Proteomics of Transcription Factors

Nikolai Mischerikow

ISBN 978-90-6464-464-1

Printed by GVO drukkers & vormgevers B.V. | Ponsen & Looijen

The research presented in this thesis was perfomed in the Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research, at the University Utrecht, and the Department of Molecular Cancer Research at the University Medical Center Utrecht. The research was financially supported by the Netherlands Genomics Initiative Horizon Program, grant number 050-71-050.

Proteomics of Transcription Factors

Proteomics van Transcriptiefactoren
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op woensdag 13 april 2011 des ochtends te 10.30 uur

door

Nikolai Mischerikow
geboren op 10 februari 1978, te Alfeld (Leine), Duitsland

Promotor Prof.dr. A.J.R. Heck

Table of contents

CHAPTER 1

GENERAL INTRODUCTION

1 Protein analysis by mass spectrometry

Protein analysis by mass spectrometry (MS) is becoming increasingly integrated with molecular biology and is nourished by the technological advances in what is known as proteomics. In allusion to genomics, proteomics aims at the qualitative and quantitative characterization of the proteome, which is generally understood as the whole of proteins present in a defined entity of living matter, for example organisms, organs, tissues, or cells. This is a formidable task when apprehending that one genome can express many different proteomes. One example is the overlap between proteomes of different cell types within one organism due to different gene expression patterns. Another example is the perturbation of a cellular proteome upon an external stimulus, which in first instance often leads to post-translational modifications (PTMs) of the existing proteome rather than changes in gene expression. Besides that, the abundance distribution of proteomes is usually extremely large, spanning multiple orders of magnitude.

To date, only mass spectrometry has proven capable of providing comprehensive profiles of proteomes. During the past decade, a key concept has been optimized which is fundamentally based on the controlled decomposition of a proteome into peptides and their analysis by mass spectrometry. One idea behind this is that compared to an ensemble of proteins its representative ensemble of peptides has a much narrower distribution of physicochemical properties, which in addition are analytically and preparatively much less challenging. The other idea is that the fragmentation of peptides by tandem mass spectrometry (MS/MS) is well understood and yields fragmentation spectra that can be used to identify the peptides and with this ultimately the represented proteins. The concept therefore allows comprehending very different proteomes by generic analytical workflows with peptide mass spectrometry at the core. The technological advances made in optimizing all stages of these generic proteomic workflows, notably mass spectrometric instrumentation, automated peptide identification and the establishment of quantitative techniques, have greatly benefited mass spectrometry-based protein analysis.

2 Technology for proteomics

Mass spectrometers ultimately measure the mass-to-charge ratio (m/z) of ionized molecules in the gas phase. Therefore they are built from two integral components: the

ion source, which charges the analyte and transfers it into the gas phase, and the mass analyzer that separates ions based on their m/z to acquire the mass spectrum. To provide collision induced dissociation (CID) fragmentation data, a mass spectrometer must be able to provide MS/MS capability. In recent years, the gas-phase ion-molecule reaction of electron transfer dissociation (ETD) has been added to the standard repertoire of many commercial proteomics-grade mass spectrometers to generate sequence-informative peptide fragmentation spectra. Regarding the ion source, the choice of ionization techniques is limited to those that deposit the energy for ionization and phase transition so gently that the analyte does not dissociate during ionization. Commonly, electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) sources are used. The following section will give a brief overview of the operational principles and key features of these technologies.

2.1 Electrospray ionization [1-3]

Electrospray ionization is ideally suited for the analysis of biomolecules because it can generate positively and negatively charged analyte ions gently enough to analyze proteins, carbohydrates, oligonucleotides and peptides without dissociation and can even be applied to study non-covalent protein complexes [4-6]. It is performed directly from a dilute solution of the non-volatile analyte at atmospheric ambient pressure. It can be operated continuously and can therefore be conveniently interfaced to liquid chromatography- (LC) based separation techniques [7,8]. Albeit the actual ESI process might seem relatively simple compared to other ionization techniques, it is in practice a very delicate process with profound effects on the overall performance of the mass spectrometer. A stable electrospray (Figure 1) producing a continuous beam of ions is prerequisite to precisely control the number of ions sampled into the mass spectrometer, which is one of the key capabilities of modern proteomics-grade instruments to control performance characteristic like mass resolution and analyte sensitivity. It also improves total ion current- (TIC) and selected reaction monitoring- (SRM) based quantification methods. A well-adjusted electrospray prevents artificial oxidation of analytes with a low reduction potential, for example the methionine residue in peptides and proteins.

Electrospray operating principles. To produce an electrospray, a dilute solution of analyte is pumped at low flow rate through an electrically conductive capillary opening. The electrospray is generated by the application of an electrical potential difference between the capillary opening and the orifice of the mass spectrometer. The electrical field

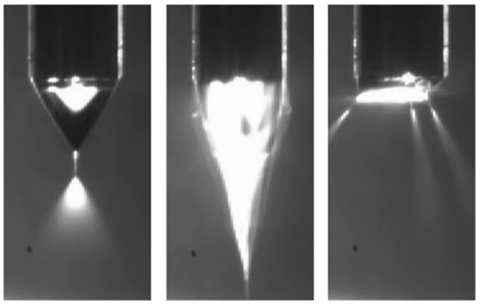


Figure 1. Picture of the stable cone jet mode of the ESI process as it can be observed at an LC-MS interface of LTQ Orbitrap instruments (left). The ESI process can also adopt the pulsed cone jet (middle) and the multi-jet mode (right); the stable cone jet mode, however, is the desired mode for ESI. Video stills from http://www.eng.qmul.ac.uk/subsites/electrospray/videos.php

gradient results in the polarization of the surface of the solution as it protrudes from the capillary. If the electrical potential of the capillary is positive, negatively charged ions near the surface of the solution are dragged towards the capillary and positively charged ions accumulate at the surface. The force exerted by the field on the polarized surface, which pulls the solution towards the mass spectrometer, and the cohesive force of the solvent that works towards a minimal solution surface area, overlay and result in a characteristic conical distortion of the surface area with the apex pointing towards the mass spectrometer. If the potential difference is sufficiently high, a positively charged liquid jet emerges from the apex. This jet quickly becomes instable and under charge conservation fragments into fine droplets. Solvent evaporates during the movement of the droplets along the field gradient and the droplet size decreases. This results in an increased charge density and, when electrostatic repulsion exceeds droplet cohesion, fission of the droplet. This process is repetitive and has been observed for larger droplets. It may also apply to the final transition of the analyte from nanometer-sized droplets into the gas phase. An alternative model assumes that in this late phase the charged analyte enters the gas phase directly from the droplet through evaporation. Effectively, ionization by electrospray produces solvent-free, charged analyte ions which can enter the mass spectrometer or discharge at the orifice. In order to maintain the electrical potential difference during continuous charge transport from the solution to the mass spectrometer, electrochemical processes continuously replenish the net charge exiting the capillary. In positive ion mode, the capillary functions as anode at which oxidative processes generate electrons. The identity of the reaction with the highest reduction potential depends mainly on the material of the capillary and the composition of the analyte solution.

Analyte properties determining the efficiency of ESI. Positive mode ESI is very sensitive for peptides and proteins due to the highly basic guanidino and primary amino

moieties, which enable positive net charging of these analytes by protonation during the polarization in the initial phase of the electrospray process. These moieties also have high gas phase basicities which results in retention of the attached protons during proton transfer reactions (PTR) occurring in the gas phase during the late phase of the electrospray process. PTRs may even lead to additional charging of peptides and proteins because the amide nitrogen of the peptide bond has a high basicity in the gas phase but a very low basicity in solution. Next to basicity, polarity is the other important chemical property of an analyte that determines its response in ESI [9,10]. Analytes bearing hydrophobic structural portions have a higher affinity for solution surfaces than polar analytes. Thus they tend to carry a greater fraction of the charge produced in the electrospray process and ultimately display a higher response. This effect can be very pronounced, for example when electrospraying peptides contaminated with detergent, where it leads to suppression of the peptide analyte.

Nanoelectrospray ionization (nanoESI) and interfacing with LC. Electrospray ionization is frequently used to sample peptides from an LC column via a spray emitter directly into the mass spectrometer (LC-MS). Among the different available LC methods for peptide separation, reversed phase (RP) LC (RP-LC) is commonly used because the composition of the liquid phase, an acidified mix of water and volatile organic solvent, is compatible with electrospray. Reducing the flow rate at which the peptides elute from the column as well as the diameter of the emitter opening increases the sensitivity of the electrospray process [11]. This is caused by an increased efficiency of ion creation and transfer into the mass spectrometer. Flow rates in the range of tens to hundreds nanoliter per minute and emitter openings of a few micrometer are commonly used. To suite these dimensions, RP-LC is performed with capillary columns of tens of micrometer inner diameter. Particle sizes below five micrometer are frequently used to achieve high peak capacities resulting in pressures still accommodated by high pressure liquid chromatography (HPLC) systems. Specialized emitters have also been developed for further improved nanoESI [12,13].

2.2 Mass analyzers [14]

Mass analyzers can be classified into beam-type devices like quadrupole and time-of-flight analyzers, and trapping devices like ion trap, orbitrap and ion cyclotron resonance analyzers. Trapping devices achieve m/z separation by resonant interaction while beam-type devices separate m/z either based on trajectory stability or on flight time. Mass

analyzers are also used in other functions, for example to produce CID or ETD fragmentation spectra. The following section describes mass analyzers commonly encountered in proteomics and descriptively summarizes selected operational modes.

Three-dimensional quadrupole assemblies

3D quadrupole ion trap mass analyzer (QIT) [15]. This quadrupole ion trap mass analyzer consists of three electrodes as depicted in Figure 2A. A three-dimensional periodic quadrupolar field is established between the ring electrode and the two end-cap electrodes by applying a radio frequency (RF) potential to the ring electrode. Ions of a broad m/z range enter the QIT axially through a bore in one end-cap electrode and are excited into stable, three-dimensional oscillatory motions in the center of the electrode assembly. Electrostatic repulsion within the trapped ion cloud that would lead to ion loss is counteracted by operation at an elevated gas density so as to dissipate kinetic energy by non-fragmenting collisions. Mass filtering is achieved by resonance ejection or by ejection at the stability limit. In both modes, ions leave the quadrupolar field selectively in axial direction through the bores in the end cap electrodes and are detected by an ion detector. For ejection at the stability limit the RF amplitude at the ring electrode is gradually increased, which forces ions with increasing m/z values onto instable axial trajectories, leading to ejection from the trap. For resonance ejection, an additional alternating current (AC) potential is applied to the end cap electrodes while scanning the RF amplitude. Effectively, this forces ions with increasing m/z onto instable trajectories at the point where their axial oscillation frequency matches the AC frequency. With both ejection methods, the mass spectrum is generated by continuous scanning of the RF amplitude.

QIT as collision cell and reaction chamber. To produce a fragment ion spectrum, the precursor ion is selectively retained in the QIT while ions with other m/z values are ejected by one of the method described above. The precursor is then fragmented by collisions with the neutral gas molecules present in the trap. To provide sufficient energy to fragment the precursor ion, the kinetic energy of the precursor ion is increased by resonant excitation. This is achieved by increasing the RF amplitude to match the axial oscillation frequency of the precursor to the frequency of the AC potential applied on the end cap electrodes. The AC potential is sufficiently small to excite axial oscillation without ejection. The generated fragment ions are trapped in the QIT and subsequently mass-selectively analyzed to create the fragment ion spectrum. Fragment ions with low m/z

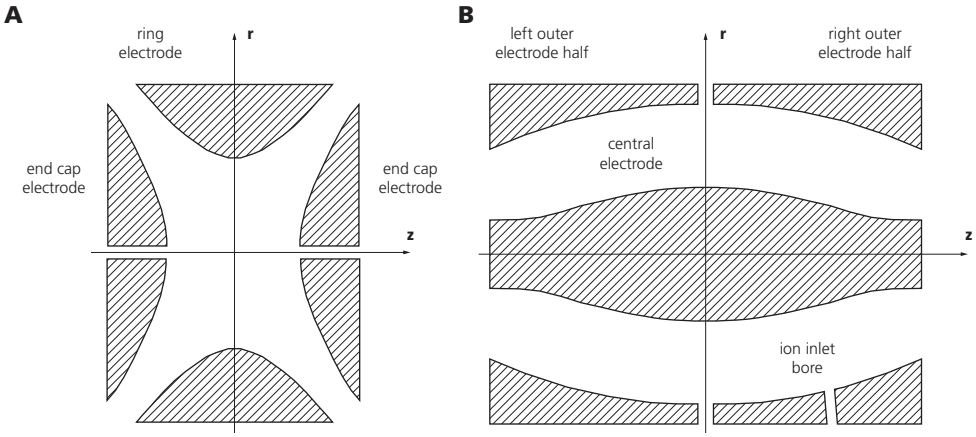


Figure 2. Schematics of the 3D quadrupole ion trap (A) and the orbitrap (B) mass analyzer.

are not trapped because they have instable trajectories at the applied RF amplitude. Although a lower AC frequency would allow fragmentation at lower RF amplitude, the RF amplitude cannot be arbitrarily lowered because the trapping efficiency increases with the square of the RF amplitude. The applied RF amplitude is therefore a compromise between trapping efficiency and lower m/z detection limit. The QIT can also function as reaction chamber to provide gas phase ion-ion reactions like ETD or PTR [16].

Linear quadrupole assemblies [17]

Linear quadrupole mass analyzer (Q). The quadrupole mass analyzer consists of four hyperbolic rods aligned in parallel around a central axis. Each pair of opposing rods is electrically connected and a periodic quadrupolar field perpendicular to the central axis is established by a combined direct current (DC) and RF potential between the two rod pairs. Ions which axially move through the quadrupole are excited into complex radial oscillations by the periodic quadrupolar field. The equation of motion shows that these radial oscillations can be stable if its amplitude is finite. If the amplitude is sufficiently small, the ions pass the quadrupole on a stable trajectory. The radial oscillations can also be instable if the amplitude of oscillation increases exponentially with time; the ions then collide with the rods and discharge. The stability depends on the m/z of the ion and the amplitudes of the applied DC and RF potential. These amplitudes can be set such that only ions with one m/z ratio have stable trajectories and ions with all other m/z ratios have instable trajectories, *i.e.* the quadrupole acts as mass filter. When the DC and RF amplitudes are systematically increased, ions with increasing m/z will be se-

quentially transmitted through the quadrupole to produce a mass spectrum.

Linear quadrupole collision cell (q) and ion guide. When a linear quadrupole is operated without a DC potential, it transmits ions of a broad m/z range and thus can be used as ion guide to transport ions between different instrument parts. The ion guide can also focus the ion beam when operated at an elevated inert gas density. Ions will lose kinetic energy through non-dissociative collisions with gas molecules which effectively dampens their radial oscillation amplitude; this is called collisional cooling. When ions enter an inert gas-filled RF-only quadrupole with sufficiently high kinetic energy, they will fragment by collision with the gas molecules. Mass-selective filtering of the fragments can, however, not be achieved, so to produce a MS/MS spectrum the quadrupole collision cell has to be combined with another mass analyzer downstream.

Linear quadrupole ion trap mass analyzer (LIT) [18]. A linear quadrupole ion trap mass analyzer is build from a linear quadrupole by applying switchable stopping potentials to electrodes at the entrance and exit. When the quadrupole is operated without DC potential, ions with a broad m/z range can be axially confined in the quadrupolar field. Collisional cooling is used to dampen the motion of the ions. Trapped ions are ejected in radial or axial direction by a variety of excitation methods some of which resemble those used in QITs. Two commercial LITs are frequently used in proteomics. One uses mass-selective axial ejection [19,20] while the other uses mass-selective radial ejection through slits in two opposing rods for mass filtering [21].

LIT as collision cell and reaction chamber. The operating sequence for LITs as collision cell is the same as for the QIT. Precursor ions are isolated and fragmented and then the fragment ions are mass-selectively ejected from the trap, either by radial or axial ejection, to produce the fragment ion spectrum. Both axial and radial ejection can be applied. Like the QIT, the LIT can also operate as reaction chamber for ETD; the negatively charged ETD reagent can be introduced into the trap axially from either side of the LIT [22,23]. One commercial setup introduces the reagent from the back end into the trap, after the precursor has been isolated and stored in the front part of the LIT. The precursor is then released to the center of the trap where it reacts with the ETD reagent. The reaction is terminated by axial ejection of the reagent and subsequent mass-selective radial ejection of fragment ions. A short pulse of resonant excitation is often applied after reagent ejection to dissociate fragment ion clusters. This technique is called supplemental activation [24].

Orbitrap mass analyzer and C-trap [25]

The orbitrap mass analyzer is build from a spindle-shaped central electrode confined by a split, barrel-shaped outer electrode as shown in Figure 2B. Unlike the other mass analyzers portrayed so far, a DC-only potential is used for trapping and m/z analysis. The DC potential is applied to the central electrode to establish a quadro-logarithmic field. Ions injected into the orbitrap can adopt stable trajectories involving orbiting motions around the central electrode and oscillations along the axis of the central electrode. The shape of the electrodes and the resulting field geometry causes these motions to be independent of each other. The m/z ratio of the oscillating ions is proportional to the frequency of the axial oscillation. This frequency is measured in the time domain as transient of the image current induced in the outer electrode halves. The signal is discretized and Fast Fourier-transformed to produce the frequency spectrum and with this the mass spectrum. Although the orbital motion of the ions is not used for mass analysis, it has to be stable and coherent during measurement of the image current. This is achieved by time-controlled switching of the DC potential during ion injection, which however requires a narrow spatial and temporal distribution of the injected ion cloud. As the ion cloud is not produced by the orbitrap itself, the operation of the orbitrap critically depends on an outside component. Different injection methods have been described of which ion cloud formation by a curved linear quadrupole ion trap is implemented in the only commercially available orbitrap instrument. This C-trap collects ions in an arched geometry and ejects them on a focusing trajectory towards the orbitrap. The trap geometry and ion ejection process are designed such that the ion beam condenses inside the orbitrap. Like LITs and QITs the orbitrap can also be used to fragment trapped ions; however, this function is not commonly used in proteomics.

Ion cyclotron resonance (ICR) mass analyzer [26]

The ion cyclotron mass analyzer is based on the interaction of ions with a strong and constant magnetic field. It is build from a superconducting magnet and a box-shaped cell stretched perpendicular to the magnetic field. The cell consists of two pairs of opposing metal plates that are pair-wise electrically coupled. Due to the magnetic field, ions inside the cell are trapped on stable orbital oscillations with an angular oscillation frequency depending on the magnetic field strength and their m/z. Next to the magnetic field, a pulsed RF potential is applied to one pair of opposing metal plates to resonantly excite the oscillating ions. The excitation results in an increase of the oscil-

Commercial instrument	Mass analyzer	Resolution	Mass accuracy	Sensitivity	Dynamic range	Scan speed
LTQ	LIT	2,000	100 ppm	femtomole	10,000	fast
TSQ	QqQ	2,000	100 ppm	attomole	1,000,000	moderate
LTQ Orbitrap	LIT orbitrap	100,000	2 ppm	attomole	10,000	moderate
LTQ FT	LIT ICR	500,000	2 ppm	femtomole	10,000	slow
Q-TOF	Qq TOF	10,000	5 ppm	femtomole	1,000,000	moderate

Table 1. Performance characteristics of selected mass spectrometers, adapted from [102].

lation radius and in spatial coherence of the oscillation. When the excitation potential is switched off, the oscillation frequency can be recorded as the image current induced in the other pair of metal plates. When the resonant excitation pulse is implemented as fast frequency sweep over a range of oscillation frequencies (broadband excitation), which results in similar oscillation radii of ions with different m/z, the image current can be converted into a mass spectrum by Fourier transformation. ICR mass analyzers can be used for analyzing ion fragmentation; however, due to the great resolving power but low scan speed they are commonly used for precursor spectrum acquisition in LC-MS.

Time-of-flight (TOF) mass analyzer

Time-of-flight analyzers separate ions based on their transit times through a field-free region. At time zero, ions with a broad range of m/z are accelerated by an electric field and allowed to drift through space. The arrival time of the ions is related to their m/z and the zero time point can is defined by pulsing the electric field or by pulsing the introduction of ions into a constant field. The mass spectrum is obtained by measuring the arrival time at fixed flight distance by an ion detector. The technical design of the TOF mass analyzer depends on the coupling to the ion source, but often the flight path involves an ion reflector which compensates for the initial spread of kinetic energy of the ions.

Hybrid mass analyzers

Each of the selected mass analyzers sketched above has its unique performance characteristic which is defined by variables such as analysis speed, mass resolution, mass accuracy, sensitivity and dynamic range. Some feature unique analysis modes such as

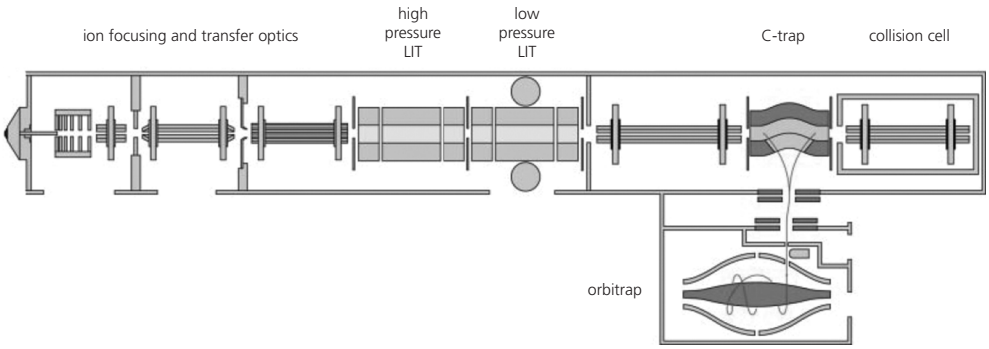


Figure 3. Schematic of the LTQ Orbitrap Velos instrument, adapted from [29].

ETD fragmentation. Nearly all modern mass spectrometers are hybrid instruments that combine multiple mass analyzers to satisfy specific analytical requirements. Certain instrument configurations have become widely accepted in the proteomics community and are commercialized. The highest overall performance is currently achieved by linear quadrupole assemblies (Q, q or LIT) alone or in combination with TOF, and LIT in combination with the orbitrap. Another high-performance hybrid is LIT with ICR. Some performance characteristics are summarized in Table 1.

The only commercial orbitrap instrument, the LTQ Orbitrap (Thermo Fisher Scientific) [27], is probably the best performing mass spectrometer for peptide analysis currently available. It combines a linear ion trap that operates fast and with high sensitivity, with an orbitrap that that performs with high mass accuracy, high resolution and high dynamic range. The LIT is used to generate the fragment ion spectrum by CID or ETD, while the orbitrap is used to acquire a full mass spectrum. Both acquisitions operate in parallel. The instrument can also fragment precursors by higher-energy collisional dissociation (HCD) in a dedicated collision cell and analyze the fragments in the orbitrap [28]. The latest model of the instrument series, the LTQ Orbitrap Velos, features two LITs in series, one which is optimized for peptide fragmentation and the other which is optimized for fragment ion readout (Figure 3) [29].

2.3 Peptide sequencing and data analysis

The initial step of the analytical workflow is the controlled production of peptides from the protein sample, which is most commonly done enzymatically. Subsequently, the produced peptide mixture is separated and analyzed by LC-MS to acquire the mass and

the fragmentation spectrum of each peptide. These two characteristic data are finally used to identify the peptides and ultimately the proteins. Both peptide and protein identification are commonly performed automatically by database search engines.

Protein digestion

Most commonly peptides are generated from proteins by enzymatic proteolysis using highly site-specific proteases. Prior to proteolytic cleavage, proteins are denatured and disulfide bonds are reduced to make accessible as many cleavage sites as possible. Many proteases are potentially useful for protein digestion but the choice is limited when considering that for mass spectrometry and data analysis peptides with a length of around 6 to 20 amino acids are optimal. Peptides of this length are well generated by Trypsin. Trypsin is also beneficial for the downstream processes because it hydrolyzes the peptide bond C-terminally of lysine or arginine residues [30]. This specificity generates peptides that contain a primary amino group on the N-terminus and a lysine or arginine on the C-terminus, all three of which are highly basic both in solution and in the gas phase. These peptides are therefore at least doubly charged in the gas phase with one positive charge residing on each terminus which makes them ideally suited for CID. Tryptic peptides are less suited for ETD, which performs well on longer and higher-charged peptides. These peptides are to a lower extent also generated by digestion with Trypsin but can also be generated by specific proteases. Examples are Lys-C and Lys-N, which cleave the amide bond at the C- or N-terminus of lysine residues, respectively [31]. While peptides generated with Lys-C have the same charge distribution as tryptic peptides, peptides produced with Lys-N carry two positive charges at the N-terminus, which leads to a specific fragmentation behavior.

Peptide fragmentation

Peptides are commonly fragmented in LITs or QITs or linear quadrupole collision cells, using CID or ETD. Both methods preferably cleave bonds within the peptide backbone, which is prerequisite to produce a sequence-informative fragment ion spectrum.

Collision induced dissociation. CID of doubly charged tryptic peptides produces mainly fragment ions of the b- and y-series (Figure 4). It is performed by colliding accelerated peptide ions with inert gas molecules such as nitrogen or argon. In this process, fractions of the kinetic energy are transformed into internal energy. If the collision energy is large enough the excitation will eventually lead to bond cleavage. The dissociation reaction is

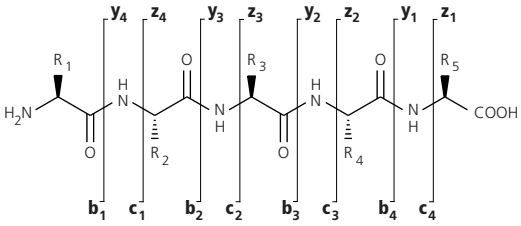


Figure 4. Nomenclature for peptide fragment ions showing the major ion series generated by CID (HCD) and ETD.

mainly charge-directed and can be qualitatively comprehended with the mobile proton model [32]. Upon peptide excitation the attached protons will translocate to other basic moieties, among them the amide nitrogen atom of the peptide bond. Protonation of this atom will destabilize the amide bond and lead to its cleavage. In general, the proton translocation will lead to a distribution of protonated states. The model can distinguish two different classes of peptides. In tryptic peptides with a single charge residing on one of the basic moieties on either peptide terminus, the translocation to the other basic terminus is energetically favored and this state becomes populated upon excitation. The population of other states, including those that potentially lead to peptide bond dissociation, is energetically less favored. Singly charged tryptic peptides are therefore deemed to be in a non-mobile proton situation. In doubly or higher charged tryptic peptides, both terminal basic moieties are protonated and other protonated states are therefore energetically easier accessible upon excitation. Destabilizing protonated states have a higher chance of being populated. These peptides are considered to be in the mobile proton situation and fragment easier in CID than peptides with non-mobile protons. Doubly charged tryptic peptides are therefore well-suited for fragmentation with CID. Additionally, they produce mainly singly charged fragment ions, which helps spectral interpretation.

In terms of impulse duration and energy conversion the collision induced excitation in linear quadrupole collision cells is different from that in ion traps. Due to the resonance excitation process, collisions in LITs and QITs transform lower amounts of kinetic energy in longer time compared to collisions in linear quadrupole collision cells. These different energy and time regimes result in different populations of protonated states. In general, low-energy ion trap CID gives fragmentation spectra resulting from a broad and diverse number of excited states, while higher-energy CID yields simpler spectra. HCD represents a recent development of CID which produces fragmentation patterns of the higher-energy collision regime [28].

Electron transfer dissociation [33,34]. ETD of tryptic peptide produces mainly fragment

ions of the c- and z-series (Figure 4). It is based on the capture of a thermal electron by the charged precursor. The electron is provided by the ETD reagent, a negatively charged molecule anion with a high electron transfer potential. The reaction occurs by mixing of the ETD reagent, which in commercial instruments is often fluoranthene, with the peptide for a defined amount of time. The transfer of the electron results in a charge-reduced, odd-electron peptide cation which rearranges and fragments by radical directed dissociation of the bond between the peptide amide nitrogen and the a-carbon atom. The selectivity for this bond is the major benefit of ETD because it makes it ideally suited for the analysis of important classes of post-translationally modified peptides, for example glycosylated and phosphorylated peptides. In CID, these peptides produce fragment ion spectra that contain very little sequence information because they are dominated by the loss of the phosphate or sugar moiety, which is the energetically preferred dissociation pathway. ETD fragment spectra, in contrary, do not display these fragments but rather sequence-informative ion series. Another benefit of ETD is that it efficiently fragments higher-charged peptides, which complements the charge preference of CID. Moreover, ETD of doubly charged peptides generated with Lys-N produces predominantly ions of the c-series because of the charge accumulation on the N-terminus [31]. The major current disadvantage of ETD is the low conversion rate from precursor ion to fragment ions. One of the reasons for this is that the electron transfer often results only in charge reduction but not fragmentation, such that the ETD fragment ion spectra are strongly dominated by the precursor and charge-reduced precursor ions.

Peptide identification [35]

Peptide spectrum matching by database searching. The most commonly used method for automated peptide identification from high-throughput LC-MS experiments is based on both peptide mass and fragmentation data. Database search algorithms such as Mascot [36] and SEQUEST [37] are used to annotate every fragmentation spectrum with a peptide sequence by comparing the experimentally acquired spectrum against theoretically generated spectra stored in a database. The theoretical spectra are generated from *in silico* digestion and fragmentation of a protein database representing the analyzed proteome. As the database is typically derived from an annotated genome database, the number of theoretical spectra is very large. Therefore, the search algorithm limits the theoretical spectra to those whose calculated peptide mass coincide with the measured peptide mass within a suitable mass window. The mass window is determined by the

mass accuracy and mass precision of the mass spectrometer. The measured spectrum is then compared to all candidate theoretical spectra and a score is reported for every peptide spectrum match (PSM) that reflects the quality of the correlation.

PSMs, however, are rarely ideal, for example due to imperfect peptide separation resulting in simultaneous isolation and fragmentation of more than one peptide, incorrect reproduction of the fragmentation pattern *in silico*, or due to the absence of the peptide in the searched database [35]. The applied scoring model is therefore *per se* susceptible to generating incorrect PSMs and the search algorithm is required to assess the significance of all PSMs. Different methods are implemented in different search engines to statistically analyze the distribution of correct PSMs, which tend to be high scoring, and incorrect PSMs, which have rather low scores. Mascot, for example, estimates the probability that the observed PSM is a random event, *i.e.* not correct. For every experimental spectrum it calculates a probability-based score threshold that depends on the number of candidate spectra, such that every match to this experimental spectrum with a score above the threshold is considered to be significant at the desired probability. Mascot also calculates a more relaxed second score threshold for every PSM, which in some cases may be smaller than the first mentioned score threshold. It also calculates an expectation value that is related to the distance between the score threshold and the score and reflects the number of times one would find this score by chance with the desired probability [38]. From the dependence of the first score threshold on the number of candidate spectra it is immediately evident that the significance of PSMs benefits from high mass accuracy and high mass precision.

Validation of peptide identifications from database searching. Although the scoring models of Mascot, SEQUEST as well as other search engines discriminate between significant and insignificant PSMs within the collection of PSMs, a statistical assessment of the PSM collection, independent of the applied scoring model, is often used to characterize the performance of the search engine. The performance is usually evaluated in the context of binary classification testing and is a balance between specificity, *i.e.* the fraction of incorrect PSMs classified as incorrect, and sensitivity, *i.e.* the fraction of correct PSMs classified as correct (Figure 5A).

An established model to evaluate the error level associated with a collection of PSMs is the false discovery rate (FDR), which is defined as the proportion of incorrect PSMs (false positives) that can be expected within the collection of accepted PSMs (true positives and false positives together). While the latter quantity is the immediate output

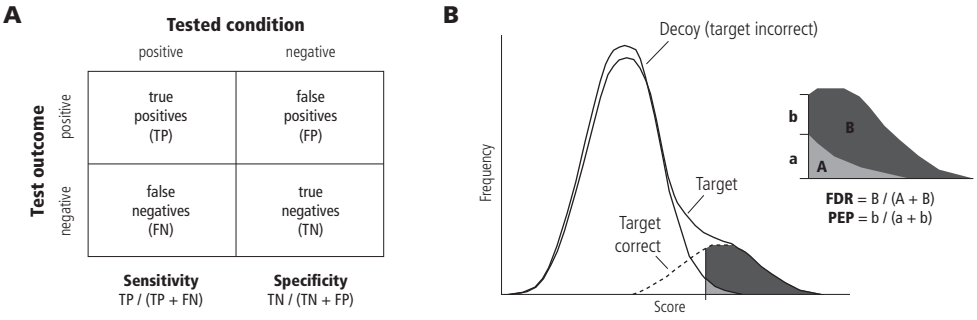


Figure 5. A, binary classification testing. B, FDR and PEP.

of the search engine, one common way to address the number of false positives is by modeling the score distribution or by searching decoy databases [39,40]. The concept of the latter strategy is that all PSMs identified in the decoy search are true negative PSMs and can be used to estimate the false positive PSMs of the target search (Figure 5B). The optimal conception of this target decoy search strategy is still debated [41-43]. A common way to generate the decoy database is to reverse all protein sequences of the target database, thereby conserving the database size and the distribution of peptide lengths, which are two important parameters for the validity of the methods [44]. However, randomized or otherwise built databases may also be advantageous [42]. The target and decoy databases can be concatenated and searched simultaneously, or both databases can be searched separately and the process of concatenation is simulated afterwards [44,45]. Rather than reporting an FDR for a collection of PSMs, usually the PSM score threshold is varied until the FDR reaches the acceptable desired value (Figure 5B). While the FDR is a property of a collection of PSMs, it can also be used to calculate q-values for every PSM, which indicate the minimum FDR threshold at which this specific PSM will be reported. The target decoy search strategy also allows calculating the posterior error probability (PEP) for every PSM, which is the probability that a specific PSM is incorrect [46].

True and false PSMs can also be discriminated by other parameters besides peptide spectrum match quality. This notion has been integrated into validation tools which factor up to 13 parameters into one discriminant score used for PSM classification [39,40,47]. The weight of each component of the discriminant score is determined by evaluating two training sets of PSMs with known classification. In case of [47] this is achieved by a support vector machine algorithm that describes every PSM as a point in a multidimen-

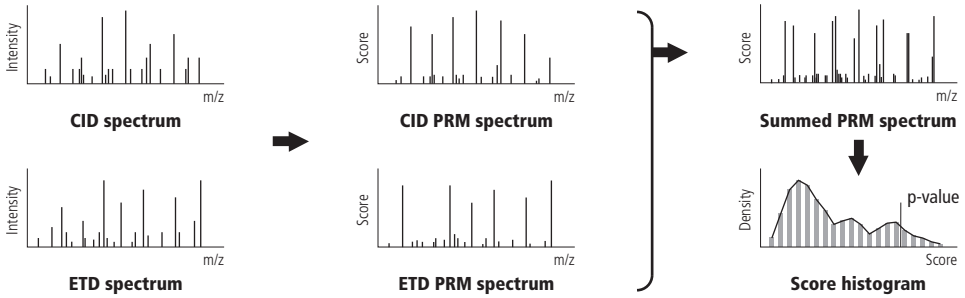


Figure 6. Illustration of the MS-GFDB workflow for analyzing paired CID/ETD spectra using summed PRMs, adapted from [49].

sional space and then calculates the hyperplane separating both true and false PSMs with the greatest distance to both. This method has shown great power in separating the two PSMs distributions and has recently been improved by dynamically calculating the weights of the parameters for every analyzed dataset using the target decoy strategy and by including a larger number of features [41]. The algorithm, called Percolator, can be used as postprocessor for a number of search engines including SEQUEST [41] and Mascot [48]. Up to 30 parameters besides the PSM score are available in Mascot Percolator, covering very diverse features such as precursor mass deviation, percentage of matched fragment ion intensity, number of modified residues, and retention time. The training set to establish the classifier weights of the false PSMs is inferred from the total of all PSMs identified in the decoy search, while the weights for the true PSMs is inferred from a selection of highest scoring PSMs in the target search. The weights are then iteratively varied until the FDR converges. Instead of classifying only the highest-ranking PSM, Percolator can also classify lower-ranking PSMs and re-rank them after classification. As an FDR-based method, Percolator also calculates the q-value and PEP for every PSM [46].

A novel database searching and PSM validation approach is MS-GFDB, which calculates an individual probability for each PSMs instead of deducting it from a global FDR [49]. MS-GFDB integrates and develops various measures from *de novo* peptide sequencing algorithms [50,51], the generating function approach [52], and the FDR-based approaches described above. The scoring model of MS-GFDB is based on assigning probabilities to the individual fragment peaks of an experimental spectrum as described in [50]. It uses offset frequency functions to deduce ion types from a training set of PSMs and to calculate binned intensity and rank mass error scores for every ion type reflecting

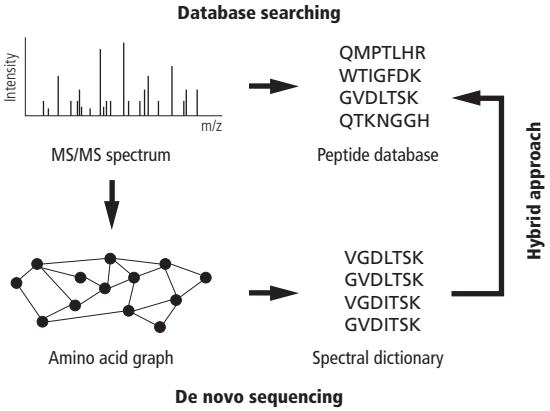


Figure 7. Scheme of *de novo* peptide sequencing and the hybrid approach in combination with database searching. Adapted from [103].

the ratio of the probability that the ion type is specific to the probability that it is noise [50,51]. The training set can, for example, be derived from top scoring PSMs identified by another search algorithm at very low FDR determined by the target decoy strategy [52]. The derived scoring model is applied to convert every experimental spectrum into a prefix residue mass (PRM) spectrum, which is matched with the theoretical spectra in the database. The framework of an exponential generating function, which also considers the frequency of individual amino acids, is used to calculate individual p-values for every PSM which are independent of the database, therefore abolishing the need for the target decoy strategy [49]. As the PRMs are representations of experimental spectra that do not depend on the fragmentation technique or instrument type, they allow combining different spectra of the same peptide, for example CID and ETD spectra, on the PRM spectrum level (Figure 6). MS-GFDB has been shown to be outperform a large number of common search and validation algorithms, including Mascot Percolator, and has been successfully applied to the analysis of Lys-N derived peptides fragmented with ETD [49].

De novo peptide sequencing. An alternative way of peptide sequence analysis is peptide *de novo* sequencing. The approach originates from manual spectra interpretation, *i.e.* direct inference of the peptide sequence from its fragment ion spectrum. The automated approach is to find the best interpretation of a fragment ion spectrum among all possible interpretations (Figure 7). This is usually based on probabilistic scoring models as outlined above for SHERENGA/PepNovo. Peptide *de novo* sequencing has potential for identifying peptides not contained in the searched database. These peptides are often highly post-translationally modified, expressed from genes with single nucleotide polymorphisms, or are derived from spliced variants or specifically cleaved proteins. In

case of peptides with multiple PTMs the search space will become so large that only few PSMs can be confidently identified. On the other hand, peptide *de novo* sequencing is computationally very time-consuming and has a much higher error rate than database searching, in particular for larger peptides.

Validation of sites of post-translational modifications. A frequently occurring situation when performing database searches with posttranslationally modified peptides is that the PTM can reside on more than one amino acid of the peptide. Typically, only very few specific ions, normally derived from cleavages around the potential modification sites, can unambiguously reveal the correct PTM site. The obvious way of resolving this ambiguity in favor of one site is to determine the delta score: Any database search algorithm will likely report positional isomers because they are isobaric and usually have similar fragmentation spectra. The score difference – the delta score – of the highest ranking peptide identification to its second ranking positional isomer is a measure of the probability of the position of the highest ranking peptide. However the sensitivity of this approach can still be improved by applying dedicated scoring models, as shown in [53]. These models usually calculate probability based scores for every potential site using the intensity-weighted site-determining fragment ions. Examples for these algorithms are Ascore [53] and PTM scoring [54], which have been developed specifically for phosphorylated peptides but can in principle be used for any PTM.

2.4 Quantitative proteomics [55-57]

The ability to quantify abundance differences between proteomes is probably the most important feature of contemporary proteomics and is widely used in biological and clinical research. Several different approaches have been developed to obtain quantitative information. They can be categorized into methods that introduce a mass difference between proteomes using stable isotope labeling, and methods that are based on a direct comparison of LC-MS signals.

Stable isotope-based quantification [58]

Mass spectrometry is not an accurate quantitative technique *per se* because its response differs with the analyte. Coupled to LC, the response is also influenced by the composition of the sample at the time of analyte elution. For this reason, the most accurate quantification is achieved when it is based on the comparison of individual peptides

under identical analytical conditions. One way to achieve this is the incorporation of heavy stable isotope atoms into the peptides derived from one proteome and the incorporation of light isotope atoms into the peptides derived from another proteome. The labeled peptides then differ only by mass but not by chemical properties and can be analyzed together to eliminate all analyte response differences. ^{13}C, ^{15}N and ^{18}O are heavy isotopes commonly used for labeling and are compared against the naturally highest abundance isotopes ^{12}C, ^{14}N and ^{16}O. Deuterium is also frequently used but displays subtle chemical differences from hydrogen which result in retention time shifts during LC. Normally, multiple atoms are replaced to result in a mass shift large enough to distinguish the labeled and unlabeled peptide in the mass spectrum.

The label can be introduced at various points of the proteomics workflow (Figure 8). Since every preparative step is associated with a loss of analyte, the most accurate quantification can be expected if the label is introduced metabolically. This is commonly done by substituting all ^{14}N sources with ^{15}N [59,60] or by substituting one or more amino acids with their ^{13}C,^{15}N-labeled counterparts during cell culture [61]. When metabolic labeling is not available, for example in case of human biopsies in biomedical research, chemical labeling of peptides is the method of choice. In this case the naturally occurring isotope is hard to replace, and therefore the peptide pools from the compared proteomes are reacted with a light or heavy labeled reagent. Two principally different strategies can here be distinguished. One strategy, for example dimethylation labeling which is based on the incorporation of ^{13}C,D-labeled methyl groups [62], result in a peptide mass difference just like metabolic labeling. Enzymatic incorporation of ^{18}O during proteolysis also falls into this category [63]. Third, peptides can also be quantified against a heavy labeled synthetic peptide that is added prior to analysis. This approach also allows an absolute quantification of the peptide if the concentration of the synthetic peptide is precisely known. The other principally different strategies comprise methods like isotope-coded affinity tagging (ICAT) [64] and isobaric tagging for relative and absolute quantification (iTRAQ) [65]. Here, the isotopic label together with a balance moiety is reacted with the peptides such that the heavy and light labeled peptides are isobaric and therefore indistinguishable in the mass spectrum. The label dissociates only upon fragmentation, releasing light and heavy reporter ions; quantitative information is thus obtained from the fragmentation spectrum.

Metabolic labeling is very accurate but requires the studied system to be amendable to labeling. In particular ^{15}N-labeling, but also labeling with heavy amino acids was successfully applied to higher organisms like for example fruit flies, rats and mice. How-

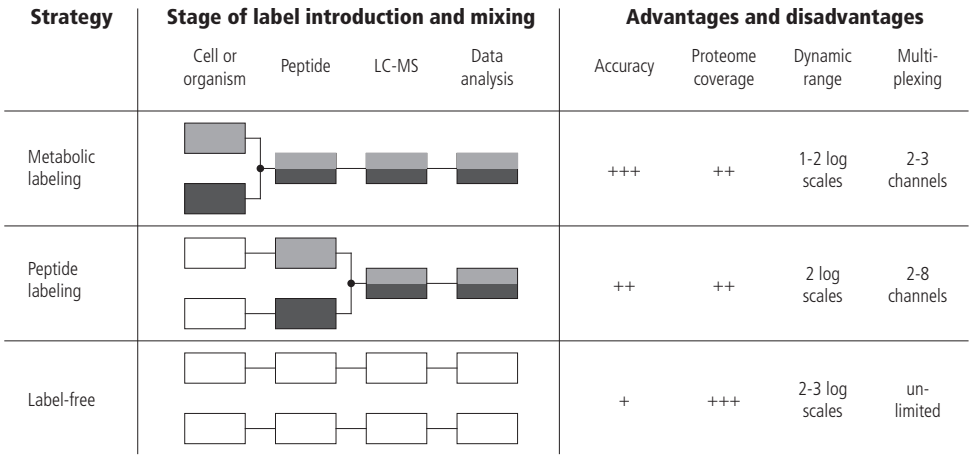


Figure 8. Isotopic labeling strategies in quantitative proteomics. The flow chart illustrates the point of label introduction (white, unlabeled; light grey, light isotope label; dark grey, heavy isotope label) and the point of sample mixing (black square). Adapted from [57].

ever, the culturing conditions for isolated cell lines grown *in vitro* may not permit any metabolic labeling. Moreover, metabolic labeling is limited in terms of multiplexing capabilities, which is not only due to limited possibilities of label introduction. Each label doubles the complexity of the mass spectrum, which results in decreased peptide identification rates. Chemical labeling, which can be applied to virtually any peptide sample, also suffers from this limitation when the quantification is MS-based. When it is based on the fragment ion spectrum, multiplexing can be increased up to 8-fold but sensitivity may then be a limitation. These and other advantages and disadvantages are summarized in Figure 8.

