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

Novel Concepts for Identifying Protein-Protein Interactions and Unusual Protein Modifications

Eva Christina Keilhauer

aus
Mindelheim, Deutschland

2015

Erklärung

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

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 15.05.2015

Eva Christina Keilhauer

Dissertation eingereicht am 15.05.2015

1. Gutachter: Prof. Dr. Matthias Mann

2. Gutachter: PD Dr. Sandra Hake

Mündliche Prüfung am 22.06.2015

Abstract

About fifteen years ago, the complete sequence of the human genome had been decoded. Great hopes were pinned on this major achievement of modern science. However, in fact genes are merely the building plan of a cell, and it is their products, the proteins, that execute all functions in biological processes. Hence now *proteomics*, the branch of science investigating proteins, is a great new hope.

During the last couple of years, mass spectrometry has evolved to be the main workhorse of proteomics research. *Mass spectrometry-based proteomics* has developed into a versatile tool for investigating a wide variety of questions. Next to studying the protein inventory of cells, mass spectrometry can be used to measure the quantity of each protein, to decode interaction networks between proteins, to detect chemical modifications attached to proteins, and much more.

In this PhD work, I have applied mass spectrometry-based proteomics to all of the aforementioned applications. In my first and main project, I have developed a new concept for efficiently mapping protein-protein interactions in yeast. Following up on this work, I have contributed to a collaborative effort to further develop the yeast interaction method into a high-throughput pipeline. Furthermore, I have successfully applied my knowledge about interactomics in a collaboration project on human histone variants. I have also applied mass spectrometry to explore protein modifications. In the first such project, I showed that mass spectrometry even allows to unravel completely new and previously unknown modifications, by discovering the modification that activates elongation factor P in certain bacteria. Finally, I investigated glycation, a protein modification relevant in diabetes, following the recent trend of mass spectrometry moving into clinical applications.

Zusammenfassung

Vor ungefähr fünfzehn Jahren wurde die Sequenz des menschlichen Genoms entschlüsselt. Diese Errungenschaft der modernen Wissenschaft war mit vielen Hoffungen verbunden. Allerdings sind die Gene nur der Bauplan einer Zelle. Tatsächlich werden alle Funktionen in biologischen Prozessesen von den Produkten der Gene, den Proteinen ausgeführt. Daher ist nun die *Proteomforschung*, die sich mit der Untersuchung von Proteinen beschäftigt, ein großer neuer Hoffnungsträger.
Während der letzten Jahre hat sich die *Massenspektrometrie* zur meistgenutzten Methode in der Proteomforschung herausgebildet. *Massenspektrometrie-basierte Proteomforschung* is heute eine flexible Technologie, mit der eine Vielzahl von Fragen beantwortet werden kann. Diese Technik kann nicht nur dazu verwendet werden, das Protein-Inventar einer Zelle zu bestimmen, sondern auch die Menge jedes einzelnen Proteins. Außerdem können Interaktions-Netzwerke zwischen Proteinen entschlüsselt werden, chemische Modifikationen an Proteinen entdeckt werden, und vieles mehr.
In der vorliegenden Doktorarbeit habe ich massenspektrometrische Methoden in allen zuvorgenannten Einsatzgebieten angewendet. In meinem Hauptprojekt habe ich ein Konzept zur effizienten Analyse von Protein-Protein Interaktionen in Hefe entwickelt. In einer Weiterführung dieser Arbeit trug ich dazu bei, die Hefe-Interaktionsmethode in eine Hochdurchsatzmethode weiterzuentwickeln. Des weiteren habe ich mein Wissen über Proteininteraktionen erfolgreich in ein Kollaborationsprojekt über menschliche Histonvarianten eingebracht. Neben der Analyse von Proteininteraktionen habe ich massenspektrometrische Methoden auch zur der Analyse von Proteinmodifikationen verwendet. Im ersten derartigen Projekt habe ich die Modifikation entdeckt, durch die Elongationsfaktor P in bestimmten Bakterien aktiviert wird. Somit konnte ich zeigen, dass die Massenspektrometrie es sogar ermöglicht bis dato unbekannte Modifikationen zu erforschen. Schließlich habe ich eine Proteinmodifikation namens Glykierung untersucht, die zur Pathologie von Diabeteserkrankungen beiträgt. Dieses Projekt folgt dem derzeitigen Trend, massenspektrometrie-basierte Proteomforschung nun auch zur Beantwortung von klinischen Fragestellungen anzuwenden.

Contents

Abbreviations

2D-GE	Two-dimensional gel electrophoresis
E.coli	*Escherichia coli*
AE-MS	Affinity enrichment mass spectrometry
AP-MS	Affinity purification mass spectrometry
BAC	Bacterial artificial chromosome
CID	Collision-induced dissociation
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
ECD	Electron-capture dissociation
EF-P	Elongation factor P
eFT	enhanced Fourier Transformation
EGF	Epidermal growth factor
ELISA	Enzyme-linked immunosorbent assay
emPAI	Exponentially modified protein abundance index
ESI	Electrospray ionization
ETD	Electron-transfer dissociation
FDR	False discovery rate
FFPE	Formalin fixed paraffin embedded

FT	Fourier transformation
FT-ICR	Fourier-transform ion cyclotron resonance
GFP	Green fluorescent protein
HbA1c	Glycated hemoglobin
HCD	Higher-energy collisional dissociation
HDAC	Histone deacetylase
HPLC	High pressure liquid chromatography
IA-MS	Immunoaffinity-based mass spectrometry
iBAQ	Intensity-based absolute quantification
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LFQ	Label-free quantification
LTQ	Linear trap quadrupole
m/z	Mass-over-charge ratio
MALDI	Matrix-assisted laser desorption/ionization
MS	Mass spectrometry
PAI	Protein abundance index
ppm	parts-per-million
PTM	Posttranslational modification
QUBIC	Quantitative BAC-GFP interactomics
RNA	Ribonucleic acid
SC	Spectral counting

SILAC	Stable isotope labeling with amino acids in cell culture
TAP	Tandem affinity purification
TGF-β	Transforming growth factor β
TOF	Time-of-flight
UHPLC	Ultra high performance liquid chromatography
Y2H	Yeast two-hybrid

1 Introduction

1.1 Proteins –The functional units of life and their investigation

The term 'protein' was first introduced by Gerardus Johannes Mulder in 1839, who in turn got the suggestion for this naming from Jöns Jakob Berzelius. Based on his experiments on the composition of various proteins like albumin, fibrin and casein, Mulder had developed the theory that there must be a common substance in all proteins that otherwise differ only in their sulfur and phosphorus content. Berzelius suggested to call this common substance 'protein' as deduced from the Greek word '$\pi\rho\omega\tau\varepsilon\iota\varsigma$' meaning 'fundamental'. Although the conclusions that Mulder drew from his experiments were incorrect owing to the limited methodologies at that time, the term 'protein' is actually very well deserved. Indeed proteins are fundamental for all living organisms. They provide structure, transport other molecules, catalyze reactions, forward signals, and basically execute or at least participate in every single biological process in living organisms. The totality of proteins that a particular cell, tissue or organism is expressing at a given point in time and under given conditions has been termed the *proteome*, in analogy to the *genome*, the totality of genes in a given organism. The branch of science that is trying to investigate the whole protein content of cells, tissues and organisms is called *proteomics*.

Owing to the complex and dynamic nature of the proteome, its investigation is inherently complicated and needs specialized methods [1]. These methods need to fulfill two main criteria to successfully identify proteome changes between samples. Firstly, they need to provide sufficient depth in order to access the whole dynamic range of existing proteins. The dynamic range is defined as the difference in abundance between the lowest and the highest abundant protein and reaches around seven orders of magnitude in cells [2]. Secondly, they need to be of a quantitative nature, as often not the identity but the amount of individual proteins changes in a cell upon perturbation [3]. The first technique trying to monitor whole proteome changes was two-dimensional gel electrophoresis (2D-GE) [4]. However, with 2D-GE only the most abundant proteins could be detected, meaning such screens could not really decode a proteome with all its components. Back then, determining the identity of the protein spots in the gel was also cumbersome as no fast and sensitive method for protein identification existed [1]. Hence, the introduction of *mass spectrometry* (MS) into the field of proteomics in the 1990s

presented a giant leap forward. Due to the development of soft ionization methods like matrix-assisted laser desorption/ionization (MALDI) [5] and particularly electrospray ionization (ESI) [6], proteins could now be analyzed by MS, which had not been possible before. This major achievement in protein science was also recognized by the Nobel committee, which awarded part of the 2002 Nobel prize in chemistry for those protein ionization techniques.
After many genomes of lower complex organisms had already been sequenced, around the year 2000 the entire human genome sequence was finally decoded in a large collaborative effort [7, 8]. The resulting protein sequence databases are now the basis for identifying proteins in a rapid and routine manner by combining MS data acquisition with database searching.
The introduction of mass spectrometry into the field of proteomics enabled the detection first of hundreds then of thousands of proteins in a single experiment, and the numbers have been growing ever since.

