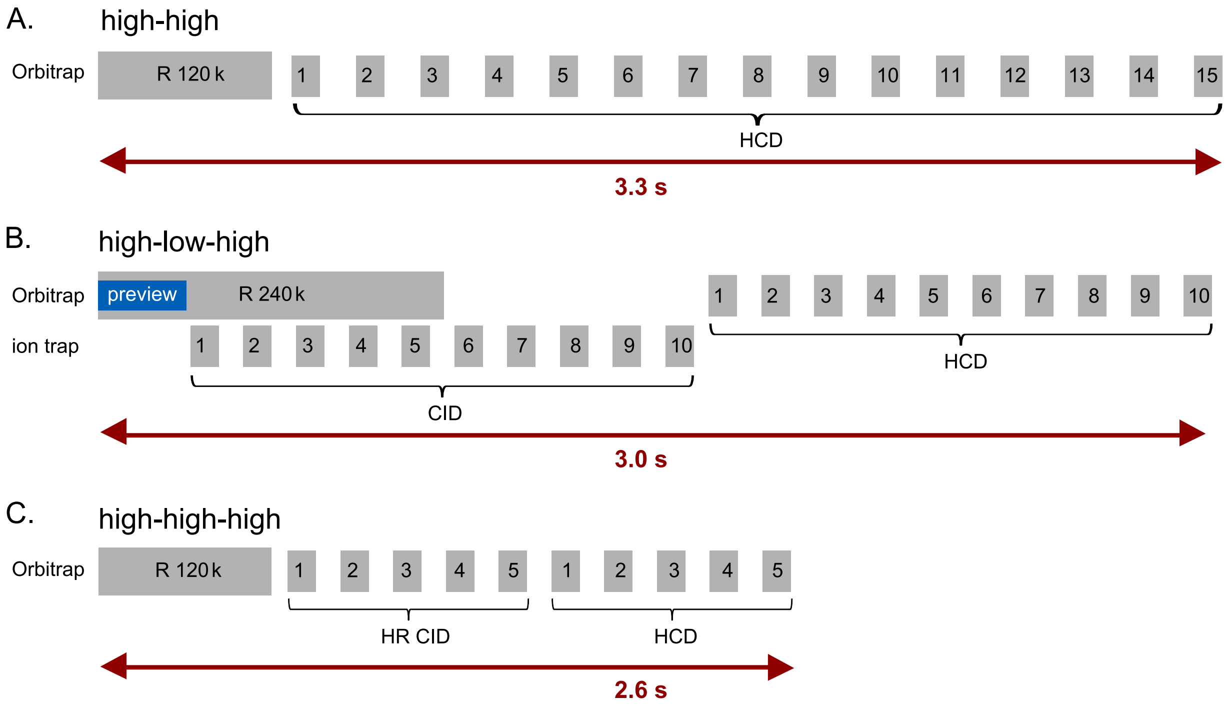


FIG. 5. **Combinations of fragmentation modes.** *A*, Sequential mode in which a high resolution survey scan with 120,000 resolution (384 ms) is followed by 15 HCD scans at 15,000 resolution (48 ms transients). *B*, Parallel and sequential mode in which an ultra high survey scan with 240,000 resolution is acquired in parallel with 10 rCID spectra in the linear ion trap and in sequence with 10 HCD scans that are analyzed in the Orbitrap analyzer. The same precursors are analyzed by rCID and HCD. *C*, Double sequential mode in which a 120,000 resolution survey scan is followed by 5 high resolution CID and 5 HCD spectra of the same precursors.

TABLE III
*Protein and peptide identification from HeLa duplicate analysis of a top15 HCD method*

| | MS spectra | MSMS spectra | Identifications [%] | Unique peptides | Proteins | Isotope clusters |
|---|---|---|---|---|---|---|
| HCDtop15(1) | 1556 | 23,211 | 45.77 | 10,847 | 2082 | 116,632 |
| HCDtop15(2) | 1538 | 22,930 | 45.82 | 10,633 | 2064 | 115,370 |

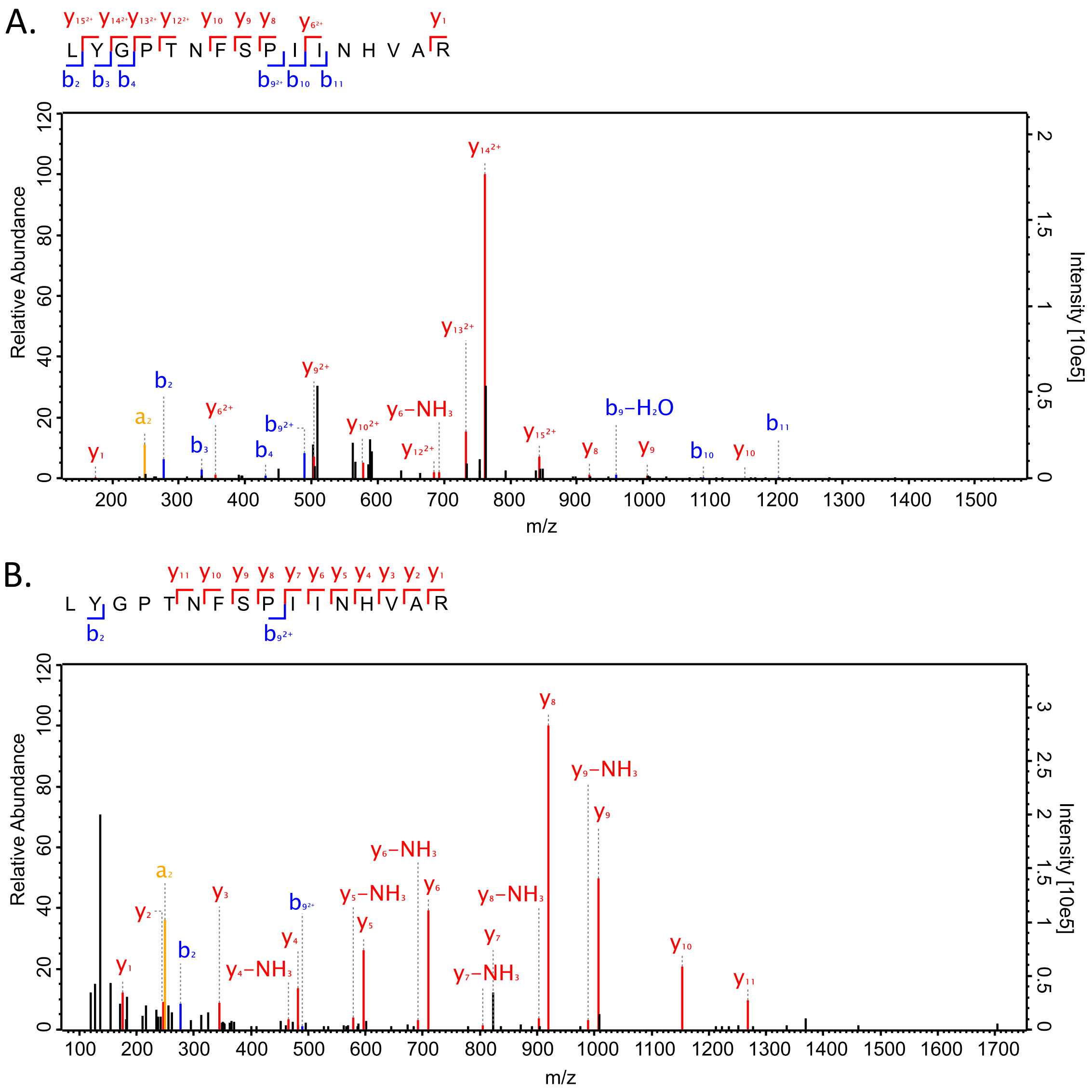


FIG. 6. **Complementary CID + HCD MS/MS spectra.** Example spectra using the parallel and sequential fragmentation mode depicted in Fig. 5*B*. *A*, The rCID spectrum features more b-ions than the HCD spectrum. *B*, HCD spectrum with nearly complete y-ion series. Mass accuracy in *B* but not in *A* is in the ppm range (absolute average deviation of 3.67 ppm. *versus* 0.06 Th).

LTQ Orbitrap Velos measurements (Table III). This high-high HCD mode offers high sequencing speed and high mass accuracy MS/MS spectra and is therefore well suited to the analysis of complex mixtures. In comparison to the CID method described above, it achieves similar numbers of peptide fragmentation events because it is slightly faster than CID in the linear ion trap but does not have parallel operation. Target values for fragmentation are somewhat higher in HCD (40,000 *versus* 5,000 ions).

The high scan speeds of this instrument also makes more complex scan modes possible. For instance, the parallel low resolution scan mode can be followed by high resolution, sequential HCD Orbitrap analyzer scans (Fig. 5*B*). To test this mode, we selected up to 10 precursors for fragmentation by CID and repeated fragmentation of the same candidates by HCD. As can be seen in the schematic, this mode makes particularly good use of the hybrid instrument's capabilities. The initial parallel operation with CID fragmentation allows ample time to perform a survey scan with 240,000 resolution without affecting total cycle time. In principle, it results in two fragmentation spectra for each of the peptides. The advantages are illustrated in Fig. 6, the *upper* panel of which shows the CID spectrum of the peptide LYGPTNFSPIINHYAR, while the *lower* panel shows the corresponding HCD spectrum acquired subsequently in the same cycle. The CID spectrum has the typical mixture of b- and y-ions, whereas the HCD spectrum contains a nearly complete series of y-ions but few b-ions. Both spectra together account for all possible y-ions as well as a large proportion of all b-ions of this peptide. Importantly, this dual fragmentation information comes at an acceptable cost in cycle time and sequencing speed. This parallel top10 CID + HCD method had a total cycle time of 3.0 s, very similar to the top15 HCD method. When targeting the same precursors, it sacrifices some sequencing speed but gains in complementary fragmentation information. Thus it may be particularly interesting in applications where relatively high sequencing speed is important but where peptide identification is challenging, for instance in the analysis of post-translational modifications at a large scale.

The final method depicted in Fig. 5C is a further step in the same direction. Here, two fragmentation modes are applied but all MS/MS scans are performed in the Orbitrap analyzer. Therefore this mode is completely sequential and does not use the linear ion trap as a scanning mass spectrometer. A 120,000 resolution survey scan followed by sequential top5 CID + HCD scans takes about 2.6 s. When targeting the same precursors, it sacrifices sequencing events compared with the above methods. However, high resolution MS/MS spectra are obtained by two different fragmentation methods, yielding maximum information of the primary structure. Therefore the sequential top5 CID + HCD scan mode would be very attractive for applications with limited peptide complexity but high demands on peptide characterization. This could be the case in traditional, single protein applications, in proteomics when no complete database is available, or generally when unusual modifications are expected.

## CONCLUSIONS AND OUTLOOK

Here we have described the Orbitrap Elite, a mass spectrometer that achieves fourfold improved resolving power by increasing the electric field strength in the Orbitrap analyzer and by enhanced Fourier Transformation. The high resolving power enables ready isotopic resolution of proteins in the BSA mass range as well as characterization of their fragments in a chromatographic time scale. In bottom-up proteomics, the instrument allows high-low topN CID methods featuring ultra high resolution survey scans and a large number of parallel MS/MS experiments in the linear ion trap. For instance, we demonstrated the combination of a survey scan of 240,000 resolution with 20 CID scans all within a 2.7 s cycle time. Remarkably, this high resolution routinely enabled resolving isobars of sulfur-containing peptides. We also explored the acquisition of CID and HCD spectra of the same precursor ions, with either parallel or sequential analysis of the fragmentation spectra. Although not shown here, CID and HCD fragmentation events could also be distributed to different peptide classes. Further fragmentation modes, such as ETD, can also be incorporated. Moreover, the product ions of one fragmentation method could be dissociated again by the same or other fragmentation methods—for example, CID could be followed by HCD and resulting fragments could be recorded in the Orbitrap analyzer—all at a rapid time scale. All these methods are possible in principle and it will be interesting to develop them for a wide range of proteomic and other applications.

*Acknowledgments*—We thank our colleagues at Thermo Fisher Scientific, and at the Max Planck Institute for help and fruitful discussions. We also thank Dr. Phil Compton from Northwestern University for valuable discussions.

* This work was supported by the European Commission's 7th Framework Programme (grant agreement HEALTH-F4-2008-201648/PROSPECTS).

S This article contains supplemental Tables S1 to S4 and Fig. S1.

‖ To whom correspondence should be addressed: Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. E-mail: mmann@biochem.mpg.de.

** Contributed equally.

Data availability: Supplementary data is available with this publication at the MCP web site. Raw MS files are uploaded to Tranche (www.proteomecommons.org) as "Michalski et al. Elite".

## REFERENCES

1. Scigelova, M., Hornshaw, M., Giannakopulos, A., and Makarov, A. (2011) Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **10,** M111.009431
2. Kaiser, N. K., Savory, J. J., McKenna, A. M., Quinn, J. P., Hendrickson, C. L., and Marshall, A. G. (2011) Electrically compensated fourier transform ion cyclotron resonance cell for complex mixture mass analysis. *Anal. Chem.* **83,** 6907–6910
3. Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72,** 1156–1162
4. Scigelova, M., and Makarov, A. (2006) Orbitrap mass analyzer—overview and applications in proteomics. *Proteomics* **6** Suppl **2,** 16–21
5. Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., and Horning, S. (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **78,** 2113–2120
6. Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005) The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40,** 430–443
7. Makarov, A., Denisov, E., and Lange, O. (2009) Performance evaluation of a high-field Orbitrap mass analyzer. *J. Am. Soc. Mass Spectrom.* **20,** 1391–1396
8. Vining, B. A., Bossio, R. E., and Marshall, A. G. (1999) Phase correction for collision model analysis and enhanced resolving power of fourier transform ion cyclotron resonance mass spectra. *Anal. Chem.* **71,** 460–467
9. Beu, S. C., Blakney, G. T., Quinn, J. P., Hendrickson, C. L., and Marshall, A. G. (2004) Broadband phase correction of FT-ICR mass spectra via simultaneous excitation and detection. *Anal. Chem.* **76,** 5756–5761
10. Lange, O., Damoc, E., Wieghaus, A., and Makarov, A. (2011) Enhanced Fourier Transform for Orbitrap Mass Spectrometry. *Proc. 59th Conf. Amer. Soc. Mass Spectrom., Denver, June 5* **9,** 2011
11. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* 10(9):M111.011015. Epub 2011 Jun 3
12. Kelleher, N. L. (2004) Top-down proteomics. *Anal. Chem.* **76,** 197–203A
13. McLafferty, F. W. (2011) A century of progress in molecular mass spectrometry. *Annu. Rev. Anal. Chem.* **4,** 1–22
14. Breuker, K., Jin, M., Han, X., Jiang, H., and McLafferty, F. W. (2008) Top-down identification and characterization of biomolecules by mass spectrometry. *J. Am. Soc. Mass Spectrom.* **19,** 1045–1053
15. Parks, B. A., Jiang, L., Thomas, P. M., Wenger, C. D., Roth, M. J., Boyne, M. T., 2nd, Burke, P. V., Kwast, K. E., and Kelleher, N. L. (2007) Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers. *Anal. Chem.* **79,** 7984–7991
16. Macek, B., Waanders, L. F., Olsen, J. V., and Mann, M. (2006) Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-

orbitrap mass spectrometer. *Mol. Cell. Proteomics* **5,** 949–958

17. Waanders, L. F., Hanke, S., and Mann, M. (2007) Top-down quantitation and characterization of SILAC-labeled proteins. *J. Am. Soc. Mass Spectrom.* **18,** 2058–2064
18. Bondarenko, P. V., Second, T. P., Zabrouskov, V., Makarov, A. A., and Zhang, Z. (2009) Mass measurement and top-down HPLC/MS analysis of intact monoclonal antibodies on a hybrid linear quadrupole ion trap-Orbitrap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **20,** 1415–1424
19. Théberge, R., Infusini, G., Tong, W., McComb, M. E., and Costello, C. E. (2011) Top-Down Analysis of Small Plasma Proteins Using an LTQ-Orbitrap. Potential for Mass Spectrometry-Based Clinical Assays for Transthyretin and Hemoglobin. *Int. J. Mass Spectrom.* **300,** 130–142
20. Cravatt, B. F., Simon, G. M., and Yates, J. R., 3rd (2007) The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000
21. Cox, J., and Mann, M. (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Ann. Rev. Biochem.* **80,** 273–299
22. Beck, M., Claassen, M., and Aebersold, R. (2011) Comprehensive proteomics. *Curr. Opinion Biotechnol.* **22,** 3–8
23. Köcher, T., Swart, R., and Mechtler, K. (2011) 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
24. Michalski, A., Cox, J., and Mann, M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793
25. Mann, M., and Kelleher, N. L. (2008) Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 18132–18138
26. 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., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8,** 2759–2769
27. Savitski, M. M., Nielsen, M. L., Kjeldsen, F., and Zubarev, R. A. (2005) Proteomics-grade de novo sequencing approach. *J. Proteome Res.* **4,** 2348–2354
28. Fälth, M., Savitski, M. M., Nielsen, M. L., Kjeldsen, F., Andren, P. E., and Zubarev, R. A. (2007) SwedCAD, a database of annotated high-mass accuracy MS/MS spectra of tryptic peptides. *J. Proteome Res.* **6,** 4063–4067
29. Swaney, D. L., McAlister, G. C., and Coon, J. J. (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **5,** 959–964
30. Second, T. P., Blethrow, J. D., Schwartz, J. C., Merrihew, G. E., MacCoss, M. J., Swaney, D. L., Russell, J. D., Coon, J. J., and Zabrouskov, V. (2009) Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. *Anal. Chem.* **81,** 7757–7765
31. Wouters, E. R., Splendore, M., Mullen, C., Schwartz, J. C., Senko, M., and Dunyach, J. J. (2009) Implementation of a progressively spaced stacked ring ion guide on a linear ion trap mass spectrometer. *57th Amer. Soc. Mass Spectrom. Annual Conf. on Mass Spectrometry & Allied Topics, Philadelphia, PA, May 31–June 5 (2009*)
32. McAlister, G. C., Phanstiel, D. H., Brumbaugh, J., Westphall, M. S., and Coon, J. J. (2011) Higher-energy collision-activated dissociation without a dedicated collision cell. *Mol. Cell. Proteomics* **10,** O111.009456
33. Makarov, A. (2009) Practical Aspects of Trapped Ion Mass Spectrometry. In: March, R. E., and Todd, J. F. J., eds. *Theory and Instrumentation*, CRC Press (Taylor & Francis)
34. Makarov, A., and Denisov, E. (2009) Dynamics of ions of intact proteins in the Orbitrap mass analyzer. *J. Am. Soc. Mass Spectrom.* **20,** 1486–1495
35. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
36. LeDuc, R. D., Taylor, G. K., Kim, Y. B., Januszyk, T. E., Bynum, L. H., Sola, J. V., Garavelli, J. S., and Kelleher, N. L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res.* **32,** W340–345
37. Leduc, R. D., and Kelleher, N. L. (2007) Using ProSight PTM and related tools for targeted protein identification and characterization with high mass accuracy tandem MS data. *Current protocols in bioinformatics/ editoral board, Andreas D. Baxevanis. [et al.]* Chapter 13, Unit 13 16
38. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10,** 1794–1805
39. Solouki, T., Emmett, M. R., Guan, S., and Marshall, A. G. (1997) Detection, number, and sequence location of sulfur-containing amino acids and disulfide bridges in peptides by ultrahigh-resolution MALDI FTICR mass spectrometry. *Anal. Chem.* **69,** 1163–1168

# 4 Extended Data Analysis Tools for Shotgun Proteomics

**Article 4** **Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment**

Jürgen Cox, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema,
Jesper V. Olsen, and Matthias Mann, *J Proteome Res,* 2011, 10 (4), 1794.

**Article 5** **Software Lock Mass by Two-Dimensional Minimization of Peptide Mass Errors**

Jürgen Cox, Annette Michalski, and Matthias Mann,
*J Am Soc Mass Spectrom*, 2011, 22(8), 1373.

## Prologue

Modern proteomics datasets contain tens of thousands of mass spectra that make manual interpretation impractical. It would not only be too time-consuming, but also prone to a bias introduced by the person performing the analysis. Identification of peptides, reassembly of proteins as well as scoring and filtering of the results are highly dependent on reproducible criteria and application of statically valid models.

The commercial search engine Mascot has evolved into a gold standard for peptide identification[110]. Unfortunately, the underlying algorithms are inaccessible to the proteomics community, which prohibits any modification or adaption to specific problems. After using the server-based Mascot search engine in conjunction with the MaxQuant software in our laboratory for several years, we wished to gain greater flexibility in programming as well as to switch the entire data-processing pipeline to desktop computers. A novel search engine, *Andromeda*, was therefore developed and integrated into MaxQuant. It features the ability to work with arbitrarily high fragment mass accuracy. Like Mascot, it applies a probabilistic scoring model and achieves similar overall performance on large-scale shotgun proteomics datasets. During extensive tests on large-scale datasets and in agreement with our Expert System project (Articles 6 and 7), we found it beneficial to supplement the ion types used for peptide identification by water and ammonia losses of specific amino acids. Integrating more detailed knowledge of peptide fragmentation into the scoring algorithm remains a project for the future.

In this thesis we evaluated the success of a second peptide algorithm that was implemented in order to tackle the problem of co-fragmenting precursor ions revealed in Article 1. We quantified the benefit of identifying more than one peptide from each tandem mass spectrum as a function of the width of the isolation window; 4 Th were found to give best results (Article 4). Larger isolation windows dilute the targeted precursor ion too strongly and smaller windows permit fewer extra identification. Finding second peptides was usually possible when there were indeed only one or two extra precursor ions in the isolation window. In contrast, our strategy did not help if the remaining precursor intensity was spread over many very low abundant peaks.

High quality peptide identifications are often associated with perfectly annotatable tandem mass spectra; however, they can also greatly rely on the accurate precursor mass. Due to the very large number of peptide sequences obtained from proteins in databases, the experimental mass accuracy of the precursor plays a crucial role in reducing the number of candidates. A prerequisite for high mass accuracy is resolution of isotopic patterns and ideally of overlaying clusters of different precursor ions, possibly even with different charge states. Mass accuracy, however, also strongly depends on the calibration of the instrument.

We therefore investigated two approaches that were applied to high resolution data at different stages of the workflow: At first, we investigated the lock mass feature of the LTQ Orbitrap, where polydimethylcyclosiloxane that is present in the laboratory air is deliberately added to each spectrum for internal calibration[58]. The second strategy was novel and entirely software based. It can be applied to any dataset of complex mixtures and comes at no experimental cost. This *software lock mass* makes use of sample-inherent information: highly confident peptide identifications obtained from a first search with larger mass tolerance (20 ppm or more), implemented in Andromeda, are utilized to find a non-linear recalibration function that corrects mass errors both on the retention time scale and on the m/z axis. The software lock mass proved to be at least as successful as the physical lock mass and is therefore now routinely used in our and other laboratories (Article 5).

Journal of proteome research

ARTICLE

pubs.acs.org/jpr

# Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment

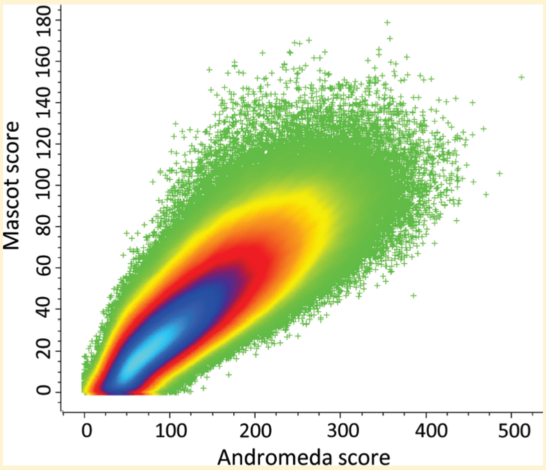
Jürgen Cox,*[†] Nadin Neuhauser,[†] Annette Michalski,[†] Richard A. Scheltema,[†] Jesper V. Olsen,[‡] and Matthias Mann*[†,‡]

[†]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[‡]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark

S *Supporting Information*

**ABSTRACT:** A key step in mass spectrometry (MS)-based proteomics is the identification of peptides in sequence databases by their fragmentation spectra. Here we describe Andromeda, a novel peptide search engine using a probabilistic scoring model. On proteome data, Andromeda performs as well as Mascot, a widely used commercial search engine, as judged by sensitivity and specificity analysis based on target decoy searches. Furthermore, it can handle data with arbitrarily high fragment mass accuracy, is able to assign and score complex patterns of post-translational modifications, such as highly phosphorylated peptides, and accommodates extremely large databases. The algorithms of Andromeda are provided. Andromeda can function independently or as an integrated search engine of the widely used MaxQuant computational proteomics platform and both are freely available at www.maxquant.org. The combination enables analysis of large data sets in a simple analysis workflow on a desktop computer. For searching individual spectra Andromeda is also accessible via a web server. We demonstrate the flexibility of the system by implementing the capability to identify cofragmented peptides, significantly improving the total number of identified peptides.



**KEYWORDS:** tandem MS, search engine, spectrum scoring, post-translational modifications, mass accuracy, collision induced dissociation, higher-energy collisional dissociation, Orbitrap

## INTRODUCTION

Mass spectrometry (MS)-based proteomics is becoming a commonly used technology in a wide variety of biological disciplines.[1−6] In a "shotgun" format, very complex peptide mixtures are produced by enzymatic digestion of protein mixtures, which are analyzed by liquid chromatography followed by tandem mass spectrometry.[7,8] Per LC−MS/MS run, thousands of MS and MS/MS scans are acquired, often producing gigabytes of high resolution data per day and per mass spectrometer. Computational proteomics has become a key research area, dealing with the challenges of how to most efficiently extract protein identification and quantification results from the raw data. Both the proteomics community and the bioinformatics community have dealt with many areas of this novel field, and there is already a large literature outlining and reviewing the general tasks involved,[9−17] particular computational aspects of the field[18−22] and integrated data analysis pipelines.[23−30]

In this context, our group has developed the MaxQuant environment, a computational proteomics workflow that addresses the above tasks with a focus on high accuracy and quantitative data. It includes peak detection in the raw data, quantification, scoring of peptides and reporting of protein groups.[31] MaxQuant takes advantage of high resolution data such as those obtained by the linear ion trap−Orbitrap instruments and employs algorithms that determine the mass precision and accuracy of peptides individually. This leads to greatly enhanced peptide mass accuracy that can be used as a filter in database searching.[32] MaxQuant was also specifically designed to achieve the highest possible quantitative accuracy in conjunction with stable isotope labeling with amino acids in cell culture (SILAC).[33,34] Using high resolution data combined with individualized mass accuracies and robust peptide and protein scoring results in high peptide identification rates of typically 50% and even higher on SILAC peptide pairs.[31] This was an important foundation for the quantification of the first complete model proteome, that of budding yeast.[35]

**Received:** October 23, 2010
**Published:** January 21, 2011

The MaxQuant environment originally used the Mascot peptide search engine[36] to match tandem mass spectra to possible peptide sequences. Mascot together with SEQUEST[37] are commonly used search tools in proteomics today. However, there are many others including Protein prospector,[38] ProbID,[39] X!Tandem;[40] OMSSA,[41] ProSight[42] and Inspect[43] (see Nesvizhskii et al. for a review[14]). Mascot takes a probability based approach to match sequences from a database to tandem mass spectra.[36] Because it is a commercial program the exact algorithms it employs are neither known nor available for modification. Furthermore, Mascot is implemented in a client-server configuration, which imposes practical restrictions for some applications such as real-time searches. We therefore set out to develop a new search engine that would be free of these restrictions. We aimed at performance at least on par with Mascot, which has become a "gold standard" in proteomic analysis, and robustness for scaling up to extremely large and complex data sets. In combination with MaxQuant, the new search engine would then enable analysis of complex data sets on desktop machines by any proteomics researcher or biologist wishing to employ proteomics.

Database searching with fragment mass spectra typically follows one of three approaches:[44,45] (i) deriving a partial or full peptide sequence with associated mass information (first implemented by PeptideSearch[46] and graph theory based *de novo* methods[47]), (ii) autocorrelation between the experimental and a calculated spectrum (first used in SEQUEST) or (iii) calculating a probability that the observed number of matches between the calculated and measured fragment masses could have occurred by chance (pioneered in Mascot). We chose the probability based approach based on the binominal distribution probability and started from a score that we had originally developed for analyzing $MS^3$ data for which no search software was available at the time.[48] This score has already been used for ranking the peptides in MaxQuant searches from the beginning and it also determines the localization probability of modifications in peptides.[48]

In this paper, we describe the architecture of the Andromeda search engine and its scoring function. We perform a rigorous comparison against the Mascot search engine on several large-scale data sets. The ability of Andromeda to accurately handle many modifications of the same peptide is demonstrated. Due to the complexity of peptide mixtures in shotgun proteomics and the relatively low resolution of precursor isolation, two peptides are frequently 'cofragmented' and there are algorithms that try to identify them from mixture spectra.[49−52] We demonstrate the flexibility of the Andromeda search engine by implementing a novel second peptide identification algorithm.

## ■ MATERIALS AND METHODS

### Benchmark Data Sets

Raw data from 84 LC−MS runs was taken from Luber et al,[53] a label-free proteome study of mouse dendritic cells to a depth of 5780 proteins. Cell subpopulations were obtained by FACS sorting, proteins were separated by 1D SDS-PAGE and digested with trypsin. Peptides from the gel pieces were analyzed on a nanoflow HPLC system connected to a hybrid LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific).

As a phosphoproteomics benchmark data set we took the raw data from 117 LC−MS runs produced in a phosphatase knockdown analysis.[54] Drosophila Schneider SL2 cells were differentially SILAC labeled as pairs with Lys-8/Arg-10 and Lys-0/Arg-0. Proteins were separated by 1D SDS-PAGE and digested with trypsin or in solution digested without gel separation. Peptides were subjected to $TiO_2$ chromatography and strong cation exchange chromatography and analyzed on a nanoflow HPLC system connected to a hybrid LTQ-Orbitrap (Thermo Fisher Scientific). For the analysis, we used only those MS/MS spectra that were acquired on a recognized SILAC pair. Modifications due to labeling with Lys-8 and Arg-10 can then be taken as fixed.

The benefits of second peptide analysis were investigated using data that was acquired on an LTQ-Orbitrap Velos. Briefly, HeLa cell lysate was in solution digested with trypsin, the peptide mixture was separated on a nanoflow HPLC system and analyzed using a data-dependent "top 10" method. Several runs were acquired with varying isolation windows. The precursor ions were isolated in selection windows of 1, 2, 4, 8, 16, and 32 Th followed by HCD fragmentation and high resolution data acquisition of the MS/MS spectra in the Orbitrap.

### Data Preparation

MaxQuant, version 1.1.1.25, generated peak lists from the MS/MS spectra for the database searches. For the low-resolution MS/MS spectra recorded in "centroid" mode the 6 most abundant peaks per 100 Th mass intervals are kept for searching. High-resolution profile MS/MS data is deconvoluted (deisotoping and transfer of all fragment ions to single charge state) before extraction of the ten most abundant peaks per 100 Th. All statistical filters in MaxQuant like peptide and protein false discovery rates and mass deviation filters were disabled in order to score all submitted MS/MS spectra. Peptide masses were recalibrated by MaxQuant prior to both Andromeda and Mascot searches. For the Mascot search (using Mascot server version 2.2.04), peak lists written out by MaxQuant were converted to mgf format, the standard Matrix Science data format. Oxidation of methionine and N-terminal protein acetylation were used as variable modifications for all searches. A mass tolerance of 6 ppm was used for the peptide mass. To make Mascot and Andromeda searches comparable, we did not use the individual peptide mass tolerances in MaxQuant. A tolerance of 0.5 Th was used for matching fragment peaks produced by CID. The HCD fragment ion data used in the co-fragmentation study were searched with a 20 ppm window in Andromeda. A maximum of two missed cleavages were allowed in all searches. The "instrument" parameter was set to "ESI-TRAP" in the Mascot search. Mascot and Andromeda scores were matched to each other based on raw file name and scan number.

The search was performed against a concatenated target-decoy database with modified reversing of protein sequences as described previously.[31] Mouse and human data was searched against the respective IPI databases,[55] version 3.68, while the drosophila data was searched against protein sequences from flybase[56] version 5.24.

### Search Engine Configuration

In Andromeda, the user specifies allowed peptide and protein modifications, enzymes used for protein cleavages and the protein sequence databases to be searched in the program AndromedaConfig.exe. Modifications are specified by their elemental composition. Neutral losses and diagnostic ions can be specified separately for each type of amino acid with the modification in question. Modifications that are interpreted as labels by MaxQuant can be defined here, such as SILAC labels. Searches with semispecific enzymes are supported as well, where

Table 1. Most Important Regular Expressions Defining How Protein Identifiers Are Extracted from the Headers of Fasta File Entries

<table>
<tr><th>regular expression</th><th>description</th></tr>
<tr><td>>(.*)</td><td>Everything after “>”</td></tr>
<tr><td>>([^ ])</td><td>Up to first space</td></tr>
<tr><td>>IPI:([^\| .]*)</td><td>IPI accession</td></tr>
<tr><td>>(gi\|[0−9]*)</td><td>NCBI accession</td></tr>
<tr><td>>([^\t]*)</td><td>Up to first tab character</td></tr>
<tr><td>>.*\|(.*)\|</td><td>Uniprot identifier</td></tr>
</table>

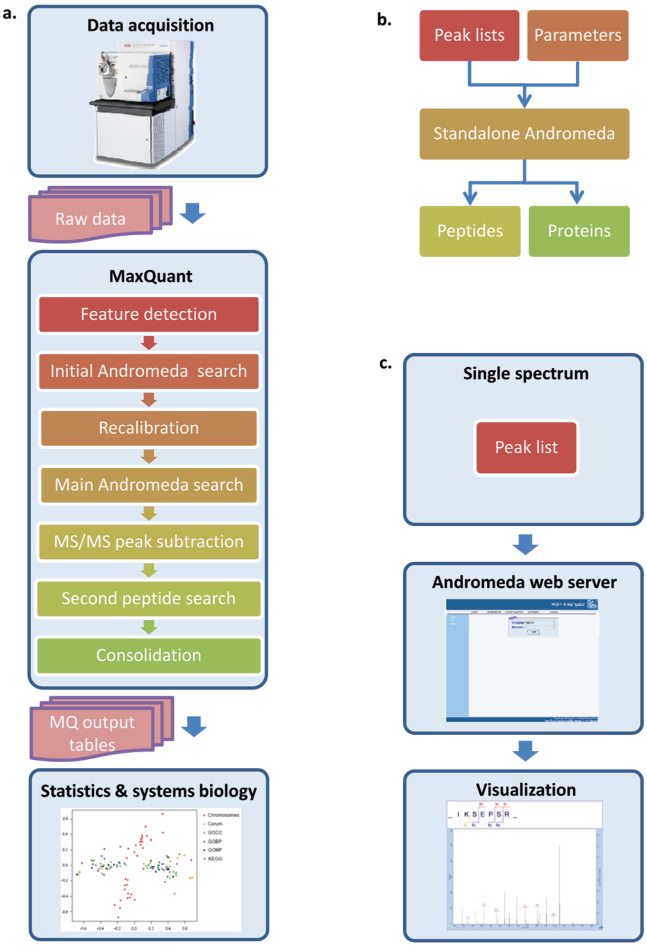


Figure 1. Three Andromeda configurations: (a) integrated in MaxQuant, (b) standalone search engine, and (c) web server.

only one peptide terminus needs to be a cleavage site according to the given protease digestion rule while the other terminus can be an arbitrary position in the protein. An unspecific search is also supported where both of the peptide termini can be arbitrary positions in a protein. Parse rules for regular expressions as defined in the Microsoft .NET framework (msdn.microsoft.com/en-us/library/az24scfc.aspx) are used to define how a protein identifier is extracted from the header line of a FASTA database file entry. Some of the most important regular expressions can be found in Table 1.

## Input and Output Formats

Input files for peak lists and parameter values as well as output files for peptide identifications and a tentative protein list are all human-readable text files. Parameter files have the ending “.apar” and contain a list of key-value pairs where each pair is separated by a “=” sign. Expressions used for modifications, labels, enzymes and databases must have been defined previously in the AndromedaConfig.exe program. Peak list files have the extension “.apl” and can consist of arbitrarily many spectra, one following the other, each spectrum entry being enclosed by “peaklist start” and “peaklist end” lines. Some key-value pairs with peaklist-specific parameters are followed by two columns of numbers containing the $m/z$ and intensity values. The peptide result files (“.res”) contain up to 15 candidate peptide matches for each peak list. For each candidate the peptide sequence, modification state, score, mass, mass deviation and all corresponding protein IDs are given.

## Software Availability

MaxQuant with Andromeda as the integrated search engine can be downloaded from www.maxquant.org. A standalone version of Andromeda is available at www.andromeda-search.org. The source code is provided as Supporting Information 1. Both applications require Microsoft .NET 3.5, which is either already installed with Microsoft Windows or can be installed as a free Windows update. The Andromeda web server can be accessed at www.biochem.mpg.de/mann/tools/ for a limited number of submissions of MS/MS spectra. Andromeda has been written in the programming language C#, using the Microsoft .NET framework version 3.5.

# ■ RESULTS

Andromeda is a search engine based on a probability calculation for the scoring of peptide−spectrum matches. A version of it is fully integrated into the MaxQuant quantitative proteomics platform. Hence, all the data processing from the acquired raw data to the list of quantified peptides and proteins can be performed in a single end-to-end workflow (Figure 1a). In addition to the regular search Andromeda can be used in different contexts: for example in MaxQuant it is used for determining the mass-dependent recalibration function based on a preliminary database search, and for the identification of one or more cofragmented peptides (see below). We also provide a standalone version of Andromeda that produces scored peptide candidates, given a collection of MS/MS peak lists and a parameter file (Figure 1b). In this option, many of the statistical processing algorithms that are part of MaxQuant are not applied to the data and the reported list of identified proteins is only tentative without rigorous control of protein false discovery rate (FDR). The output consists of a raw list of scored peptide candidates per spectrum together with the protein list. Furthermore, there is a web server version of Andromeda for the submission of a limited set of spectra (Figure 1c), www.biochem.mpg.de/mann/tools/. In addition to the scoring results of the 15 best peptide candidates, the annotated spectrum can be inspected for the highest scoring and all other candidate peptide sequences. Despite these alternative uses, we anticipate that Andromeda will most commonly be employed as the search engine for MaxQuant.