Label-free quantification

In this strategy, the compared proteomes are subjected to separate LC-MS analyses while keeping the analytical conditions as similar as possible. A quantitative comparison is established based on peptide intensities between the different analyses. One of the biggest advantages of label-free quantification is the potentially unlimited multiplexing capability and the depth of quantification due to the fact that no increase in sample complexity is required. The major difficulty is the variation of chromatographic profiles because the peaks intensity comparison is guided by the peak retention time. The main disadvantage is the low accuracy which is introduced by the inherent variability of the analytical workflow. Another strategy to quantify protein abundances between differ-

ent runs is spectral counting, which correlates the number of identified fragmentation spectra assigned to a protein with the abundance of the protein. The approach is suitable for spotting large quantitative differences of highly abundant proteins which give rise to a large number of identified spectra. Due to big uncertainties, it is not suitable to estimate small quantitative changes of highly abundant proteins. It is also very inaccurate for low abundant proteins with low spectral counts and is therefore very sensitive to fluctuations in LC-MS performance.

3 The transcription factors SAGA and TFIID

3.1 Structure of genetic information [66,67]

The variety of cellular proteomes of an identical genome is to a great extent based on differential gene expression. In eukaryotes, heritable material is present as chromatin, a complex mainly of DNA and histones. The information is stored as DNA in units of genes, regions of DNA that govern the production of proteins and functional RNA. Genes are structured into the coding sequence, which contains the information about the amino acid or base sequence of the functional product, and a variety of non-coding elements with essential or accessory function for gene expression. The essential key element for the regulation of DNA transcription is the gene promoter which determines if the gene is transcribed and to what extent [68]. The promoter of eukaryotic DNA is commonly structured into a core region, a proximal region, a distal region, and enhancer or silencer elements. The core promoter is localized around the transcriptional start site (TSS) and serves as an assembly platform for RNA polymerase and a number of general transcription factors (GTFs) that are relatively unspecific for a particular gene and can be found on many promoters. The proximal and the distal promoter are localized upstream of the core promoter and contain response elements for gene-specific transcription factors. The enhancer and silencer elements also bind gene-specific transcription factors but are not fixed in their localization within the genome. While the diversity of response elements of the proximal and distal promoter region reflects the specificity of transcription factors for small sets of genes, the relatively high degree of conservation of core promoter elements throughout the genome reflects the common function of GTFs for the initiation of transcription.

The association of DNA with histones in eukaryotes provides another level of structural organization of the genetic material. Chromatin is based on a repeating structural unit

of one octamer of H2A, H2B, H3 and H4 which is enfolded by around 145 bp of DNA [69]. This arrangement, together with additional proteins like H1, is responsible for the compaction of DNA into fibers, higher-ordered structures and ultimately the chromosome [70]. The association with nucleosomes profoundly affects all DNA related processes including gene transcription. One regulatory effect of chromatin is that the condensed nucleosomal DNA is less accessible to the protein machineries involved in these processes, for example RNA polymerase. Specialized protein machineries, chromatin remodeling complexes, can alter chromatin structure by moving or ejecting nucleosomes or by changing the nucleosome composition through the incorporation of histone isoforms [71]. The other regulatory effect is related to the extensive post-translational modifications of nucleosomal histones [72]. The detected abundance and combinatorial variety of histone PTMs is extremely large with acetylation and methylation of lysine residues being among the most abundant PTMs. While histone acetylation has long been linked to the compaction grade of chromatin and with this its transcriptional activity, it emerged only in the past decade that the distribution of nucleosomes with distinct histone modification patterns is correlated with the functional state of the bound DNA segment on a genome-wide level even down to single nucleosomes [73]. These chromatin marks are specifically set and deleted by chromatin modifying complexes or chromatin modifying modules of related complexes. In general, chromatin modifications are mechanistically linked to the functional state of the DNA by serving as binding sites for proteins with the appropriate recognition modules so that the presence or absence of specific PTMs can specifically induce or hinder binding of protein machineries [74]. These interactions play an important role in the regulation of gene transcription [75].

3.2 Factors of transcription initiation [76,77]

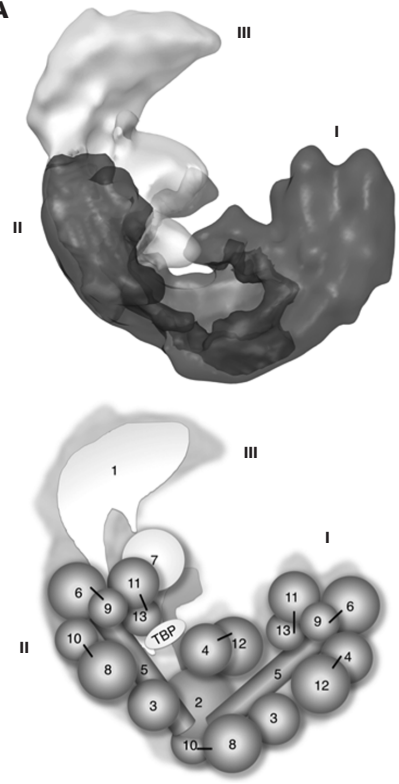
The transcription of an inactive, mRNA-coding gene requires a number of factors besides RNA polymerase II (Pol II). In general, transcription starts with the binding of a transcriptional activator to its response element upstream of the promoter. The bound activator results in the recruitment of transcriptional coactivators that alter the chromatin structure surrounding the promotor and facilitate binding of GTFs to the core promotor. The GTFs are required for recognition and preparation of the promoter for Pol II binding. Together with Pol II they assemble into the pre-initiation complex (PIC) from which Pol II can proceed into the elongation phase. GTFs comprise the class of basal transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH which are sufficient to

establish a PIC competent for basal transcription on nucleosome-free DNA *in vitro*, and a number of transcriptional cofactors required for regulated transcription, for example TATA-binding protein (TBP) associated factors (TAFs), Mediator, SAGA, SLIK, BTAF1 and NC2. Within the PIC, basal transcription factors execute rather fundamental steps of transcription initiation like TSS selection by recognition of and binding to core promoter elements and melting of the double-stranded promoter DNA. One of the most prominent GTFs is TBP which is a stable component of TFIID but also associates with other GTFs like SAGA, BTAF1 and NC2 [78]. TBP nucleates PIC formation at the core promoter by selective interaction with the TATA box, a core promoter element found in around 10-20% of all Pol II-dependent promoters. On promoters without a TATA element, TBP binding is mediated by interactions of TFIID subunits with other core promoter elements. TBP binding to the TATA box induces a drastic conformational change of the promoter DNA which has been linked to the differences in TBP turnover between promoter with and without a TATA element [78]. The basic functionality of TBP is further underlined by the fact that TBP, as subunit of the SL1 and TFIIIB complexes, is also involved the transcription initiation of RNA polymerase I- and III- dependent genes, unlike other GTFs related to Pol II-dependent genes. Transcriptional cofactors often function by bridging the interaction between the basal transcription factors and proximally bound activators and facilitate or repress PIC formation by interaction with basal transcription factors. Other transcriptional coactivators can post-translationally modify nucleosomes around the promoter region and thereby influence PIC formation. While some GTFs like TFIIB or BTAF1 are single polypeptides, others like TFIIH, TFIID or SAGA can be purified as large heteromers that unite a variety of enzymatic activities as well as binding modules for protein-protein and protein-DNA interactions. Most GTFs are evolutionarily conserved throughout the eukaryotic kingdom.

TFIID

TFIID is a heteromeric complex of TBP and 13-14 stably associated TAFs which are widely conserved within the eukaryotic kingdom. The structures of TFIID from *Saccharomyces cerevisiae* and human cultured cells have been studied by electron microscopy (EM) and show remarkable similarity [79,80]. The shape of TFIID is reminiscent of a clamp with three lobes joined at a central intersection (Figure 9A) [81]. An important structural feature of many TAFs is the histone fold domain (HFD), a primary sequences stretch with homology to the helix-turn-helix domain of core histones [82]. The HFD enables the heterodimerization of TAFs into the pairs TAF4-TAF12, TAF6-TAF9, TAF8-TAF10, TAF3-

A



B

Yeast	Human
Gcn5	GCN5/PCAF
Ada1	ADA1/STAF42
Ada2	ADA2b
Ada3	ADA3
Spt8	?
Spt20	SPT20/FAM48A/P38IP
Spt7	SPT7/STAF65γ
Spt3	SPT3
TAF5	TAF5L
TAF6	TAF6L
TAF9	TAF9/TAF9b
TAF10	TAF10
TAF12	TAF12
Tra1	TRRAP
Sgf29	SGF29
Ubp8	USP22
Sgf11	ATXN7L3S
Sus1	ENY2
Sgf73	ATXN7

Figure 9. A, structure of yeast TFIID with the three lobes indicated I-III (upper panel) and with subunit localization indicated (lower panel, TAF heterodimers indicated). Reproduced with permission from [81]. B, subunit composition of Saccharomyces cerevisiae and human SAGA. From human cells different complexes with similar composition can be purified which may be collectively termed SAGA. Adapted from [98].

TAF10 and TAF11-TAF13. Together with TAF5, which is probably present as homodimer associated via another domain, these TAFs are present in at least two copies within TFIID and localize to two of the lobes [83]. TAF1, TAF2, TAF7 and TBP, which are present as single copy, localize to the third lobe and the central intersection with the other lobes [84]. TAF4, TAF5 and TAF10 are critical for the stability of TFIID, while TAF2 is relatively loosely associated with TFIID. TFIID displays structural heterogeneity [85] which is influenced by TAF2 [86] and the replacement of TAFs with TAF isoforms [87]. TFIID contains a number of recognition capabilities for the core promoter and other proteins. Besides TATA box recognition by TBP, TFIID can also bind to the core promoter elements DPE (downstream promoter element) and Inr (initiator) through TAF6-TAF9 and TAF1-TAF2, respectively. It can also selectively bind to nucleosomal H3 trimethylated at lysine at po-

sition 4 (H3K4me3), a chromatin modification found at actively transcribed promoters [75], by the plant homeodomain (PHD) finger of TAF3 [88]. The tandem bromodomain (BD) [89,90] of TAF1 confers TFIID the ability to bind H4K5acK12ac and H4K8acK16ac [91], marks that are also associated with active promoters [75,92], and likely H3K9acK-14ac in which case a cooperative effect with the TAF3 PHD finger might exist. TFIID also displays coactivator functionality as it can directly bind transcriptional activators bound to upstream promoter elements.

SAGA

SAGA (Spt-Ada-Gcn5 histone acetyltransferase) is a large heteromeric complex in *Saccharomyces cerevisiae*, comprising around 20 subunits [93]. The subunit composition and the subunits themselves are less conserved in higher eukaryotes that TFIID; complexes with similar subunit composition and functionality in metazoans are TFTC and STAGA [94] which may be collectively called SAGA (Figure 9B). The structure of SAGA has been approached by immunolabeling EM and it appears that the complex is organized into modules harboring different functions [95]. The general function of SAGA is that of a transcriptional coactivator. It is recruited to gene promoters by binding to activators through Tra1 and acetylates nucleosomal H3 around the promoter, thereby marking the promoter as active [92]. This histone acetyltransferase activity with Gcn5 as the catalytic subunit is the main activity of SAGA. Gcn5 maps to a central module of the SAGA structure and co-localizes with subunits essential for structural integrity [96]. Next to SAGA-specific subunits like Spt7 and Ada1, the stability is conferred by TAF5, TAF6, TAF9, TAF10 and TAF12. In human TFTC and STAGA, TAF5 and TAF6 are replaced by the homologous proteins TAF5L and TAF6L [94]. The HFD has a structural function within SAGA as well and is present in the heterodimers TAF6-TAF9, TAF10-Spt7 and TAF12-Ada1 [82]. The BD is another prominent domain within SAGA and present within Gcn5 and Spt7. While the BD of Gcn5 confers SAGA the ability to bind acetylated nucleosomes, the function of the BD of Spt7 is unclear [97]. The second enzymatic functionality of SAGA is deubiquitinylation with Ubp8 as the catalytic core [98]. This activity is also related to the coactivator function, as it targets ubiquitinylated nucleosomal H2A and H2B which repress transcription at the promoter, *i.e.* the removal of the repressive marks activates the gene [99]. Ubp8 forms a stable module with Sgf11 and Sus1, which is stably associated with SAGA through Sgf73. SAGA also interacts with TBP via Spt3 and Spt8 [100,101], which localize together with the structurally essential subunit Spt20, but the interaction with TBP is only transient.

4 References

[1] Kebarle, P., *et al., Mass Spectrom Rev* **2009**, *28*, 898

[2] Cech, N. B., *et al., Mass Spectrom Rev* **2001**, *20*, 362

[3] Cole, R. B., *J Mass Spectrom* **2000**, *35*, 763

[4] Dole, M., *et al., J Chem Phys* **1968**, *49*, 2240

[5] Yamashita, M., *et al., J Phys Chem* **1984**, *88*, 4451

[6] Fenn, J. B., *et al., Science* **1989**, *246*, 64

[7] Whitehouse, C. M., *et al., Anal Chem* **1985**, *57*, 675

[8] Makarov, A., *et al., J Chromatogr A* **2010**, *1217*, 3938

[9] Cech, N. B., *et al., Anal Chem* **2000**, *72*, 2717

[10] Cech, N. B., *et al., Anal Chem* **2001**, *73*, 4632

[11] Wilm, M., *et al., Anal Chem* **1996**, *68*, 1

[12] Gibson, G. T. T., *et al., Mass Spectrom Rev* **2009**, *28*, 918

[13] Covey, T. R., *et al., Mass Spectrom Rev* **2009**, *28*, 870

[14] Hoffmann, E. d., *et al. Mass spectrometry: principles and applications*; 3rd edition; John Wiley & Sons, 2007

[15] March, R. E., *Mass Spectrom Rev* **2009**, *28*, 961

[16] Hartmer, R., *et al., Int J Mass Spectrom* **2008**, *276*, 82

[17] Douglas, D. J., *Mass Spectrom Rev* **2009**, *28*, 937

[18] Douglas, D. J., *et al., Mass Spectrom Rev* **2005**, *24*, 1

[19] Hager, J. W., *Rapid Comm Mass Spectrom* **2002**, *16*, 512

[20] Le Blanc, J. C. Y., *et al., Proteomics* **2003**, *3*, 859

[21] Schwartz, J. C., *et al., J Am Soc Mass Spectrom* **2002**, *13*, 659

[22] Syka, J. E. P., *et al., P Natl Acad Sci USA* **2004**, *101*, 9528

[23] Xia, Y., *et al., Anal Chem* **2008**, *80*, 1111

[24] Swaney, D. L., *et al., Anal Chem* **2007**, *79*, 477

[25] Perry, R. H., *et al., Mass Spectrom Rev* **2008**, *27*, 661

[26] Marshall, A. G., *et al., Int J Mass Spectrom* **2002**, *215*, 59

[27] Makarov, A., *et al., Anal Chem* **2006**, *78*, 2113

[28] Olsen, J. V., *et al., Nat Meth* **2007**, *4*, 709

[29] Olsen, J. V., *et al., Mol Cell Proteom* **2009**, *8*, 2759

[30] Olsen, J. V., *et al., Mol Cell Proteom* **2004**, *3*, 608

[31] Taouatas, N., *et al., Nat Meth* **2008**, *5*, 405

[32] Paizs, B., *et al., Mass Spectrom Rev* **2005**, *24*, 508

[33] Mikesh, L. M., *et al., BBA Proteins Proteom* **2006**, *1764*, 1811
[34] Wiesner, J., *et al., Proteomics* **2008**, *8*, 4466
[35] Nesvizhskii, A. I., *et al., Nat Meth* **2007**, *4*, 787
[36] Perkins, D. N., *et al., Electrophoresis* **1999**, *20*, 3551
[37] Eng, J. K., *et al., J Am Soc Mass Spectrom* **1994**, *5*, 976
[38] Brosch, M., *et al., Mol Cell Proteom* **2008**, *7*, 962
[39] Keller, A., *et al., Anal Chem* **2002**, *74*, 5383
[40] Nesvizhskii, A. I., *et al., Anal Chem* **2003**, *75*, 4646
[41] Kall, L., *et al., Nat Meth* **2007**, *4*, 923
[42] Choi, H., *et al., J Proteom Res* **2008**, *7*, 47
[43] Fitzgibbon, M., *et al., J Proteom Res* **2008**, *7*, 35
[44] Elias, J. E., *et al., Nat Meth* **2007**, *4*, 207
[45] Kall, L., *et al., J Proteom Res* **2008**, *7*, 29
[46] Kall, L., *et al., J Proteom Res* **2008**, *7*, 40
[47] Anderson, D. C., *et al., J Proteom Res* **2003**, *2*, 137
[48] Brosch, M., *et al., J Proteom Res* **2009**, *8*, 3176
[49] Kim, S., *et al., Mol Cell Proteom* **2010**, *9*, 2840
[50] Dancik, V., *et al., J Comput Biol* **1999**, *6*, 327
[51] Frank, A., *et al., Anal Chem* **2005**, *77*, 964
[52] Kim, S., *et al., J Proteom Res* **2008**, *7*, 3354
[53] Beausoleil, S. A., *et al., Nature Biotechnol* **2006**, *24*, 1285
[54] Mortensen, P., *et al., J Proteom Res* **2009**, *9*, 393
[55] Heck, A. J. R., *et al., Expert Rev Proteom* **2004**, *1*, 317
[56] Ong, S. E., *et al., Nat Chem Biol* **2005**, *1*, 252
[57] Bantscheff, M., *et al., Anal Bioanal Chem* **2007**, *389*, 1017
[58] Gouw, J. W., *et al., Mol Cell Proteom* **2010**, *9*, 11
[59] Wu, C. C., *et al., Anal Chem* **2004**, *76*, 4951
[60] Krijgsveld, J., *et al., Nature Biotechnol* **2003**, *21*, 927
[61] Ong, S. E., *et al., Mol Cell Proteom* **2002**, *1*, 376
[62] Boersema, P., *et al., Mol Cell Proteom* **2009**, S47
[63] Miyagi, M., *et al., Mass Spectrom Rev* **2007**, *26*, 121
[64] Gygi, S. P., *et al., Nature Biotechnol* **1999**, *17*, 994
[65] Ross, P. L., *et al., Mol Cell Proteom* **2004**, *3*, 1154
[66] Campos, E. I., *et al., Annu Rev Genet* **2009**, *43*, 559
[67] Rando, O. J., *et al., Annu Rev Biochem* **2009**, *78*, 245
[68] Georges, A. B., *et al., Faseb J* **2010**, *24*, 346

[69] Khorasanizadeh, S., *Cell* **2004**, *116*, 259
[70] Bassett, A.*, et al., Curr Opin Genet Dev* **2009**, *19*, 159
[71] Clapier, C. R.*, et al., Annu Rev Biochem* **2009**, *78*, 273
[72] Kouzarides, T., *Cell* **2007**, *128*, 693
[73] Berger, S. L., *Nature* **2007**, *447*, 407
[74] Taverna, S. D.*, et al., Nat Struct Mol Biol* **2007**, *14*, 1025
[75] Li, B.*, et al., Cell* **2007**, *128*, 707
[76] Thomas, M. C.*, et al., Crit Rev Biochem Mol* **2006**, *41*, 105
[77] Sikorski, T. W.*, et al., Curr Opin Cell Biol* **2009**, *21*, 344
[78] Tora, L.*, et al., Trends Biochem Sci* **2010**, *35*, 309
[79] Andel, F.*, et al., Science* **1999**, *286*, 2153
[80] Brand, M.*, et al., Science* **1999**, *286*, 2151
[81] Cler, E.*, et al., Cell Mol Life Sci* **2009**, *66*, 2123
[82] Gangloff, Y. G.*, et al., Trends Biochem Sci* **2001**, *26*, 250
[83] Leurent, C.*, et al., Embo J* **2002**, *21*, 3424
[84] Leurent, C.*, et al., Embo J* **2004**, *23*, 719
[85] Grob, P.*, et al., Structure* **2006**, *14*, 511
[86] Papai, G.*, et al., Structure* **2009**, *17*, 363
[87] Liu, W. L.*, et al., Mol Cell* **2008**, *29*, 81
[88] Vermeulen, M.*, et al., Cell* **2007**, *131*, 58
[89] Yang, X. J., *Bioessays* **2004**, *26*, 1076
[90] Mujtaba, S.*, et al., Oncogene* **2007**, *26*, 5521
[91] Jacobson, R. H.*, et al., Science* **2000**, *288*, 1422
[92] Shahbazian, M. D.*, et al., Annu Rev Biochem* **2007**, *76*, 75
[93] Baker, S. P.*, et al., Oncogene* **2007**, *26*, 5329
[94] Nagy, Z.*, et al., Oncogene* **2007**, *26*, 5341
[95] Wu, P. Y. J.*, et al., Mol Cell* **2004**, *15*, 199
[96] Timmers, H. T. M.*, et al., Trends Biochem Sci* **2005**, *30*, 7
[97] Hassan, A. H.*, et al., Cell* **2002**, *111*, 369
[98] Rodriguez-Navarro, S., *Embo Rep* **2009**, *10*, 843
[99] Pijnappel, W.*, et al., Mol Cell* **2008**, *29*, 152
[100] Sermwittayawong, D.*, et al., Embo J* **2006**, *25*, 3791
[101] Mohibullah, N.*, et al., Genes Dev* **2008**, *22*, 2994
[102] Yates, J. R., *et. al.*, *Annu Rev Biomed Eng* **2009**, 11, 49
[103] Kim, S., *et al., Mol Cell Proteom* **2009**, 8, 53

Chapter 2

Targeted large-scale analysis of protein acetylation

Nikolai Mischerikow[1,2] and Albert J. R. Heck[1,2]

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

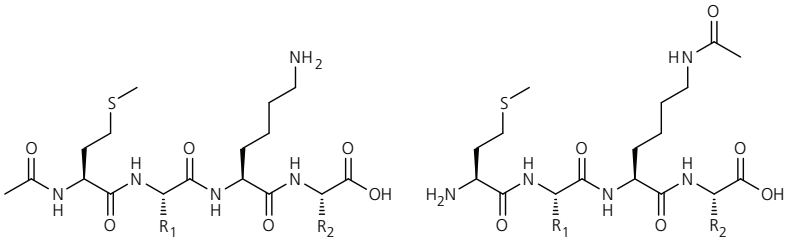
[2] Netherlands Proteomics Centre, Padualaan 8, 3584 CH Utrecht, The Netherlands

1 Summary

Protein modifications are biologically important events that may be studied by mass spectrometry- (MS) based high-throughput proteome analyses. In recent years, several new technologies have emerged that have widened and deepened the targeted analysis of one important, albeit functionally ill-defined modification, namely protein acetylation. This modification can take place both co- and post-translationally by the transfer of acetyl groups catalyzed by acetyltransferases. The acetyl group can modify either the α-amino group at the N-terminus, so-called N-terminal acetylation, or the ε-amino group on the side chain of lysine residues. Here, we review several emerging targeted technologies to chart both N-terminal acetylation as well as acetylation of the lysine side chain, on a proteome-wide scale, highlighting in particular studies that have expanded the biological knowledge of the appearance and function of these common but functionally still less investigated co- and post-translational modifications.

2 Introduction

Proteome diversity extends far beyond that of the genome and it has been estimated that proteomes may be 2-3 orders of magnitude more complex (>1,000,000 molecular species of proteins) than the encoding genomes would predict [1]. There are two major mechanisms for expanding the coding capacity of genomes to generate diversity in the corresponding proteomes: at the transcriptional level, for example by mRNA splicing, and by covalent modification of proteins. To focus on the latter, proteins are decorated, often reversibly, by a large variety of co- and post-translational modifications (collectively abbreviated in this chapter as PTMs), which all may influence their functional behavior. More than 300 different modifications have been described [1-5]. Although probably most abundant and diverse in eukaryotes, they occur in all kingdoms of life [6,7]. PTMs can play a role in cell signaling, protein-protein interactions, protein stability, as well as activation/deactivation of enzymatic activity. In recent years, MS -based proteomics has made huge contributions to the analysis of protein modifications, as this technique is ideally suited not only to identify PTMs, but also to pinpoint the exact site of the modification, and to differentially and sometimes even absolutely quantify the modification. As co- and post-translational modifications often occur at a low stoichiometry and therefore need to be detected in a massive background of unmodified proteins or proteolytic peptides, specific approaches targeting a particular PTM are of-



N_α-acetylated **Met-X-Lys-X** N_ε-acetylated **Met-X-Lys-X**

Figure 1.

ten essential to increase their coverage by proteomics. Great advances have been made in phosphoproteomics [8,9], glycomics [10-12], and also several promising approaches are surfacing to tackle less abundant or more complex protein modifications such as ubiquitination [13], sumoylation [14] and palmitoylation [15,16].

In this review we will focus on recent advances in mass spectrometric detection of another protein modification, namely protein acetylation. This modification can occur both co- and post-translationally by the transfer of acetyl groups from acetyl-CoA catalyzed by acetyltransferases [17]. The acetyl group can become attached to either the α-amino group at the protein N-terminus or the ε-amino groups of lysine residues (Figure 1). Protein N-terminal acetylation is thought to be primarily a co-translational process occurring during protein synthesis when the N-terminus of a newly synthesized protein extrudes from the ribosomal exit channel [18-22]. Acetylation of the ε-amino group of lysine is a reversible process and occurs post-translationally. In both cases, the positive charge of the amino group is eliminated. In the last couple of years several influential new approaches using MS-based proteomics as readout have been introduced and subsequently important studies have appeared through which our knowledge of the N-terminal and lysine acetylome has significantly expanded beyond that described in related reviews [5,17,23,24]. In this review we will first describe background information on the occurrence and function of protein N-terminal and lysine acetylation and then give a description of several important approaches and methods targeting these modifications, highlighting a few case studies that have expanded our biological knowledge on the appearance and function of these modifications.

2.1 Acetylation of protein N-termini

N-terminal acetylation is one of the most common and abundant protein modifications in eukaryotes, estimated to occur in about 85% of mammalian proteins and 60%

of yeast proteins and significantly less frequent in prokaryotes [23,25-27]. N-terminal acetylation predominantly takes place at the N-terminal methionine residue of nascent polypeptide chains, when the first 20 to 50 amino acid residues extrude from the ribosome [18-22]. Alternatively, the N-terminal methionine can be cleaved off by aminopeptidases followed by acetylation of the then exposed second residue. The tendency for methionine aminopeptidases to cleave the methionine residue is higher if the second amino acid is less bulky [23,28]. Therefore, next to methionine, amino acid residues such as serine, alanine and threonine account for the majority of experimentally observed N-terminal acetylated residues [23,28-30]. N-terminal methionine excision also is a co-translational process which involves two types of aminopeptidases, namely methionine aminopeptidase type I (MetAP1) and type II (MetAP2) [31-33]. MetAP1 is further divided in three subclasses A, B and C of which subclasses A and C are present in prokaryotes, while subclass B is present in eukaryotes [34,35]. These two types of methionine aminopeptidases have clearly diverged early in evolution, with MetAP1 and MetAP2 of eukaryotes being more closely related to eubacterial and archaeal homologues than to each other [36].

Independent of whether protein N-terminal methionine excision occurs, N-terminal protein acetylation is performed by one of several specific N-terminal acetyltransferase machineries (Figure 2) [23,25]. An important approach to study sequence specificity of the different N-terminal acetyltransferase machineries is the deletion of N-acetyltransferase genes, which can be easily accomplished in *Saccharomyces cerevisiae*. From these studies, substrate specificities of Naa10, Naa20 and Naa30, the catalytic subunits of the three yeast N-terminal acetyltransferases (NATs) termed NatA, NatB and NatC have been deduced by observing specifically diminished protein N-terminal acetylation in the corresponding deletion strains [29,37]. In addition to the catalytic subunit, each of these N-acetyltransferases contains one or more auxiliary subunits. The Nat complexes are known to associate with the ribosome and the polysome [21,22]. Recently, two other NATs have been described in yeast, termed NatD and NatE, with Naa40 and Naa50 as catalytic subunits. NatD is responsible for N-terminal acetylation of histones H4 and H2A [38] and the substrate specificity of NatE is still largely unknown; it was found to associate with NatA [21,39]. NatA is probably the best characterized N-terminal acetyltransferase complex, consisting of the catalytic subunit Naa10 in complex with Naa15 [40]. Both subunits are equally essential for acetyltransferase activity as yeast strains lacking either one of these subunits display the same phenotype, indicating that these proteins function together to catalyze N-terminal acetylation of a subset of proteins

	NatA	NatB	NatC	NatD	NatE
Catalytic subunit	Naa10	Naa20	Naa30	Naa40	Naa50
Auxiliary subunit	Naa15	Naa25	Naa35 Naa38		
Specificity	Ala Gly Ser Thr	Met-Asp Met-Glu Met-Asn	Met-Phe Met-Leu Met-Trp Met-Ile	H2A, H4, H2A.Z	#

Figure 2. N-terminal acetyltransferase machineries. #, the substrate specificity of Naa50 was recently studied with recombinant human Naa50, and it was found that Naa50 preferentially acetylates peptides with an N-terminus similar to that of human heterogeneous nuclear ribonucleoprotein F (hnRNP F). Adapted from [37].

[39]. The NATs possess different substrate specificities. NatA tends to acetylate proteins with serine, alanine, glycine or threonine at the N-terminus, NatB acts on proteins with methionine followed by glutamic acid, aspartic acid, or asparagine at the N-terminus, and NatC preferably acts on methionine followed by isoleucine, leucine, tryptophane, and phenylalanine at the N-terminus [22,37,41,42].

Higher eukaryotes have NAT genes which are homologous to the yeast genes, indicating that large parts of the different N-terminal acetylation machineries are conserved [26,43,44]. Three NAT complexes have been described in human, termed hNatA [45], hNatB [46] and hNatC [47]. Human homologues of NatD and NatE have been predicted; however, these complexes have still not been well characterized. Not surprisingly N-terminal acetylation patterns in humans are similar to those in yeast; however, it seems that human proteins are much more prone to N-terminal acetylation [29,41]. By knockdown studies the phenotypes, substrate specificities and expression patterns of the three human NATs have been described and appear to be similar to those in yeast. However, still little is known about the function and regulation of the human N-terminal acetylation system. Arnesen *et al.* recently identified HYPK as a novel interactor of the human NatA [48]. HYPK, Naa10 and Naa15 were associated with polysomes indicating a function of NatA-associated HYPK during protein translation. Furthermore, they demonstrated that HYPK is required for N-terminal acetylation of the NatA substrate PCNP, indicating that the physical association of this novel interactor HYPK and NatA is of functional relevance for N-terminal acetylation.

Although not generally accepted, it has been argued for already a long time that protein N-terminal acetylation would protect proteins from premature degradation [49-51], for instance by N-terminal ubiquitinylation [52,53]. In contrast, Varshavsky *et al.* [54] recently proposed that N-terminal acetylated Met residues may act as a degrada-

tion signal, targeted by the Doa10 ubiquitin ligase, and showed that Doa10 can also act on other N-terminal acetylated residues like alanine and serine. It remains to be seen whether these results are generic or specific for *Saccharomyces cerevisiae*. In general, the interplay of N-terminal acetylation and protein turnover affects cellular processes in very different ways and no convergent picture has emerged so far [41].

2.2 Acetylation of lysine residues

Acetylation of the ε-amino group of lysine is a reversible process and occurs post-translationally. It was originally discovered as a PTM of histones, were it very frequently occurring and part of a balanced network of chromatin modifications that regulate transcription-related processes [55-57]. Apart from histones, a large number of non-histone proteins have been found to be lysine-acetylated as well. Transcription-related factors are probably among the most extensively studied proteins regulated by lysine acetylation; however, no convergent function of lysine acetylation has been found on the cellular and molecular level [58]. Nonetheless, newer views suggest that the occurrence of multiple acetylated residues within short sequence stretches or its co-occurrence with other PTMs on the same or nearby residues might reflect functional patterns, at least on the molecular level [59].

Unlike N-terminal acetylation, lysine acetylation is reversible *in vivo* in the sense that both lysine acetyltransferases (KATs, or HATs, from histone acetyltransferases [60]) and lysine deacetylases (KDACs or HDACs) exists, which determine the acetylation state of lysine residues in a highly dynamic fashion. The importance of lysine acetylation for the regulation of intracellular processes is reflected by the existence of a relatively large number of evolutionarily well-conserved KATs and KDACs in eukaryotes [61] as well as by the existence of a protein domain which specifically binds acetylated lysine residues, the bromodomain [62,63].

Most nuclear KATs can be grouped into the MYST, GNAT and p300/CBP families based on sequence homology of their catalytic domain, while some KATs do not display sufficient homology to either of these classes [61]. Although different in primary sequence, the catalytic domains of MYST, GNAT and p300/CBP KATs are structurally similar and effectively catalyze the acetyl group transfer to the free ε-amino group by utilization of acetyl-CoA via a sequential catalytic mechanism involving lysine ε-amino group deprotonation followed by a nucleophilic attack on the acetyl-CoA cofactor; however, the mechanistic details are disputed for p300/CBP family members [64-66]. KDACs, which

catalyze the reverse reaction, are more diverse in mammals than in lower eukaryotes. They are commonly grouped into the classes I, II, III, and IV [61,67,68]. KDACs of classes I and II are Zn^{2+}-dependent enzymes which hydrolyze the acetyl-lysine amide bond yielding acetate and the deacetylated substrate [69]. Class III KDACs, also called sirtuins, utilize nicotinamide adenine dinucleotide (NAD^+) as cofactor for the deacetylation reaction, yielding acetyl-ADP-ribose, nicotinamide, and the deacetylated substrate [70]. By this, sirtuins couple protein deacetylation to NAD^+ hydrolysis. Alternatively, sirtuins can transfer the ADP-ribosyl moiety of NAD^+ to the substrate, yielding the mono-ADP-ribosylated protein and nicotinamide [71]. Many KATs, and to a lesser extent KDACs, exert their function as part of multi-subunit protein complexes [72,73]. The associated proteins alter the substrate specificity and modulate the specific activity compared to the free KAT or KDAC. Another way of regulating KAT/KDAC activity is the localization of the complexes to their site of action by the numerous domains mediating protein-protein and protein-chromatin interactions, provided both by the KAT/KDAC and the associated complex subunits [74,75]. This modulation is of particular importance for polymeric substrates like chromatin, where lysine acetylation can be highly specific on the level of regions of chromosomes, genes, gene regions, nucleosomes and lysine residues within the nucleosomes [56,76].

Lysine residues are not exclusively subject to acetylation but can also be modified by other PTMs such as methylation, ubiquitination and sumoylation. Therefore, the acetylation state of a lysine residue can also be influenced by the responsible enzymes which generate these PTMs. Moreover, the post-translational modification of amino acids adjacent to a lysine residue, such as phosphorylation of serine or threonine, can regulate the acetylation state of the lysine residue. These interplays between different PTMs have mostly been studied for histones [77-79]; however, they also exist on non-histone proteins where they have shown to display important functional roles [68,80].

3 Methods to probe N-terminal acetylation

3.1 2D gel electrophoresis

The earliest proteomics-based approaches to study protein N-terminal acetylation were performed by using 2D gel electrophoresis, often in combination with the use of mutants lacking acetyltransferases, making use of the fact that N-terminal acetylation often affects the electrophoretic mobility of the intact protein on the gel [81,82]. A number of

studies, especially in *Saccharomyces cerevisiae*, have made use of genetic deletions in subunits of various *NAT* genes to investigate the N-terminal acetylation status of abundant proteins or those involved in protein complexes such as the ribosome [42,83] or the proteasome [84,85]. Using three N-acetyltransferase-deficient yeast strains, Polevoda *et al.* most extensively explored the substrate specificities of NatA, NatB and NatC by 2D gel electrophoresis [82]. Although this method has provided clear information about the N-terminal acetylation status of particular proteins, it is relatively cumbersome and generates data at relatively low throughput. Detailed information about protein N-terminal acetylation obtained from many of such 2D gel electrophoresis-based studies have been nicely compiled by Polevoda *et al.* in a series of reviews that largely cover the knowledge of protein N-terminal acetylation up to a few years ago [23,29,44].

3.2 Non-targeted approaches

In recent years, peptide-centric approaches have taken center stage in high-throughput proteomics. Proteins are extracted from lysate, digested by a protease (primarily Trypsin), and the resulting complex mixture of often thousands of peptides is separated by multidimensional chromatography before being subjected to MS analysis for identification by database search strategies. As previously pointed out by Meinnel *et al.* [5], protein N-termini are often underrepresented in such high-throughput, non-targeted MS analysis. Theoretically, it may be argued that about 10% of the detected peptides should involve the N-terminus, but deposited data indicate that N-termini are identified much less frequently. There may evidently be several reasons for that. Protein N-termini being extensively processed may be present in different forms (unmodified, acetylated, methionine cleaved) and when performing database searches, all modification states should be allowed, increasing the search space. Additionally, N-terminal acetylated peptides are atypical tryptic peptides that display a different charge distribution and are more hydrophobic. This may lead to negative discriminative effects for N-terminal acetylated peptides in both the peptide separation technologies typically used as well as in the ionization and subsequent fragmentation process, which are primarily optimized for doubly charged, tryptic peptides. N-terminal acetylated peptides may also fragment differently than regular tryptic peptides, and as most search algorithms have been developed using fragmentation rules of tryptic peptides, this may also have a negative effect on the identification probability of N-terminal acetylated peptides.

It is evident that targeted proteomic approaches are beneficial for the analysis of acety-

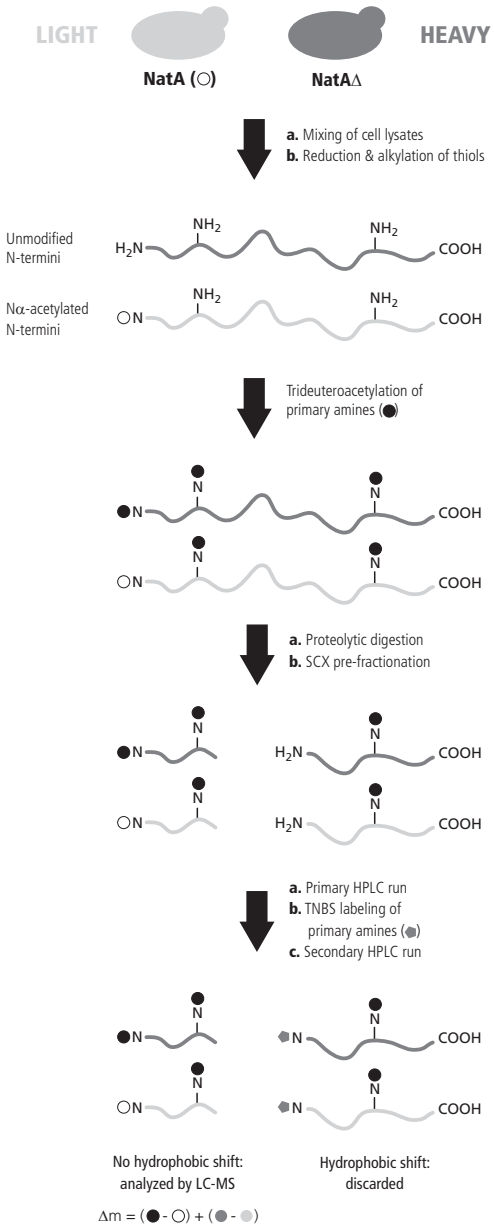
lated protein termini to reduce the peptide complexity of the original sample. Enrichment for specific classes of peptides, in this case N-terminal acetylated peptides, may also increase the confidence in protein identifications. This may sound counterintuitive as co- or post-translationally modified peptides, and especially peptides representing the N-terminus of a protein, are by nature 'one-hit-wonders' [86,87], but they can boost the identification of proteins identified by single peptides from the bulk of unmodified peptides in the flow-through fractions of the enrichment. A complicating issue in the analysis of not specifically enriched N-terminal acetylated peptides is the competitive occurrence of N-terminal acetylation and lysine acetylation, in particular when the N-terminal amino acid is a lysine. Therefore, in these cases, manual validation of the identifications is still required for quality control [87].

Studies of the Oesterhelt group [7,27] on the analysis of N-terminal peptides from the archaea *Halobacterium salinarum* and *Natronobacterium pharaonis* illustrate the significant gains obtained by using targeted approaches. Their work is of great interest as it represents some of the first large-scale studies of N-terminal peptides from prokaryotes, showing that protein N-terminal acetylation is also reasonably abundant in these species. From a technology viewpoint the publication [7] is of great interest as it compares the results of several targeted and non-targeted methods allowing the characterization of protein N-terminal peptides, namely combined fractional diagonal chromatography (COFRADIC) and strong cation exchange (SCX) chromatography, both known to enrich for N-terminally blocked peptides, but also N-terminal peptide identifications that were extracted from a large collection of proteome-wide high-throughput data gathered from the cytosolic, membrane and low molecular weight proteomes of the studied prokaryotes. In general, their work confirmed that the use of specific methods in the identification of N-terminal peptides resulted in a dramatic increase in number of detected N-termini. Also Goetze *et al.* [43] compared data for N-terminal peptides (both unmodified and N-acetylated) obtained from classical shotgun proteomics approaches with COFRADIC-enriched fractions and reported an increase of detected N-termini of roughly one order of magnitude.

3.3 Targeted approaches

Negative selection by COFRADIC

Over the years, the groups of Gevaert and Vandekerckhove have introduced several elegant approaches based on two-dimensional diagonal chromatography termed CO-



FRADIC [88-90]. COFRADIC aims at the negative selection of specific classes of peptides from complex mixtures, *i.e.* the quantitative removal of all unwanted peptides to yield the target class of peptides. It typically consists of two or more consecutive chromatographic separations, with a chemical or biochemical modification step targeted to a subset of peptides between the two separations. The modified peptides obtain different chromatographic properties and segregate from the bulk of unaltered peptides in the second separation. The final analysis is limited to the targeted class of peptides, thereby significantly reducing the complexity but keeping all of the characteristics of the proteome. The application of the method ranges from the selective isolation of peptides containing methionine residues, phosphopeptides and N-terminal peptides. The application of COFRADIC for the enrichment of N-terminal peptides is illustrated in Figure 3. First, all

Figure 3. Schematic overview of the COFRADIC workflow applied for determining NatA-dependent N-terminal acetylation sites. Proteins from unlabeled wild-type and isotope-labeled NatAΔ yeast are quantitatively trideuteroacetylated at free α-and ε-amino groups, digested, and subjected to SCX to enrich for N-terminally acetylated peptides. N-terminal peptides are then further separated from internal and C-terminal peptides by labeling of newly generated free α-amino groups with 2,5,6-trinitrobenzene sulfonic acid in combination with two RP-HPLC runs. Peptide pairs are identified by LC-MS and quantified using the mass difference introduced by the metabolic label and/or the trideuteroacetyl group. Adapted from [26].

free amino groups of denatured yet undigested proteins are blocked, for example by chemical acetylation. The subsequent proteolytic digestion generates a mixture of protein-internal and protein C-terminal peptides with unblocked N-termini and protein N-terminal peptides with blocked termini. The free amino groups are then trinitrophenylated with 2,4,6-trinitrobenzenesulfonic acid (TNBS) which in a reversed phase (RP) separation causes a hydrophobic retention, while protein N-terminal peptides which did not react with TNBS will be separated normally. In more recent applications, peptide separation was further enhanced by SCX before the TNBS-modification step to allow chromatographic sorting [88,91,92]. When the amino groups are blocked by isotopically coded acetyl groups, this technique also allows distinguishing and quantitatively analyzing *in vivo* acetylated versus non-acetylated N-termini. The technique has been successfully used to characterize N-terminal acetylation in different proteomes [7,43] and also to evaluate the orthologous function of the yeast and human NatA complex [26]. In the latter study it was shown that in yeast a loss of both catalytic subunits (Naa10 and Naa15) can be complemented by the two human homologues but not by a combination of yeast and human subunits. This indicates that the function of N-acetyltransferases is highly conserved, although apparent structural differences do not allow a direct combination of subunits from different species.