1.2 Mass spectrometry-based proteomics

Mass spectrometry is an analytical method that determines the mass of ionized analytes in the gas phase. A mass spectrometer consists of five basic parts: (1) An ion source transferring the analyte into the gas phase (2) some kind of ion optics guiding the ions through the mass spectrometer, (3) a fragmentation device, (4) a mass analyzer that determines the *mass-over-charge ratio* (m/z) of the ionized analyte, and (5) a detector that measures the number of analytes at each m/z ratio. For a long time, mass spectrometers were successfully used in the analysis of small molecules. However, to use mass spectrometry for the analysis of proteins a major obstacle had to be overcome. Being fairly large and heat-labile biomolecules, proteins were incompatible with MS using the ionization techniques available at that time. However, this issue was solved by the introduction of new ionization strategies called MALDI and ESI. Proteins could now successfully be transferred from a solid matrix or a liquid, respectively, into the gas phase and hence analyzed by MS. ESI in particular was a huge success story for proteome analysis as it allowed on-line coupling of high pressure liquid chromatography (HPLC) to mass spectrometry, which in turn allows 'presorting' of peptides and therefore much greater analysis depth. Since then, many technical developments have been implemented to

improve mass spectrometry-based proteomics, and a variety of different methodologies and applications have evolved [9].

1.2.1 Basic principles

Two important basic concepts in MS-based proteomics are top-down and bottom-up analysis. In the *top-down* approach, full length proteins are ionized and analyzed by tandem MS. The proteins can be measured in a denatured state, or as natively folded proteins, which even enables the analysis of whole protein complexes (Native MS [10]). Top-down techniques have two major advantages: (1) The whole protein sequence is accessible for analysis, hence full sequence coverage can in principle be achieved. (2) All protein isoforms can be detected, i.e. all splice variants and all modifications with their localization. However, mass spectra of full length proteins are exceedingly complicated and not easy to interpret, therefore generally top-down approaches can only be applied to purified proteins or very low complexity mixtures. It is difficult to couple top-down MS with liquid chromatography, because whole proteins require long analysis times in the mass spectrometer in order to achieve the mass accuracy and resolution needed for their identification [11]. Furthermore, the size of the protein itself is a limiting factor: the larger the protein, the harder it is to analyze by MS. Despite these issues, some remarkable results have recently been achieved using standard (recent review: [12]) and native top-down methods (recent review: [13]).

The second, easier and much more widely used MS-based proteomic approach is *bottom-up* or *shotgun* proteomics (see Figure 1, page 4). In this concept, the proteins are cleaved into smaller peptides using specific proteases. The most commonly used protease for this task is trypsin [9]. Tryptic peptides are of an ideal length for HPLC separation, MS analysis and efficient fragmentation. Trypsin specifically cleaves on the C-terminal side of lysine and arginine residues, two basic amino acids that can carry a positive charge on their side chains [14]. Therefore digestion with trypsin facilitates efficient ionization of the peptides, which enables their analysis by MS. Some other proteases are also suitable for cutting proteins into small peptides, e.g. LysC, GluC, chymotrypsin, AspN and ArgC [15–17]. Finally, other proteases like outer membrane protease T cut proteins into relatively long peptides or protein fragments, an approach in between the other two called *middle-down* proteomics [18].

Bottom-up proteomics is typically used to analyze highly complex mixtures, like whole

cell lysates. Of course digesting such complex protein mixtures results in even more complex peptide mixtures, therefore HPLC is essential to reduce this complexity. For this purpose, the peptide mixture is loaded onto a chromatographic column packed with a material the peptides interact with, usually a hydrophobic *reverse phase* material e.g. C_{18}. The peptides bind to the C_{18} material with different strengths according to their chemical properties, and can gradually be released from the column by increasing the organic content of the chromatography solvent.

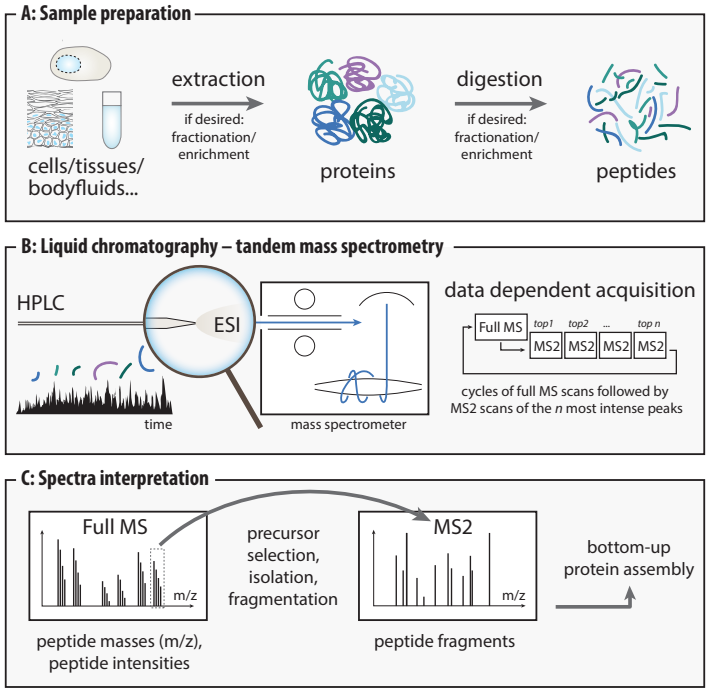


Figure 1: Standard shotgun proteomics workflow. A Every proteomics experiment starts with the extraction of proteins from the corresponding sample. Depending on the application, the resulting protein mixture can be fractionated or enriched for a certain protein population. In the next step, the proteins are digested into peptides, usually using trypsin. The resulting peptide mixture can also be fractionated or enriched for a certain population, e.g. modified peptides. **B** Since the resulting peptide mixture is still highly complex, peptides are usually separated by HPLC. From the chromatography column, the peptides can be directly sprayed into the mass spectrometer and transferred to the gas phase by ESI. MS data is usually acquired in a data-dependent fashion. **C** From the full scan, the *top n* peptide features are selected for fragmentation and further analyzed in MS2 scans. Computational software can then identify the measured peptides from the acquired data and reconstruct the corresponding proteins. Adapted from [19].

Nevertheless, not one but many peptides elute from the chromatography column at a given point in time. To identify the eluting peptides, in most cases *data-dependent ac-*

quisition (DDA) is used, which works as follows: First the mass spectrometer acquires a full scan (also called MS scan or survey scan) monitoring all peptide features that elute from the HPLC column at a given point in time. The full scan is usually recorded at high resolution and it hence measures very accurately the mass and intensity of every peptide. The accurate mass of a peptide is however not sufficient for its identification, since different amino acid sequences can result in the same peptide mass. Therefore the peptides have to be fragmented in order to acquire sequence information. For that purpose in DDA the *top n* most abundant peptides features from the full scan are sequentially isolated and fragmented, acquiring so called MS2 scans or fragmentation scans. The MS2 scans can be recorded either at low resolution e.g. in a linear ion trap (*high-low strategy*), or at high resolution e.g. in an Orbitrap analyzer (*high-high strategy*). After all *top n* features have been analyzed, the instrument records the next full scan and the cycle starts anew. To fragment as many peptides as possible, the cycle time should be adapted to the standard peak width of the chromatography setup. The number of *top n* features that can be fragmented within one cycle depends strongly on the instrument speed. For this reason, the scan speed of a mass spectrometer is of crucial importance for proteomic analysis of complex samples. Even though DDA methods are very efficient for analyzig complex mixtures, they still miss many lower abundant peptide features, a phenomenon known as the *undersampling problem*. The number of detectable peptides in a standard single-shot measurement of a HeLa cell lysate was estimated to be more than 100.000, however only around 10.000 of them could be identified (see Figure 2) [20].

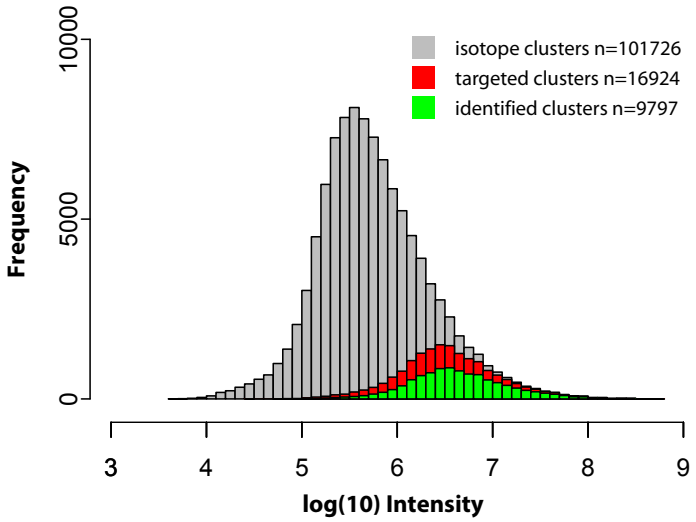


Figure 2: Undersampling in data-dependent acquisition The grey histogram depicts all peptide features detected in digested HeLa cell lysate. The red histogram shows those features that were targeted for fragmentation using a *top 10* method. Finally, the green histogram shows those peptides features that could be identified in the end. Adapted from [20].