## Indexing Peptides and Proteins

To efficiently score an MS/MS spectrum it is important to be able to quickly retrieve all candidate peptides that have a suitable calculated precursor mass within a given tolerance. First we generate a list of all peptides obtained by the specified digestion

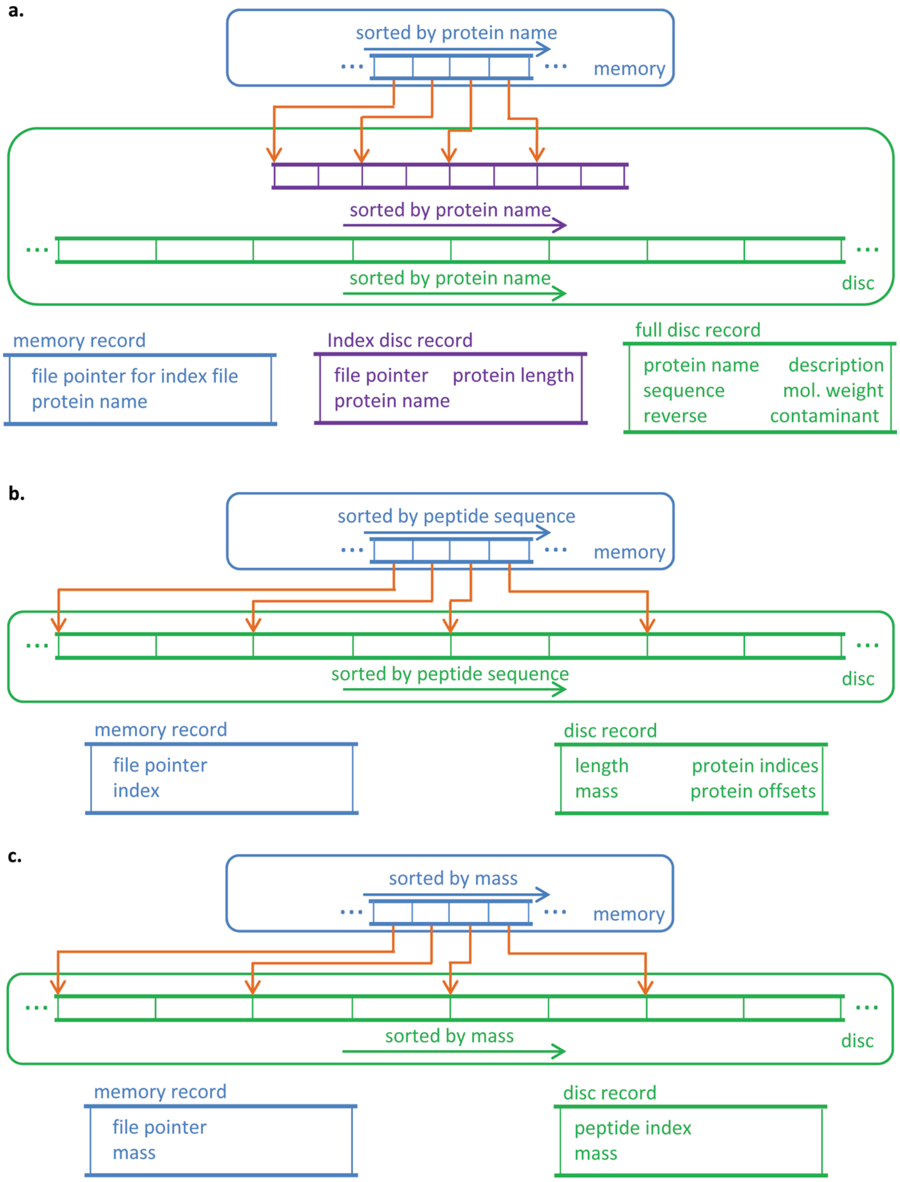


**Figure 2.** Memory and disk structure. (a) Protein list has a two-layer index structure. One small index is kept in memory whose entries point to blocks of multiple entries in the secondary index that is kept on disk. Each entry of the disk index points to the position of the protein entry in the file containing the complete information for each protein including the amino acid sequence. The protein lists are sorted alphabetically by the protein names. (b) Peptide index that resides in memory points to equally sized blocks of peptide entries, which are kept on disk. (c) Similar structure for the list of all combinations of peptide sequence and variable modifications. Index and disk entries are sorted by the peptide mass to allow for quick retrieval of all peptide candidates within a given mass interval.

rule from the protein sequences considering all possible combinations of preset variable modifications. At this stage we are only interested in the peptide masses, therefore only the number but not the positions of the modifications are important. The list of all of these peptides is sorted by mass for quick search access, which only grows slowly with increasing size (proportional to the log of the number of peptides for a binary search). The number of peptides with specific modifications can become very large, either when searching in an extended protein sequence database or by specifying many variable modifications. One common setting is to search the human IPI database including reverse sequences and common contaminants digested with trypsin and allowing for up to two missed cleavages. The number of modifications to consider can also grow rapidly. For example, in a phosphoproteomic experiment with triple SILAC labeling of lysine and arginine, one may simultaneously deal with phosphorylation of serine, threonine and tyrosine, Lys4, Lys8, Arg6, Arg10 and oxidation of methionine as variable modifications. (This is the case for those MS/MS spectra where the SILAC state could not be determined prior to the database search; otherwise the modification state of Arg and Lys are set by MaxQuant.) For the human IPI database and including the reversed sequences, this corresponds to a list of 174 618 protein sequences resulting in 7 837 653 peptide sequences and 76 937 183 modification-specific peptides (without taking modification positioning into account). These numbers can become even larger, for example in cases where one wants to search against a six-frame translation of the whole genome.

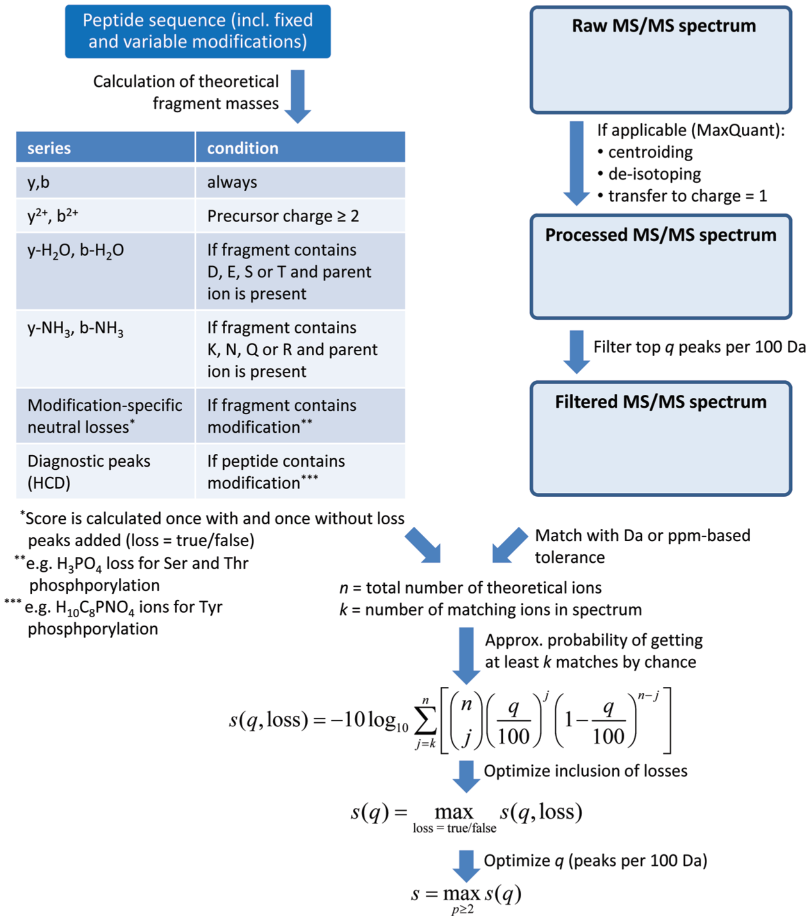

Peptide sequence (incl. fixed and variable modifications)
Calculation of theoretical fragment masses
series | condition
y,b | always
$y^{2+}$, $b^{2+}$ | Precursor charge ≥ 2
y-$H_2O$, b-$H_2O$ | If fragment contains D, E, S or T and parent ion is present
y-$NH_3$, b-$NH_3$ | If fragment contains K, N, Q or R and parent ion is present
Modification-specific neutral losses* | If fragment contains modification**
Diagnostic peaks (HCD) | If peptide contains modification***
*Score is calculated once with and once without loss peaks added (loss = true/false)
**e.g. $H_3PO_4$ loss for Ser and Thr phosphporylation
***e.g. $H_{10}C_8PNO_4$ ions for Tyr phosphporylation
Raw MS/MS spectrum
If applicable (MaxQuant):
• centroiding
• de-isotoping
• transfer to charge = 1
Processed MS/MS spectrum
Filter top $q$ peaks per 100 Da
Filtered MS/MS spectrum
Match with Da or ppm-based tolerance
$n$ = total number of theoretical ions
$k$ = number of matching ions in spectrum
Approx. probability of getting at least $k$ matches by chance
$s(q,\text{loss}) = -10\log_{10}\sum_{j=k}^{n}\left[\binom{n}{j}\left(\frac{q}{100}\right)^{j}\left(1-\frac{q}{100}\right)^{n-j}\right]$
Optimize inclusion of losses
$s(q) = \max_{\text{loss = true/false}} s(q,\text{loss})$
Optimize $q$ (peaks per 100 Da)
$s = \max_{p\geq 2} s(q)$

**Figure 3.** Schematic of the peptide scoring algorithm. The upper left branch shows the calculation of the theoretical fragment ion masses while the right branch indicates the processing of the experimental MS/MS spectra. In particular, all ion types that are used for the scoring can be found in the table on the left. The final score involves an optimization of the number of highest intensity peaks that are taken into account per 100 Da $m/z$ interval and over the inclusion of modification-specific neutral losses.

We therefore wished to be able to handle protein sequence information without limitation on the sizes of calculated protein and peptide lists. Our goal was to work within the memory limits of 32-bit operating systems, which is around 1.6 GB from within the Microsoft .NET framework. The data structures for the search engine have to have an even smaller memory footprint since other data might be required to be in memory at the same time. Obviously the full modification-specific peptide list is too large to keep in memory and it has to reside on the hard disk (or solid state disk for improved performance). This is also true for the peptide and protein lists because unlimited scalability is desired. Only an index for each of the files is kept in memory, which contains positions of the records relative to the beginning of the file. These memory indices can already exceed the memory limitations for very large numbers of peptides. Therefore the index points to beginnings of blocks of elements in the file with a suitably chosen block size such that the lengths of the indices in memory never exceed a fixed size. In Figure 2, the structure of these lists and the relationships between memory and files residing on the hard disk are shown for proteins, peptides and modification-specific peptides. The records always contain indices to the respective items in the hierarchy above, assuring easy navigation from a candidate peptide to all the proteins that it occurs in. The modification-specific peptide list is the one that is directly accessed in database searches. It is sorted by mass, which allows quick retrieval of peptides within the given mass window. Protein and peptide list are instead sorted alphabetically by protein name and peptide sequence, respectively.

## Scoring Model

The probabilistic score employed in Andromeda is derived from the p-score that was introduced for the identification of $MS^3$ spectra.[48] Given a peptide sequence together with a configuration of fixed and variable modifications for that peptide, first the theoretical fragment ions are calculated (Figure 3). For CID and HCD the list of theoretical fragment ion masses always contains the singly charged b- and y-ions. If the precursor charge is greater than one, the doubly charged b- and y-ions are added. In case of low resolution ion trap MS/MS spectra the charge state of fragments usually cannot be determined. The calculated doubly charged $m/z$ values are then added explicitly if it is desired to match more highly charged fragments. For high-resolution MS/MS the charge state can be assigned to a fragment if more than one isotopic peak is detected. For these cases we remove peaks of fragments with charge higher than 1 from the spectrum and reintroduce them into the spectrum as singly charged fragment ions. If there are several charge states for a fragment their intensities are added, taking account of the fact that signal is proportional to charge in the Orbitrap analyzer. We noticed that even for high-resolution MS/MS data, where charge state detection is possible in general, it is beneficial to consider doubly charged b- and y-ions as well. This is because for lower mass fragments sometimes only the monoisotopic peak is detectable precluding charge state determination and hence also the transformation to charge state one. For example assuming that the elemental composition of fragments follows the averagine model[57] the ratio between the $^{13}$C and monoisotopic peak intensities for a fragment of 400 Da is 4.6:1. For less abundant fragments this can obviously lead to nondetection of the $^{13}$C peak while the monoisotopic peak is above the noise level.

Calculated peaks corresponding to water and ammonia losses are only offered for matching as singly charged ions in those cases where the main b- and y-ion fragment is present and contains the amine-, amide- or hydroxyl-containing amino acid side-chains that tend to lead to the respective side chain loss. Modification-specific losses are configurable in the program AndromedaConfig, which is included in the MaxQuant distribution. The above-mentioned modification-specific neutral losses, as well as ions that are diagnostic for the presence of a particular modification of an amino acid type can be freely configured there. For example, the loss of phosphate from a phosphorylated serine or threonine is much more likely than from a tyrosine, which instead produces a highly specific immonium ion at mass 216.0426 (see, e.g., Steen et al.[44]). If Andromeda is used within MaxQuant, the report for each modification site includes presence or absence of a diagnostic peak in the MS/MS spectrum. The score is calculated once including configurable neutral losses and once excluding them and the maximum of the two scores is chosen. (Note that all scoring procedures are carried out identically for sequences from the reverse database, so they do not introduce a bias.)

The first step in the actual calculation of the score is to count the number of matches $k$ between the $n$ theoretical fragment masses and the peaks in the spectrum. The higher $k$ is compared to $n$, the lower the chance that this happened by chance.[48] Because there are many signals in MS/MS spectra, including many low intense noise signals, the number of peaks in a defined mass interval—here 100 Th, which is the typical distance between consecutive members of fragment series (average mass of amino acids)—are limited to a maximum number. The parameter $q$ is defined as the number of allowed peaks in the mass interval and it is needed to calculate the probability of a single random match. If the difference between calculated and measured masses is less than a predefined value, a match is counted. This can be done with an absolute mass tolerance window specified in Th or a relative mass window specified in ppm. While the former is appropriate for ion trap spectra, the latter is more suitable for high-resolution FT-ICR or Orbitrap spectra.

The Andromeda score is calculated as −10 times the logarithm of the probability of matching at least $k$ out of the $n$ theoretical masses by chance as shown in Figure 3. This is slightly different from Olsen et al.,[48] where the probability of matching exactly $k$ out of $n$ theoretical masses is determined. The formula used here is more similar to a definition of a $p$ value for the null hypothesis that there is no similarity between the theoretical mass list and list of the spectrum masses. In particular, the score has the desirable property to vanish for $k = 0$. The calculation of the probability is only approximate since the probability for a single random match is taken to be $q/100$, which is exact if there was only one possible match per nominal mass. For high resolution MS/MS data the true random match probability is considerably less than this and the true score would be higher but more complicated to calculate. However, this simplification is conservative as it decreases the calculated score and is justified by the excellent performance of the search algorithm on high-accuracy MS/MS data.

The intensities of the peaks in the MS/MS spectra are indirectly taken into account by calculating the score for all values for $q$ (number of peaks per 100 Th) up to the specified maximum. The best of these scores for varying $q$ is selected. Therefore two spectrum-sequence comparisons with the same values for $n$ and $k$ can result in different scores depending on the intensities of the matched peaks. Generally, the score is higher if the matches are among the more intense peaks because the optimal value of $q$ will be lower (see formula in Figure 3). However, we have found it crucial that this intensity weighting is not done on the overall intensity scale over the whole spectrum, but that it is restricted to local mass regions (e.g., the 100 Th mass range intervals.). This compensates for underlying global peak density distributions which typically favor small fragment masses.

The inclusion of additional information like peptide length, number of modifications or of missed cleavages can aid the specificity of peptide assignments to spectra. Ideally this is done in a data-dependent manner in which different weights for different classes of peptides can be derived from the data by machine learning in a Bayesian framework. We wished to include such a weighting of peptide classes into the score while retaining a basic search engine score that is deterministic and only depends on the spectrum being scored rather than the ensemble of all other spectra. To capture the dependence of the score on peptide mass and on the number of modifications we introduced a fixed additive component to the Andromeda score, which depends on the number of modifications and is a linear function of the mass. The specific values are determined in a manner that adjusts the distributions of reverse hits from a target-decoy search so that they become equal. The net effect of this procedure is to minimize the FDR for a given cutoff value, because it does not depend on peptide mass and modification state any longer. We used a large data set of MS/MS spectra and incorporated the specific weights into the scoring function. A data-dependent Bayesian scoring can still be applied to the output of the Andromeda search engine. For instance, MaxQuant additionally performs a peptide length dependent Bayesian analysis in a data dependent manner.[31]

### Comparison to the Mascot Search Engine

Mascot[36] is a widely used standard for database searching and most other search engines have been compared to Mascot. Therefore we investigated how Andromeda compares to Mascot in terms of scoring of peptide-spectrum matches. As the exact details of the Mascot scoring system are not known, we compared the performance of Andromeda vs Mascot empirically on very large sets of proteomic data.

In Figure 4a, we plot the Mascot score against the Andromeda score for a data set of 732 287 MS/MS spectra derived from a label-free mouse proteome measurement as described in Materials and Methods. For each MS/MS spectrum the highest scoring peptide is taken which is not necessarily the same for the Mascot and the Andromeda scoring. In Figure 4b, the fraction of cases for which the top-scoring Andromeda and Mascot peptide sequences coincide is displayed as a histogram depending on the Andromeda score. As can be seen, above an Andromeda score of 100 the top-scoring peptides coincide in almost all cases.

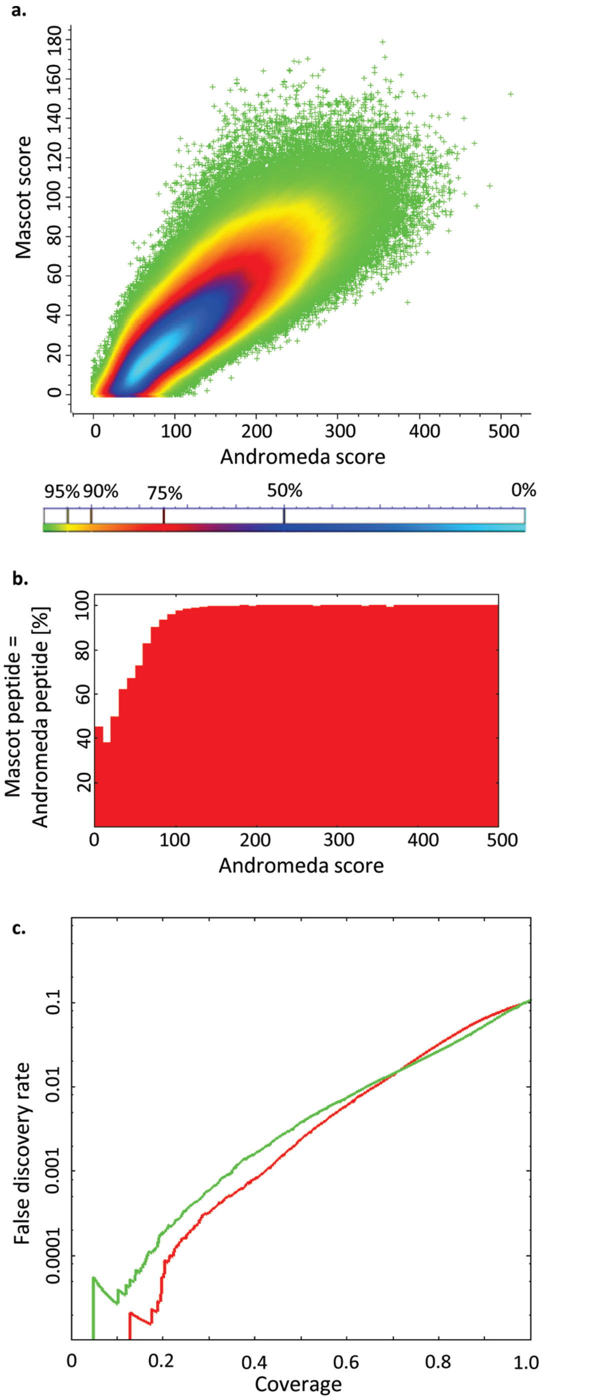


**Figure 4.** (a) Andromeda vs Mascot score for a data set of 732 287 MS/MS spectra derived from a label-free mouse proteome measurement.[53] The score for the top-scoring peptide for each MS/MS spectrum is shown which is not necessarily the same peptide sequence for the Mascot and the Andromeda identification. The color code indicates the percentage of points that are included a region of a specific color. (b) Histogram of the percentage of cases in which the top-scoring Andromeda and Mascot peptide sequences are equal as a function of Andromeda score. For the comparison leucine and isoleucine were treated as the same amino acid. (c) False discovery rate as a function of coverage for the same data set calculated based on the reverse hits from the target-decoy search.

Of the recorded MS/MS spectra, 89.1% correspond to unmodified peptides and most of the identified modified peptides have an oxidized methionine. The point density is indicated by the color code in Figure 4a which encodes the percentage of points that are included a region of a specific color. For example, the yellow line in Figure 4a encloses 95% of all data points. This visualization allows the visual detection of outliers (like a two-dimensional data plot), while at the same time retaining information about the density of points that would normally only be visible in a 3D data plot. It is immediately apparent from the figure that the scores correlate well overall. There are no distinct populations of peptides that are only identified by one of the search engines. A linear regression results in the equation $M = 0.311 * A - 32.231$, where $M$ is the Mascot score and $A$ the Andromeda score, with an $R^2$ value of 0.708. This indicates that Andromeda scores are generally about 3-fold larger than Mascot scores. However, this does not indicate a 3-fold larger confidence. The statistical power is better determined by calculating coverage and false discovery rates as a function of score threshold as is done below. A rough conversion between Andromeda and Mascot scores can be performed by a division by three or application of the regression line. Note that there are only very few and dispersed outliers on either side; of the order of tens of spectra out of the total of more than 700 000. Furthermore, there are virtually no high-scoring outliers near either axis, indicating an absence of spectra that were ranked highly with one method but scored close to zero with the other. This demonstrates that no populations of peptides would be lost entirely by employing one score or the other.

Next we compare the performance of the Andromeda and Mascot search engines as a function of False Discovery Rates estimated as the number of hits from the reverse database divided by the number of forward hits at any given minimum score. The sensitivity of the database search is defined as the number of accepted forward hits relative to the total number of forward hits at the same score. Mascot and Andromeda have very similar characteristics over the whole range of FDRs, in particular including the often used 1% FDR rate (Figure 4b). This shows that the two scores are very close in discriminatory power.

## Scoring of Phosphopeptides

Figure 5a shows the same type of plot as in Figure 4a but for a data set that is enriched for phosphopeptides. Of the recorded 586 883 MS/MS spectra in Figure 5a, 27.4% have one or more phosphorylations. Outliers are visible in the region of high Andromeda and low Mascot score and most of them correspond to peptides with three to five phosphorylation events. Figure 5b displays the MS/MS spectrum of a peptide with five phosphorylation sites that has a Mascot score of 5.2 and an Andromeda score of 199.3. The y-series coverage is almost complete with most fragments occurring with a neutral loss of a phosphate molecule. An FDR coverage curve for the phosphopeptide data set is depicted in Figure 5c. The performances of Mascot and Andromeda are similar over the entire range with an advantage for Andromeda in the high specificity region. At the typical operation point of 1% FDR results are very close. We speculate that the better scoring in the region of higher specificity may be due to a better matching of spectra of phosphopeptides in Andromeda due to more comprehensive combinatorics of positioning of phospho-groups on the available serine, threonine and tyrosine sites in the peptide sequences, including a more complete offering of neutral losses. During the Andromeda search we offer up to 1000 positionings of variable modifications within any given peptide which is exhaustive for most situations.

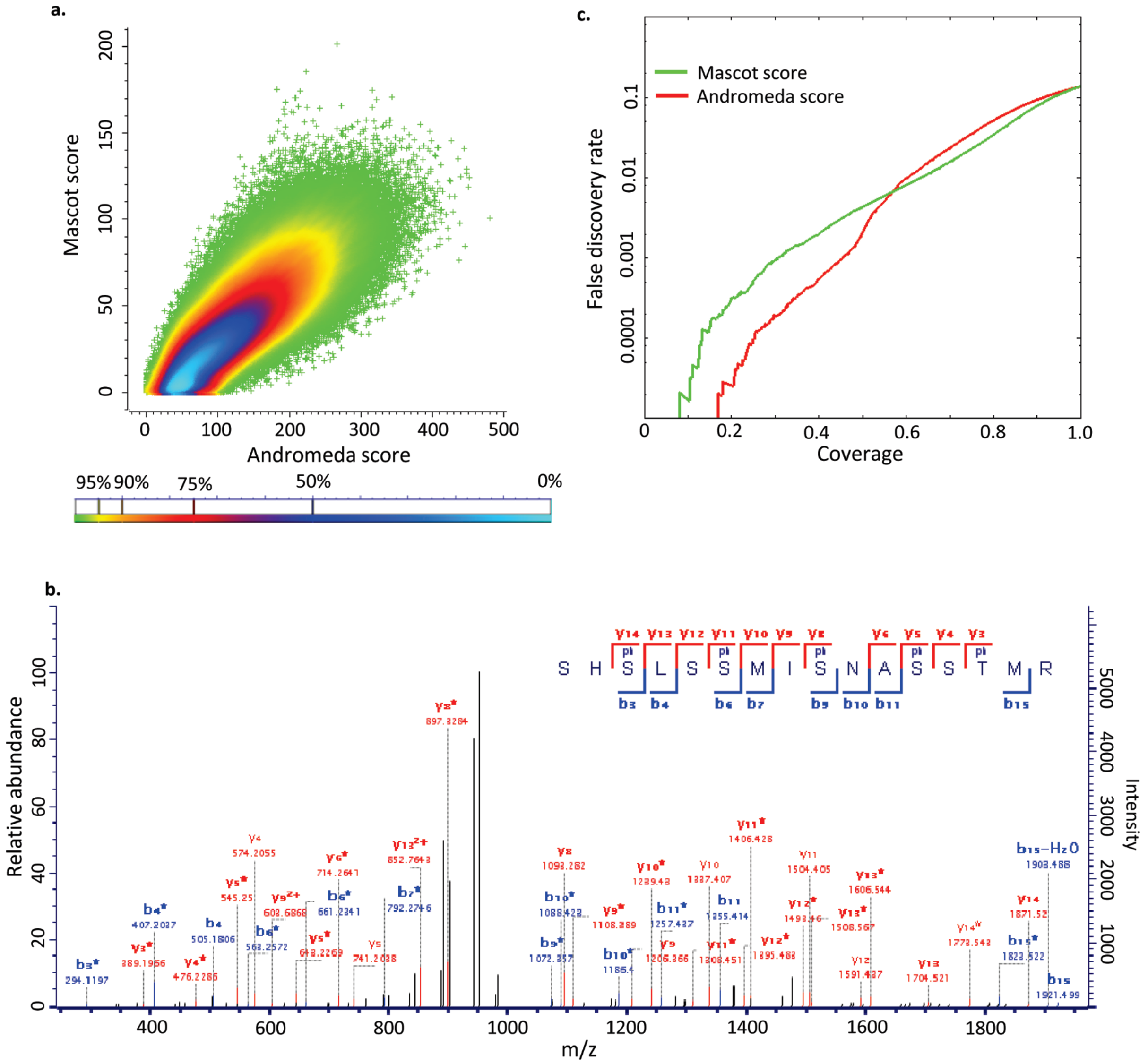


**Figure 5.** (a) Andromeda vs Mascot score for 586,883 MS/MS spectra from the phospho-proteome data by Hilger et al.[54] (b) Annotated MS/MS spectrum of the peptide SHpSLSpSMIpSNApSSpTMR. Mascot and Andromeda produce the same top-scoring peptide sequence with a Mascot score of 5.2 and an Andromeda score of 199.3. (c) False discovery rate as a function of coverage for the same data set calculated in the same way as in Figure 4c.

In MaxQuant, the top-scoring peptide is furthermore rescored with essentially exhaustive positioning of modifications. We merely restrict the combinatorics to 100 000 possibilities to exclude the rare instances where single peptides cause long calculation times due to "combinatorial explosion". In Supplementary Figure 1 (Supporting Information), the same data as in Figure 5a is shown six times—each time highlighting another population of top-scoring peptides with a fixed number of phosphorylations. Peptides with higher phosphorylations tend to have many data points in the high Andromeda score but low-to-moderate Mascot score region further indicating that Andromeda performs better on highly phosphorylated peptides.

## Identification of Second Peptides

Even in high-resolution MS, the selection of the precursor ion for fragmentation is always performed with low resolution (typically a few Th) to ensure adequate sensitivity for MS/MS. In complex mixtures, this results in frequent cofragmentation of coeluting peptides with similar masses. These 'chimerical' MS/MS spectra[52] can be detrimental for identification of the peptide of interest, especially if the cofragmented peptide is of comparable intensity. Co-fragmentation generally reduces the number of peptides identified in database searches and poses special problems for reporter fragment based quantification methods because both peptides contribute to the measured ratios.

However, this situation can be turned to an advantage if both peptides can be identified. In particular, this presents the opportunity to identify peptides that have not been targeted for MS/MS and to obtain two or more peptide identifications from a single MS/MS spectrum. Although this problem has been addressed before,[49−52] to our knowledge it has not been adopted in mainstream search engines yet. Here we describe a second peptide identification algorithm that we have integrated into the Andromeda/MaxQuant workflow.

To illustrate the principles of our algorithm, Figure 6a shows an LC−MS map, where 3D peaks are indicated as lines marking the peak boundaries. The blue isotope pattern has been selected for fragmentation at the position of the cross on the

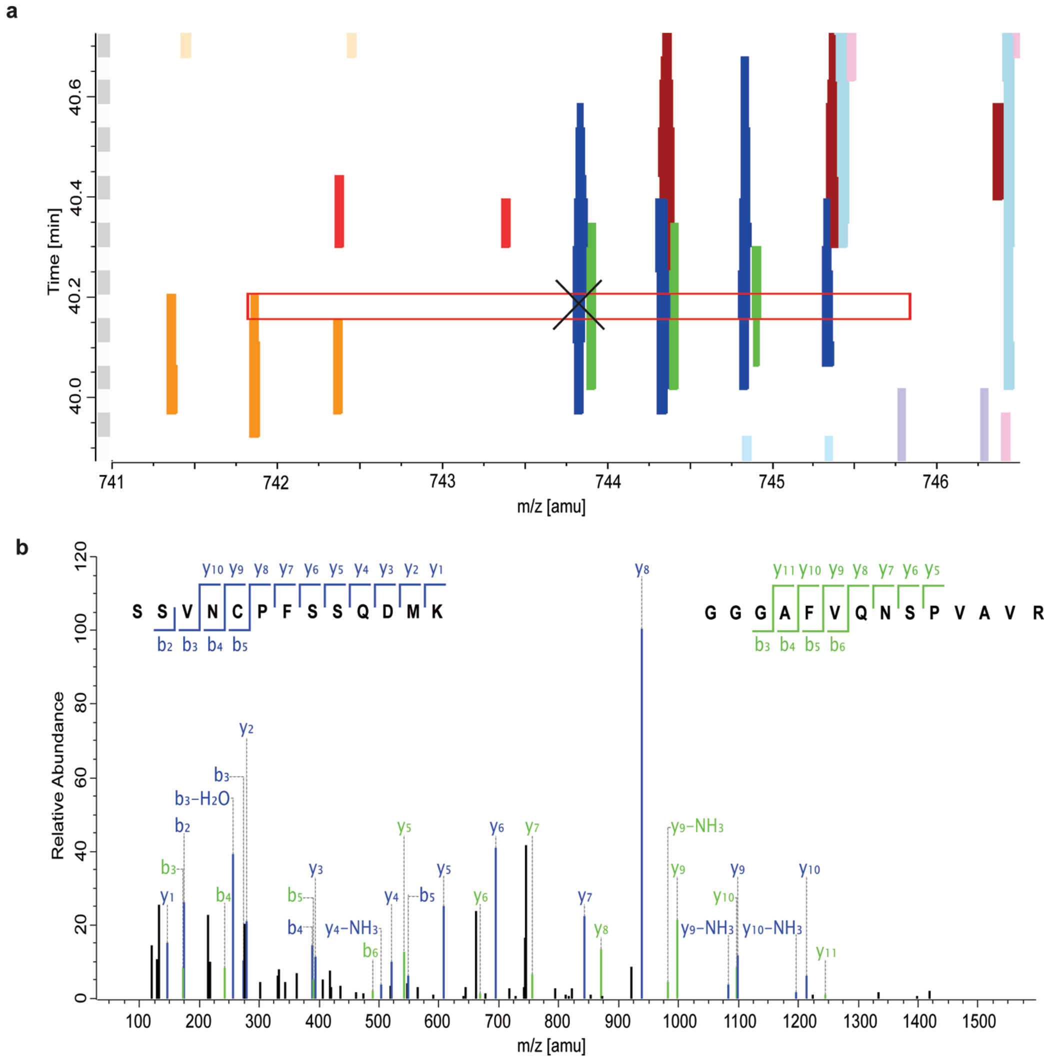


**Figure 6.** Second peptide identification. (a) LC−MS map of the sequenced (blue) and cofragmented (green) peptide described in the main text. The blue peptide has been selected for fragmentation at the position of the cross. The red rectangle indicates the isolation window. (b) MS/MS spectrum leading to the identification of both peptides. Fragments of the two peptides are indicated in blue and green, respectively. The blue peptide is identified in the conventional database search while the green peptide has been identified as "second peptide".

monoisotopic peak of that peptide. The red rectangle indicates the region from which ions have been isolated for fragmentation. Clearly the peptide corresponding to the green isotope pattern that has not been selected for sequencing intersects with the isolation rectangle. Therefore its fragments should be present in the MS/MS spectrum as well. The actual fragment spectrum is shown in Figure 6b where the fragments originating from the targeted and identified peptide (blue isotope pattern) are indicated in blue. This process is repeated for the entire LC−MS/MS run. For every 3D MS isotope pattern that has not been selected for sequencing the algorithm checks whether it intersects with the isolation window of any MS/MS spectrum. If this is the case then the fragments in this MS/MS spectrum that have already been assigned to a peptide sequence during the main Andromeda search are subtracted. The remaining fragments are submitted to a new database search with the precursor mass from the peptide that was not targeted for MS/MS. The collection of these "subtracted" peak lists is submitted to Andromeda in the same way as in a conventional search. However, the results of this second peptide search are further processed with their own peptide length based posterior error probability and precursor mass filtering. Since these spectra are on average of lower quality than the original MS/MS spectra we have found it to be crucial that they have their own data-dependent statistical model for peptide identification. The resulting peptides are then accepted up to a 1% FDR and integrated into the usual protein identification and quantification workflow.

The HCD data set used for testing (see Materials and Methods) was acquired with a total isolation width of 4 Th for every MS/MS spectrum. The identification rate for the set of second peptide spectra is much lower compared to the normal MS/MS identification rate of 50%. Nevertheless, since the number of the second peptide spectra is quite high compared to normal MS/MS spectra considering cofragmentation still leads to a considerable increase in peptide identifications. In our example, the number of identified peptide features increased by 10.7% by the inclusion of second peptide identifications. The gain in the number of identified peptides depends on the isolation width for the acquisition of MS/MS spectra. For instance, at an isolation width of 2 Th we observe that the increase in identified peptides through second peptide identifications is only 5.7%. The relative gain is larger at increased isolation width because the average number of additional peptides within the window increases. However, the chance to identify the main peptide decreases due to the mixing of the spectrum with fragments from other peptides. The dependence of the number of peptide identifications for conventional and second peptides is shown in Supplementary Figure 2 (Supporting Information).

## DISCUSSION AND OUTLOOK

Here we have described Andromeda, a novel search engine for matching MS/MS spectra to peptide sequences in a database. Andromeda can either be used in a stand-alone mode or—more typically—as part of the MaxQuant environment. Apart from an optimal scoring model our intention was to develop a very robust architecture with unlimited scalability. We have demonstrated this on large scale data sets with hundreds of thousands of spectra. Andromeda has been "stress tested" in ongoing studies and has been the default search engine in our laboratory for some time. A practical advantage of the MaxQuant/Andromeda combination is that it runs locally on the user's computer. This eliminates client-server set up and communication issues. The computational proteomics pipeline starting from raw data files to reported protein groups and their quantitative ratios now appears unified to the user. Despite the local search architecture, processing speeds are generally not different from the previous MaxQuant/Mascot environment in which Mascot was run on an external server. Furthermore, we have added a separate module called Perseus (www.maxquant.org), which performs bioinformatic analysis of the output of the MaxQuant/Andromeda workflow. Perseus is already available and in use[58] and completes the pipeline for computational proteomics analysis but will be described in a future publication (Cox et al., in preparation).

The scoring function at the heart of Andromeda is built on a simple binominal distribution probability formula (Figure 3), which we have previously used in scoring $MS^3$ spectra and localizing PTMs.[59] Andromeda divides the MS/MS spectrum into mass ranges of 100 Th. In each of these ranges the number of experimental peaks offered for matching is dynamically tested in an intensity prioritized manner.

False discovery rates for the same initial probability score can still depend on the number of modifications and on the mass of the peptide. This is accounted for in Andromeda by an additive component to the score. Comparison to Mascot on very large data sets reveals very few outliers—in particular almost no peptides are exclusively identified by one of the two search engines. Furthermore, the coverage of identified peptides at any given FDR is likewise similar, including at the generally used operating point of 1% expected false positives. We did notice improved identification of heavily modified peptides in Andromeda compared to Mascot, which we attribute to the more exhaustive combinatorial analysis of placing PTMs on all possible amino acids. As the Mascot search engine has become one of the standards in proteomics, equivalent performance fulfills the goal that we had set for the development of Andromeda and likely implies favorable comparison to other search engines as well. Apart from describing the score we have also made the actual code used in Andromeda available for inspection with this publication (Supporting Information 1).

A key advantage of Andromeda is its extensibility. For example, proteomics with high accuracy MS and MS/MS data (high−high mode[60]), is becoming increasingly common. Andromeda, in contrast to Mascot, allows arbitrarily accurate MS/MS requirements specified in ppm. Similarly, Mascot precludes identification of SILAC pairs if the same amino acid can bear a fixed and a variable modification. This causes a substantial loss of quantification information, for example in the analysis of lysine acetylated peptides[61] because all MS/MS spectra of lysine-acetylated peptides that were sequenced on the heavy SILAC partner will not be identified by Mascot. All these quantitative ratios are retrieved in the MaxQuant/Andromeda workflow.

More generally, additional scoring modes can be added to Andromeda. We demonstrated this by implementing a second peptide identification algorithm into the MaxQuant/Andromeda workflow. For each isotope cluster that is detected in the LC−MS data but that was not targeted for fragmentation the algorithms checks if the precursor isotope pattern intersects the selection window of any MS/MS event. If so, fragment ions belonging to the identified peptide are subtracted and the search is repeated with the cofragmented peptide in a statistically rigorous way. As demonstrated here, this leads to an appreciable increase in peptide and protein identifications in complex mixtures. As another example, special algorithms are necessary for peptide identification in data independent MS/MS where the whole mass range is fragmented.[62,63] Using the MaxQuant/Andromeda infrastructure our group recently developed an implementation of this principle on the Exactive instrument, which consists only of an Orbitrap analyzer with HCD capability.[64]

In conclusion, we have developed, described and tested a robust and scalable search engine that in combination with MaxQuant represents a powerful and unified analysis pipeline for quantitative proteomics, which is freely available to the community.

## ASSOCIATED CONTENT

### S Supporting Information

Supplemental figures and materials. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*J.C. e-mail cox@biochem.mpg.de, fax +49 (89) 8578 3209, phone +49 (89) 8578 2088 or M.M. e-mail mmann@biochem.mpg.de, fax +49 (89) 8578 3209, phone +49 (89) 8578 2557.

## ACKNOWLEDGMENT

We thank other members of our groups in Martinsried and Copenhagen for fruitful discussion and help. We also thank early adopters of Andromeda in other laboratories for helpful feedback. This work was partially supported by the European Union seventh Framework Program (HEALTH-F4-2008-201648/PROSPECTS).

## REFERENCES

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

(2) Yates, J. R., 3rd; Gilchrist, A.; Howell, K. E.; Bergeron, J. J. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6* (9), 702–14.

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

(4) Cox, J.; Mann, M. Is proteomics the new genomics? *Cell* **2007** *130* (3), 395–8.

(5) Vermeulen, M.; Selbach, M. Quantitative proteomics: a tool to assess cell differentiation. *Curr. Opin. Cell Biol.* **2009**, *21* (6), 761–6.

(6) Choudhary, C.; Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell. Biol.* **2010**, *11* (6), 427–39.

(7) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–82.

(8) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–7.

(9) Nesvizhskii, A. I.; Aebersold, R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discovery Today* **2004**, *9* (4), 173–81.

(10) Listgarten, J.; Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (4), 419–34.

(11) Chalkley, R. J.; Hansen, K. C.; Baldwin, M. A. Bioinformatic methods to exploit mass spectrometric data for proteomic applications. *Methods Enzymol.* **2005**, *402*, 289–312.

(12) Colinge, J.; Bennett, K. L. Introduction to computational proteomics. *PLoS Comput. Biol.* **2007**, *3* (7), e114.

(13) Matthiesen, R. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics* **2007**, *7* (16), 2815–32.

(14) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4* (10), 787–97.

(15) Deutsch, E. W.; Lam, H.; Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **2008**, *33* (1), 18–25.

(16) Mueller, L. N.; Brusniak, M. Y.; Mani, D. R.; Aebersold, R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.* **2008**, *7* (1), 51–61.

(17) Matthiesen, R.; Jensen, O. N. Analysis of mass spectrometry data in proteomics. *Methods Mol. Biol.* **2008**, *453*, 105–22.

(18) Bandeira, N.; Clauser, K. R.; Pevzner, P. A. Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol. Cell. Proteomics* **2007**, *6* (7), 1123–34.

(19) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. Clustering millions of tandem mass spectra. *J. Proteome Res.* **2008**, *7* (1), 113–22.

(20) Choi, H.; Ghosh, D.; Nesvizhskii, A. I. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **2008**, *7* (1), 286–92.

(21) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 47–50.

(22) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **2008**, *7* (1), 245–53.

(23) Rauch, A.; Bellew, M.; Eng, J.; Fitzgibbon, M.; Holzman, T.; Hussey, P.; Igra, M.; Maclean, B.; Lin, C. W.; Detter, A.; Fang, R.; Faca, V.; Gafken, P.; Zhang, H.; Whiteaker, J.; States, D.; Hanash, S.; Paulovich, A.; McIntosh, M. W. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **2006**, *5* (1), 112–21.

(24) Rinner, O.; Mueller, L. N.; Hubalek, M.; Muller, M.; Gstaiger, M.; Aebersold, R. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat. Biotechnol.* **2007**, *25* (3), 345–52.

(25) Park, S. K.; Venable, J. D.; Xu, T.; Yates, J. R., 3rd A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **2008**, *5* (4), 319–22.

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

(27) May, D.; Law, W.; Fitzgibbon, M.; Fang, Q.; McIntosh, M. Software platform for rapidly creating computational tools for mass spectrometry-based proteomics. *J. Proteome Res.* **2009**, *8* (6), 3212–7.