Goetze *et al.* [43] used COFRADIC to enrich and identify amino-terminal peptides from proteins extracted from membrane, cytoplasmic, and nuclear fractions of *Drosophila melanogaster* Kc167 cells. Overall, they detected more than 1,200 protein N-termini and could show that N-terminal acetylation occurs in *Drosophila* with a frequency somewhere between yeast and humans. From this large dataset they concluded that acetylation will never occur when the adjacent amino acid is proline.

Negative selection by beads and polymers

Another elegant way to enrich *in vivo* N-acetylated peptides uses non-chromatographic negative selection. Like COFRADIC, this method also requires the prior blocking of all free amino groups, which can be performed, for instance, by chemical acetylation [93,94] or by reductive dimethylation [95,96]. Following proteolysis, internal peptides containing free amines can be tagged through biotinylation, for example with a NHS ester derivative, and depleted using streptavidin [94,97,98]. Alternatively, amine reactive matrixes, either coupled to beads or to polymers, are used to selectively couple and capture the vast amount of free amine-containing peptides allowing the negative selec-

tion of N-terminal acetylated peptides and other peptides with blocked N-termini in the flow-through [93,94,96,99]. Several groups have explored reaction methods that are specific towards the α-amino group, excluding the ε-amino group, such as isocyanate-coupled or cyanogen bromide-activated beads [100,101], which evidently is beneficial for even more targeted enrichment strategies.

These negative selection approaches can be very useful as the complexity of the sample is hugely diminished. Application of such an approach lead Zhang *et al*. [101] to the identification of 588 *in vivo* N-terminally modified peptides, including a few propionylated peptides. The group of Overall recently reported the use of hyperbranched polyglycerols functionalized with aldehyde moieties. These polymers react very efficiently with all unblocked internal and C-terminal tryptic peptides, resulting in a more than 10-fold improvement in capacity over other amine-reactive resins [96]. Although their elegant approach can be used for the analysis of protein N-terminal acetylation, their reported work focuses primarily on the detection on neo-termini of proteins generated by specific proteolytic cleavage.

Strong cation exchange

SCX is a chromatographic separation technology often used in proteomics, most typically to pre-fractionate peptides prior to further separation by RP chromatography [102]. The separation mechanism in SCX is primarily based on the charge of the peptides in solution which is determined by its sequence, in particular by the number of basic and acidic residues. Ignoring miscleavages, tryptic peptides have a lysine or arginine residue at the C-terminus and one free amino group at the N-terminus. Therefore the bulk of the peptides generated by Trypsin contain two positive charges at pH <3, which means that they largely co-elute in SCX. A typical SCX chromatogram of a tryptic digest shows a very strong peak corresponding to the doubly charged peptides, with some additional shoulders eluting at a later stage formed by triply and higher charged peptides, as a result of miscleaved peptides and/or other peptides containing more than one basic residue such as histidine. There are also classes of peptides that contain less than two charges; in a tryptic digest they comprise blocked N-terminal peptides. At pH <3 the blocked N-terminal peptides can be distinguished and enriched from the tryptic peptides on the basis of charge difference (Figure 4). Additionally, when a tryptic peptide is modified by serine, threonine or tyrosine phosphorylation, its net charge will also be diminished by one and thus will have the same charge as N-terminal blocked

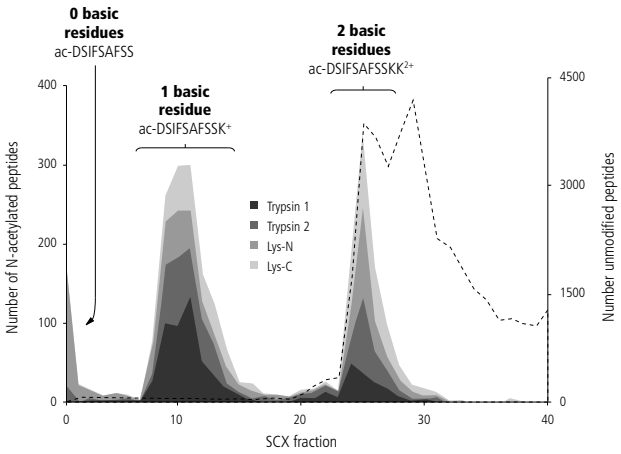


Figure 4. SCX-based enrichment of N-terminal acetylated peptides generated by digestion with different proteases. The elution profiles of N-terminally acetylated peptides show complete separation into peptides containing 0, 1 or 2 basic residues. N-terminal acetylated peptides with 2 basic residues overlap with the abundant non-acetylated peptides with 1 basic residue (dashed line, values on secondary axis).

peptides. Therefore, SCX chromatography can be used to specifically pre-fractionate phosphopeptides and N-terminal blocked peptides. The use of SCX for the enrichment of phosphopeptides, pioneered at the proteome-wide scale by Beausoleil *et al*. [103] and Gruhler *et al.* [104] is nowadays well established and essential for complete mapping of phosphoproteomes. In these phosphoproteomic studies the samples are often first pre-fractionated by SCX and thereafter the early-eluting fractions, containing the majority of phosphorylated peptides, are further enriched by more specific enrichment strategies [105,106]. Such approaches can nowadays result in the identification of over 10,000 phosphopeptides from cell lysates [8,107-110].

N-terminal acetylated peptides also require significant enrichment to be detected by mass spectrometry approaches in large enough numbers, and as pointed out above SCX may be used for that purpose. Dormeyer *et al.* were among the first to explore the capacity of SCX to enrich peptides originating from *in vivo* acetylated protein N-termini in a cell membrane-enriched fraction of human embryonic carcinoma cells [111]. An in-depth liquid chromatography-coupled MS (LC-MS) screen revealed that the bulk of N-terminal acetylated peptides eluted in two consecutive early fractions from the SCX column. 116 N-terminal acetylated and 26 protein C-terminal peptides were identified in these two fractions representing 92% of the total number of unique peptides in these fractions. The beauty of this approach is the simple sample preparation protocol involved, which does not require any chemical derivatization. The study pointed out that the purity and high enrichment grade allowed the identification of 87 unpredicted acetylated N-termini. These previously unannotated N-acetylated peptides were conform to the criteria for *in vivo* N-terminal acetylation, *i.e.* enriched in methionine,

alanine and serine N-terminal residues. The significant number of protein N-termini suggests a high degree of unannotated gene boundaries or specific protein processing.

Next to Trypsin, other proteases are increasingly used in proteomics, for example Lys-C, Chymotrypsin, Glu-C, and Lys-N. They exhibit quite different cleavage specificities, and some of them are rather unspecific. A high specificity is beneficial in proteomics as it assists peptide identifications by database searches. With the advent of electron capture dissociation (ECD) and electron transfer dissociation (ETD) there has been a growing need to generate proteolytic peptides that are both longer in length and carry more charges, as ETD then outperforms collision induced dissociation (CID) [112,113]. Multiply charged peptides can be generated by Trypsin when the sample is incubated with the enzyme for only a limited time, allowing extensive miscleavages, or in a more controlled manner by using enzymes like Lys-C and Lys-N, which cleave C- and N-terminally of a lysine residue, respectively. Multiply charged peptides can also be fractionated by SCX, and for instance van den Toorn *et al.* optimized peptide identification in multidimensional protein identification technology (MudPIT) experiments combining SCX fractionation with CID on the earlier fractions and ETD on the later fractions [112].

Lys-N is an enzyme with interesting properties as it exhibits high thermo-stability, high tolerance towards detergents, and a proteolytic activity in a broad pH range, and can also be used for in-gel digestion [114,115]. Most importantly, it was shown to have a very significant specificity for cleaving before lysine [116,117]. Lys-N is also interesting when used in combination with SCX, especially when focusing on peptides that elute in early SCX fractions. When Lys-N cleaves peptides that are N-terminally blocked, these peptides acquire a net charge of zero. In contrast to Trypsin, Lys-N-generated protein C-terminal peptides still contain a lysine residue and will therefore elute later in the SCX run. Therefore, especially the combination of SCX with Lys-N facilitates the pre-fractionation and specific enrichment of N-terminal acetylated peptides [118,119]. This feature was first evaluated and explored by Taouatas *et al.* who subjected Lys-N-generated peptides to low-pH SCX, obtaining fractionation profiles in which peptides from different functional categories could be separated [118]. The four categories they were able to distinguish and to separate to near completion were (a) acetylated N-terminal peptides, (b) singly phosphorylated peptides containing a single basic residue, (c) peptides containing a single basic residue, and (d) peptides containing more than one basic residue. They further assessed whether CID or ETD was better for the analysis of these peptide pools. The N-terminal acetylated peptides in category (a) could be identified confidently either by both CID and ETD with spectra dominated by sequence-informative z-ions.

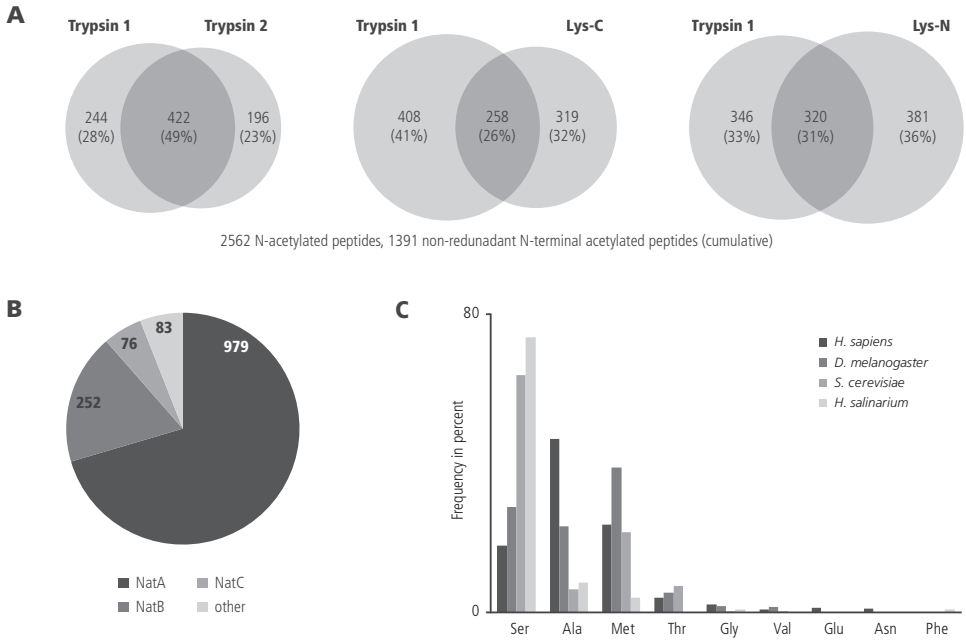


Figure 5. A, Redundancy, reproducibility and complementarities observed in a SCX-based approach to analyze *in vivo* acetylated termini in HEK293 cells, enabling the identification of around 1,400 N-acetylated proteins. B, distribution of the protein termini to the various NAT machineries when applying the NAT substrate selectivity rules as summarized in Figure 2. C, distribution of N-terminal amino acid residues detected in large-scale proteome analysis of N-terminal acetylated peptides in digests from a human cell line, cells from fruitfly, yeast and halobacterium. Amino acid residues with an occurrence of <1% are omitted for clarity.

For the phosphorylated peptides in category (b) and the regular, single lysine-containing peptides in category (c), ETD provided straightforward sequence ladders of c-ions, facilitating sequencing and the determination of the exact location of possible phosphorylation sites [114]. Although the results of Taouatas *et al.* already indicated that the use of Lys-N can be beneficial for the analysis of acetylated protein N-termini, Gauci *et al.* [119] extended this observation by comparatively assessing the performance of Trypsin, Lys-C and Lys-N in the analysis of protein phosphorylation and N-terminal acetylation. They found that the overlap between the peptide datasets generated with different proteases was marginal compared to the large overlap between two datasets generated with Trypsin. While Gauci *et al.* focused primarily on phosphopeptides, using an identical experimental approach Helbig *et al.* identified 1,400 N-terminal acetylated protein N-termini (Figure 5) [120]. Most of these acetylated protein N-termini could be classified following the substrate sequences patterns of the known N-acetyltransferases, with

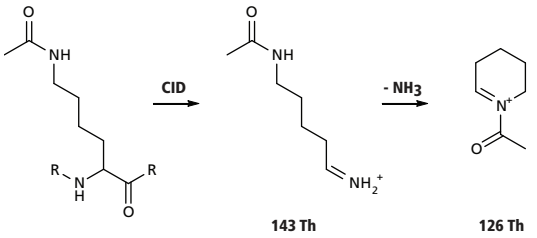


Figure 6. Chemical structures of immonium ions generated during CID of a lysine acetylated peptide that may be used as marker ions.

about 25% of the detected peptides starting with methionine, 45% with alanine and 20% with serine. The study also discovered an exceptionally higher frequency of alanine residues at the second position of human proteins. Further genome-wide comparative analyses revealed that this effect is not related to different processing of protein N-termini, but can be traced back to species-specific characteristics of the genome. These data expand on the observations made by Falb *et al*. [27] who referred to them as archaeal-specific patterns for N-terminal acetylation, as they observed N-terminal acetylated serine and alanine but no N-terminal acetylated threonine, although mature proteins frequently start with this amino acid.

4 Methods to probe lysine acetylation

Although lysine acetylation is emerging as a frequently occurring PTM not only in histones, it is also relatively low abundant and thus requires enrichment over non-acetylated peptides for detection in high-throughput proteomics. However, no chromatographic enrichment technique specific for lysine acetylated peptides, which would greatly facilitate large-scale studies, has been devised so far. Therefore, many studies have characterized *in vivo* protein lysine acetylation sites by targeting a protein complex or protein directly by affinity purification [121]. In combination with common protein or peptide pre-fractionation techniques this usually sufficiently reduces sample complexity to detect acetylated peptides. Besides the lack of chromatographic pre-fractionation techniques, the highly divergent cellular and molecular function of lysine acetylation might require other mass spectrometric techniques than just peptide sequencing. Therefore it is common to couple *in vitro* acetylation assays with MS-based approaches. Recently it has been shown that the use of $^{12}C_2$,H_3-acetyl-CoA and $^{13}C_2$,D_3-acetyl-CoA in an equimolar mixture aids the identification of acetylation sites in peptide sequencing due to the observed mass difference of 5 Da per added acetylation site with equal peptide intensities of light and heavy in MS spectra [122]. Since the input material for *in vitro* acetylation assays is usually relatively pure, MS of protein or protein fragments

can be used to determine the number of acetylation sites and the population of each acetylation state. *In vitro* acetylation assays have also been performed on protein microarrays that cover nearly the complete yeast proteome, allowing the discovery of new KAT target candidates [123].

In non-targeted approaches, accurate mass measurement of the precursor ion greatly facilitates differentiation of the acetyl lysine moiety from the nearly isobaric trimethylated lysine residue, which has a mass difference of 0.03 Da compared to its acetylated counterpart. For typical tryptic peptides this requires a mass accuracy in the range of 20-30 ppm, which is commonly achieved by state-of-the-art time of flight (TOF), ion cyclotron resonance (ICR) and orbitrap instruments [124]. In addition to this, acetylated lysine residues generate diagnostic ions upon fragmentation by CID, namely the immonium ion at 143.118 Th and its cyclic derivative at 126.091 Th, which can be used to confirm the presence of acetylated lysine residues and to distinguish it from trimethylated lysine residues (Figure 6) [124,125]. The 143 Th immonium ion is less specific for the acetylated lysine residue than the 126 Th ion because some frequently occurring dipeptide motifs can generate internal fragment ions which are almost isobaric to the 143 Th ion [125,126]. The 126 Th itself is highly specific but has a poor sensitivity for acetylated lysine, partly because the formation of the 143 Th ion and hence also the 126

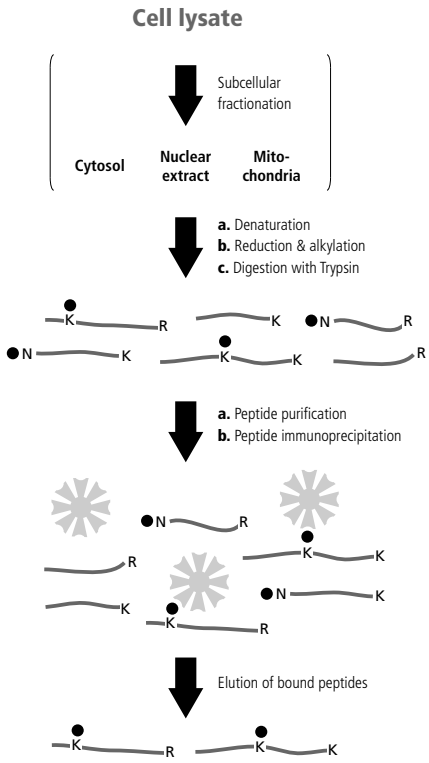


Figure 7. Peptide immunoprecipitation for the enrichment of lysine acetylated peptides using immobilized pan-specific anti-acetyl lysine antibodies. The input for the immunoprecipitation are peptides prepared by digestion from cell lysate. An optional organelle isolation and/or extraction may aid in depleting highly abundant, strongly lysine acetylated proteins such as histones which might hamper the detection of low-abundant lysine acetylated peptides. Lysine acetylated peptides are typically eluted from the immobilized antibodies using acidic conditions, which facilitates subsequent C18-purification and direct LC-MS analysis.

Th ion is strongly depending on the position of the acetyl lysine within the peptide: The closer to the N-terminus the higher the signal of the immonium ions [126]. Both diagnostic ions have been used to specifically select for lysine acetylated peptides in multiple reaction monitoring- (MRM)-based peptide sequencing [127-129].

An immunological technique to enrich for lysine acetylated peptides can be achieved by the use of pan-specific antibodies (Figure 7). Compared to N-acetylated peptides, the side chain of lysine is relatively bulky and the acetylated ε-amino group is chemically so different from the free primary amine that the generation of pan-specific antibodies specifically recognizing the acetyl lysine side chain is possible and mono- and polyclonal pan-specific antibodies are commercially available [130-132]. These antibodies recognize acetyl-lysine residues relatively independently of the flanking amino acids and therefore facilitate the enrichment of lysine acetylated peptides from complex peptide mixtures by peptide immunoprecipitation, a method employed in several recent proteomic studies [133-138]. The drawback of using polyclonal antibodies is that the immunized animals do not produce the same antibodies with repeated immunization. Moreover, acetylation site consensus sequences built from large datasets of acetylated peptides have still to be taken with care, as they might not only reflect a specificity of the KAT but also include that of the used antibodies. It is beneficial, however, that this enrichment technique provides a negative selection against N-acetylated and trimethylated peptides, which facilitates data analysis.

5 Large-scale studies of lysine acetylation

Peptide enrichment using pan-specific antibodies followed by tandem mass spectrometry analysis has so far been the only method employed in studies targeting lysine acetylation on a proteome-wide scale [133-138]. In a first survey study, Kim *et al*. analyzed lysine acetylation in cytosolic and nuclear extracts from HeLa S3 cells treated with TSA and sirtinol, two KDAC inhibitors, as well as mitochondria lysates of liver tissue from fed and fasted mice [133]. Using low-resolution, low mass accuracy ion trap mass spectrometry, the authors reported 57/78 unique lysine-acetylated peptides from the cytosolic/nuclear HeLa cell extract and 188/219 from fed/fasted liver mitochondria lysates. In all four datasets, the cellular functions of the identified acetylated proteins are scattered across divergent functional classes and do not display a common trend, which is in line with the knowledge accumulated in numerous particulate studies on

the function of lysine acetylation of non-histone proteins [58]. However, lysine acetylation is very common in the mitochondria of both fed and fasted mice, with more than 20% of known mitochondrial proteins being acetylated. This figure increases to up to 65% in proteins involved in mitochondrial energy metabolism, such as the citric acid cycle, oxidative phosphorylation and fatty acid oxidation. Acetylation is also frequent in mitochondrial dehydrogenases, with 44% of all mitochondrial dehydrogenases being acetylated. Next to these important observations, the authors also reported an overlap of acetylation sites with the sumoylation site consensus sequence and with predicted nuclear localization signals.

Lysine acetylation can also be detected in *Escherichia coli*, which was shown by two separate groups reporting 138 and 128 acetylation sites, respectively, from whole-cell lysates [134,138]. Like in mammalian mitochondria, many metabolic enzymes were acetylated, among them enzymes involved in carbohydrate metabolism, glycolysis and gluconeogenesis, the citric acid cycle, as well as amino acid and nucleotide metabolism. Moreover, the acetylation patterns were found to adapt significantly upon switching from aerobic to anaerobic conditions, suggesting that lysine acetylation is involved in the regulation of metabolism [134]. This hypothesis was further consolidated by proteome-wide analysis of lysine acetylation in a human liver tissue [135] and in *Salmonella enterica* [137], again using antibody-based enrichment (with a different antibody than the one used in the studies mentioned above) but in combination with high-resolution, high mass accuracy MS. In mitochondrial and cytosolic fractions of human liver tissue, 1047 acetylated proteins were identified, with many of them involved in fundamental pathways of intermediary metabolism [135]. Extensive biochemical analyses in cultured cells and *in vitro* of selected metabolic enzymes of fatty acid catabolism (EHHADH), the citric acid cycle (MDH2), the urea cycle (ASL) and gluconeogenesis (PCK1) showed that acetylation of these enzymes regulates their enzymatic activity either positively or negatively and that the state of acetylation is coupled to the availability of extracellular nutrients. In *Salmonella enterica* lysate, 191 proteins were acetylated [137]. Like in human and murine liver tissue, enzymes of key metabolic pathways are represented; in fact, nearly all enzymes of glycolysis, gluconeogenesis and the citric acid cycle are acetylated. Since only a single KAT (Pat) and a single KDAC (CobB, an NAD^+-dependent KDAC) have been reported for *Salmonella*, the authors used Δ*pat* and Δ*cobB* mutants to show by gas chromatography- (GC) coupled MS that the flux of metabolites through central metabolic pathways depends on the activity of the KAT/KDAC pair. Moreover, they demonstrated biochemically for enzymes of glycolysis (GAPDH), the citric acid cycle

(IDH) and the glyoxylate pathway (ICL) that changes in metabolic flux are indeed mediated by enzyme acetylation.

A different approach was reported by Lin *et al.* who used a protein array comprising 5,800 proteins from *Saccharomyces cerevisiae* to screen for targets of the essential KAT Esa1 *in vitro* [123]. Purified NuA4, which is the protein complex that harbors Esa1 *in vivo*, acetylated 91 proteins of the proteome array, among them many metabolic enzymes. The acetylation sites were identified by MS. Most prominently, the authors demonstrated that the acetylation state of a single lysine residue of PCK1, the key enzyme of gluconeogenesis, is determined by Esa1 and the KDAC Sir2 *in vivo* and regulates the enzymatic activity of PCK1. These studies consolidate the finding that lysine acetylation is involved in regulating the enzymatic activity of core metabolic enzymes, which emerged by the observation that the *Salmonella enterica* acetyl-CoA synthetase ACS, which catalyzes the formation of acetyl-CoA from ATP, CoA and acetate, is activated upon deacetylation by CobB, and inactivated by acetylation by PatB [139,140]. This acetylation switch is conserved in mammalian cells, which have a cytosolic (AceCS1) and a mitochondrial (AceCS2) enzyme, whose enzymatic inactivation is relieved by the deacetylase activity of the sirtuins SIRT1 and SIRT3, respectively [141,142]. Importantly, through acetyl-CoA as cofactor for KATs and NAD^+ as cofactor for sirtuin-class KDACs, the acetylation state may be directly coupled to the intracellular levels of these key metabolites. Acetylation switches on key transcription factors also regulate metabolic pathways [143]. In this context, lysine acetylation, and the KATs and KDACs regulating these events have been closely linked to ageing and age-related diseases [144-146].

The largest dataset on lysine acetylation has been reported by Choudhary *et al.* who cumulatively identified 3,600 unique acetylation sites in several experiments with three human cell lines, untreated as well as treated with specific histone deacetylase inhibitors (MS275 and SAHA), mapping to 1,750 acetylated proteins [136]. In addition to the above described proteomic analyses, this study highlighted also the abundance of lysine acetylation in the nucleus, were a large number of acetylated proteins with functions such as DNA replication, DNA damage repair, RNA splicing, transcription, cell cycle regulation, and chromatin remodeling were present. Many of them are subunits of larger complexes and acetylation sites are accumulated on subunits of methyltransferases, ubiquitin ligases and deubiquitinylases, and chromatin remodeling complexes. Interestingly, many key nuclear KAT complexes carry a number of acetylated subunits as well.

6 Propionylation and butyrylation of lysine

Lysine propionylation and butyrylation, which are much less common than acetylation, were first detected as PTMs of histones [133,147,148] when Chen *et al.* [147,149] showed that p300/CBP can also catalyze propionyl and butyryl group transfer to histones. More recently, these PTMs have also been reported to occur on non-histones proteins such as p53 and p300/CBP [150]. p300/CBP probably use propionyl-CoA and butyryl-CoA as cofactors in these reactions. Moreover, SIRT1 was shown to be able to act as a depropionylase [150]. Garrity *et al.* [151] reported that the propionyl-CoA synthetase PrpE of *Salmonella enterica* can be reversibly propionylated *in vivo*. Propionylation was removed by CobB from which they concluded that propionylation/depropionylation is a regulatory mechanism. Based on these observations, it appears that some enzymes are common to lysine propionylation and lysine acetylation regulatory pathways. Whether lysine propionylation and butyrylation play a specific role in regulatory processes remains elusive. Propionylation and butyrylation have also been reported to minor extent at the N-terminus of proteins (Figure 8) [101,111]. As the propionylated substrates were in some cases detected in parallel to the analog N-terminal acetylated peptides, it has been suggested that both N-terminal modifications are performed by the same enzyme [101].

7 Conclusions and future outlook

The proteome-wide characterization of N-terminal and lysine acetylated peptides has been greatly facilitated by the development of specific enrichment techniques in combination with advances in LC-MS, in particular the improvement in mass accuracy and mass resolution as implemented on current instruments [152]. N-terminal acetylated peptides appear to fragment differently from their unmodified counterparts in CID while lysine acetylated peptides apparently do not [153]. Generic advances in MS instrumentation may further increase detection of these modifications. Higher-energy collisional dissociation (HCD) greatly improves peptide sequencing as it allows a CID regime on the commercial LTQ Orbitrap that is similar to that in TOF instruments, with fragment ion readout in the orbitrap analyzer, thereby generating high mass accuracy MS/MS spectra with the sensitivity and speed of comparable to a linear ion trap [154]. HCD also overcomes the limited trapping capabilities of ion traps for low m/z fragment ions and so enables the utilization of immonium ions. The implementation of

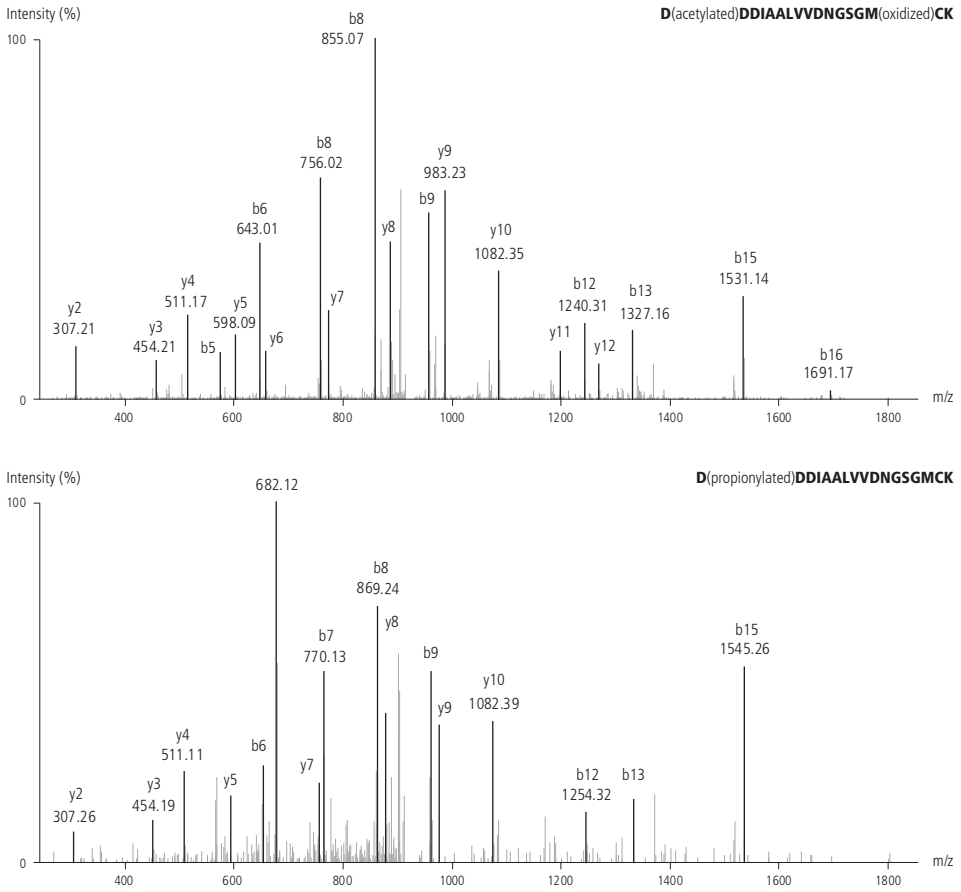


Figure 8. CID tandem mass spectra of the acetylated (upper spectrum) and propionylated (lower spectrum) N-terminus of human actin (UniProt accession number P60709). The similarity of the spectra assists in the genuine identification of N-propionylation as PTM of actin.

ion funnels to improve transmission of the analyte from the electrospray source into the mass spectrometer has greatly increased the sensitivity of various proteomics grade mass spectrometers [155]. Despite these advances, the dynamic range reached by modern mass spectrometers is still insufficient to comprehensively and quantitatively detect acetylated peptides in complex mixtures and therefore targeted peptide-based enrichment is often still essential.

A current limitation for the detection of lysine acetylation, for which antibodies are at present the only choice for enrichment, is the high dynamic range of lysine acetylation in the cell which may obscure the detection of low abundant acetylated peptides due

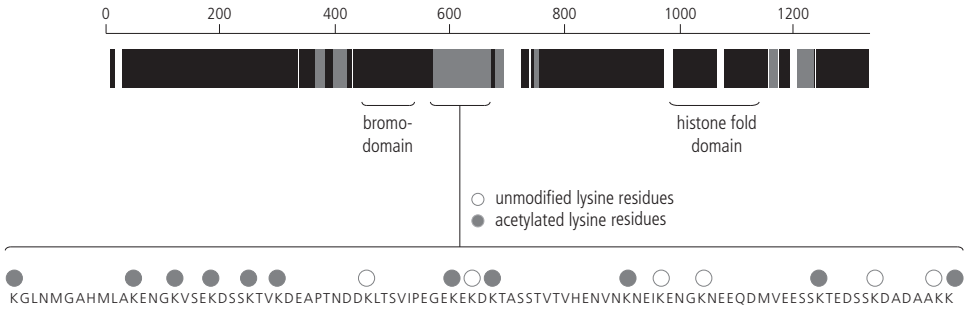


Figure 9. Lysine acetylation on Spt7, a subunit of the Gcn5-containing KAT complex SAGA, detected by a non-targeted approach. Most prominent is a hyperacetylated stretch directly adjacent to the bromodomain. In this stretch, 11 out of 17 lysine residues are acetylated.

to the relatively high abundance of others, for example of histones. This affects both peptide immunoprecipitation and MS analysis and might require sub-cellular fractionation of the analyzed material prior to enrichment. Similarly, short acetylated peptides which do not lead to successful peptide identifications may also be problematic during the immunoprecipitation by blocking binding sites for detectable acetylated peptides. ECD and ETD have been extensively used in the analysis of highly modified histone tail peptides [156,157]. Similarly, ETD might find a niche in the characterization of larger peptides resulting from hyperacetylated protein stretches, such as the auto-regulatory loop of p300/CBP, which is critical for enzymatic activation [158], or the functionally uncharacterized hyperacetylated stretch of Spt7, a subunit of the SAGA complex in *Saccharomyces cerevisiae* (Figure 9) [121]. These stretches contain many closely neighbored lysine residues in either unmodified or acetylated form, and the acetylation state of these stretches might be functionally relevant.

The proteome-wide analysis of protein N-terminal acetylation, vastly benefiting from chromatographic separation techniques, is far more advanced. The use of alternative proteases and alternative fragmentation techniques may relief detection problems related to peptide size and amino acid composition. Although the protein machineries responsible for N-terminal acetylation have been linked to specific substrates, the function of N-terminal acetylation is still not well known. Functional conclusions are particularly difficult to draw for this PTM because although the protein machineries are very well conserved, the substrates – protein N-termini – are not, even for proteins with high sequence homology. A protein which is a NatB substrate in yeast may be a NatA substrate in human. However, the detailed investigation of the function of N-terminal

acetylation may become accessible by analyzing point mutants, for example as done by Goetze *et al.* [43] who genetically engineered the N-termini of proteins with proline at the second position which prevents N-terminal acetylation.

Recent proteomics studies on N-terminal and lysine acetylation have revealed that these modifications are much more abundant and occurring on many more proteins than previously anticipated. The major bottleneck in the near future will not be the detection of these PTMs but rather the understanding of their biological significance, a stage of research where phosphoproteomics was several years ago. Relatively soon, a better view will emerge on which acetyltransferase and deacetylases can be linked to specific substrates. It is evident that further functional annotation of protein acetylation will greatly benefit from the new technologies reviewed in this work.

8 Acknowledgements

We would like to acknowledge all members of the Heck group for their contributions, in particular Shabaz Mohammed and Reinout Raijmakers. We acknowledge financial support from the NGI Horizon Program (grant number 050-71-050). The Netherlands Proteomics Centre, embedded in the Netherlands Genomics Initiative, is kindly acknowledged for financial support. AJRH likes to acknowledge the Aebersold group for supporting his sabbatical stay as a guest professor at the Department of Biology at the ETH Zürich and the Netherlands Genomics Initiative for granting him a Distinguished Visiting Scientist Stipend.

9 References

[1] Walsh, C. T.*, et al.*, *Angew Chem Int Ed Engl* **2005**, *44*, 7342
[2] Witze, E. S.*, et al.*, *Nat Methods* **2007**, *4*, 798
[3] Mann, M.*, et al.*, *Nat Biotechnol* **2003**, *21*, 255
[4] Jensen, O. N., *Curr Opin Chem Biol* **2004**, *8*, 33
[5] Meinnel, T.*, et al.*, *Proteomics* **2008**, *8*, 626
[6] Eichler, J.*, et al.*, *Microbiol Mol Biol Rev* **2005**, *69*, 393
[7] Aivaliotis, M.*, et al.*, *J Proteome Res* **2007**, *6*, 2195
[8] Grimsrud, P. A.*, et al.*, *ACS Chem Biol* **2010**, *5*, 105
[9] Rogers, L. D.*, et al.*, *Mol Biosyst* **2009**, *5*, 1122

[10] Zaia, J., *Chem Biol* **2008**, *15*, 881
[11] An, H. J.*, et al.*, *Curr Opin Chem Biol* **2009**, *13*, 601
[12] Zielinska, D. F.*, et al.*, *Cell* **2010**, *141*, 897
[13] Kirkpatrick, D. S.*, et al.*, *Nat Cell Biol* **2005**, *7*, 750
[14] Matic, I.*, et al.*, *Mol Cell Proteomics* **2008**, *7*, 132
[15] Martin, B. R.*, et al.*, *Nat Methods* **2009**, *6*, 135
[16] Yang, W.*, et al.*, *Mol Cell Proteomics* **2010**, *9*, 54
[17] Polevoda, B.*, et al.*, *Genome Biol* **2002**, *3*
[18] Strous, G. J.*, et al.*, *Biochem Biophys Res Commun* **1974**, *58*, 876
[19] Pestana, A.*, et al.*, *Biochemistry* **1975**, *14*, 1404
[20] Palmiter, R. D.*, et al.*, *Proc Natl Acad Sci USA* **1978**, *75*, 94
[21] Gautschi, M.*, et al.*, *Mol Cell Biol* **2003**, *23*, 7403
[22] Polevoda, B.*, et al.*, *J Cell Biochem* **2008**, *103*, 492
[23] Polevoda, B.*, et al.*, *J Biol Chem* **2000**, *275*, 36479
[24] Nakazawa, T.*, et al.*, *Proteomics* **2008**, *8*, 673
[25] Driessen, H. P.*, et al.*, *CRC Crit Rev Biochem* **1985**, *18*, 281
[26] Arnesen, T.*, et al.*, *Proc Natl Acad Sci USA* **2009**, *106*, 8157
[27] Falb, M.*, et al.*, *J Mol Biol* **2006**, *362*, 915
[28] Tsunasawa, S.*, et al.*, *J Biol Chem* **1985**, *260*, 5382
[29] Polevoda, B.*, et al.*, *J Mol Biol* **2003**, *325*, 595
[30] Flinta, C.*, et al.*, *Eur J Biochem* **1986**, *154*, 193
[31] Bradshaw, R. A.*, et al.*, *Trends Biochem Sci* **1998**, *23*, 263
[32] Giglione, C.*, et al.*, *Cell Mol Life Sci* **2004**, *61*, 1455
[33] Li, X.*, et al.*, *Proc Natl Acad Sci USA* **1995**, *92*, 12357
[34] Addlagatta, A.*, et al.*, *Biochemistry* **2005**, *44*, 14741
[35] Addlagatta, A.*, et al.*, *Biochemistry* **2005**, *44*, 7166
[36] Walker, K. W.*, et al.*, *Protein Sci* **1998**, *7*, 2684
[37] Polevoda, B.*, et al.*, *BMC Proc* **2009**, *3 Suppl 6*, S2
[38] Song, O. K.*, et al.*, *J Biol Chem* **2003**, *278*, 38109
[39] Mullen, J. R.*, et al.*, *EMBO J* **1989**, *8*, 2067
[40] Park, E. C.*, et al.*, *EMBO J* **1992**, *11*, 2087
[41] Starheim, K. K.*, et al.*, *BMC Proc* **2009**, *3 Suppl 6*, S3
[42] Arnold, R. J.*, et al.*, *J Biol Chem* **1999**, *274*, 37035
[43] Goetze, S.*, et al.*, *PLoS Biol* **2009**, *7*, e1000236
[44] Polevoda, B.*, et al.*, *Biochem Biophys Res Commun* **2003**, *308*, 1

[45] Arnesen, T., *et al.*, *Biochem J* **2005**, *386*, 433
[46] Starheim, K. K., *et al.*, *Biochem J* **2008**, *415*, 325
[47] Starheim, K. K., *et al.*, *Mol Cell Biol* **2009**, *29*, 3569
[48] Arnesen, T., *et al.*, *Mol Cell Biol* **2010**
[49] Arfin, S. M., *et al.*, *Biochemistry* **1988**, *27*, 7979
[50] Varshavsky, A., *Proc Natl Acad Sci USA* **1996**, *93*, 12142
[51] Giglione, C., *et al.*, *EMBO J* **2003**, *22*, 13
[52] Meinnel, T., *et al.*, *Biochimie* **2005**, *87*, 701
[53] Meinnel, T., *et al.*, *Biol Chem* **2006**, *387*, 839
[54] Hwang, C. S., *et al.*, *Science* **2010**, *327*, 973
[55] Shahbazian, M. D., *et al.*, *Annu Rev Biochem* **2007**, *76*, 75
[56] Kouzarides, T., *Cell* **2007**, *128*, 693
[57] Campos, E. I., *et al.*, *Annu Rev Genet* **2009**, *43*, 559
[58] Glozak, M. A., *et al.*, *Gene* **2005**, *363*, 15
[59] Yang, X. J., *et al.*, *Mol Cell* **2008**, *31*, 449
[60] Allis, C. D., *et al.*, *Cell* **2007**, *131*, 633
[61] Kimura, A., *et al.*, *J Biochem* **2005**, *138*, 647
[62] Yang, X. J., *Bioessays* **2004**, *26*, 1076
[63] Mujtaba, S., *et al.*, *Oncogene* **2007**, *26*, 5521
[64] Berndsen, C. E., *et al.*, *Curr Opin Struct Biol* **2008**, *18*, 682
[65] Hodawadekar, S. C., *et al.*, *Oncogene* **2007**, *26*, 5528
[66] Wang, L., *et al.*, *Curr Opin Struct Biol* **2008**, *18*, 741
[67] Gregoretti, I. V., *et al.*, *J Mol Biol* **2004**, *338*, 17
[68] Yang, X. J., *et al.*, *Oncogene* **2007**, *26*, 5310
[69] Hernick, M., *et al.*, *Arch Biochem Biophys* **2005**, *433*, 71
[70] North, B. J., *et al.*, *Genome Biol* **2004**, *5*, 224
[71] Hassa, P. O., *et al.*, *Microbiol Mol Biol Rev* **2006**, *70*, 789
[72] Yang, X. J., *et al.*, *Nat Rev Mol Cell Biol* **2008**, *9*, 206
[73] Lee, K. K., *et al.*, *Nat Rev Mol Cell Biol* **2007**, *8*, 284
[74] Seet, B. T., *et al.*, *Nat Rev Mol Cell Biol* **2006**, *7*, 473
[75] Taverna, S. D., *et al.*, *Nat Struct Mol Biol* **2007**, *14*, 1025
[76] Li, B., *et al.*, *Cell* **2007**, *128*, 707
[77] Berger, S. L., *Nature* **2007**, *447*, 407
[78] Bhaumik, S. R., *et al.*, *Nat Struct Mol Biol* **2007**, *14*, 1008
[79] Ruthenburg, A. J., *et al.*, *Nat Rev Mol Cell Biol* **2007**, *8*, 983

[80] Sims, R. J., 3rd*, et al.*, *Nat Rev Mol Cell Biol* **2008**, *9*, 815
[81] Klier, H.*, et al.*, *Biochim Biophys Acta* **1996**, *1280*, 251
[82] Polevoda, B.*, et al.*, *EMBO J* **1999**, *18*, 6155
[83] Takakura, H.*, et al.*, *J Biol Chem* **1992**, *267*, 5442
[84] Kimura, Y.*, et al.*, *J Biol Chem* **2000**, *275*, 4635
[85] Kimura, Y.*, et al.*, *Arch Biochem Biophys* **2003**, *409*, 341
[86] Veenstra, T. D.*, et al.*, *Electrophoresis* **2004**, *25*, 1278
[87] Helsens, K.*, et al.*, *Mol Cell Proteomics* **2008**, *7*, 2364
[88] Gevaert, K.*, et al.*, *Nat Biotechnol* **2003**, *21*, 566
[89] Gevaert, K.*, et al.*, *Anal Biochem* **2005**, *345*, 18
[90] Van Damme, P.*, et al.*, *BMC Proc* **2009**, *3 Suppl 6*, S6
[91] Gevaert, K.*, et al.*, *Biochim Biophys Acta* **2006**, *1764*, 1801
[92] Staes, A.*, et al.*, *Proteomics* **2008**, *8*, 1362
[93] McDonald, L.*, et al.*, *Nat Protoc* **2006**, *1*, 1790
[94] McDonald, L.*, et al.*, *Nat Methods* **2005**, *2*, 955
[95] Boersema, P. J.*, et al.*, *Nat Protoc* **2009**, *4*, 484
[96] Kleifeld, O.*, et al.*, *Nat Biotechnol* **2010**, *28*, 281
[97] Yamaguchi, M.*, et al.*, *Rapid Commun Mass Spectrom* **2008**, *22*, 3313
[98] Yamaguchi, M.*, et al.*, *Rapid Commun Mass Spectrom* **2007**, *21*, 3329
[99] Auf dem Keller, U.*, et al.*, *Mol Cell Proteomics* **2010**
[100] Mikami, T.*, et al.*, *Anal Chem* **2007**, *79*, 7910
[101] Zhang, X.*, et al.*, *J Proteomics* **2009**, *73*, 240
[102] Wolters, D. A.*, et al.*, *Anal Chem* **2001**, *73*, 5683
[103] Beausoleil, S. A.*, et al.*, *Proc Natl Acad Sci USA* **2004**, *101*, 12130
[104] Gruhler, A.*, et al.*, *Mol Cell Proteomics* **2005**, *4*, 310
[105] Pinkse, M. W.*, et al.*, *J Proteome Res* **2008**, *7*, 687
[106] Thingholm, T. E.*, et al.*, *Nat Protoc* **2006**, *1*, 1929
[107] Villen, J.*, et al.*, *Nat Protoc* **2008**, *3*, 1630
[108] Van Hoof, D.*, et al.*, *Cell Stem Cell* **2009**, *5*, 214
[109] Bodenmiller, B.*, et al.*, *Mol Syst Biol* **2007**, *3*, 139
[110] Lemeer, S.*, et al.*, *Curr Opin Chem Biol* **2009**, *13*, 414
[111] Dormeyer, W.*, et al.*, *J Proteome Res* **2007**, *6*, 4634
[112] van den Toorn, H. W. P.*, et al.*, *J Proteom Bioinform* **2008**, *1*, 379
[113] Swaney, D. L.*, et al.*, *Nat Methods* **2008**, *5*, 959
[114] Taouatas, N.*, et al.*, *Nat Methods* **2008**, *5*, 405

[115] Taouatas, N., *et al.*, *J Proteome Res* **2010**
[116] Nonaka, T., *et al.*, *J Biochem* **1995**, *118*, 1014
[117] Nonaka, T., *et al.*, *J Biol Chem* **1997**, *272*, 30032
[118] Taouatas, N., *et al.*, *Mol Cell Proteomics* **2008**, *1*, 190
[119] Gauci, S., *et al.*, *Anal Chem* **2009**, *81*, 4493
[120] Helbig, A. O., *et al.*, *Mol Cell Proteomics* **2010**
[121] Mischerikow, N., *et al.*, *J Proteome Res* **2009**, *8*, 5020
[122] Wu, H. Y., *et al.*, *Anal Chem* **2008**, *80*, 6178
[123] Lin, Y. Y., *et al.*, *Cell* **2009**, *136*, 1073
[124] Zhang, K., *et al.*, *Proteomics* **2004**, *4*, 1
[125] Kim, J. Y., *et al.*, *Anal Chem* **2002**, *74*, 5443
[126] Trelle, M. B., *et al.*, *Anal Chem* **2008**, *80*, 3422
[127] Couttas, T. A., *et al.*, *J Proteome Res* **2008**, *7*, 2632
[128] Griffiths, J. R., *et al.*, *J Am Soc Mass Spectrom* **2007**, *18*, 1423
[129] Niggeweg, R., *et al.*, *Proteomics* **2006**, *6*, 41
[130] Qiang, L., *et al.*, *J Immunoassay Immunochem* **2005**, *26*, 13
[131] Komatsu, Y., *et al.*, *J Immunol Methods* **2003**, *272*, 161
[132] Iwabata, H., *et al.*, *Proteomics* **2005**, *5*, 4653
[133] Kim, S. C., *et al.*, *Mol Cell* **2006**, *23*, 607
[134] Zhang, J., *et al.*, *Mol Cell Proteomics* **2009**, *8*, 215
[135] Zhao, S., *et al.*, *Science*, *327*, 1000
[136] Choudhary, C., *et al.*, *Science* **2009**, *325*, 834
[137] Wang, Q., *et al.*, *Science*, *327*, 1004
[138] Yu, B. J., *et al.*, *J. Microbiol. Biotechnol.* **2008**, *18*, 1529
[139] Starai, V. J., *et al.*, *Science* **2002**, *298*, 2390
[140] Starai, V. J., *et al.*, *J Mol Biol* **2004**, *340*, 1005
[141] Schwer, B., *et al.*, *Proc Natl Acad Sci USA* **2006**, *103*, 10224
[142] Hallows, W. C., *et al.*, *Proc Natl Acad Sci USA* **2006**, *103*, 10230
[143] Schwer, B., *et al.*, *Cell Metab* **2008**, *7*, 104
[144] Lin, S. J., *et al.*, *Curr Opin Cell Biol* **2003**, *15*, 241
[145] Longo, V. D., *et al.*, *Cell* **2006**, *126*, 257
[146] Saunders, L. R., *et al.*, *Oncogene* **2007**, *26*, 5489
[147] Chen, Y., *et al.*, *Mol Cell Proteomics* **2007**, *6*, 812
[148] Liu, B., *et al.*, *J Biol Chem* **2009**, *284*, 32288
[149] Zhang, K., *et al.*, *J Proteome Res* **2009**, *8*, 900

[150] Cheng, Z.*, et al.*, *Mol Cell Proteomics* **2009**, *8*, 45
[151] Garrity, J.*, et al.*, *J Biol Chem* **2007**, *282*, 30239
[152] Hardman, M.*, et al.*, *Anal Chem* **2003**, *75*, 1699
[153] Li, Y.*, et al.*, *Int J Mass Spectrom* **2009**, *281*, 24
[154] Olsen, J. V.*, et al.*, *Nat Methods* **2007**, *4*, 709
[155] Ibrahim, Y.*, et al.*, *Anal Chem* **2007**, *79*, 7845
[156] McAlister, G. C.*, et al.*, *J Proteome Res* **2008**, *7*, 3127
[157] Garcia, B. A.*, et al.*, *Curr Opin Chem Biol* **2007**, *11*, 66
[158] Thompson, P. R.*, et al.*, *Nat Struct Mol Biol* **2004**, *11*, 308

Chapter 3

In-depth profiling of post-translational modifications of the related transcription factor complexes TFIID and SAGA

Nikolai Mischerikow[1,2], Gianpiero Spedale[2,3], A. F. Maarten Altelaar[1,2], H. T. Marc Timmers[2,3], W. W. M. Pim Pijnappel[2,3], and Albert J. R. Heck[1,2,4]

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

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

[3] Department of Molecular Cancer Research, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

[4] Centre for Biomedical Genetics, Padualaan 8, 3584 CG Utrecht, The Netherlands

Supplementary material referred to in this chapter can be found accompanying the publication of this work in *Journal of Proteome Research*, 2009, volume 8, issue 11, pages 5020-5030.