One approach to overcome the undersampling issue and to achieve a higher dynamic range and sensitivity is *data-independent acquisition* (DIA) [21]. However, also DIA approaches can naturally not achieve a higher dynamic range than what is given by the instrument. In DIA, not only a selection, but in fact every detectable feature is fragmented (*all-ion-fragmentation*). The resulting MS2 spectra are inherently highly complex and multiplexed, hence DIA requires very sophisticated data analysis. To at least partially reduce this complexity, one of the most well-known DIA approaches called 'SWATH' fragments mass windows of a fixed width instead of the entire mass range [22]. This is done by rapidly scanning through the whole mass range in consecutive mass isolation windows of typically 25 Da. In SWATH, the data-independent acquisition is then combined with targeted data analysis. This means only a certain number of peptides with known fragmentation behavior (which has to be established beforehand) is effectively followed. In this way, the method achieves high quantification accuracy for the targeted peptides. However, so far SWATH cannot compete with DDA for whole proteome analyses in terms of proteome coverage, which also holds true for all other DIA approaches.

Fragmentation techniques

Several methods can be used to fragment peptides into smaller parts in order to obtain sequence information. In the most classical fragmentation technique, collision-induced dissociation (CID), the peptide ions are accelerated to high kinetic energy and then collided with an inert gas like nitrogen, helium or argon [23]. CID was initially performed in triple quadrupole instruments (*beam-type CID*). Later also linear trap quadrupole (LTQ) cells were used for CID fragmentation, however, this *trap-type CID* approach suffers from a low mass cutoff problem. Because the generation and the detection of the ions happens within the same device (*tandem-in time principle*), not all created peptide fragments can be efficiently stabilized after fragmentation and product ions below a certain mass cutoff are lost. Higher-energy collisional dissociation (HCD) is an advanced version of CID, featured in Orbitrap instrumentation and typically performed in a specialized octopole collision cell [24]. In principle HCD strongly resembles beam-type CID and hence also allows low mass fragment ions to be observed, because the generation and the detection of the ions is separated in space (*tandem-in space principle*). Both CID and HCD techniques preferentially fragment peptides at the peptide bonds, leading to the formation of so called *b*-and *y*-ions (see Figure 3A, page 7).

Electron-capture dissociation (ECD) on the other hand induces fragmentation by letting peptide ions interact with free electrons [25]. Finally, electron-transfer dissociation (ETD), an advanced version of ECD, induces fragmentation by colliding the peptides with anthracene or fluoranthene anions [26]. The latter two techniques primarily lead to the formation of *c*- and *z*-ions (see Figure 3A).

While ECD and ETD have advantages for longer peptides or even entire proteins (top-down) or peptides carrying labile modifications, CID and HCD are more powerful for effective fragmentation of short tryptic peptides and peptides carrying stable modifications [27]. Hence CID and HCD are the most widespread techniques for fragmenting peptides in shotgun proteomics. HCD is becoming more widely adapted, because the resulting spectra additionally contain the informative low-mass region. This is particularly important for reporter-based quantification techniques that require this low mass range to be observed (see section 1.2.4).

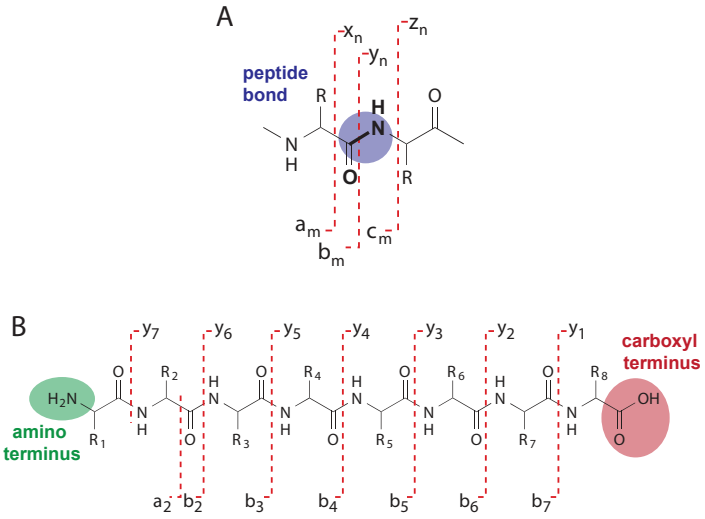


Figure 3: Peptide fragmentation. **A** The most sequence informative fragments are obtained by peptide backbone fragmentation. Depending on the fragmentation technique, different fragments are observed. Corresponding to the cleavage site and which terminus they retain, they are designated as *a*, *b*, *c* and *x*, *y*, *z*-ions (Roepstorff-Fohlmann-Biemann nomenclature [28, 29]). **B** Theoretically, complete sequence coverage can be obtained by fragmenting a peptide, in this example by HCD, resulting in a complete *y*-and *b*-ion series. The *y*-ions are numbered consecutively from the original C-terminus, the *b*-ions are numbered consecutively from the original N-terminus. The difference between consecutive ions (b_m and b_{m+1} and y_n and y_{n+1},respectively) yield the masses of the corresponding amino acids. Next to the *y* and *b* ions, in HCD often an a_2 ion is observed. Adapted from [30].

Fragmentation behavior of peptides in HCD

The fragmentation of peptides in HCD is charge directed and results in a variety of fragment types. Some of the most relevant fragments observed in HCD will be explained in the following.
In general, the most sequence informative fragments are those obtained by peptide backbone fragmentation designated *a*, *b* and *c* for those retaining the N-terminus and *x*, *y* and *z* for those retaining the C-terminus (see Figure 3A, page 7) [30]. In low-energy fragmentation techniques such as in CID/HCD, the lowest energy pathway is naturally favored, which is the breakage of the amide bonds leading to the formation of *b*-and *y*-ions. In principle it is possible to determine the complete amino acid sequence of a peptide in this way, provided breakage occurs at every amide bond (*de-novo sequencing* by MS; see Figure 3B, page 7). However, since every molecule is ideally breaking only once, a relatively high number of peptide ions needs to be collected and fragmented for this purpose. Hence in practice, complete *b*-and *y*-ion-series are rarely observed in high complexity samples, necessitating the use of database searching for peptide identification.
How the *b*-and *y*-ions are created can at least qualitatively be explained by the *mobile proton model* (see Figure 4, page 9) [31, 32]. The prerequisite for peptide fragmentation is protonation during ionization. The proton(s) are initially sitting on basic residues of the peptide, e.g. the terminal amino group or arginine and lysine side chains. Hence tryptic peptides usually carry at least two charges [30]. After ionization, the protons are initially quite tightly bound to the basic residues. However, during fragmentation the peptide ions are excited, and as their internal energy increases, one proton becomes 'mobile' and can move to energetically less favored protonation sites, such as the nitrogens of the backbone amide bonds (see Figure 4, page 9). Protonation on the amide nitrogen leads to a considerable weakening of the amide bond, however, direct bond cleavage is disfavored in the low energy fragmentation regime. Instead, the dissociation of the peptide into two parts occurs via more complex rearrangement reactions. Protonation on the amide nitrogen makes the carbon atom of the amide bond a likely target for nucleophilic attack. Consequently it is attacked by the oxygen of the N-terminally neighboring amide bond (red arrow in Figure 4, page 9). This leads to the formation of an oxazolone ring structure and dissociation of the peptide bond, resulting in a *b*- and a *y*-fragment. Which one of the two is actually observed depends on which one retains the proton; this in turn depends on the proton affinity of the two fragments.

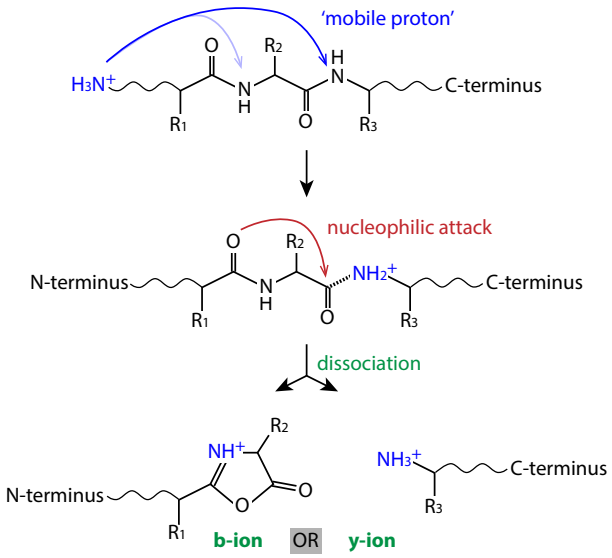


Figure 4: The b-y fragmentation pathway as explained by the mobile proton model.
Adapted from [32].

Other 'sequence informative' backbone fragments can also occur in HCD, e.g. *a*-ions that are generated from *b*-ions via the loss of CO. However, this mostly occurs for the b_2-ion leading to a characteristic a_2/b_2-fragment ion pair in HCD spectra [33]. Many other fragments are produced by losing small neutral molecules, mostly water and ammonia [32]. This leads to the formation of $[MH\text{-}H_2O]^+$, $[MH\text{-}NH_3]^+$, $[y_n\text{-}H_2O]^+$, $[y_n\text{-}NH_3]^+$,$[b_m\text{-}H_2O]^+$, and $[b_m\text{-}NH_3]^+$ ions. Water can be lost from the C-terminal COOH group or the aspartic and glutamic acid COOH groups, and from serine and threonine side chains; ammonia can be lost from the N-terminus and the side chains of arginine, asparagine and glutamine. Other frequently observed neutral losses in HCD spectra include CH_4SO from oxidized methionine, CH_3NO from glutamine or asparagine, and C_2H_4O from threonine [33]. Furthermore, HCD fragmentation produces internal fragments that result from *b*-or *y*-ions undergoing a second cleavage. Such internal fragments are characteristic for HCD, a *beam type* fragmentation method, and occur much less in *trap-type* CID.