(28) Deutsch, E. W.; Shteynberg, D.; Lam, H.; Sun, Z.; Eng, J. K.; Carapito, C.; von Haller, P. D.; Tasman, N.; Mendoza, L.; Farrah, T.; Aebersold, R. Trans-Proteomic Pipeline supports and improves analysis of electron transfer dissociation data sets. *Proteomics* **2010**, *10* (6), 1190–5.

(29) Mortensen, P.; Gouw, J. W.; Olsen, J. V.; Ong, S. E.; Rigbolt, K. T.; Bunkenborg, J.; Cox, J.; Foster, L. J.; Heck, A. J.; Blagoev, B.; Andersen, J. S.; Mann, M. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res.* **2010**, *9* (1), 393–403.

(30) Kumar, C.; Mann, M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett.* **2009**, *583* (11), 1703–12.

(31) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–72.

(32) Cox, J.; Mann, M. Computational Principles of Determining and Improving Mass Precision and Accuracy for Proteome Measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1477–85.

(33) Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7* (12), 952–8.

(34) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002**, *1* (5), 376–86.

(35) de Godoy, L. M.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008**, *455* (7217), 1251–4.

(36) 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* **1999**, *20* (18), 3551–67.

(37) 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.* **1994**, *5*, 976–89.

(38) Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **1999**, *71* (14), 2871–82.

(39) Zhang, N.; Aebersold, R.; Schwikowski, B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2* (10), 1406–12.

(40) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–7.

(41) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958–64.

(42) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W701–6.

(43) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally

modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626–39.

(44) Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell. Biol.* **2004**, *5* (9), 699–711.

(45) Sadygov, R. G.; Cociorva, D.; Yates, J. R., 3rd Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (3), 195–202.

(46) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390–9.

(47) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **1999**, *6* (3—4), 327–42.

(48) Olsen, J. V.; Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (37), 13417–22.

(49) Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005**, *5* (16), 4096–106.

(50) Bern, M.; Finney, G.; Hoopmann, M. R.; Merrihew, G.; Toth, M. J.; MacCoss, M. J. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **2010**, *82* (3), 833–41.

(51) Wang, J.; Perez-Santiago, J.; Katz, J. E.; Mallick, P.; Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* **2010**, *9* (7), 1476–85.

(52) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152–60.

(53) Luber, C. A.; Cox, J.; Lauterbach, H.; Fancke, B.; Selbach, M.; Tschopp, J.; Akira, S.; Wiegand, M.; Hochrein, H.; O'Keeffe, M.; Mann, M. Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **2010**, *32* (2), 279–89.

(54) Hilger, M.; Bonaldi, T.; Gnad, F.; Mann, M. Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Mol. Cell. Proteomics* **2009**, *8* (8), 1908–20.

(55) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **2004**, *4* (7), 1985–8.

(56) Tweedie, S.; Ashburner, M.; Falls, K.; Leyland, P.; McQuilton, P.; Marygold, S.; Millburn, G.; Osumi-Sutherland, D.; Schroeder, A.; Seal, R.; Zhang, H. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.* **2009**, *37* (Database issue), D555–9.

(57) Senko, M. W.; Beru, S. C.; McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229–233.

(58) Geiger, T.; Cox, J.; Mann, M. Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet.* **2010**, *6* (9), e1001090.

(59) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, In Vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127* (3), 635–48.

(60) Mann, M.; Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (47), 18132–8.

(61) Choudhary, C.; Kumar, C.; Gnad, F.; Nielsen, M. L.; Rehman, M.; Walther, T. C.; Olsen, J. V.; Mann, M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **2009**, *325* (5942), 834–40.

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

(63) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77* (7), 2187–200.

(64) Geiger, T.; Cox, J.; Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell. Proteomics* **2010**, *9* (10), 2252–61.

© The Author(s) 2011. This article is published with open access at Springerlink.com

J. Am. Soc. Mass Spectrom. (2011) 22:1373–1380
DOI: 10.1007/s13361-011-0142-8

FOCUS: INTERDISCIPLINARY BIOLOGICAL MS: RESEARCH ARTICLE

# Software Lock Mass by Two-Dimensional Minimization of Peptide Mass Errors

Jürgen Cox, Annette Michalski, Matthias Mann

Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

**Abstract**

Mass accuracy is a key parameter in proteomic experiments, improving specificity, and success rates of peptide identification. Advances in instrumentation now make it possible to routinely obtain high resolution data in proteomic experiments. To compensate for drifts in instrument calibration, a compound of known mass is often employed. This 'lock mass' provides an internal mass standard in every spectrum. Here we take advantage of the complexity of typical peptide mixtures in proteomics to eliminate the requirement for a physical lock mass. We find that mass scale drift is primarily a function of the *m/z* and the elution time dimensions. Using a subset of high confidence peptide identifications from a first pass database search, which effectively substitute for the lock mass, we set up a global mathematical minimization problem. We perform a simultaneous fit in two dimensions using a function whose parameterization is automatically adjusted to the complexity of the analyzed peptide mixture. Mass deviation of the high confidence peptides from their calculated values is then minimized globally as a function of both *m/z* value and elution time. The resulting recalibration function performs equal or better than adding a lock mass from laboratory air to LTQ-Orbitrap spectra. This 'software lock mass' drastically improves mass accuracy compared with mass measurement without lock mass (up to 10-fold), with none of the experimental cost of a physical lock mass, and it integrated into the freely available MaxQuant analysis pipeline (www.maxquant.org).

**Key words:** Mass accuracy, MaxQuant, Proteomics, Database search, Orbitrap, Peptide mass measurement, Lock mass

## Introduction

Mass spectrometry (MS)-based proteomics [1–3] greatly benefits from high resolution and high mass accuracy measurements [4]. For example, resolving co-eluting peptides of similar mass is a prerequisite for their accurate quantification, and high accuracy measurement of peptide masses greatly aid in their identification by providing stringent filters on possible candidates. Several definitions of mass accuracy are commonly used, and this important parameter is often only assessed anecdotally [5]. In proteomics, the operationally important definition is the best mass estimate from the MS measurement together with a statistical confidence interval. This interval can then be used as the basis for setting a permissible mass deviation window for peptide identification in databases. Such confidence intervals can be assigned to each peptide separately. They are obtained from the measured values from consecutive scans and isotope states, weighted by the signal for each data

**Electronic supplementary material** The online version of this article (doi:10.1007/s13361-011-0142-8) contains supplementary material, which is available to authorized users.

*Correspondence to:* Jürgen Cox; *e-mail:* cox@biochem.mpg.de or Matthias Mann; *e-mail:* mmann@biochem.mpg.de

Received: 16 February 2011
Revised: 16 March 2011
Accepted: 16 March 2011
Published online: 22 April 2011

point. We have previously described principles of extracting these mass values from large scale data sets and implemented the corresponding algorithms in the MaxQuant computational proteomics analysis pipeline [4, 6]. As a result of applying these computational algorithms, these peptide mass accuracies are frequently improved to the sub-ppm range. This makes the precursor mass value an important search parameter and allows a corresponding drop in the required quality of the MS/MS spectra while still maintaining a 1% false discovery rate for peptide identifications.

A precondition for the above analysis was the elimination of the systematic mass drift by using a lock mass [7, 8]. A lock mass is a defined compound of known composition that is added to the MS analysis. Some instruments feature a separate electrospray source, which is used to spray the reference compound [9, 10]. Alternatively, the reference compound could be mixed into the analyte directly, but this has disadvantages because the compound may interfere with analysis of low abundance samples or it may not be detectable in high abundance samples. In electrospray, charged droplets are formed in laboratory air and analyte ions are desorbed from them. However, these charged droplets can also absorb and ionize background chemicals that are always present in laboratory air [11, 12]. On the LTQ-Orbitrap family of instruments these ions, specifically polycyclodimethylsiloxanes, can be separately isolated in the linear ion trap and injected into the C-trap, which is an intermediate storage trap [13]. In the C-trap the lock mass ions are mixed with the MS or MS/MS ions to be analyzed and co-injected into the Orbitrap analyzer. The ion is recognized by the data system in real time and the mass scale is automatically adjusted. While this procedure is sufficiently fast to be routinely applicable in proteomic experiments, there is some time requirement for isolating the lock mass ions, adding to overall MS and MS/MS cycle times. In addition, it can often be desirable to suppress background ions in laboratory air (i.e., by the ABIRD device: www.esisourcesolutions.com). This has the side effect that the lock mass is no longer available. For these reasons, an alternative to the lock mass would be beneficial.

In proteomics experiments, typically hundreds or thousands of peptides are identified in every LC-MS/MS run. Many of these peptides have very information-rich MS/MS spectra, and they can be unambiguously identified even with large mass tolerances. We have previously made use of this fact by implementing a two-pass search, where the top identified peptides serve as mass references for calibration [14]. However, this recalibration was done globally for the entire LC-MS run, was only applicable to time of flight data and did not attempt to reach sub-ppm mass accuracy. For Orbitrap data, the simple mass scale adjustments [14] would not be applicable. In this paper we set out to develop algorithms to replace the physical lock mass with a software algorithm that performs at least as well in global recalibration of Orbitrap data.

## Methods

### *Protein Digestion*

Total HeLa cell lysate was treated with a urea (6 M) and thiourea (2 M) solution followed by reduction with dithiothreitol (DTT) (1 mM) for 30 min and alkylation with iodoacetamide (IAA) (55 mM) for 20 min at room temperature. The proteins were digested with Lys-C (1 μg/50 μg protein) (Wako, Neuss, Germany) for 3 h at room temperature. The mixture was diluted with water (1:4) before incubation with trypsin (1 μg/50 μg protein) (Promega, Mannheim, Germany) for 12 h at room temperature. The digestion was stopped by addition of formic acid (3%) and the samples stored on StageTips [15].

### *LC-MS/MS Analysis*

The peptide mixture was loaded onto a $C_{18}$-reversed phase column (15 cm long, 75 μm i.d.) that was packed in-house with ReproSil-Pur C18-AQ 3 μm resin (Dr. Maisch) in buffer A (0.5% acetic acid). The peptide mixture was separated with a linear gradient of 5%–60% buffer B (80% ACN and 0.5% acetic acid) at a flow rate of 250 nL/min on a nanoflow HPLC (Proxeon Easy HPLC; Thermo Fisher Scientific). On-line coupling of the HPLC system to an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) was achieved using a nanoelectrospray ion source (Proxeon Biosystems, now Thermo Fisher Scientific). Data were acquired in a data-dependent 'top5' format, selecting the most abundant precursor ions from the survey scan (mass range 300–1650 Th) in order to isolate them in the linear ion trap and fragment them by CID with a normalized collision energy of 35 eV. Survey scans were acquired with a resolution of 60,000 at *m*/*z* 400 and with a target value of $10^6$ in the Orbitrap analyzer. The MS/MS scans were acquired with unit mass resolution in the LTQ using 3000 as target value. Dynamic exclusion was defined by a list size of 500 features and exclusion duration of 90 s. Early expiration was set to expiration count 3 and S/N threshold 3. The lower threshold for targeting a precursor ion in the MS scans was 1000 counts. Three technical replicates were acquired without using the lock mass option in Xcalibur. In three separate technical replicates protonated polycyclodimethylsiloxane (PCM-6) with exact *m*/*z* 445.1200 Th was selected as lock mass for the measurement [13].

Data was analyzed by MaxQuant [4] using the Andromeda search engine [16]. The IPI human data base was used for peptide identification in the IPI human data base (containing 87,061 entries) combined with 262 common contaminants and concatenated with the reversed versions of all sequences. Enzyme specificity was set to trypsin, allowing cleavage N-terminal to proline. Further modifications were cysteine carbamidomethylation (fixed) as well as protein N-terminal acetylation and methionine oxidation (variable).

### *Computational Methods*

Data were analyzed with the MaxQuant framework [4], which is written in C# in the Microsoft .NET environment. Algorithmic parts of MaxQuant are available as source code and the entire program can be freely downloaded as well from www.maxquant.org. Detailed instructions for installation and support programs are also available [17].

## Results and Discussion

### *Time and m/z Dependence of the Mass Error*

We start by illustrating global features of the mass error distributions in liquid chromatography-tandem mass spectrometry runs. We performed 2 h LC-MS/MS runs of a HeLa lysate acquired on an LTQ Orbitrap without enabling the lock mass feature. Peptides were identified with a suitably large tolerance for the peptide mass, sufficient to include possible deviations due to instrumental drift (20 ppm in this case). We then calculated the mass error for each peptide. Figure 1 shows the results of these measurements for the elution profiles (MS-level isotope patterns) of four peptides. Indeed, all four peptides in Figure 1 are shifted by approximately 5 to 6 ppm because of the lack of calibration and because the lock mass feature was not applied. These mass deviations are far in excess of the sub-ppm accuracy that the instrument is capable of [4]. It would now be interesting to determine if the masses are off due to a global shift, due to statistical fluctuations, or if the mass error is a function of either retention time or mass, or if it depends on both. The two doubly charged peptides in Figure 1a and b have approximately the same mass but differ in their retention times. One finds that the mass errors of the two peptides differ from each other by more than 1 ppm, suggesting a time dependence of the mass error. Likewise, we can compare the two triply charged peptides in Figure 1c and d, which have similar retention time but differ in mass. Again, the difference in the mass errors is more than 1 ppm, indicating an *m/z* dependence of the mass error.

To investigate time dependence of the mass error in a systematic manner we plotted the ppm mass error as a function of retention times in Figure 2a (red data points). Clearly, there are systematic effects in the mass error distribution. There is a tendency for the error to be slightly greater at larger elution times. In addition there is a wave-like pattern on the timescale of 10 to 20 min. Figure 2b shows a zoom of the same data into a smaller retention time interval. This figure reveals that the red curve has systematic structures on smaller timescales of about 1 min as well. The blue data points in Figure 2a and b are mass errors from the corresponding LC-MS/MS run in which the lock mass feature was enabled. As expected, the average mass error is now much closer to zero. However, also here larger deviations on smaller timescales can be seen. For instance, at t=80 min and t=

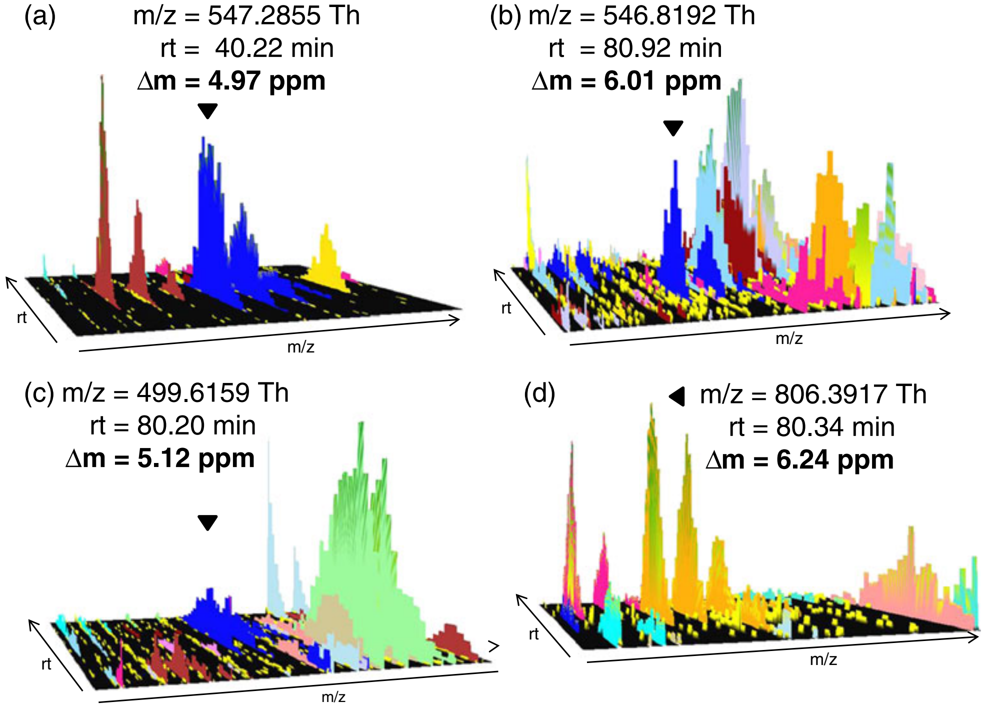


**Figure 1.** Three-dimensional views of three MS isotope patterns corresponding to peptides. Peptides (**a**) and (**b**) have similar mass but different retention times. Their mass errors differ by more than one ppm. Peptides (**c**) and (**d**) have similar retention time but differ in *m/z*. They also require different mass recalibrations

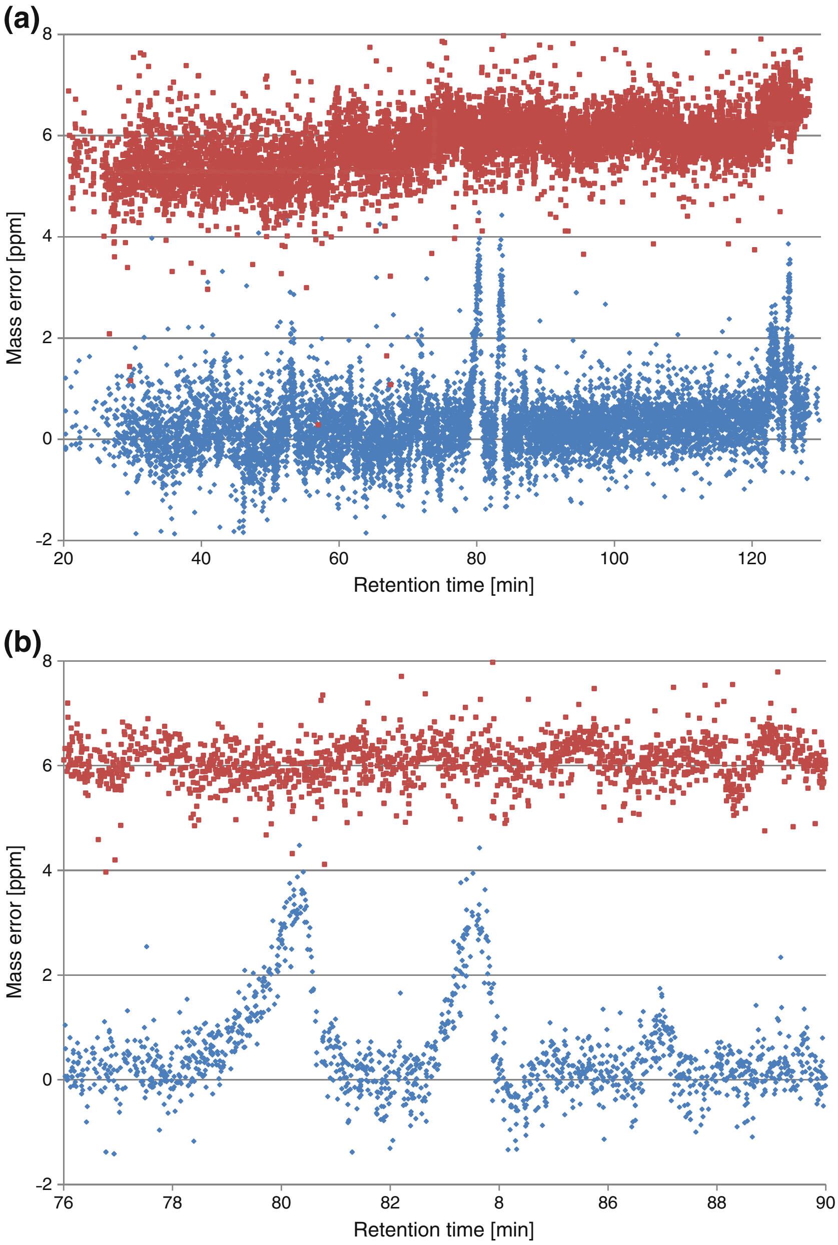


**Figure 2.** (**a**) Mass error in ppm for the peptides identified in two LC-MS runs as a function of retention time. Blue points originate from an LC-MS run in which the lock mass feature has been used while the red points are from an LC-MS run without lock mass. (**b**) Same data zoomed in the time window from min 76 to 90 min

83.5 min, the mass error rises for short times to 4 ppm. Inspection of the data files reveals that this is due to loss of the lock mass in these time intervals.

Figure 3a is a plot of the mass error as a function of *m*/*z* instead of elution time. Again a systematic nonlinear dependence can clearly be seen. These systematic variations seem to be only on larger scales of 100 Da without an indication of systematic effects on lower *m*/*z* scales. Histograms of these mass deviations are shown in Figure 3b for the data with lock mass (blue) and without lock mass (red). The lock mass helps in keeping the deviations near zero but does not completely center them there. This is partially due to the mass dependence of the error, which is not eliminated by the lock mass. The tail to the right of the distribution is mainly derived from the time intervals where the lock mass has not been found. The absolute average mass deviation of the lock mass data is 0.53 ppm. In the data that were acquired without lock mass the errors are centered at ~6 ppm. The full width half maximum of both distributions is similar and, in both cases, around 1 ppm.

## *The Software Lock Mass Optimization Problem*

As we have seen the mass error is a function of (at least) two variables, time and *m*/*z*, and projections onto each of them display rich structure and clear functional dependencies. It is

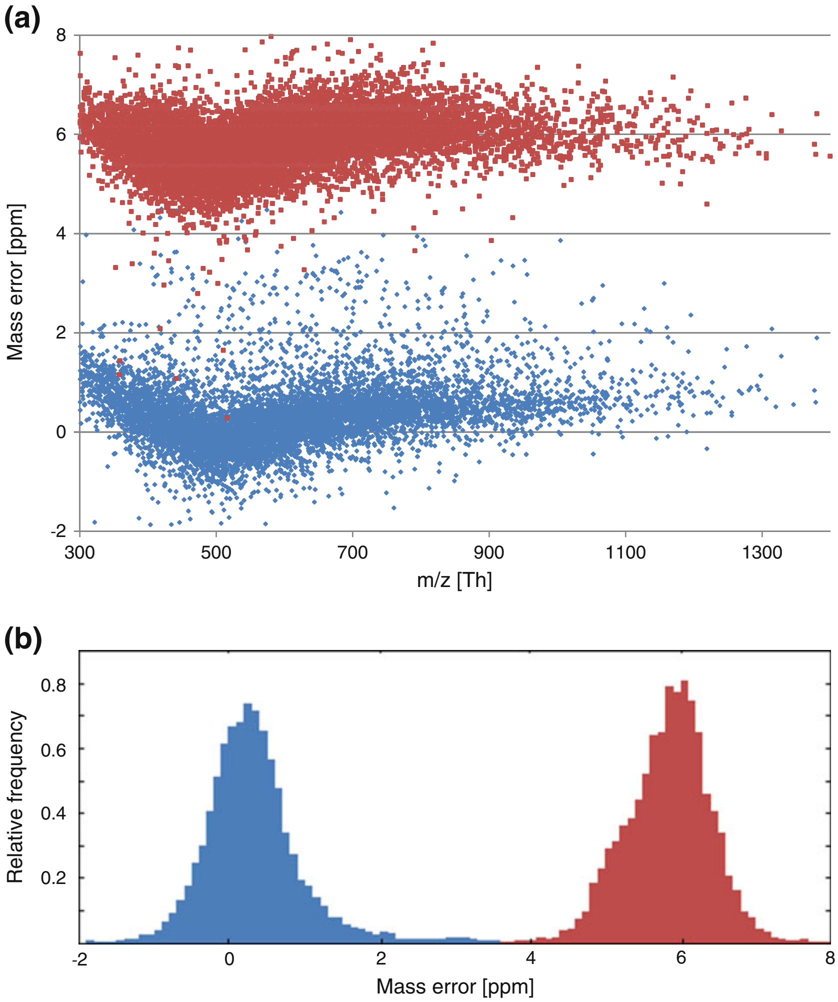


**Figure 3.** (**a**) Same data as in Figure 2a but plotted as a function of *m/z*. Red and blue points originate from LC-MS runs without and with lock mass, respectively. (**b**) Histograms of the mass errors shown in (**a**)

a reasonable assumption that the mass error depends on these two parameters in an additive way, i.e., the nonlinearities in the mass scale should be independent of elution time. This assumption makes sense from a physical point of view. The *m/z*-dependent error is determined by static properties of the mass spectrometer that do not vary with time; for instance, imperfections in the geometry of the Orbitrap cell. In addition to this static error, a dynamic component is superimposed that is caused by any kind of disturbance that happens during the chromatographic time scale, with temperature shifts being a prominent example. This means that it should be possible to parameterize the mass error function as the sum of two terms, f and g, each depending on only one of the two variables and each being parameterized by sets of parameters $\theta_f$ and $\theta_g$:

$$\Delta m = f(t, \theta_f) + g(m/z,\ \theta_g). \qquad (1)$$

Note that this equation does not assume linearity in any of the variables or parameters, but only that the contribution of each variable can be represented as a sum of nonlinear terms. The explicit form of the parameterization of the functions f and g are described below.

To determine the functions f and g in the above equation (equation 1), we generate number triples ($\Delta m_j$, $t_j$, $m/z_j$) by performing a first peptide database search with the Andromeda search engine [16], which is integrated into the MaxQuant software package (see Figure 4). For this purpose, we allow a large tolerance of the peptide mass of 20 ppm. This initial tolerance can be set by the user. While we have found 20 ppm to be a good value for routine operation on this instrument class, this number can and should be increased in cases where the calibration is off by more than 20 ppm. All peptide identifications that have an Andromeda score of at least 80 are accepted. The mass error is then calculated based on the elemental composition of the identified peptides and the experimentally measured masses. Note that in MaxQuant the measured peptide mass is derived from the entire three-dimensional isotope pattern that the MS/MS spectrum was associated with [4]. MS/MS spectra not associated with a three-dimensional MS isotope pattern

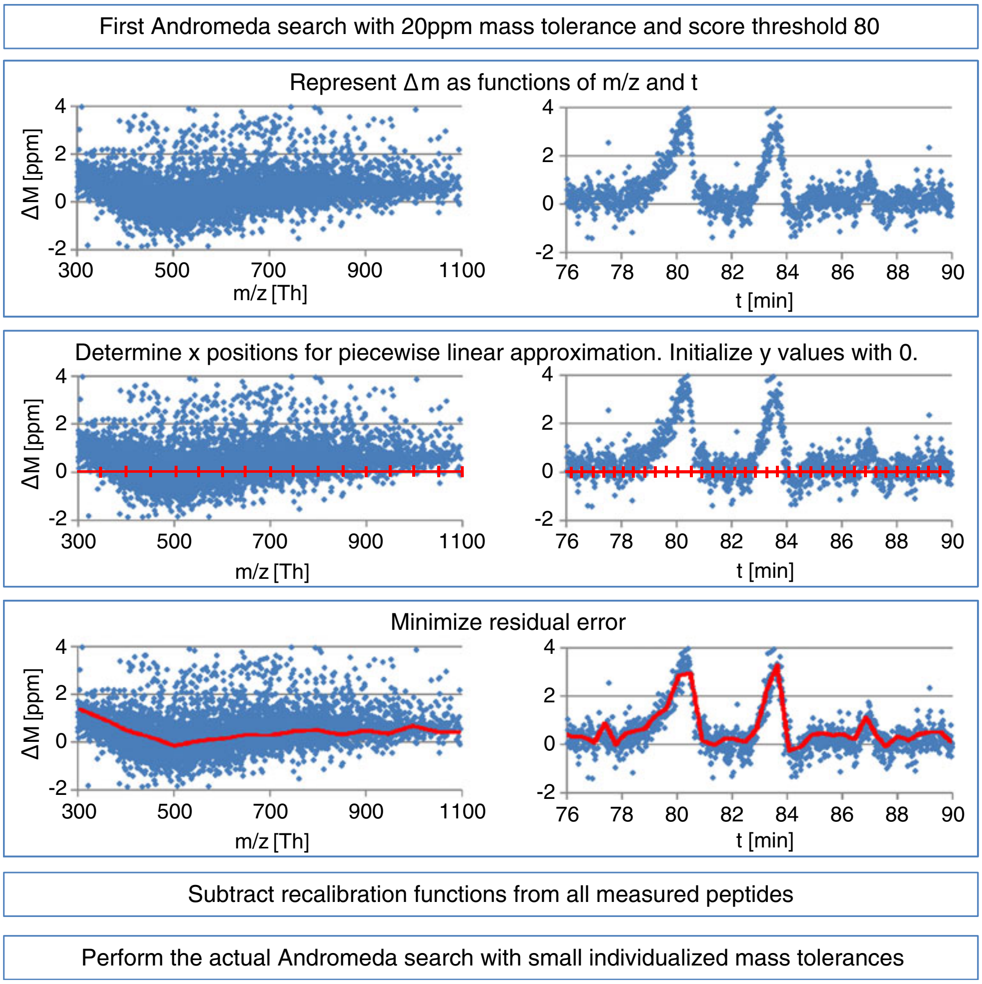


**Figure 4.** Algorithmic steps of the software lock mass workflow

are discarded. The retention time is estimated as the intensity-weighted time average over the elution profile of the peptide. The computational task is now to determine the functions f and g in such a way that their sum best approximates the calculated mass error. To achieve this, we minimize the sum of squares of the residual errors (equation 2).

$$\sum_{j}\left(\Delta m_j - f\left(t,\theta_f\right) - g\left(\frac{m}{z},\theta_g\right)\right)^2 \qquad (2)$$

For this purpose, the functions f and g have to be parameterized in a suitable way. We use piecewise linear functions for f and g. First, the x-positions of these functions are adapted to the data, and they are then treated as constant during the minimization. The number of x-positions and their exact location are chosen such that the number of degrees of freedom adjusts itself to the complexity of the data. Roughly speaking, the more data are available the more complex the parameterizations of the functions can be. The x positions are chosen such that there are at least 80 data points per x position in the *m*/*z* direction and 50 data points per x position in the time direction. Furthermore, the x positions have to be at least 50 Th apart in the *m*/*z* direction. The numbers that are being determined during the optimization are the y-values at these fixed x positions, typically several dozens or hundreds of coefficients. The functions f and g are linearly interpolated between these positions. One of the y-values has to be fixed to an arbitrary value since the system otherwise has a zero mode. The numerical solution of this minimization problem is obtained by the Levenberg-Marquard method (see, e.g., reference [18] for an introduction). After the parameters of f and g have been determined, we can subtract the systematic mass error from the measured mass of each MS isotope pattern in the LC-MS run. Subsequently, the actual database search (second pass search) is performed with individualized peptide mass tolerances inside the MaxQuant framework as before.

## *Performance of the Software Lock Mass*

Figure 5 depicts the mass error distribution after recalibration and second pass Andromeda search. Figure 5a and b show the dependence on *m*/*z* while Figure 5c and d show the

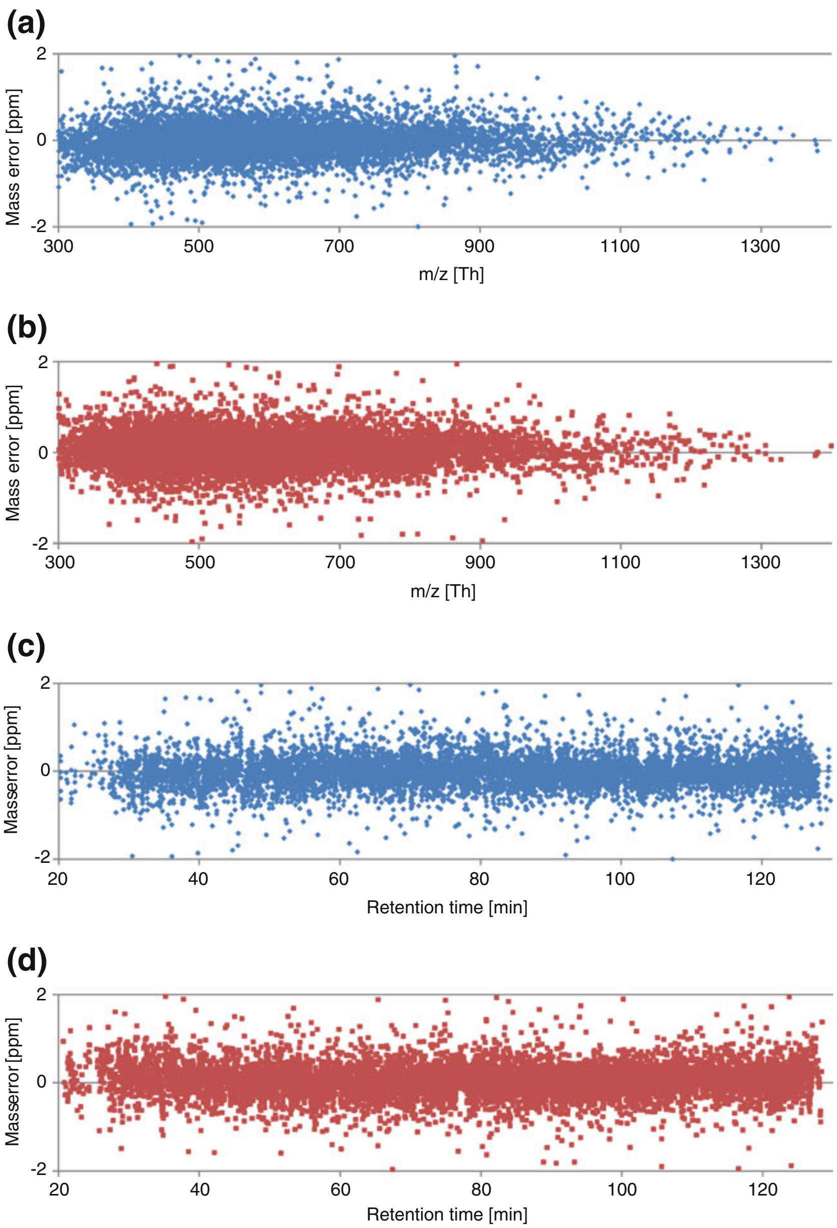


**Figure 5.** Mass errors after recalibration. (**a**) Lock mass; plotted against *m/z*. (**b**) No lock mass; plotted against *m/z*. (**c**) Lock mass; plotted against retention time. (**d**) No lock mass; plotted against retention time

time dependence. For the data with and without lock mass our algorithm has removed all systematic effects from the data. Figure 6 shows histograms of the mass error after recalibration. The absolute average mass deviations are 0.29 ppm for the data with lock mass and 0.27 ppm for the data without lock mass. The corresponding mass standard deviations are 0.42 and 0.39 ppm, respectively. This indicates that when using the software lock mass workflow, the mass accuracy is as good as for data that were acquired with lock mass. While shown here for a particular example, we have found this to be true in general. As an example, Supplemental Figure 1 shows a very challenging LC-MS/MS run acquired with lock mass feature in which the lock mass was lost and found again several times. Panel a shows

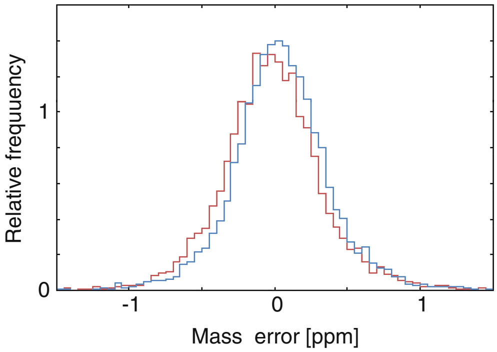


**Figure 6.** Histograms of the mass errors after recalibration for data acquired with lock mass (blue), and without lock mass (red)

the time-dependence of the mass error before recalibration. The time series of the mass error is bi-stable, flipping back and forth between zero and 7 ppm. Nevertheless, this very difficult case is reliably recalibrated by the software lock mass (Suppl. Figure 1).

## Conclusion

Here we have investigated the concept of a software lock mass, a replacement for its physical version, which is integrated into the MaxQuant/Andromeda computational proteomics workflow. We have demonstrated that it performs as least as well as the physical lock mass on typical complex proteome data. Even data that were acquired with a lock mass may benefit from the application of our recalibration workflow, especially in cases where the lock mass performance was not optimal. In contrast to the hardware lock mass option, the software lock mass can correct nonlinearities in the mass scale. Here, we have demonstrated the method on an Orbitrap instrument. However, we speculate that other instrument types would also benefit from the software lock mass approach. For instance, mass calibration drift typically is an issue of practical importance for time of flight instruments. Furthermore, while shown here for MS spectra, the benefits of the software lock mass also carry over to high-resolution MS/MS spectra.

Importantly, use of the software lock mass is completely free from an experimental point of view. All it requires is a peptide mixture of sufficient complexity. In contrast, a physical lock mass, even if derived from laboratory air, always has some experimental cost, such as additional hardware, influence on the spectra, or a slight increase in cycle time. Since the software lock mass is an unmitigated benefit, it can be adopted for all proteomics experiments, as we have done for some time in our laboratory.

## Acknowledgments

The authors thank all the other members of the Proteomics and Signal Transduction group for help and discussions. This work was partially supported by the European Union 7th Framework Program (HEALTH-F4-2008-201648/PROSPECTS).

## Open Access

This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

1. Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003)
2. Yates, J.R.I.I.I.: Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 297–316 (2004)
3. Yates III, J.R., et al.: Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* **6**, 702–714 (2005)
4. 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**, 1367–1372 (2008)
5. Zubarev, R., Mann, M.: On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteom.* **6**, 377–381 (2007)
6. Cox, J., Mann, M.: Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* **20**(8), 1477–1385 (2009)
7. Muddiman, D.C., Oberg, A.L.: Statistical evaluation of internal and external mass calibration laws utilized in fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **77**(8), 2406–2414 (2005)
8. Williams Jr., D.K., Muddiman, D.C.: Parts-per-billion mass measurement accuracy achieved through the combination of multiple linear regression and automatic gain control in a Fourier transform ion cyclotron resonance mass spectrometer. *Anal. Chem.* **79**(13), 58–63 (2007)
9. Hannis, J.C., Muddiman, D.C.: A dual electrospray ionization source combined with hexapole accumulation to achieve high mass accuracy of biopolymers in Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.* **11**(10), 876–883 (2000)
10. Nepomuceno, A.I., Muddiman, D.C., Bergen, H.R. 3rd, Craighead, J. R., Burke, M.J. , Caskey, P.E., Allan, J.A.: Dual electrospray ionization source for confident generation of accurate mass tags using liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **75**(14), 3411–3418 (2003)
11. Fenn, J.: Proceedings of the 34th ASMS, Cincinnati, OH; p. 507–509 (1986)
12. Schlosser, A., Volkmer-Engert, R.: Volatile polydimethylcyclosiloxanes in the ambient laboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry. *J. Mass Spectrom.* **38** (5), 523–525 (2003)
13. Olsen, J.V., de Godoy, L.M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., Mann, M.: Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteom.* **4**(12), 2010–2021 (2005)
14. Mortensen, P., Gouw, J.W., Olsen, J.V., Ong, S.E., Rigbolt, K.T., Bunkenborg, J., Cox, J., Foster, L.J., Heck, A.J., Blagoev, B., Andersen, J.S., Mann, M.: MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res.* **9** (1), 393–403 (2010)
15. Rappsilber, J., Ishihama, Y., Mann, M.: Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**(3), 663–670 (2003)
16. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., Mann, M.: Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**(4), 1794–1805 (2011)
17. Cox, J., Matic, I., Hilger, M., Nagaraj, N., Selbach, M., Olsen, J.V., Mann, M.: A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat. Protoc.* **4**(5), 698–705 (2009)
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, Cambridge, 683–688 (2007)

# 5 A Global View of Peptide Fragmentation Chemistry

**Article 6** **A Systematic Investigation into the Nature of HCD Spectra**

Annette Michalski, Nadin Neuhauser, Jürgen Cox, and Matthias Mann
*J Proteome Res*, 2012, submitted.

**Article 7** **Expert System for Computer-Assisted Annotation of MS/MS Spectra**

Nadin Neuhauser*, Annette Michalski*, Jürgen Cox, and Matthias Mann
*Mol Cell Proteomics*, 2012, in revision.

(*) equal contribution

## Prologue

Since the early days of tandem mass spectrometry, fragmentation chemistry has been of great interest to researchers as the data can be used to determine structure and composition of the substances under investigation. Especially for unknown small molecule species, tandem mass spectra often appear to some extent random, despite a number of common fragmentation rules, well-described rearrangements or typical neutral losses that are summarized in entire books. In contrast, the chemistry involved in peptide fragmentation is much less complex, because the regular peptide backbone structure results in quite predictable fragmentation spectra. Usually, the peptide bonds are the weakest ones of the molecule and are therefore cleaved first if appropriate collision energy is applied. As a result, the fragmentation spectra of peptides feature sequence ladders from N- to C-terminus and in the opposite direction. Reading out the amino acid sequence is in principle relatively straightforward.

In practice, however, peptide fragmentation is an extensive area of research by itself, because the fragmentation spectra contain many more peaks than just backbone fragments. Furthermore, fragmentation pathways are of scientific interest as are the energetic aspects of the fragmentation process in the gas phase. For collision-induced fragmentation of protonated peptides, one result of such studies was the *mobile proton model*[117]. It provides a framework for understanding the necessity of a certain number of protons for peptides with a certain number of basic amino acids for the generation of interpretable fragmentation spectra.