1 Summary

This chapter describes the profiling of post-translational modifications (PTMs) of the general transcription factors (GTFs) TFIID and SAGA from *Saccharomyces cerevisiae*. A multi-protease approach using Trypsin, Chymotrypsin and Glu-C in conjunction with high resolution, high mass accuracy mass spectrometry (MS) allowed the coverage most TFIID and SAGA subunit sequences to near completion, with peptide identifications established at a false discovery rate (FDR) of < 1%. The analysis resulted in the mapping of 118/102 unique phosphorylated and 54/61 unique lysine acetylated sites to TFIID/SAGA. The SAGA subunits Sgf73 and Spt7 displayed a particularly large number of lysine acetylation sites with peculiar clustering. Spectral counting showed that the common subunit TAF5 is phosphorylated to a greater and acetylated to a lesser extent in SAGA than in TFIID. Finally, the analysis provides evidence for the existence of a truncated form of Spt7 related to the SAGA-like complex SLIK.

2 Introduction

Proteome-wide studies of protein-protein interactions in *Saccharomyces cerevisiae* have substantiated the view that most proteins are organized into multimeric protein complexes [1,2]. These studies also show that a large set of proteins can be assembled into several different protein complexes. For instance, the TATA-box binding protein (TBP) associated factors (TAFs) TAF5, TAF6, TAF9, TAF10 and TAF12 are known to be present in both the basal transcription factor TFIID and the histone acetyl transferase complex SAGA [3-7]. However, it is still poorly understood how shared subunits incorporate into one or the other protein complex. Both TFIID and SAGA are large and heterogeneous complexes, containing 15 and 19 different subunits, respectively. Although most of the subunits of the TFIID and SAGA complexes have been known for several years, proteomics approaches still identified new SAGA subunits including Ubp8, Sgf73 and Sgf29 as well as Sgf11 [8] [9,10]. At present, the view is that besides these four subunits and the five shared TAFs, SAGA canonically contains the SPT class gene products Spt3, Spt7, Spt8 and Spt20, the ADA class gene products Ada1, Ada2, Ada3, Gcn5, and Tra1 and Sus1 [11-13]. TFIID is composed of TBP as well as the 14 TAFs TAF1-TAF14 [14]. Electron microscopy combined with immunolabeling suggests that certain subunits of TFIID and SAGA are likely present in more than one copy [15,16]. Functionally, both TFIID and SAGA are involved in regulated transcription of RNA polymerase II dependent

genes [17,18]. With Gcn5 the SAGA complex also harbors one of the major nuclear histone acetyl transferase activities that acetylates histones H3 and H4. Closely linked to the SAGA complex is the SAGA-like complex SLIK or SALSA [19,20]. SLIK contains all SAGA subunits except Spt8 which together with Spt3 is involved in TBP binding [21]. Instead of full length Spt7, SLIK contains a shorter form of Spt7 that lacks a C-terminal portion of the protein which is required for Spt8 binding. The site of truncation has been mapped in a deletion study to somewhere between amino acids 1125 and 1151 [22].

In general, little is known about the PTMs of SAGA and TFIID. The publicly available data are not comprehensive as they are derived from proteome-wide studies. To increase confidence about the PTM state of SAGA and TFIID, we decided to analyze both complexes by an in-depth liquid chromatography-coupled MS- (LC-MS) based peptide sequencing approach. A viable approach for this is to use various proteases in parallel to generate different sets of peptides from the same sample [23-28]. The accompanying variation of physicochemical properties of the generated peptides usually increases protein sequence coverage, as the fraction of peptides that are successfully sequenced is constrained for each peptide set by peptide solubility, ionizability and fragmentation behavior. The approach can also benefit the identification of PTMs if they can be established from spectra originating from different proteolytic peptides covering the same sequence stretch. A major requirement to approach comprehensive sequence coverage at the protein level is the reduction of the complexity of the peptide mixture provided to the mass spectrometer. This is generally achieved by using protein and/or peptide pre-fractionation techniques in the first separation dimension and reversed-phase (RP) LC-MS in a second dimension [29] [30-32] [29]. As a result of extensive pre-fractionation, peptides are more frequently sequenced, thus enabling spectral-counting based relative quantification [33-35].

This study describes an in-depth analysis of SAGA and TFIID purified from *Saccharomyces cerevisiae* and digested using Trypsin, Chymotrypsin or Glu-C. The multi-protease approach was combined with extensive pre-fractionation by SDS polyacrylamide gel electrophoresis (SDS PAGE) or strong cation exchange (SCX) to achieve separation orthogonality to the RP separation used in LC-MS. This strategy enabled the generation of nearly complete sequence maps of most subunits and revealed a number of phosphorylated and acetylated sites within TFIID and SAGA, including some present in both complexes. Spectral counting hinted at differential abundance of PTMs on the shared subunit TAF5. The study also identified a C-terminal portion of Spt7 which was instru-

mental to the mapping of the truncation site within Spt7 that is potentially linked to a transition from SAGA to SLIK.

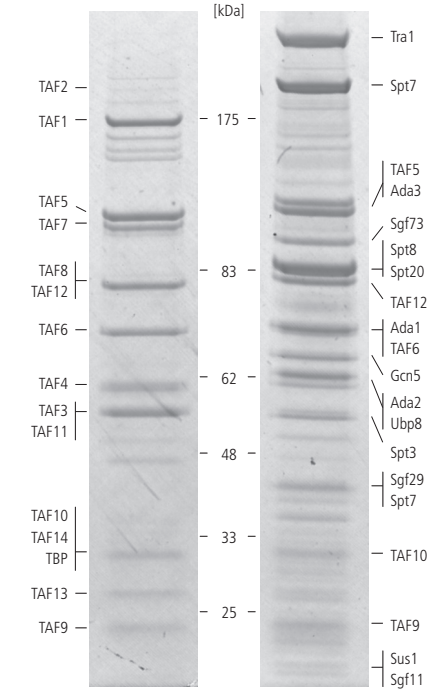
3 Results and discussion

To comprehensively map PTMs of SAGA and TFIID, especially with respect to the TAFs that occur in both complexes, both SAGA and TFIID were tandem affinity purified using Spt7 and TAF1 as tagged subunits, respectively, and separated by SDS PAGE, digested with Trypsin and analyzed by high-resolution, high mass accuracy LC-MS. As concluded from the gel pattern, both purifications were strongly enriched for their respective canonical subunits (Figure 1A). The high degree of enrichment of both complexes was also reflected by the high numbers of spectra detected for the SAGA/TFIID subunits, ranging in the hundreds for the larger components, as well as the high sequence coverage (Figures 1B and 2). More importantly, the amount of TFIID co-purified with SAGA and *vice versa* the amount of SAGA co-purified with TFIID, was around 5% as estimated from normalized spectral counts (Figure 2). This is particularly important for a differential comparison of the PTM state of the subunits present in both complexes.

Although the median of sequence coverage of SAGA and TFIID subunits was already above 70% from the tryptic digests, we also digested both purifications with both Lys-C followed by Chymotrypsin and Lys-C followed by Glu-C to further enhance sequence coverage and to independently verify PTMs. Lys-C was chosen because of its ability to cleave under strong denaturing conditions [36]. Indeed, when projecting the identified peptides from all three analyses onto their respective subunits, around 90% (median) of primary sequence could be covered (Figures 1B and 2).

The other gain of using different proteases is an improved confidence in PTM identification. In some cases, PTMs could only be identified from Chymotrypsin- or Glu-C-derived peptides. TAF1, for example, was found to be phosphorylated at multiple residues. Phosphorylated S34 and S189 were exclusively identified on peptides with a chymotryptic N-terminus because the corresponding tryptic peptides are likely too large for detection and fragmentation. In other cases, a PTM could be identified from different proteolytic peptides with different precursor m/z and different fragment ion spectra patterns, which increases confidence about the identity of the observed PTMs. For example, phosphorylated TAF1 T355 was detected in all three digests and could be inferred from a fully Lys-C derived peptide of 22 residues. In addition, phosphorylated

A



B

TFIID		
Subunit	MW [kDa]	Sequence coverage
TAF2	163	65
TAF1	121	86
TAF5	89	94
TAF7	68	66
TAF12	61	96
TAF8	58	55
TAF6	58	92
TAF4	42	92
TAF11	41	96
TAF3	40	91
TAF14	27	59
TBP	27	49
TAF10	23	93
TAF13	19	74
TAF9	17	96

SAGA		
Subunit	MW [kDa]	Sequence coverage
Tra1	435	84
Spt7	153	91
TAF5	89	93
Ada3	79	90
Sgf73	73	79
Spt20	68	80
Spt8	66	89
TAF12	61	93
TAF6	58	89
Ubp8	55	92
Ada1	55	87
Ada2	51	89
Gcn5	51	92
Spt3	39	90
Sgf29	30	88
TAF10	23	88
TAF9	11	100
Sgf11	17	84
Sus1	11	96

Figure 1. A, gradient SDS PAGE gel of tandem affinity purified TFIID and SAGA with the detected subunits indicated. B, overview of detected TFIID/SAGA subunits with their sequence coverage (in %) when adding up the individual sequence coverages from each digest.

T355 was also detected on a smaller peptide of 10 amino acids with a chymotryptic N-terminus. The large peptide contained 4 serine and 4 threonine residues so the small peptide significantly increased the confidence in the site assignment. Other examples of novel phosphorylation sites of SAGA and TFIID subunits, not previously detected in large-scale phosphoproteomic datasets [37-40], were phosphorylated residues of Spt7 (S100, S341, T367, S636, S637), Spt8 (S25, S372, S385, S409, S506, S545), Ada3 (S134) and TAF4 (S301, S321, S338), illustrating the analysis depth reached here. In total for TFIID/SAGA, we detected 481/624 spectra related to serine or threonine phosphorylated peptides, and 69/232 peptide queries related to lysine acetylated peptides (see supplementary material). These spectra correspond to 183/200 phosphopeptides and 56/108 lysine acetylated peptides (see supplementary material), mapping 118/102 serine or threonine phosphorylation sites and 54/61 lysine acetylation sites of TFIID/SAGA. Information on these peptides can be found in the supplementary material.

TFIID purification

Protein	Assigned spectra			Spectral abundance		
	T	C	G	T	C	G
TAF2	85	145	53	9E-6	3E-5	1E-5
TAF1	990	648	344	1E-4	2E-4	1E-4
TAF5	713	523	297	1E-4	2E-4	1E-4
TAF7	465	315	141	1E-4	1E-4	7E-5
TAF12	529	155	208	1E-4	8E-5	1E-4
TAF8	36	15	11	1E-5	8E-6	7E-6
TAF6	396	383	111	1E-4	2E-4	7E-5
TAF4	439	264	211	2E-4	2E-4	2E-4
TAF11	256	311	184	1E-4	2E-4	2E-4
TAF3	417	195	68	2E-4	2E-4	6E-5
TAF14	5	31	21	3E-6	4E-5	3E-5
TBP	24	21	21	2E-5	2E-5	3E-5
TAF10	115	171	70	8E-5	2E-4	1E-4
TAF13	103	41	17	9E-5	7E-5	3E-5
TAF9	96	113	54	9E-5	2E-4	1E-4
Tra1	9	26	26	3E-7	2E-6	2E-6
Spt7	23	17	10	3E-6	3E-6	2E-6
Ada3	2	7	2	4E-7	3E-6	9E-7
Sgf73	0	5	6	0	2E-6	3E-6
Spt20	3	4	9	7E-7	2E-6	5E-6
Spt8	1	0	0	3E-7	0	0
Ubp8	1	1	1	3E-7	6E-7	6E-7
Ada1	12	0	1	4E-6	0	7E-7
Ada2	4	1	3	1E-6	6E-7	2E-6
Gcn5	0	7	4	0	4E-6	3E-6
Spt3	5	6	0	2E-6	5E-6	0
Sgf29	0	3	5	0	3E-6	6E-6
Sgf11	0	0	0	0	0	0
Sus1	0	1	0	0	3E-6	0

SAGA purification

Protein	Assigned spectra			Spectral abundance		
	T	C	G	T	C	G
Tra1	669	1043	933	2E-5	2E-5	4E-5
Spt7	1162	973	695	8E-5	5E-5	8E-5
TAF5	405	388	381	5E-5	4E-5	8E-5
Ada3	285	225	286	4E-5	2E-5	7E-5
Sgf73	355	117	251	5E-5	1E-5	6E-5
Spt20	450	172	275	7E-5	2E-5	7E-5
Spt8	614	840	591	1E-4	1E-4	2E-4
TAF12	461	88	210	8E-5	1E-5	6E-5
TAF6	334	265	140	6E-5	4E-5	4E-5
Ubp8	98	101	169	2E-5	2E-5	6E-5
Ada1	301	162	158	6E-5	3E-5	5E-5
Ada2	195	77	126	4E-5	1E-5	4E-5
Gcn5	257	122	175	5E-5	2E-5	6E-5
Spt3	161	162	63	4E-5	3E-5	3E-5
Sgf29	221	57	122	8E-5	2E-5	8E-5
TAF10	99	128	98	5E-5	5E-5	8E-5
TAF9	69	93	65	4E-5	4E-5	7E-5
Sgf11	13	20	22	1E-5	1E-5	4E-5
Sus1	7	56	25	7E-6	4E-5	4E-5
TAF2	0	4	12	0	2E-7	1E-6
TAF1	2	28	31	2E-7	2E-6	5E-6
TAF7	11	13	14	2E-6	2E-6	4E-6
TAF8	4	1	0	7E-7	1E-7	0
TAF4	209	7	20	5E-5	1E-6	9E-6
TAF11	21	11	18	5E-6	2E-6	8E-6
TAF3	34	2	3	9E-6	4E-7	1E-6
TAF14	6	8	8	2E-6	2E-6	5E-6
TBP	3	5	8	1E-6	2E-6	5E-6
TAF13	5	1	3	3E-6	4E-7	3E-6

Figure 2. Abundance of SAGA subunits in the TFIID purification (left panel) and TFIID subunits in the SAGA purification (right panel), estimated in terms of the number of assigned spectra as well as in terms of normalized spectral abundance. T, C and G indicate the three analyses using Trypsin, Chymotrypsin and Glu-C, respectively.

TAF5, TAF6, TAF9, TAF10 and TAF12 are present in both SAGA and TFIID and some of them were also found to be phosphorylated (Figure 3). Three of these stretches, which in two cases contain multiple adjacent serine or threonine residues, namely TAF5 S411, S414 and S415, TAF10 S58, and TAF12 S286, S287, S288, S290 and T291, were identified by a large number of spectra, which allowed a estimation of the phosphorylation levels based on spectral counts. While the number of phosphorylated peptide queries assigned to TAF10 S58 (25 in SAGA versus 31 in TFIID) and to the phosphorylated sequence stretch in TAF12 (48 in SAGA versus 51 in TFIID) was found to be similar in both complexes, the phosphorylated stretch in TAF5 has a roughly 10-fold higher number of assigned phosphorylated spectra in SAGA (59 spectra) than in TFIID (6 spectra). In addi-

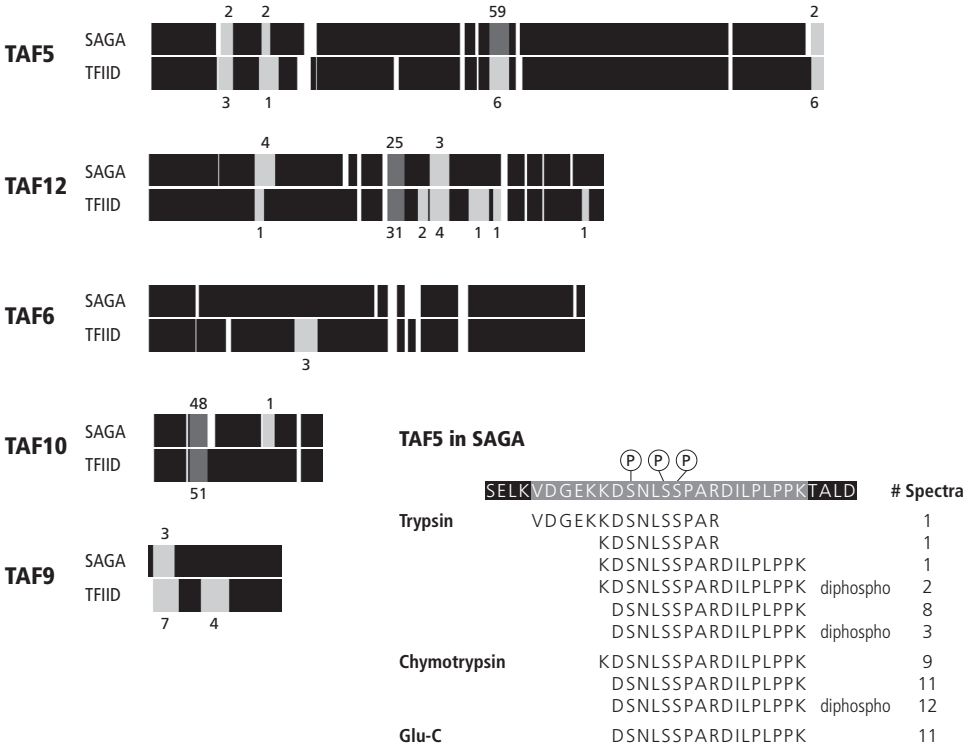


Figure 3. Cumulative sequence coverage maps of the TAFs present in both SAGA and TFIID. Black sequences stretches indicate protein sequence coverage by non-modified peptides, grey stretches indicate coverage by a larger number (>25, dark grey) or smaller number (<10, light grey) of phosphorylated peptides with the number of spectra indicated, and white indicates no sequence coverage. Using TAF5 from the SAGA purification as example, the lower right inset illustrates that the length of a grey stretch does not necessarily correlate with the number of phosphorylated residues, but rather with the length of the phosphorylated peptides that map this stretch.

tion, we found TAF5 to be differentially acetylated as well. In the cumulative data from the SAGA preparation, TAF5 K103 was identified as acetylated in only a single query, while in TFIID 23 spectra mapped this residue.

Both differential phosphorylation and acetylation of TAF5 could be confirmed by extracted ion chromatograms (Figure 4). These were generated from the analyses of the two TAF5-containing gel bands of the TFIID and SAGA preparation. In case of the phosphorylated stretch of TAF5, the most prominent (as judged by spectral counts, see inset of Figure 3) singly phosphorylated peptide DSNLSSPARDILPLPPK was found to be more abundant in the context of SAGA then in TFIID. For TAF5 K103, which is exclusively

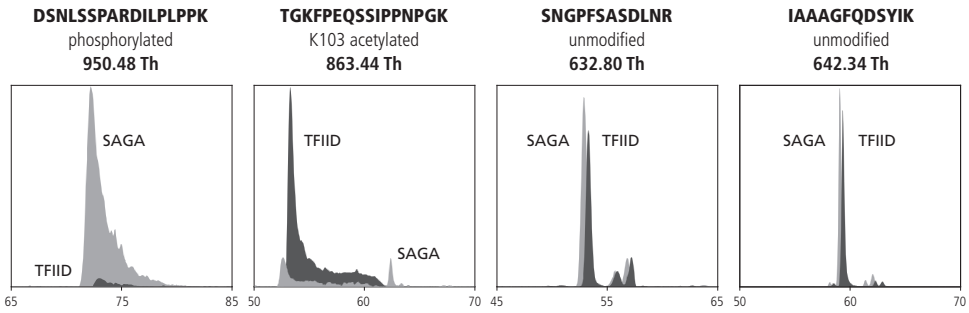


Figure 4. Extracted ion chromatograms of TAF5-derived peptides mapping the S411/S414/S415 phosphorylation (peptide DSNLSSPARDILPLPPK) and K103 acetylation (TGKFPEQSSIPPNPGK) correlate with the results obtained by spectral counting. Two unmodified peptides are plotted as control.

identified on the acetylated peptide TGKFPEQSSIPPNPGK, this relation was opposite. For an estimation of differences in between runs, several TAF5-derived peptides were selected from the pool of non-modified, fully cleaved peptides, of which two are shown in Figure 4 as well. All phosphorylation sites of TAF5, TAF12 and TAF10 have been reported previously in large-scale studies [37-40], but to our knowledge this is the first time that the phosphorylation state can be linked to the presence of the TAFs in either SAGA or TFIID, which is impossible in analyses at the lysate level without prior purification.

To test whether TAF5 phosphorylation is essential for the integrity or function of SAGA, we mutated TAF5 S411, 414 and 415 to alanine (alone or in combination) and tested the mutant TAF5 strains for known SAGA functions. Mutants in SAGA subunits are able to suppress Ty insertions (SPT phenotype) [22,41,42]. The SPT phenotype was tested by assessing growth of mutant TAF5 strains carrying the *his4-917*Δ and *lys2-17R2* alleles in medium lacking histidine and by assessing inhibition of growth in medium lacking lysine. In addition, SAGA mutants lack the ability to grow in medium containing alternative carbon sources like galactose and to grow in medium lacking inositol [22,41,42]. None of these growth phenotypes were observed in single, double, or triple mutant TAF5 strains, whereas the *spt3*Δ control displayed all of these phenotypes (supplementary material). This suggests that TAF5 phosphorylation at S411, S414 and S415 in the context of SAGA is not essential for these SAGA functions and for SAGA integrity. It remains possible that phosphorylation of these residues has alternative (for example gene-specific) functions not tested here, or that these phosphorylations are redundant

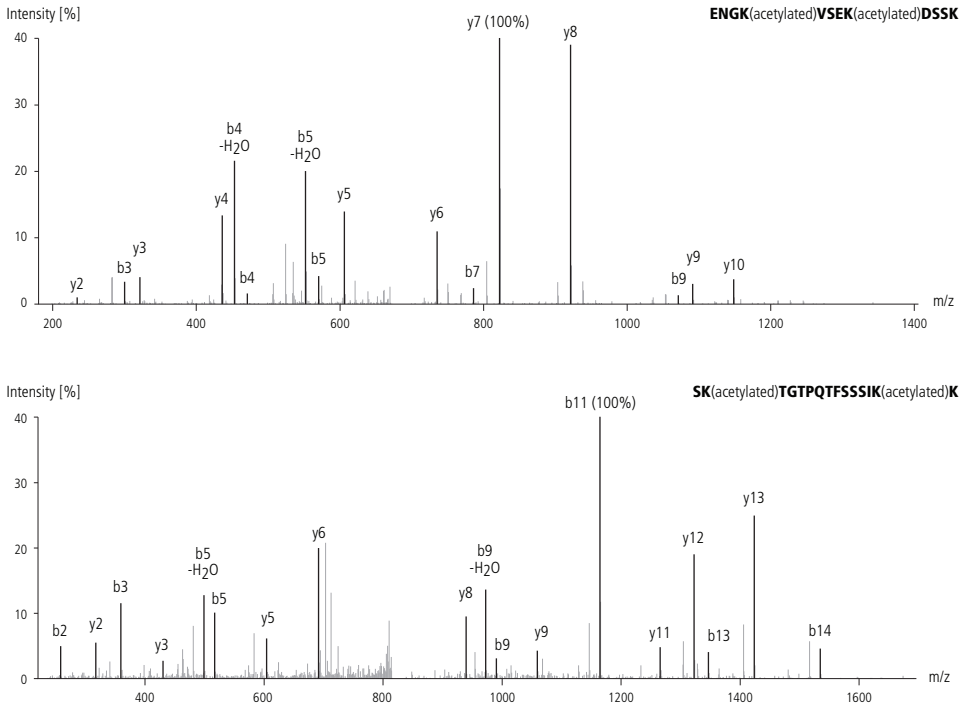


Figure 5. Fragment ion spectra of Spt7 doubly lysine acetylated peptides originating from Spt7 (upper spectrum) and Sgf73 (lower spectrum). Main fragment ion series and a few prominent water losses are annotated; many non-annotated peaks originate from additional neutral losses.

with other PTMs present on TAF5 or on other SAGA subunits.

The SAGA subunits Spt7 and Sgf73 were found to be significantly lysine acetylated. Although other SAGA subunits were found to be acetylated as well, Spt7 and Sgf73 stand out by the high number of spectra and unique peptides, as well as by the quality of the spectra, enabling confident site assignments (Figure 5). Moreover, the high degree of lysine acetylation of Spt7 and Sgf73 could be confirmed by Western blotting using a pan-specific anti-acetyl-lysine antibody.

We found that Spt7 is acetylated at multiple lysine residues that group to mainly two sequence stretches adjacent to the bromodomain (Figure 6A). In the stretch that precedes the bromodomain, between K379 and K410, three out of five lysine residues present were detected in the acetylated form (K379, K400, K410), and in the sequence stretch that follows the bromodomain, between K584 and K680, notably ten out of seventeen lysine residues (K584, K599, K603, K607, K610, K629, K633, K647, K667, K680) were

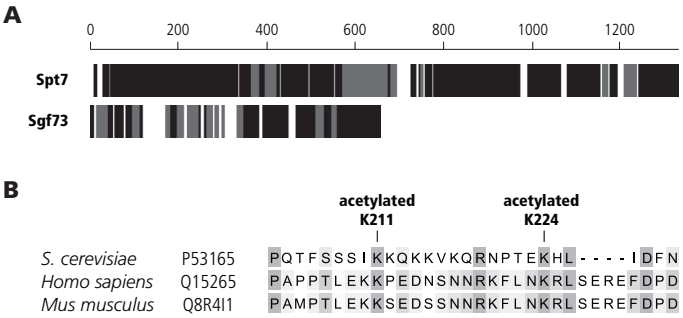


Figure 6. A, cumulative sequence coverage maps of Spt7 and Sgf73 with black indicating unmodified sequence, grey indicating sequence containing acetylated lysine residues, and white no coverage. B, sequence alignment of Sgf73 to its mouse and human homologues.

detected in their acetylated state. In the case of K599, K603, K607 and K610, lysine residues are so close together that two acetylated residues could be detected on a single peptide, indicating that a single Spt7 molecule can have at least two acetylated lysine residues at once. Sgf73 was found to be acetylated at multiple lysine residues (K171, K199, K211, K224, K288, K300) in a region adjacent to the second Zn-finger motif. As with Spt7, two lysine residues (K199 and K211) were found to be acetylated on a single peptide, showing that also a single Sgf73 molecule can be multiply acetylated. Moreover, phosphorylated T212 could be detected together with acetylated K224 on the same peptide.

From our data alone, the possible implication of this significant lysine acetylation for Spt7 or Sgf73 function can just be speculated, although it is well known that lysine acetylation of transcription factors generally affects protein function in very diverse ways [43]. One possibility for Spt7 acetylation could be the regulation of the binding affinity of the Spt7 or the Gcn5 bromodomain, similar to the auto-regulative effect that the acetylation of Rsc4 K24 has on the affinity of the Rsc4 bromodomain [44]. Concerning Sgf73, a regulatory effect on Gcn5 activity can be envisaged too, as Sgf73 has been shown to regulate SAGA and SLIK acetyl transferase activity [45]. By comparative genome analysis we could not show a conservation of any of the lysine residues found to be acetylated in Spt7 throughout fungal Spt7 homologues (data not shown). However, in Sgf73, which is generally a much more conserved protein than Spt7, K211 and K224 are homologous to K321 and K334 in human Ataxin-7 and K307 and K320 in mouse

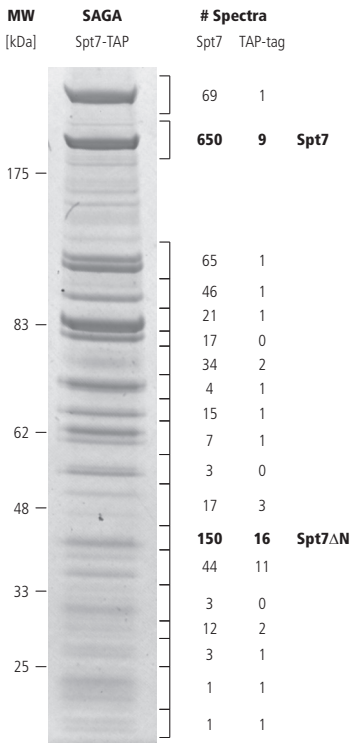
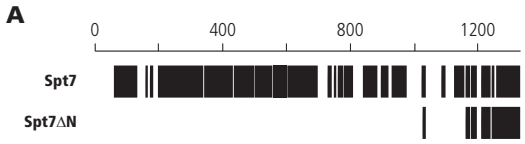


Figure 7. Spectra mapping Spt7 and the tandem affinity purification- (TAP)-tag identified in the purification of SAGA via Spt7-TAP. Next to the gel the cutting scheme for the in-gel digestion, and the number of spectra mapping Spt7 and the TAP-tag that were identified in each gel piece. The second maximum of Spt7- and tag-derived spectra between 33 and 48 kDa is annotated as Spt7ΔN.

Ataxin-7 (Figure 6B). The function of this region is unknown; it is not implicated in the binding of Sgf73 to SAGA [46].

Spt7, besides being assembled into SAGA in its full-length form, can also be present in a truncated form that lacks a C-terminal part of approximately 200 amino acids as revealed by a deletion study [22]. This remaining N-terminal part of Spt7, which we here name Spt7ΔC, is uniquely present in the SLIK (or SALSA) complex [19,20]. We co-purified the putative complementary piece to Spt7ΔC, which we name Spt7ΔN, along with the SAGA complex. We conclude this in first instance from the shape of the distribution of the number of Spt7-derived peptides over the gel lane (Figure 7). This distribution has two peaks, one in the MW region of full-length Spt7 at around 180 kDa and another one, unexpectedly, in the low MW region at around 40 kDa. Second, in contrast to the 650 Spt7-derived peptides identified in the 180 kDa MW region of the gel, which map the full-length Spt7 sequence, the 150 Spt7-derived peptides from the 40 kDa region of the gel exclusively map the C-terminal 174 amino acids of Spt7 (Figure 8A). Third, if the hypothetical Spt7ΔN was co-purified with the SAGA complex, it had to carry the tandem affinity purification- (TAP-) tag since the yeast cells express only Spt7 TAP-tagged at its C-terminus. In this case, peptides derived from the TAP-tag of Spt7ΔN should be detected equally well as peptides originating from full-length Spt7. This hypothesis was tested by observing the shape of the distribution of the number of TAP-tag derived peptides over the gel lane (Figure 7). The distribution has two peaks that coincide perfectly with the two peaks observed for the distribution of the number of Spt7-derived peptides over the gel lane. Taken this evidence together, we conclude that we have indeed co-purified Spt7ΔN.



B

Name	N-terminal sequence	# Spectra
Spt7_C192	LLTTQFQTVDGETK...	1
Spt7_C191	LTTQFQTVDGETK...	28
Spt7_C190	TTQFQTVDGETK...	10
Spt7_C189	TQFQTVDGETK...	1

Figure 8. A, peptide identifications made in the Spt7 and Spt7ΔN gel piece as indicated in Figure 7. All Spt7-derived peptides except one in Spt7ΔN map exclusively to the last 175 amino acids. B, mapping of the truncation site of Spt7 as revealed by a database search with consecutive N-terminally truncated forms of Spt7. The number of N-terminal peptides unique to each truncated form clearly peaks at Spt7_C191 (amino acids 1142-1332) and Spt7_C190 (1143-1332) with only a few semitryptic peptides spectra assigned to shorter or longer truncated forms.

As the length of Spt7ΔN is in very good agreement with the one proposed from the genetic deletion study mentioned above, we decided to try mapping the N-terminus of Spt7ΔN to the exact residue. For this purpose, we generated a database that contained 313 potential C-terminal cleavage products of Spt7 ranging between 138 amino acids (Spt7 residues 1195-1332, named Spt7_C138) and 450 (Spt7 residues 883-1332,

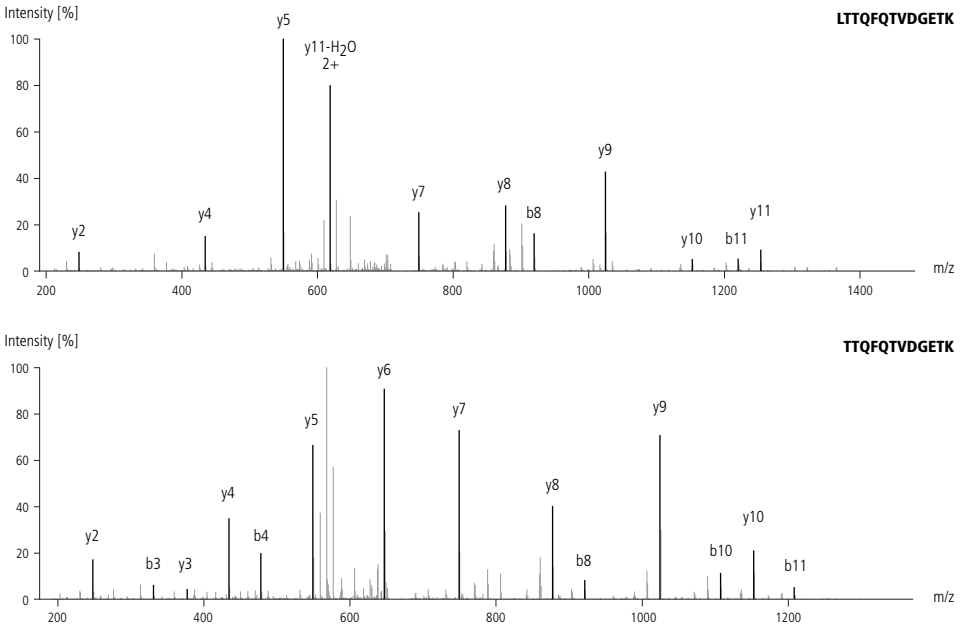


Figure 9. Fragment ion spectra of semi-tryptic peptides originating from Spt7_C191 (upper spectrum) and Spt7_C190 (lower spectrum). Only main fragment ion series and a few prominent water losses are annotated.

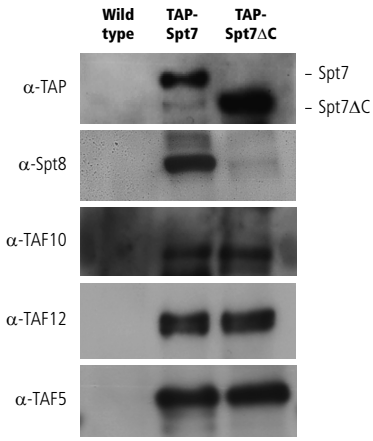


Figure 10. Analysis of immunoprecipitates from wild type, TAP-Spt7 and TAP-Spt7ΔC strains with antibodies against SAGA subunits (indicated on the left) shows that the truncation of Spt7 at the mapped cleavage site results in specific loss of Spt8 from the SAGA complex.

named Spt7_C450) and all differing by a single N-terminal amino acid in length. We searched the in-gel digestion dataset against this database for peptides with a non-tryptic N-terminus and indeed found a significant number of N-terminally semi-tryptic peptides that could be uniquely assigned to Spt7_C191 and Spt7_C190 (Figure 8B). These unambiguous peptide identifications (with 28 and 10 spectra, respectively) were found to be well above the normal background of 1-2 semi-tryptic peptides that were sometimes assigned to other Spt7 variants in the database. Representative fragment ion spectra characterizing N-terminus of the truncated forms Spt7_C191 and Spt7_C190 are shown in Figure 9, showing that the terminus of Spt7ΔN is likely either Spt7 L1142 or T1143. This result strongly hints to a protease that specifically processes Spt7 by cleaving it between L1141 and L1142 or L1142 and T1143.

To test whether cleavage of Spt7 at this position results in formation of SLIK rather than SAGA, we generated N-terminally TAP tagged strains of full-length Spt7 (TAP-Spt7) or to Spt7 truncated at the C-terminus at L1141 (TAP-Spt7ΔC) as depicted in Figure 10A. TAP followed by immunoblotting confirmed that the C-terminal truncation at L1141 indeed results in a specific loss of Spt8 from SAGA, indicative of formation of SLIK (Figure 10B). In addition during SDS PAGE, SptΔC present in TAP-Spt7ΔC migrated at the same position as endogenous Spt7ΔC. As expected, shared SAGA/SLIK subunits including Spt7, TAF10, TAF12 and TAF5 were present at similar levels. The genetic truncation of Spt7 at L1141 results in SLIK formation and therefore validates the identified cleavage site.

The cleavage site as well as a large sequence stretch preceding it is highly conserved in fungal Spt7 homologues (Figure 11) and also shows conservation in human STAF65γ [47]. As the formation of Spt7ΔC or its incorporation into SAGA is an important step in the transition from SAGA to SLIK, a protease catalyzing this step would have an important regulatory function. Previous studies suggest that this protease might be an

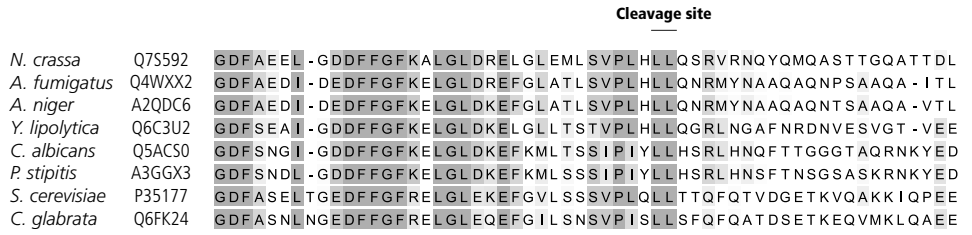


Figure 11. Sequence alignment of Spt7 to its fungal homologues.

upstream element of the retrograde response pathway [5]. Another possibility is that Spt7ΔN has a transcriptional role by itself, as it represents the portion of Spt7 that interacts with Spt8 and through this subunit with TBP. Our results should aid in the further verification of these hypotheses.

4 Experimental procedures

Cell extraction. Cell extracts were prepared from *Saccharomyces cerevisiae* strains FY2031 (genotype *MAT**a** HA-SPT7-TAP::TRP1 ura3Δ0 leu2Δ1 trp1Δ63 his4-917Δ lys2-173R2*) [22] and SC1064 (genotype *MAT**a** TAF1-TAP::URA3 ade2 arg4 leu2-3,112 trp1-289 ura3-52*) [1] as described previously with minor modifications [48]. In brief, cells were harvested from suspension culture in yeast extract peptone dextrose (YEPD) medium during early log phase, disrupted by glass bead homogenization in extraction buffer (20 mM HEPES-NaOH pH 8.0, 150 mM sodium chloride, 10% (v/v) glycerol, 0.1% (v/v) Tween-20, 1 µg/mL pepstatin, 0.5 µg/mL leupeptin, 2 µg/mL aprotinin, 1 mM PMSF, 1% (v/v) phosphatase inhibitor cocktails #1 and #2 (Sigma-Aldrich), 10 mM sodium butyrate) and cleared by ultracentrifugation.

Tandem affinity purifications. TAPs of the were carried out as previously published with the following specifications [49]. 200 µL 50% (v/v) immunoglobulin G sepharose (Sigma-Aldrich) were incubated for 2 h with 10 mL cell extract and subsequently washed with 30 mL extraction buffer and 10 mL cleavage buffer (10 mM Tris-HCl pH 8.0, 150 mM sodium chloride, 0.5 mM EDTA pH 8.0, 0.1% (v/v) Tween-20, 1 mM DTT). The protein A moiety of the tag was cleaved off with 10 µL TEV protease (Invitrogen) in 1 mL cleavage buffer for 2 h. Binding to calmodulin agarose (Stratagene) for 1 h in binding buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM magnesium acetate, 1 mM imidazole, 2 mM calcium chloride, 10% (v/v) glycerol, 0.1% (v/v) Tween-20, 10 mM

mercaptoethanol), followed by washing with 30 mL binding buffer and elution in 300 µL elution buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM magnesium acetate, 1 mM imidazole, 2 mM EGTA pH 8.0, 10% (v/v) glycerol, 0.1% (v/v) Tween-20, 10 mM mercaptoethanol).