As mentioned before, another characteristic of HCD is a number of fragments in the low mass range, that cannot be observed in trap-type CID. These include immonium ions originating from arginine, lysine, phenylalanine, tryptophane, tyrosine, histidine, glutamine, and glutamic acid. A special case is the immonium ion generated by phosphotyrosine, which can be used as a reporter ion to verify the existence of this PTM [33]. Finally, side chain fragments of several amino acids can be observed in the low mass

range.
Detailed knowledge about peptide fragmentation behavior can be used to explain more peaks in fragmentation spectra and gain more confidence in identifications. This was recently demonstrated by developing an 'expert system' for computer-assisted annotation of MS2 spectra [34]. In this study, including other fragment types just discussed in addition to the classic *b*-and *y*-ions increased the intensity coverage of fragment peaks from 56% to 86% in a typical shotgun experiment.

Mass analyzers

Five basic types of mass analyzers are used for proteomics experiments: Quadrupole analyzers, ion trap analyzers, time-of flight (TOF) analyzers, Fourier-transform ion cyclotron resonance (FT-ICR) analyzers [35], and Orbitrap analyzers [36]. While quadrupole and TOF analyzers continuously scan incoming ions (*beam-type analyzers*), ion trap, FT-ICR and Orbitrap analyzers capture certain ion populations and perform sequential processes on them (*trap-type analyzers*). The different mass analyzers have different properties, with the key parameters being resolution, sensitivity, mass accuracy and speed.
The *resolution*, calculated as the m/z value divided by the the width of the peak at half of its height, is a measure of how well two different peaks of slightly different m/z ratios can be detected as such. Ion traps and quadrupoles typically have a low resolution (~1000), TOF instruments perform better (>10.000), however, by far the highest resolving power is provided by FT-ICR and Orbitrap analyzers (>100.000). For the latter two, this high resolution can be achieved because both measure frequencies of circulation ions, which can be measured in a highly accurate fashion [30]. The Orbitrap mass analyzer provides this high resolution at a much lower price and footprint than the FT-ICR instruments, which are equipped with expensive ultra strong magnets. Therefore nowadays the Orbitrap is the preferred high-resolution analyzer.
The *sensitivity* of a mass analyzer is dependent on the detection principle. Standard ion traps and linear ion traps [37] employ electron multipliers as detectors, which are capable of detecting single ions and therefore highly sensitive. Detection based on Fourier transformation usually requires a few more charges to distinguish a signal from the noise. However, in the Orbitrap analyzer single ion detection is in principle feasible due to improved electronics and thermal stability [38].
The *mass accuracy* describes how far an experimentally determined mass deviates from

the real (theoretically calculated) mass. In general, it depends on the resolution of a mass analyzer, hence high-resolution analyzers can achieve parts-per-million (ppm) mass accuracy. The Orbitrap has been reported to even achieve sub-ppm mass accuracy [39].
The *scan speed* is roughly inversely correlated with resolution. Therefore, FT-ICR analyzers are usually slowest, ion traps and Orbitraps are much faster, and beam-type analyzers (quadrupoles and TOFs) are the fastest.
Nowadays, the dominant types of mass spectrometers for shotgun proteomics are *hybrid instruments* combining several mass analyzers. The most frequently used mass spectrometers of this type are quadrupole TOF instruments, quadrupole ion trap instruments and finally quadrupole Orbitrap instruments. Quadrupole Orbitrap mass spectrometers are particularly powerful, because the quadrupole supplies fast and accurate mass selection, and the Orbitrap mass analyzer combines outstanding resolution with high sensitivity, mass accuracy, and scanning speed.

1.2.2 Orbitrap mass spectrometry

The Orbitrap mass analyzer was used in all of the following work, hence this section will provide more details about this particular device and the instrument family that evolved around it. The Orbitrap was invented by Russian physicist Alexander Makarov and first described in 2000. Its working principle is based on trapping ions by making them orbit around and along a central spindle-shaped electrode (*electrostatic ion trap*, see Figure 5 A, page 12) [40]. While the frequency of rotation around the central electrode is dependent on several factors, like the initial ion velocity and the initial radius, the frequency of the harmonic oscillations along the the axis of the field (designated *z-axis* in Figure 5A) is only dependent on the m/z value. This axial frequency can be measured using image current detection on the segmented outer electrodes in a highly accurate fashion, and by Fourier Transformation (FT) transformed into a mass-to charge signal (see Figure 5B, page 12) [41]. Already the first Orbitrap from 2000 provided high mass resolution (up to 150.000), extremely high mass accuracy (around 5 ppm) as well as high dynamic range [40]. In the following years, the Orbitrap was even further improved. In FT-ICR analyzers, the resolution can only be increased by using a stronger magnet, which quickly raises the prices for such instruments. In an electrostatic trap like the Orbitrap on the other hand, the field strength can be increased by either applying higher voltages, or by changing the geometry of the trap [42]. The first Orbitrap cell being improved by one of

these principles was the 'High-Field Orbitrap' cell first incorporated into an instrument in 2012 [43]. This high-field Orbitrap is smaller than the standard Orbitrap (20 mm inner diameter vs. 30 mm inner diameter) and features around two-fold higher resolution. Along with the improved high-field Orbitrap cell, an enhanced Fourier Transform (eFT) algorithm was introduced that further improves resolving power. Starting in the same year, the standard Orbitrap became available with a higher central electrode voltage (now 5 kV instead of 3.5 kV). Finally, in 2014 both improvements were combined in the 'Ultra High Field Orbitrap' that is now available in the latest generation of instruments and achieves very high resolution at a very high scan speed (Specified resolution at m/z 200 is 240.000 with a transient length of 512 ms [44]).

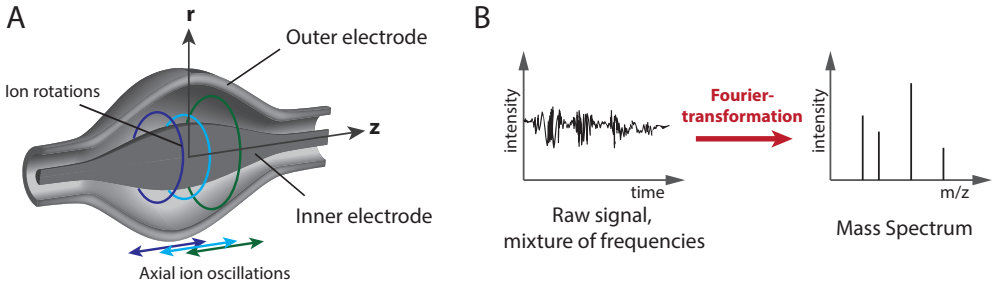


Figure 5: The basic principles of the Orbitrap mass analyzer. A Cross-section schematic representation of the Orbitrap analyzer. The Orbitrap consist of an outer barrel-shaped electrode and an inner spindle-shaped electrode. Ion populations are injected tangentially to the outer electrode, stabilized by the electric field and forced into circular motion around the inner electrode in ion packages (colored circles). The ions also oscillate harmonically from right to left within the Orbitrap along the *z-axis* (colored arrows). Adapted from [36, 40]. **B** Fourier Transformation converts the detected image current first into frequencies and then into m/z signals.

The first instrument featuring an Orbitrap mass analyzer was introduced by Thermo Fisher in 2006, and since then a whole family of Orbitrap mass spectrometers has evolved. The data in this thesis were produced on four different machines, which represent the progress of the Orbitrap instrumentalization over the last years: The 'LTQ Orbitrap' [45], the 'Orbitrap Elite' [43], the 'Q Exactive' [46] and finally the 'Q Exactive HF' [44, 47]. The four different machines and the differences between them are explained in Figure 6 on page 13.

In general, Orbitrap instrumentation is exceptionally well suited for investigating complex proteome samples, and together with improvements on the chromatography side this mass analyzer has advanced the entire MS-based proteomics field, making high-resolution mass spectrometry a standard in many laboratories.