Systematic investigations of fragment ions obtained from specifically designed synthetic peptides reveal patterns that keep occurring in fragmentation spectra, e.g. the exceptional existence of $b_1$

ions or high abundance of fragment ions that contain an N-terminal proline[118,119]. Experienced mass spectrometrists are familiar with these processes and are able to explain nearly all fragment ion peaks of a peptide tandem mass spectrum. Often it is very helpful to try to resolve the entire spectrum in this way, because reporter ions or specific neutral losses reveal important information that make peptide identification more reliable.

Despite decades of extensive research, the greatest number of peptide fragmentation spectra has probably been obtained very recently in modern large-scale proteomics experiments. Here, however, the focus is usually on biological questions and due to the overwhelming number of spectra, automated data analysis software and search engines are inevitably applied. Moreover, sophisticated software tools such as MaxQuant perform very accurate mass calibration to the data and apply statistical concepts to ensure proper data analysis. Conversely, search engines only incorporate very basic knowledge of peptide fragmentation for the identification. As a result, visual inspection of the results may leave researchers puzzled due to many unannotated peaks.

The recent success of the high-high strategy in shotgun proteomics now allows explaining the origin of fragment ion peaks much more confidently. We were motivated by the excellent quality of HCD fragmentation spectra as routinely acquired on the Orbitrap instruments (LTQ Orbitrap Velos, Q Exactive and Orbitrap Elite) since 2009, to revisit the extensively studied subject of peptide fragmentation chemistry. The aim of this project was to iteratively perform a large-scale statistical investigation into the ion types occurring in collision induced fragmentation spectra and develop a computer-assisted Expert System to automatically annotate fragment ion spectra with information beyond that provided by standard peptide search engines. This Expert System features a knowledgebase that contains all fragmentation patterns that we collected and evaluated based on our experience. It uses a rule engine to apply this codified experience to identified spectra. While novice users primarily benefit from the comprehensive annotation with which they can quickly become experts themselves, advanced scientists can modify and extend the rule set according to their specific question (Article 7). We then employed this novel software tool for a broad inquiry into the nature of fragment ion peaks of tryptic peptides. A detailed comparison of the ion types in HCD and high resolution CID reveals a greater variety of fragment ions in HCD. Nevertheless, after applying the Expert System the overall percentage of explained or annotated MS/MS intensity is very comparable at a median of about 85% for both. This corresponds to a 35% increase compared to the annotations resulting from standard database identification. We further demonstrated the flexibility of the system by adapting the rule set to phosphorylation events (Article 6).

# A systematic investigation into the nature of tryptic HCD spectra

Annette Michalski, Nadin Neuhauser, Jürgen Cox and Matthias Mann*

From the Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried, Germany

**Keywords**: tandem mass spectrometry, fragmentation mechanisms, shotgun proteomics, ion types, CID, HCD, Expert System, spectrum annotation

*Corresponding author:

Matthias Mann
Department of Proteomics and Signal Transduction
Max-Planck Institute of Biochemistry
Am Klopferspitz 18
D-82152 Martinsried
Germany

Email: mmann@biochem.mpg.de
Fax: +49 89 8578 2219

Abbreviations: CID, Collision Induced Dissociation; ETD, Electron Transfer Dissociation; FDR, False Discovery Rate; FT, Fourier Transform; HCD, Higher Energy Collisional Dissociation; HPLC, High Performance Liquid Chromatography; ICR, Ion Cyclotron Resonance; IM, Immonium Ion; IPI, International Protein Index; LTQ, Linear Trap Quadrupole; MS/MS, Tandem mass spectrometry; PIF, Precursor Intensity Fraction; Q TOF, Quadrupole Time Of Flight instrument; TOF, Time Of Flight.

## Abstract

**Modern mass spectrometry – based proteomics can produce millions of peptide fragmentation spectra, which are automatically identified in databases using sequence specific b- or y-ions. Proteomics projects have mainly been performed with low resolution collision induced dissociation (CID) in ion traps, but higher energy collisional dissociation (HCD) now routinely provides high mass accuracy and full mass range fragmentation. To systematically study the nature of HCD spectra, we made use of a large scale dataset of tryptic peptides identified with an FDR of 0.0001 from which we extract a subset of more than 16,000 that have little or no contribution from co-fragmented precursors. We employed a newly developed computer assisted Expert System, which distills our experience and literature knowledge about fragmentation pathways. It aims to automatically annotate the peaks in high mass accuracy fragment spectra while strictly controlling the false discovery rate. Using this Expert System we determined that sequence specific regular ions covering the entire sequence were present for almost all peptides with up to 10 amino acids (median 100%). Peptides up to 20 amino acid length contained sufficient fragmentation to cover 80% of the sequence. Internal fragments are common in HCD spectra but not in high resolution CID spectra (10% vs. 1%). The low mass region contains abundant immonium ions (6% of fragment ion intensity), the characteristic $a_2$, $b_2$ ion pair (72% of spectra), side chain fragments and reporter ions for peptide modifications such as tyrosine phosphorylation. B- and y-ions account for only 20% of fragment ions by number but 53% by ion intensity. Overall, 84% of the fragment ion intensity was unambiguously explainable. Thus high mass accuracy HCD and CID data are near comprehensively and automatically interpretable.**

## Introduction

Rapid technological development of mass spectrometric instrumentation in conjunction with advanced bioinformatics analysis capabilities now allow relatively streamlined and in depth analysis of proteomic samples[1-3]. Modern proteomics projects routinely generate millions of fragmentation spectra, making entirely automated software tools a necessity. These include search engines that match MS/MS spectra to the most probable peptide sequence in a database - typically relying on sequence specific backbone fragments, referred to as 'regular ions' in this article, as well as associated neutral losses[4]. However, there are many other fragment ions in tandem mass spectra and it has been argued that detailed interpretation of these of at least the more abundant peaks should be a requirement for confident peptide assignment[5]. Likewise, detailed understanding of the fragmentation process and discovery of potential new fragment types require knowledge of the identity of the majority of fragmentation peaks.

While there are many different ways to fragment peptides, in proteomics collision induced dissociation (CID) has by far been the most frequently used technique (for a recent tutorial of peptide fragmentation and spectrum interpretation see[6]). While there are differences between the types of CID used, the general ion types and fragmentation modes are the same and are summarized in Figure 1. The backbone fragments are designated as a, b, c for N-terminal and x, y, z for C-terminal types depending on the cleavage position on the peptide backbone[7-10]. A full series of either b- or y-type ions in principle allows reading out the entire amino acid sequence from a fragment ion spectrum. In collision-induced fragmentation techniques cleavage of the peptide bond is preferred but labile post translational modifications such as phosphorylation or glycosylation also partially or (rarely) completely detach. While the chemistry involved in peptide fragmentation is still not completely understood, the mobile proton model is currently the most widely accepted framework to describe the dissociation process[11, 12]. Moreover, different fragmentation pathways of protonated peptides have been extensively investigated and modeled with respect to both kinetic and thermodynamic aspects[13].

In addition to the standard backbone ions, tandem mass spectra can contain many additional fragment ions[14]. Numerous studies of peptide dissociation behavior have been carried out to investigate the abundance and structure of ion types such as internal ions, immonium ions or neutral losses from these (Fig. 1)[15, 16] and some programs such as Protein Prospector provide comprehensive lists of produced ion types for different fragmentation mechanisms and instrument types and consider the latter for scoring of tandem mass spectra[17]. Furthermore, special types of ions have been characterized, for instance $b_1$ ions of N-terminally acetylated ions[18], $c_1$ ions in case glutamine is the second amino acid from the N-terminus[19, 20], specific side chain losses such as from oxidized methionine[21] and many more. Finally, novel fragmentation processes continue to be discussed controversially, such as the extent of scrambling of b-ions due to their formation of a cyclic peptide structures followed by random cleavage, which could interfere with the determination the correct amino acid sequence from the data[22-25].

Another important parameter influencing the types of fragment ions observed in a tandem mass spectrum is the instrument type itself. Triple quadrupole and quadrupole time of flight (TOF) fragmentation are beam-type dissociation processes[26], where primary fragments retain kinetic energy and are therefore more likely to fragment again in the multiple collision conditions typical of these instruments. In 3D or 2D ion traps the excitation and activation step is only applied to the selected precursor mass. Any primary fragmentation product is off-resonance with the applied radio-frequency and therefore usually remains intact. When CID is performed in ion traps, the low mass fragments are typically not retained, leading to a low mass cut-off in the tandem mass spectra[27].

Higher energy collisional dissociation (HCD), first described in 2007, made beam type fragmentation available on the Orbitrap analyzer platforms[28]. HCD fragments have been analyzed at low resolution in an ion trap[29, 30] but are in general always detected in the Orbitrap analyzer at high resolution and mass accuracy. Since the introduction of the LTQ Orbitrap Velos mass spectrometer, which features improved sensitivity and HCD capability compared to its predecessors, routine acquisition of tandem mass spectra in the Orbitrap analyzer has become feasible[31]. This approach is termed 'high-high' strategy because both the full scans (MS) and the fragment ion scans (MS/MS) have high resolution and high mass accuracy in comparison to previous strategies with acquisition of CID scans (MS/MS) in the ion trap ('high-low')[32]. Note that high-high strategies have been the default in quadrupole-TOF instruments for many years – however, this did not necessarily imply high mass accuracy in the MS/MS mode; primarily due to issues with ion statistics. Due to the dedicated collision cell, HCD fragment ion spectra cover nearly the entire mass range and are therefore particularly suitable for observing the low mass region which contains $a_2$/$b_2$ pair, immonium ions, fragments resulting from the amino acid side chains as well the reporter ions[18] used for quantification in the TMT or iTRAQ methods[33-35]. Importantly, high mass accuracy of fragment ions helps to unambiguously annotate the fragment ion peaks. Especially in the low mass region, an accurate mass measurement may even uniquely determine the elemental composition of the fragment.

In contrast to ion trap CID data, high resolution HCD has been relatively little studied. Although HCD ion types are expected to recapitulate fragmentation rules known from older CID type instruments, those have not been tested on large-scale and high accuracy data. Here, we wished to take advantage of the excellent signal to noise, dynamic range and mass accuracy of HCD spectra on the Orbitrap analyzer to systematically investigate features of HCD spectra. This was facilitated by a rule-based Expert System, which was developed in an iterative manner with this study and is described elsewhere[36]. This Expert System synthesizes well-established knowledge about peptide fragmentation pathways mechanisms. It is capable of annotating large-scale MS/MS data sets based on the rules chosen by the researcher. We apply the Expert System for a comprehensive statistical investigation into the nature of HCD tandem mass spectra of tryptic peptides.

## Experimental Procedures

**Sample preparation.** Total cell extracts of *E.coli*, yeast and HeLa cells were separated by 1D-SDS PAGE (4-12% Novex mini-gel, Invitrogen) in three separate lanes. Colloidal Coomassie (Invitrogen) was used for staining of the proteins before each lane was cut into 8 or 10 slices. All gel slices were subjected to reduction of the proteins with 10 mM DTT in 50 mM ammonium bicarbonate and subsequently alkylated with 55 mM IAA in 50 mM ammonium bicarbonate. In-gel digestion with 12.5 ng/µl trypsin (Promega) in 50 mM ammonium bicarbonate was carried out at 37°C for 12h followed by extraction of the tryptic peptides with 3% TFA in 30% ACN[37]. Peptides were loaded on $C_{18}$ StageTips[38] before eluting them with 80% ACN in 0.5% acetic acid prior to analysis.

HeLa cell lysate was digested according to the filter-aided sample preparation (FASP) method[39]. Briefly, the lysate was solubilized in SDS-containing buffer and loaded onto Microcon YM-30 devices (Millipore, Billerica, MA, USA) to remove SDS and exchange it by urea. The protein mixture was alkylated with 50 mM iodoacetamide before urea was replaced with 20 mM ammonium bicarbonate. The proteins were digested overnight at 37°C with trypsin (Promega) (1 ug trypsin: 100 ug protein). Peptides were collected from the filter after centrifugation. For enrichment of phosphorylated peptides, the mixture was acidified with trifluoroacetic acid to pH 2.7 and ACN was added to a final concentration of 30%. Incubation with $TiO_2$ beads [40] (MZ Analysentechnik, Germany) prepared in 30 mg ml−1 solution of dihydrobenzoic acid (Sigma) was carried out for 30 mins, before washing the beads with 30% ACN and 3% TFA (twice) followed by two washes with 75% ACN and 0.3% TFA. The phosphopeptides were eluted with buffer containing 15% ammonium hydroxide and 40% ACN. Finally, the eluted phosphopeptides were loaded on $C_{18}$ StageTips before eluting them with 60% ACN in 0.5% acetic acid prior to analysis.

**LC-MS/MS analysis.** For the analysis of proteome samples, the peptide mixture was separated on a C18-reversed phase column (15 cm, 75 µm ID, packed in-house with ReproSil-Pur $C_{18}$-AQ 3 µm resin, Dr. Maisch GmbH). An Easy-nLC (Thermo Scientific, Odense) with IntelliFlow system was used for sample loading and operated at a constant flow rate of 250 nl/min during the 110 min linear gradient of 8-60% buffer B (80% ACN and 0.5% acetic acid). A nano-electrospray ion source (Thermo Scientific, Odense) was used for on-line coupling to the LTQ Orbitrap Velos mass spectrometer[31]. Mass spectra were measured in positive ion mode applying a data-dependent 'top 10' method for the acquisition of a survey scan followed by MS/MS spectra of the ten most abundant precursors. High resolution data was acquired in the Orbitrap analyzer with a resolution of 30,000 (m/z 400) for MS and 7,500 (m/z 400) for MS/MS scans. For peptide fragmentation higher energy collisional dissociation (HCD) was used applying a normalized collision energy of 40 eV. The minimal signal threshold required was set to 5,000. The target value in the Orbitrap analysis was 1e6 for the MS scans and 5e4 for the MS/MS scans with 2 Da isolation window and the first mass was set to 80 Th for HCD spectra. Fragmented precursors were dynamically excluded from targeting for 90 seconds. High resolution CID data was acquired on an Orbitrap Elite (Thermo Scientific) the same parameters, however the resolution for MS scans was 120,000 (m/z 400) and for MS/MS scans 15,000 (m/z 400); the normalized collision energy was set to 35 eV.

For the phosphoproteome data, the enriched peptide mixtures were separated on a $C_{18}$-reversed phase column (20 cm, 75 µm ID, packed in-house with ReproSil-Pur $C_{18}$-AQ 1.8 µm resin, Dr. Maisch GmbH) applying a 90 min linear gradient of 5-30% buffer B (80% ACN and 0.1% formic acid) and analyzed on the Orbitrap Elite instrument [41] that was online-coupled to an Easy-nLC 1000 (Thermo Scientific, Odense). The MS data was acquired with resolution of 120,000 (m/z 400) and target value of 1e6 and MS/MS (HCD fragmentation) with resolution of 15,000 (m/z 400) and target value of 5e4 in a data-dependent 'top 15' method with a dynamic exclusion of 30 s. The signal threshold was set to 5,000 for an isolation window of 2 Th and the first mass of HCD spectra to 80 Th. The collision energy was set to 35 eV.

**Data analysis.** All spectra were processed with MaxQuant[42] version 1.2.5.2 using the Andromeda search engine[43] to search the MS/MS spectra with trypsin specificity  against the IPI human data base (version 3.68, 87,061 entries) combined with 262 common contaminants. We allow for up to 2 missed cleavages and N-terminal acetylation and methionine oxidation were selected as variable, carbamidomethylation of cysteine was selected as fixed modification. For MS spectra an initial mass accuracy of 7 ppm was allowed and the MS/MS tolerance was set to 20 ppm for fragment detection in the Orbitrap analyzer for high resolution CID and HCD. A sliding mass window was applied to filter the MS/MS spectra for the 10 most abundant peaks in 100 Th. For identification, the peptide FDR was set to 0.0001. (The protein FDR remained at the standard setting of 0.01, but protein identifications were not directly used in this paper.) The shortest peptide length was set to 6 amino acids and the Max Quant feature to treat the isobaric amino acids leucine and isoleucine as indistinguishable for improved statics was disabled. This setting ensures that either amino acid matches the fragmentation spectrum as HCD in our set-up cannot distinguish them; however, side chain losses can then be assigned correctly because the isoleucine/leucine ambiguity is absent after database search. MaxQuant and Andromeda data processing provides access to the peptide sequences that were identified from the MS/MS spectra.  Detailed annotation of the MS/MS spectra was then carried out using the Expert System[36]. Results were further analyzed within the R scripting and statistical environment[44]. MaxQuant output tables are provided with this manuscript (Supplementary Tables 1-16 and column description in Supplementary Figure 2.) Raw mass spectrometric data are available at Tranche (www.proteomecommons.org) using the following hash code:

pI2oaLaSi7gPxUWNbesdXCgR17sWvMY6qVkHL+MtWA0Q5sqn/UxZVSjk3KpFTfrmDYpf3y/Iv6WfaAi6
HaILdZL0YocAAAAAAAAT7Q==

## Results and Discussion

**Generation of a high quality dataset.** To produce a diverse set of fragmentation spectra of tryptic peptides, we separated proteomes of *E.coli*, yeast and HeLa cells by one dimensional gel electrophoresis, excised eight slices and in-gel digested them (EXPERIMENTAL PROCEDURES). This generated a total of 24 complex peptide mixtures, which were analyzed using a 'high-high' strategy on a linear ion trap – Orbitrap instrument (LTQ Orbitrap Velos) using HCD as the fragmentation method. For a smaller number of fractions we also employed CID fragmentation followed by high resolution detection of fragments in the Orbitrap analyzer (EXPERIMENTAL PROCEDURES).

We wished to work with an extremely high quality set of fragmentation spectra in order to enable us to unambiguously attribute the observed fragments to the precursors. Therefore, we set the false discovery rate (FDR) for peptide identification by MaxQuant using the Andromeda search engine[42, 43] to 0.0001 rather than the customary 0.01. From our data set we obtained more than 100,000 MS/MS spectra that were identified with this very stringent criterion. We and others have recently introduced the notion of the Precursor Intensity Fraction (PIF)[45], chimeric or mixture MS/MS spectra[46, 47], which refers to the fact that precursor ions are frequently co-fragmented unintentionally in the analysis of complex peptide mixtures. For our purposes we needed to minimize the occurrence of co-eluting precursor ions in the isolation window so that they could not 'contaminate' the MS/MS spectra with unassignable peaks. This was achieved by only retaining spectra with a PIF greater than 0.95. If there was more than one spectrum for a particular sequence the one with the highest PIF was kept. Furthermore, we required the peptide length to be smaller than 26 amino acids and the charge state to be 2+, 3+ or 4+. These filters reduced the number of MS/MS spectra to about 16,000, which were nearly free of any contaminating peaks and which represented a broad sampling of typical tryptic peptides.

**Computer assisted annotation by the Expert System.** We recently developed a computer Expert System[36], which is now integrated into the Viewer component of the MaxQuant software environment. Briefly, the Expert System features a knowledgebase that was supplied with peptide fragmentation mechanisms described in the literature (see introduction) and with knowledge gained from manual evaluation of small and large-scale HCD data sets. These facts are implemented in a rule-engine that assigns annotations to the peaks in the MS/MS spectra. In order to avoid incorrect assignments, the Expert System follows strict dependencies among its rules. We derived a rigorous FDR for peak annotation, which made it possible to derive a minimal yet relative comprehensive set of rules[36].

Some MS/MS peaks can have an elemental composition that corresponds to more than one ion type and we have developed a strict ranking of the possible annotations to address this particular issue (Fig. 2A). Based on the identified peptide sequence, regular ions that result from cleavage of peptide bonds (b- and y-type ions), a-type ions that derive from the corresponding b-type ion by losing CO and c-type ions that occur in specific cases[20], are assigned the highest priority for annotation. The chemical structures of regular ions and immonium ions are different and as a consequence there is no possible overlap between them. Therefore the order of assignment is of no consequence and they are treated with the same priority. The second step covers annotations of neutral losses and internal fragment ions; these types derive from regular backbone ions. Importantly, neutral losses are specific to N- or C-termini of fragments or to a single or several amino acids. These are required to be contained in the peptide sequence to allow an annotation. Internal fragment ions originate from regular ions that have undergone a second cleavage of the peptide backbone. The side chains of the amino acids tryptophan (W), arginine (R) and lysine (K) are prone to produce specific fragment ions that can carry a proton due to the heteroatom in their chemical structure. Their mass is sufficiently large (> 100 Da) that they are recorded in HCD fragment ion spectra as side chain fragment ions. They are assigned a low priority because they are independent of any other ion type. Finally, incomplete fragmentation results in protonated precursor ions remaining in the MS/MS spectra, which are annotated as $[M+nH]^{n+}$.

The Expert System greatly improves on the number and intensity coverage of assigned peaks in the fragmentation spectra calculated by adding the signal for the ten largest peaks per sliding 100 Th window. Standard annotation by the Andromeda search engine results in an intensity coverage of up to 58% for pure spectra (PIF > 0.95; highlighted in gray in Figure 2B). Including the additional ion types that are covered by the Expert System increased the intensity coverage to 84%. With the Expert System in hand, we next annotated all of the about 100,000 high scoring fragment spectra in the initial set. This showed that even for impure MS/MS spectra (PIF less than 0.5), the intensity coverage of assigned peaks in MS/MS spectra was still above 50%. A typical MS/MS spectrum with a high PIF precursor that was comprehensively annotated by the Expert system is displayed in Figure 2C. Virtually all major peaks are correctly annotated and fragment intensity coverage reaches 87 %. The figure also illustrates the mass accuracy typically achieved in our experiment. Even though the lock mass feature during data acquisition was enabled[48], data analyzed with Max Quant is routinely independently recalibrated[49].

**Sequence related information content of HCD spectra.** The most important information imbedded in tandem mass spectra relates directly to the amino acid sequence of the peptide. Cleavage of all peptide bonds, resulting in b- and y-type ion series would in principle allow read out of the peptide sequence from the MS/MS spectrum in two directions starting from the N- or the C-terminus, respectively. Moreover, combining the band y-ion series highlights complementary b- and y-type ions pairs that together match the mass of the unfragmented peptide. Complementary pairs provide strong constraints for correct peptide identification and can be used in scoring algorithms even of multiplexed spectra[50].

In our large collection of HCD data we found nearly universal evidence for such pairs. Typical spectra have the prominent $a_2/b_2$ pair (observed in 72% of the peptide sequences) followed by at least a few more b-ions. Y-ion series were very abundant in our spectra, especially in the middle mass range (450 to 800 Da). For peptides that were not too long (< 20 amino acids), the low mass b-ion series almost always had a corresponding, complementary y-ion series of high intensity. These trends are well known from triple quadrupole and quadrupole time-of-flight spectra.

We next evaluated all 16,000 HCD spectra in the collection (Fig. 3A). Remarkably, for peptides up to 12 amino acids the y-ion series alone almost always provided complete sequence coverage (median 100%), indicating that complete sequencing of such peptides even in routinely acquired large-scale datasets is in principle possible. This includes the order of the two first amino acids, which is normally inaccessible because of the missing $y_{n-1}$ and $b_1$ ions (see below). With increasing peptide length the amino acid coverage slowly drops to a median of 50% at a peptide length of 25 amino acids, which was the upper limit in our collection (Fig. 3A). The b-ion series, in contrast, remains at a constant level, providing about 30% amino acid coverage independent of the peptide length. Taking both ion series together yields median amino acid coverage of 80% percent even for a peptide length of 20 AA.

Besides the percentage of the sequence that is covered by backbone fragmentation, another important parameter is the length of amino acids that can be read out from the MS/MS spectrum as an uninterrupted part of the sequence, i.e. the maximum sequence tags length[51]. A sequence tag of six amino acids is generally unique in the human genome even without added peptide mass information[6, 52]. In addition to peptide identification, such stretches are useful for partial de novo sequencing or homology searching. The lower panel in Figure 3A depicts the median sequence tag lengths based of the two different ion series. Peptides up to 10 amino acids contain a complete y-ion based sequence tag but above this length the $y_n$-1 ion is often of too low intensity to be recorded. Even small peptides contain short sequence tags of three amino acids, which are sufficient for peptide identification. When combined with the y-ion series, the b-ion series helps to increase the sequence tag length for peptides larger than 14 amino acids. The largest median sequence tag length is about 12 amino acids and it starts to drop from a peptide length of 16 amino acids.

We next compared the sequence related information content of HCD with that of high resolution CID spectra. A prominent difference is the much larger contribution of the b-ion series in CID spectra (Fig. 3B). This is due to the higher stability of b-ions in ion trap fragmentation processes. Although lower than the y-ion series, the b-ion series continued to provide a median of more than 50% sequence coverage up to a peptide length of 19 amino acids. Nevertheless, the combined contribution from y-ions and b-ions was slightly higher for HCD than for CID, which reflects the more extensive fragmentation in beam type instruments. Maximum sequence tag length was likewise higher in HCD spectra.

We have previously investigated maximum sequence tag lengths in low resolution CID spectra. In more than 85% of the identified spectra sequence tags of at least three amino acids and only in half of the spectra sequence tags of six or more amino acids were detected[52]. Despite the potential for over-counting due to the lower mass accuracy, these sequence tags were substantially shorter than tags from either high resolution HCD or high resolution CID.

**Neutral loss fragments in HCD.** During collision induced dissociation processes, peptides can follow numerous fragmentation pathways and consequently give rise to various ion types beyond those produced by the typical peptide backbone cleavage. A large class of such ions are those involving neutral losses from different fragment species. These occur from nearly all ion types, however, the chemical structures of the diverse ion types as well as the amino acid side chains allow specific neutral losses (Fig. 2A). In some cases, these can result either from the peptide terminus or from one of the side chains of the amino acids and localization of the origin is not straightforward. However, such losses can still be unambiguously assigned to the fragment ion. We carried out a systematic study considering 45 possible chemical compositions that could formally occur as neutral losses from amino acid residues. We then used our large scale dataset to determine the primary neutral losses for all of the fragments in the collection that contained the amino acid in question. The median absolute mass accuracy of all neutral losses is 2.7 ppm with 97.5% of the peaks within 5 ppm, therefore they can unambiguously be connected to their precursor fragments. Only first neutral losses which happened in at least 5% of the cases were considered and encoded in the Expert System[36]. Table 1 summarizes the observed frequencies of the primary neutral losses that occur in different combinations in more than 270,000 fragments. While b-type ions frequently loose a water molecule, the chemical structure of y-type ions allows both water and ammonia losses. These are by far the most frequent neutral losses. Furthermore, acidic amino acids as well as serine and threonine are likely to lose water. However, it was possible in about 48% of the cases to assign the neutral loss to either a specific amino acid or the C-terminus of the fragment, because there was only one possible origin for the water loss. At least 33% of the spectra from sequences that contain glutamic acid, serine or threonine exhibit water losses from those amino acids. This is the case in only 29% of spectra where the water loss can be confidently assigned to aspartic acid. The rate of ammonia losses is comparable to water losses and this also holds true for confidently assignable losses from glutamine (29%), asparagine (30%) and arginine (21%). Further frequently observed neutral losses that are specific to certain amino acids include $CH_3NO$ from glutamine (20%) and from asparagines (29%) or $C_2H_4O$ from threonine (26%). While other neutral losses may exist, our large data set suggests that they are unlikely to occur at substantial frequencies in HCD spectra.

**Internal fragments**. Internal fragments in the MS/MS spectra are characteristic of beam-type fragmentation because these result from ions undergoing a second cleavage resulting in a C-terminal carboxyl-group and an N-terminal oxazolone structure[13, 53]. In our large-scale data set, the length of internal fragments varied between two and more than ten amino acids, depending on peptide length. The majority of internal fragments, however, are shorter than five amino acids. Proline is most often the first amino acid of an internal fragment since N-terminal cleavage is very pronounced at this amino acid – this is called the proline effect[54]. However, we found that on the basis of peak presence, rather than peak intensity, proline initiated internal sequences were more than four times as common as those of the median of other amino acids (Supplementary figure 1A). For cleavage at the C-terminal amino acid of an internal fragment there is a slight preference for aspartic acid, glutamic acid, glutamine, tryptophan and histidine (Supplementary figure 1B). Proline is the least common amino acid at the C-terminus of internal ions.

**Low mass region.** HCD fragmentation takes place in a dedicated collision cell and is not subject to the low mass cut-off of ion trap CID spectra, therefore in principle allowing observation of the entire mass range. In practice, HCD spectra are normally acquired from m/z 100, but for a more extensive investigation of the low mass region we acquired data in our study from m/z 80, which was the lowest practical m/z without reducing the scan speed of the instrument. Therefore our dataset does not contain immonium ions with an m/z lower than 80 Th.

Figure 4B displays the frequency of immonium ions in the MS/MS spectra. The most prominent immonium ions originate from phenylalanine (F), tryptophan (W) and tyrosine (Y) and can be observed in at least 84% of all peptide sequences containing the respective amino acid. This is due to their chemical structure containing both a heteroatom and an aromatic system that are prone to stabilize a positive charge and for the same reason, the immonium ion of histidine (H) is often present (70%). Carbamidomethylated cysteine (caC), glutamine (Q) and glutamic acid (E) (61%, 52% and 37%, respectively) immonium ions can also be found relatively  in the spectra. Aspartic acid (D) and asparagine (N) produce a significantly lower rate of immonium ions. Interestingly, immonium ions of isoleucine (I) and leucine (L) are detected in the MS/MS spectra with different frequencies. Immonium ions of glycine (G), alanine (A), serine (S), proline (P), valine (V) and threonine (T) are not observed

in our data as their m/z is below 80 Th. Arginine (R) and lysine (K) represent special cases due to their position at the N-termini of tryptic peptides. A very frequently observed ion is the immonium ion of lysine with an ammonia loss (IM K – $NH_3$). In fact, this ammonia loss often occurs even without immonium ion and this was therefore implemented as an exception to the strict requirement for a detected precursor fragment in the Expert System. Immonium ions can be used to support the peptide sequence assignment. In special cases, such as phosphotyrosine (pY), immonium ions can be used as reporter ions to verify the existence and the nature a phosphorylation site (see below)[55, 56].
Another fragment ion type in the low mass region are fragment ions that result from cleavage of amino acid side chains in which the molecular structure can stabilize a proton. This is the case for some of the amino acids that contain a nitrogen atom, such as arginine, lysine and tryptophan. The chemical compositions of the side chain fragments and their frequency of occurrence is displayed in figure 5C. Note that these side chain fragments are different from the v-, w- and d-type ions from high energy CID dissociation carried out on TOF/TOF instruments[57, 58]. In addition to the general ion types, certain amino acid side chains follow different fragmentation pathways resulting into unusual ion types. Lehmann and co-workers observed $c_1$ ions resulting from the N-terminal amino acid of the peptide, if the second amino acid is glutamine (Q)[19, 20]. Along these lines we investigated asparagine and carbamidomethylated cysteine and found the same behavior for these two candidates. Furthermore, $b_1$ ions are usually not observed because of their chemical instability. However, we did observe $b_1$ ions from acetylated of methionine, serine or alanine at the protein N-terminus. This is thought to be due to stabilization of this fragment by the acetyl group[18, 59]. Besides the qualitative information contained in the variety of ion types of natural peptides, the low mass region in HCD fragmentation also gives access to the reporter ions for the TMT[60] and iTRAQ[33] quantification methods[34, 35]. The reporter ions of TMT and iTRAQ are at m/z 126.1277, 127.1248, 128.1344, 129.1315, 130.1411, 131.1382 and m/z 114.1112, 115.1146, 116.1116, 117.1150, respectively. Investigation of our large-scale and high accuracy data set revealed no interfering ions of the same m/z. Therefore problems in quantification by these methods are confined to co-fragmentation of other labeled peptides rather than other ion types that have the same mass as these reporter ions.

**Global composition of tryptic HCD spectra.** The different ion types covered by the Expert system, such as regular ions, neutral losses, internal fragments, immonium ions, side chain fragments and the intact peptide mass $[M+nH]^{n+}$ by their nature occur in MS/MS spectra with different frequencies (Figure 5A). However, for high confidence of peptide identification it is predominantly the highly abundant MS/MS peaks that are of interest. Figure 5B displays the contribution of each of the ion types to the overall intensity coverage:

Regular ions (a, b, c and y) account for 54% of total MS/MS spectra intensity and peaks that result from neutral losses for a further 15%. Immonium ions can originate from several amino acids and these signals are added as singly charged peaks at defined masses in the low mass region. Together, their mean contribution to the total intensity coverage is 6%. Unlike immonium ions, internal fragments are spread over the low to middle mass range of the MS/MS spectrum because they can be generated by any two cleavages of the peptide backbone and hence they are not as obvious in tandem mass spectra. As described above, in HCD internal fragment ions are frequently observed. However, their abundance is lower than that immonium ions or y-ions and together they contribute 10% to the total fragment intensity. The protonated unfragmented peptide precursor only has an average intensity coverage less than 1% in our dataset. Side chain fragments account for only 0.1% of total peaks and an intensity coverage of less than 0.1% and are therefore not displayed in the pie chart. The fraction of unannotated peaks accounts for 44% on the basis of the ten largest peaks per hundred Th but only for 15% with regard to total intensity coverage. This provides evidence that remaining peaks are mainly of low abundance. Note that those, beyond potentially being noise peaks, could also result from combinations of multiple neutral losses without precursor fragments or similar, which were not allowed by the Expert System to maintain a strict false positive rate. Furthermore, co-fragmentation of other precursors still occurs in our data set to some degree. Together, our data suggests that nearly all fragment peaks in HCD are explainable on the basis of current understanding of fragmentation pathways.

We next repeated the same analysis as above for high resolution CID spectra which resulted in quite similar findings for the number of peaks. As expected, the number of immonium ions and internal fragments was drastically reduced since ion trap fragmentation is not capable of retaining the low mass region of the tandem mass spectra and their formation requires double cleavage. Together with the higher preponderance of high mass b-ions, this has the effect of increasing the fraction of regular ions to 32% as compared to the 20% of HCD fragmentation. On the basis of intensity coverage, this effect is less pronounced (72% for CID compared to 54% for HCD). Interestingly, using the Expert System the fraction of unannotated peaks by intensity is very similar between CID (17%) and HCD (15%).

**Characteristics of phosphorylated peptides.** Protein phosphorylation is among to most important and best studied post-translational modifications and is almost always located at serine, threonine or tyrosine in mammalian cells. Due to its chemical nature, the phosphogroup easily detaches from serine and threonine during collision induced fragmentation processes resulting in very characteristic and abundant neutral loss peaks such as $HPO_3$ and $H_3PO_4$. Furthermore, as already mentioned above, phosphortyrosine leads to a unique and characteristic immonium ions with m/z 216.0426.

We investigated large scale phosphorylation data with the Expert System, incorporating rules for the above mentioned phosphospecific fragment ions. We found that the occurrence of both neutral losses from phosphorylated serine is about four times as high (65% for $HPO_3$ and 49% for $H_3PO_4$) as from threonine (18% and 12%, respectively). Table 2 summarizes the frequencies of these neutral losses. Their absolute number reveals an average of three $H_3PO_4$ losses and two $HPO_3$ losses per spectrum.

Finally, we investigated the frequency of $x_n$ ions pinpointing the localization of a serine or threonine phosphor site in the peptide sequence very recently described by Kelstrup et al.[61]. Our dataset consisting of 1157 spectra of phosphorylated peptide sequences contains this characteristic $x_n$ ion in 279 of the fragmentation patterns (24%).

## Conclusion and Outlook

HCD is one of the beam type collision induced dissociation methods and due to its implementation on the popular Orbitrap mass spectrometers, it is becoming very frequently used. In our group, for instance, both proteome and PTM-based investigations are routinely done with HCD rather than low or high resolution CID. This was one reason why it was important to investigate the ion types that are produced by HCD. However, even though the general dissociation mechanisms operative in CID have been studied for decades, it has not previously been possible to study large datasets with very high quality thresholds. This was made possible here by very stringent filtering of peptide fragment spectra on the basis of identification score as well as near absence of co-fragmenting peaks. Most importantly, we developed and made use of an Expert System, which annotated peptide peaks with high comprehensiveness but low false positive rates.

Our investigation of HCD yielded a broad and quantitative overview of the ion types produced. It turns out that HCD spectra are somewhat more complex than CID spectra but that the peaks are assignable to the same degree. The low mass region is particularly straightforward to interpret given the very high resolution of the Orbitrap analyzer in this region, coupled to the high mass accuracy, which generally allows determination of the chemical composition of these fragments. The information content of HCD spectra is mostly related to very extensive series of y-ions, supplemented by relatively short series of low mass b-ions. This is in contrast to ion trap CID spectra, in which the high mass b-ions are also very prominent. Nevertheless, the coverage of peptide sequence overall and in particular with continuous ion series is somewhat higher in HCD than it is in CID. Remarkably, for tryptic peptides up to 15 amino acids, the fragment contents is almost complete, meaning that there is sufficient information in principle for de novo sequencing or at least very long sequence tags.

Our quantification of the overall contribution of different ion types to the entire MS/MS spectrum revealed that only a relatively small proportion remains unassigned by the rules that we have implemented into the Expert System. This proportion would further shrink if noise and remaining co-fragmentation was further reduced and if the rules of the Expert System were relaxed. This means that the ion types produced in HCD and by extension by CID are already very well understood. New fragmentation pathways of standard peptides could of course be discovered in the future but it is unlikely that such ions would contribute very much to the overall ion intensity. For modified peptides, our Expert System and quantification of fragmentation frequencies could help to discover potential new fragment types. In this connection we have already demonstrated straightforward extension of our approach to phosphorylated peptides. In conclusion, we have here reported the most extensive investigation into HCD of peptides and hope that the results will be useful for both small and large scale investigation of the proteome.

**Acknowledgments***:* We thank our colleagues at the Max Planck Institute of Biochemistry for help and fruitful discussions, in particular Kirti Sharma for the preparation of the phosphoproteome samples. The research leading to these results has received funding from the European Commission's 7th Framework Programme (grant agreement HEALTH-F4-2008-201648 / PROSPECTS).

## TABLE 1

| | | | | | |
|---|---|---|---|---|---|
| **NH3** | 45% (N-term) | 30% (N) | 29% (Q) | 21% (R) | |
| **H2O** | 48% (C-term) | 37% (S) | 44% (T) | 21% (D) | 33% |
| **CO** | 84% (internal)* | | | | |
| **CO2** | 5% (D) | | | | |
| **CH2N2** | 8% (R) | | | | |
| **CH3NO** | 29% (N) | 20% (Q) | | | |
| **CH4O** | 5 % (S) | | | | |
| **CH4SO** | 89% (Mox) | | | | |
| **C2H4** | 5 % (I) | | | | |
| **C2H5NO** | 9% (N) | 6% (Q) | | | |
| **C2H4O** | 26% (T) | | | | |
| **C2H4O2** | 6 % (D) | 6% (E) | | | |
| **C3H6** | 6 % (L) | | | | |
| **C3H9N3** | 6% (R) | | | | |
| **C3H6SO** | 6 % (Mox) | | | | |
| **C3H8SO** | 12 % (Mox) | | | | |
| **C4H8** | 5 % (L) | | | | |
| **C8H7N** | 6 % (W) | | | | |
| **C9H9N** | 12 % (W) | | | | |

TABLE 1: Neutral losses considered by the Expert System and fraction of spectra that contain the loss from the corresponding amino acid. Only examples allowing unambiguous localization of the origin of the neutral loss were considered. * This ion is formally equivalent to an a-type internal fragment.

## TABLE 2

| | **$-HPO_3$** | **$-H_3PO_4$** | **pS** | **pT** | **$x_n$ (S,T)** |
|---|---|---|---|---|---|
| **S (1094)** | 65% (713) | 49% (540) | 29 | | 279 |
| **T (585)** | 18% (103) | 12% (68) | | 3 | |

TABLE 2: Fraction of 1157 spectra of modified sequences (phospho STY) containing neutral losses, reporter ions from phosphorylated serine (S) and threonine (T) or the characteristic x ion at least once. The first column lists the total number of sequences that contain the amino acid S or T at least once.

## References:

1. Ahrens, C. H.; Brunner, E.; Qeli, E.; Basler, K.; Aebersold, R., Generating and navigating proteome maps using mass spectrometry. *Nature reviews. Molecular cell biology* **2010,** 11, (11), 789-801.

2. Mallick, P.; Kuster, B., Proteomics: a pragmatic perspective. *Nature biotechnology* **2010,** 28, (7), 695-709.

3. Cox, J.; Mann, M., Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry* **2011,** 80, 273-99.

4. Nesvizhskii, A. I.; Vitek, O.; Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods* **2007,** 4, (10), 787-97.

5. White, F. M., The potential cost of high-throughput proteomics. *Science signaling* **2011,** 4, (160), pe8.

6. Ma, B.; Johnson, R., De novo sequencing and homology searching. *Molecular & cellular proteomics : MCP* **2012,** 11, (2), O111 014902.

7. Biemann, K., Contributions of mass spectrometry to peptide and protein structure. *Biomedical & environmental mass spectrometry* **1988,** 16, (1-12), 99-111.

8. Oomens, J.; Young, S.; Molesworth, S.; van Stipdonk, M., Spectroscopic evidence for an oxazolone structure of the b(2) fragment ion from protonated tri-alanine. *Journal of the American Society for Mass Spectrometry* **2009,** 20, (2), 334-9.