Protein pre-fractionation, digestion and desalting. Proteins were precipitated from the eluate with methanol/chloroform. The total protein content measured after precipitation was around 5 µg as determined by a micro-BCA assay (Pierce). Proteins were either reduced, alkylated and digested with Trypsin in the gel matrix after visualization by SDS PAGE (NuPAGE Novex Bis-Tris 4-12% gradient gel, Invitrogen), or digested in solution using Lys-C under denaturing conditions followed by digestion with Chymotrypsin or Glu-C (all proteases from Roche Applied Science). For the in solution digestion, the precipitate was suspended in 20 µL 50 mM Tris-HCl pH 8.0, 8 M urea and 2.5 mM DTT, incubated at 55°C for 10 min and alkylated by addition of 5 µL of 50 mM Tris-HCl pH 8.0, 8 M urea and 20 mM iodoacetamide. Subsequently, 20 µL diluent (50 mM Tris-HCl pH 8.0) were added followed by addition of 5 µL (0.5 µg) Lys-C and incubation at 37°C for 4 h. For the second digestion, 145 µL diluent and 5 µL (0.5 µg) Chymotrypsin or Glu-C were added and the mixture was incubated at 25°C for 18 h. The digest was dried in a vacuum centrifuge, solubilized in 20 µL 10% (v/v) formic acid and desalted using C18 membrane (Empore C18 extraction disk, 3M) packed into a pipette tip.

Peptide pre-fractionation by SCX. In solution digests were pre-fractionated by SCX, which was performed using a Shimadzu LC-9A binary pump with two ZORBAX BioSCX-Series II columns (50 mm x 0.8 mm, 3.5 µm particle size, Agilent Technologies) in series, connected to a SPD-6A UV-detector (Shimadzu) and a FAMOS autosampler (Dionex). The desalted digest was dissolved in 10 µL 10% (v/v) formic acid and loaded onto the columns at 50 µL/min solvent A (0.05% (v/v) formic acid in 1:4 acetonitrile : water) for 10 min, followed by linear gradient elution of 1.3% (v/v) 1/min solvent B (500 mM NaCl in solvent A) at 50 µL/min. A total of 50 SCX fractions (1 min each) were collected and dried in a vacuum centrifuge.

Liquid chromatography-coupled mass spectrometry. Vacuum dried digests and SCX fractions were dissolved in 20 µL 10% (v/v) formic acid and 10 µL of each sample was analyzed on a LTQ Orbitrap (Thermo Fisher Scientific). An Agilent 1200 series HPLC system was equipped with a 20 mm Aqua C18 (Phenomenex) trapping column (packed in house, 100 µm inner diameter, 5 µm particle size) and a 200 mm ReproSil-Pur C18-AQ (Dr. Maisch GmbH) analytical column (packed in house, 50 µm inner diameter, 3 µm

particle size). Trapping was performed at 5 µL/min solvent C (0.1 M acetic acid in water) for 10 min, and elution was achieved with 10 to 40% (v/v) solvent D (0.1 M acetic acid in 1:4 acetonitrile : water in either 22 or 35 min, followed by 38 to 100% (v/v) solvent D in 3 min and 100% solvent D for 2 min. The flow rate was passively split from 0.45 mL/min to 100 nL/min as previously described [50]. Nano-electrospray was achieved using a distally coated fused silica emitter (360 µm outer diameter, 20 µm inner diameter, 10 µm inner diameter of the emitter tip, New Objective) biased to 1.7 kV. The LTQ Orbitrap was operated in the data dependent mode to automatically switch between MS and MS/MS. Survey full scan MS spectra were acquired from m/z 150 to m/z 1500 in the Orbitrap with a resolution of 60,000 at m/z 400 after accumulation to a target value of 500,000 in the linear ion trap. The two most intense ions at a threshold of above 500 were fragmented in the linear ion trap using CID at a target value of 30,000.

Data processing and analysis. Spectra were processed with Bioworks 3.3 (Thermo Fisher Scientific) and the subsequent data analysis was carried out using Mascot 2.2.1 (Matrix Science) licensed in house. Mascot was set up to search the Saccharomyces genome database (SGD) with carbamidomethyl cysteine as fixed modification and the following variable modifications: oxidation of methionine, acetylation of lysine, phosphorylation of serine and threonine, methyl esterification of aspartic and glutamic acid (only for samples digested in gel) and carbamylation of the peptide N-terminal amino group (only for samples digested in solution). Up to two missed cleavages were allowed. The mass tolerance of the precursor ion was set to 5 ppm and that of fragment ions to 0.6 Da. In all six datasets, protein identifications were established at a significance threshold <0.01 which corresponds to individual Mascot ion scores >31 and false discovery rates <1% as determined by decoy database searching as implemented in Mascot. Calculation of sequence coverage was performed using SpectraMapper [26]. Fragment ion spectra were visualized using Scaffold 2.1.03 (Proteome Software). Fragment ion spectra and tables of fragment ion matches of peptides mapping modified amino acids which are explicitly discussed in the text can be found in the supplementary material. RAW data files are available on the public repository Tranche (https://proteomecommons.org/) using the following hash code: VYyYWAkbdUaOOrcKoz1eQ XHQSxxUHv1oJL6nFxmMB-1J0M9NgyFStOC9Fqp8BYx87Vx8XF6I7POD9/oZDFid5ZM XACM8AAAAAAACk1Q==

Spectral abundance was calculated similar to a method previously published by the Washburn group [33,34]. In detail, for each of the six digests, peaklists from all MS runs were combined and searched with Mascot. In the search results, the number of spectra assigned to a protein was divided by the molecular weight of the protein and the total

number of queries that were submitted. Extracted ion chromatograms were generated using Xcalibur 2.0.5 (Thermo Fisher Scientific). For every peptide precursor EIC the first two isotopes were considered using an m/z window of 0.02 or 0.03 Th. Gaussian smoothing was applied over nine m/z data points. Protein sequences of fungal Spt7 homologues found in the UniProt database (http://www.uniprot.org) were aligned with MUSCLE as implemented on the EBI website (http://www.ebi.ac.uk/tools/muscle). The full alignments of Spt7 and Sgf73 can be found in FASTA format in the supplementary material. Aligned residues were colored in JalView 2.4 (http://www.jalview.org) by conservation at a visibility threshold of 30 using BLOSUM62 homology definitions.

Analysis of TAF5 mutants and of truncated Spt7. Yeast strains for the analysis of TAF5 mutants are listed in the supplementary material. They were created in pRS415-*TAF5* by site-directed mutagenesis using oligonucleotides listed in the supplementary material [42]. pRS415-*TAF5* mutant plasmids were transformed into YJR501 (*mat**a** taf5::TRP1* [pRS316-*TAF5*] *his4-917Δ lys2-17R2 leu2 ura3-52*) and plated onto leucine dropout plates. Two independent clones were grown overnight in YEPD medium, plated onto 5-fluoro-orotic acid, and grown for three days at 30°C. Two independent clones were verified by sequence analysis and used in the spot assay. For the spot assay, cells were grown to saturation overnight in YEPD medium and spotted in a dilution series from 1E8 to 1E5 cells/mL. The plates were incubated at 30°C for the following times: YEPD, 2 days; YEP with galactose, 3 days; synthetic complete (SC) medium without lysine, 3 days; SC without histidine, 4 days; SC without inositol, 3 days. For immunoblotting of TAF5 mutants, extracts were prepared as described previously [51]. Samples were blotted on PVDF membrane and probed with a TAF5 antibody kindly provided by Jerry Workman. Yeast strains used for the analysis of truncated Spt7 were W303-1B derivatives and are listed in the supplementary material. N-terminal TAP tagging and C-terminal truncation of the genomic copy of *SPT7* were performed using standard methods [49,52]. Strains were grown to mid-log phase, whole extracts were prepared as above, and TAP-tagged Spt7 was immunoprecipitated with IgG affinity resin. Elution was performed in SDS PAGE loading buffer for 5 min at 100°C. Eluted proteins were separated by SDS PAGE, blotted to a PVDF membrane, and analyzed for SAGA/SLIK subunits. PAP antibody (Sigma Aldrich) was used to detect the TAP tag. Antibodies recognizing Spt8, TAF10, TAF12 and TAF5 were kindly provided by Jerry Workman.

5 Acknowledgements

The authors thank Shabaz Mohammed for helpful discussion, Joseph Reese for providing yeast strains and plasmids, Jerry Workman for providing antibodies, Simone Lemeer and Lars Meijer for sharing anti-acetyl lysine antibodies, and Robert Kerkhoven for the SpectraMapper software. This work was financed by the NPC Horizon Program (grant number 050-71-050) and by the Netherlands Proteomics Centre.

6 References

[1] Gavin, A. C*., et al.*, *Nature* **2006**, *440*, 631
[2] Gavin, A. C*., et al.*, *Nature* **2002**, *415*, 141
[3] Timmers, H. T. M*., et al.*, *Trends Biochem Sci* **2005**, *30*, 7
[4] Thomas, M. C*., et al.*, *Crit Rev Biochem Mol Biol* **2006**, *41*, 105
[5] Baker, S. P*., et al.*, *Oncogene* **2007**, *26*, 5329
[6] Lee, K. K*., et al.*, *Nat Rev Mol Cell Biol* **2007**, *8*, 284
[7] Green, M. R., *Trends Biochem Sci* **2000**, *25*, 59
[8] Sanders, S. L*., et al.*, *Mol Cell Biol* **2002**, *22*, 4723
[9] Powell, D. W*., et al.*, *Mol Cell Biol* **2004**, *24*, 7249
[10] Lee, K. K*., et al.*, *Biochem Soc Transac* **2004**, *32*, 899
[11] Grant, P. A*., et al.*, *Genes Dev* **1997**, *11*, 1640
[12] Grant, P. A*., et al.*, *Cell* **1998**, *94*, 45
[13] Kohler, A*., et al.*, *Mol Biol Cell* **2006**, *17*, 4228
[14] Sanders, S. L*., et al.*, *Mol Cell Biol* **2002**, *22*, 6000
[15] Wu, P. Y. J*., et al.*, *Mol Cell* **2004**, *15*, 199
[16] Leurent, C*., et al.*, *Embo J* **2004**, *23*, 719
[17] Lee, T. I*., et al.*, *Nature* **2000**, *405*, 701
[18] Hahn, S., *Cell* **1998**, *95*, 579
[19] Pray-Grant, M. G*., et al.*, *Mol Cell Biol* **2002**, *22*, 8774
[20] Sterner, D. E*., et al.*, *Proc Natl Acad Sci USA* **2002**, *99*, 11622
[21] Sermwittayawong, D*., et al.*, *Embo J* **2006**, *25*, 3791
[22] Wu, P. Y. J*., et al.*, *Mol Cell Biol* **2002**, *22*, 5367
[23] MacCoss, M. J*., et al.*, *Proc Natl Acad Sci USA* **2002**, *99*, 7900
[24] Wu, S. L*., et al.*, *J Proteom Res* **2005**, *4*, 1155
[25] Schlosser, A*., et al.*, *Anal Chem* **2005**, *77*, 5243

[26] Mohammed, S., *et al.*, *Anal Chem* **2008**, *80*, 3584
[27] Wu, C. C., *et al.*, *Nat Biotechnol* **2003**, *21*, 262
[28] Gauci, S., *et al.*, *Anal Chem* **2009**, *81*, 4493
[29] Washburn, M. P., *et al.*, *Nat Biotechnol* **2001**, *19*, 242
[30] Cargile, B. J., *et al.*, *J Proteom Res* **2004**, *3*, 112
[31] Krijgsveld, J., *et al.*, *J Proteom Res* **2006**, *5*, 1721
[32] Boersema, P. J., *et al.*, *J Proteom Res* **2007**, *6*, 937
[33] Florens, L., *et al.*, *Methods* **2006**, *40*, 303
[34] Paoletti, A. C., *et al.*, *Proc Natl Acad Sci USA* **2006**, *103*, 18928
[35] Lu, P., *et al.*, *Nat Biotechnol* **2007**, *25*, 117
[36] Link, A. J., *et al.*, *Nat Biotechnol* **1999**, *17*, 676
[37] Chi, A., *et al.*, *Proc Natl Acad Sci USA* **2007**, *104*, 2193
[38] Smolka, M. B., *et al.*, *Proc Natl Acad Sci USA* **2007**, *104*, 10364
[39] Albuquerque, C. P., *et al.*, *Mol Cell Proteom* **2008**, *7*, 1389
[40] Li, X., *et al.*, *J Proteom Res* **2007**, *6*, 1190
[41] Sterner, D. E., *et al.*, *Mol Cell Biol* **1999**, *19*, 86
[42] Durso, R. J., *et al.*, *Mol Cell Biol* **2001**, *21*, 7331
[43] Glozak, M. A., *et al.*, *Gene* **2005**, *363*, 15
[44] VanDemark, A. P., *et al.*, *Mol Cell* **2007**, *27*, 817
[45] McMahon, S. J., *et al.*, *Proc Natl Acad Sci USA* **2005**, *102*, 8478
[46] Kohler, A., *et al.*, *Nat Cell Biol* **2008**, *10*, 707
[47] Martinez, E., *et al.*, *J Biol Chem* **1998**, *273*, 23781
[48] Logie, C., *et al.*, *Methods Enzymol* **1999**, *304*, 726
[49] Puig, O., *et al.*, *Methods* **2001**, *24*, 218
[50] Pinkse, M. W. H., *et al.*, *Anal Chem* **2004**, *76*, 3935
[51] Kushnirov, V. V., *Yeast* **2000**, *16*, 857
[52] Lorenz, M. C., *et al.*, *Gene* **1995**, *158*, 113

Chapter 4

Analysis of the general transcription factor TFIID from mouse embryonic fibroblasts and mouse embryonic stem cells

Nikolai Mischerikow[1,2], A. F. Maarten Altelaar[1,2], H. T. Marc Timmers[2,3], W. W. M. Pim Pijnappel[2,3], and Albert J. R. Heck[1,2,4]

Work in mouse embryonic fibroblasts was done in collaboration with:

Mohamed-Amin Choukrallah[5], Dominique Kobi[5], Igor Martianov[5], and Irwin Davidson[5]

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

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

[3] Department of Molecular Cancer Research, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

[4] Centre for Biomedical Genetics, Padualaan 8, 3584 CH Utrecht, The Netherlands

[5] Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, 1 Rue Laurent Fries, 67404 Illkirch Cédex, France

1 Introduction

Proteins within cells commonly do not exert their function alone, but in concert with other biomolecules with which they stably or transiently associate. Within these complexes, the functions of single proteins are often altered such that the function of the complex is of primary interest. The dissection of the interactions is therefore essential to investigate the function of a protein complex. Mass spectrometry- (MS) based proteomics is ideally suited to study protein complexes because it provides a comprehensive picture of all subunits including unknown interactors [1,2]. MS-based protein complex characterizations require the enrichment of the targeted complex from its native environment, which is achieved by affinity purification (AP) using a tagged complex subunit, or by immunoprecipitation (IP) with an antibody directed against one subunit. Both methods can also be performed in tandem which results in stronger enrichment of specifically co-purifying proteins over unspecific background which binds to the affinity resin. In contrast to antibody-based protein detection methods, MS-based protein detection can clearly discriminate complex subunits from unspecific background protein.

The classical approach of analyzing protein complexes by MS is the combination of AP or IP with subsequent protein or peptide pre-fractionation and liquid chromatography-coupled MS (LC-MS). Unspecific interactors are revealed from a control purification using an untagged bait protein or, in case of IPs, the affinity resin without antibody. A strong enrichment is crucial to identify specific interactors and therefore the method is nearly always combined with tandem affinity purifications (TAP) or tandem immunoprecipitations. Specific and unspecific binders are then discriminated by spectral counting or more elaborate label-free data analysis techniques. To obtain additional certainty in characterizing the protein-protein interactions that establish a protein complex, the complex can be purified by targeting multiple subunits in different experiments, an approach that has been successfully applied to characterize protein complexes on a proteome-wide level [3,4]. Drawbacks of this method are that TAPs usually result in loss of weak and transient interactors, and that the LC-MS analysis usually is quite extensive due to the additional protein or peptide pre-fractionation which is commonly used for complexity reduction of peptides sampled into the mass spectrometer. An important improvement to the classical approach is the application of stable isotope labeling of the AP/IP input material. Here, the purification and the control purification are encoded with different isotopes so that background binders, which are equally present in both purifications, and specific interactors, which are enriched in the purification over its

control, can be discriminated by quantitative proteomics. This approach greatly decreases the need for sample complexity reduction prior to LC-MS, which affects both the enrichment methods of the targeted protein complex as well as pre-fractionation prior to LC-MS. Another continuous improvement is the development of AP or IP systems with enhanced binding and elution characteristics. Among many different strategies, an antibody-based GFP-binding protein has recently been commercialized [5] that has a sufficiently high binding affinity and specificity for GFP to allow replacing TAPs by single step APs for a quantitative analysis of protein-protein interactions [6].

Here, we present two analyses of the general transcription factor (GTF) TFIID from mouse embryonic fibroblasts (MEFs) and mouse embryonic stem cells (mESCs), a well-studied protein complex composed of TBP and 13-14 TBP associated factors (TAFs) [7].

The first analysis, a study of TBP-interacting proteins in MEFs, was performed as a classical tandem IP using FLAG and HA epitopes in combination with peptide pre-fractionation using strong cation exchange (SCX) resulting in extensive LC-MS analysis. The FLAG-HA-tagged bait protein was TBP, which is not only stably incorporated into TFIID, but also into smaller protein complexes like for example the binary B-TFIID complex consisting of BTFA1 and TBP, the binary TFIIIB complex composed of TBP and BRF1, and the SL1 complex consisting of TBP and four associated TAFs, all of which have different functions in transcription initiation [8]. Regular TBP was compared with selected TBP mutants that among many other, mainly single amino acid substitution mutants have been used to define TBP interaction epitopes by mapping protein and DNA interactions of TBP to its solvent-exposed surface and to study functional effects of the mutations in large screens mainly conducted *in vitro* and to some extent also *in vivo* [9,10]. The specific loss of TBP interactions in these mutants was characterized by spectral counting, a method that is well applicable if a large number of fragment spectra are assigned to a protein and the differences of theses number in between samples are large.

The second analysis was the isolation and characterization of TFIID from mESCs, which was conducted as highly efficient single step AP using GFP in conjunction with SILAC, a strategy which has recently been termed QUBIC when combined with bacterial artificial chromosome- (BAC) recombineering for transgene generation [6]. No additional pre-fractionation was applied prior to quantitative LC-MS analysis. The GFP-tagged bait was TAF6 which is only incorporated into TFIID. Due to the high efficiency of the analysis strategy, we also tried to study perturbations of the complex composition during *in vitro* transcription (IVT) reactions [11].

		Analysis batch #1			Analysis batch #2							
		C	WT	R188E	C	WT	V162A	Q242A	K243E	K265A	R318A	SPM3
TFIID	TBP	0	40	27	0	11	11	10	32	30	14	17
	TAF1	0	209	168	0	99	79	77	169	138	133	73
	TAF2	0	99	82	0	40	38	43	64	67	56	26
	TAF3	0	118	104	0	60	52	44	83	56	66	24
	TAF4a/b	0	276	228	0	110	86	106	171	184	165	79
	TAF5	0	168	127	5	68	54	92	139	154	129	81
	TAF6	0	204	137	0	117	90	85	145	143	130	63
	TAF7	0	77	71	0	45	38	26	63	49	44	33
	TAF8	0	48	53	0	31	26	18	34	21	15	13
	TAF9a/b	0	130	117	0	62	49	41	67	61	48	25
	TAF10	0	48	30	0	17	14	15	23	20	18	7
	TAF11	0	24	18	0	10	12	8	19	14	9	11
	TAF12	0	50	33	0	13	10	13	13	17	14	5
	TAF13	2	22	30	0	9	11	7	13	13	9	8
B-TFIID	BTAF1	0	52	0	0	29	20	20	0	4	14	31
TFIIIB	BRF1	0	18	17	0	7	5	3	4	27	26	20
SL1	TAFIA	0	4	4	0	9	5	5	4	0	0	10
	TAFIB	0	6	5	0	5	7	6	3	6	2	11
	TAFIC	0	12	12	0	15	17	3	2	0	4	16
	TAFID	0	7	8	0	2	6	4	1	4	2	13
TFIIA	TFIIAa/b	0	4	0	0	1	0	0	0	3	2	2
	TFIIAg	0	12	1	0	4	3	0	0	4	1	0

Figure 1. The interactome of selected TBP point mutant with the specific mutation denoted on the top (C, background control; WT, wild type). The table is split according to the two batches of analyses that were conducted separately. The table entries represent the number of spectra assigned to the identified proteins denoted in the left column, which are grouped by protein complex. Clearly visible is the loss of BTAF1 interaction in TBP R188E and K243E.

2 Results and discussion

To analyze how mutations of TBP affect the stable association with TBP- interacting proteins *in vivo*, tagged wild type or mutant human TBP (hTBP) was expressed in MEFs which had both endogenous TBP alleles inactivated (*Tbp*$^{-/-}$). The expression levels of the different transgenes were similar to or slightly higher than the expression level of endogenous TBP. All wild type and mutant hTBP transgenes were N-terminally tagged with a FLAG-HA tandem epitope tag that was used to co-immunoprecipitate factors stably associated with hTBP from nuclear extracts of the corresponding cells lines. These TBP interactomes were precipitated, digested with Trypsin and analyzed by LC-MS. Although TAPs and tandem IPs like FLAG-HA drastically reduce sample complexity, we chose to extensively pre-fractionate the peptides using SCX in combination with short

Mutation	Affected interaction
V162A	DNA, BTAF1, TFIIB
R188E	BTAF1, NC2
Q242A	?
K243E	BTAF1, NC2
K265A	DNA
R318A	DNA
SPM3	DNA

Figure 2. Known effects of TBP mutations on protein-protein and protein-DNA interactions.

LC-MS analyses of each fraction. MS was performed on a high resolution, high mass accuracy LTQ FTICR instrument. In addition to the reduction of sample complexity, the extensive orthogonal pre-fractionation resulted in highly abundant proteins including TBP interactors being identified by a large number of spectra. This did not only increase confidence in protein identification, but also allowed the estimation of quantitative differences between samples by spectral counting. Spectral counting gives reliable estimates if the samples have a very similar overall composition and if the number of identified spectra for one or few single proteins differs largely in between different samples, a situation encountered in this experiment. The analysis of the hTBP interactome was carried out in two different sets of experiments each with its own controls. Tandem IPs from nuclear extracts of *Tbp*$^{-/-}$ MEFs expressing wild type hTBP and tandem IPs from nuclear extracts of MEFs expressing one allele of endogenous TBP but no hTBP (*Tbp*$^{lox/-}$) were used as controls. The first analysis was that of hTBP R188E and the second analyses comprised hTBP V162A, Q242A, K243E, K265A, R318A, and SPM3. On average, around 2,500 spectra assigned to 400 unique proteins were cumulatively identified in each experiment, with around 30% of all spectra assigned to known TBP interactors in wild type or mutant hTBP samples. Known interactors that were identified in both wild type hTBP samples included all TFIID subunits, BTAF1, BRF1, and all four SL1 subunits including the recently discovered TAFID/JOSD3, while all these interactors were absent in the background control (Figure 1). Proteins known to interact transiently with TBP *in vivo*, like NC2 or TFIIB, were not identified in the analyses and only a very low number of spectra was identified for TFIIA. Besides the known TBP interactors, no other protein displayed a pattern consistent with specific TBP association, *i.e.* absent in the background control and present in the wild type and most mutant hTBP samples. No difference in terms of spectral abundance could be observed between wild type TBP and most of the mutants, but hTBP R188E and K243E displayed a complete loss of BTAF1 association while all other interactions were maintained. This observation is in line with previous experimental results [9,10] summarized in Figure 2.

Further investigations exploited this selective disruption of the interaction of hTBP R188E and K243E with BTAF1 *in vivo* without having to affect intracellular BTAF1 or

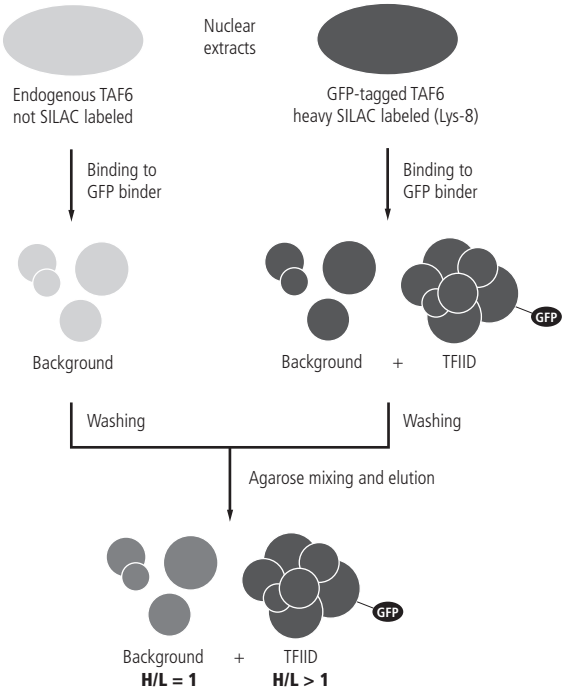


Figure 3. Illustration of the the application of quantitative proteomics to discriminate the TFIID complex from background proteins during affinity purification.

TBP levels. These experiments showed, in summary, that in MEFs expressing hTBP R188E and K243E the global genomic distribution of TBP at genes promoters was not affected in general and only a small number of genes showed altered expression. For some of these genes with increased expression levels this increase could be linked to the increased recruitment of TFIID and RNA polymerase II to the promotor, while altered promoter occupancy of TBP itself, as well as the GTFs TFIIB and Mediator, could be excluded.

A different approach was taken to analyze the TFIID composition in mESCs. Instead of a tandem IP to enrich TFIID sufficiently above background, a highly efficient single step AP was combined with SILAC to discriminate specifically co-purifying proteins from unspecific background. The selected bait was TAF6, which unlike its homologue in *Saccharomyces cerevisiae* is solely incorporated into TFIID but not into other related complexes like TFTC/STAGA [12]. The chosen affinity tag was eGFP, for which a highly affine binder is commercially available [5]. mESCs stably expressing C-terminally eGFP-tagged TAF6 were generated using BAC recombineering [13]. The BAC contained the native TAF6 gene including its endogenous promotor and other regulatory elements to achieve a level of transcriptional control of the transgene similar to that of endogenous TAF6, and indeed the selected clonal line expressed TAF6-eGFP at a level close to endogenous TAF6. Stable isotope labeling of the TAF6-eGFP mESC line was achieved with $^{13}C_6$,$^{15}N_2$-lysine (Lys-8) SILAC. The experimental design to discriminate specific from unspecific TAF6 interactors, *i.e.* TFIID from background, is described in Figure 3. In brief, two APs from nuclear extracts of Lys-8 labeled TAF6-eGFP mESCs and of unlabeled wild type

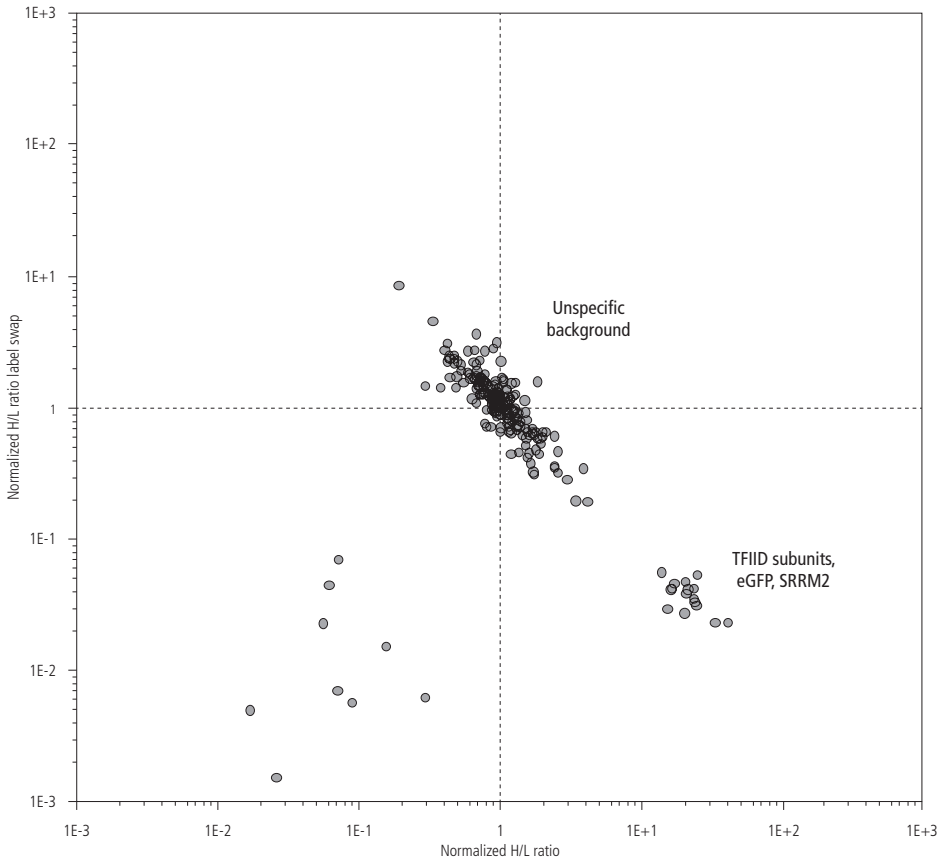


Figure 4. Summary of the TFIID purification using the quantitative strategy. Each dot represents a quantified protein. The horizontal axis indicates the normalized H/L ratio when the APs were isotope-coded as illustrated in Figure 3, and the vertical axis indicates the H/L ratio when the isotope coding of the APs was swapped.

mESCs were performed in parallel under identical conditions and the light and heavy labels were combined during elution. Background proteins which were present in both unlabeled and Lys-8 labeled nuclear extracts eluted in equal amounts and therefore adopted a heavy/light ratio (H/L) of 1. TAF6-eGFP and its specific interactors were only present with a heavy label in the eluate and therefore adopted an H/L of >1. The eluates were digested with Lys-C using the FASP protocol [14] and analyzed using LTQ Orbitrap LC-MS. Quantification was carried out without further validation using MaxQuant [15].

Using this strategy, we were able to purify the TFIID complex from TAF6-eGFP expressing mESCs with the same subunit composition as the TFIID complex purified from MEFs. All TFIID subunits as well as eGFP had an H/L between 10 and 50 and grouped clearly

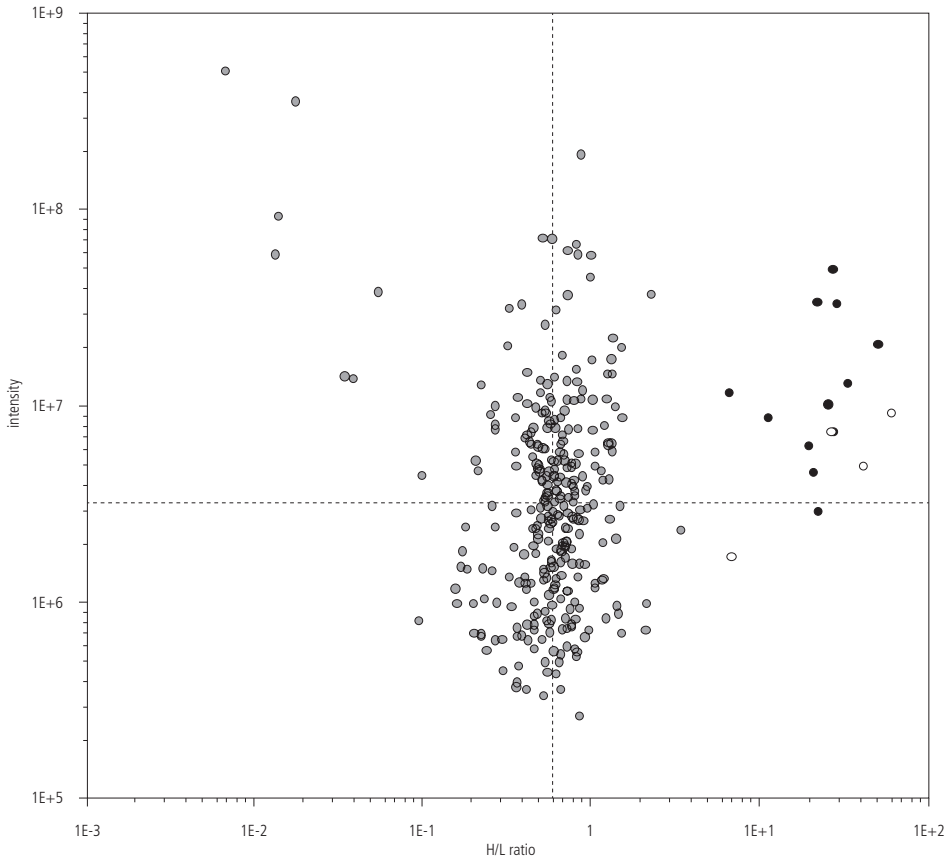


Figure 5. Result of the TFIID purification using 2 min binding to the affinity resin. Each dot represents a quantified protein, with grey dots indicating background binders, black dots indicating TFIID components, and black circles indicating TFIID subunits with a single H/L count. The horizontal axis indicates the H/L and the vertical axis the protein intensity as reported by the quantification software. Dashed lines indicate intensity and H/L medians.

distinct from the around 300 quantifiable background proteins with an H/L of around 1 (Figure 4). The protein intensities, *i.e.* the summed intensities of all assigned heavy and light peaks, of all TFIID components were well above the intensity median of all quantified proteins, which indicates a robust detection during LC-MS. The identified TFIID subunits included TBP, TAF1-TAF13, and the TAF4a related TAF4b. The TAF9-related TAF9b, which was identified as stable component of TFIID in MEFs, was not reproducibly detected in all analyses, which is likely related to the high degree of sequence identity of the two isoforms resulting in only very few peptides being able to discriminate the two isoforms. Another issue related to the analytical strategy was the quantification of TBP

and TAF12, which in some replicate analyses was only based on a single ratio count. In addition to the known TFIID subunits, the potentially novel TFIID interactor SRRM2 was clearly enriched as well. SRRM2 was not detected when ethidium bromide was added to the nuclear extracts during binding to the affinity resin, indicating that the SRRM2-TAF6/TFIID interaction is mediated by double-stranded DNA or RNA.

Two remarkable technical aspects highlight the sensitivity of the analytical strategy. First, the APs were performed with only very little input material corresponding to around half a large (15 cm diameter) tissue culture dish of mESCs. Second, the binding of the TAF6-eGFP from the nuclear extracts to the affinity resin was achieved in only 10 min. Interestingly, a longer binding duration of 60 or 90 min resulted in a larger number of background proteins. While we did not investigate this aspect further, it is apparent that

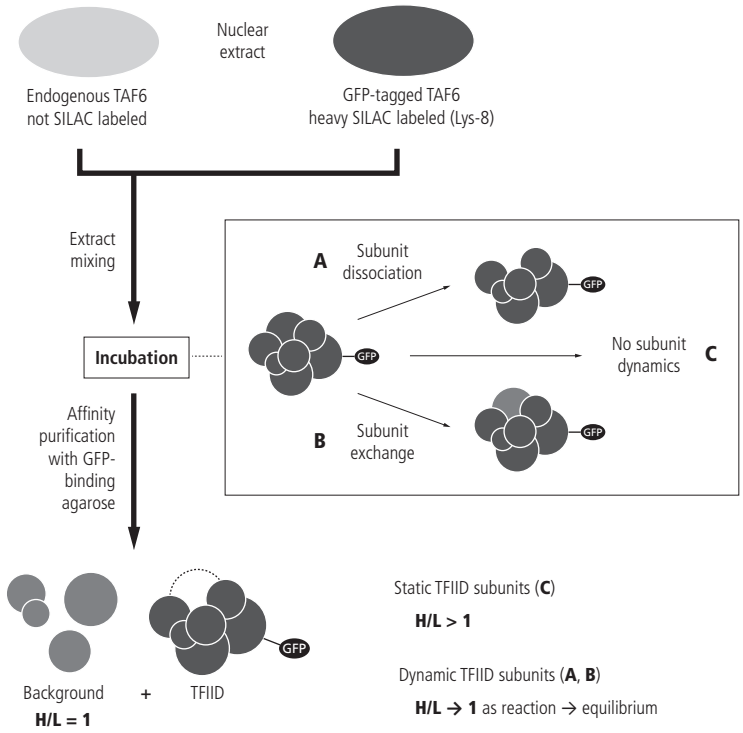


Figure 6. Adaptation of the workflow illustrated in Figure 3 to measure in vitro subunit dynamics of TFIID. Isotope-coded extracts are mixed, incubated under different conditions, and affinity purified. The discrimination of static TFIID subunits (C) from background is identical as described in Figure 3. TFIID subunits that exchange with the surrounding nuclear extract or dissociate (A, C) will approach an H/L of 1 as the exchange reaction proceeds into equilibrium, as long as they are still identified (case A).

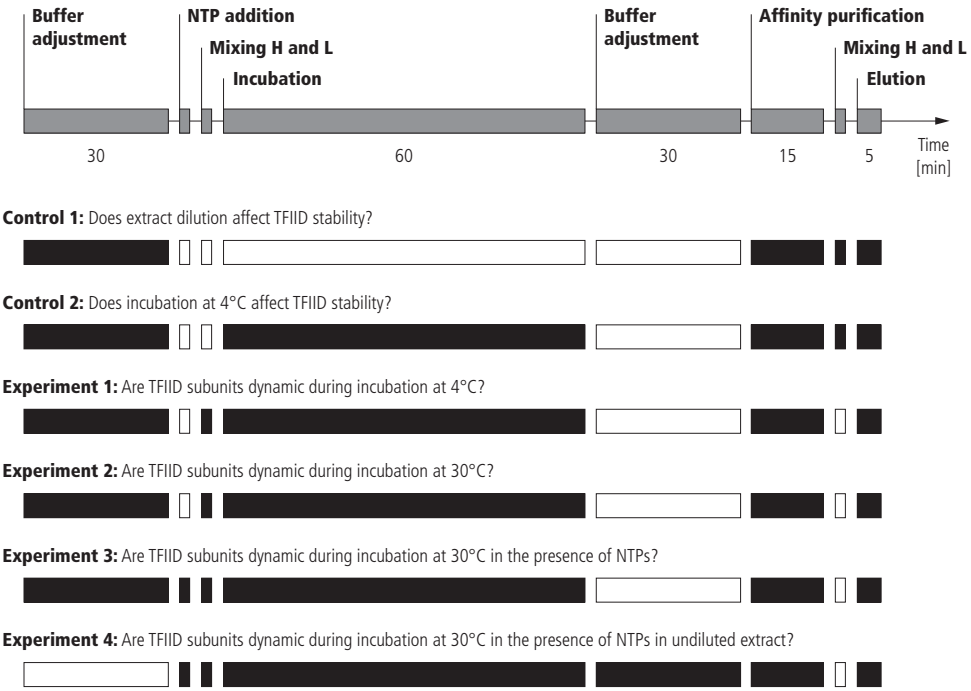


Figure 7. Illustration of the experimental workflows to investigate TFIID stability under different incubation conditions. The upper time bar in grey shows the general sequence of possible experimental steps with the time required/allowed for each step. The lower black/white panels indicate time bars of real experiments and controls, with black bars indicating that a possible step has been realized and white bars indicating that the step has been omitted.

an increased purification background requires longer LC gradients to reach the same sampling depth with respect to the targeted complex. The binding duration could even be lowered to 2 min while maintaining detection and quantification of all TFIID subunits. However, at this short binding time, the quantification of TBP, TAF8, TAF10, and TAF12 were based on single ratio counts (Figure 5). We therefore kept 10 min binding duration as standard for all experiments described in the following.

The speed and efficiency of the AP was exploited to test whether TFIID subunits display dynamic behavior during incubations *in vitro*. For this purpose, the AP was adopted as described in Figure 6. In contrast to the workflow described above, the Lys-8 labeled TAF6-eGFP containing nuclear extract and the unlabeled endogenous TAF6 containing nuclear extract were combined and incubated under various different conditions before the mixture was subjected to the affinity purification. Background binders then adopt an H/L of 1 and stable TFIID subunits adopt an H/L of >1. Dynamic subunits approach

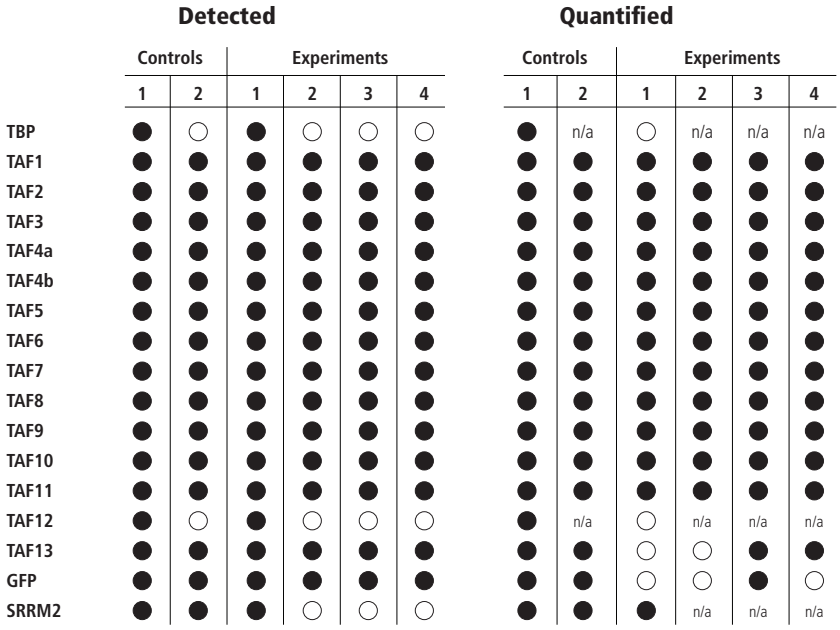


Figure 8. Summary of the results of the TFIID stability experiments as illustrated in Figure 7. Except for TBP and TAF12, all subunits are identified with more than 1 unique peptide (black dots) in both controls and all four experiments (black circles indicate that the protein was not identified). Most identified subunits could also be quantified (black dots) and only few subunits had a single H/L count (black circles).

the H/L of the component of the nuclear extract they exchange with, which in this case is 1 due to equimolar extract mixing, while the exchange reaction runs into equilibrium. Subunits that dissociate from TFIID also approach an H/L of 1, however only as long as they are not fully dissociated because in this case they are likely not detected and therefore not quantified.

The first sets of experiments were carried out to test whether TFIID can be purified from mixtures incubated over a prolonged time, at low salt conditions, elevated temperatures, and in the presence of nucleotide triphosphates (NTPs), which are conditions applied during *in vitro* transcription reactions. The experimental setups for these controls are depicted in Figure 7 and the results are summarized in Figure 8. None of the tested conditions affected TFIID stability and all TFIID subunits were enriched at H/L of >1 together with TAF6-eGFP, although the detection and quantification of TAF12 and TBP again was not robust in most experiments.