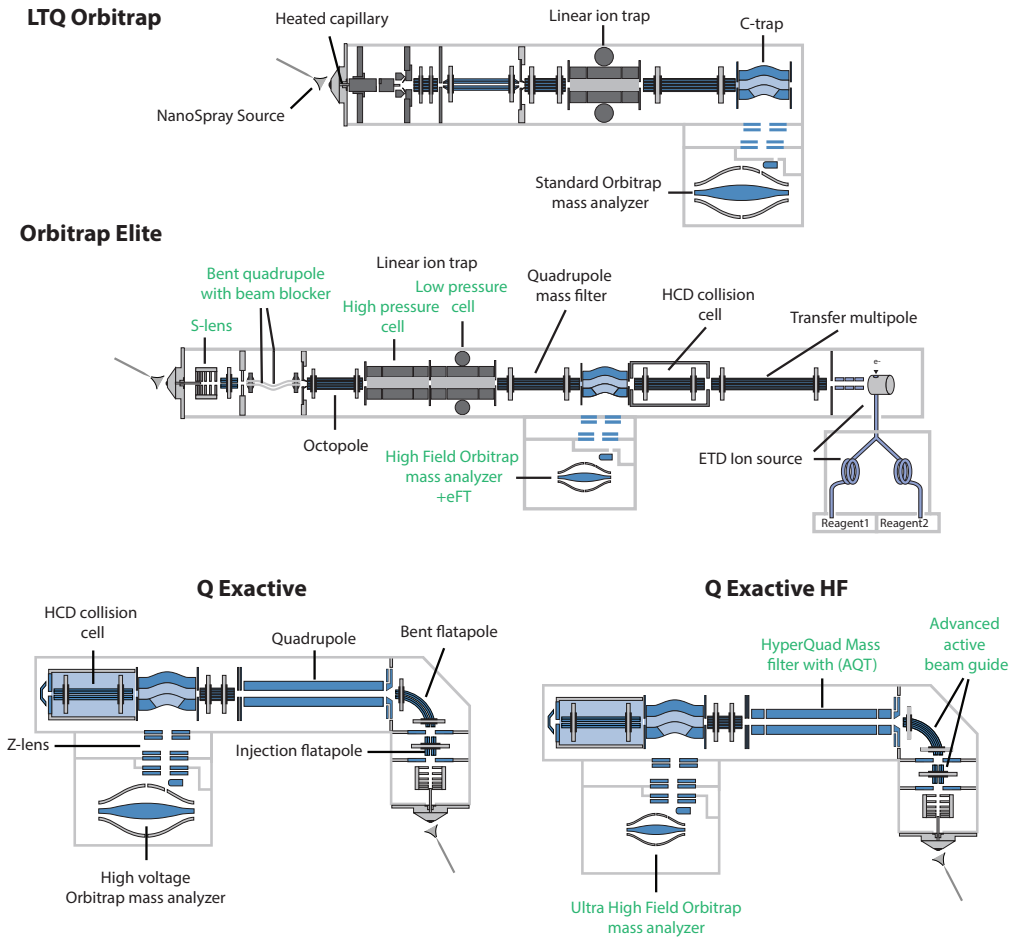


Figure 6: Schematic representations of the four types of Orbitrap mass spectrometers used in this work. The first instrument ever to feature an Orbitrap mass analyzer was the LTQ Orbitrap [45]. It is equipped with a linear ion trap for the generation of peptide fragments by CID. The fragmentation spectra are also acquired in the linear ion trap and have low resolution. The second machine used in this work was the Orbitrap Elite. It features a High-Field Orbitrap with eFT, an improved dual linear ion trap, improved ion optics, and HCD and ETD as additional fragmentation techniques [43]. The next machine, the Q Exactive, is a much simpler and smaller (benchtop) mass spectrometer, yet extremely powerful. It features a high voltage Orbitrap analyzer and HCD as the only fragmentation technique. Therefore in this instrument both full scans and fragmentation scans are always read-out in the Orbitrap with high resolution [46]. Finally, the most recently developed machine is the Q Exactive HF. Compared to the classic Q Exactive, it features improved ion optics, an improved selection quadrupole, and most importantly an Ultra High Field Orbitrap analyzer [44, 47].

1.2.3 Bioinformatic data analysis and computational proteomics

High-resolution mass spectrometry-based proteomics experiments produce a tremendous amount of data. Already a single two hour measurement of a digested HeLa lysate contains around 90.000 spectra, and the corresponding raw file has a size of around 2 GB. Therefore efficient data analysis software is required to analyze and interpret the data. Regardless of the software that is used, the bioinformatic workflow can be divided into several parts.

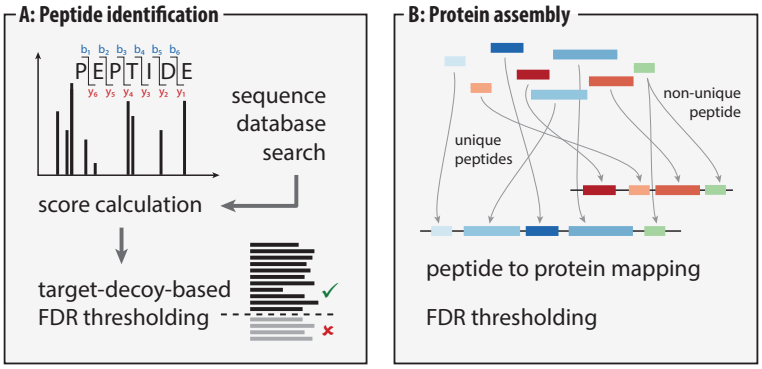


Figure 7: Computational steps to **A** identify peptides by a database search and **B** assemble proteins from the peptide identifications. Adapted from [19]

Peak detection: First of all, the peptides features in the full scans have to be identified, which requires sophisticated 3D peak detection algorithms. For each detected peptide feature, the mass-over-charge ratio and the intensity are determined.
Peptide identification: To identify peptides, the corresponding fragmentation spectra are used. However, the sequence information in these spectra is in most cases not sufficient to directly read out the peptide sequence, therefore basically all bottom-up approaches make use of protein sequence databases, e.g. FASTA files [48] obtained from UniProt [49]. Theoretical peptide lists are generated by *in-silico* digesting the proteins in the appropriate database with the same protease that has been used in the experiment. The obtained theoretical peptides are then *in-silico* fragmented using the appropriate fragmentation method. Experimentally obtained peptide and fragment m/z values are then compared to the theoretical ones. Several algorithms are available for this purpose, the most commonly used ones are SEQUEST [50], Mascot [51] and the Andromeda search engine [52] integrated in the MaxQuant environment [53]. In most cases, the search employs a *target-decoy* principle, by searching not only against the real database, but also a decoy database that contains reversed nonsense versions of the true peptide sequences

[54]. The hits to both databases are then sorted according to their score, and a cutoff is placed at a point where a certain number of hits to the decoy database have accumulated (see Figure 7 A, page 14). Typically this cutoff is set at 1% of hits to the reverse database leading to a 1% false discovery rate (FDR) at the peptide level.
Protein assembly: The next step is to assemble the identified peptide sequences into proteins (see Figure 7 B, page 14). This step is of crucial importance and at the same time not trivial, as the same peptide sequence can be present in different proteins and especially in different isoforms of the same protein [55]. Such peptides are referred to as *non-unique peptides*, while the ones unequivocally identifying a protein are referred to as *unique peptides*. There are different ways to deal with this problem. In the MaxQuant software for example, two proteins are joined into a 'protein group' whenever the set of identified peptides is the same or completely contained within the set of peptides from the other protein, because there is not enough evidence to report them as separate proteins [53]. Ideally, an FDR cutoff of 1% is also applied on the protein level, again by using a target-decoy principle, to avoid reporting of false positive protein identifications as much as possible.
Protein quantification: Proteins are quantified using different strategies that will be discussed in the following section (Section 1.2.4).

Next to these standard steps, the data can also be searched for posttranslational modifications (PTMs). Usually the modification one wants to look for is known, and usually the residues at which this modification is naturally attached to are also known. In this case, the corresponding mass difference introduced by the PTM is considered in the database search as a *variable modification*. Some artificial modifications are deliberately introduced during sample preparation, e.g. disulfide bridges are usually reduced and subsequently alkylated. This leads to all cysteine residues being modified by carbamidomethylation, which is then considered as a *fixed modification* in the search. The identification of peptides carrying such known modifications is usually straightforward. Most of the time, also the exact modification site can be determined by examining the fragmentation spectra of the modified peptides (see also section 1.3.2)
A completely different and unbiased strategy to identify PTMs was introduced with the ModifiComb algorithm [56] and is available in the MaxQuant software as an option called *dependent peptide search*. In this approach, no information about any modification is passed on to the search engine, hence in the first instance all modified peptides are not identified (see Figure 8, page 16). After this first search, the algorithm compares

all identified peptides with all unidentified peptides, and determines the mass difference ΔM for each peptide pair. Now the fragmentation spectra of the peptide pairs are compared. If some of the peaks in the unidentified spectrum are shifted by the exact same peptide mass difference ΔM, while some other fragments are identical, the unidentified peptide is assumed to be a modified version of the identified peptide. Depending on which peaks are shifted, the most likely position for the modification can be determined. Because the identification of the modified peptide is dependent on the identification of the unmodified counterpart, it is called *dependent peptide*, while the unmodified peptide is called the *base peptide*. I have used the dependent peptide search approach in this thesis to identify the modification that activates elongation factor P (see chapter 2.2.1).