9. Bythell, B. J.; Somogyi, A.; Paizs, B., What is the structure of b(2) ions generated from doubly protonated tryptic peptides? *Journal of the American Society for Mass Spectrometry* **2009,** 20, (4), 618-24.

10. Perkins, B. R.; Chamot-Rooke, J.; Yoon, S. H.; Gucinski, A. C.; Somogyi, A.; Wysocki, V. H., Evidence of Diketopiperazine and Oxazolone Structures for HA b(2)(+) Ion. *Journal of the American Chemical Society* **2009,** 131, (48), 17528-17529.

11. Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A., Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of mass spectrometry : JMS* **2000,** 35, (12), 1399-406.

12. Boyd, R.; Somogyi, A., The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *Journal of the American Society for Mass Spectrometry* **2010,** 21, (8), 1275-8.

13. Paizs, B.; Suhai, S., Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* **2005,** 24, (4), 508-48.

14. Medzihradszky, K. F., Peptide sequence analysis. *Methods in enzymology* **2005,** 402, 209-44.

15. Papayannopoulos, I. A., The Interpretation of Collision-Induced Dissociation Tandem Mass-Spectra of Peptides. *Mass spectrometry reviews* **1995,** 14, (1), 49-73.

16. Falick, A. M.; Hines, W. M.; Medzihradszky, K. F.; Baldwin, M. A.; Gibson, B. W., Low-Mass Ions Produced from Peptides by High-Energy Collision-Induced Dissociation in Tandem Mass-Spectrometry. *Journal of the American Society for Mass Spectrometry* **1993,** 4, (11), 882-893.

17. Chalkley, R. J.; Baker, P. R.; Huang, L.; Hansen, K. C.; Allen, N. P.; Rexach, M.; Burlingame, A. L., Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset

acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. *Molecular & cellular proteomics : MCP* **2005,** 4, (8), 1194-204.

18. Hung, C. W.; Schlosser, A.; Wei, J. H.; Lehmann, W. D., Collision-induced reporter fragmentations for identification of covalently modified peptides. *Anal Bioanal Chem* **2007,** 389, (4), 1003-1016.

19. Winter, D.; Lehmann, W. D., Sequencing of the thirteen structurally isomeric quartets of N-terminal dipeptide motifs in peptides by collision-induced dissociation. *Proteomics* **2009,** 9, (8), 2076-84.

20. Lee, Y. J.; Lee, Y. M., Formation of c1 fragment ions in collision-induced dissociation of glutamine-containing peptide ions: a tip for de novo sequencing. *Rapid communications in mass spectrometry : RCM* **2004,** 18, (18), 2069-76.

21. Reid, G. E.; Roberts, K. D.; Kapp, E. A.; Simpson, R. J., Statistical and mechanistic approaches to understanding the gas-phase fragmentation behavior of methionine sulfoxide containing peptides. *Journal of proteome research* **2004,** 3, (4), 751-759.

22. Harrison, A. G.; Young, A. B.; Bleiholder, C.; Suhai, S.; Paizs, B., Scrambling of sequence information in collision-induced dissociation of peptides. *Journal of the American Chemical Society* **2006,** 128, (32), 10364-5.

23. Bleiholder, C.; Osburn, S.; Williams, T. D.; Suhai, S.; Van Stipdonk, M.; Harrison, A. G.; Paizs, B., Sequence-scrambling fragmentation pathways of protonated peptides. *Journal of the American Chemical Society* **2008,** 130, (52), 17774-89.

24. Goloborodko, A. A.; Gorshkov, M. V.; Good, D. M.; Zubarev, R. A., Sequence scrambling in shotgun proteomics is negligible. *Journal of the American Society for Mass Spectrometry* **2011,** 22, (7), 1121-4.

25. Yu, L.; Tan, Y.; Tsai, Y.; Goodlett, D. R.; Polfer, N. C., On the relevance of peptide sequence permutations in shotgun proteomics studies. *Journal of proteome research* **2011,** 10, (5), 2409-16.

26. Xia, Y.; Liang, X. R.; McLuckey, S. A., Ion trap versus low-energy beam-type collision-induced dissociation of protonated ubiquitin ions. *Anal Chem* **2006,** 78, (4), 1218-1227.

27. Schwartz, J. C.; Senko, M. W.; Syka, J. E. P., A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry* **2002,** 13, (6), 659-669.

28. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* **2007,** 4, (9), 709-12.

29. McAlister, G. C.; Phanstiel, D. H.; Brumbaugh, J.; Westphall, M. S.; Coon, J. J., Higher-energy collision-activated dissociation without a dedicated collision cell. *Molecular & cellular proteomics : MCP* **2011,** 10, (5), O111 009456.

30. Horner, J. A.; Remes, P.; Biringer, R.; Huhmer, A.; Specht, A., Achieving Increased Coverage in Global Proteomics Survey Experiments Using Higher-Energy Collisional Dissociation (HCD) on a Linear Ion Trap Mass Spectrometer. *Application note 538, Thermo Fisher Scientific, San Jose, CA, USA* **2011**.

31. 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 : MCP* **2009,** 8, (12), 2759-69.

32. Mann, M.; Kelleher, N. L., Precision proteomics: the case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** 105, (47), 18132-8.

33. Ross, P. L.; Huang, Y. L. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics* **2004,** 3, (12), 1154-1169.

34. Bantscheff, M.; Boesche, M.; Eberhard, D.; Matthieson, T.; Sweetman, G.; Kuster, B., Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Molecular & cellular proteomics : MCP* **2008,** 7, (9), 1702-13.

35. Pichler, P.; Kocher, T.; Holzmann, J.; Mohring, T.; Ammerer, G.; Mechtler, K., Improved precision of iTRAQ and TMT quantification by an axial extraction field in an Orbitrap HCD cell. *Anal Chem* **2011,** 83, (4), 1469-74.

36. Neuhauser, N.; Michalski, A.; Cox, J.; Mann, M., Expert System for Computer Assisted Annotation of MS/MS Spectra. *Molecular & cellular proteomics : MCP* **2012**, in revision.

37. Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M., In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature protocols* **2006,** 1, (6), 2856-60.

38. Rappsilber, J.; Ishihama, Y.; Mann, M., Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **2003,** 75, (3), 663-70.

39. Wisniewski, J. R.; Zougman, A.; Mann, M., Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *Journal of proteome research* **2009,** 8, (12), 5674-8.

40. Pinkse, M. W.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J., Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal Chem* **2004,** 76, (14), 3935-43.

41. Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, 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 : MCP* **2012,** 11, (3), O111 013698.

42. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **2008,** 26, (12), 1367-72.

43. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **2011,** 10, (4), 1794-805.

44. Ihaka, R.; Gentleman, R., R: A Language for Data Analysis and Graphics. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS* **1996,** 5, (3), 16.

45. 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. *Journal of proteome research* **2011,** 10, (4), 1785-93.

46. Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M., Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of proteome research* **2010,** 9, (8), 4152-60.

47. Wang, J.; Bourne, P. E.; Bandeira, N., Peptide Identification by Database Search of Mixture Tandem Mass Spectra. *Molecular & Cellular Proteomics* **2011,** 10, (12).

48. Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M., Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular & cellular proteomics : MCP* **2005,** 4, (12), 2010-21.

49. Cox, J.; Michalski, A.; Mann, M., Software lock mass by two-dimensional minimization of peptide mass errors. *Journal of the American Society for Mass Spectrometry* **2011,** 22, (8), 1373-80.

50. Ledvina, A. R.; Savitski, M. M.; Zubarev, A. R.; Good, D. M.; Coon, J. J.; Zubarev, R. A., Increased throughput of proteomics analysis by multiplexing high-resolution tandem mass spectra. *Anal Chem* **2011,** 83, (20), 7651-6.

51. Mann, M.; Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* **1994,** 66, (24), 4390-9.

52. Cox, J.; Hubner, N. C.; Mann, M., How much peptide sequence information is contained in ion trap tandem mass spectra? *Journal of the American Society for Mass Spectrometry* **2008,** 19, (12), 1813-20.

53. Ballard, K. D.; Gaskell, S. J., Sequential Mass-Spectrometry Applied to the Study of the Formation of Internal Fragment Ions of Protonated Peptides. *Int J Mass Spectrom* **1991,** 111, 173-189.

54. Harrison, A. G.; Young, A. B., Fragmentation reactions of deprotonated peptides containing proline. The proline effect. *Journal of mass spectrometry : JMS* **2005,** 40, (9), 1173-86.

55. Steen, H.; Kuster, B.; Fernandez, M.; Pandey, A.; Mann, M., Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Anal Chem* **2001,** 73, (7), 1440-8.

56. Boersema, P. J.; Mohammed, S.; Heck, A. J., Phosphopeptide fragmentation and analysis by mass spectrometry. *Journal of mass spectrometry : JMS* **2009,** 44, (6), 861-78.

57. Johnson, R. S.; Martin, S. A.; Biemann, K., Collision-Induced Fragmentation of (M+H)+Ions of Peptides - Side-Chain Specific Sequence Ions. *Int J Mass Spectrom* **1988,** 86, 137-154.

58. Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick, A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L., The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal Chem* **2000,** 72, (3), 552-8.

59. Yalcin, T.; Khouw, C.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G., Why are B ions stable species in peptide spectra? *Journal of the American Society for Mass Spectrometry* **1995,** 6, (12), 1165-1174.

60. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **2003,** 75, (8), 1895-904.

61. Kelstrup, C. D.; Hekmat, O.; Francavilla, C.; Olsen, J. V., Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. *Journal of proteome research* **2011,** 10, (7), 2937-48.

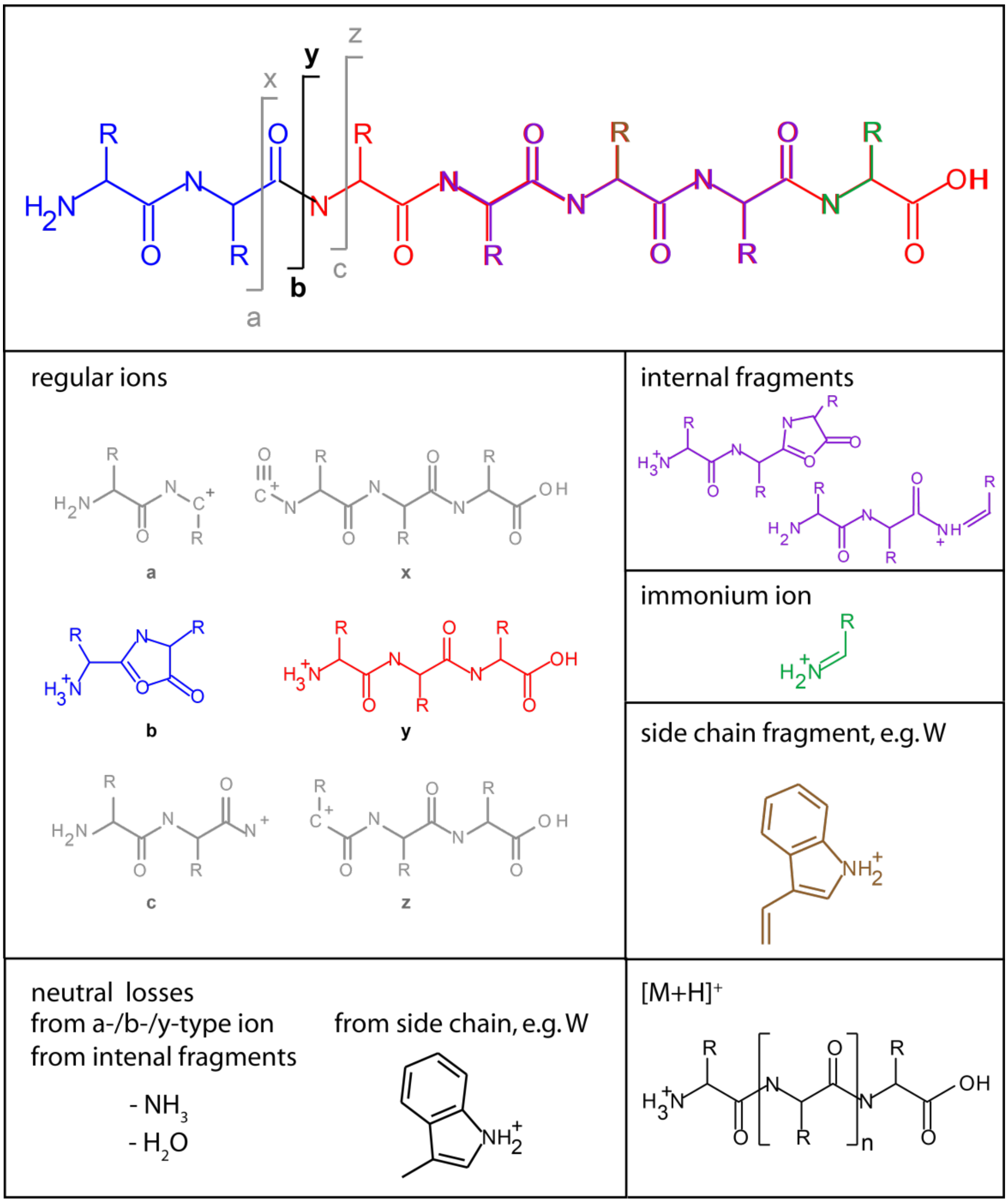


**FIGURE 1:** Cleavage sites of the peptide backbone giving rise to N-terminal a-, b- or c-type ions and the corresponding C-terminal x-, y- or z-type ions, respectively. The most prominent cleavage in CID and HCD fragmentation happens at the peptide bond. The boxes below, represent the most frequent ion types of collision induced fragmentation processes; the color code is providing their origin in the peptide sequence.

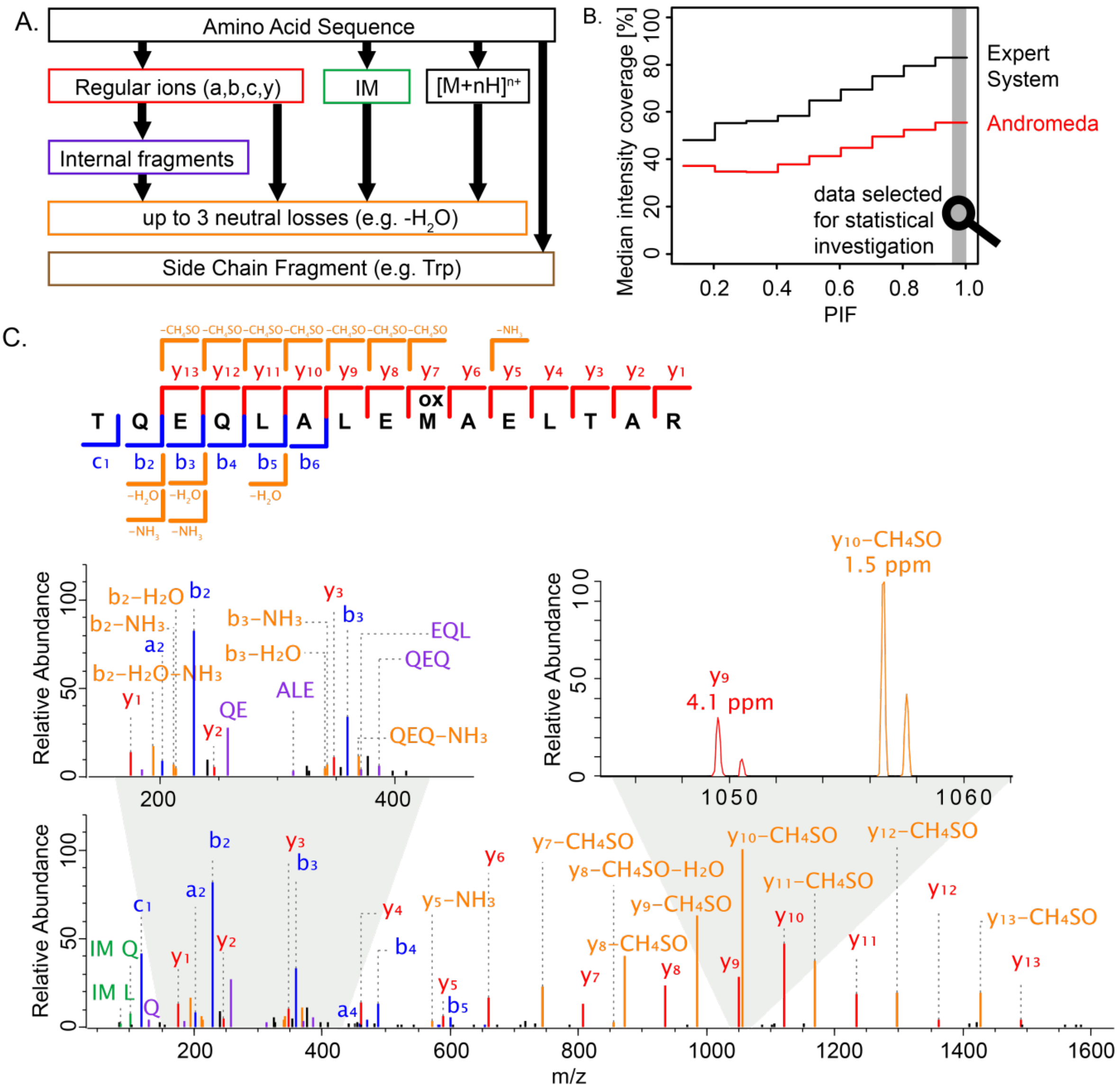


**FIGURE 2:** Peak annotation by the expert System **A.** Ranking of the six major ion types: intact precursor mass $[M+nH]^{n+}$, regular ions, immonium ions (IM), internal fragments, neutral losses and side chain fragments that are considered for peak annotation by the Expert System. **B.** Average intensity coverage of the total intensity of > 100,000 MS/MS spectra by standard search engine (Andromeda, red line) and by the expert system (black line) vs. the precursor intensity fraction 'PIF' provides a measure for the purity of precursor isolation. The high quality dataset (16,000 spectra) that was selected for statistical investigation is highlighted in grey. **C.** Typical MS/MS spectrum with PIF 0.99 annotated by the Expert System reaching an intensity coverage of 87%. A zoom window displays the high mass accuracy of two fragment ion peaks.

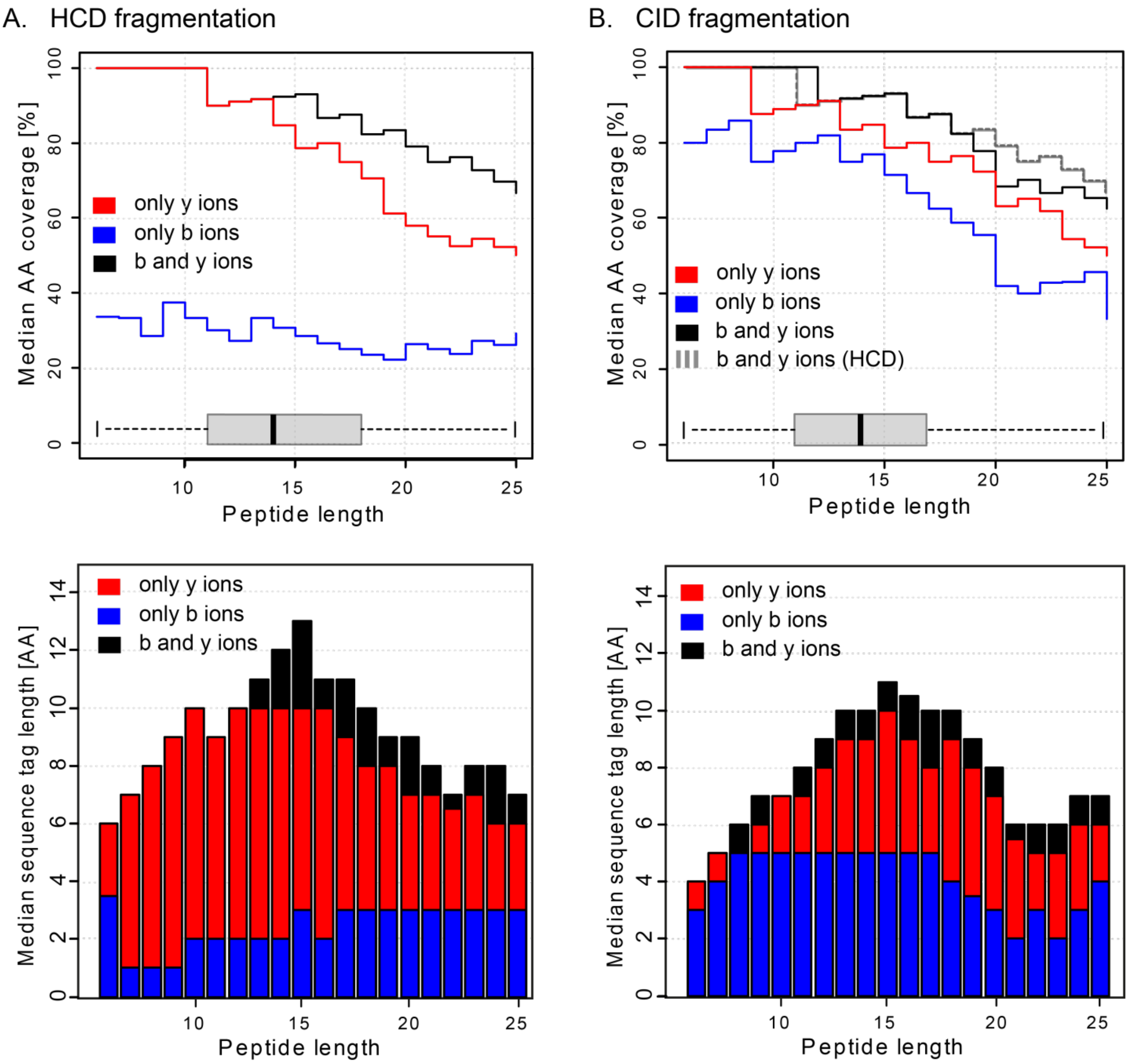


**FIGURE 3:** Sequence information content in HCD and CID spectra A. Median coverage of amino acids by y-type ions (red), b-type ions (blue) and both together (black) in the upper panel. The boxplot displays the distribution of the peptide length within the dataset (> 16,000 spectra); The lower panel shows the median length of the longest sequence tag based on y-type ions (red), b-type ions (blue) and both together (black). B. Same as (A) for a dataset of 3290 high resolution CID spectra. The dashed gray line in the upper panel repeats the median amino acid coverage in HCD from figure (A) for comparison.

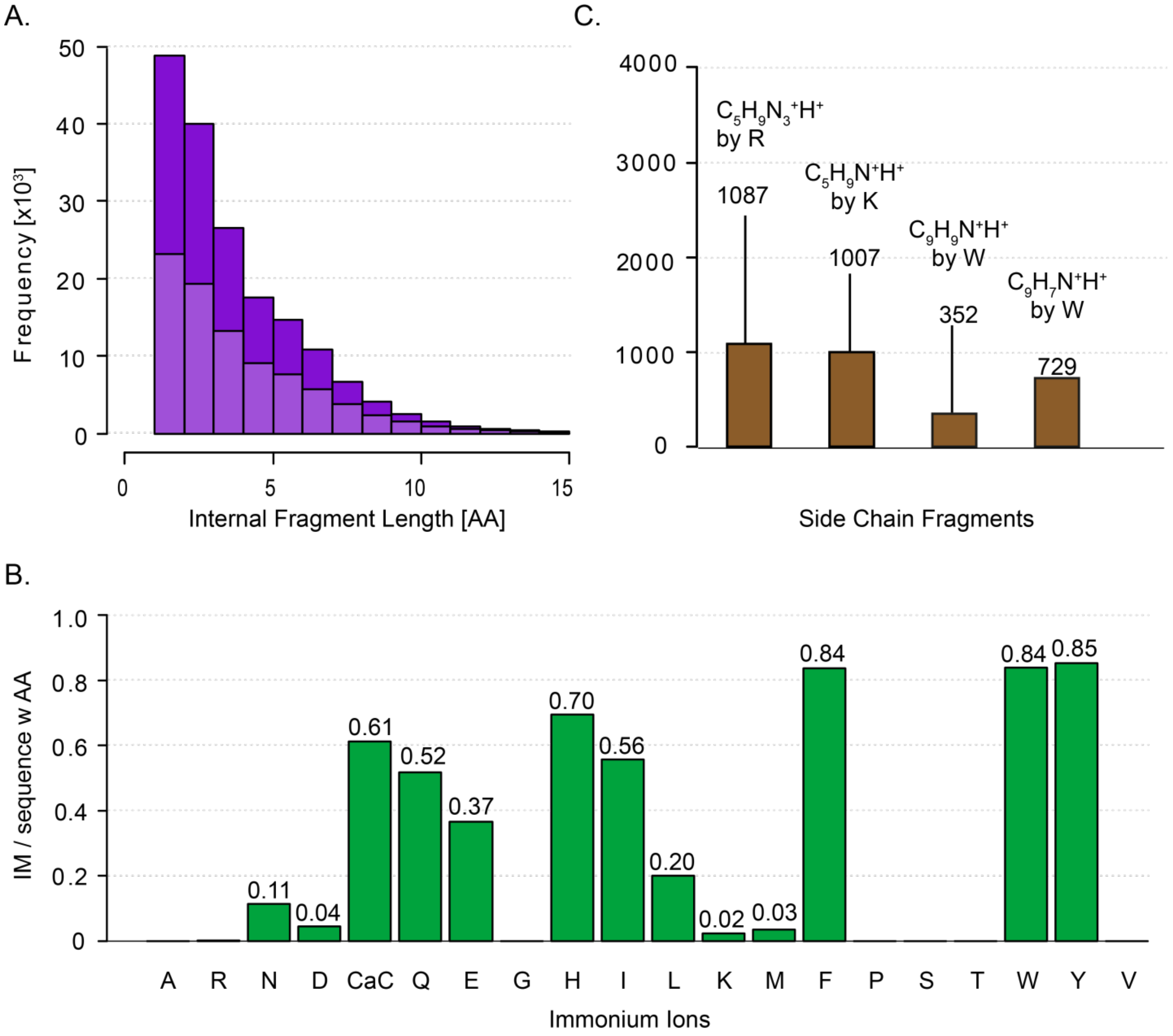


**FIGURE 4:** Statistics on the low mass region fragment ions from 16,000 MS/MS spectra **A.** Histogram of the length of all internal fragment ions in purple; the fraction of internal fragment ions starting with proline is highlighted by light color. **B.** Percentage of immonium ion (IM) occurrence if the amino acid corresponding amino acid was at least once contained in the peptide sequence; Immonium ions of Alanine, Glycine, Proline, Serine and Threonine were not considered, because their m/z value is lower than 80 Th. **C.** Bar plot displaying the five most abundant side chain fragment ions that are automatically assigned by the Expert System with their total number of occurrence within the dataset and their chemical structures.

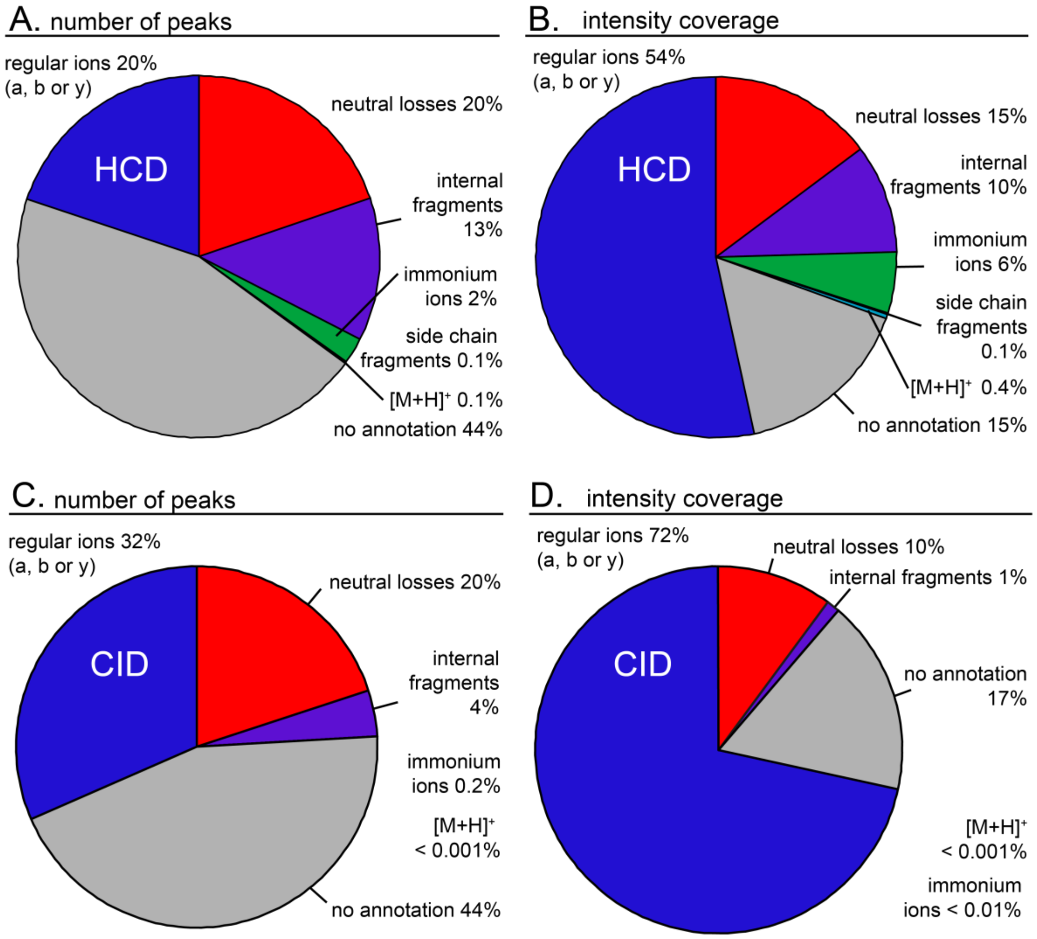


**FIGURE 5:** Intensity distribution of different ion types **A.** Average proportions of the six major ion types in HCD spectra by peak count based on a sliding mass window filtering for the 10 most abundant peaks per 100 Da; > 16,000 tandem mass spectra. **B.** same as (A) but referring to the intensity coverage of the MS/MS spectrum. **C.** and **D.** same as (A) and (B) for > 3,200 high resolution CID tandem mass spectra for comparison to the HCD ion type distribution.

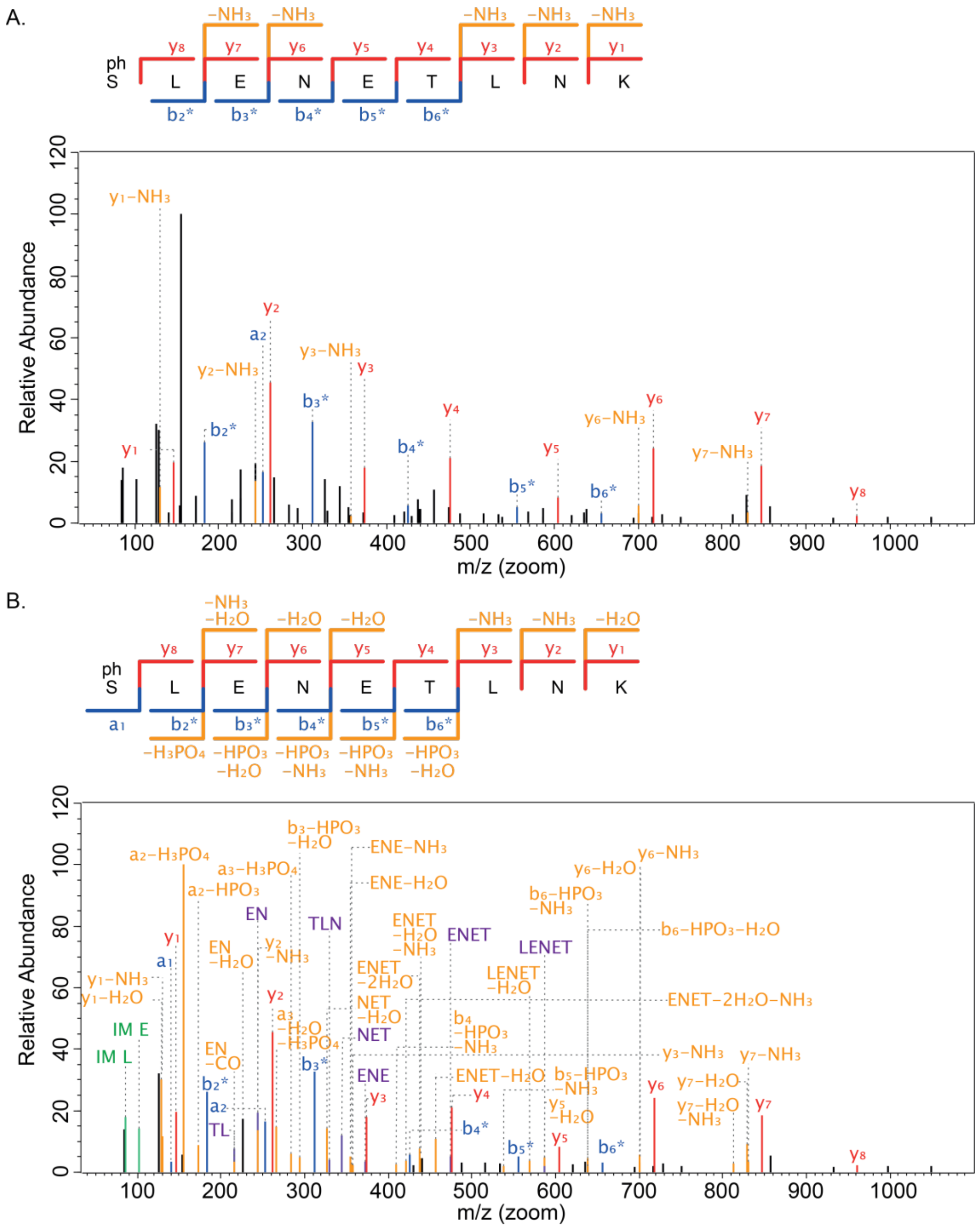


**FIGURE 6:** Annotated spectrum of phosphorylated peptide fragmented with HCD **A.** The phosphorylated peptide phSLENETLNK was identified and annotated by the Andromeda search engine assigning regular ions and single neutral losses. B. The Expert System was modified for phosphorylated peptides to enable comprehensive annotation: Several additional neutral losses, internal fragments and immonium ions increase the intensity coverage to 82%.

# Expert System for Computer Assisted Annotation of MS/MS Spectra

Nadin Neuhauser[&], Annette Michalski[&], Jürgen Cox and Matthias Mann[§]*

[&]These authors contributed equally

[§]Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

*To whom correspondence may be addressed. Email: mmann@biochem.mpg.de

**Running title**: Expert system for annotation of MS/MS spectra

**Abbreviations**: **FDR**, False Discovery Rate; **MS/MS**, Tandem mass spectrometry; **HCD**, Higher Energy Collisional Dissociation; **PIF**, Precursor Intensity Fraction; **PEP**, Posterior Error Probability; **PSM** peptide spectra match; **PDF**, Portable Document Format; **IM**, immonium ion; **SC**, side chain fragment ion; **Th**, Thomson

## Summary

**An important step in mass spectrometry (MS)-based proteomics is the identification of peptides by their fragment spectra. Regardless of the identification score achieved, almost all MS/MS spectra contain remaining peaks that are not assigned by the search engine. These peaks may be explainable by human experts but the scale of modern proteomics experiments makes this impractical. In computer science, expert systems are a mature technology to implement a list of rules generated by interviews with practitioners. We here develop such an expert system, making use of literature knowledge as well as a large body of high mass accuracy and pure fragmentation spectra. Interestingly, we find that even with high mass accuracy data, rule sets can quickly become too complex, leading to over-annotation. Therefore we establish a rigorous false discovery rate, calculated by random insertion of peaks from a large collection of other MS/MS spectra, and use it to develop an optimized knowledge base. This rule set correctly annotates almost all peaks of medium or high abundance. For high resolution HCD data, median intensity coverage of fragment peaks in MS/MS spectra increases from 58% by search engine annotation alone to 86%. The resulting annotation performance surpasses a human expert, especially on complex spectra such as those of larger phosphorylated peptides. Our system is also applicable to high resolution CID data. It is available both as a part of MaxQuant and via a webserver that only requires an MS/MS spectrum and the corresponding peptides sequence, and which outputs publication quality, annotated MS/MS spectra (www.biochem.mpg.de/mann/tools/). It provides expert knowledge to beginners in the field of MS-based proteomics and helps expert users to focus on unusual and possibly novel types of fragment ions.**

In MS-based proteomics, peptides are matched to peptide sequences in databases using search engines [1-3]. Statistical criteria are established for accepted vs. rejected peptide spectra matches (PSM) based on the search engine score, and usually a 99% certainty is required for reported peptides. The search engines typically only take sequence specific backbone fragmentation into account (i.e. a, b- and y-ions) and some of their neutral losses. However, tandem mass spectra – especially of larger peptides – can be quite complex and contain a number of medium or even high abundance peptide fragments that are not annotated by the search engine result. This can result in uncertainty for the user – especially if only relatively few peaks are annotated – because it may reflect an incorrect identification. However, the most common cause of unlabeled peaks is that another peptide was present in the precursor selection window and was co-fragmented. This has variously been termed 'chimeric spectra' [4-6], or the problem of low Precursor Ion Fraction (PIF) [7]. Such spectra may still be identifiable with high confidence. The Andromeda search engine in MaxQuant, for instance, attempts to identify a second peptide in such cases [8, 9]. However, even 'pure' spectra (those with a high PIF) often still contain many unassigned peaks. These can be due to different fragment types, such as internal ions, single or combined neutral losses as well as immonium and other ion types in the low mass region. A mass spectrometric expert can assign many or all of these peaks, based on expert knowledge of fragmentation and manual calculation of fragment masses, resulting in a higher degree of confidence for the identification. However, there are more and more practitioners of proteomics without in depth training or experience in annotating MS/MS spectra and such annotation would in any case be prohibitive for hundreds of thousands of spectra. Furthermore, even human experts may wrongly annotate a given peak – especially with low mass accuracy tandem mass spectra – or fail to consider every possibility that could have resulted in this fragment mass.

Given the desirability of annotating fragment peaks to the highest degree possible, we turned to 'Expert Systems', a well-established technology in computer science. Expert Systems achieved prominence in the 1970s and 1980s and were meant to solve complex problems by reasoning about knowledge [10, 11]. Interestingly, one of the first examples was developed by Nobel Prize winner Joshua Lederberg more than 40 years ago, and dealt with the interpretation of mass spectrometric data. The program's name was Heuristic DENTRAL [12], and it was capable of interpreting the mass spectra of aliphatic ethers and their fragments. The hypotheses produced by the program described molecular structures that are plausible explanations of the data. To infer these explanations from the data, the program incorporated a theory of chemical stability that provided limiting constraints as well as heuristic rules.

In general, the aim of an Expert System is to encode knowledge extracted from professionals in the field in question. This then powers a rule-based system that can be applied broadly and in an automated manner. A rule-based expert system represents the information obtained from human specialists in the form of IF-THEN rules. These are used to perform operations on input data to reach appropriate conclusion. A generic Expert system is essentially a computer program that provides a framework for performing a large number of inferences in a predictable way, using forward or backward chains, backtracking and other mechanisms [13]. Therefore, in contrast to statistics based learning, the 'expert program' does not know what it knows through the raw volume of facts in the computer's memory. Instead, like a human expert, it relies on a reasoning-like process of applying an empirically derived set of rules to the data.

Here we implemented an expert system for the interpretation for high mass accuracy tandem mass spectrometry data of peptides. The expert system was developed in an iterative manner together with human experts on peptide fragmentation, using the published literature on fragmentation pathways as well as large data sets of HCD [14] and CID based peptide identifications. Our goal was to achieve an annotation performance similar or better than experienced mass spectrometrists, thus making comprehensively annotated peptide spectra available in large scale proteomics.

## EXPERIMENTAL PROCEDURES

The benchmark data set is from the accompanying paper by Michalski et al. [15]. Briefly, E. coli, yeast and HeLa proteomes were separated on 1D gel electrophoresis and in gel digested [16]. Resulting peptides were analyzed by LC MS/MS on a linear ion trap – Orbitrap instrument (LTQ Velos [17] or ELITE [18], Thermo Fisher Scientific). Peptides were fragmented by Highr Energy Collisional Dissociation (HCD) [14] or by Collisional Induced Dissociation (CID), but in either case fragments were transferred to the Orbitrap analyzer to obtain high resolution tandem mass spectra (7,500 at m/z 400). We scanned tandem mass spectra already from m/z 80 to capture immonium ions as completely as possible. Data analysis was performed by MaxQuant using the Andromeda search engine [8, 9]. Maximum initial mass deviation for precursor peaks was 6 ppm and maximum deviation for fragment ions for both the search engine and for the expert system was 20 ppm. MaxQuant preprocessed the spectra to be annotated by the Expert System in the same way as it does for the Andromeda search engine: Peaks were filtered to the 10 most abundant ones in a sliding 100 m/z window, de-isotoped and shifted to charge one where possible. Raw data are available at Tranche (www.proteomecommons.org) as indicated in ref [15]. From this data, sequence-spectra pairs were selected that had PIF values [7] larger than 95% and that were sequence unique (more than 16,000 peptides).