We then tested TFIID stability during (IVT) reactions [11]. In this assay, a transcription-

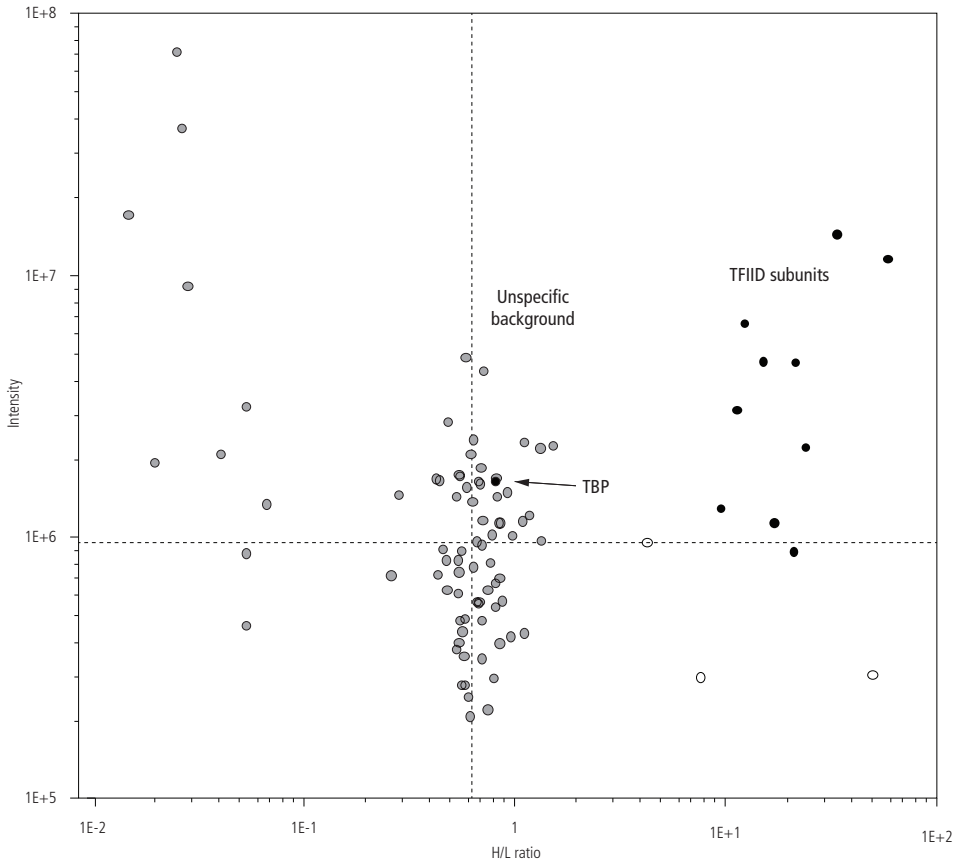


Figure 9. Result of the TFIID purification after incubation under IVT conditions. The axis descriptions, dashed lines and dot color codes are identical as in Figure 5. Clearly visible is that TBP as the only TFIID subunit groups together with the background binders, while the other TFIID subunits have an elevated H/L of >1.

competent nuclear extract containing RNA polymerase II and all GTFs is supplemented with template DNA, NTPs, and an ATP regenerating system and incubated to allow RNA polymerase II-dependent transcription of the template DNA. The assay readout is the RNA production from the template DNA, which usually is monitored by ^{32}P-NTP radioactive labeling or RT-qPCR. Different perturbations, for example variations of the promotor DNA or mutations or depletion of the GTFs present in the nuclear extract, can be used to study the molecular mechanisms of transcription *in vitro*.

Here, we have tested whether standard IVT reactions conditions can introduce dynamics of subunits of TFIID. As a complex enzymatic assay, IVT reactions are very sensitive

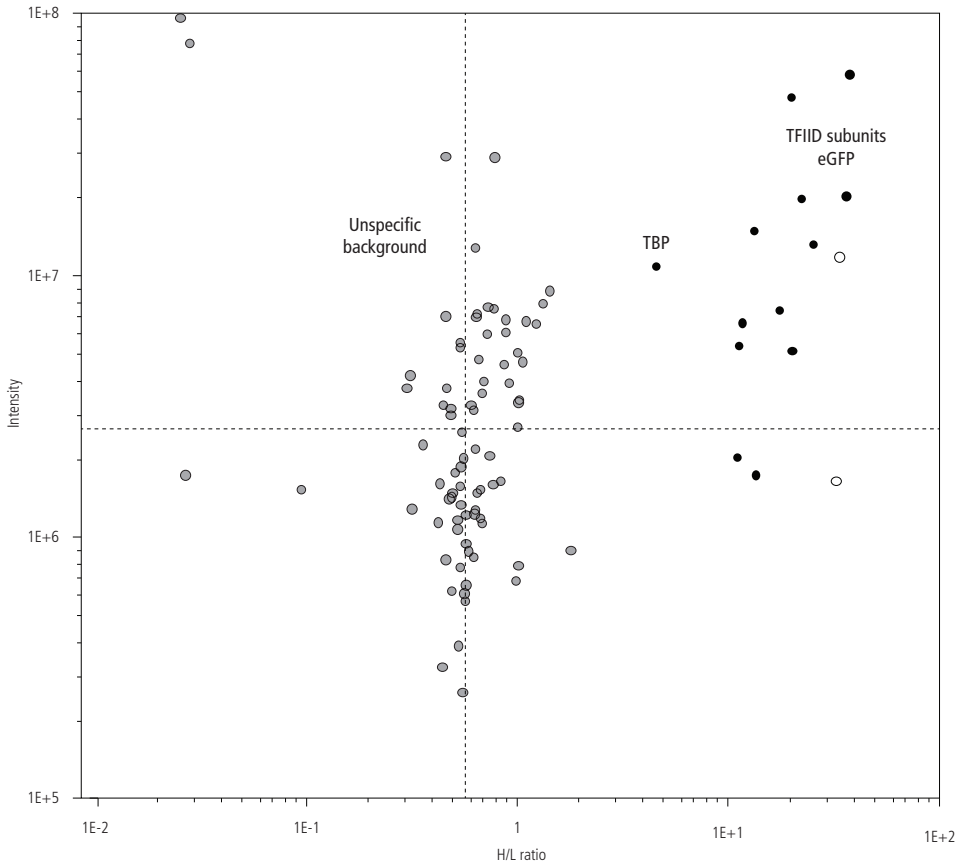


Figure 10. Result of the TFIID purification after incubation under IVT conditions without template DNA. The axis descriptions, dashed lines and dot color codes are identical as in Figure 5. In contrast to the experiment including template DNA, TFIID groups in between TFIID and the background binders.

to the buffer composition, in particular with respect to pH and the concentrations of Mg^{2+} and monovalent cations. Therefore, instead of diluting nuclear extracts to approximate the desired buffer composition as in the previously described experiments, we microdialyzed the nuclear extracts against a suitable buffer [11]. The reactions were set up from equal volumes of heavy and light isotope-coded extracts and the required IVT reaction components. As a control experiment, the IVT reaction was conducted without template DNA which should abolish IVT completely. After incubation, TAF6-eGFP was affinity purified and the eluate subjected to on-filter digestion followed by LC-MS analysis on a LTQ Orbitrap.

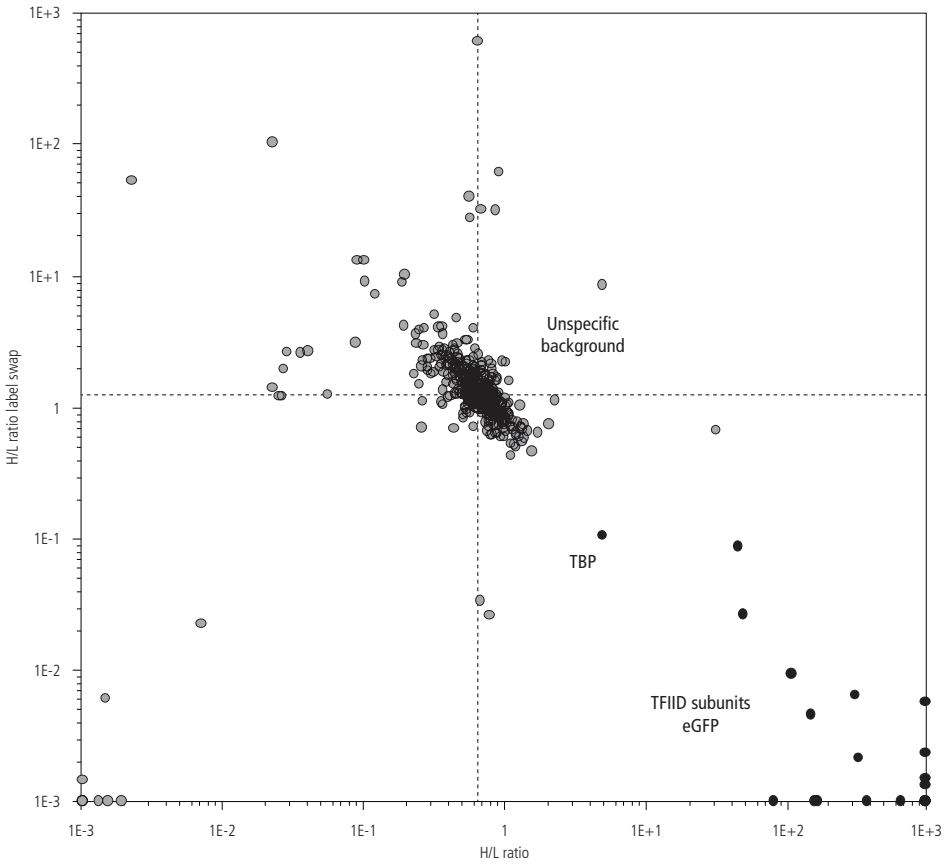


Figure 11. Result of the TFIID purification after incubation under IVT conditions. The experiment was conducted in replicate using a label swap design as described in Figure 4. In contrast to the IVT experiment described in Figure 9, TBP adopts an intermediate position between the position of the background and TFIID.

In both APs, around 200 proteins were identified and around 100 proteins could be quantified. All TFIID subunits except TAF10 and TAF12 were detected and group together at an elevated H/L clearly distinct from noise (Figure 9). The absence of TAF10 and TAF12 might point towards their complete dissociation from TFIID; however, the low number of detected proteins and their low absolute intensities rather indicate that the experiment was conducted at the limit of detection and/or quantification.

The only subunit that clearly displayed an altered H/L is TBP, which groups together with the background proteins, suggesting that the heavy labeled TBP had been dynamically exchanged into full equilibrium with a 1:1 mixture of labeled and unlabeled TBP during

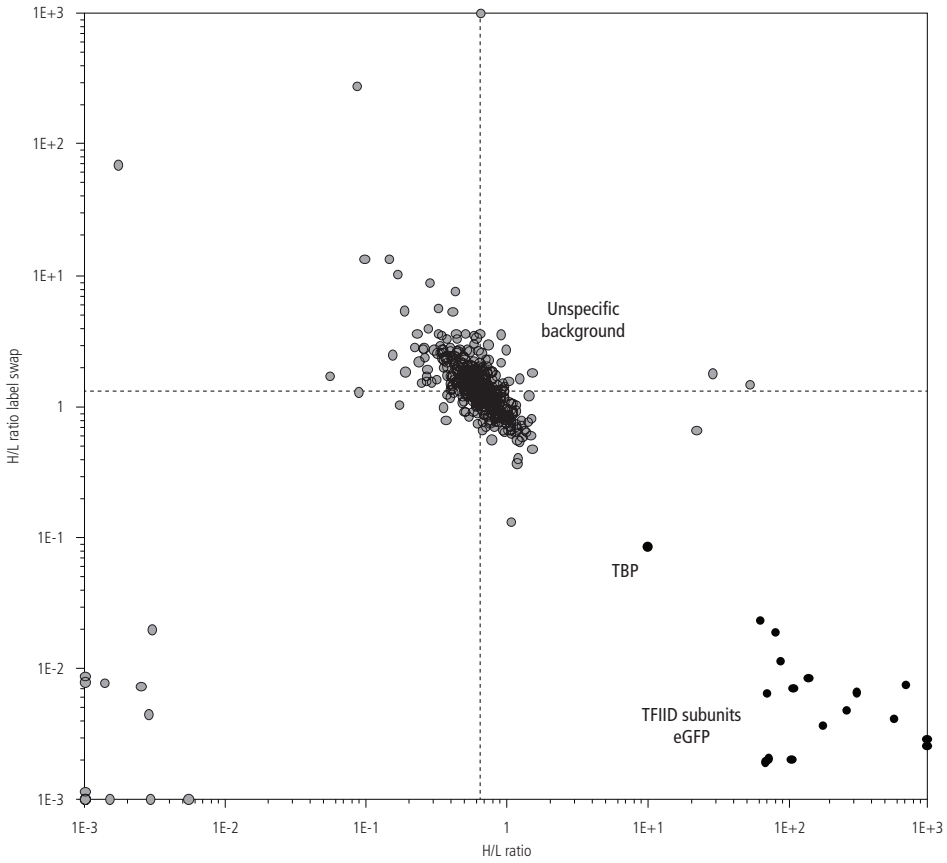


Figure 12. Result of the TFIID purification after incubation under IVT conditions without template DNA. The experiment was conducted in replicate using a label swap design as described for Figure 4. TBP assumes a nearly identical position as in the experiments including template DNA (Figure 11).

the incubation. The high absolute intensity of TBP suggests that TBP did not only dissociate from the complex, which would have rendered it a background binder, but was indeed re-associating with TFIID. In the affinity purification from the control IVT reaction without template DNA, TBP grouped at an intermediate H/L between the stable TFIID subunits and the background binders (Figure 10).

Due to the somewhat unclear grouping of TBP in the control IVT reaction, the experiments were repeated with a higher amount of input material and LC-MS analysis on the then available LTQ Orbitrap VELOS which has an increased detection sensitivity compared to a regular LTQ Orbitrap. Peptide fragmentation was performed using higher-

Figure 13. Sequence alignment of TAF7 (upper sequence, Q9R1C0) and TAF7L (lower sequence, Q9D3R9) with all unique peptides for TAF7L mapped to the sequence. Indicated below each peptide is the number of spectra detected using HCD (cumulative from all four analyses), the isotope labeled versions detected, and the maximum ion score observed.

energy collisional dissociation (HCD) which due to fragment ion readout in the Orbitrap analyzer improves the confidence of peptide identifications. Both the experiment and the control (without template DNA) were performed in duplicate in a label swap design. The results are shown in Figures 11 and 12. In all experiments, TBP displayed an H/L between background binders and other TFIID subunits; no difference could be observed between the IVT reaction and the control. A comparison of the intermediate isotope ratios of TBP in all four experiments with the results obtained in the previous experiments (Figures 9 and 10) could indicate that the IVT reactions were impaired and therefore did not occur. In fact, an inhomogeneous (as analyzed by SDS PAGE) proteinaceous precipitate had formed during nuclear extract dialysis and IVT reactions which was not observed in the previous experiments. In conclusion, our experiments so far indicate that during IVT reactions TBP in TFIID displays dynamic behavior which can be stimulated by the addition of large amounts of TATA-containing promoter DNA. However, our approach cannot clearly distinguish between dissociation of TBP from TFIID and exchange with TBP in the extracts, so further validation would certainly be required.

The LTQ Orbitrap VELOS analyses also revealed that the TAF7 isoform TAF7L is clearly a component of mESC TFIID. The two experiments and two controls identified TAF7L with 8 unique peptides in total which map to TAF7L as depicted in Figure 13. Several of these peptide identifications were confirmed in re-analyses of the experiments using electron transfer dissociation (ETD) as dissociation method, further bolstering confidence in TAF7L identification. The presence of TAF7L in mESC TFIID is a novel finding as TAF7L has previously been shown to be expressed in testis but not ovary, brain, liver, lung or kidney [16]. In testis, TAF7L expression is upregulated (while TAF7 expression is downregulated) during spermatocyte differentiation, a switch required for spermatogenesis but apparently no other differentiation processes which was shown by TAF7L knockout mice [17]. Although TFIID harboring TAF7L has so far not been characterized, TAF7L has been shown to associate with most TFIID subunits including TBP and TAF6 in spermatocytes, replacing TAF7 from the immunoprecipitate [16]. Our finding that mESC TFIID contains both TAF7 and TAF7L raises a number of questions, for example regarding the genes regulated by TAF7L, the composition of TAF7L-containing TFIID, *etc.* The immediate general question, however, is if TAF7L has a specific role in maintaining the ESC-like state of the used mESC line. In our opinion, the fact that TAF7L is only activated within in a short developmental window during spermatogenesis and besides that appears to be silent would make it worthwhile pursuing this task.

3 Experimental procedures

Generation of transgenic mESCs and MEFs. mESCs expressing GFP-labeled TAF6 next to endogenous TAF6 were generated as described in the literature [13]. In brief, a localization and affinity purification (LAP) tag comprising TY, SBP and eGFP was C-terminally fused to TAF6 using BAC recombineering, and the transgene was stably transfected into E14 mESCs under G418 selection. All experiments were performed with a clonal line expressing the TAF6 transgene at a near-endogenous level. MEFs expressing wild type or mutant FLAG-HA-hTBP in a TBP knockout background were generated from *Tbp*$^{lox/-}$ mESCs [18,19]. In brief, MEFs were isolated from *Tbp*$^{lox-/-}$ mice, immortalized with the SV40 large T-antigen, and transfected with retroviral vectors expressing N-terminally FLAG-HA-tagged hTBP. Selected clonal lines were then transfected with Cre recombinase to inactivate the floxed TBP allele, resulting in *Tbp*$^{-/-}$ MEFs expressing FLAG-HA-hTBP at near-endogenous levels or higher, depending on the specific mutation.

Culturing of mESC and MEFs. mESCs were cultured in DMEM (with 4.5 g/L glucose and 2 mM L-glutamine, LONZA) and 15% (v/v) fetal bovine serum (FBS) (Hyclone), supplemented with 100 units/mL penicillin, 100 units/mL streptomycin, 1 mM sodium pyruvate, 1x non-essential amino acids (LONZA), 100 µM β-mercaptoethanol, and recombinant leukemia inhibitory factor. Transgene expression was maintained with 250 µg/mL G418. For the SILAC procedure, mESCs were cultured in DMEM (with 4.5 g/L glucose and 2 mM L-glutamine, without L-lysine, L-leucine and L-arginine, PAA Laboratories) and 15% (v/v) mESC serum substitute (Thermo Fisher Scientific), supplemented as described above plus 74 mg/L L-lysine, 52 mg/L L-leucine, and 105 mg/L arginine. The stable isotope label used was $^{13}C_6$,$^{15}N_2$-lysine (Lys-8, ≥ 99% isotope enrichment, Cambridge Isotope Laboratories). All mESCs were grown on gelatin-coated dishes using standard culturing procedures. MEFs were cultured using standard procedures as described previously [19].

Preparation of crude nuclear extracts. mESCs were harvest at approximately 70% confluence by trypsinization and nuclear extraction was carried out as described in the literature [11]. In brief, cells were washed with DPBS, swollen in 10 mM HEPES-NaOH pH 8.0, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 1% (v/v) protease inhibitor cocktail (Sigma) and homogenized by gentle passing through a small gauge hypodermic needle. Nuclei were sedimented from the lysate by centrifugation and extracted with 20 mM HEPES-NaOH pH 8.0, 25% (v/v) glycerol, 420 mM NaCl, 1.5 mM MgCl2, 0.2 mM EDTA, 0.5 mM DTT, 1% (v/v) protease inhibitor mix (Sigma). The extracts were cleared by centrifugation, aliquoted and snap frozen in liquid nitrogen. Nuclear extracts from MEFs were prepared as described previously [19].

In vitro transcription reactions. To establish suitable buffer conditions for IVT reactions, nuclear extracts were dialyzed against 20 mM HEPES-NaOH pH 8.0, 20% (v/v) glycerol, 100 mM KCl, 0.5 mM EDTA, 0.5 mM DTT, 0.5 mM PMSF using a 10 kDa MWCO microdialysis device (Pierce). Precipitates were removed by brief centrifugation and IVT reactions were set up with the cleared supernatant as described in the literature [11]. In summary, IVT reactions were set up by adding equal volumes of unlabeled and Lys-8 labeled dialyzed nuclear extracts to a reaction premix to obtain a final composition of 12 mM HEPES-NaOH pH 8.0, 12% (v/v) glycerol, 60 mM KCl, 5 mM $MgCl_2$, 0.3 mM EDTA, 0.5 mM DTT, 125 µM NTP mix (Invitrogen), 5 mM creatine phosphate, 1 unit/µL murine RNase inhibitor (New England Biolabs), 1x protease inhibitor mix (Roche Applied Science) and 10 ng/µL template DNA. The template DNA was plasmid pML(C_2AT) which contains a G-less transcription cassette under control of the TATA-containing adenovi-

rus major late promotor [20]. The reactions were incubated for 60 min at 30°C, then briefly cooled on ice, supplemented with 50 µg/mL ethidium bromide and 0.1% (v/v) Nonidet P-40 and immediately subjected to AP.

Buffer adjustments. For testing mESC TFIID stability without establishing IVT reaction conditions, nuclear extracts were slowly (30 min) stepwise diluted with Δ[NaCl] of around 50 mM using suitable dilution buffers to a final buffer composition of 20 mM HEPES-NaOH pH 8.0, 20% (v/v) glycerol, 150 mM NaCl, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% (v/v) Nonidet P-40, 0.1 mM DTT and 50 µg ethidium bromide.

Affinity purifications, immunoprecipitations and protein digestion. APs of GFP-tagged TAF6 from mESC nuclear extracts, diluted nuclear extracts, or IVT reactions were carried out using a GFP binder immobilized on agarose (ChromoTek) as published recently [21]. 0.5 -1.0 mL input was added to 20 µL anti-GFP agarose equilibrated with DPBS containing 0.25% (v/v) Nonidet P-40 and the mixture was gently rotated for 2-90 min. The agarose was then settled by brief centrifugation and washed 4 times with DPBS containing 0.1% (v/v) Nonidet P-40. Bound proteins were eluted with 50 µL 100 mM glycine-HCl pH 2.0 and neutralized with 10 µL 1 M Tris-HCl pH 8.5. Eluted proteins were proteolytically digested on centrifugal filter as described in the literature [14]. In brief, eluates were loaded on 30 kDa filter units (Millipore) under denaturing conditions achieved with 8 M urea, washed, reduced with DTT, alkylated with iodoacetamide, and digested with Lys-C (WACO). The generated peptides were eluted from the membrane, purified using C18 membrane (3M) packed into pipette tips as described previously [22], and freeze-dried prior to injection into the LC-MS system. IPs of FLAG-HA-tagged hTBP from MEF nuclear extracts were performed in tandem using anti-FLAG followed by anti-HA affinity resin in conjunction with competitive epitope elution. The protein eluates were precipitated with methanol/chloroform [23], reduced with DTT and alkylated with iodoacetamide under denaturing conditions, and digested in solution with Lys-C (Roche Applied Science) and Trypsin (Roche Applied Science). The generated peptides were subjected to C18 purification as described above and freeze-dried prior to pre-fractionation by strong cation exchange (SCX).

Strong cation exchange. Peptides generated from IPs were dissolved in solvent A (0.05% (v/v) formic acid in 1:4 acetonitrile:water) and fractionated over two ZORBAX BioSCX-Series II columns (50 mm x 0.8 mm, 3.5 µm particle size, Agilent) in series. Peptides were loaded at 50 µL/min solvent A for 10 min and eluted using a linear gradient of 1.3% (v/v) 1/min solvent B (500 mM NaCl in solvent A) at 50 µL/min. 40 fractions of 1

min each were collected and freeze-dried prior to LC-MS.

LC-MS analysis of immunoprecipitations from MEFs. SCX fractions were dissolved in 5% (v/v) formic acid and analyzed by nanoflow RP LC coupled to an LTQ FTICR mass spectrometer (Thermo Fisher Scientific). The LC was set up on an Agilent 1200 series HPLC system equipped with a 20 mm Aqua C18 (Phenomenex) trapping column (100 µm inner diameter, 5 µm particle size) and a 200 mm ReproSil-Pur C18-AQ (Dr. Maisch GmbH) analytical column (50 µm inner diameter, 3 µm particle size). Trapping was performed at 5 µL/min solvent C (0.1 M acetic acid in water) for 10 min and elution was achieved with 10 to 40% (v/v) solvent D (0.1 M acetic acid in 1:4 acetonitrile:water in either 22 or 35 min (45 and 60 min analysis time in total), followed by 38 to 100% (v/v) solvent D in 3 min and 100% solvent D for 2 min. The flow rate was passively split from 0.45 mL/min to 100 nL/min [24]. Electrospray was achieved using a distally coated fused silica emitter (360 µm outer diameter, 20 µm inner diameter, 10 µm inner diameter of emitter tip, New Objective) biased to 1.7 kV. The LTQ FTICR was operated in the data dependent mode to automatically cycle between MS and MS/MS. MS spectra in the range of 150 or 350 to 1,500 Th were acquired in the ICR analyzer with a resolution of 100,000 at 400 Th after accumulation to an automatic gain control (AGC) target value of 2,000,000 in the LTQ. The two most intense precursor ions with a charge >1+ and above an intensity of 500 were selected for collision induced dissociation (CID) under dynamic exclusion reflecting the chromatographic peak width. CID was performed in the LTQ after accumulation to an AGC target value of 10,000.

LC-MS analysis of affinity purifications from mESCs. Peptides were dissolved in 10% formic acid and analyzed by nanoflow LC coupled to an LTQ Orbitrap or LTQ Orbitrap VELOS mass spectrometer (Thermo Fisher Scientific). The LC and LC-MS interface were set up as described above, with the difference that peptides were eluted using longer gradients of 90, 120 or 180 min total analysis time. For LTQ Orbitrap analyses, MS spectra were acquired from 350 to 1,500 Th in the Orbitrap at a resolution of 60,000 at 400 Th after accumulation to an AGC target value of 500,000 in the LTQ. CID spectra were recorded in the LTQ at an AGC target of 30,000. For LTQ Orbitrap VELOS analyses, MS spectra were recorded the same way but peptides were fragmented with HCD or ETD. HCD was carried out at 35% NCE and fragments were analyzed in the Orbitrap. ETD was performed in the high pressure LIT at a reaction time of 50 ms using supplemental activation after accumulation to an AGC target value of 50,000.

Analysis of LTQ FTICR data. Raw data files were processed with Bioworks (version 3.1,

Thermo Fisher Scientific) and the subsequent data analysis was carried out using Mascot (version 2.2.1, Matrix Science). Mascot was set up to search the IPI mouse database (version 3.34, www.ebi.ac.uk/ipi) with carbamidomethyl cysteines as fixed modification and oxidation of methionines and carbamylation of peptide N-termini as variable modifications. Trypsin was specified as the proteolytic enzyme and up to two missed cleavages were allowed. The mass tolerance of the precursor ion was set to 15 ppm and the tolerance of fragment ions to 0.9 Da. Mascot results files were filtered to contain only peptides above a Mascot ion score of 15 using in house written software. Scaffold (version 2.01.02, Proteome Software) was used to validate protein identifications. Protein identifications were accepted if they could be established at greater than 99.9% probability and contained at least 2 identified peptides. Human TBP and its variants were identified following essentially the same procedure but using a manually generated database containing sequences of mouse TBP (SwissProt accession number P29037), human TBP (SwissProt accession number P20226), and those of the TBP mutants used in this study.

Analysis of LTQ Orbitrap data. Raw data files were processed with the *quant.exe* module of MaxQuant (version 1.0.13.8) [15] with Lys-C as the proteolytic enzyme, SILAC Lys-8 as isotope labeling strategy, maximally 2 missed cleavages and 3 isotopically labeled amino acids, and 6 peaks per 100 Th. The extracted peak lists **.iso*, **.sil0* and **.sil1* were searched with Mascot (version 2.2.1, Matrix Science) against a concatenated reverse version of the IPI mouse database (version 3.63, www.ebi.ac.uk/ipi) including common protein contaminants generated with MaxQuant. Oxidation of methionines was set as variable and carbamidomethyl cysteine as fixed modification; $^{13}C_6$,$^{15}N_2$-lysine was set to suite the searched peak list. The fragment ion tolerance was 0.5 Da and the precursor tolerance was set to reflect the instrument performance, which varied between 5 and 20 ppm. Quantification was carried out without further manual validation using the MaxQuant *ident.exe* module using a peptide false discovery rate (FDR) of 0.01%. A minimal ratio count of 2 was required for quantification and the re-quantify option was used.

Analysis of LTQ Orbitrap VELOS data. Raw data files acquired on the LTQ Orbitrap VELOS were analyzed with Proteome Discoverer (version 1.2, Thermo Fisher Scientific). For HCD data, Mascot (version 2.3, Matrix Science) was used to search the IPI mouse database (version 3.75, www.ebi.ac.uk/ipi) specifying Lys-C as enzyme with 1 missed cleave allowed. The instrument type was set to ESI-QUAD-TOF. The Mascot-internal target-decoy strategy was used for FDR calculation. The precursor mass tolerance was set to 5

ppm and the fragment ion tolerance to 0.05 Da. Cysteine carbamidomethylation was specified as static and $^{13}C_6$,$^{15}N_2$-lysine as dynamic modification. For ETD data, a non-fragment ion filter was applied which removes the precursor peak within a window of 4 Da and charge-reduced precursors and neutral losses of charged-reduced precursors up to 120 Da within a window of 2 Da. The Mascot search was performed as described for HCD data with the instrument type set to ETD trap and the fragment ion tolerance set to 0.5 Da. Both HCD and ETD data were quantified without further manual validation using a customized SILAC 2plex quantification scheme with the following settings: mass precision 2 ppm, 1 min retention time tolerance for isotope pattern doublets, 1 missing channel allowed. For H/L calculation a maximal allowed fold change of 1,000 was allowed, single peak *quan* channels were allowed, and missing *quan* values were allowed to be replaced with the minimally determined *quan* value. The results were filtered for a FDR of 1%.

4 Acknowledgements

The authors thank Michiel Vermeulen for sharing the GFP affinity purification protocol and for help with the FASP procedure, and Marijke Baltissen for help with mESC cultivation. This work was financed by the NPC Horizon Program (grant number 050-71-050) and by the Netherlands Proteomics Centre which is part of the Netherlands Genomics Initiative.

5 References

[1] Kocher, T.*, et al.*, *Nat Methods* **2007**, *4*, 807
[2] Gingras, A. C.*, et al.*, *Nat Rev Mol Cell Biol* **2007**, *8*, 645
[3] Gavin, A. C.*, et al.*, *Nature* **2006**, *440*, 631
[4] Krogan, N. J.*, et al.*, *Nature* **2006**, *440*, 637
[5] Rothbauer, U.*, et al.*, *Mol Cell Proteom* **2008**, *7*, 282
[6] Hubner, N. C.*, et al.*, *J Cell Biol* **2010**, *189*, 739
[7] Cler, E.*, et al.*, *Cell Mol Life Sci* **2009**, *66*, 2123
[8] Thomas, M. C.*, et al.*, *Crit Rev Biochem Mol Biol* **2006**, *41*, 105
[9] Bryant, G. O.*, et al.*, *Genes Dev* **1996**, *10*, 2491
[10] Klejman, M. P.*, et al.*, *Nucl Acid Res* **2005**, *33*, 5426

[11] Dignam, J. D.*, et al.*, *Nucl Acid Res* **1983**, *11*, 1475
[12] Nagy, Z.*, et al.*, *Oncogene* **2007**, *26*, 5341
[13] Poser, I.*, et al.*, *Nat Methods* **2008**, *5*, 409
[14] Wisniewski, J. R.*, et al.*, *Nat Methods* **2009**, *6*, 359
[15] Cox, J.*, et al.*, *Nat Biotechnol* **2008**, *26*, 1367
[16] Pointud, J. C.*, et al.*, *J Cell Sci* **2003**, *116*, 1847
[17] Cheng, Y.*, et al.*, *Mol Cell Biol* **2007**, *27*, 2582
[18] Martianov, I.*, et al.*, *Mol Cell* **2001**, *7*, 509
[19] Mengus, G.*, et al.*, *Embo J* **2005**, *24*, 2753
[20] Sawadogo, M.*, et al.*, *Proc Natl Acad Sci USA* **1985**, *82*, 4394
[21] Vermeulen, M.*, et al.*, *Cell* **2010**, *142*, 967
[22] Rappsilber, J.*, et al.*, *Nat Prot* **2007**, *2*, 1896
[23] Wessel, D.*, et al.*, *Anal Biochem* **1984**, *138*, 141
[24] Pinkse, M. W. H.*, et al.*, *Anal Chem* **2004**, *76*, 3935

Chapter 5

Gaining efficiency by parallel quantification and identification of iTRAQ-labeled peptides using HCD and decision tree guided CID/ETD on an LTQ Orbitrap

Nikolai Mischerikow[1,2], Pim van Nierop[3], Ka Wan Li[3], Hans-Gert Bernstein[4], August B. Smit[3], Albert J. R. Heck[1,2,5], and A. F. Maarten Altelaar[1,2]

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

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

[3] Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, VU University Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

[4] Department of Psychiatry, University of Magdeburg, Leipziger Strasse 44, 39120 Magdeburg, Germany

[5] Centre for Biomedical Genetics, Padualaan 8, 3584 CH Utrecht, The Netherlands

Supplementary material referred to in this chapter can be found accompanying the publication of this work in *Analyst*, 2010, volume 135, issue 10, pages 2643-2652.

1 Summary

Isobaric stable isotope labeling of peptides using isobaric tagging for relative and absolute quantification (iTRAQ) is an important method for mass spectrometry- (MS) based quantitative proteomics. Traditionally, quantitative analysis of iTRAQ-labeled peptides has been confined to beam-type instruments because of the weak detection capabilities of ion traps for low mass ions. Recent technical advances in fragmentation techniques on linear quadrupole ion trap mass analyzers (LITs) and the commercial hybrid LIT-orbitrap, the LTQ Orbitrap (Thermo Fisher Scientific), allow circumventing this limitation. Namely, pulsed Q collision induced dissociation (PQD) and higher energy collisional dissociation (HCD) facilitate iTRAQ analysis on these instrument types. Here we report a method for iTRAQ-based relative quantification on the ETD-enabled LTQ Orbitrap XL, which is based on parallel peptide quantification and peptide identification. iTRAQ reporter ion generation is performed by HCD, while CID and ETD provide peptide identification data in parallel in the LIT. This approach circumvents problems accompanying iTRAQ reporter ion generation with ETD and allows quantitative, decision tree based CID/ETD experiments. Furthermore, the use of HCD solely for iTRAQ reporter ion readout significantly reduces the number of ions needed to obtain informative spectra, which significantly reduces the analysis time. Finally, we show that integration of this method, both with existing CID and ETD methods as well as with existing iTRAQ data analysis workflows, is simple to realize. By applying our approach to the analysis of the synapse proteome from human brain biopsies, we demonstrate that it outperforms a latest generation MALDI TOF/TOF instrument, with improvements in both peptide and protein identification and quantification. Conclusively, our work shows how HCD, CID and ETD can be beneficially combined to enable iTRAQ-based quantification on an ETD enabled LTQ Orbitrap XL.

2 Introduction

The relative quantitative comparison of differences in peptide and protein abundances, between two or more samples, has become an important experiment type in MS-based biomedical research [1,2]. For this purpose, numerous analytical strategies have been devised that are based on the incorporation of stable heavy isotopes into proteins or peptides of one sample, which is then mixed and analyzed in parallel with a sample labeled with light isotopes so to provide internal reference to one another. Both meta-

bolic labeling of proteins, such as ^{15}N-labeling and stable isotope labeling with amino acids in cell culture (SILAC), which incorporate stable isotopes in cell culture [3] or in the whole organism [4-7], as well as chemical labeling of peptides after digestion, such as isotope-coded affinity tagging (ICAT) [8], iTRAQ [9] or dimethylation labeling [10,11], are frequently used. Of these, the fully tandem mass spectrometry- (MS/MS) based iTRAQ quantification strategy allows between 4- and 8-fold multiplexing, which is not easily achieved by any other method, making it a popular application particular in clinical proteomics. Quantification by iTRAQ differs from other quantitative proteomics approaches, which are based on precursor ion intensities, by the production of iTRAQ-specific reporter ions during fragmentation. The different iTRAQ labels are isobaric as they consist of a reporter and a balance moiety, which together result in an equal mass for every label. Peptides from multiple samples are differentially labeled and are selected for MS/MS as a single precursor ion. Upon fragmentation by CID, 4-8 specific reporter ions are released and used for quantification. Advantages of the iTRAQ strategy are beneficial signal intensity on the MS level, as the label is isobaric, as well as good quantification accuracy due to low sample noise on the MS/MS level.

From the analytical perspective, iTRAQ is traditionally performed on beam type rather than trapping type mass spectrometers. The latter instruments cannot trap iTRAQ reporter ions generated during MS/MS of typical proteolytic peptides (600-800 Th with 2-3 charges) because the voltage commonly applied for optimal resonant excitation of the precursor shifts the low mass-to-charge ratio (m/z) fragments, including the iTRAQ reporter ions, beyond the stability limit, resulting in their ejection from the trap. However, two technical developments offer the possibility to do iTRAQ on an ion trap instrument. One is the implementation of PQD [12] on the commercial LIT, the LTQ (Thermo Fisher Scientific) and the hybrid LTQ Orbitrap [13] to overcome the detection limit for low mass ions posed by regular resonant CID. PQD is an alternative fragmentation method to CID in LITs allowing low m/z fragment ions to be trapped [14]. A more recent technical advance, which lifts this limitation in a different way, is the implementation of HCD on the LTQ Orbitrap. HCD allows MS/MS fragmentation in a dedicated collision cell comparable to time-of-flight (TOF) precursor fragmentation [15].

The applicability of both PQD and HCD for iTRAQ-based quantification has been the subject of several recent studies. Griffin *et al.* [16] showed that PQD on a LTQ performs comparable to CID on the commercial QSTAR hybrid quadrupole TOF instrument (Applied Biosystems) with respect to quantification performance. Bantscheff *et al.* [17] described the implementation of PQD on a LTQ Orbitrap for the purpose of iTRAQ

analysis, and showed that it outperforms the commercial Q-Tof Ultima quadrupole TOF hybrid instrument (Micromass) in the number of quantified proteins. They also showed that PQD is faster and more sensitive than HCD on a LTQ Orbitrap equipped with a prototype HCD cell. Another study by Zhang *et al.* [18] demonstrated that iTRAQ-labeled phosphorylated peptides can be analyzed quantitatively by HCD on a LTQ Orbitrap. In these studies, the need for optimization of PQD or HCD to yield both sequence-informative fragment ions and iTRAQ reporter ions in a single fragmentation event, results in a compromise between optimal settings for identification and optimal settings for quantification. Although in the above studies HCD and PQD methods were able to outperform older generation TOF instruments, recent studies showed that PQD and HCD are significantly slower, less sensitive and produce less informative fragment ion spectra than CID analysis on current ion trap instruments [19-21]. Therefore, an alternative strategy was developed combining quantification by HCD with identification using CID [19,20].

Recently, electron transfer dissociation (ETD) of peptides was introduced as an alternative peptide fragmentation method [22,23]. In ETD, the interaction between near-thermal electrons and positively charged peptides results in dissociation of the peptide N-Cα bond, producing *c*- and *z*-ions. Since ETD prefers larger and more basic peptides, which attain multiple charges during electrospray ionization (ESI), as compared to CID, the two techniques are considered largely complementary [24,25]. Reduced fragmentation efficiency of ETD for doubly charged peptides is compensated by the use of supplemental activation (ETcaD) [26]. With the implementation of ETD, two groups studied the dissociation of iTRAQ-labeled peptides upon ETD. Both Han *et al.* [27] and Phanstiel *et al.* [28] found that the iTRAQ label does not release reporter ions as straightforward upon ETD as it does upon CID, which is caused by the different cleavage site within the iTRAQ label. While in CID only unique reporter ions are observed after cleavage of the iTRAQ label, N-Cα cleavage upon ETD releases multiple fragment ions. Although unique iTRAQ reporters are observed with ETD, the high multiplexing capability of iTRAQ can only be used upon ETD followed by CID [29]. Circumventing these issues, a recent workflow by Yang *et al.* [30] described sequential PQD and ETD on a LTQ for the quantification of phosphorylated peptides with, whereby both PQD and ETD contribute peptide sequence information but only PQD contributing to iTRAQ quantification.

Here, we present a parallel acquisition setup which implements ETD for iTRAQ-based quantification on a HCD- and ETD-enabled LTQ Orbitrap XL. iTRAQ reporter ions are generated using HCD while sequence information is gathered in parallel using ETD or

CID. The selection of ETD or CID is made on the fly according to the m/z of the precursor [31]. We demonstrate the performance of our workflow by the analysis of iTRAQ labeled peptides derived from synaptic preparations of human brain biopsies, and we show that in terms of identified and quantified peptides we outnumber the commercial 4800 Proteomics Analyzer MALDI TOF/TOF instrument (AB SCIEX), which has recently been shown to have equal performance as the modern ESI-Q-TOF platform QSTAR Elite (Applied Biosystems) [32].

Our workflow requires only very little optimization, which is a benefit compared to the PQD-based approach, and allows on the fly, decision tree-based CID/ETD analysis, since the same iTRAQ reporter ions are generated irrespective of CID or ETD precursor fragmentation. Furthermore, the low number of ions needed for the generation of the iTRAQ reporter ions allows the experiment to be conducted on similar time scales as PQD or HCD alone. Our method results in a significant gain in efficiency in protein quantification by conducting iTRAQ quantification in the HCD collision cell and proficient parallel identification using decision tree-guided CID/ETD peptide fragmentation.

3 Results and Discussion

To implement iTRAQ-based peptide quantification on an ETD- and HCD-enabled LTQ Orbitrap, we performed peptide quantification in parallel with peptide identification. Peptide identification was performed by the common combination of precise precursor m/z determination using the orbitrap analyzer and precursor sequencing by CID or ETD in the LIT (Figure 1). For peptide quantification, the same precursors selected for fragmentation by CID or ETD were selected for fragmentation by HCD in the octopole collision cell, and fragment ions were analyzed in the orbitrap. For this purpose, the instrument was operated in data-dependent acquisition mode with three HCD and in parallel CID or ETD data-dependent MS/MS scans succeeding the MS scan. The HCD scans were relatively fast because we set the instrument to an automatic gain control (AGC) target value of only 50,000 for HCD fragmentation, 4 to 6 times smaller than previously reported [19,20]. A higher AGC target value was not required because the HCD spectrum was only used to read out iTRAQ reporter ions and not to provide high quality fragment ion information. Furthermore, since in the orbitrap mass analyzer scan time increases proportionally with increasing resolution and resolution increases proportionally with decreasing m/z, HCD fragment ion read out with a full width at half maximum

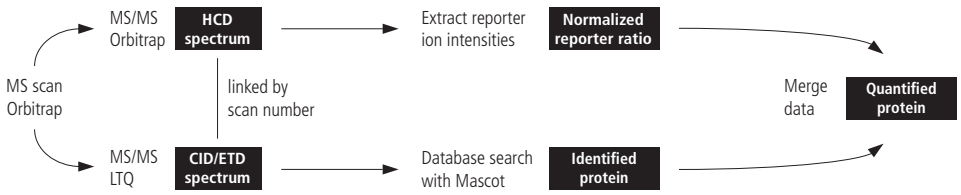


Figure 1. Workflow illustrating how a precursor selected for fragmentation from the MS scan is fragmented by both HCD and CID/ETD and how sequence-informative and quantitative information is processed separately leading to a protein quantification.

(FWHM) resolution of 7,500 at 400 Th also reduced the scan time, while still obtaining a resolution of 16,000 at the iTRAQ reporter ions m/z. This high resolution readout is important to be able to resolve the iTRAQ reporters in the presence of possible interfering compounds [33]. The MS scan was acquired at 30,000 FWHM resolution, typically leading to scan times of around 750 milliseconds.

To assess our method in analytical terms, we analyzed three relatively simple peptide mixtures, namely tryptic BSA peptides labeled with 4 out of 8 iTRAQ 8-plex channels (113, 115, 117 or 119), mixed in ratios 1:1:1:1, 1:1:2:4 or 2:2:3:4. As our experimental setup is based on separate identification and quantification events, we did not have to tune the HCD event to yield balanced quantification and identification information from the HCD spectrum, but rather varied HCD collision energy to produce mainly iTRAQ reporter ions. For this purpose, we analyzed each of the three ratios at 35, 45, 55 and 65% normalized collision energy (NCE). All other parameters were held constant, including the AGC target value for HCD, which we set to 50,000 with a single microscan. HCD spectra were recorded in profile mode. For peptide identification in these experiments, we used CID.

A typical spectrum pair on the MS/MS level, consisting of a CID and a HCD spectrum, recorded for the BSA peptide LGEYGFQNALIVR, is shown in Figure 2. While the CID spectrum provides the expected b- and y-type ion series used for peptide identification, the HCD spectrum contains almost solely iTRAQ reporter ions. All other spectra displayed similar characteristics, with the HCD spectra ideal for detecting iTRAQ reporter ions, and the CID/ETD spectra ideal for parallel identification.