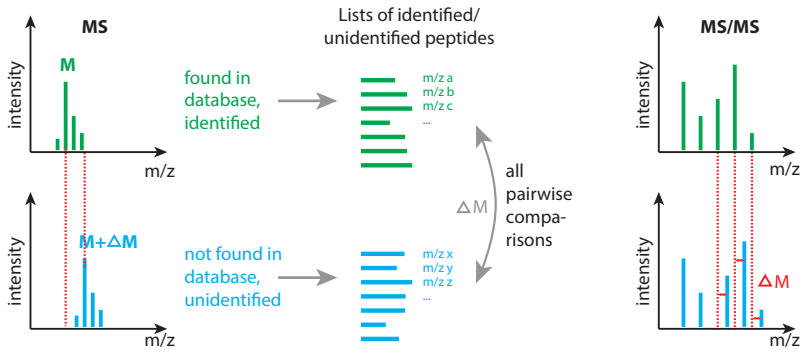


Figure 8: In the 'dependent peptide' search approach, no *a priori* information about potential modifications is passed on to the software. Hence in the first instance, all modified peptides are not identified. By performing pairwise comparisons between all peptide features, some of the unidentified peptides can be determined to be modified versions of identified peptides. Adapted from [56].

1.2.4 Protein quantification by mass spectrometry

As outlined before, it is of crucial importance to not only determine the identity of proteins present in a particular sample, but also the amount of each protein. Only if this quantitative information is acquired, biologically meaningful statements about proteome changes between samples of any kind can be made. Therefore, several MS-based quantification methods have been developed, either for relative quantification comparing protein amounts between samples, or for absolute quantification measuring the exact absolute amount of a protein in a sample.

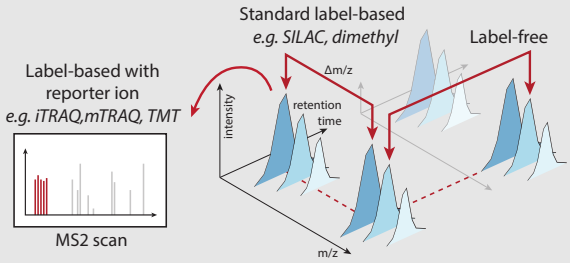
The challenge for all MS-based quantification approaches is that MS by itself is not a quantitative technique. Due to varying length and amino acid composition, tryptic peptides have different chemical properties and charge states. This results in different 'flyabilities' i.e. different ionization efficiencies and behavior in the MS. Therefore the intensity in the mass spectrometer is not directly proportional to the peptide amount, and quantification cannot rely on comparing different peptides with each other. Instead, quantification is based on comparing the intensities of identical peptides originating from different samples representing e.g. a diseased state and the corresponding control. In the bottom-up approach, quantitative values are initially obtained for peptides, but the quantitative information about all peptides originating from one protein can then be combined to obtain a quantitative value for this protein. In this way one can determine proteins changing in abundance between two samples, and hence identify important players in the process under investigation.

Overview of quantification approaches

An overview of the most important quantification approaches is given in Box 1 (page 18). There are two basic principles for MS-based protein quantification. In *label-based quantification* approaches the sample and the control are differentially labeled. For this purpose, a different number of stable (i.e. nonradioactive) 'heavy' isotopes are introduced, mostly ^{13}C, ^{15}N and $^{2}H/D$, resulting e.g. in a 'light' control and a 'heavy' sample. Most importantly, the introduction of stable isotopes does not change the physiochemical properties of the peptides, but only their mass. Hence they behave the same as their natural counterparts in the cell, during sample preparation and during HPLC separation, but can be distinguished during MS measurement. After the labeling step, the samples can be mixed and analyzed together in one liquid chromatography-tandem mass spectrometry (LC-MS/MS) run. Depending on the labeling strategy and the available number of differential labels, more or less samples can be combined (called *multiplexing*), and up to 54 different samples have already been multiplexed [57]. Dependent on the number of samples to be compared, the labeling results in multiple peaks for every peptide, separated by a characteristic mass difference, with each peak originating from one sample. In this way, the intensities of these peaks can easily be compared for each peptide.

Box1: Quantification Approaches

Approach	Principle	Name	Type	Reference
Label-based	Metabolic labeling	SILAC	relative	[58]
		Super-SILAC	relative	[59]
		^{15}N	relative	[60, 61]
	Chemical labeling	TMT	relative	[62]
		iTRAQ	relative	[63]
		mTRAQ	relative	[64]
		Dimethyl	relative	[65]
	Spike-in	Absolute-SILAC	absolute	[66]
		PrESTs	absolute	[67]
		AQUA	absolute	[68]
		QconCAT	absolute	[69]
		PSAQ	absolute	[70]
		FlexiQuant	absolute	[71]
Label-free		Spectral counting	relative	[72]
		Intensity-based	relative	[73, 74]
		iBAQ	absolute	[75]



The table on top lists the most important quantification approaches with the corresponding references. Quantification approaches can be divided in label-based and label-free approaches. Label-based approaches can be further subdivided into metabolic labeling, chemical labeling and spike-in techniques. Quantification can either be relative, comparing relative protein amounts between different samples, or absolute, determining the absolute protein concentration in a sample. The figure illustrates how quantitative information is extracted in the different quantification approaches (Adapted from [19]). In standard label-based methods, the intensities of differentially labeled peptides is compared within the same LC-MS/MS run. In label-based methods that use reporter techniques, the labels are 'isobaric' i.e. indistinguishable in the full scan, however create reporter ions of differential mass upon fragmentation, whose intensities are extracted from the MS2 scans. Finally, in label-free methods, the intensity of the same peptide in different LC-MS/MS runs is compared.

The isotopic labels can be introduced in two ways, either metabolically or chemically. In *metabolic labeling*, the stable isotopes are introduced in the living cell or organism through its metabolism, by feeding heavy-isotope-modified amino acids. Our group has pioneered the most well-known of the metabolic labeling approaches, called stable isotope labeling with amino acids in cell culture (SILAC) [58]. In *chemical labeling* techniques, the stable isotopes are added in chemical reactions during sample preparation either at the protein or the peptide level. In general, isotopic labeling approaches are very robust and accurate, especially metabolic labeling approaches where samples are combined at the very beginning of the workflow and hence no artificial changes are introduced (see Figure 9).

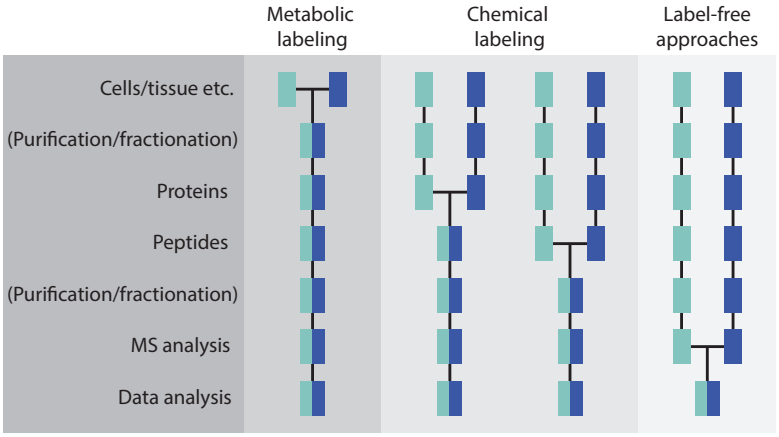


Figure 9: Parallel/separate sample processing in the three major quantitative workflows. In metabolic labeling approaches, samples are combined at the very first step of the sample processing workflow, directly after culturing/obtaining the samples. Therefore, all experimentally introduced changes affect both samples in the same way, leading to highly accurate quantification results. Chemical labeling is either performed at the protein or at the peptide level, therefore samples have to be processed in parallel for some steps, resulting in less accurate quantification. In label-free quantification procedures, the complete sample preparation is done in parallel, hence a highly reproducible workflow is a prerequisite to obtain reliable quantification results. Adapted from [76].

The second approach for MS-based quantification is called *label-free quantification*. As the name suggests, no labels are introduced, therefore samples can naturally not be mixed. Instead the control and the sample are analyzed in separate runs, and after MS measurement, peptides abundances are calculated from the acquired data using different approaches. Because samples are only combined at the stage of data analysis (see Figure 9), it is considered the least accurate quantification approach. However, relatively high accuracy quantification can still be achieved if highly reproducible sample prepar-

ation and LC-MS/MS measurement are ensured and powerful data analysis algorithms are applied that correct for the remaining variability. The work presented in this thesis was exclusively acquired using label-free quantification, hence label-free approaches will be discussed in more detail in the next section.