The Expert System was written in the programming language C#, using the Microsoft .NET framework version 3.5 and the Workflow.Activities library, which contains a rule engine to implement an Expert System (Microsoft Corporation, Redmond, USA).

MaxQuant contains the Expert System as an integrated option in its Viewer – the component that allows visualization of raw and annotated MS data. MaxQuant can freely be downloaded from www.maxquant.org. It requires Microsoft .NET 3.5, which is either already installed with Microsoft Windows or can be installed as a free Windows update. In our group we have implemented the expert system both on a Windows cluster and in a desktop version. Additionally we provide an Expert System web server which can be accessed at www.biochem.mpg.de/mann/tools/. While MaxQuant allows the Expert System annotation of arbitrary numbers of MS/MS spectra, the webserver is currently limited to the submission of one MS/MS spectrum at a time. After upload of a list of peaks with m/z value and their intensities – together with the corresponding peptide sequence – the spectrum with all annotations is displayed. This can then be exported in different graphical formats.

## RESULTS AND DISCUSSION

*Construction of the expert system* – Human experts perform a generic set of tasks when solving problems such as the interpretation of an MS/MS spectrum. These rules have to be codified in the Expert System, mainly in the form of a series of IF-THEN rules. Figure 1 shows the major steps involved in building and using the Expert System. It is important to acquire all relevant rules to interpret MS/MS spectra as comprehensively as possible. However, to avoid over-annotation leading to false positives (see below), the number of rules and their interactions should not become too large. This balance was struck by evaluating the performance of different set of rules on large datasets in conjunction with human experts.

Rules were encoded in a table-like structure, where they could be activated, deactivated or modified. To create the knowledge base, the extent of interactions of the rules also had to be determined - for instance, which combination of neutral losses to allow. After iterative construction of the knowledge base, the rule engine then applied the encoded knowledge to MS/MS spectra and displayed the result to the user (Fig. 1A). The processing steps that are performed on the raw MS and MS/MS spectra are shown in Figure 1B (see also EXPERIMENTAL PROCEDURES). Note that the workflow is entirely automated and that user interaction is possible but not required. Arbitrary numbers of annotated spectra of peptides of interest can be produced as interactive screen images or high resolution, printable PDF files. The expert system is very fast, and 16,000 spectra can be annotated in less than four hours on a desktop system.

The IF-THEN constraints of our Expert System can be divided into four major parts (Fig. 2). At first the Expert System calculates any specific backbone fragments (a, b and y-ion series), the charged precursor ion, the immonium ions as well as side chain fragments in the low-mass region and places them into a queue. In the second part of the workflow every element in this queue is filtered with respect to the actual MS/MS spectrum. Even if there is a peak corresponding to a calculated item in the queue, it may still be filtered out (symbolized by missing annotations after the filter in Fig. 2). For instance, a b1 ion is only allowed in very restricted circumstances (for explanation see ref [15]).

In the 3rd step, neutral losses and internal fragments for the filtered values are calculated and added to the queue. They are then subjected to the same filtering rules as in step 2. Step 3 is iterative, as several subsequent neutral losses may be allowed.

In the 4th and last step each potential annotation is given a priority. If there is more than one possible annotation, the one with the highest priority is chosen (i.e. the one that triggered the rules with higher priority). However, in this case the Expert System provides a pop-up (or 'tool-tip') containing the other possibility when hovering the mouse over the peak. (This can still happen if the FDR is properly controlled and is then typically caused by two different chemical designations for the same ion; or by different ions with the same chemical composition, such as small internal fragments with different sequence but the same amino acids).

*Determining an FDR for peak annotation* – Use of a very high threshold for peptide identification ensured that virtually none of the peptides in our collection should be misidentified. However, when building the Expert System, we noticed that it was still possible to over-interpret the MS/MS spectra. This was initially surprising to us because our large scale data set had good signal to noise and peaks are only candidates for annotation when their calculated mass was less than 20 ppm from the observed mass. The over-interpretation became apparent through conflicting annotations for the same peak, and was typically due to a combination of rules, such as several neutral losses from major sequence specific backbone or internal ions. Because conflicting or wrong annotations would undermine the entire rational for the Expert System, we devised a scheme to stringently control the false discovery rate for peak annotation.

The FDR is meant to represent the percent probability that a fragment peak is annotated by chance because its mass fits one of the Expert System rules for the peptide sequence. To calculate a proper FDR, we therefore needed to provide a set of background peaks that would represent false positives when they are labeled by the Expert System. Producing realistic background peaks turned out to be far from trivial because they need to have possible masses that can in principle be generated from peptide sequences and they need to be independent of the sequence of the peptide in question. The principle of our solution to this problem is shown in Figure 3A. From the large data set underlying this study, we collect the m/z values of all annotated peaks, except those coming from immonium or side chain ions. They were stored in a large peak collection of several million entries, together with the respective peptide sequences and the relative intensity of the peak. For each spectrum in which we wanted to determine the FDR, we then inserted a random set of 10 peaks from the collection, where after we checked if the sequence of the selected peaks is independent from the sequence of the current spectrum.  If one of the inserted peaks overlapped with an existing peak, it was discarded. By definition these 10 peak masses represent possible peptide fragments and, because they are chosen randomly from millions of other peaks, they collectively represent a good approximation to a true background set. This would not be the case for permutation of the sequence of the precursor in question, for instance, because many of the fragment peaks in permutated sequences are identical. Whenever the Expert System annotated one of these peaks, it is counted as a false positive. To find the number of repeats necessary to obtain a stable FDR for this procedure, we chose a number of spectra and simulated a thousand times of each one. We found that the FDR was constant after 500 iterations. For the final FDR calculation, for each spectrum we added a different set of 10 random peaks from the collection and repeated this 500 times. This was then applied to each of the more than 16,000 pure (high PIF) spectra in the large scale data set.

Beyond providing a solid FDR estimate for each rule set, this procedure also allowed us to identify the rules or rule combinations that were responsible for miss-annotation – i.e. the rules that falsely annotated the inserted peaks. These mostly turned out to be chains of subsequent neutral losses. In conjunction with detailed evaluation of the frequency of ion types [15], we iteratively designed an optimal rule set (Supplementary Table 1). For instance, neutral losses from a particular amino acid were allowed if they occurred in more than five percent of the fragment sequences that contained that amino acid. Likewise, of a set of about 42 possible neutral side chain losses, only six were sufficiently important to retain them in the Expert System. The Figures 3B-D show the results of the median FDR as a function of the peptide length based on this final rule set. The overall FDR - indicated in red - is the same in all plots and shows a clear growing trend in the number of false positives with the length of the peptides. For small peptides of 12 amino acids or less, the FDR was less than 2.1 % and all peptides in the range investigated had a peak annotation FDR of less than 5%. With these settings, the annotations are correct in more than 97% the cases for the vast majority of MS/MS spectra. The Expert System could of course be pruned to provide a lower FDR by narrowing the mass tolerance window; however, this would come at the expense of discarding correct annotations. To explore the influence of mass accuracy on potential false positive annotations, we repeated these calculations with required mass deviations no larger than 5 ppm or no larger than 10 ppm. As can be seen in Figure 3B, this further reduced possible errors to less than 1%, or less than 0.3 %, respectively. This highlights the value of high mass accuracy in unambiguously identifying fragment mass identity.

Furthermore, peaks with a low signal to noise are more likely to be miss-annotated than more intense peaks. In Figure 3C we sorted the peak intensity of the false positives into three intensity classes (Fig. 3C). The median FDR of peaks with high or medium abundance are only 0.1 or 0.5 %. For low abundance peaks it is higher but still with a median of no more than 2.1 %.

Next we separately investigated the FDR as a function of peptide length for the different fragment ion types. As can be seen in Figure 3C, regular ions and internal fragments contribute very little to overall false annotation (0.4 and 0.5 %), whereas neutral loss ions are wrongly annotated in 1.8 % of the case or even more.

*Performance of the Expert System* - Figure 4 shows an illustrative example of an HCD fragmented peptide before and after Expert System evaluation. The peptide was identified with an Andromeda score of 136 and Posterior Error Probability (PEP) of 1.1E-21 (the corresponding Mascot score was 83). The spectrum features an uninterrupted b-ion series from $b_2$ to $b_9$ and an uninterrupted y-ion series from $y_1$ to $y_{12}$, together covering the entire peptide sequence. Despite this unambiguous identification, the peaks used by the search engine to identify the peptide only accounted for 35% of the summed intensity of the peaks in the fragmentation spectrum. Coverage by number of explained peaks was even lower at 24% (allowing up to 10 peaks per 100 Th in the measured spectrum see EXPERIMENTAL PROCEDURES). There is a series of high abundance, high m/z fragments as well as a large number of low abundance peaks in the low and medium m/z range that are unexplained by the search engine. After annotation by the Expert System, this situation changes entirely. The high m/z series is revealed to be due to prominent loss of $CH_4SO$ from oxidized methionine. The low mass ions are due to neutral losses, internal fragments and combinations between them and they were unambiguously and correctly assigned. Altogether, the expert system accounted for almost all prominent ions and explained a total of 88% of the ion current. Manual annotation of this spectrum would have been possible but would have been very time consuming.

Interpretation of phosphorylated peptides – especially large ones – is more difficult than that of unmodified peptides. Furthermore, accurate placement of the phosphorylation site can be challenging. We used literature knowledge [19, 20] and the results of a large-scale investigation into the fragmentation of phosphorylated peptides to derive suitable fragmentation rules for the Expert System [15]. This led to an additional six rules, which were easily integrated - illustrating the extensibility of the Expert System. Figure 4B depicts an example annotation of the relatively complex fragmentation spectra typical of phosphorylated peptides. The large ion series from the low mass range to about mass 1000 is due to an extensive and uninterrupted internal ion series starting from the proline in the 2nd position of the peptide sequence. As these internal fragments contain several glutamines, they lead to additional water and ammonia losses. However, there are also newly annotated fragments resulting from neutral losses in addition to loss of the phosphorylation site. Moreover, the neutral loss of $HPO_3$ is annotated.

*Large-scale evaluation of the performance of the Expert System* – We used the population of 16,000 spectra with high PIF - identified regarding a false discovery rate of 0.01% by the search engine - and annotated them automatically using the Expert System. For each spectrum we calculated the intensity coverage obtained by the fragments used by the search engine and the fragments explained by the Expert System. Higher scoring fragmentation spectra would be expected to have a larger fraction of their ion current annotatable than lower scoring peptides. Figure 5A shows a plot of the median of these values for all search engine scores. A total of 95% of these Andromeda scores are within a range of 96 to 138. Here the median intensity coverage by standard annotation varies from 55% at 96 to 64% at 138. The Expert System, in contrast, annotated between 86% and 89% of the total ion current in the fragment spectra of the same peptides. This represents an average increase of 28%. There was only a small percentage of peptides that were lower scoring than 96 and for these the increased annotation percentage due to the Expert System was even larger (34%). Interestingly, even in very high scoring HCD fragment spectra there are still many peaks not directly annotated by the search engine. For these, the average increase of annotated ion current due to the Expert System was still 23%.

The rule set of the expert system was derived from HCD data. However, HCD and CID appear to produce similar ion types, although with different abundances [15]. We therefore tested if the derived rule set was also applicable to high resolution CID data. This was indeed the case, and a total of 85% of the ion current in high resolution CID spectra explained by the Expert System, although in CID spectra a higher percentage (79%) of the peaks are already accounted for by standard ion types. Therefore we conclude that the Expert System can be used equally well for high resolution HCD and CID data although the benefits are not as large as they are for HCD data.

*Webserver for Expert System annotation of spectra* - The Expert System is now part of the viewer component of MaxQuant, which is freely available at www.maxquant.org. In this environment, the Expert System can annotate arbitrarily large data sets of identified peptides and visualize and export them in different graphical formats such as PDF. Additionally, we established a webserver to make the Expert System available to any proteomics scientist, regardless of the computational workflow that he or she is using. The webserver is located at http://www.biochem.mpg.de/mann/tools/ and its graphical interface is shown in Figure 6. The user needs to supply a mass spectrum in the form of an m/z and peak intensity list as well as the sequence of the identified peptide (Fig. 6A, 6B). Common modifications and their position in the sequence can also be specified. The webserver then provides an annotation of the spectrum within the stated mass tolerance as shown in Figure 6C. The graph is scalable to enable detailed study of complex fragmentation spectra. Mass deviations in ppm (calculated mass – measured mass) can also be depicted. This annotated spectrum can be downloaded in a number of graphical formats for use in publications.

## CONCLUSION AND OUTLOOK

Here we have made use of Expert Systems - a well-known technology in computer science – to automatically but accurately interpret the fragmentation spectra of identified peptides. We have shown that the Expert System performs very well on high mass accuracy data, annotating the large majority of medium to high abundance peaks. For HCD spectra it explains on average 28% more of the peak intensities than the search engine results alone. We derived a rigorous false positive rate, ensuing that less than 5% of peaks can be miss-annotated – this rate is even lower for spectra with at least median scores and fragment ion intensities of at least moderate abundance. The rule set was derived by iterative interpretation of large HCD dataset but we show that the Expert System is equally applicable to high resolution CID spectra.

We envision different uses for the Expert System: For beginners in MS-based proteomics, it enables efficient training in the interpretation of MS/MS spectra without requiring much input from a specialist. For experts, it allows focusing on unusual and potentially novel types of fragments. One caveat is that the Expert System currently cannot explain fragmented peaks that belong to co-fragmented precursors; a very common occurrence that we deliberately avoided here by selecting only pure MS/MS spectra. This limitation can be addressed if both precursors are identified and communicated to the Expert System. Such a feature might be particularly useful for instruments that allow deliberate multiplexing of precursors, which leads to complex MS/MS spectra [21].

The Expert System has been in routine use in our laboratory for a number of months. During this time we have found that it provides helpful confirmation of the identification of the peptide and the identity of the previously unlabeled fragment ions. This is particularly welcome in the case of complicated spectra of important peptides, such as the ones regulated in the biological function in question. Compared to a human expert, the principal advantages of the Expert System are its speed, its ability to check for all supplied rules in a consistent manner as well as its rigorously controlled false positive rate. Obviously, the Expert System is limited to the knowledge supplied whereas an experienced mass spectrometrist can go beyond these rules and discover the origin of novel fragmentation mechanisms.

As we have shown here, Expert Systems can readily be applied to problems in computational proteomics. Given their relative ease of implementation, they may become useful in other areas in MS-based proteomics, too.

# REFERENCES

1. Steen H and Mann M (2004) The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol 5:699-711.

2. Nesvizhskii AI, Vitek O and Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 4:787-97. doi: 10.1038/nmeth1088

3. Granholm V and Kall L (2011) Quality assessments of peptide-spectrum matches in shotgun proteomics. Proteomics 11:1086-93. doi: 10.1002/pmic.201000432

4. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG and Old WM (2010) Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. J Proteome Res 9:4152-60. doi: 10.1021/pr1003856

5. Zhang N, Li XJ, Ye M, Pan S, Schwikowski B and Aebersold R (2005) ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics 5:4096-106. doi: 10.1002/pmic.200401260

6. Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ and MacCoss MJ (2010) Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. Anal Chem 82:833-41. doi: 10.1021/ac901801b

7. Michalski A, Cox J and Mann M (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res 10:1785-93. doi: 10.1021/pr101060v

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

9. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV and Mann M (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10:1794-805. doi: 10.1021/pr101065j

10. Giarratano JC and Riley G (2005) Expert systems : principles and programming. PWS Pub. Co., Boston.

11. Liao SH (2005) Expert system methodologies and applications - a decade review from 1995 to 2004. Expert Systems with Applications 28:93-103. doi: DOI 10.1016/j.eswa.2004.08.003

12. Schroll G, Duffield AM, Djerassi C, Buchanan BG, Sutherland GL, Feigenbaum EA and Lederberg J (1969) Applications of artificial intelligence for chemical inference. III. Aliphatic ethers

diagnosed by their low-resolution mass spectra and nuclear magnetic resonance data. Journal of the American Chemical Society 91:7440-7445.

13. Russell SJ, Norvig P and Davis E (2010) Artificial intelligence : a modern approach. Prentice Hall, Upper Saddle River.

14. Olsen JV, Macek B, Lange O, Makarov A, Horning S and Mann M (2007) Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods 4:709-12.

15. Michalski A, Neuhauser N, Cox J and Mann M (2012) A systematic investigation into the nature of HCD spectra. Mol Cell Proteomics:submitted.

16. Shevchenko A, Tomas H, Havlis J, Olsen JV and Mann M (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 1:2856-60.

17. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M and Horning S (2009) A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. Mol Cell Proteomics 8:2759-69.

18. Michalski A, Damoc E, Lange O, Denisov E, Nolting D, Muller M, Viner R, Schwartz J, Remes P, Belford M, Dunyach JJ, Cox J, Horning S, Mann M and Makarov A (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol Cell Proteomics 11:O111 013698. doi: 10.1074/mcp.O111.013698

19. Boersema PJ, Mohammed S and Heck AJ (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. J Mass Spectrom 44:861-78. doi: 10.1002/jms.1599

20. Kelstrup CD, Hekmat O, Francavilla C and Olsen JV (2011) Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. J Proteome Res 10:2937-48. doi: 10.1021/pr200154t

21. Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M and Horning S (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics 10:M111 011015. doi: 10.1074/mcp.M111.011015

*Acknowledgments* – We thank Forest White for critical comments on this manuscript. This work was supported by funding from the European Union 7th Framework project PROSPECTS (Proteomics Specification in Time and Space, grant HEALTH-F4-2008-201645)

# FIGURES

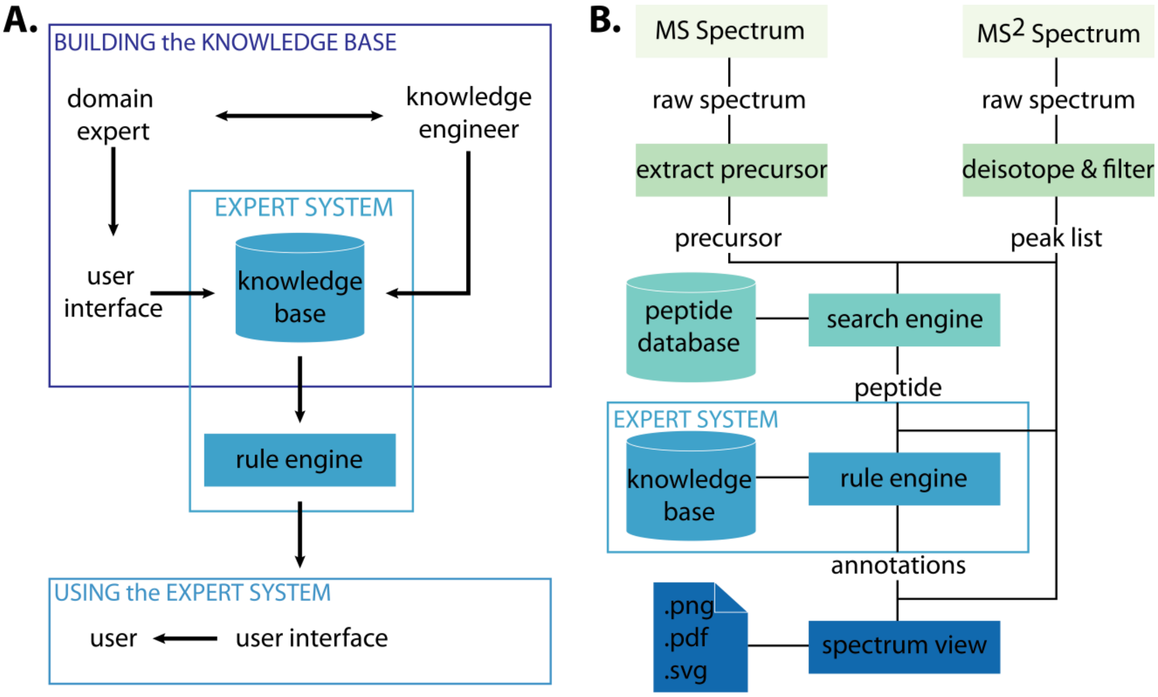


**FIGURE 1** **Basic concept of the Expert System.** A. An Expert System is constructed by interviewing an expert in the domain (here peptide fragmentation and the accumulated literature) and devising a set of rules with associated priority and dependence on each other. The knowledge base contains the rules whereas the rule engine is generic and applies the rules to the data. B. Data are automatically processed following the steps depicted.

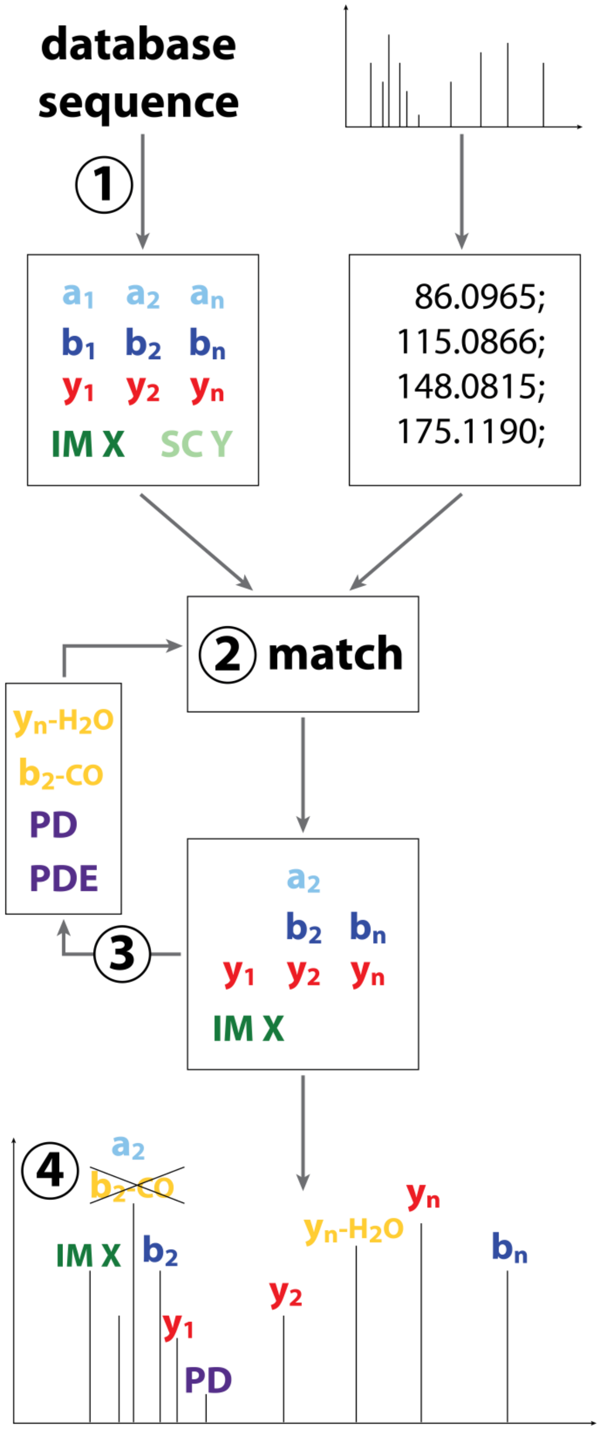


**FIGURE 2** **Work flow of the Expert System.** (1) From the database sequence of the peptide identified by the search engine, a list of possible fragment ions is created. (2) Peaks from the measured spectrum are compared with the possible fragments and preliminarily annotated if they pass the rules of the Expert System. (3) Neutral losses and internal fragments are generated from the candidate, annotated peaks and exposed to the Expert System rules. (4) Potential conflicts are resolved via the priority of the annotations and peaks are labeled.

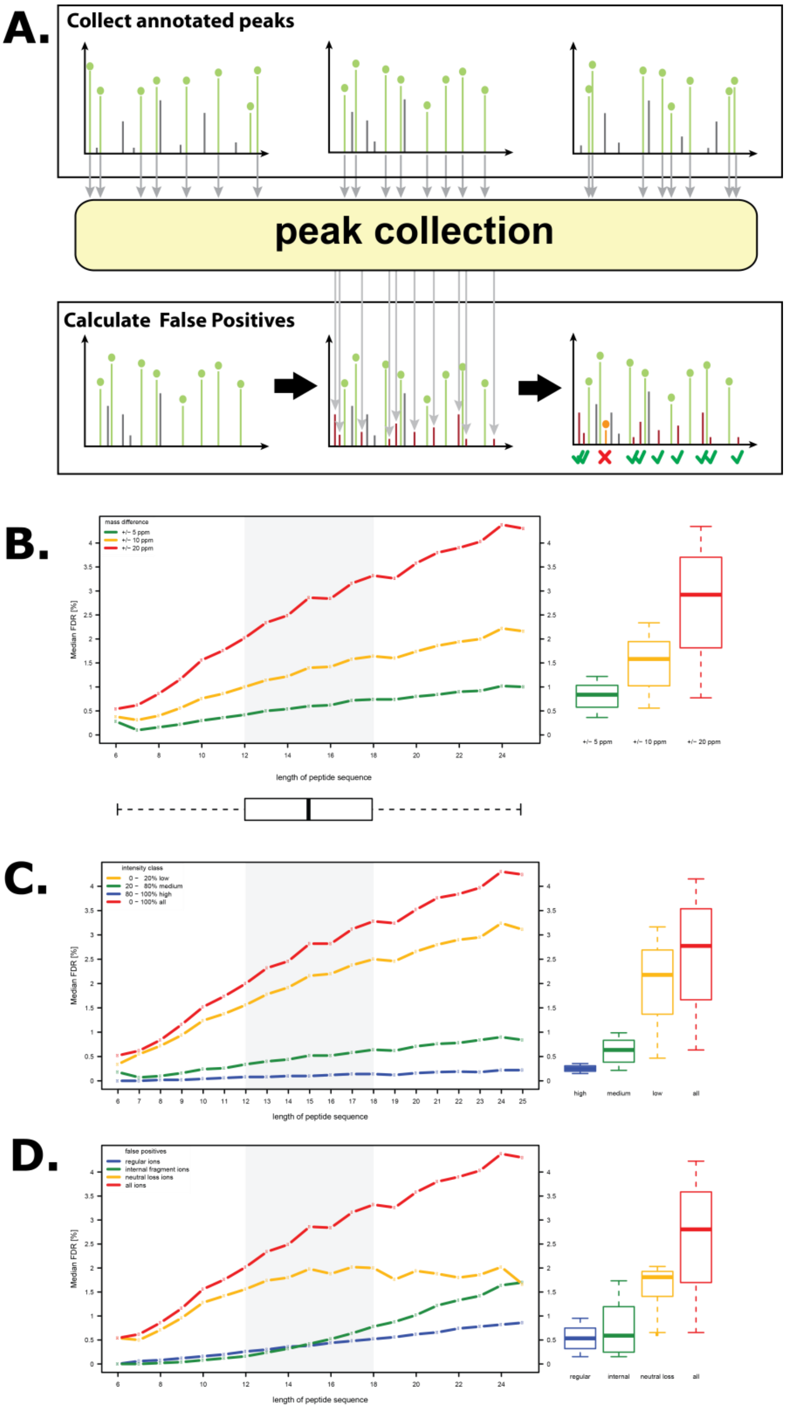


**FIGURE 3 Calculation of false discovery rate for peak annotations.** A. The upper panels represent a large number of identified MS/MS spectra from which annotated peaks are drawn to form a large peak collection of possible fragment masses. From each identified spectrum in the data set, 10 random fragments are inserted and the number of annotations by the Expert System is counted. This process is repeated 500 times for each peptide. B. Median FDR as determined in A as a function of peptide length distinguished by the mass difference of fragment ion and theoretical mass. The FDR for peak annotation rises with peptide length and is strongly dependent on the mass difference. Box plot at the bottom shows that 50 % of the peptides were between 12 and 18 amino acids long. The box plots on the right summarize the range of FDR values regardless of peptide length. C. Graph of the median FDR as a function of peptide length but separated by intensity classes of the false annotated fragment peaks. Most false positives come from the low abundant peaks (blue) rather than the medium (green) or high abundance fragment peaks (yellow). D. Same plot as above but differentiated by the fragment ion type of the false positives. Low number of false positives from regular fragment annotations (blue), compared with internal fragment (green) and neutral loss annotations (yellow).

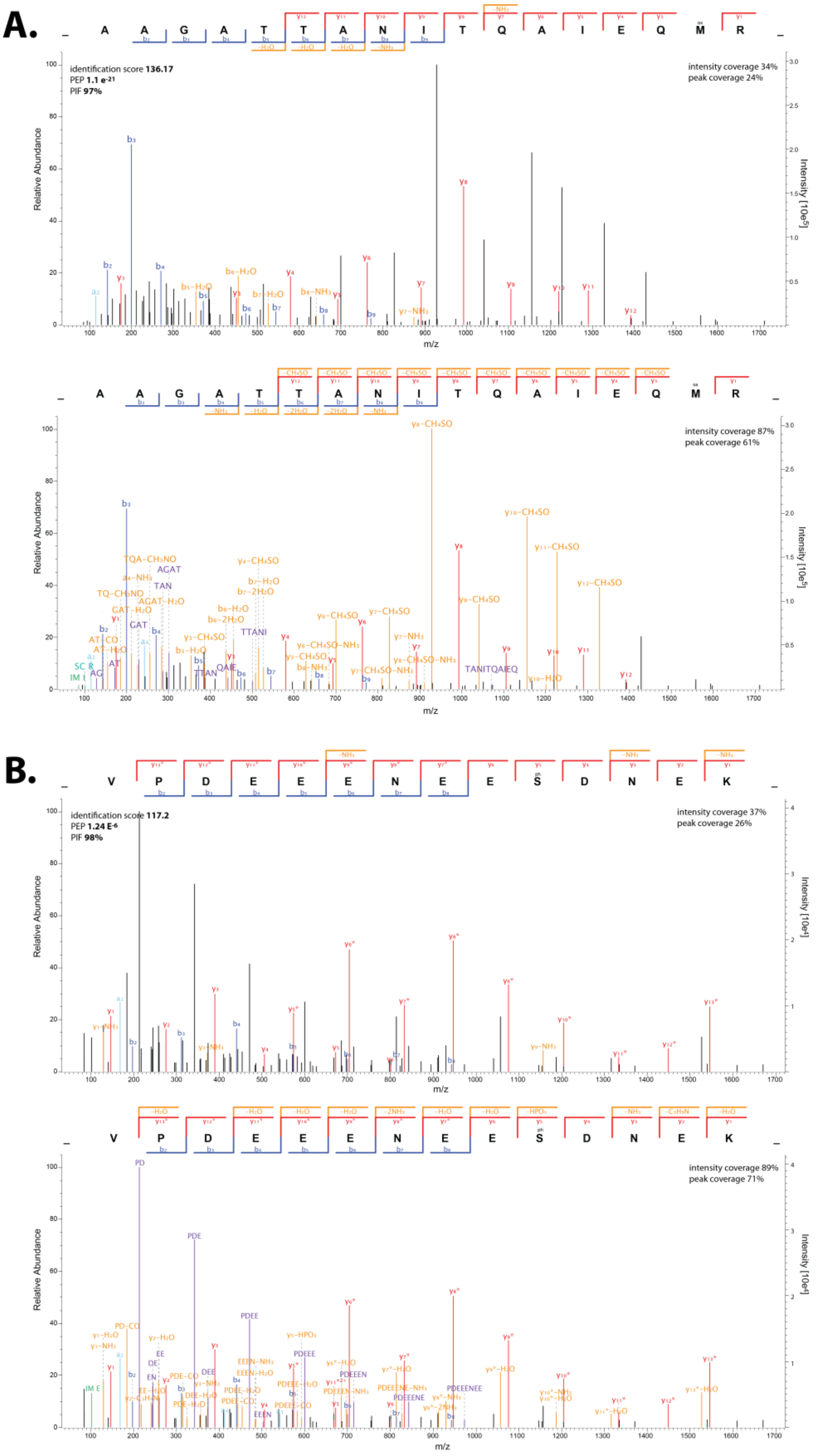


**FIGURE 4 Example spectra before and after Expert System annotation.** A. Based on the search engine result, 34% of the fragments by peak intensities and 24% by peak number are explained, whereas the Expert System almost completely annotates the spectrum (for further explanation see main text). Posterior Error Probability (PEP) a statistical expectation value for the peptide identification in Andromeda. Apart from the huge fraction of a-, b- and y-ions (pale blue/dark blue/red) and ions with neutral losses (orange), you can find internal fragment ions (purple) and in the low mass region one immonium ion of Isoleucine (green) and a side chain loss from Arginine (turquoise). B. Expert System annotation of a phosphorylated peptide. Apart from the internal ions, several phosphorylation-related fragment ions were found. The asterix (*) denotes loss of H3O4P with a delta mass of 97.9768 from the phosphorylated fragment ion.

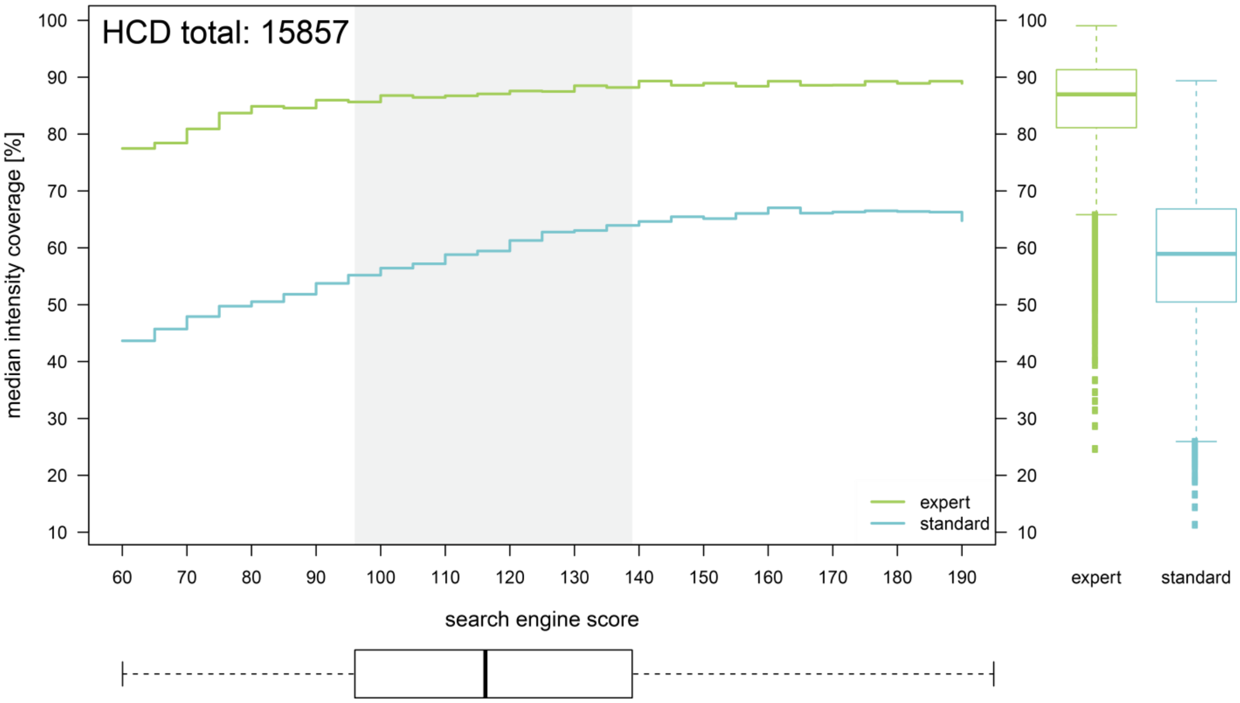
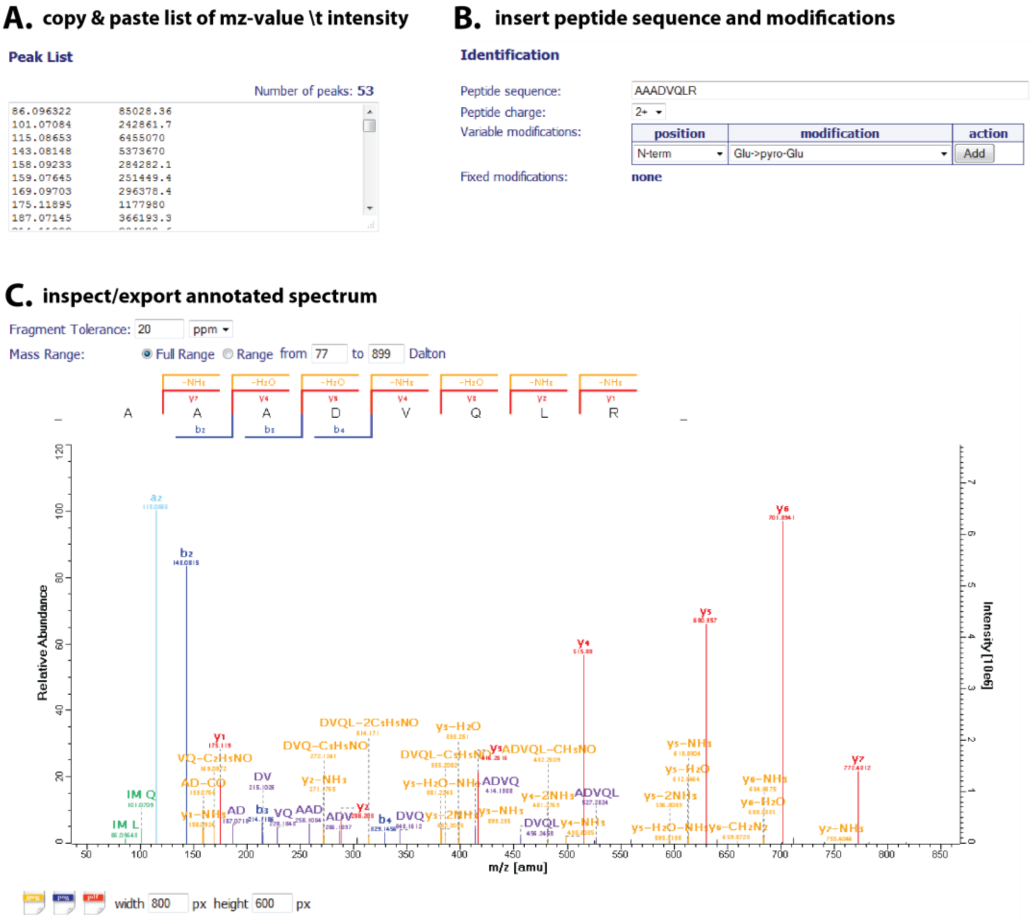


**FIGURE 5** **Expert System** performance on a large data set. A. Median sequence coverage by summed fragment ion intensity is plotted as a function of identification score. Statistics is based on more than 16,000 spectra. For every identification score, the Expert System adds a large proportion of explainable peaks. Box plot below the graph indicates that 50% of peptides in the set have an Andromeda score between 98 and 140. Box plots on the right indicate the range of values for the intensity coverage for standard and Expert System annotation.



**A.** copy & paste list of mz-value \t intensity

Peak List

Number of peaks: 53

```
86.096322       85028.36
101.07084       242861.7
115.08653       6455070
143.08148       5373670
158.09233       284282.1
159.07645       251449.4
169.09703       296378.4
175.11895       1177980
187.07145       366193.3
```

**B.** insert peptide sequence and modifications

Identification

Peptide sequence: AAADVQLR

Peptide charge: 2+

Variable modifications:

| position | modification | action |
|---|---|---|
| N-term | Glu->pyro-Glu | Add |

Fixed modifications: none

**C.** inspect/export annotated spectrum

Fragment Tolerance: 20 ppm

Mass Range: ● Full Range ○ Range from 77 to 899 Dalton

width 800 px height 600 px

**FIGURE 6** **Web interface for the Expert System.** A. Text field to paste the spectrum in text format (m/z value; intensity in arbitrary units). B. Form to enter the peptide sequence, modifications and their positions. C. Detected backbone fragments and their neutral losses are indicated in the peptide logo. Scalable spectrum annotated by the Expert System. Note that neutral loss peaks are very small compared to the major backbone fragments. The spectrum can be downloaded with the desired resolution and in the desired graphical format.

# 6 Mass spectrometry-based Proteomics meets Ribosome Profiling

**Article 8** **Decoding human cytomegalovirus**

Noam Stern-Ginossar, Ben Weisburd, Annette Michalski*, Vu Thuy Khanh Le,
Hartmut Hengel, Matthias Mann, Nicholas T. Ingolia, and Jonathan S. Weissman*

*Science*, 2012, in revision.