The analysis of the BSA LC-MS runs, which are shown for ratios 1:1:1:1 and 1:1:2:4 in Figure 3, demonstrate that the variation of collision energy has little or no influence on the relative variance of reporter ion intensities (Figures 3A and 3B), which is mea-

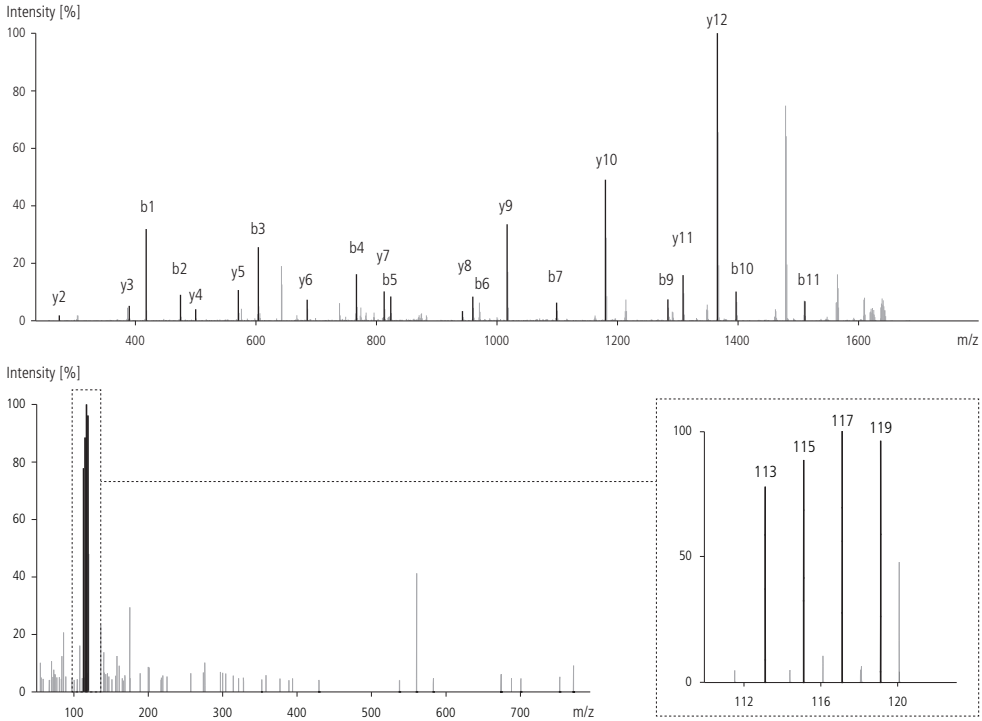


Figure 2. Fragment ion spectra of the BSA peptide LGEYGFQNALIVR. The CID spectrum (top) contains the sequence-informative fragment ions, while the HCD spectrum (bottom), produced at 55% NCE and recorded in the orbitrap at an effective FWHM resolution greater than 16,000 around 100 Th, contains only the iTRAQ reporter ions (inset).

sured as standard deviation of the normalized reporter ion intensities of all quantified peptides. The variance is around 10% for every iTRAQ channel, at all collision energies measured. These variances propagate into 10% uncertainty in reporter ion ratios (Figures 3C-F), again independent of collision energy. Note that the measured mixing ratios deviate from the nominal mixing ratios of 1:1:1:1 (observed as 1:1.2:1.6:1.4) and 1:1:2:4 (observed as 1:1.2:3.1:6) due to differences in the concentration of iTRAQ-labeled preparations. When dividing experimental ratios displayed in Figure 3F by the experimental ratios shown in Figure 3E the ratios are 1:1:2:4 as expected. In a typical large scale quantitative proteomics experiment such as described in the next section, the influence of differences in overall peptide concentration between iTRAQ-labeled samples on the observed protein regulation is removed by a normalization procedure that has the critical assumption that the majority of proteins does not differ between samples [34]. Because BSA is the only protein, normalization was not applied in this

A Reporter ion intensities 1:1:1:1

NCE	113	115	117	119
35	0.80 ± 0.10	1.02 ± 0.16	1.14 ± 0.10	1.10 ± 0.07
45	0.79 ± 0.07	0.99 ± 0.08	1.17 ± 0.08	1.11 ± 0.10
55	0.78 ± 0.11	1.01 ± 0.09	1.19 ± 0.23	1.11 ± 0.11
65	0.76 ± 0.06	0.96 ± 0.06	1.21 ± 0.11	1.15 ± 0.09

B Reporter ion intensities 1:1:2:4

NCE	113	115	117	119
35	0.48 ± 0.05	0.58 ± 0.02	1.36 ± 0.07	2.65 ± 0.07
45	0.45 ± 0.08	0.58 ± 0.08	1.46 ± 0.20	2.75 ± 0.32
55	0.46 ± 0.08	0.58 ± 0.05	1.39 ± 0.19	2.79 ± 0.28
65	0.46 ± 0.07	0.55 ± 0.08	1.45 ± 0.15	2.83 ± 0.24

C Reporter ion Ratios 1:1:1:1

NCE	113/113	115/113	117/113	119/113
35	1.00 ± 0.15	1.28 ± 0.19	1.42 ± 0.14	1.37 ± 0.12
45	1.00 ± 0.09	1.25 ± 0.11	1.48 ± 0.11	1.40 ± 0.12
55	1.00 ± 0.16	1.30 ± 0.14	1.52 ± 0.26	1.42 ± 0.16
65	1.00 ± 0.09	1.26 ± 0.09	1.59 ± 0.13	1.52 ± 0.11

D Reporter ion Ratios 1:1:2:4

NCE	113/113	115/115	117/113	119/113
35	1.00 ± 0.06	1.20 ± 0.05	2.83 ± 0.09	5.50 ± 0.08
45	1.00 ± 0.11	1.30 ± 0.11	3.28 ± 0.21	6.16 ± 0.33
55	1.00 ± 0.12	1.28 ± 0.10	3.06 ± 0.21	6.13 ± 0.29
65	1.00 ± 0.10	1.21 ± 0.11	3.18 ± 0.17	6.22 ± 0.25

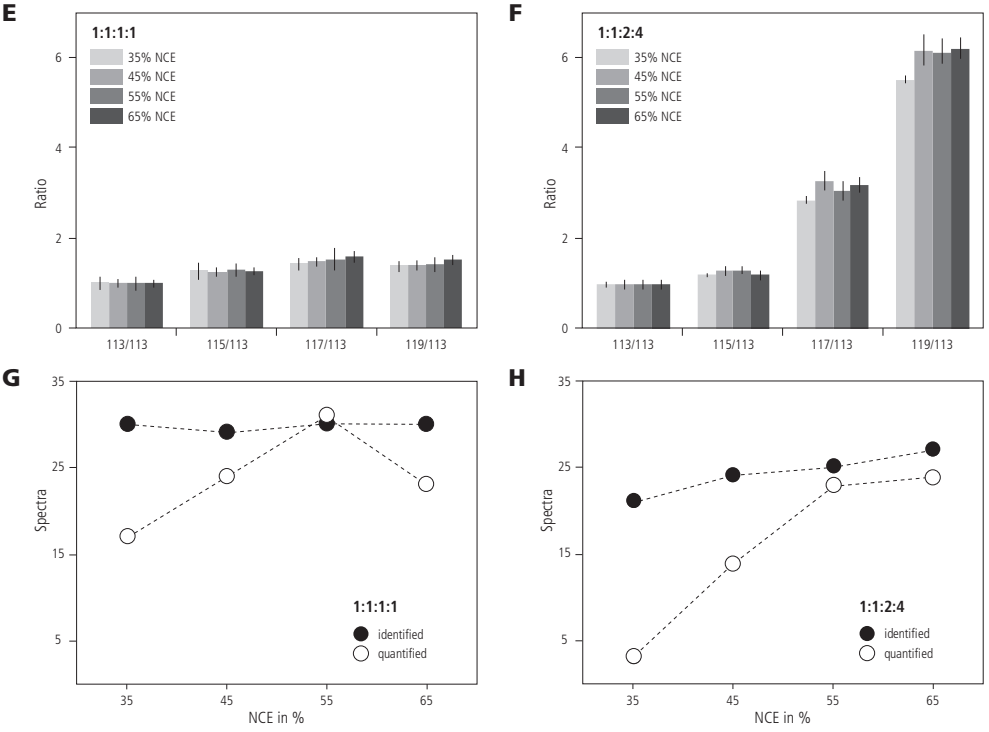


Figure 3. HCD CID analysis of 4-plex iTRAQ-labeled BSA mixed in ratios 1:1:1:1 (panels on the left) and 1:1:2:4 (panels on the right), at different HCD collision energies. See text for details.

experiment since this by definition would remove any observations of changes in BSA abundance. What is affected by collision energy variation is the number of quantified spectra (Figures 3G and 3H). This number increases with collision energy, because the number of spectra that lack a reporter ion decreases. These spectra, although the other

channels are present, were excluded from the quantification. At all ratios tested, 55% NCE during HCD resulted in the highest number of quantifiable spectra. In contrast, the number of identified spectra, as it is not based on the HCD event, did not change with collision energy.

To demonstrate that ETD can be easily integrated into our workflow, we also analyzed the BSA mixtures with the HCD CID/ETD method. For peptide identification we used either CID or ETD, whereby the fragmentation type was chosen for every single precursor according to the decision tree logic that was recently published by Swaney *et al.* [31]. This logic is based on the observation that peptides with a charge greater than 2+ up to a certain m/z threshold are more likely to be identified from their ETD rather than their CID spectrum. For peptides above these m/z thresholds, and for all unmodified 2+ peptides, irrespective of their m/z value, CID provides better sequence information [31,35]. In comparison, both methods, HCD CID as well as HCD CID/ETD, performed comparably well on this simple peptide mixture.

To test our methods in a biologically relevant experimental setting, we analyzed peptide mixtures generated by tryptic digestion of synaptic preparations of four *post mortem* human brain samples with both HCD CID and HCD CID/ETD. The four peptide mixtures were labeled with the 4 out of 8 iTRAQ 8-plex channels 114, 116, 118 and 121, pooled, and then pre-fractionated by SCX prior to MS analysis. Selected SCX fractions were analyzed by LC-Ms on the LTQ Orbitrap using the HCD CID and HCD CID/ETD method, and for a one-to-one comparison also by off-line LC followed by MALDI MS analysis on the 4800 Proteomics Analyzer.

Data processing was performed using the iTRAQ analysis platform described in the experimental section. This software, which had initially been developed for iTRAQ data acquired on the 4800 Proteomics Analyzer, was adapted to handle HCD data as well. The software was linked to a Mascot server for peptide identification. In the first round of data processing, proteins were identified based on the peptide sequence matches generated by Mascot (peptide ion score > peptide homology threshold established at $p < 0.01$) using relatively strict criteria; namely at least two unique peptides were required per protein and these peptides could not be assigned to any other protein. Using these criteria, these peptides were called identified peptides. In the second round of data processing, proteins were quantified using a nearly identical repetition of the first round, with two changes. The major adaptation was that the peptide ion score was not required to be bigger than the peptide homology threshold, which effectively yields a

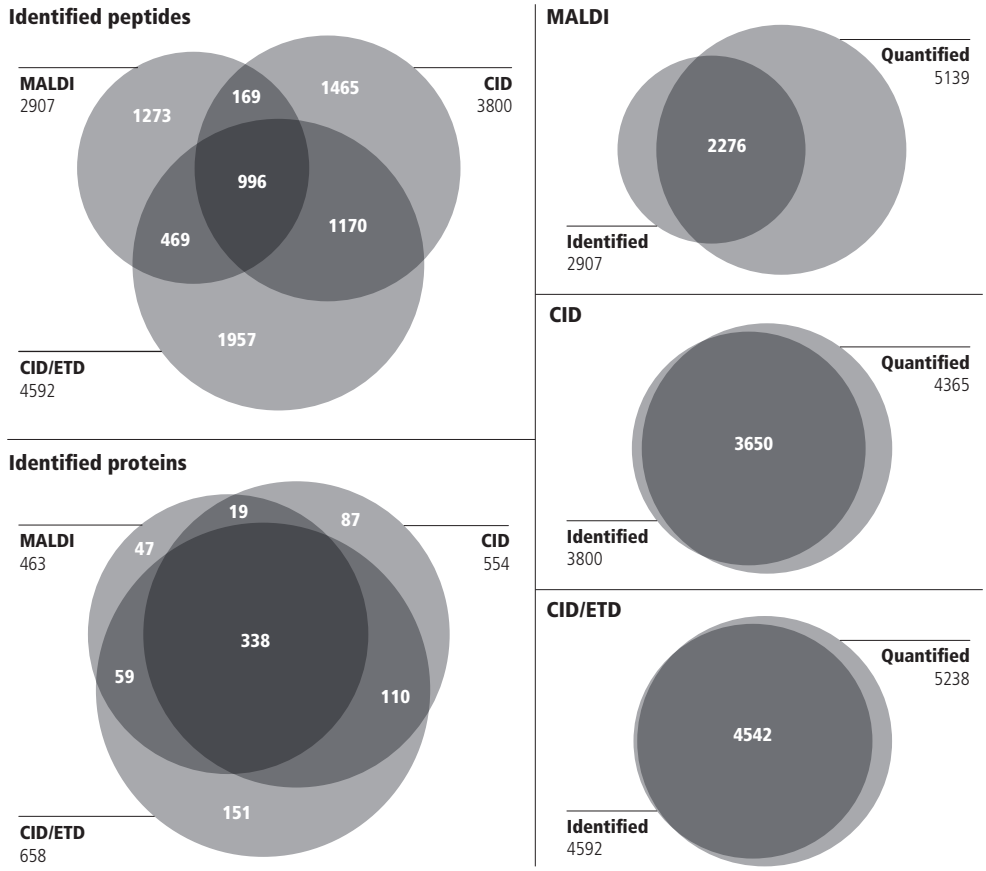


Figure 4. Comparison of the identification and quantification results of the three methods. As expected the identification overlap between the three different methods at the protein level is much larger than at the peptide level. The overlap between identified and quantified peptides for each method (right panels) reveals that both HCD methods outperform the MALDI method based on the higher identification efficiency. The quantification efficiency is close to 100% for both HCD methods. Note that the number of quantified peptides exceeds the number of identified peptides because of the different criteria applied for the identification and quantification process (see experimental section for details).

larger number of peptides. The minor adaptation was that in the HCD spectrum, which accompanied every CID or ETD spectrum and which was used to read out reporter ion intensities, the reporter ion intensities were required to be greater than 5000. Using these criteria, the peptides that could be quantified and assigned to a protein were called quantified peptides. Finally, the identified peptide pool was used to create a list of proteins, which were then quantified using the quantified peptide pool. The rationale behind this data analysis strategy was to achieve very strict protein identifications, but

once a protein was identified to gather as much quantitative information as possible by using more relaxed criteria for selecting peptides for quantification. This procedure implies that the number of quantifiable peptides is potentially larger than the number identifiable peptides.

We confidently identified 2,907 peptides by MALDI, 3,800 peptides by the HCD CID method, and 4,592 peptides by the HCD CID/ETD method. Figure 4 shows that for any of the three analyses, around 40% of the identified peptides were unique in a single analysis, and around 1,000 peptides were identified by all three methods. It also shows that HCD CID/ETD identified most peptides which were not identified by the other analyses. One of the advantages of the HCD CID and even more so the HCD CID/ETD experiment over the MALDI experiment is the possibility to identify peptides in the later SCX fractions. While the MALDI analysis does not add peptide identifications after SCX fraction 23, because of the size and properties of the peptides, especially our HCD CID/ETD approach identified peptides up to SCX fraction 27, which consisted mainly of peptides with 4, 5 and more charges. As illustrated in Figure 4, for the MALDI method around 80% of the peptides (2,276), which fit the identification criteria could also be quantified, as compared to more than 95% of the peptides for both the HCD CID (3,650) and the HCD CID/ETD (4,542) methods, showing a higher efficiency in quantification for the latter two methods. At the level of protein identification, the overlap between the three analyses is much greater, where 463 (by MALDI), 554 (by HCD CID) and 658 (by HCD CID/ETD) proteins were identified, with 338 proteins found in all three analyses. For nearly all identified proteins quantitative information could be inferred, with only 11 proteins not quantifiable by MALDI and 1 protein not quantifiable with HCD CID (see supplementary material). Although the majority of proteins detected in the arbitrarily selected combinations of brain samples were found to be equally abundant, we identified a relatively small number of proteins (72 by MALDI, 55 by HCD CID, and 115 by HCD CID/ETD) with a significantly different abundance ($p < 0.01$, Figure 5A). We used these data to investigate whether the three methods used here gave comparable quantitative results. To compare proteins found to be significantly different in two analyses, the logarithmic ratios of two arbitrarily selected label combinations, namely 114:116 and 118:121, were compared pairwise (Figures 5B-D). These ratios correlate very well, both when comparing HCD CID to HCD CID/ETD (Figure 5B) and when comparing either of the two orbitrap analyses to the MALDI TOF/TOF analysis (Figures 5C and 5D). In all three cases, the correlation coefficient was found to be around 0.95.

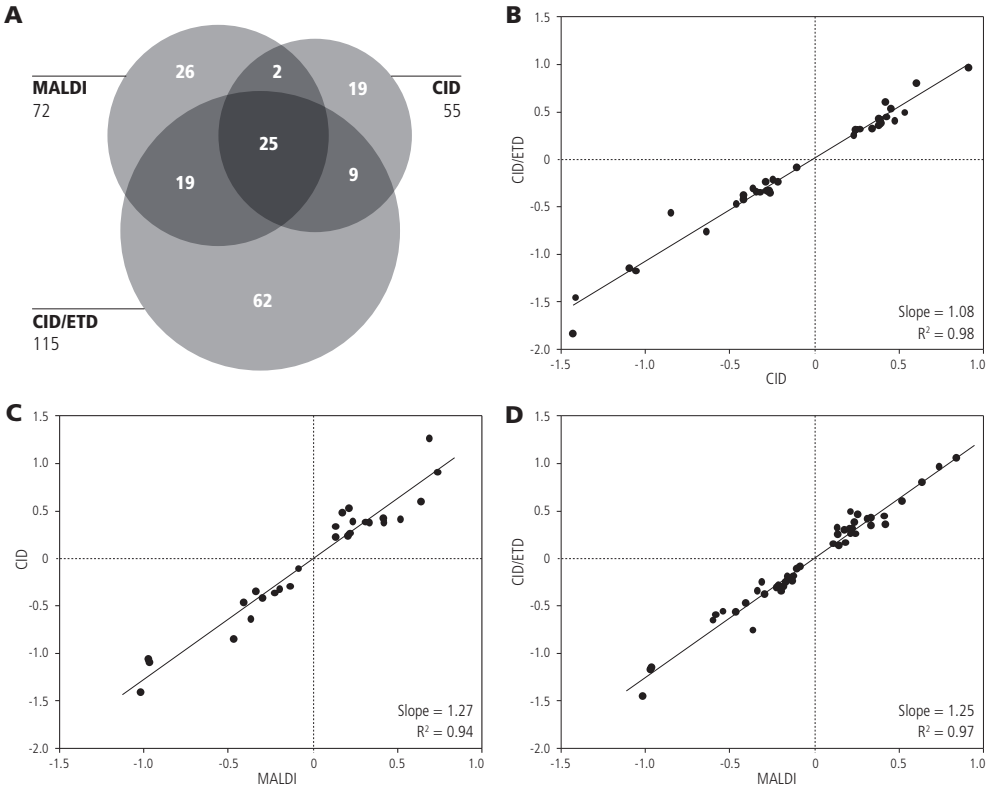


Figure 5. A, number of proteins with different abundance (measured with a significance threshold below 0.01) in the arbitrarily selected channels 114 versus 116 and 118 versus 121. B, C and D, pair-wise comparison of protein abundances expressed in logarithmic ratios. Each dot in a diagram represents the logarithmic ratios 114/116 or 118/121 of a single protein found to be significantly differentially abundant in two of the three analyses, which can be either HCD CID/ETD versus HCD CID (B), HCD CID versus MALDI (C), or HCD CID/ETD versus MALDI (D). In all three diagrams, a very strong linear correlation between the quantification methods can be observed.

From our protein identification data we can conclude, as expected, that the human synapse proteome resembles the rodent synapse proteome [34]. Proteins that are known to be present in the pre-synaptic terminal and post-synaptic compartment were well represented in our data, including synaptic vesicle proteins, adhesion molecules, scaffolding proteins, receptors, ion channels, signaling proteins and proteins involved in trafficking. Mitochondrial proteins, metabolic proteins and structural proteins that are commonly observed in rodent synapse proteomes were also detected. Furthermore, the observed protein expression profiles between the four human brain samples were very comparable with only a low number of proteins (around 10%) showing significant differences

between samples, mainly clustered around a logarithmic ratio of 0.5 and a maximum at 1.5. These results indicate that the brain biopsies as used here in this study might be valuable control samples in future studies of human brain disease. Protein and peptide identifications can be found in the supplementary material.

To conclude, we present here a workflow that allows quantitative analysis of iTRAQ labeled peptides, by parallel peptide quantification and peptide identification. The workflow is easily set up on an ETD-enabled LTQ Orbitrap XL and allows decision tree-guided CID/ETD for peptide identification. We show that our method outperforms a MALDI TOF/TOF instrument in terms of peptide and protein identification as well as protein quantification. Utilizing HCD solely for the determination of the iTRAQ reporter ions circumvents long ion accumulation times, generates the same iTRAQ reporter ions in both CID and ETD peptide identification events and leads to very high quantification efficiency. Finally, we demonstrate that our workflow is applicable to the quantitative analysis of the synapse proteomes from *post mortem* human brains. In principle, this method can also be used to quantitatively characterize the human synapse proteome isolated from patients with brain disorders, which should give insights into the molecular mechanisms of these disorders.

4 Experimental Procedures

Preparation of BSA samples. 500 µg BSA were solubilized in 100 µL 0.5 M triethylammonium bicarbonate pH 8.5 with 10 µL 50 mM tris-(2-carboxyethyl) phosphine. After incubation for 1 h at 55°C, 5 µL 200 mM methyl-methanethiosulfonate was added and mixed for 10 min. Subsequently, Trypsin was added and incubated overnight at 37°C. The sample was split into four equal aliquots, and isopropanol plus 1 unit of iTRAQ 8-plex reagent were added (channels 113, 115, 117 and 119). After incubation for 2 h at room temperature, samples were centrifuged and the supernatant was dried in a vacuum centrifuge. BSA peptides labeled with 4 out of 8 iTRAQ 8-plex channels (113, 115, 117 or 119) were diluted to 10 fmol/µL and mixed in ratios 1:1:1:1, 1:1:2:4 or 2:2:3:4, with the lowest concentration being 20 fmol of labeled BSA injected on column.

Preparation of human brain samples. The research presented here was given written approval by the ethics commission of the University of Magdeburg, Germany. Samples from the left dorsolateral frontal cortex were obtained from four psychiatric healthy

persons as described [36]. The specimens were collected at different times *post mortem* (12, 17, 21 and 22 h), quickly frozen in liquid nitrogen, and stored at -80°C until further processing. Isolation of synaptic membranes from these brains as well as labeling with iTRAQ reagents was performed according to the same protocol used for the preparation of rodent samples [34]. In brief, synaptic membranes were solubilized in 30 µL 0.85% (w/v) RapiGest (Waters). After reduction and alkylation of cysteine residues, the proteins were digested with Trypsin at 37°C overnight. The four peptide samples were tagged with 4 out of 8 iTRAQ 8-plex reagents 114, 116, 118 and 121, respectively, pooled, and then fractionated by SCX on a 2.1 x 150 mm polysulfoethyl A column (PolyLC) with a linear gradient of 0-500 mM KCl in 20% (v/v) acetonitrile and 10 mM KH_2PO_4 pH 2.9 over 25 min at a flow rate of 200 µL/min. SCX fractions were collected (36 in total) and further analyzed without further treatment by online LC-MS/MS or by off-line LC separation followed by MS/MS.

RP LC and MALDI TOF/TOF analyses. For the MALDI MS experiment, SCX fractions were separated on an analytical capillary C18 column (150 mm x 100 µm inner diameter) at 400 nL/min using a linear increase in concentration of acetonitrile from 5 to 50% (v/v) in 90 min and to 90% in 10 min. The eluent was mixed with matrix (7 mg of re-crystallized α-cyano-hydroxycinnamic acid in 1 mL 50% (v/v) acetonitrile, 0.1% (v/v) trifluoroacetic acid, 10 mM ammonium dicitrate) delivered at a flow rate of 1.5 µL/min, deposited off-line to a metal target (Applied Biosystems) every 15 seconds for a total of 384 spots. The sample was analyzed on an 4800 Proteomics Analyzer (Applied Biosystems) and the data were processed as previously reported [34].

RP LC and LTQ Orbitrap XL analyses. Analysis of SCX fractions was performed on a LC-coupled LTQ Orbitrap XL, equipped with an ETD source (Thermo Fisher Scientific). An Agilent 1200 series HPLC system was equipped with a 20 mm Aqua C18 (Phenomenex) trapping column (100 µm inner diameter, 5 µm particle size) and a 400 mm ReproSil-Pur C18-AQ (Dr. Maisch GmbH) analytical column (50 µm inner diameter, 3 µm particle size). Trapping was performed at 5 µL/min solvent A (0.1 M acetic acid in water) for 10 min, and elution was achieved with a gradient of 10 to 35% (v/v) solvent B (0.1 M acetic acid in 1:4 acetonitrile : water) in a total analysis time of 120 min. For analysis of BSA digests, the total analysis time was shortened to 45 min. During elution, the flow rate was passively split to 100 nL/min. Fused silica emitters (New Objective, 10 µm tip inner diameter) biased to 1.7 kV were used for nano-electrospray. The LTQ Orbitrap was operated in data dependent mode to automatically switch between MS and MS/MS. MS spectra in the range of m/z 350-1500 were acquired in the orbitrap at a FWHM

resolution of 30,000 after accumulation to an AGC target value of 500,000 in the linear ion trap with 1 microscan. The three most abundant precursor ions were selected for fragmentation by HCD followed by CID or ETD with an isolation width of 3 Th. If not stated differently, HCD was performed at 55% NCE in the dedicated collision cell at the back end of the C-trap, after accumulation to an AGC target value of 50,000 in the linear ion trap with 1 microscan, and fragments were read out in the Orbitrap at a FWHM resolution of 7,500, resulting in an average scan time of 0.7 s. CID and ETD were performed in the linear ion trap after accumulation to an AGC target value of 50,000 with 1 microscan (average total scan time 0.3 and 0.5 s. respectively). HCD and CID reaction time was set to 30 ms. The ETD reagent was fluoranthene which was accumulated in the ion trap to an AGC target value of 20,000. Charge state-dependent ETD reaction time was set to 50 ms for 2+ precursor ions, and supplemental activation [26] was used. The LTQ Orbitrap XL was set to automatically switch between CID and ETD, based on a combination of peptide m/z and charge state. In all experiments, fragmented precursors were dynamically excluded from further fragmentation for 30 seconds within a mass window of 60 ppm.

Data processing. In case of LTQ Orbitrap XL data, raw files were processed with Proteome Discoverer 1.0 (Thermo Fisher Scientific) to extract separate DTA files for HCD, CID and ETD spectra. In case of ETD spectra, peaks resulting from non-fragmented precursor, charge-reduced precursors, and neutral losses from the charge reduced precursors were automatically removed by the software. In case of 4800 Proteomics Analyzer data, monoisotopic peak lists were extracted using TS2 Mascot software (Matrix Science) at a signal to noise level of 5. Peak lists from both instruments used for protein identification were filtered to include only the eighty most intense peaks prior to database searching. Database searches were performed using Mascot 2.2.1 (Matrix Science). In case of human samples, Mascot was set up to search the SwissProt database (version 56.2) with taxonomic restriction to *Homo sapiens* (20,407 sequences), using iTRAQ 8-plex as quantification mode and methionine oxidation as additional variable modification. In case of BSA, the IPI bovine database (version 3.22 with 32,915 sequences) was searched. The mass tolerance of precursor ions was set to 5 ppm (for LTQ Orbitrap XL data) or 300 ppm (for 4800 Proteomics Analyzer data) and that of fragment ions to 0.6 Da. Up to 2 missed cleavages were allowed and Trypsin was specified as enzyme. Peptide identifications were established at a significance of at least 0.01, resulting in false discovery rates (FDRs) of 2-3% for all three datasets, as reported by the Mascot search engine which uses a target decoy strategy. For subsequent data analysis,

only peptides that uniquely map to one of the identified proteins in the experiment were included. Protein identification was based on presence of two or more significant (ion score greater than or equal to homology threshold) unique peptides that map to a protein. iTRAQ reporter ion intensities were extracted from peak lists with a 0.2 Da (for 4800 Proteomics Analyzer data) or 0.006 Da (for LTQ Orbitrap XL data) mass window using custom software. HCD quantification data were associated to CID and ETD peptide annotations based on scan numbers provided by the instrument. Quantification data were corrected for iTRAQ isotope impurities, and label-specific bias was removed by logarithmic transformation of the data and subtraction of the mean reporter ion intensity of the respective label in the experiment. Finally, reporter intensities were standardized by expressing each relative to the mean reporter intensity within the spectrum. Differences in protein abundance were determined by comparing mean logarithmic reporter ion intensities of spectra associated with a particular protein, and the significance was assessed using standard t-statistics. For a spectrum to be used for quantification it was required that every reporter ion could be measured, and that the highest reporter ion intensity in the spectrum exceeded 2000 (for 4800 Proteomics Analyzer data) or 5000 (for LTQ Orbitrap XL data). There was no requirement on the significance level of peptide annotation for spectra to contribute to protein quantification.

5 Acknowledgements

The authors wish to thank Shabaz Mohammed for critical discussion and Patricia Klemmer for the preparation of iTRAQ labeled samples. ABS and KWL are supported by grants of the Center for Medical Systems Biology. NM and AJRH are supported by the NGI Horizon Program grant number 050-71-050, and NM, AJRH and AFMA additionally by the Netherlands Proteomics Centre.

6 References

[1] Aebersold, R., *et al.*, *Nature* **2003**, *422*, 198

[2] Heck, A. J., *et al.*, *Expert Rev Proteomics* **2004**, *1*, 317

[3] Ong, S. E., *et al.*, *Mol Cell Proteom* **2002**, *1*, 376

[4] Oda, Y., *et al.*, *P Natl Acad Sci USA* **1999**, *96*, 6591

[5] Krijgsveld, J., *et al.*, *Nat Biotechnol* **2003**, *21*, 927

[6] Wu, C. C., *et al.*, *Anal Chem* **2004**, *76*, 4951
[7] Kruger, M., *et al.*, *Cell* **2008**, *134*, 353
[8] Gygi, S. P., *et al.*, *Nat Biotechnol* **1999**, *17*, 994
[9] Ross, P. L., *et al.*, *Mol Cell Proteom* **2004**, *3*, 1154
[10] Boersema, P. J., *et al.*, *Proteomics* **2008**, *8*, 4624
[11] Boersema, P. J., *et al.*, *Nat Protoc* **2009**, *4*, 484
[12] Meany, D. L., *et al.*, *Proteomics* **2007**, *7*, 1150
[13] Hardman, M., *et al.*, *Anal Chem* **2003**, *75*, 1699
[14] Schwartz, J. C., *et al.* In *53rd ASMS Conference on Mass Spectrometry,* San Antonio, Texas, 2005
[15] Olsen, J. V., *et al.*, *Nat Methods* **2007**, *4*, 709
[16] Griffin, T. J., *et al.*, *J Proteom Res* **2007**, *6*, 4200
[17] Bantscheff, M., *et al.*, *Mol Cell Proteom* **2008**, *7*, 1702
[18] Zhang, Y., *et al.*, *J Am Soc Mass Spectrom* **2009**
[19] Kocher, T., *et al.*, *J Proteome Res* **2009**
[20] Dayon, L., *et al.*, *Journal of Proteomics* **2010**, *73*, 769
[21] Savitski, M. M., *et al.*, *J Am Soc Mass Spectrom*, *21*, 1668
[22] Coon, J. J., *et al.*, *P Natl Acad Sci USA* **2005**, *102*, 9463
[23] Pitteri, S. J., *et al.*, *Anal Chem* **2005**, *77*, 1831
[24] Toorn van de, H. W. P., *et al.*, *J Proteomics Bioinform* **2008**, *1*, 379
[25] Molina, H., *et al.*, *Anal Chem* **2008**, *80*, 4825
[26] Swaney, D. L., *et al.*, *Anal Chem* **2007**, *79*, 477
[27] Han, H., *et al.*, *J Proteom Res* **2008**, *7*, 3643
[28] Phanstiel, D., *et al.*, *J Am Soc Mass Spectrom* **2008**, *19*, 1255
[29] Phanstiel, D., *et al.*, *Anal Chem* **2009**, *81*, 1693
[30] Yang, F., *et al.*, *Anal Chem* **2009**
[31] Swaney, D. L., *et al.*, *Nat Methods* **2008**, *5*, 959
[32] Kuzyk, M. A., *et al.*, *Proteomics* **2009**, *9*, 3328
[33] Ow, S. Y., *et al.*, *J Proteom Res* **2009**
[34] Li, K. W., *et al.*, *J Proteom Res* **2007**, *6*, 3127
[35] Taouatas, N., *et al.*, *Mol Cell Proteom* **2009**, *8*, 190
[36] Smalla, K. H., *et al.*, *Mol Psychiatry* **2008**, *13*, 878

Chapter 6

Comparative assessment of site assignments in CID and ETD spectra of phosphopeptides discloses limited relocation of phosphate groups

Nikolai Mischerikow[1,2], A. F. Maarten Altelaar[1,2], J. Daniel Navarro[1,2], Shabaz Mohammed[1,2], and Albert J. R. Heck[1,2,3]

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

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

[3] Centre for Biomedical Genetics, Padualaan 8, 3584 CH Utrecht, The Netherlands

Supplementary material referred to in this chapter can be found accompanying the publication of this work in *Molecular and Cellular Proteomics*, 2010, volume 9, issue 10, pages 2140-2148.

1 Summary

In mass spectrometry- (MS) based phosphoproteomics a current bottleneck is the unambiguous assignment of the phosphorylation site of a phosphopeptide. Additionally it has been reported that in the low-energy energy regime of ion trap collision induced dissociation (CID) the phosphate group may migrate to a nearby phosphate group acceptor, thus causing ambiguity in site assignment. Here, we describe the analysis of a statistically significant number of phosphopeptides generated from human cell lysate by proteolytic digestion using Trypsin or Lys-N followed by phosphopeptide enrichment by strong cation exchange (SCX). The nearly pure phosphopeptide fractions were analyzed by liquid chromatography-coupled MS (LC-MS) using an LTQ Orbitrap equipped with electron transfer dissociation (ETD) functionality. The instrument was set up to sequentially acquire both CID and ETD spectra for every precursor. We exploited the resistant nature of ETD towards phosphate group rearrangements to evaluate if phosphate group relocation occurs during CID. We evaluated a number of peptide and spectral annotation properties and found that for approximately 75% of the sequenced phosphopeptides the assigned phosphorylation site was unmistakably identical when inferred from the ETD and CID spectrum. For the remaining 25% of the sequenced phosphopeptides we also did not observe evident signs of relocation. These peptides exhibited signs of ambiguity in site localization predominantly induced by factors such as poor fragmentation, sequences causing inefficient fragmentation, and generally poor spectrum quality. Our data lets us derive the conclusion that for Trypsin- and Lys-N-generated peptides there is little relocation of phosphate groups occurring during CID.

2 Introduction

Signaling events, many regulated by dynamic protein phosphorylation, form the basis of inter- and intracellular communication. For instance, cellular activation through membrane receptors often induces changes in the phosphorylation state of the membrane receptor but also of hundreds of other proteins connected to that receptor in downstream pathways. Obtaining phosphopeptide profiles consisting of significant numbers of sites mapping to thousands of proteins has now become accessible to the specialized proteomics community [1]. A number of key technologies and strategies have led to this breakthrough. Peptide complexity reduction and phosphopeptide enrichment can be considered to be the most improved aspects of these phosphoproteome profiling

methodologies. A classic strategy for a phosphoproteome screen is to perform a fractionation of the sample followed by enrichment using a reagent with affinity for the phosphate group [2]. Fractionations based on (SCX) [3-5] or SDS polyacrylamide gel electrophoresis (SDS PAGE) [6,7] are the most common, although hydrophilic interaction chromatography (HILIC) [8,9] and isoelectric focusing (IEF) [10,11] provide attractive alternatives. In the case of SCX and HILIC, one can obtain nearly pure phosphopeptide enrichments when appropriate conditions and materials are chosen [5,12]. The classic choice for a second enrichment step has been immobilized affinity chromatography (IMAC) [13], which can have its selectivity improved even further through the use of additives [14] or methyl esterification of the peptide population [15]. In recent years, one of the most popular technologies used for enrichment of phosphopeptides has been titanium dioxide [16] often with additives [17,18] and applied either in an offline form or online as part of the final nanoflow LC-MS [19-21]. Alternative metal oxides such as those of zirconium [22] have proven to be quite successful too. In the specific case of phosphotyrosine profiling, antibodies against phosphotyrosine used at the peptide level have been demonstrated to be very effective [23-25].

The increase in throughput of phosphoproteomics has caused the community to focus on high-throughput analysis of the mass spectrometric data in order to expedite interpretation. One aspect that has received much attention is the dominant neutral loss peak in the fragmentation spectra of phosphopeptides obtained by traditional CID experiments [26,27]. The neutral loss peak can often suppress sequence diagnostic ion peaks causing identification of the peptide to become difficult or even impossible. This problem is somewhat exacerbated for ion trap-based CID primarily due to the energetic regime employed and the need to use a discrete limited ion population [28]. Since the use of ion traps currently represents the most common way of performing phosphoproteome screens, there have been various attempts to alleviate this specific problem. Modified fragmentation regimes have been introduced such as neutral loss-triggered MS^3 [2,4,29] or multistage activation [30]. Both of these methods fragment the neutral loss peak of the precursor ion further in order to generate more backbone cleavages which are the source for peptide sequencing. Alternatively, there has also been a re-exploration of higher-energy CID in the form of HCD [31], which exhibits lower levels of neutral loss since the activation step occurs at higher energies and on shorter timeframes [28]. ETD and electron capture dissociation (ECD) have also shown great promise since the phosphate group remains attached during and after activation [32-35].

Many detected phosphopeptides contain multiple serine/threonine/tyrosine residues representing the possibility that there is more than one location for the phosphorylation site. The abundant neutral loss observed in low energy CID can hamper the correct assignment of the phosphorylation site in such peptides. Therefore, a concerted effort has been made to understand in detail the rules of phosphopeptide fragmentation. Currently, there are a number of tools that allow automated site assignment [36,37] including some that can exploit MS^3 [38] and ETD data [39]. Palumbo *et al.* recently investigated phosphopeptide fragmentation from a more mechanistic point of view [40]. Interestingly, their results demonstrated that the phosphate group neutral loss pathway observed in low-energy CID operates through a nucleophilic substitution and not through a elimination reaction. This charge-directed mechanism requires proton abstraction as an initiation step and participation of neighboring groups. Further mechanistic work using synthetic phosphopeptides to explore the phosphate neutral loss pathway allowed the identification of rearrangement reactions involving the phosphate group during low-energy CID fragmentation. The rearrangement takes the form of a relocation of the phosphate group to an alternative hydroxyl group within the peptide. Thus when performing a comprehensive annotation of the spectra one would be able to observe multiple possibilities for the location of the phosphate group, referred to as phosphate group scrambling [41]. The authors note that the issue is more apparent when there are no mobile protons [42] and that the scrambling will be more dominant when low energy millisecond timeframes are used for CID such as in an ion trap. These observations evidently have raised concerns regarding the possible accuracy of phosphorylation site assignments.

Here, we investigate the extent of rearrangement reactions involving the phosphate group during low-energy CID fragmentation by generating a large scale phosphoproteomic dataset consisting of phosphopeptides fragmented sequentially by low-energy CID and ETD using an ETD-enabled LTQ Orbitrap. We utilized the ETD spectrum to allocate the 'correct' phosphate group position and compare it with the position allocated by the CID spectrum.

3 Results and discussion

To study the potential effect of phosphate group rearrangement on the identification of a phosphorylation site, we performed in a single run sequentially both CID and ETD [43]

on all peptides. The experiment was carried out using the following proteomics workflow. The phosphopeptide mixture was generated by digestion of a lysate of human cells with Lys-N [44] or Trypsin followed by SCX which was tuned for the enrichment of singly phosphorylated peptides as described previously [5]. Selected SCX fractions were further separated by online reversed phase (RP) LC-MS using an LTQ Orbitrap ETD instrument configured to fragment every peptide precursor with both CID and ETD. The time required for the full CID or ETD scan, typically a few hundred milliseconds, was short compared to the elution period of a peptide which is approximately 1 minute. Thus, the peptide precursor isolated for fragmentation by ETD was considered identical to the precursor isolated for fragmentation by CID. Two different peaklists, comprising CID and ETD fragmentation data, were generated for each analysis (Trypsin and Lys-N) and searched separately with Mascot. After database searching, the identifications obtained from CID and ETD peaklists were merged, resulting in a Lys-N and a Trypsin dataset.

The identified spectra and peptides are summarized in Figure 1. In the Trypsin dataset, 2,511 phosphorylated peptides (of these 911 unique) were identified by CID and 1,831 (740 unique) were identified by ETD. From these two groups, 1,464 phosphorylated peptides (593 unique) were identified from two consecutive spectra, *i.e.* CID and ETD spectra. For these peptides, for which sequence information derived from complementary fragmentation modes is available, 1,316 peptides (561 unique) or 90% had been assigned identical phosphate group positions by Mascot without any further site validation or manual inspection. The remaining 148 spectra (of these 93 unique) or 10% had been assigned different phosphate group positions. To obtain a higher level of certainty about the localization of the phosphate group when derived from the CID spectrum, assuming the ETD spectrum provided the correct site assignment, we also evaluated the phosphate group position by PTM scoring as implemented in MSQuant [45]. When analyzing the 1,464 phosphopeptides with this algorithm, 1,089 peptides or 74% had been assigned identical phosphate group positions and 138 peptides or 9% had been assigned different positions. The remaining 237 peptides or 17% could either not been scored by PTM scoring, or the position of the phosphate group in the CID spectrum could not be unambiguously determined, which was reflected by multiple highest scoring phosphate group positions with identical PTM scores. This group of peptides with ambiguous phosphate group positions might potentially contain spectra with highly abundant fragment ions indicative of phosphate group rearrangement; this is discussed in more detail below.

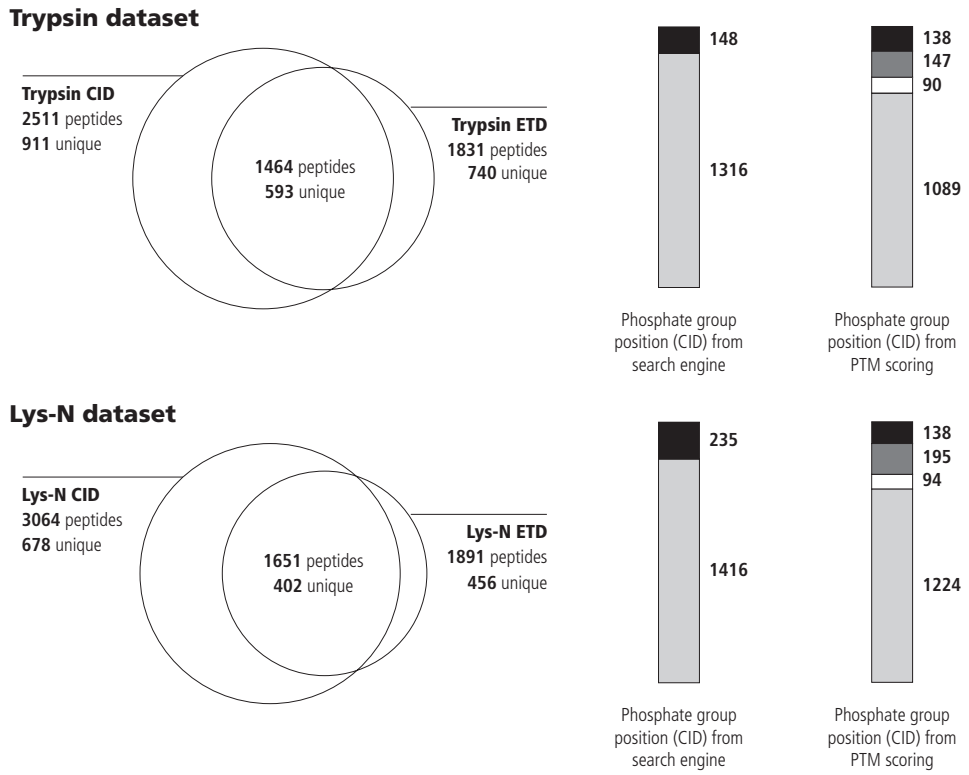


Figure 1. Number of (unique) phosphorylated peptides identified in the Trypsin (upper panels) and Lys-N (lower panels) datasets. The left diagrams show the identifications obtained with CID and ETD with the overlap representing identifications when the ETD scan directly followed the CID scan, i.e. from the same precursor. The bar diagrams break theses overlaps down into CID-ETD pairs with identical (light grey) and different (black) phosphate group positions. They also illustrate how these numbers changed when the phosphate group positions obtained by CID was validated by PTM scoring. Peptides for which PTM scoring did not unambiguously report one site for the CID spectrum (dark grey) and for which the CID spectrum was not scored (white) added to the other two classes.