Label-free quantification

Label-free quantification (LFQ) is a purely computational approach, which has several basic advantages. (1) It can be applied to virtually any kind of sample, even clinical samples derived from patients [77]. (2) It is experimentally much easier to apply as no labeling has to be performed, and usually it is also less expensive. (3) It can be applied to an unlimited number of samples, while the highest number with labeling approaches for direct comparisons is 10 conditions in parallel with TMT 10-plex [78]. On the other side, due to the completely parallel sample processing, LFQ approaches usually require more replicates (at least triplicates), excellent reproducibility of the whole LC-MS/MS pipeline and sophisticated data analysis to yield accurate results.
The first, relatively simple LFQ approach was described in 2004 and used the number of acquired fragmentation spectra as a semi-quantitative approximation for protein abundance [72]. Although the number of acquired MS2 spectra is indeed directly related to the abundance of a peptide, this *spectral counting* (SC) approach suffers from several weaknesses. First of all, protein size naturally introduces a bias, as large proteins produce more peptides than smaller ones. Furthermore, the chromatography and the resulting peak width have a strong impact on the results. Small differences are hard to detect in SC approaches, and a relatively high number of MS2 spectra per protein is required for this purpose [79]. Finally, in shotgun proteomics usually *dynamic exclusion* is used to prevent highly abundant peptides from being sequenced over and over again, however, this leads to underestimation of these highly abundant proteins in the SC approach.
Another counting approach is based on the number of identified peptides per protein [80]. In this protein abundance index (PAI) approach, the results are normalized for the protein size, by calculating the ratio of experimentally observed peptides to the theoretically observable ones for each protein. The PAI method was later extended resulting in the 'exponentially modified PAI' (emPAI) approach ($emPAI = 10^{PAI-1}$), in which it was empirically found that the resulting score is directly proportional to the protein abundance and can hence even be used for estimating absolute protein amounts [81].
The above-mentioned methods based on counting MS2 spectra or peptides naturally

result in discrete numbers, and completely ignore a wealth of information contained within the measured peptide intensities. When low-resolution mass spectrometry was the standard, counting approaches yielded valuable results. However, today more promising intensity-based approaches have become feasible, yielding more accurate quantification results. Intensity-based LFQ approaches require high resolution both in the time and in the mass dimension, because the different peptide peaks have to be clearly resolved. Hence they benefit greatly from the recently developed nano-scale ultra high performance liquid chromatography (UHPLC) platforms and high-resolution mass analyzers like the Orbitrap.

Intensity-based LFQ approaches rely on the fact that the peak intensities of individual peptide signals are linearly correlated with the peptide concentration over a wide range of concentrations [73]. The first intensity-based approaches simply used the summed peak area of all peptides belonging to one particular protein [73]. A more sophisticated intensity-based LFQ approach is the MaxLFQ algorithm [74] available in the MaxQuant software environment [53]. In this approach, first all peptide features are detected in all LC-MS/MS runs to be compared. Then, the retention times of all runs are aligned to make them comparable at all and correct for small variations in chromatography. After that, *matching between runs* can be performed, an operation transferring identifications from one run where a peptide feature was identified to another run where the same peptide feature was also present, but not selected for fragmentation and hence not identified [82]. Matching between runs requires highly accurate masses and corrected retention times of the peptides, and resolves some of the stochastic nature of sequencing in shotgun approaches. This allows to extract the maximum amount of quantitative information available on peptide level, which is highly beneficial for following protein quantification. The raw intensities are then normalized, to correct for small differences introduced during the parallel sample handling. For this purpose, a certain number of proteins that are assumed to be unchanging between all samples is required. Finally, protein intensities are calculated from the peptide intensities by taking all available pairwise peptide ratios between all samples into account. After that, the resulting 'LFQ intensities' represent excellent approximations for the protein amounts observed in the different samples, and MaxLFQ has proven to be superior to spectral counting and summed intensity approaches [74].

Label-free quantification can also be used for absolute quantification. In principle, even spectral counting approaches and the emPAI method can be used for this purpose. A

more sophisticated method combining intensity-based and peptide counting approaches is intensity-based absolute quantification (iBAQ) [75]. In this approach, the protein intensity is normalized by the number of theoretically observable peptides, and scaled using spiked-in commercially available protein standards. The newest method for absolute LFQ is the so-called 'proteomic ruler' concept. In principle it is similar to the iBAQ method, but instead of spiked-in protein standards the data is scaled using histones whose signal is proportional to the DNA content and hence the number of cells in the sample [83].
In general, newly improved label-free quantification approaches provide a good accuracy despite the parallel sample handling, and present a viable alternative to all label-based approaches. This will have a large impact on MS-based proteomics in general, because quantitative data can now easily be achieved without large efforts, even in large-scale studies comparing hundreds to thousands of conditions or patients.

1.3 Applications of mass spectrometry-based proteomics

Mass spectrometry-based proteomics can be applied to a wide range of biological questions, the most obvious one being investigating whole proteomes. The first proteome ever that came close to being complete was that of budding yeast published in 2008, where the authors found evidence for 4399 proteins in haploid and diploid yeast [84]. Three years later, proteomes of a human cancer cell line were published, containing more than 10.000 proteins [15, 85]. Recently, two big drafts of the entire human proteome were published, acquired by combining data from many different human cell types and claiming to have evidence for up to 18.000 proteins [86, 87]. However, the notion of a 'complete' human proteome is inherently difficult. To detect all existing proteins with all their isoforms is nearly impossible, because some of them will only be expressed at very specific cases under very specific conditions and in very specific cell types.
Many scientific problems indeed benefit from investigating a specific sub-proteome instead of the whole proteome. This sub-proteome can consist of a specific cellular compartment. For example, for investigating nuclear processes often nuclear extracts are prepared. Another obvious application that requires the extraction of a sub-proteome is *interaction proteomics*. For investigating protein-protein interactions one wants to extract the part of the proteome that interacts with the corresponding protein of interest. Finally if one is interested in a particular posttranslational modification (PTM), extrac-

tion of the part of the proteome that bears this particular modification is of course highly beneficial.
Mass spectrometry-based proteomics has matured during the last years, and has now reached a stage where it can be applied to actual patient material. This field, called *clinical proteomics*, tries to assist in diagnosis and even treatment of various diseases. Although being one of the most challenging applications of mass-spectrometry-based proteomics, it also offers the most reward and promise for the future.
Three of the just described applications, namely interaction proteomics, PTM-related proteomics and clinical proteomics are part of this PhD work. Therefore they will now be described in more detail.

1.3.1 Investigating protein-protein interactions by mass spectrometry

The interaction of proteins with other proteins, but also with nucleic acids, lipids, metabolites, small molecules etc. is the basis of life at the molecular level. This section will mainly focus on the investigation of protein-protein interactions, however examples for the other types of interactions will also be shortly discussed at the end.
Proteins interact with each other to form sometimes small and defined, sometimes very large multiprotein complexes. Investigating these protein complexes and also their interconnections can provide meaningful information about biological processes and the functions of proteins inside the cell. Next to more indirect techniques such as phage display [88] and protein-fragment complementation assays such as the yeast two-hybrid (Y2H) approach [89–91], mass spectrometry has become the method of choice for analyzing protein complexes under near physiological conditions.
To investigate protein-protein interactions by mass spectrometry, the protein of interest first has to be immobilized. It is bound to a certain matrix, usually via an antibody either directed against the protein itself, or against a protein tag that has been fused to it. Subsequently, the bound protein is used as a *bait* to fish for *prey* interacting proteins by incubating it e.g. with cell lysate. After a washing step that removes some of the unspecific binders, the remaining bound proteins are released from the matrix and identified by LC-MS/MS. This workflow is commonly known as affinity purification mass spectrometry (AP-MS) [92].
When AP-MS was first developed, quantitative MS measurements were not yet available. Therefore the affinity purification workflow relied on dual affinity tags like the tandem

affinity purification (TAP) tag, which enabled two consecutive rounds of purification [93]. Combined with specific elution steps and stringent washing, this procedure yielded relatively clean protein complexes, which was pivotal as all proteins subsequently identified by MS were considered to be interactors. The TAP technique was applied to generate the first large-scale interaction datasets in the model organism budding yeast [94–96]. However, the TAP technique and similar procedures suffer from several issues. Firstly, due to the stringent nature of the two-step purification, weak or transient interactors are mostly lost in the process. Secondly, unspecific binders can still not completely be removed, hence a high false-positive rate has to be expected. To deal with these false positives, proteins appearing in empty control pulldowns or in more than a certain percentage of all pulldowns were often simply put on a 'contaminant blacklist'. This naturally decreases the true positive rates as many proteins are simply not considered, and is also not a suitable technique to efficiently identify unspecific binders.

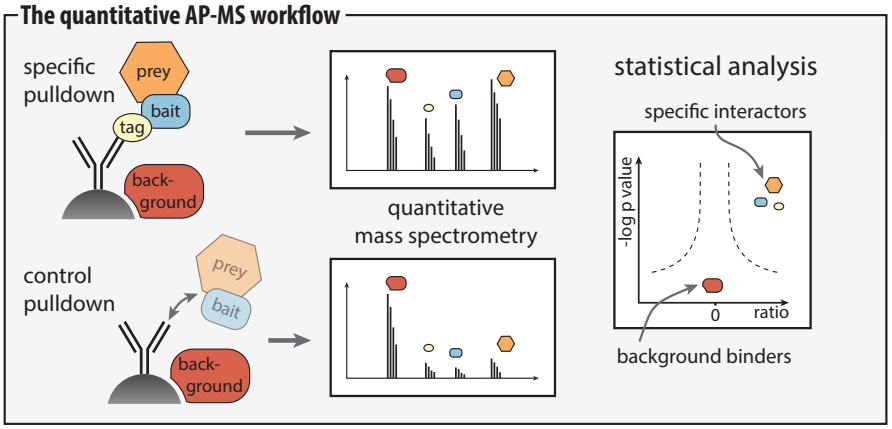


Figure 10: Investigating protein-protein interactions by quantitative AP-MS. In quantitative AP-MS strategies, a specific pulldown is compared with a control pulldown where the bait protein is not tagged. Through a washing step, the interactors of the bait are enriched in the specific pulldown. This enrichment is then reflected in the quantitative MS readout. Using statistical tests like the *t*-test, the enriched interactors can easily be identified as such and distinguished from the unchanging background binders centered around zero. Adapted from [19].