(*) corresponding authors

## Prologue

Mass spectrometry-based proteomics aims to characterize the complement of all proteins of a cell. However, the underlying information on expression of different proteins and their abundances are at least partially encoded in messenger RNAs (mRNA) that function as templates. Translation from mRNA to protein occurs at the ribosomes and generally is a well-understood process. Nevertheless, revealing details of the translation process remains challenging. Furthermore, mRNA levels provide only limited insight into the amount and precise nature of the corresponding proteins expressed in a cell. This is primarily due to the fact that not every mRNA is translated into a protein and sophisticated mechanisms control the actual abundance[120,121]. Therefore, systematic investigation of *in vivo* protein translation not only requires comprehensive measurement of the mRNAs present in the cell, but also benefits from *RNA footprinting*, i.e. monitoring the mRNA sequences that are covered by a ribosome for translation at any given time.

Our collaborators, Jonathan Weissman's group at the University of California in San Francisco, recently established a deep sequencing-based technique, referred to as *ribosome profiling*, to investigate ribosome-protected mRNA fragments with very high accuracy and through-put[122]. Briefly, eukaryotic cells are treated with a translation inhibitor such as cycloheximid to 'freeze' the state of translation prior to cell lysis. Unprotected mRNA is removed by nuclease digestion. The ribosome-covered mRNA fragments that consist of about 30 nucleotides are of very suitable size for RNA-Seq analysis, which is then used to obtain the ribosome footprints of the cell (Figure 11).

Advances in deep sequencing fuel this novel method that allows answering fundamental biological questions in large-scale studies. Ribosome profiling experiments are ideally complemented by mass spectrometry-based proteomics, i.e. by investigation of the end product of gene expression, to cross-validate results.

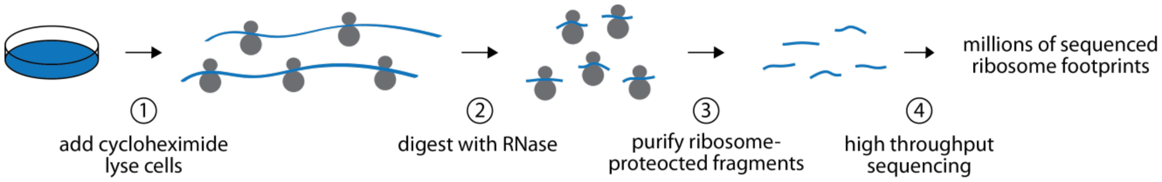

millions of sequenced ribosome footprints
①
add cycloheximide lyse cells
②
digest with RNase
③
purify ribosome-proteocted fragments
④
high throughput sequencing

Figure 11: Schematic workflow of ribosome profiling; adapted from [122].

The project described in this thesis applied ribosome profiling to human fibroblast cells infected with human cytomegalovirus for systematic analysis of the viral transcripts. Intriguingly, our collaborators discovered a large number of novel open reading frames that would not be expected to be protein-coding genes, especially because many encode for very short proteins or feature non-AUG start codons. Even though the human cytomegalovirus genome was sequenced and annotated before, the complexity of the virus genome appears to have been vastly underestimated up to now[123].

During the collaboration, we tested several strategies to detect proteins corresponding to the novel open reading frames predicted by ribosome profiling. We found that the latest generation mass spectrometry instrumentation, Q Exactive, gave a significant boost to the number of novel proteins identified, partially due to its very high sequencing speed. The complete lysate of virus-infected human cells presented a very high complexity background that we reduced by gel fractionation. Additionally, we tested a strategy using size exclusion filters to enrich for smaller proteins after cell lysis with urea, which unfortunately did not result in the expected molecular weight cut-off to reduce sample complexity. Besides tryptic digestion and shotgun proteomics, we also evaluated a top-down strategy. Especially for the very short proteins consisting of only 10-30 amino acids, this approach should have been very advantageous compared to bottom-up shotgun proteomics, as the sequences often lack a tryptic cleavage site or the obtained peptides are too short. However, we found that all of these additional strategies were inferior to the conventional shotgun proteomics workflow. Remarkably, that approach allowed us to confirm more than 50% of the novel open reading frames that did not overlap with any previously known sequence and which were more than 55 amino acids in length. The detection of shorter proteins in complex mixtures remains object of further method development.

In summary, this collaborative project demonstrated the scope of powerful complementary *'omics'* technologies, and provides an example of setting system-wide biological studies into a broader context.

# Decoding human cytomegalovirus

Noam Stern-Ginossar[1], Ben Weisburd[1], Annette Michalski[2]*, Vu Thuy Khanh Le[3], Marco Y. Hein[2], Sheng-Xiong Huang[5], Ming Ma[5], Ben Shen[5,6,7], Shu-Bing Qian[8], Hartmut Hengel[3], Matthias Mann[2], Nicholas T. Ingolia[1,4] and Jonathan S. Weissman[1]*

[1]Department of Cellular and Molecular Pharmacology, Howard Hughes Medical Institute, University of California, San Francisco, CA 94158, USA.

[2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, D-82152, Germany

[3]Institut für Virologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[4]Present address: Department of Embryology, Carnegie Institute for Science, Baltimore, MD 21218, USA.

[5]Department of Chemistry, [6]Department of Molecular Therapeutics, and [7]Natural Products Library Initiative at The Scripps Research Institute, The Scripps Research Institute, 130 Scripps Way #3A2, Jupiter, FL 33458

[8]Division of Nutritional Sciences, Cornell University, Ithaca, NY 14853, USA.

* To whom correspondence should be addressed. A.M. (michalsk@biochem.mpg.de) J.S.W. (weissman@cmp.ucsf.edu)

**Abstract**:

The human cytomegalovirus (HCMV) genome was sequenced 20 years ago. However, like other viruses, our understanding of its protein coding potential is far from complete. Here we use ribosome profiling and transcript analysis to experimentally define the HCMV translation products and follow their temporal expression. We identified several hundred open reading frames not previously suspected to be coding and confirmed a fraction by mass spectrometry. We found that regulated use of alternative transcript start sites plays a role in enabling tight temporal control of HCMV protein expression and allowing multiple distinct polypeptides to be generated from a single genomic locus. Our results reveal an unanticipated complexity to the HCMV proteome and identify regulated changes in transcript start sites as a mechanism for generating this complexity.

**One Sentence Summary:**

Ribosome profiling reveals an unanticipated complexity to HCMV translation.

The herpesvirus, human cytomegalovirus (HCMV), infects the majority of the world's population, leading to severe disease in newborns and immunocompromised adults (*1*). The HCMV genome contains ~240kb  (nearly half the size of the simplest free-living bacterium), making it the largest known human viruses (*2*). Recent estimates of the number of protein coding regions in the HCMV genome have ranged from 165 (*3*) to 252 ORFs (*4*). Yet it is likely that these annotations (which typically rely on conservation and ORF length) do not capture the full complexity of the HCMV proteome (*5*). Indeed, the few genomic regions that have been studied in detail have exposed noncanonical translational events including regulatory (*6*) and overlapping ORFs (*7-11*). Moreover, analysis of cDNA libraries (*12*) and more recently deep sequencing of HCMV RNA(*13*) revealed a highly intricate viral transcriptome. Similar complexity was also observed for Murine cytomegalovirus(*14*) and other herpesviruses (*15, 16*) suggesting that the current genomic maps underestimate the complexity of these viruses

Defining the full set of translation products is critical for identifying the functional proteins encoded by HCMV. Additionally, translational events that do not lead to stable, functional polypeptides could be important as they can regulate translation of a down stream ORF (*6*), contribute to the antigenic potential of the virus (*17, 18*), regulate mRNA stability by inducing nonsense mediated decay (*19*) and even serve as source of proto-proteins enabling the evolution of novel proteins (*20*)

To identify the range of HCMV translated ORFs and monitor their temporal expression, we infected human foreskin fibroblasts (HFF) with the clinical HCMV strain Merlin and harvested the cells at 5, 24 and 72 hours post infection (hpi) using three approaches to generate libraries of ribosome-protected mRNA fragments (Fig. 1a and table S1). The first two measured the overall *in vivo* distribution of ribosomes on a given message; infected cells were either pre-treated with the translation elongation inhibitor cycloheximide or, to exclude drug artifacts, lysed without drug pre-treatment (no drug). In the third approach, cells were first pre-treated with harringtonine, which specifically inhibits translation initiation by preventing 80S ribosomes formed at translation initiation sites from elongating (*21*). This treatment leads to strong accumulation of ribosomes precisely at the sites of translation initiation and depletion of ribosomes over the body of the message, facilitating identification of translation start sites (*22*) (Fig. 1a).

Additionally, we monitored the abundance and ends of mRNA transcripts using a modified RNA-Seq protocol that helps in identification of 5' ends of transcripts ; mRNA is subjected to partial fragmentation followed by size selection of small fragments (50 to 80bp), conversion to DNA and sequencing (Fig. 1b). This results in coverage along the body of messages but also leads to a strong overrepresentation of fragments that start at the 5' end of messages, because every message yields a fragment that contains its 5' end, whereas breaks along the body of the message accrue substoichiometrically (Fig. 1b and (*23*)). We identified the 3' end of messages by searching for the final nucleotide preceding the start of polyadenylation tracks from the sequenced mRNA fragments (*23*).

The power of these approaches to provide a comprehensive view of a gene organization is illustrated for the expression of the UL25 ORF at 72hpi. Here a single transcript start site is found upstream of the ORF (Fig. 1c, mRNA panel). Harringtonine marks a single translation initiation site at the first AUG downstream of the transcript start (Fig. 1c, Harr panel). Ribosome density is seen over the ORF body ending at the first in-frame stop codon (Fig. 1c, CHX and no-drug panels). In the no-drug sample, an excess density of ribosomes is seen at the stop codons, which is a characteristic feature of translation termination (Fig. 1c, no-drug panel) (*22*).

Examination of the full range of HCMV translation products, as reflected by the ribosome footprints, revealed many putative novel ORFs that originate from a variety of sources: short uORFs that lie directly upstream of canonical ORFs (Fig. 2a); ORFs within anti-sense regions of canonical ORFs (Fig. 2b and (*13*)); novel internal ORFs that lie within existing ORFs either in frame, resulting in N-terminally truncated alternative translation products (Fig. 2c), or out of frame, resulting in entirely novel polypeptides (Fig. 2d). Finally, we found novel short ORFs that reside within distinct transcripts (Fig. 2e). For all of these categories, we also observed ORFs that start at near-cognate codons (i.e., codons differing from AUG by one nucleotide), especially CUG (Fig. 2f). Interestingly, while the ribosome density across the body of most of these near-cognate derived ORFs was depleted following harringtonine treatment, a subset was not, suggesting the use of an alternative, harringtonine-resistant, initiation mechanism (Fig. 2f and (*24*))

HCMV encodes for several long RNAs that lack canonical ORFs and thus were thought to be non-coding (*13*), including β2.7, a highly abundant viral message that inhibits apoptosis (*25*). Interestingly, in agreement with previous reports showing that β2.7 transcript is polysome-associated (*26*), our data indicate that multiple short ORFs are translated from this RNA (Fig. 2g and Fig. S1). We detected the presence of the corresponding protein for two of these ORFs by high-resolution mass spectrometric measurements (Fig. 2g). Although the translation efficiency (calculated by dividing footprints by mRNA densities) of these ORFs is low, four of these ORFs show high degree of amino acid conservation in other HCMV strains (Table S2). We characterize three similar polycistronic coding RNAs (including RNA1.2 and RNA4.9) in the HCMV genome and two short proteins encoded by these RNAs were confirmed by mass spectrometry (Fig. S2). The extent to which these transcripts act on the RNA level and/or through their translation products remains an open question.

This striking apparent complexity of translation products prompted us to try and define systematically all the HCMV translated ORFs using the ribosome profiling data. To enable these efforts, we first annotated the HCMV splice junctions using our RNA measurements and two spliced-read mapping tools, TopHat (*27*) and HHMSplicer (*28*). This analysis identified 88 splice junctions, of which all but 6 were recently reported by Gatherer et al.(*13*) (Table S3).

We next exploited the harringtonine-induced accumulation of ribosomes at translation start site to identify ORFs systematically. Automated start site identification was facilitated by a support vector machine (SVM)-based machine learning strategy using a set of canonical ORFs as a training set (*22, 29*). Although all potential translation initiation sites were examined, we observed a strong enrichment for AUG (33- fold) and near cognate codons (Fig. 3a), which provides strong evidence that we captured real translation initiation events. After identifying initiation sites, we generated a list of the HCMV translated ORFs by finding the next in-frame stop codon down-stream of each of the initiating codons taking into account any intervening splice junctions. Visual inspection confirmed the SVM-identified ORFs and identified an additional 30 previously annotated genes and 33 novel translation products that were not identified by our stringent automated approach (Table S4). The large majority of these ORFs, including all of the manually identified ones, where identified by SVM analysis of an independent biological replicate (Table S5). To exclude the possibility that the predicted initiation sites resulted from an artifact caused by harringtonine treatment, we have employed a second drug lactimidomycin (LTM) (*30*), which acts by a distinct mechanism to cause ribosome to accumulate at the sites of translation initiation (e.g. harringtonine binds to the 60S prior to subunit joining (*21, 31*) whereas LTM binds to the E site in the 80S(*30*)). LTM treatment resulted in the expected accumulation of ribosomes at translation initiation sites (fig. 2 and supplementary file 1). Analysis of the LTM data confirmed the vast majority (>98%) of the translation start sites identified by harringtonine (Fig. 3b and (*23*)).

In total we identified 751 translated ORFs that were supported by both the LTM and harringtonine treatments (Table S5, Table S6 and supplementary file 1). The footprint density measurements for these ORFs were highly reproducible both globally ($R^2$ = 0.981, between independent biological replicates, Fig S3) and for individual ORFs (Fig. S4). 147 of these ORFs correspond to ORFs that were previously suggested to be coding (Fig. 3c). We did not find strong evidence of translation for 25 previously annotated ORFs (Table S7) although these proteins may well be expressed under different conditions or by different HCMV strains.

Many of the newly identified ORFs are very short (245 ORFs < 20 codons, Fig. 3d) and are found upstream of longer ORFs. It is possible that some of these play roles in regulating translation of the downstream ORF (*32*) (see below). We also identified many (249) short ORFs (21-80 codons, Fig. 3c) some of which could have a non-regulatory function. Indeed, the idea that a small ORF may encode for a peptide in herpesvirus genomes has been raised previously (*33, 34*). Lastly, we identified 129 novel ORFs that are longer than 80 amino acids. These are primarily internal ORFs, ORFs that contain splice junctions and adjustments or alternative 5’ ends of previous annotations.

Several lines of evidence support the validity of our ORF identification approach. First, we exploited the observed excess of ribosomes on termination codons in the no-drug sample (Fig. 1c and (*22*)) to evaluate our predictions. We found that, as seen for the previously annotated ORFs, the newly identified ORFs show a significant ($P< 10^{-70}$, *Ks*) test) excess of ribosome footprints at the predicted stop codon (Fig. S5). Because our ORF predictions were based on translation initiation sites found in the harringtonine and LTM samples, the fact that these accurately predicted downstream stop codons in an independent untreated sample provides strong validation for our annotations. Second, the ribosome-protected footprints from the cycloheximide-treated sample displayed a 3-nt periodicity (reflecting the movement of ribosomes along mRNAs) that was in phase with the predicted start site- a hallmark of translation. This periodicity was globally similar for footprints within the newly predicted and previously annotated ORFs (Fig. 3e). By analyzing the 3-nt periodicity in specific ORFs that contain an internal out of frame ORF, we were able to detect the localized shift in translation frame (Fig. S6). Third, we found that ribosome density was depleted from the body of the large majority of the predicted ORFs following short inhibition of translation initiation using an eIF4A (a component of the eIF4F complex needed for engagement of the 40s ribosome subunit) inhibitor PateamineA (PatA) (*35*)(Fig. S7). The observation that ribosomes run-off after short inhibition of translation indicates that they were engaged in active elongation rather than being stalled. The newly identified ORFs also exhibit a distribution of expression and translation efficiencies levels that is similar to that of the previously annotated canonical ORFs (Fig. S8). Finally many of the newly identified ORFs show a high degree of conservation among other HCMV strains (Table S2). These properties of the newly identified ORFs underscore their potential to be an important part of the HCMV proteome but future work will have to delineate which of the HCMV translated ORFs are functional during productive infection.

We also conducted several lines of experiments to provide support directly at the protein level for our annotations. First, we performed high-resolution tandem mass spectrometric measurements on virally infected cells. We applied stringent automated criteria (e.g., 99% statistical confidence) (*23, 36*) and manually verified the identity each of the spectra (supplementary file 2). Using these criteria, out of the 96 genomic loci that contain at least one unique novel protein that is longer than 55 amino acids and does not overlap with any of the previously annotated ORFs, we unambiguously detected 53 novel proteins (table S8 and Supplementary file 3). This likely largely underestimates the true fraction of novel ORFs products that are present during infection as the detection of short viral proteins in the background of a total cellular proteome is very challenging.

For classes of new ORFs that were difficult to monitor by mass spectrometry (i.e., truncated forms of longer proteins or short proteins (*23*)), we used a tagging approach. For two N-terminally truncated proteins (derived from UL16 and UL38), we confirmed the appearance of alternative shorter transcripts and detected the expected full length and truncated tagged protein products (Fig. S9). The N-terminally truncated protein that derives from UL16 was confirmed by analysis of genetically modified HCMV in which the UL16 loci was tagged with an HA epitope (Fig. S10) and we confirmed a splice variant of UL138 using an antibody (Fig. S10). For five short ORFs (ranging from 25-80 amino acids, including two initiated at near cognate start sites), we fused the ORFs in frame to a GFP coding region (in their otherwise native transcript context) and identified protein products of the expected sizes. Frame shift experiments confirmed that we correctly identified the sites of translation initiation (Fig. S11). We also show that one of these short proteins (US33A- 57aa), which we were not able to identify by mass spectrometry but was also recently predicted to be coding by transcript analysis (*13*), is expressed in the context of the native virus (Fig 3f and Fig. S10). Additionally, we focused on the very short (as few as one codon) near cognate driven uORFs that lie directly upstream of UL119 and US9, whose inclusion changes during infection as a result of changes in the 5’ end of the transcripts. We found that these uORFs impact translation of a downstream reporter gene. Mutation of the near cognate translation start site codons eliminated the observed translation effect confirming that we correctly identified the translation initiation sites (Fig. S12).

Finally, we examined the subcellular localization of 18 of the newly identified ORFs (size range 40-160 AAs, 11 of which were detected by mass-spectrometry, table S9) using transient expression of GFP-tagged proteins. Notably, 10 out of the 15 proteins we detected showed highly specific subcellular localization patterns: six in mitochondria, three in the endoplasmic reticulum and one in the nucleus (Fig. 3g and Fig. S13).

We performed immunoprecipation and mass spectrometry experiments on two of these GFP-tagged proteins; ORF359W (ER localized) and US33A (mitochondrially localized) both of which are predicted to be transmembrane and are highly conserved (Table S2). This analysis identified a few specific interacting proteins that we confirmed by western blot analysis including TAP1 (ORF359W) and the mitochondrial inner membrane transport TIM machinery (US33A) (Fig. S14).

It is possible that a significant subset of the short ORFs we have identified (especially those expressed only late during infection) are rapidly degraded and do not act as functional polypeptides. Nonetheless, these could still be an important part of the immunological repertoire of the virus as MHC class I bound peptides are generated at higher efficiency from rapidly degraded polypeptides (17).

HCMV genes are expressed in a temporally regulated cascade traditionally divided into three major waves of viral gene expression, immediate early (IE), early and late. IE genes are defined as those that are transcribed following infection in the absence of de-novo protein synthesis, and “early” and “late” genes are divided based on the insensitivity (early) versus sensitivity (late) to viral DNA replication inhibitors. Although microarray-based approaches have previously been used to characterize HCMV gene expression during infection (*37, 38*), those studies were limited by the incomplete annotation of HCMV transcripts and ORFs. Our data thus provides an opportunity to monitor viral protein translation throughout infection. Notably, most of the viral genes (canonical and newly identified putative ORFs) showed tight temporal regulation of protein synthesis levels; 82% of ORFs varied by at least 5-fold including 66% of ORFs which showed >10-fold changes. Hierarchical clustering of viral coding regions by their footprints densities during infection (a measure of the relative amount of protein being produced) revealed several dominant clusters (Fig. S15). One cluster contains genes with strong translation at 5 hpi (Fig. S15, cluster 1). This includes a limited number of IE genes (UL123, US3, UL36 and UL37) that have been identified and 25 additional ORFs (Table S10). Although our data is limited to three infection time points, the division of viral genes into three categories does not capture the full range of expression signatures (Fig S15).

How is such tight temporal regulation of the range of different ORFs, including many encoded by overlapping genomic regions, achieved? Examination of the viral transcripts during infection revealed the pervasive use of alternative 5’ ends that appeared to be critical to the tight temporal regulation of viral genes during infection.

A notable example is the US18-US20 locus. At 5 hpi, there is one main transcript that starts just upstream of US20 enabling US20 translation. At 24 hpi, a shorter version of the transcript is detected starting immediately upstream of US18 enabling its translation. A third novel version of the transcript that starts within the US18 coding sequence emerges at 72 hpi. This transcript results in the translation of a truncated version of US18 (ORFS346C.1), which is unique to this time point (Fig. 4a and b).

Another detailed example is illustrated in Fig. S16 and we have qualitatively identified similar phenomena in 61 viral loci that showed clear temporal regulation of their 5’ ends that was reproducible between biological replicates and affected their coding potential during infection (Fig S17, Fig. S18 and table S11). Time dependent changes in transcripts derived from six of these genomic regions were confirmed by Northern blot analysis (Fig. 4b, FigS9 and Fig. S19). Thus our studies reveal a pervasive mode of viral gene regulation in which dynamic changes in 5’ ends of transcripts control which proteins are expressed from overlapping coding regions. Just as alternative splicing (a process in which a single gene codes for multiple proteins) has an established role in expanding protein diversity, alternative transcript start sites may provide a broadly used mechanism, for generating complex proteomes from a limited number of genes.

The genomic era began with the sequencing of the bacterial DNA virus, phi X, in 1977 (*39*) and the mammalian DNA virus, Simian virus 40 (*40, 41*), the following year. Since then, extraordinary advances in sequencing technology have enabled the determination of a vast array of viral genomes. However, due to the high-density nature of these genomes, deciphering their protein coding potential remains a great challenge.  Here we present the first experimentally-based analysis of the expressed proteome of a complex DNA virus, HCMV, using both next generation sequencing and high-resolution proteomics. Our work provides a framework for studying HCMV by establishing the viral proteome and its temporal regulation, providing a context for mutational studies and revealing the full range of HCMV antigenic potential. More broadly, our work establishes a model for mapping and deciphering complex genomes.

## Figure legends:

**Fig. 1**. Ribosome profiling of HCMV infected cells during the course of infection.

**(a)** Schematic representation of the experimental set-up. HFF cells were infected with HCMV and harvested at different times after infection for ribosome footprint analysis using cycloheximide or no-drug to map translation and harringtonine to map translation initiation. Samples were also collected for mRNA measurements using random fragmentation.

**(b)** Schematic representation of RNA-Seq protocol that leads to mRNA 5’ end enrichment. Isolated mRNA was randomly fragmented and then size selected to include only RNA fragments smaller than 80bp (marked with star). This procedure generates enrichment in fragments that start at the 5’ end of transcripts which is represented by plotting the 5’end of each sequencing read.

**(c)** Example of a viral ORF. The ribosome occupancies of the various treatments; cycloheximide (CHX), no-drug and harringtonine (Harr) together with mRNA profiles of the UL25 gene at 72 hpi is shown. The mRNA profile shows only the 5’ end of each sequencing read and an arrow marks the mRNA start.

**Fig. 2**. Many ribosome footprints do not correspond to previously annotated viral ORFs

**(a)** Example of short uORFs. The ribosome occupancy profiles are shown for the leader region of UL139 gene.

**(b)** Example of antisense ORF. The ribosome occupancy profiles of plus and minus strands (red and blue respectively) are shown for the region of UL150 gene. Notice that the ORF is being spliced (marked by a dashed line).

(**c, d)** Examples of internal ORFs. The ribosome occupancy profiles are shown for the UL38 (c) and UL10 (d) genes. In both examples there are internal initiations that generate internal ORFs, either in frame (truncated protein) (c) or out of frame (d). The grey area symbolizes a low complexity region to which less reads are aligned uniquely.

**(e)** Example of a novel short ORF. The mRNA and ribosome occupancy profiles are shown for a novel short ORF.

**(f)** Example of a short ORF that initiates at a CUG codon. The ribosome occupancies around the genomic locus of a novel CUG starting ORF are shown.

**(g)** Example of translation from long RNA that lack canonical ORFs. The ribosome occupancy profiles are shown around lncRNA2.7. Mass spectrometric analysis and the ribosome occupancies show that 10 short ORFs are translated. The upper panels show the annotated MS/MS spectra of the unique peptides KLQTFGYISFPITR and VYVCIVSPPSR originating from the 86aa and 84aa proteins. Median absolute fragment mass deviation is 1 ppm.

**Fig. 3.** Annotating HCMV translated ORFs.

(a) Fold enrichment of AUG and near-cognate codons at predicted sites of translation initiation compared to the overall codon distribution in the genome.

(b) The distribution of ribosome occupancy at start codons after LTM treatment. The ribosome footprints density at each start codon (relative to the median density across the gene) is depicted for the previously annotated ORFs (blue) and for the newly identified ORFs (red and empty red for ORFs that were removed). As a control, the occupancy of a codon 5 positions down stream of the start codon was calculated in the same way (green).

(c) Venn diagram comparing the HCMV translated ORFs identified in the present study with previously annotated ORFs.

(d) The length distribution of HCMV ORFs. The lengths of newly identified putative ORFs (red) are compared to the lengths of previously annotated ORFs (blue).

(e) Frame bias seen in the position of the 30-nt ribosome footprints relative to the reading frame in the newly identified ORFs (red) compared to previously annotated ORFs (blue).

(f) MRC-5 cells were mock-treated or infected with TB40-US33A-HA. Protein lysates were generated after indicated times of infection and analyzed by western blotting using antibodies recognizing HA, pp65 and beta-actin. (g) Subcellular localization of two HCMV GFP-tagged short ORFs. HeLa cells were transfected with GFP fusion proteins together with an ER marker (KDEL-mCherry) or stained with MitoTracker Red and imaged by confocal microscopy.

**Fig. 4**. A major source of ORFs diversity during infection originates from alternative transcripts starts.

**(a)** An example of changes in transcripts 5’ ends during infection that lead to protein diversity. The mRNA and ribosome occupancy profiles are shown around US18-US20 loci at different time points during infection (marked on the left). Small arrows denote the different mRNA starts and the corresponding mRNAs are illustrated (upper part). Note that changes in the transcript 5’ ends expose the downstream ORF (US18) at 24 hpi and an internal ORF (ORFS346C.1) at 72 hpi. The lower panel shows an expanded view of the US18 locus at 72 hpi and includes the harringtonine and LTM data that illustrate the internal initiation (marked with a star).

**(b)** Total RNA extracted at different time points along infection was subjected to Northern blotting for ORFS346C.1

## References and Notes:

1. E. S. Mocarski, T. Shenk, R. F. Pass, in *Fields Virology,* B. N. Fields, D. M. Knipe, P. M. Howley, Eds. (Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, 2007).
2. M. S. Chee *et al.*, Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Current topics in microbiology and immunology* **154**, 125 (1990).
3. A. J. Davison *et al.*, The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *The Journal of general virology* **84**, 17 (Jan, 2003).
4. E. Murphy, I. Rigoutsos, T. Shibuya, T. E. Shenk, Reevaluation of human cytomegalovirus coding potential. *Proc Natl Acad Sci U S A* **100**, 13585 (Nov 11, 2003).
5. E. Murphy, T. Shenk, Human cytomegalovirus genome. *Current topics in microbiology and immunology* **325**, 1 (2008).
6. J. Cao, A. P. Geballe, Inhibition of nascent-peptide release at translation termination. *Mol Cell Biol* **16**, 7109 (Dec, 1996).
7. T. Stamminger *et al.*, Open reading fram UL26 of human cytomegalovirus encodes a novel tegument protein that contains a strong transcriptional activation domain. *Journal of virology* **76**, 4836 (May, 2002).
8. B. J. Biegalke, E. Lester, A. Branda, R. Rana, Characterization of the human cytomegalovirus UL34 gene. *Journal of virology* **78**, 9579 (Sep, 2004).
9. H. Isomura *et al.*, Noncanonical TATA sequence in the UL44 late promoter of human cytomegalovirus is required for the accumulation of late viral transcripts. *Journal of virology* **82**, 1638 (Feb, 2008).
10. Z. Qian, B. Xuan, T. T. Hong, D. Yu, The full-length protein encoded by human cytomegalovirus gene UL117 is required for the proper maturation of viral replication compartments. *Journal of virology* **82**, 3452 (Apr, 2008).
11. L. Grainger *et al.*, Stress-inducible alternative translation initiation of human cytomegalovirus latency protein pUL138. *Journal of virology* **84**, 9472 (Sep, 2010).
12. G. J. Zhang *et al.*, Antisense transcription in the human cytomegalovirus transcriptome. *Journal of virology* **81**, 11267 (Oct, 2007).
13. D. Gatherer *et al.*, High-resolution human cytomegalovirus transcriptome. *Proc Natl Acad Sci U S A* **108**, 19755 (Dec 6, 2011).
14. P. Lacaze *et al.*, in *Journal of virology*. (2011), vol. 85, pp. 6065-76.
15. B. Y. H. Cheng *et al.*, in *Journal of virology*. (2012), vol. 86, pp. 4340-57.

16. L. R. Dresang *et al.*, in *BMC Genomics*. (2011), vol. 12, pp. 625.

17. J. W. Yewdell, Plumbing the sources of endogenous MHC class I peptide ligands. *Current opinion in immunology* **19**, 79 (Feb, 2007).

18. S. R. Schwab, K. C. Li, C. Kang, N. Shastri, in *Science*. (2003), vol. 301, pp. 1367-71.

19. T. Hirose, M.-D. Shu, J. A. Steitz, in *Proc Natl Acad Sci USA*. (2004), vol. 101, pp. 17976-81.

20. A. R. Carvunis *et al.*, Proto-genes and de novo gene birth. *Nature*, (Jun 24, 2012).

21. F. Robert *et al.*, Altering chemosensitivity by modulating translation elongation. *PloS one* **4**, e5428 (2009).

22. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789 (Nov 11, 2011).

23. Materials and methods are available as supporting material.

24. S. R. Starck *et al.*, Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. *Science* **336**, 1719 (Jun 29, 2012).

25. M. B. Reeves, A. A. Davies, B. P. McSharry, G. W. Wilkinson, J. H. Sinclair, Complex I binding by a virally encoded RNA regulates mitochondria-induced cell death. *Science* **316**, 1345 (Jun 1, 2007).

26. P. C. Lord, C. B. Rothschild, R. T. DeRose, B. A. Kilpatrick, Human cytomegalovirus RNAs immunoprecipitated by multiple systemic lupus erythematosus antisera. *The Journal of general virology* **70 ( Pt 9)**, 2383 (Sep, 1989).

27. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105 (May 1, 2009).

28. M. T. Dimon, K. Sorber, J. L. DeRisi, HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PloS one* **5**, e13875 (2010).

29. T. Joachim, *Making large scale SVM learning practical*. Advances in Kernel Methods - Support Vector Learning (Cambridge, MIT Press, 1998).

30. T. Schneider-Poetsch *et al.*, Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nature chemical biology* **6**, 209 (Mar, 2010).

31. M. Fresno, A. Jimenez, D. Vazquez, Inhibition of translation in eukaryotic systems by harringtonine. *European journal of biochemistry / FEBS* **72**, 323 (Jan, 1977).

32. K. A. Spriggs, M. Bushell, A. E. Willis, Translational regulation of gene expression during conditions of cell stress. *Mol Cell* **40**, 228 (Oct 22, 2010).

33. Y. Xu, D. Ganem, Making sense of antisense: seemingly noncoding RNAs antisense to the master regulator of Kaposi's sarcoma-associated herpesvirus lytic replication do not regulate that transcript but serve as mRNAs encoding small peptides. *Journal of virology* **84**, 5465 (Jun, 2010).

34. S. M. Varnum *et al.*, Identification of proteins in human cytomegalovirus (HCMV) particles: the HCMV proteome. *Journal of virology* **78**, 10960 (Oct, 2004).

35. M. E. Bordeleau *et al.*, RNA-mediated sequestration of the RNA helicase eIF4A by Pateamine A inhibits translation initiation. *Chemistry & biology* **13**, 1287 (Dec, 2006).

36. J. Cox *et al.*, Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794 (Apr 1, 2011).

37. J. Chambers *et al.*, DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *Journal of virology* **73**, 5757 (Jul, 1999).

38. A. K. Cheung, A. Abendroth, A. L. Cunningham, B. Slobedman, Viral gene expression during the establishment of human cytomegalovirus latent infection in myeloid progenitor cells. *Blood* **108**, 3691 (Dec 1, 2006).

39. F. Sanger, S. Nicklen, A. R. Coulson, DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463 (Dec, 1977).

40. V. B. Reddy *et al.*, The genome of simian virus 40. *Science* **200**, 494 (May 5, 1978).

41. W. Fiers *et al.*, Complete nucleotide sequence of SV40 DNA. *Nature* **273**, 113 (May 11, 1978).

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

43. A. Dolan *et al.*, Genetic content of wild-type human cytomegalovirus. *The Journal of general virology* **85**, 1301 (May, 2004).

44. R. J. Stanton *et al.*, Reconstruction of the complete human cytomegalovirus genome in a BAC reveals RL13 to be a potent inhibitor of replication. *The Journal of clinical investigation* **120**, 3191 (Sep, 2010).

45. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

46. M. J. de Hoon, S. Imoto, J. Nolan, S. Miyano, Open source clustering software. *Bioinformatics* **20**, 1453 (Jun 12, 2004).

47. A. J. Saldanha, Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**, 3246 (Nov 22, 2004).

48. R. Atalay *et al.*, Identification and expression of human cytomegalovirus transcription units coding for two distinct Fcgamma receptor homologs. *Journal of virology* **76**, 8596 (Sep, 2002).

49. C. Sinzger *et al.*, Cloning and sequencing of a highly productive, endotheliotropic virus strain derived from human cytomegalovirus TB40/E. *The Journal of general virology* **89**, 359 (Feb, 2008).

50. M. Wagner, A. Gutermann, J. Podlech, M. J. Reddehase, U. H. Koszinowski, Major histocompatibility complex class I allele-specific cooperative and competitive interactions between immune evasion proteins of cytomegalovirus. *The Journal of experimental medicine* **196**, 805 (Sep 16, 2002).

51. V. T. Le, M. Trilling, H. Hengel, The cytomegaloviral protein pUL138 acts as potentiator of tumor necrosis factor (TNF) receptor 1 surface density to enhance ULb'-encoded modulation of TNF-alpha signaling. *Journal of virology* **85**, 13260 (Dec, 2011).

52. N. C. Hubner *et al.*, Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *The Journal of cell biology* **189**, 739 (May 17, 2010).

53. J. R. Wisniewski, A. Zougman, N. Nagaraj, M. Mann, Universal sample preparation method for proteome analysis. *Nature methods* **6**, 359 (May, 2009).

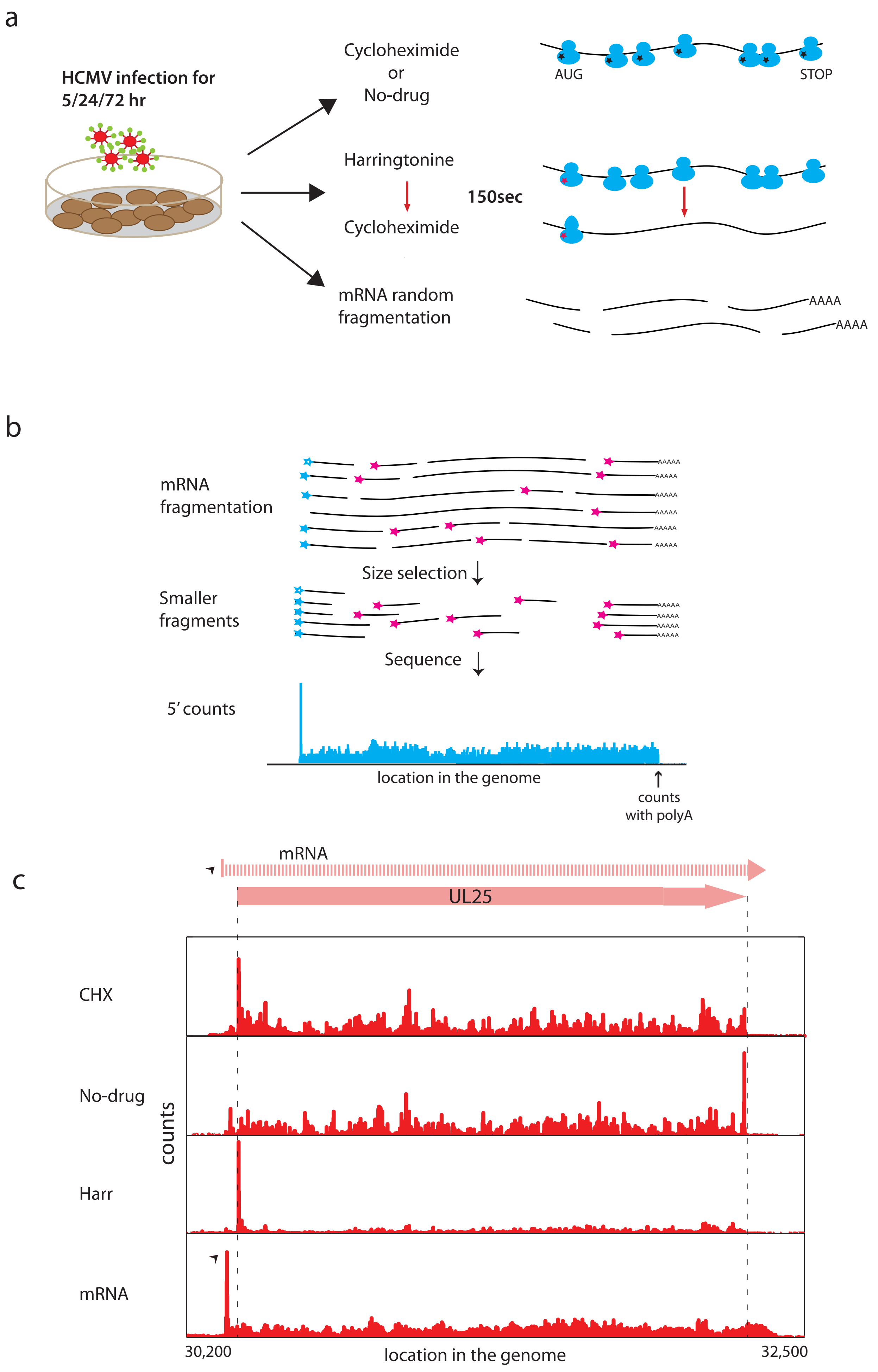
54. S. Schandorff *et al.*, A mass spectrometry-friendly database for cSNP identification. *Nature methods* **4**, 465 (Jun, 2007).

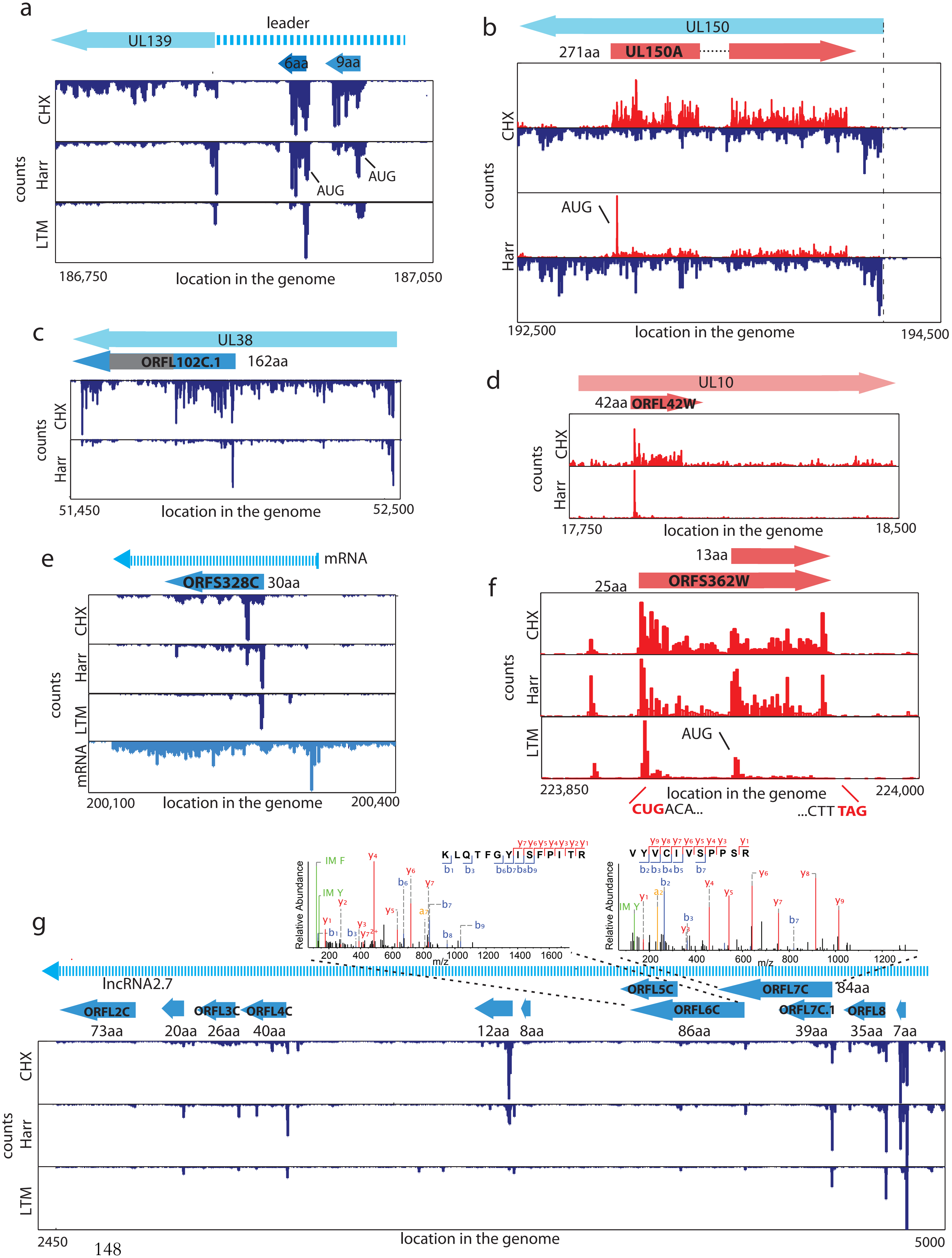
55. J. Rappsilber, Y. Ishihama, M. Mann, Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**, 663 (Feb 1, 2003).