The Lys-N dataset gave a very similar picture to that of the Trypsin dataset (Figure 1). Approximately 3,000 phosphorylated peptides (678 unique) were identified by CID and 1,891 phosphopeptides (456 unique) by ETD. The overlap is formed by 1,651 peptides (402 unique), which were identified by both CID and subsequent ETD. For 1,416 of these peptides (376 unique) or 86% an identical position of the phosphate group had been derived from both the CID and ETD spectra, while for 235 peptides (109 unique) or 14% the site was different. As for the Trypsin dataset, the CID spectra were processed with PTM scoring to validate the position of the phosphate group which resulted

in the following figures. From the total of 1,651 peptides, 1,224 or 74% had been assigned identical phosphate group positions and 138 or 8% had different phosphate group positions. The remaining 289 peptides or 18% could either not be scored or had multiple highest scoring phosphate group positions in the CID spectrum, classifying these peptides as ambiguous. Similar to the Trypsin dataset the group of phosphopeptides containing ambiguous phosphate group positions is potentially most interesting.

In both datasets the position of the phosphate group was identical for most peptides which had been identified from their CID and subsequent ETD spectra. Specifically, this was true for 86% (Lys-N) or 90% (Trypsin) of all peptides with complementary sequencing information if the phosphate group position was taken directly from the search engine and for 74% if the position derived from the CID spectrum was validated by PTM scoring. As rearrangement of phosphate groups is suspected to take place only during CID in the ion trap, but not during ETD, only CID spectra can potentially contain fragment ions indicative of gas-phase rearrangement. If these features would have been more abundant, both in terms of number of CID spectra affected and in terms of intensities of the features in the CID spectra, we would have expected to see many more peptides with disagreeing phosphate group positions. In case the CID phosphate group position was evaluated by PTM scoring, we would have expected a higher level of peptides with ambiguous phosphate group position in CID. From all our observations we conclude that if gas-phase rearrangement had taken place it did not affect the identification of the phosphate group position in the majority of cases, i.e. identical phosphate group positions were observed in CID and ETD spectra.

Irrespective of this conclusion, gas-phase rearrangement might still have taken place during CID, with or without having affected the determination of the correct phosphate group position by the search engine or by PTM scoring. One possible way is that the group of peptides for which identical phosphate group positions were obtained might still contain CID spectra with features indicative of phosphate group rearrangement, but these features might have been so low abundant that they did not hinder the search engine or PTM scoring in reporting the correct phosphate group position. Alternatively, within the groups of peptides for which different or ambiguous phosphate group positions were reported, the indicative fragment ions corresponding to phosphate relocation might have been dominant in the CID spectra such that the correct phosphate group position was not reported by the search engine or PTM scoring. In case of phosphate group position ambiguity, the PTM scoring algorithm might not have been able to distinguish between the correct and the rearranged phosphate group

position, leading to multiple highest-scoring positions.

If rearrangement had taken place, in all three peptide groups – identical position, different position, and ambiguous position – it would only have affected the CID spectra but not the ETD spectra because ETD has limited intramolecular energy dissipation [46]. Therefore the probability of a less likely peptide match with identical amino acid sequence but different phosphate group position should have increased for a peptide match when derived from the CID spectrum but not when derived from the ETD spectrum. This difference in probability is directly reflected in the Mascot delta ion score of a peptide, which is the difference of the ion score of the most likely, *i.e.* identified, peptide match to the ion score of the next ranking, sequence-identical peptide match. We used this delta ion score as a measure to compare the significance of a phosphate group positioning in CID and ETD [28]. For calculating the delta ion score we also used peptide sequence matches with a rank lower than the first two matches, if the matched amino acid sequence was identical to the highest ranking peptide sequence match. In this case, as the two peptides sequences for which the delta ion score is calculated are identical in their primary amino acid sequence and only differ in the position of the phosphate group, the delta ion score is directly related to the presence and intensity of the fragment ions that can distinguish the two possible phosphorylation sites.

In the Trypsin dataset, for 1,167 (or 90%) of the 1,374 phosphorylated peptides for which complementary sequencing information was available and which could be grouped by PTM scoring, a delta ion score could be calculated for both CID and ETD. The remaining 207 (or 10%) of peptides did not have a next lower ranking, sequence-identical peptide match in either CID or ETD or both. In the Lys-N dataset, for 1,257 (or 78%) of 1,618 peptides a delta ion score could be calculated for both CID and ETD, while the remaining 361 (or 22%) of peptides had no delta ion score in either CID or ETD or both. In both datasets, many peptides without delta ion score contain only a single serine, threonine or tyrosine residue and hence cannot show rearrangement *per se*. In the Trypsin dataset, the comparison of the delta ion score for CID versus the delta ion score for ETD, broken down to the three peptide groups that were generated by PTM scoring, showed that all peptides are relatively evenly distributed over the delta ion score space (Figure 2A). Peptides for which identical phosphate group positions had been determined showed the same distribution (Figure 2B). Peptides for which non-identical phosphate group positions had been determined showed a mild preference towards a small delta ion score in ETD coupled with a more widely distributed delta ion score in CID (Figure 2D). Logically, peptides with ambiguous phosphate group positions

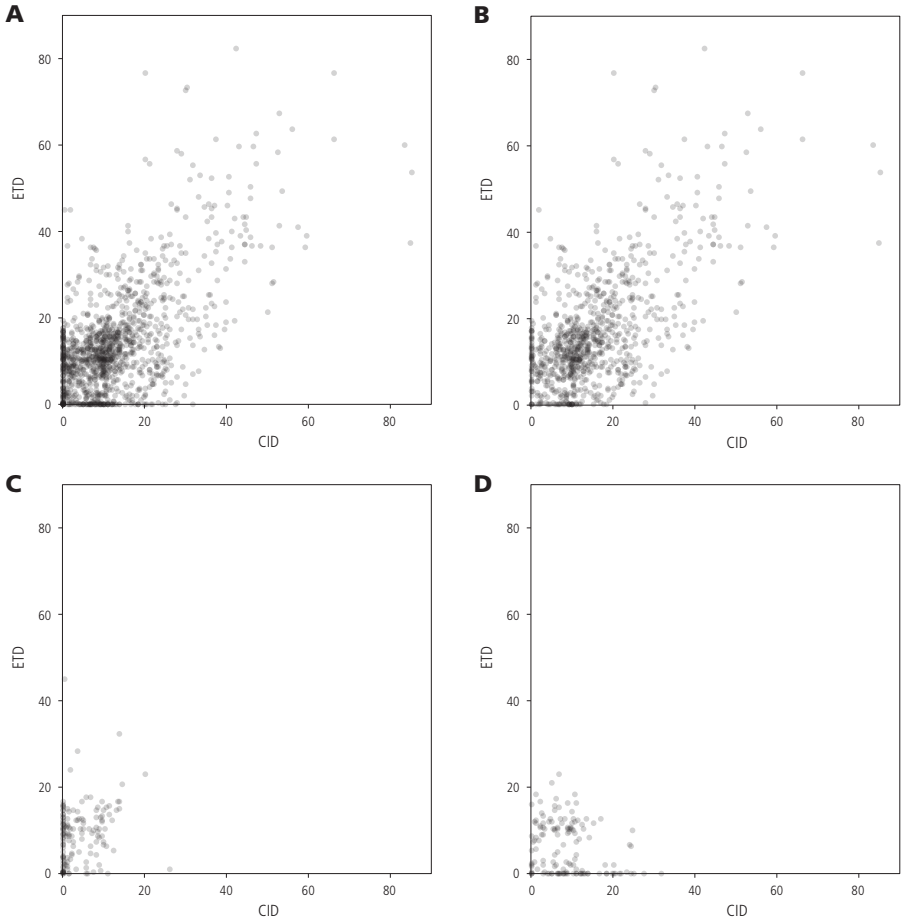


Figure 2. Delta ion scores as a measure for ambiguity of phosphate group site assignment in the Trypsin dataset. All plots display the delta ion score for ETD versus the delta ion score for CID for peptides with identical (B) or ambiguous (C) site assignments as well as for peptides with different phosphosites (D). An overlay of all three plots is shown for comparison (A).

had a mild bias towards a larger delta ion score in ETD (Figure 2C). The Lys-N dataset displayed very similar behavior in all groups of peptides, although slightly less clear cut than the Trypsin dataset (supplementary material).

For peptides with ambiguous and different phosphate group positions the uncertainty of phosphate group position, as reflected by the delta ion score, was not related to the spectral quality as reflected by the ion score itself (supplementary material). Both the CID and ETD ion scores of peptides in these two groups were evenly distributed over the

whole score range and most had a delta score below 15. This observation was also true when the CID and ETD ion scores were normalized for the peptide length (supplementary material), a normalization method that takes into account that peptide Mascot ion scores increase with peptide length even though spectral quality does not improve [34].

When manually inspecting spectra of peptides with ambiguous and different phosphate group positions, we found that many peptides had either a low CID or a low ETD ion score, reflecting either low quality CID or low quality ETD spectra. We also observed that many peptides in these groups had other properties which had hindered the clear determination of the phosphate group in either the CID or the ETD spectra. One of the most influential properties was the number of serine/threonine/tyrosine residues and their relative locations. As the number of residues increases, the possibility of unambiguously detecting the phosphate group position correctly in CID decreases. In both datasets, we observed that peptides with ambiguous or different phosphate group positions are biased towards containing higher numbers of phosphorylatable residues. Another noteworthy property which hinders the assignment of the phosphate group position particularly in ETD spectra, due to the fragmentation regime, is the localization of phosphorylatable residues close to the termini of the peptide, as fragment ions with low mass-to-charge ratio (m/z) are unlikely to be detected in the ion trap [47-49]. To conclude, when manually inspecting CID and ETD spectra of these peptides, no fragment ions which could have clearly indicated phosphate group rearrangement could be found. In fact, for every peptide investigated in these two groups, one or more of the factors prevented the unambiguous or even correct identification of the phosphate group in either the CID or ETD spectrum or both.

These observations were also true for doubly phosphorylated peptides, of which 483 (141 unique) were identified by CID and 210 (79 unique) were identified by ETD in the Trypsin dataset. In the two groups, 29 double phosphorylated peptides (17 unique) were identified by consecutive CID and ETD spectra. Of these, 26 had identical phosphorylation patterns without any further site validation using PTM scoring. Manual validation showed that most spectra were rather poor due to the generally worse fragmentation of multiply phosphorylated peptides.

In the enrichment technology used here, we have specifically generated phosphopeptide pools with two basic moieties and one phosphate group. Considering that these peptides will have a minimum of two charges in the gas phase, the majority of them are in the mobile proton situation [42]. Our finding is therefore in agreement with Pa-

lumbo *et al.* [41] who observed that fragment ions indicative of phosphate group rearrangement are more prominent for peptides in a non-mobile proton situation than for peptides in a partially mobile or, even less, mobile proton situation.

In summary, we have assessed the effect of gas phase phosphate group rearrangement under LC-MS conditions typically used in a proteomics experiment. By comparing phosphate group positions derived from CID and ETD spectra, we observed in both datasets that for 74% of all phosphopeptides for which complementary sequencing information exists, assignment of phosphate group positions were identical. Moreover, close inspection of the remaining 26% also revealed no clear evidence for significant phosphate group relocation. The spectrum quality of these 26% was generally lower, due to factors such as poor fragmentation and sequences causing inefficient fragmentation. This indicates that phosphate group rearrangement did not affect the identification of the correct phosphorylation sites in large datasets of Trypsin- or Lys-N generated peptides.

4 Experimental procedures

Digestion of cell lysate. 2 mg of human cell lysate in 8 M urea were reduced with DTT (45 mM, 30 min at 50°C) and alkylated with iodoacetamide (100 mM, 30 min incubation). For digestion with Trypsin, 1 mg of lysate was digested with Lys-C (1.25 µg, 4 h at 37°C) in 7 M urea followed by digestion with Trypsin (15 µg, incubation over night at 37°C) after dilution to 2 M urea. For digestion with Lys-N, 1 mg lysate was digested with Lys-N (5 µg, 3 h at 37°C) in 5 M urea followed by digestion with Lys-N (5 µg, incubation over night at 37°C) after dilution to 2 M urea.

Peptide pre-fractionation by (SCX) [5,50]. Peptides from each digest were loaded onto two C18 cartridges using an Agilent 1100 HPLC system. The flow rate applied was 100 µL/min using water pH 2.7 as solvent. Next, peptides were eluted from the trapping cartridges with 80% acetonitrile pH 2.7 onto a PolySULFOETHYL A 200 x 2.1 mm column (PolyLC) for 10 min at the same flow rate. Separation of different peptide populations was performed using a nonlinear 65 min gradient, 0 to 10 min 100% solvent A (5 mM KH_2PO_4, 30% acetonitrile, pH 2.7), 10 to 15 min up to 26% solvent B (5 mM KH_2PO_4, 30% acetonitrile, 350 mM KCl, pH 2.7), 15 to 40 min to 35% solvent B and from 40 to 45 min to 60% solvent B. At 49 min the concentration of solvent B was 100%. The column was subsequently washed for 6 min under high salt conditions and finally equilibrated with 100% solvent A for 9 min. The flow rate applied during the SCX

gradient was 200 µL/min. Fractions were collected in 1 min intervals for 40 min. After evaporation of the solvents, fractionated peptides were dissolved in 10% formic acid.

LC-MS. SCX fractions were analyzed on a RP LC-coupled LTQ Orbitrap ETD (Thermo Fisher Scientific). An Agilent 1200 series HPLC system was equipped with a 20 mm Aqua C18 (Phenomenex) trapping column (packed in-house, 100 µm inner diameter, 5 µm particle size) and a 400 mm ReproSil-Pur 120 C18-AQ (Dr. Maisch GmbH) analytical column (50 µm inner diameter, 3 µm particle size). Trapping was performed at 5 µL/min solvent C (0.1 M acetic acid in water) for 10 min, and elution was achieved with a gradient from 10 to 30% (v/v) solvent D (0.1 M acetic acid in 1:4 acetonitrile : water) in solvent C in 110 min, followed by a gradient of 30 to 50% (v/v) solvent D in solvent C in 30 min, followed by a gradient of 50 to 100% (v/v) solvent D in solvent C in 5 min and finally 100% solvent D for 2 min. The flow rate was passively split from 0.45 mL/min to 100 nL/min. Electrospray was achieved using a distally coated fused silica emitter (360 µm outer diameter, 20 µm inner diameter, 10 µm tip inner diameter, New Objective) biased to 1.7 kV. The LTQ Orbitrap ETD was operated in the data dependent mode to automatically switch between MS and MS/MS. Survey full scan MS spectra were acquired from 350 to 1500 Th in the orbitrap with a resolution of 60,000 at 400 Th after accumulation to a target value of 500,000 in the linear ion trap. The two most intense ions at a threshold of above 500 were fragmented in the linear ion trap using CID at a target value of 30,000 and ETD with supplemental activation at a target value of 50,000. The ETD reagent target value was set to 100,000 and the reaction time to 50 ms.

Data processing. From every raw data file recorded by the mass spectrometer, representing a single SCX fraction, two different peaklists containing either CID or ETD fragmentation data were generated using Proteome Discoverer (version 1.0, Thermo Fisher Scientific) with a signal to noise threshold of 3 and the following settings for the ETD non-fragment filter: precursor peak removal with 4 Da, charge-reduced precursor removal with 8 Da, and removal of know neutral losses from charge-reduced precursor with 8 Da within a window of 120 Da. Single-fraction peaklists of the major phosphopeptide-containing SCX fractions for Trypsin- and Lys-N-derived peptides were then merged into four larger peaklists, namely Trypsin CID, Trypsin ETD, Lys-N CID and Lys-N ETD. Mascot (version 2.2.04, Matrix Science) was used to search these peaklists against an in-house built database (150,852 entries) assembled from the IPI human database (version 3.54, http://www.ebi.ac.uk/ipi) plus all sequence-reversed entries of the latter. Peptide matches were established at a significance level of <0.05 and above or equal to

an ion score of 20. Only first ranking and bold peptides were considered for identification. The false discovery was estimated as the fraction of all peptides that match to a reversed database entry and was between 0.6% and 1.0% for all four datasets. PTM scoring was performed only on the Trypsin CID and Lys-N CID datasets using MSQuant (version 2.0a81, http://msquant.alwaysdata.net) [45]. Further data processing was performed with Excel. Scaffold data files of fragmentation spectra are publicly available in the repository Tranche (https://proteomecommons.org/) using the following hash code: yvSNWDchATkyAxILEPYxFjILNyDbSDL7BCKET/l13e9OKIpAmMgMHZP1x2kLkJliuHryd5Hf+G2HdBO5fnE2u1xw4ssAAAAAAAAFWg==

5 Acknowledgements

The authors acknowledge financial support from the NGI Horizon Program (grant number 050-71-050). This work was supported by the Netherlands Proteomics Centre, which is part of the Netherlands Genomics Initiative.

6 References

[1] Reinders, J., *et al.*, *Proteomics* **2005**, *5*, 4052
[2] Gruhler, A., *et al.*, *Mol Cell Proteom* **2005**, *4*, 310
[3] Motoyama, A., *et al.*, *Anal Chem* **2007**, *79*, 3623
[4] Beausoleil, S. A., *et al.*, *Proc Natl Acad Sci USA* **2004**, *101*, 12130
[5] Gauci, S., *et al.*, *Anal Chem* **2009**, *81*, 4493
[6] Everley, P. A., *et al.*, *J Proteom Res* **2006**, *5*, 1224
[7] Elortza, F., *et al.*, *Mol Cell Proteom* **2003**, *2*, 1261
[8] Gilar, M., *et al.*, *Anal Chem* **2005**, *77*, 6426
[9] Boersema, P. J., *et al.*, *Mol Cell Proteom* **2009**, *8*, 650
[10] Cargile, B. J., *et al.*, *Electrophoresis* **2004**, *25*, 936
[11] Krijgsveld, J., *et al.*, *J Proteom Res* **2006**, *5*, 1721
[12] McNulty, D. E., *et al.*, *Mol Cell Proteom* **2008**, *7*, 971
[13] Stensballe, A., *et al.*, *Proteomics* **2001**, *1*, 207
[14] Stensballe, A., *et al.*, *Rap Commun Mass Spectrom* **2004**, *18*, 1721
[15] Ficarro, S. B., *et al.*, *Nat Biotechnol* **2002**, *20*, 301
[16] Pinkse, M. W. H., *et al.*, *Anal Chem* **2004**, *76*, 3935

[17] Larsen, M. R., *et al.*, *Mol Cell Proteom* **2005**, *4*, 873
[18] Sugiyama, N., *et al.*, *Mol Cell Proteom* **2007**, *6*, 1103
[19] Pinkse, M. W. H., *et al.*, *J Proteom Res* **2008**, *7*, 687
[20] Cantin, G. T., *et al.*, *Anal Chem* **2007**, *79*, 4666
[21] Mohammed, S., *et al.*, *J Proteom Res* **2008**, *7*, 1565
[22] Kweon, H. K., *et al.*, *J Proteom Res* **2008**, *7*, 749
[23] Zhang, Y., *et al.*, *Mol Cell Proteom* **2005**, *4*, 1240
[24] Rikova, K., *et al.*, *Cell* **2007**, *131*, 1190
[25] Boersema, P. J., *et al.*, *Mol Cell Proteom* **2009**
[26] Huddleston, M. J., *et al.*, *J Am Soc Mass Spectrom* **1993**, *4*, 710
[27] DeGnore, J. P., *et al.*, *J Am Soc Mass Spectrom* **1998**, *9*, 1175
[28] Boersema, P. J., *et al.*, *J Mass Spectrom* **2009**, *44*, 861
[29] Benschop, J. J., *et al.*, *Mol Cell Proteom* **2007**, *6*, 1198
[30] Schroeder, M. J., *et al.*, *Anal Chem* **2004**, *76*, 3590
[31] Olsen, J. V., *et al.*, *Mol Cell Proteom* **2009**
[32] Molina, H., *et al.*, *Proc Natl Acad Sci USA* **2007**, *104*, 2199
[33] Chi, A., *et al.*, *Proc Natl Acad Sci USA* **2007**, *104*, 2193
[34] Mohammed, S., *et al.*, *Anal Chem* **2008**, *80*, 3584
[35] Sweet, S. M. M., *et al.*, *Mol Cell Proteom* **2009**, *8*, 904
[36] Olsen, J. V., *et al.*, *Cell* **2006**, *127*, 635
[37] Beausoleil, S. A., *et al.*, *Nat Biotechnol* **2006**, *24*, 1285
[38] Ruttenberg, B. E., *et al.*, *J Proteom Res* **2008**, *7*, 3054
[39] Bailey, C. M., *et al.*, *J Proteom Res* **2009**, *8*, 1965
[40] Palumbo, A. M., *et al.*, *J Proteom Res* **2008**, *7*, 771
[41] Palumbo, A. M., *et al.*, *Anal Chem* **2008**, *80*, 9735
[42] Wysocki, V. H., *et al.*, *J Mass Spectrom* **2000**, *35*, 1399
[43] Swaney, D. L., *et al.*, *Anal Chem* **2007**, *79*, 477
[44] Taouatas, N., *et al.*, *Nat Meth* **2008**, *5*, 405
[45] Mortensen, P., *et al.*, *J Proteom Res* **2009**
[46] Syka, J. E. P., *et al.*, *Proc Natl Acad Sci USA* **2004**, *101*, 9528
[47] Good, D. M., *et al.*, *Mol Cell Proteom* **2007**, *6*, 1942
[48] Molina, H., *et al.*, *Anal Chem* **2008**, *80*, 4825
[49] Hennrich, M. L., *et al.*, *Anal Chem* **2009**, *81*, 7814
[50] Taouatas, N., *et al.*, *Mol Cell Proteom* **2009**, *8*, 190

Summary

Peptide mass spectrometry (MS) is an invaluable analytical method in biological and medical research. It is the only technique that, when integrated with liquid chromatography (LC) and database search tools, allows a highly sensitive qualitative characterization and highly accurate quantitative comparison of proteomes. Although many proteomes are much more complex than their corresponding genomes, due to, for example, extreme differences in protein abundance and post-translational modifications, continuous technical advances in MS instrumentation and peptide pre-fractionation techniques lead to increasing fractions of proteomes that can be covered. Nevertheless, the targeted analysis of subsets of proteomes defined by post-translational modifications (PTMs), for example phosphorylation, acetylation, or glycosylation, using specialized enrichment techniques, is required to gain insight into cellular processes that would be inadequately covered by analysis of the full proteome alone. The technological progression in proteomics also benefits the analysis of protein complexes and other relatively small ensembles of proteins. With modern MS instrumentation, a targeted analysis is mostly not required to create a comprehensive picture of protein complexes, including PTMs and protein isoforms. Selected core technologies of proteomics are introduced in **Chapter 1**. It is mainly focused on MS instrumentation and database searching, but also covers aspects like peptide fragmentation and methods in quantitative proteomics. In this chapter we also give a brief introduction to the general transcription factors (GTFs) TFIID and SAGA and put them into their broader biological context.

Chapter 2 reviews strategies for the targeted analysis of protein acetylation, a PTM of primary amines of protein N-termini or lysine side chains. We describe strong cation exchange (SCX) and combined fractional diagonal chromatography (COFRADIC) as comparably powerful methods for the enrichment of N-terminal acetylated peptides. The enrichment of protein N-termini on a proteome-wide level has helped to shed light on the substrate specificity of NATs and also allows the identification of protein variants that are not annotated in genome databases. In this chapter we also discuss peptide immunoprecipitation as the only enrichment technique for lysine acetylated peptides. The targeted proteome-wide analysis of this PTM on non-histone proteins is relatively recent and has yielded unanticipated insights into the function of lysine acetylation in metabolic regulation.

Chapter 3 describes the analysis of the heteromultimeric GTFs SAGA and TFIID from *Saccharomyces cerevisiae*. In this study, different pre-fractionation techniques and pro-

teases were combined to achieve an in-depth characterization of the tandem affinity-purified complexes. We mapped the complete primary sequence of all known SAGA and TFIID subunits to near completion and identified a large number of phosphorylation and acetylation sites without any specialized enrichment. We confirmed published phosphorylation sites of TBP-associated factors (TAFs) that are subunits of both SAGA and TFIID with sufficiently high numbers of spectra to compare the phosphorylation levels in SAGA and TFIID. We found that TAF5 S411/414/415 is phosphorylated to greater extent when in SAGA than when in TFIID and that TAF5 K103 is acetylated in the opposite manner. All other sites were equally populated in both complexes. A second focus of the chapter is the discovery of lysine acetylation as a predominant PTM of the acetyltransferase SAGA, mainly affecting its subunits Spt7 and Sgf73. As a third result of this analysis we describe the discovery of a truncated form of Spt7, which is the C-terminal counterpart of a truncated form of Spt7 that is present in the SAGA-related protein complex SLIK, and show how we use this fragment to map the exact cleavage site within Spt7.

Chapter 4 wraps up two different approaches of interaction proteomics to the analysis of mainly TFIID from *Mus musculus*. In the first part of the chapter we describe the use of SCX to pre-fractionate digests of the TBP interactome obtained by tandem immunoprecipitation from transgenic embryonic fibroblast (MEF) lines to obtain mainly TFIID and the related B-TFIID complex formed by TBP and BTAF1. The interactomes of selected TBP mutants were compared with those of wild type TBP to show that two single amino acid substitutions, R188E and K243E, result in loss of the BTAF1 interaction while all interactions that constitute TFIID are maintained. With this we confirmed published *in vitro* interactions of the TBP mutants in MEFs *in vivo*. In the second part of this chapter we used a GFP affinity purification and stable isotope labeling to show that TFIID in embryonic stem cells (mESC) has qualitatively the same subunit composition as in MEFs. Additionally we identified the spermatogenesis-related TAF7L as potentially unique subunit isoform. Using this experimental approach we also explored the induction of dynamic exchange of TFIID subunits in mESC nuclear extracts under *in vitro* transcription reaction conditions using stable isotope ratios as readout. We identify TBP as the only potential dynamic subunit candidate and show that its exchange may be stimulated by the addition of TATA-containing promoter DNA.

Chapter 5 describes a method for tandem mass spectrometry (MS/MS) based quantitative proteomics using iTRAQ on an electron transfer dissociation (ETD) and higher energy collision dissociation (HCD) enabled LTQ Orbitrap XL. iTRAQ is widely used in

quantitative proteomics but incompatible with ETD and difficult to implement on the LTQ Orbitrap. The method fully decouples peptide identification, which is performed as regular collision induced dissociation (CID) or ETD scan, and peptide quantification using HCD. We demonstrate the optimization of the HCD collision energy for the maximization of the total number of quantifiable peptides and evaluate the method with an analysis of an 8-plex iTRAQ labeled clinical sample. We show that our method compares with the performance of a 4800 Proteomics Analyzer in terms of quantification, while in terms of identification it performs significantly better.

Chapter 6 presents a study on the influence of gas phase re-arrangement of phosphate groups of phosphorylated peptides during CID on the identification of the phosphorylation site in large scale phosphoproteomics experiments. We fragmented a statistically significant number of phosphorylated peptides with both CID and ETD, during which re-arrangement is mechanistically not favored, and compared the reported phosphorylation sites for both spectra. We found that after site validation with PTM scoring, identical sites could be observed for 75% of all phosphopeptides. For the remaining 25% of phosphopeptides we showed by manual inspection of the spectra that site ambiguity or disagreement was related to poor spectral quality in one or both spectra or large number of potential sites. No fragment ions clearly indicative for gas phase re-arrangement were observed. In this chapter we also show that the Mascot delta ion score, rather than the Mascot score, can serve as a measure for the certainty of the site assignment both in CID and ETD spectra.

Samenvatting

De analyse van peptiden met behulp van massaspectrometrie (MS) is van cruciaal belang voor biologisch en medisch onderzoek. Het is de enige techniek die, wanneer het gecombineerd wordt met vloeistofchromatografie (LC) en database analyse, het mogelijk maakt om zeer gevoelig en nauwkeurig verschillende proteomen zowel kwalitatief als kwantitatief met elkaar te vergelijken. Veel proteomen zijn complexer dan de corresponderende genomen, bijvoorbeeld door grote verschillen in de hoeveelheid van eiwitten en door post-translationele modificaties (PTMs). De constante technische verbeteringen aan MS instrumentatie en scheidingsmethoden voor peptiden maken het mogelijk een steeds groter deel van het proteoom te detecteren. Toch zijn voor de analyse van post-translationele modificaties zoals fosforylering, acetylering of glycosylering speciale verrijkingsmethoden noodzakelijk, om zo celullaire processen te analyseren die niet voldoende gedetecteerd worden als het hele proteoom ineens geanalyseerd wordt. De technologische vooruitgang in proteomics maakt het ook mogelijk om eiwitcomplexen of andere groepen eiwitten beter te bestuderen. Met moderne MS instrumenten is een gerichte analyse van die eiwitten meestal niet nodig om toch een goed overzicht te krijgen van een eiwitcomplex, inclusief modificaties en isovormen. Een selectie van in proteomics gebruikte technieken wordt uitgelegd in **Hoofdstuk 1**. Het hoofdstuk richt zich met name op MS instrumentatie en database analyse, maar beschrijft ook peptide fragmentatie en methoden gebruikt in kwantitatieve proteomics. In dit hoofdstuk geven we ook een korte introductie in basale transcriptiefactoren (GTFs), TFIID en SAGA en hun biologische context.

Hoofdstuk 2 behandelt strategieën voor de gerichte analyses van eiwit acetylering, een PTM van de primaire amines op de N-terminus of Lysine zijketens van een eiwit . We beschrijven *strong cation exchange* (SCX) and *combined fractional diagonal chromatography* (COFRADIC) als twee methoden om geactyleerde N-terminale peptiden te verrijken. De verrijking van eiwit N-termini op een proteoom-brede schaal helpt bij het begrijpen van de substraat specificiteit van NATs en maakt het mogelijk om eiwit varianten te identificeren die tot nu toe niet bekend waren. In dit hoofdstuk bediscussiëren we ook immunoprecipitatie als de enige verrijkingsmethode voor peptiden met geacetyleerde Lysine residuen. De gerichte proteoom-brede analyse van deze PTM op andere eiwitten dan histoneiwitten is een recente ontwikkeling en heeft tot onverwachte inzichten geleid in hoe acetylering van Lysine een rol speelt in metabole regulatie.

Hoofdstuk 3 beschrijft de analyse van de heteromultimere GTFs SAGA en TFIID van *Sac-*

charomyces cerevisiae. In deze studie worden verschillende fractioneringsmethoden en proteases gecombineerd in een uitgebreide analyse van de *tandem affinity* gezuiverde complexen. We brachten bijna de volledige sequentie van alle bekende SAGA en TFIID eiwitten in beeld en identificeerden een groot aantal fosforylerings- en acetyleringsplaatsen zonder gerichte verrijking. We bevestigden bekende fosforyleringsplaatsen op TBP geassocieerde factoren (TAFs), die een onderdeel zijn van zowel SAGA als TFIID. De identificatie van voldoende spectra maakte het mogelijk een kwantitatieve vergelijking te maken. We vonden dat TAF5 S411/414/415 meer gefosforyleerd was in SAGA dan in TFIID en dat voor acetylering op TAF5 K103 het omgekeerde geld. Alle andere plaatsen bleken in gelijke mate gemodificeerd te zijn in beide complexen. Een tweede aandachtspunt in het hoofdstuk is de ontdekking dat Lysine acetylering veel voorkomt in de acetyltransferase SAGA, met name op de eiwitten Spt7 en Sgf73. Tot slot ontdekten we een kort fragment van Spt7, waarvan ook een langer fragment aanwezig is in het aan SAGA gerelateerde complex SLIK, en gebruikten deze informatie om de exacte klievingsplaats in Spt7 te vinden.

Hoofdstuk 4 combineert twee verschillende methoden voor de analyse van eiwit interacties in TFIID uit *Mus musculus*. In het eerste gedeelte gebruiken we SCX voor de fractionering van digesten van het TBP interactoom, verkregen door tandem immunoprecipitatie van TFIID en het gerelateerde B- TFIID complex, bestaande uit TBP en BTAF1, uit een transgene embryonale fibroblast (MEF) cellijn. Het interactoom van geselecteerde TBP mutanten werd vergeleken met wildtype TBP. Hieruit bleek dat twee aminozuur substituties, R188E en K243E, de interactie van BTAF1 verhinderen terwijl alle TFIID interacties intact bleven. Daarmee bevestigden we gepubliceerde *in vitro* TBP interacties in MEFs *in vivo*. In het tweede gedeelte van dit hoofdstuk gebruikten we GFP affiniteitszuivering en labeling met stabiele isotopen om te laten zien dat TFIID in embryonale stamcellen (mESCs) in de zelfde samenstelling aanwezig is als in MEFs. Daarnaast identificeerden we het bij spermatogenese betrokken TAF7L als een mogelijke unieke isovorm in het complex. Met deze methode onderzochten we ook de dynamiek van de uitwisseling van TFIID eiwitten in kern extracten van mESCs tijdens *in vitro* transcriptie reacties. We vonden dat TBP het enige eiwit was dat mogelijk dynamisch gedrag vertoonde en laten zien dat dit gestimuleerd zou kunnen worden door TATA bevattende DNA promotor sequenties.

Hoofdstuk 5 beschrijft een methode voor kwantitatieve proteomics op basis van tandem massaspectrometrie (MS/MS) met behulp van iTRAQ en een LTQ Orbitrap XL massaspectrometer met *electron transfer dissocation* (ETD) en *higher energy collision*

dissociation (HCD) mogelijkheden. De methode koppelt peptide identificatie, gedaan met de standaard *collision induced dissociation* (CID) of ETD fragmentatie, los van de kwantificering met behulp van HCD. We laten zien dat optimalisatie van de HCD botsingsenergie het aantal gekwantificeerde peptiden kan vergroten en testen de methode met een 8-voudige iTRAQ labeling van een klinisch monster. We laten zien dat onze methode qua kwantificering vergelijkbaar is met een 4800 Proteomics Analyzer, maar dat het aantal identificaties aanzienlijk groter is.

In **Hoofdstuk 6** komt een studie aan bod naar de invloed van de herschikking van fosfaatgroepen op gefosforyleerde peptiden in de gas fase tijdens CID op de identificatie van de fosforyleringsplaatsen in grootschalige fosfoproteomics experimenten. We fragmenteren een statistisch significant aantal gefosforyleerde peptiden met zowel CID als ETD, waarbij geen herschikking plaatsvindt, en vergeleken de gerapporteerde fosforyleringsplaatsen voor beide spectra. We vonden dat, na validatie van de PTM scoring, identieke fosforyleringsplaatsen werden gezien in ongeveer 75% van alle fosfopeptiden. Voor de overgebleven 25% lieten we zien, na handmatige analyse van de spectra, dat de verschillen in de gevonden fosforyleringsplaatsen te wijten waren aan spectra van slechte kwaliteit of een groot aantal mogelijke fosforyleringsplaatsen in de peptiden. We vonden geen fragmentionen die bewijs gaven voor herschikking van de fosfaatgroep in de gas fase. In dit hoofdstuk laten we ook zien dat de Mascot delta ion score een betere maat is voor de zekerheid van de identificatie van de fosforyleringsplaats dan de Mascot score alleen, wanneer zowel CID als ETD spectra beschikbaar zijn.

Publications

M. A. Choukrallah, D. Kobi, I. Martianov, W. W. M. Pijnappel, N. Mischerikow, T. Ye, A. J. R. Heck, H. T. M. Timmers, and I. Davidson. **RNA polymerase II transcription in murine cells lacking the B-TFIID complex.** *Manuscript submitted*

N. Mischerikow and A. J. R. Heck. **Targeted large-scale analysis of protein acetylation.** *Proteomics, 2011, 11(4), 571*

S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. R. Heck, and P. A. Pevzner. **The generating function of CID, ETD and CID/ETD pairs of tandem mass spectra: Applications to database search.** *Molecular and Cellular Proteomics, 2010, 9(12), 2840*

N. Mischerikow, P. van Nierop, K. W. Li, H. G. Bernstein, A. B. Smit, A. J. R. Heck, and A. F. M Altelaar. **Gaining efficiency by parallel quantification and identification of iTRAQ-labeled peptides using HCD and decision tree guided CID/ETD on an LTQ Orbitrap.** *Analyst, 2010, 135(10), 2643*

G. Spedale, N. Mischerikow, A. J. R. Heck, H. T. M. Timmers, and W. W. M. Pijnappel. **Identification of Pep4p as the protease responsible for formation of the SAGA-related SLIK protein complex.** *Journal of Biological Chemistry, 2010, 285(30), 22793*

N. Mischerikow, A. F. M. Altelaar, J. D. Navarro, S. Mohammed, and A. J. R Heck. **Comparative assessment of site assignments in CID and ETD spectra of phosphopeptides discloses limited relocation of phosphate groups.** *Molecular and Cellular Proteomics, 2010, 9(10), 2140*

N. Mischerikow, G. Spedale, A. F. M. Altelaar, H. T. M. Timmers, W. W. M. Pijnappel, and A. J. R. Heck. **In-depth profiling of post-translational modifications on the related transcription factor complexes TFIID and SAGA.** *Journal of Proteome Research, 2009, 8(11), 5020*

M. L. Hennrich, P. J. Boersema, H. van den Toorn, N. Mischerikow, A. J. R. Heck, and S. Mohammed. **Effect of chemical modifications on peptide fragmentation behavior upon electron transfer induced dissociation.** *Analytical Chemistry, 2009, 81(18), 7814*

Curriculum vitae

Nikolai was born on 10 February 1978 in Alfeld, Lower Saxony, Germany, to Dr. Klaus-Dieter and Ulrike Mischerikow, and grew up with two sisters in Bonn and Hannover. He finished his secondary education in 1997 with the *Abitur*, with majors in chemistry and biology, at the humanistic Matthias-Claudius-Gymnasium, Gehrden. In 1997 he took up studies in Philosophy, until 2003, and in 1998 additionally in Biochemistry at the University of Hannover and Hannover Medical School. In 2005 he graduated as *Diplom-Biochemiker* with majors in biochemistry, biophysical chemistry and physical chemistry. During his thesis at the Institute of Biophysical Chemistry at Hannover Medical School under supervision of Prof. Dr. Dietmar Manstein he worked on the purification of the Myosin J motor domain from *Dictyostelium discoideum*, its crystallization and the transient kinetics of its dissociation from F-actin *in vitro* using caged ATP. In 2005 and 2006 he worked at the Department of Neurobiology at the Max Planck Institute for Biophysical Chemistry, Göttingen, on the cloning, expression, and purification of metazoan SNARE membrane fusion complexes under supervision of Dr. Dirk Fasshauer. In 2007 Nikolai started as *Promovendus* in the Biomolecular Mass Spectrometry and Proteomics Group at Utrecht University under supervision of Prof. Dr. Albert Heck. The research was centered around the general transcription factors TFIID and SAGA and carried out in collaboration with Prof. Dr. Marc Timmers and Dr. Pim Pijnappel at the Department of Molecular Cancer Research at the University Medical Center Utrecht. The results of his work are presented in this doctoral thesis and in co-authored publications.

Acknowledgements

First I would like to thank my promotor and supervisor, Albert. I am very grateful that you gave me the opportunity to pursue a PhD in your group, especially bearing in mind that you hired me not straight after my studies but after a little detour. It was a great experience working in this international, well-equipped and strongly funded lab. Thank you for the creative freedom you left to me and the always extremely generous support. Probably most of all – all is well that ends well – I appreciate the personal education that I directly and indirectly received from you. Logically that did not always go without friction and I am very glad that you tolerated my *eigenzinnigheid*.

I would also like to thank my collaborators and supervisors from the University Medical Center. Marc, thank you for integrating me into your group by invitations to lab excursions, group meetings and journal clubs. I always found it very stimulating to discuss science with you. Pim, thank you for the meticulous training in yeast and mammalian cell culture, metabolic labeling, cell extractions, purifications, and so much more. Especially for the introduction into ES cells. I am very grateful that you put so much time and effort in preventing me becoming *just a mass spec guy who does not know anything about biology*. Besides that I have to thank you for teaching me the language of the pigs. The way you demonstrated it was simple yet very metaphorical. I have trained it well and every time I speak it, this day in the cold room zaps into my mind and makes me smile.

Finally I would like to thank my fourth and fifth supervisors, Maarten and Shabaz. I thank both of you that I could always walk into your offices and riddle you with questions. Maarten, thank you for picking me up at the ion trap and transferring me to a real mass spectrometer. I very much appreciated your calm way of supervision. You have also been a great office mate and I enjoyed the many conversations and the excessive fun we had during this time. Shabaz, for us it took relatively long to adopt to each other, but I am glad that we were able to find a common level. A breakthrough for me was to accept that maintaining our lab operational and implementing new technologies accessible for everyone may require harsh words sometimes and a bit of a temper to add the necessary authority to them. Thank you very much for the supervision of the scrambling project and your professional explanations and discussions of so many other things as well.

Of course I also would like to thank my colleagues, mainly from the Heck and Timmers groups, for dinners and drinks, for barbecues and weekend excursions, for interesting conversations and discussions, for funny evenings and exciting nights out, for scientific advice and help, for being great office mates, and so on. You are so many! I thought how to thank you best, but as the distribution of *thank you scores* for every *person thank you match* (PTYM) is strictly unimodal – and to make matters more complicated not constant with time –, I am reluctant to estimate a cutoff score that represents a reasonable tradeoff between significant and insignificant PTYMs. So I hereby break with our unwritten tradition of reporting all PTYMs, and leave it with a

BIG THANK YOU

to all of you. I am confident – since the *person* component confers every PTYM the unique property of consciousness – that the PTYM outliers know for themselves that they are positioned in the tails of the distribution. I explicitly wish to thank Reinout for the favor of translating the summary of this thesis into Dutch, as well as Corine for her great and always friendly help with organizational matters.

Finally, I would like to thank my parents and Soudeh. Soudeh, although our relationship did not endure the hardship of living on a distance, it would be more than unjust not to mention that without your love, liveliness and spirit, and without our scheherazadian *week-ends à Paris* which gave me so much energy, my time here in Utrecht would have been unimaginably harder. My parents – your idealism and freedom have always greatly inspired me anew when I visited you. Thank you so much for your trust in me and the generous support during all these long years.

Abbreviations

Trademarks as well as names of chemicals and proteins are not included in this list.

AC	alternating current
AGC	automatic gain control
AP	affinity purification
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
BD	bromodomain
BSA	bovine serum albumin
CID	collision induced dissociation
COFRADIC	combined fractional diagonal chromatography
DC	direct current
DNA	deoxyribonucleic acid
ECD	electron capture dissociation
EM	electron microscopy
ESC	embryonic stem cell
ESI	electrospray ionization
ETD	electron transfer dissociation
FASP	filter-aided proteome preparation
FBS	fetal bovine serum
FDR	false discovery rate
FWHM	full width at half maximum
GC-MS	gas chromatography-coupled mass spectrometry
GF	generating function
GTF	general transcription factor
H/L	heavy/light ratio
HCD	higher-energy collisional dissociation
HFD	histone fold domain
HPLC	high pressure liquid chromatography
IEF	isoelectric focusing
ICAT	isotope-coded affinity tagging
ICR	ion cyclotron resonance
IMAC	immobilized metal affinity chromatography
IP	immunoprecipitation
iTRAQ	isobaric tagging for relative and absolute quantification
IVT	in vitro transcription
LAP	localization and affinity purification
LC	liquid chromatography
LC-MS	liquid chromatography-coupled mass spectrometry

LIT	linear quadrupole ion trap mass analyzer
m/z	mass-to-charge ratio
MALDI	matrix assisted laser desorption ionization
MEF	mouse embryonic fibroblast
mESC	mouse embryonic stem cell
MRM	multiple reaction monitoring
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MudPIT	multidimensional protein identification technology
MW	molecular weight
nanoESI	nanoelectrospray ionization
NCE	normalized collision energy
NTP	nucleotide triphosphates
PEP	posterior error probability
PIC	pre-initiation complex
Pol II	DNA-directed RNA polymerase II
PQD	pulsed Q collision induced dissociation
PRM	prefix residue mass
PSM	peptide spectrum match
PTM	post-translational modification
PTR	proton transfer dissociation
Q	linear quadrupole mass analyzer
q	linear quadrupole collision cell
QIT	3D quadrupole ion trap mass analyzer
QUBIC	quantitative BAC-GFP protein interactomics
RF	radio frequency
RNA	ribonucleic acid
RP	reversed phase
RP-LC	reversed phased-liquid chromatography
RT-qPCR	reverse transcription-coupled quantitative polymerase chain reaction
SC	synthetic complete
SCX	strong cation exchange
SDS PAGE	SDS polyacrylamide gel electrophoresis
SILAC	stable isotope labeling with amino acids in cell culture
SRM	selected ion monitoring
TAF	TBP associated factor
TAP	tandem affinity purification
TIC	total ion current
TOF	time-of-flight
TSS	transcriptional start site
YEPD	yeast extract peptone dextrose