56. O. R. Homann, A. D. Johnson, MochiView: versatile software for genome browsing and DNA motif analysis. *BMC Biol* **8**, 49 (2010).

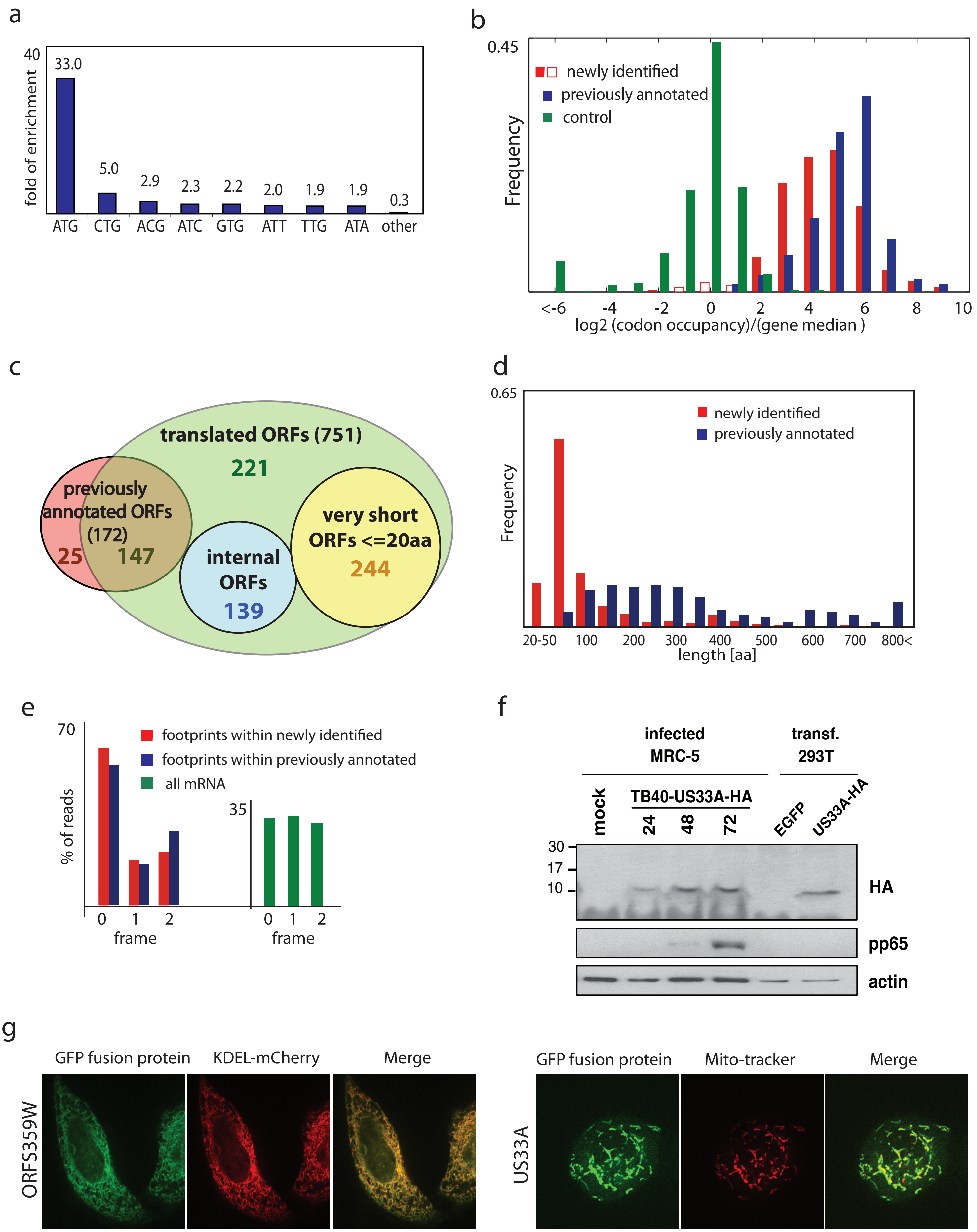
## Acknowledgements:

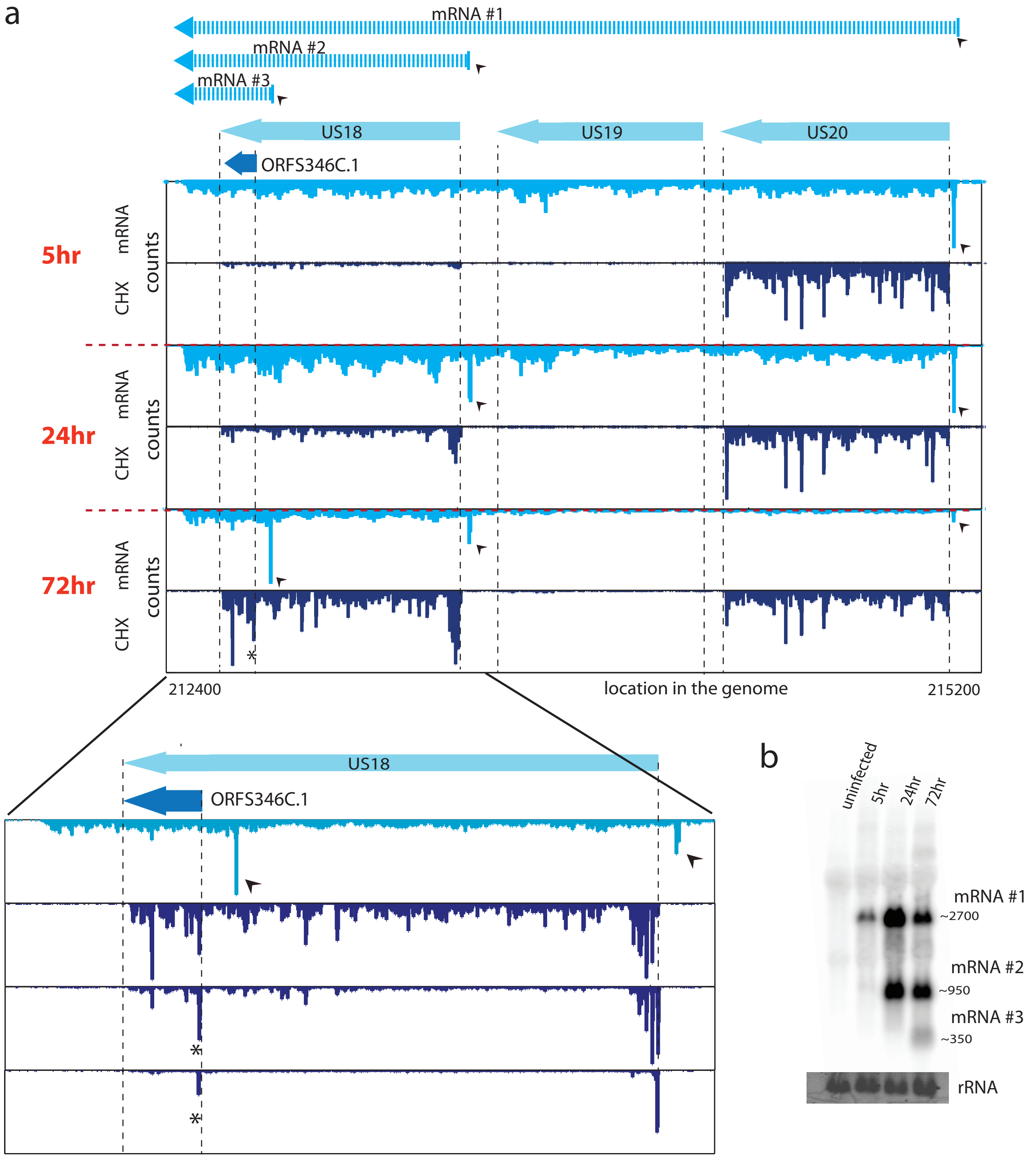
We thank O. Mandelboim, D. Wolf, C. Arias, M. Trilling, A. Lauring, S. Karniely and members of the Weissman lab for critical reading of the manuscript; C. Chu for assistance with sequencing; Pelletier, J. for providing PateamineA; N. S-G is supported by human frontier science program postdoctoral fellowship. N.T.I is supported by a Ruth L. Kirschstein NRSA (GM080853). This work was supported by HMMI funding (JSW) and the Max-Planck Society for the Advancement of Science (MM).


a
HCMV infection for
5/24/72 hr
Cycloheximide
or
No-drug
AUG
STOP
Harringtonine
150sec
Cycloheximide
mRNA random
fragmentation
AAAA
AAAA
b
mRNA
fragmentation
AAAAA
Size selection
Smaller
fragments
Sequence
5′ counts
location in the genome
counts
with polyA
c
mRNA
UL25
CHX
No-drug
Harr
mRNA
counts
30,200
location in the genome
32,500


a
UL139
leader
6aa
9aa
CHX
Harr
LTM
counts
AUG
AUG
186,750
location in the genome
187,050
b
UL150
271aa
UL150A
CHX
Harr
counts
AUG
192,500
location in the genome
194,500
c
UL38
ORFL102C.1
162aa
CHX
Harr
counts
51,450
location in the genome
52,500
d
UL10
42aa
ORFL42W
CHX
Harr
counts
17,750
location in the genome
18,500
e
mRNA
ORFS328C
30aa
CHX
Harr
LTM
mRNA
counts
200,100
location in the genome
200,400
f
13aa
25aa
ORFS362W
CHX
Harr
LTM
counts
AUG
223,850
location in the genome
224,000
CUGACA...
...CTT TAG
g
K L Q T F G Y I S F P I T R
V Y V C I V S P P S R
Relative Abundance
m/z
IM F
IM Y
200 400 600 800 1000 1200 1400 1600
200 400 600 800 1000 1200
lncRNA2.7
ORFL5C
ORFL7C
84aa
ORFL6C
86aa
ORFL2C
73aa
20aa
ORFL3C
26aa
ORFL4C
40aa
12aa
8aa
ORFL7C.1
39aa
ORFL8
35aa
7aa
CHX
Harr
LTM
counts
2450
location in the genome
5000

*Stern-Ginossar et al. Figure 3*




a
mRNA #1
mRNA #2
mRNA #3
US18
US19
US20
ORFS346C.1
5hr
24hr
72hr
mRNA
CHX
counts
*
212400
location in the genome
215200
US18
ORFS346C.1
mRNA
CHX
Harr
LTM
counts
*
*
b
uninfected
5hr
24hr
72hr
mRNA #1
~2700
mRNA #2
~950
mRNA #3
~350
rRNA

# 7 Summary & Concluding Remarks

In this thesis, the fundamental principles of shotgun proteomics were investigated from a global analytical perspective. The discipline of MS-based proteomics deals with various challenges at different levels, starting from optimal sample preparation methods and prominently including mass spectrometry technology with its associated data analysis tools. It has become clear in our and other laboratories that a strong integration and good balance between these topics is of pivotal importance for overall success. With the ultimate goal of high data quality for comprehensive proteome analysis in mind, we focused on (1) the instrumental capabilities including improvements to sensitivity, sequencing speed and resolution and (2) development of data analysis tools that translate these technological advances into the most comprehensive and confident peptide and protein identifications possible. The cumulative benefit of these contributions was demonstrated in the collaborative project *Decoding human cyclomegalovirus* in which we applied the latest technologies of mass spectrometry-based proteomics to confirm the existence of a large proportion of novel proteins predicted by ribosome profiling.

Figure 12 illustrates three levels that need to be carefully linked for a successful shotgun proteomics experiment, and it illustrates the areas in which the efforts of this thesis have made a contribution.

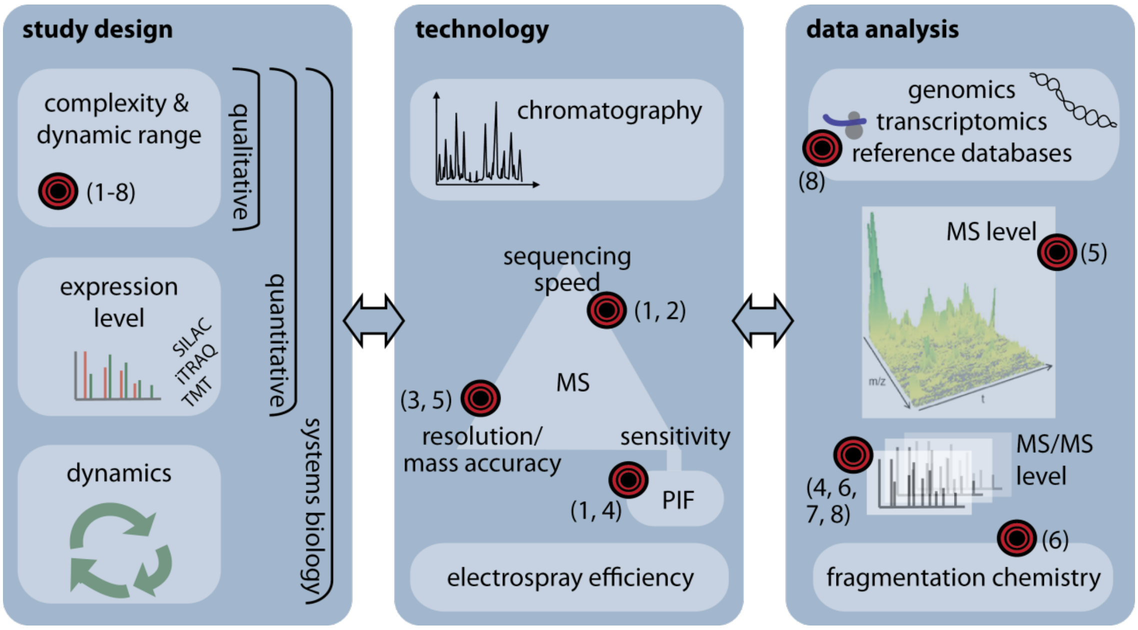


Figure 12: Different levels of shotgun proteomics research. Focus and contribution of the projects of this thesis are marked in red; numbers (1-8) indicate the corresponding articles.

Initially we performed a baseline study that outlines the requirements needed in mass spectrometric instrumentation to efficiently analyze highly complex shotgun proteomics samples. This effort provided detailed insight into the influence of the three major instrument parameters sequencing speed, sensitivity and purity of precursor isolation, and allowed judging the potential of the most commonly applied data-dependent acquisition strategy with regard to the accessibility of the *peptidome* resulting from digestion of all proteins in the sample. Due to the high complexity of shotgun proteomics samples, mere improvement of sequencing speed and sensitivity is not sufficient although highly beneficial. The purity of precursor isolation remains an inherent limitation of the data dependent peptide sequencing approach. However, the vast majority of proteomics studies do not always require complete sequence coverage of all proteins and the desired aim of complete proteome analysis as a complementing technology to next-generation sequencing is not necessarily affected.

The latest generation Orbitrap instrumentation was optimized and evaluated in terms of hardware and software as a major effort of this thesis. Especially the high sequencing speed of the quadrupole Orbitrap combination (Q Exactive) proved to be of tremendous importance for increased proteome coverage; the duty cycle of the quadrupole Orbitrap instrument is increased by a factor 2-3 over its predecessor (Article 2). This substantial advancement does not come at the expense of lower quality mass spectra as it appears to do in quadrupole TOF type instruments, but is largely facilitated by parallel ion accumulation and scanning capability of the instrument and by an enhanced Fourier transformation algorithm, which effectively doubles the resolution or decreases the transient times by half. Perhaps most importantly for the future, this platform makes targeted proteomics feasible in various applications. Already, we have demonstrated that multiplexed SIM scans can increase the sensitivity up to 50-fold. Targeted approaches are complementary to data-dependent methods that primarily aim at high duty cycles in peptide sequencing. In the future, however, both strategies may be combined to ensure the measurement of specific peptides and comprehensive proteome analysis in shotgun approaches.

We found that the best use of the ultra-high resolution featured by the compact high-field Orbitrap analyzer incorporated into the Orbitrap Elite instrument was to translate resolution into a speed advantage for shotgun proteomics. This is done by selecting shorter transient lengths or by using the high resolution Orbitrap analyzer in parallel with the linear ion trap for tandem mass spectra acquisition. Beyond shotgun proteomics, ultra-high resolution proved to be extremely beneficial for top-down proteomics where it achieved baseline resolution of the isotope patterns of purified proteins of medium to even large size and improved the sequence coverage in their tandem mass spectra. Furthermore, it became possible to investigate intact proteins by MS and MS/MS on a chromatography time-scale (Article 3).

While high resolution is a system-inherent parameter to the mass spectrometer, high mass accuracy is enabled by high resolution, but also depends on the calibration procedure. We performed a comparison between the mass accuracy achieved by internal calibration with the so-called *lock mass* procedure and with a non-linear recalibration algorithm applied to the data during post-processing. The *software lock mass* matched the maximum hardware capabilities of the instrument in terms of mass accuracy, which clearly demonstrated that addressing the problem of calibration at a later point in the shotgun proteomics pipeline is highly practical and beneficial (Article 5).

The high mass accuracy MS/MS data obtained with HCD fragmentation from the dual linear ion trap Orbitrap platform introduced in 2009 was a clear example of superior data quality compared to previous technological platforms. Visual inspection of the tandem mass spectra immediately convinced the trained expert, however, the commonly used commercial peptide database search tools such as Mascot were not able to access the full potential of this data. A novel search engine, *Andromeda*, developed by Jürgen Cox as an addition to the MaxQuant software package, features similar performance as Mascot, but is capable of searching fragment ion spectra with arbitrarily high mass accuracy. In this context, we also implemented a novel algorithm for a second database search that attempts to identify a second peptide in the fragmentation window and thereby takes maximum advantage of the high mass accuracy. We found that Andromeda retrieved about 10% additional identifications of unique peptides. This does not fully solve the inherent problem of co-fragmenting precursor ions (Article 1), but it often turns the apparent disadvantage into a benefit, i.e. a second peptide identification (Article 4).

Independent of high mass accuracy and regardless of successful search engine identification, many tandem mass spectra contain highly abundant peaks that lack an assignment. This situation can be confusing when visually inspecting results. The advent of the *high-high* strategy now makes it possible to unambiguously explain most of such unassigned peaks. The basic chemistry of peptide fragmentation in the gas-phase has been studied in great detail for decades. However, mass spectrometrists usually had to rely on very low numbers of peptides investigated with older mass spectrometer types or somewhat larger numbers of low resolution MS/MS spectra acquired in ion traps. In this thesis, we have critically reviewed, brought together, checked and partially extended the knowledge on peptide fragmentation mechanisms and ion types using our large collections of high resolution collision induced fragmentation spectra. To make this information easily accessible and enable comprehensive annotation of large-scale datasets, we have established a computer-assisted Expert System that provides annotations of tandem mass spectra in publication quality (Article 7). An in-depth statistical investigation of the ion types found in higher energy collisional dissociation spectra was iteratively carried out between Expert System development and human expert annotation. We also used the automated annotation to compare the nature of HCD and ion trap fragmentation spectra of tryptic peptides (Article 6).

The technological improvements outlined in this thesis then culminated in the successful identification of novel proteins of human cytomegalovirus despite the highly complex background from a human cell line. High sequencing speed in conjunction with high MS/MS data quality and comprehensive automated annotation were the key-features in this success. The mass spectrometry-based proteomics data convincingly complement the ribosome profiling findings of the same cellular system, which not only supports the biological evidence of the results on new open reading frames, but also underlines the maturity and potential of both technologies (Article 8).

Taken together, shotgun proteomics technology continues to present interesting challenges inherent to the nature of proteomics samples and to the data-dependent analysis, which despite alternative developments is still by far the most successful strategy. Remaining areas for development are clearly related to the accessibility of the peptidome, because the proteome can often be covered in sufficiently high depth already. Other fundamental challenges include the electrospray ionization process that is still very poorly understood. For instance, ion suppression is known to have negative effects on instrument sensitivity and signal-to-noise ratios. Improvements in these basic issues therefore may translate into extended mass and fragment mass measurements with a remarkable potential for better data quality and thus for the depth of biological conclusions based on proteomics studies. Furthermore, it became clear during this thesis that there are more than enough peptide precursor ions available (Article 1). Due to the high complexity of overlapping elution profiles, however, it is not trivial to decide on the fly which precursor ions to target for fragmentation. It would be desirable to select each precursor ion only once at optimal quality, in the case of SILAC pairs just the more abundant partner, and to schedule the fragmentation event at the apex of the chromatographic peak. These and even more sophisticated approaches may in the future be enabled by *intelligent data acquisition* software. This was recently principally demonstrated[124] and would allow to optimally benefit from the enhanced instrument capabilities. Finally, in order to successfully complement next-generation DNA and RNA sequencing, it is of primary importance to focus on the robustness of proteomics technology and to establish fast, inexpensive and user-friendly workflows and instrumentation. The more and more streamlined proteomics pipeline, including the latest benchtop format and largely automated data analysis tools, provide a promising perspective for moving mass spectrometry-based proteomics technology into biological laboratories and into the clinic where it can be used by an increasing number of scientists.

# References

1 Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/technology* **14**, 61-65 (1996).

2 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207, doi:10.1038/nature01511 (2003).

3 Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G. *et al.* Nucleotide-Sequence of Bacteriophage-Phi-X174. *Journal of molecular biology* **125**, 225-246 (1978).

4 Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695 (1977).

5 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467 (1977).

6 Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).

7 Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science* **270**, 397-403 (1995).

8 Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B. *et al.* Life with 6000 genes. *Science* **274**, 546, 563-547 (1996).

9 Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

10 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

11 Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 19428-19433, doi:10.1073/pnas.0709013104 (2007).

12 Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. *Science* **270**, 467-470 (1995).

13 Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology* **14**, 1675-1680, doi:10.1038/nbt1296-1675 (1996).

14 Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. M. Expression profiling using cDNA microarrays. *Nature genetics* **21**, 10-14, doi:10.1038/4434 (1999).

15 Mardis, E. R. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics* **9**, 387-402, doi:10.1146/annurev.genom.9.081307.164359 (2008).

16 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63, doi:10.1038/nrg2484 (2009).

17 Yates, J. R., 3rd. Mass spectrometry. From genomics to proteomics. *Trends in genetics : TIG* **16**, 5-8 (2000).

18 Nesvizhskii, A. I. Protein identification by tandem mass spectrometry and sequence database searching. *Methods in molecular biology* **367**, 87-119, doi:10.1385/1-59745-275-0:87 (2007).

19 Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry* **80**, 273-299, doi:10.1146/annurev-biochem-061308-093216 (2011).

20 de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251-1254, doi:10.1038/nature07341 (2008).

21 Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7**, 548, doi:10.1038/msb.2011.81 (2011).

22 Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A. *et al.* The quantitative proteome of a human cell line. *Molecular systems biology* **7**, 549, doi:10.1038/msb.2011.82 (2011).

23 Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs (vol 28, pg 503, 2010). *Nature biotechnology* **28**, 756-756, doi:Doi 10.1038/Nbt0710-756b (2010).

24 Haas, B. J. & Zody, M. C. Advancing RNA-Seq analysis. *Nature biotechnology* **28**, 421-423, doi:10.1038/nbt0510-421 (2010).

25 Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515, doi:10.1038/nbt.1621 (2010).

26 Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annual review of medicine* **63**, 35-61, doi:10.1146/annurev-med-051010-162644 (2012).

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

28 Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-2301 (1988).

29 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).

30 Sekioka, T., Terasawa, M. & Awaya, Y. Ion Storage in Kingdon Trap. *Radiat Eff Defect S* **117**, 253-259 (1991).

31 Knight, R. D. Storage of Ions from Laser-Produced Plasmas. *Appl Phys Lett* **38**, 221-223 (1981).

32 Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem* **72**, 1156-1162 (2000).

33 Hardman, M. & Makarov, A. A. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal Chem* **75**, 1699-1705 (2003).

34 Schwartz, J. C., Senko, M. W. & Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry* **13**, 659-669 (2002).

35 Scigelova, M. & Makarov, A. Orbitrap mass analyzer--overview and applications in proteomics. *Proteomics* **6 Suppl 2**, 16-21, doi:10.1002/pmic.200600528 (2006).

36 Mamyrin, B. A., Karataev, V. I., Shmikk, D. V. & Zagulin, V. A. Mass-Reflectron a New Nonmagnetic Time-of-Flight High-Resolution Mass-Spectrometer. *Zh Eksp Teor Fiz+* **64**, 82-89 (1973).

37 Okumura, D., Toyoda, M., Ishihara, M. & Katakuse, I. A compact sector-type multi-turn time-of-flight mass spectrometer 'MULTUM II'. *Nucl Instrum Meth A* **519**, 331-337, doi:DOI 10.1016/j.nima.2003.11.249 (2004).

38 Verentchikov, A. N., Ens, W. & Standing, K. G. Reflecting Time-of-Flight Mass-Spectrometer with an Electrospray Ion-Source and Orthogonal Extraction. *Anal Chem* **66**, 126-133 (1994).

39 Paul, W., Reinhard, H. P. & Vonzahn, U. Das Elektrische Massenfilter Als Massenspektrometer Und Isotopentrenner. *Z Phys* **152**, 143-182 (1958).

40 Stafford, G. C., Kelley, P. E., Syka, J. E. P., Reynolds, W. E. & Todd, J. F. J. Recent Improvements in and Analytical Applications of Advanced Ion Trap Technology. *Int J Mass Spectrom* **60**, 85-98 (1984).

41 Mayya, V., Rezaul, K., Cong, Y. S. & Han, D. Systematic comparison of a two-dimensional ion trap and a three-dimensional ion trap mass spectrometer in proteomics. *Molecular & cellular proteomics : MCP* **4**, 214-223, doi:10.1074/mcp.T400015-MCP200 (2005).

42 Hager, J. W. A new linear ion trap mass spectrometer. *Rapid Commun Mass Sp* **16**, 512-526, doi:Doi 10.1002/Rcm.607 (2002).

43 Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E. *et al.* A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Molecular & cellular proteomics : MCP* **8**, 2759-2769, doi:10.1074/mcp.M900375-MCP200 (2009).

44 Second, T. P., Blethrow, J. D., Schwartz, J. C., Merrihew, G. E., MacCoss, M. J. *et al.* Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. *Anal Chem* **81**, 7757-7765, doi:10.1021/ac901278y (2009).

45 Remes, P. M. & Schwartz, J. C. Performance Improvements in High Mass Range Modes on a Dual Pressure Linear Ion Trap. *Proceedings for 57th ASMS Meeting* (2008).

46 Comisaro.Mb & Marshall, A. G. Fourier-Transform Ion-Cyclotron Resonance Spectroscopy. *Chem Phys Lett* **25**, 282-283 (1974).

47 Lawrence, E. O. & Edlefsen, N. E. On the production of high speed protons. *Science* **72**, 376 (1930).

48 Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. Fourier transform mass spectrometry. *Molecular & cellular proteomics : MCP* **10**, M111 009431, doi:10.1074/mcp.M111.009431 (2011).

49 Schaub, T. M., Hendrickson, C. L., Horning, S., Quinn, J. P., Senko, M. W. *et al.* High-performance mass spectrometry: Fourier transform ion cyclotron resonance at 14.5 tesla. *Anal Chem* **80**, 3985-3990, doi:Doi 10.1021/Ac800386h (2008).

50 Bruker Solarix, T.

51 Leach III, F. E., Anderson, G., MNorheim, R., Futrell, J. H., Tolmachev, A. *et al.* Disign Considerations for High Field FT-ICR MS at 21T. *Proceedings for 60th ASMS Meeting* (2012).

52 Perry, R. H., Cooks, R. G. & Noll, R. J. Orbitrap Mass Spectrometry: Instrumentation, Ion Motion and Applications. *Mass spectrometry reviews* **27**, 661-699, doi:Doi 10.1002/Mas.20186 (2008).

53 Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O. *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem* **78**, 2113-2120, doi:10.1021/ac0518811 (2006).

54 Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M. *et al.* The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry : JMS* **40**, 430-443, doi:10.1002/jms.856 (2005).

55 Makarov, A. & Denisov, E. Dynamics of ions of intact proteins in the Orbitrap mass analyzer. *Journal of the American Society for Mass Spectrometry* **20**, 1486-1495, doi:10.1016/j.jasms.2009.03.024 (2009).

56 Marshall, A. G. & Hendrickson, C. L. High-resolution mass spectrometers. *Annual review of analytical chemistry* **1**, 579-599, doi:10.1146/annurev.anchem.1.031207.112945 (2008).

57 Makarov, A., Denisov, E. & Lange, O. Performance evaluation of a high-field Orbitrap mass analyzer. *Journal of the American Society for Mass Spectrometry* **20**, 1391-1396, doi:10.1016/j.jasms.2009.01.005 (2009).

58 Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P. *et al.* Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular & cellular proteomics : MCP* **4**, 2010-2021, doi:10.1074/mcp.T500030-MCP200 (2005).

59 Makarov, A., Denisov, E., Lange, O. & Horning, S. Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *Journal of the American Society for Mass Spectrometry* **17**, 977-982, doi:10.1016/j.jasms.2006.03.006 (2006).

60 Schwartz, J. C., Zhou, X. G. & Bier, M. E. U.S. Patent 5,572,022.

61 Elias, J. E., Haas, W., Faherty, B. K. & Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature methods* **2**, 667-675, doi:10.1038/nmeth785 (2005).

62 Limbach, P. A., Grosshans, P. B. & Marshall, A. G. Experimental-Determination of the Number of Trapped Ions, Detection Limit, and Dynamic-Range in Fourier-Transform Ion-Cyclotron Resonance Mass-Spectrometry. *Anal Chem* **65**, 135-140 (1993).

63 Glish, G. L. & Burinsky, D. J. Hybrid mass spectrometers for tandem mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **19**, 161-172, doi:DOI 10.1016/j.jasms.2007.11.013 (2008).

64 Glish, G. L. & Goeringer, D. E. Tandem Quadrupole-Time-of-Flight Instrument for Mass-Spectrometry Mass-Spectrometry. *Anal Chem* **56**, 2291-2295 (1984).

65 van den Heuvel, R. H., van Duijn, E., Mazon, H., Synowsky, S. A., Lorenzen, K. *et al.* Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. *Anal Chem* **78**, 7473-7483, doi:10.1021/ac061039a (2006).

66 Cornish, T. J. & Cotter, R. J. Collision-Induced Dissociation in a Tandem Time-of-Flight Mass-Spectrometer with 2 Single-Stage Reflectrons. *Org Mass Spectrom* **28**, 1129-1134 (1993).

67 Cotter, R. J., Griffith, W. & Jelinek, C. Tandem time-of-flight (TOF/TOF) mass spectrometry and the curved-field reflectron. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **855**, 2-13, doi:10.1016/j.jchromb.2007.01.009 (2007).

68 Syka, J. E., Marto, J. A., Bai, D. L., Horning, S., Senko, M. W. *et al.* Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *Journal of proteome research* **3**, 621-626 (2004).

69 Johnson, J. V., Yost, R. A., Kelley, P. E. & Bradford, D. C. Tandem-in-Space and Tandem-in-Time Mass-Spectrometry - Triple Quadrupoles and Quadrupole Ion Traps. *Anal Chem* **62**, 2162-2172 (1990).

70 Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F. & Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem* **76**, 3590-3598 (2004).

71 Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* **4**, 709-712, doi:10.1038/nmeth1060 (2007).

72 Sanger, F. & Thompson, E. O. The amino-acid sequence in the glycyl chain of insulin. *The Biochemical journal* **52**, iii (1952).

73 Sanger, F. & Thompson, E. O. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *The Biochemical journal* **53**, 366-374 (1953).

74 Sanger, F. & Thompson, E. O. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal* **53**, 353-366 (1953).

75 Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *The Biochemical journal* **49**, 481-490 (1951).

76 Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal* **49**, 463-481 (1951).

77 Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* **68**, 850-858 (1996).

78 Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal Chem* **68**, 1-8 (1996).

79 Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology* **17**, 676-682, doi:10.1038/10890 (1999).

80 Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & cellular proteomics : MCP* **3**, 608-614, doi:10.1074/mcp.T400003-MCP200 (2004).

81 van Breukelen, B., Georgiou, A., Drugan, M. M., Taouatas, N., Mohammed, S. *et al.* LysNDeNovo: an algorithm enabling de novo sequencing of Lys-N generated peptides

fragmented by electron transfer dissociation. *Proteomics* **10**, 1196-1201, doi:10.1002/pmic.200900405 (2010).

82 Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nature methods* **6**, 359-362, doi:10.1038/nmeth.1322 (2009).

83 Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Frohlich, F. *et al.* Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular & cellular proteomics : MCP* **10**, M110 003699, doi:10.1074/mcp.M110.003699 (2011).

84 Kocher, 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, doi:10.1021/ac103243t (2011).

85 Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K. *et al.* System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Molecular & cellular proteomics : MCP* **11**, M111 013722, doi:10.1074/mcp.M111.013722 (2012).

86 Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**, 663-670 (2003).

87 Wisniewski, J. R., Zougman, A. & Mann, M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *Journal of proteome research* **8**, 5674-5678, doi:10.1021/pr900748n (2009).

88 Hubner, N. C., Ren, S. & Mann, M. Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics* **8**, 4862-4872, doi:10.1002/pmic.200800351 (2008).

89 Schmidt, A., Claassen, M. & Aebersold, R. Directed mass spectrometry: towards hypothesis-driven proteomics. *Current opinion in chemical biology* **13**, 510-517, doi:10.1016/j.cbpa.2009.08.016 (2009).

90 Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **389**, 1017-1031, doi:10.1007/s00216-007-1486-6 (2007).

91 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*, doi:10.1007/s00216-012-6203-4 (2012).

92 Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics : MCP* **1**, 376-386 (2002).

93 Kruger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C. A. *et al.* SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* **134**, 353-364, doi:10.1016/j.cell.2008.05.033 (2008).

94 Geiger, T., Wisniewski, J. R., Cox, J., Zanivan, S., Kruger, M. *et al.* Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nature protocols* **6**, 147-157, doi:10.1038/nprot.2010.192 (2011).

95 Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature methods* **7**, 383-385, doi:10.1038/nmeth.1446 (2010).

96 Boersema, P. J., Mohammed, S. & Heck, A. J. Phosphopeptide fragmentation and analysis by mass spectrometry. *Journal of mass spectrometry : JMS* **44**, 861-878, doi:10.1002/jms.1599 (2009).

97 Hsu, J. L., Huang, S. Y., Chow, N. H. & Chen, S. H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* **75**, 6843-6852, doi:10.1021/ac0348625 (2003).

98 Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J. *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).

99 Ross, P. L., Huang, Y. L. N., Marchese, J. N., Williamson, B., Parker, K. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics* **3**, 1154-1169, doi:DOI 10.1074/mcp.M400129-MCP200 (2004).

100 Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods* **8**, 937-940, doi:10.1038/nmeth.1714 (2011).

101 Wenger, C. D., Lee, M. V., Hebert, A. S., McAlister, G. C., Phanstiel, D. H. *et al.* Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nature methods* **8**, 933-935, doi:10.1038/nmeth.1716 (2011).

102 Luber, C. A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M. *et al.* Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **32**, 279-289, doi:10.1016/j.immuni.2010.01.013 (2010).

103 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

104 Cox, J. & Mann, M. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *Journal of the American Society for Mass Spectrometry* **20**, 1477-1485, doi:10.1016/j.jasms.2009.05.007 (2009).

105 Mann, M. & Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 18132-18138, doi:10.1073/pnas.0800788105 (2008).

106 Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society* **120**, 3265-3266 (1998).

107 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. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9528-9533, doi:DOI 10.1073/pnas.0402700101 (2004).

108 Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology : a journal of computational molecular cell biology* **6**, 327-342, doi:10.1089/106652799318300 (1999).

109 Ma, B. & Johnson, R. De novo sequencing and homology searching. *Molecular & cellular proteomics : MCP* **11**, O111 014902, doi:10.1074/mcp.O111.014902 (2012).

110 Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999).

111 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **4**, 207-214, doi:Doi 10.1038/Nmeth1019 (2007).

112 Nagaraj, N., D'Souza, R. C., Cox, J., Olsen, J. V. & Mann, M. Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation. *Journal of proteome research* **9**, 6786-6794, doi:10.1021/pr100637q (2010).

113 Nagaraj, N., D'Souza, R. C., Cox, J., Olsen, J. V. & Mann, M. Correction to Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation. *Journal of proteome research*, doi:10.1021/pr3003886 (2012).

114 Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & cellular proteomics : MCP* **11**, O111 016717, doi:10.1074/mcp.O111.016717 (2012).

115 Panchaud, A., Scherl, A., Shaffer, S. A., von Haller, P. D., Kulasekara, H. D. *et al.* Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem* **81**, 6481-6488, doi:10.1021/ac900888s (2009).

116 Bateman, K. P., Kellmann, M., Muenster, H., Papp, R. & Taylor, L. Quantitative-qualitative data acquisition using a benchtop Orbitrap mass spectrometer. *Journal of the American Society for Mass Spectrometry* **20**, 1441-1450, doi:10.1016/j.jasms.2009.03.002 (2009).

117 Dongre, A. R., Jones, J. L., Somogyi, A. & Wysocki, V. H. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *Journal of the American Chemical Society* **118**, 8365-8374 (1996).

118 Yalcin, T., Khouw, C., Csizmadia, I. G., Peterson, M. R. & Harrison, A. G. Why are B ions stable species in peptide spectra? *Journal of the American Society for Mass Spectrometry* **6**, 1165-1174 (1995).

119 Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* **24**, 508-548, doi:10.1002/mas.20024 (2005).

120 Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58-63, doi:10.1038/nature07228 (2008).

121 Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64-71, doi:10.1038/nature07242 (2008).

122 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223, doi:10.1126/science.1168978 (2009).

123 Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R. *et al.* The DNA sequence of the human cytomegalovirus genome. *DNA sequence : the journal of DNA sequencing and mapping* **2**, 1-12 (1991).

124 Graumann, J., Scheltema, R. A., Zhang, Y., Cox, J. & Mann, M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Molecular & cellular proteomics : MCP* **11**, M111 013185, doi:10.1074/mcp.M111.013185 (2012).

# Acknowledgements

Special thanks to…

Matthias Mann not only for sharing his outstanding knowledge about mass spectrometry and his vision of the future of proteomics research, but also for constantly coming up with new challenges and ideas in the most encouraging and supportive way.

Our collaborators from ThermoScientific, Bremen, in particular Eugen Damoc, Oliver Lange, Andreas Kühn and Stevan Horning for very open discussions and great support with the novel instrumentation, Rosa Viner and Vlad Zabrouskov for their hospitality during my visit in San Jose, US.

Noam Stern-Ginossar and Jonathan Weissman from the University of California in San Francisco for the very interesting and fruitful collaboration.

Jürgen Cox for immense support during large parts of this thesis and Nadin Neuhauser for together successfully dealing with all obstacles of everyday life.

Korbinian Mayr and Igor Paron for excellent support with the instrumentation and many other members of the department for the great working atmosphere.

Stefanie Uhlmann, Rochelle D'Souza, Michal Bassani-Sternberg, Christian Eberl and Dirk Walter for proofreading and comments.

Nagarjuna Nagaraj for very enjoyable trips to mass spectrometry meetings and beyond, Tami Geiger for discussing more than just scientific problems and the girls' office for sharing insights into different cultures, tastes and ideas of life. Maybe these are the most precious experiences and memories.

Nils Konken for making life so much more colourful and pleasant.

My little sister for being around whenever needed and for her honest and critical advice.

My parents for keeping in touch over any distance in the world, for often trusting more than I do and for being generous and supportive no matter what.