Many of the aforementioned issues could be improved or even completely solved by the introduction of quantitative mass spectrometry. By comparing quantitative amounts of proteins in a specific and a control pulldown, interactors can be distinguished from unspecific binders that are unchanging between the two pulldowns (see Figure 10). The fact that unspecific binders can now easily be identified as such and do not have to be

removed from the dataset has induced a reversal of trend in the workflow. Instead of purifying protein complexes as much as possible, fast and low stringent single-step purifications are employed (for more information see also Results section 2.1.1). This in turn minimizes the loss of weak and transient interactors.

A critical point in every interaction experiment is the expression level of the bait [97]. Ideally, one should use the endogenous protein as a bait, e.g. by using an antibody directed against the protein itself. This ensures correct localization of the bait and correct amounts of the bait compared to its interaction partners. Developing specific antibodies for every bait protein of interest is however hardly feasible in large-scale studies. Hence, in most cases tagged proteins are used, enabling the use of generic purification protocols for a large number of bait proteins. Another advantage of using tagged proteins is that a very simple control is at hand: the same strain/cell line, but with an untagged bait (see Figure 10, page 24). Next to the green fluorescent protein (GFP)-tag, which can additionally be used for protein localization, popular tags in larger scale AP-MS studies are the aforementioned TAP-tag and the FLAG-tag [98, 99]. In simple organisms like yeast, the tag can be directly introduced into the endogenous locus of the protein ensuring endogenous expression. In human cells this is much more complicated, therefore mostly alternative approaches are used. One strategy to achieve at least very close to endogenous expression is to use bacterial artificial chromosomes (BAC) as vectors. BACs can accommodate large pieces of DNA, even whole human genes with all their regulatory elements, are easily modified to contain the tag, and can be stably expressed in human cells. Hence they allow expression of e.g. a GFP-tagged version of the bait under endogenous control [100], as in the quantitative BAC-GFP interactomics (QUBIC) approach [101].

Quantification in pulldown experiments can be performed as explained before, using label-based or label-free approaches. However, to capture a reasonable large part of an interactome, a large number of pulldowns has to be performed and compared on a quantitative level. Hence label-free techniques, which have no limitations regarding the number of samples, are gaining strong momentum in interaction proteomics. In pulldowns, relatively large ratios are usually expected, which can easily be detected by label-free methods. Intensity-based LFQ, in particular, has been shown to perform as well as SILAC quantification for pulldowns [101, 102]. As LFQ approaches require a highly reproducible sample preparation workflow to produce accurate quantification results, they benefit from high-throughput parallel sample processing platforms. In the QUBIC

pipeline, for example, all pipetting steps were performed on a robotic platform, ensuring very high reproducibility of the pulldowns [101]. Similar LFQ workflows for investigating protein interactions in yeast will be explained in Results sections 2.1.1 and 2.1.2. Such fast sample preparation methods now leave the mass spectrometric measurement as the bottleneck of the interaction pipeline. However, with the newest generation of Orbitrap mass spectrometers and the advancing field of single-shot proteomics, much shorter measurement times are possible, as also described in Results section 2.1.2. Very soon it will be possible to obtain high coverage interactomes of many organisms in a quantitative manner, with manageable effort and in a relatively short time. These datasets will be highly valuable resources for biology and systems biology research.

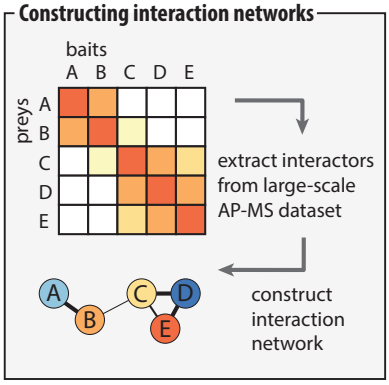


Figure 11: Constructing protein-protein interaction networks from quantitative AP-MS data The output of a quantitative AP-MS experiment is a large data matrix of all the different bait proteins vs. all the identified prey proteins. The quantitative data contained in the matrix have to be extracted with appropriate statistical methods to identify interacting proteins. From that information, interaction networks can be constructed that reveal the composition of protein complexes, their interconnectivity, and possibly also their topology. Adapted from [19].

Large-scale quantitative AP-MS experiments yield a large data matrix as schematically depicted in Figure 11, from which interacting proteins can be extracted and entire interaction networks can be generated. Such quantitative networks contain a wealth of information, for example about the interconnectivity of proteins. Some proteins in the network are 'interaction hubs', i.e. they have multiple connections to other proteins and take part in many different biological processes, while others interact only with one or few proteins hinting at a very specialized role. To some extent, is is also possible to determine the strength of the individual interactions: strong enrichment in the pulldown is indicative of a strong interaction, whereas mild enrichment is indicative of a weaker or more transient interaction (see also Results section 2.1.1). Furthermore, the stoichiometries within the complexes can be estimated by using absolute quantification methods like iBAQ [103]. Finally, if enough entry points for a complex are included in the dataset, it can be possible to determine alternative subcomplexes and complex topologies.

A different approach to decipher the complex topology is the use of chemical crosslinkers, that 'freeze' protein complexes in a certain moment in time (see Figure 12) [104–106]. The crosslinkers are bispecific and have a defined length determined by a spacer group, hence they can only link regions of proteins that are in a certain proximity to each other. The treated proteins are then digested as usual and the crosslinked peptides are identified by mass spectrometry, which determines spacial constraints that can hint at interaction surfaces within the protein complex. However, identifying crosslinked peptides is not trivial and the fragmentation spectra are inherently more complex, hence the technology needs more improvement before it can be applied in a generic way and in a large-scale manner. Crosslinking approaches using unspecific crosslinkers like formaldehyde can also be used to retain transient interactors, and therefore yield functionally relevant information that may be more difficult to obtain in standard AP-MS experiments [107].

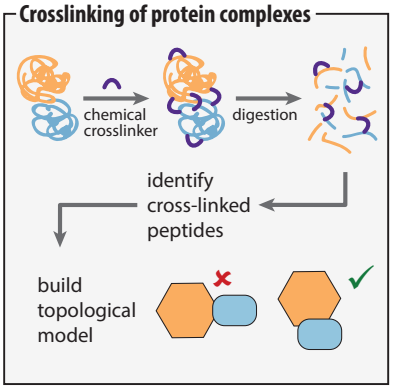


Figure 12: Investigating the interaction topology of protein complexes by chemical crosslinking and AP-MS. A bispecific chemical crosslinker targeting defined chemical groups is added to the cells to freeze complexes in their current state. The standard proteolytic digestion then results in crosslinked peptides that can be identified by MS. According to which peptides between two proteins are found to be crosslinked, a topological model can be built. Adapted from [19].

The previously described principles of AP-MS are of course not limited to investigating protein-protein interactions, but can be extended to any kind of bait that can be immobilized on a solid support. For example, peptides can be easily synthesized, linked to a matrix and used to screen for interactors. Even more interestingly, both a modified and unmodified version of the peptide can be synthesized, to screen for proteins specificly recognizing a certain PTM. This approach has for example been successfully applied to histone modifications and specific phosphorylation events [108–110]. Also DNA (e.g. in [111, 112]) and RNA (e.g. in [113, 114]) molecules can be used as baits to screen for protein interactors. Finally, in a field called *chemical proteomics*, small molecules are used as baits to screen for drug-binding proteins. Hence chemical proteomics can be used

to identify drug targets and potential off-targets leading to side effects [115], to decipher a drug's mechanism of action and even to assess binding characteristics. The inhibitor coupled to the solid support can also be unspecific e.g. targeting kinases in general [79]. In this approach the interaction with different kinase inhibitors is then tested via competition with the affinity matrix, as proteins binding to the free inhibitor will not bind to the matrix anymore, and hence be detected with lower abundance in the pulldown. A similar approach has been used to study the dissociation constants between proteins and kinase inhibitors, by adding different concentrations of free inhibitor [116]. More recently, it has also been applied to study histone deacetylase (HDAC) inhibitors and their selective targeting of different HDAC complexes [117].
In summary, MS-based proteomics has established itself as a valuable tool for investigating the interactions of proteins with all kinds of other molecules. Especially if proper MS quantification is applied, the acquired data contains a wealth of information that can help to identify previously unknown functions of proteins, or to understand biochemical processes both in a physiological and pathological state.

1.3.2 Investigating posttranslational modifications by mass spectrometry

Posttranslational modifications are important key players of cellular control. They allow the propagation of signals inside the cell so that it can react to a rapidly changing environment or to changes in the internal state. Classically, signaling was thought to occur via isolated 'pathways', i.e. linear cascades of different proteins propagating signals e.g. from the cell membrane to the nucleus. Nowadays it is acknowledged that in reality signaling pathways are extensively connected and in fact organized in incredibly complex *signaling networks* that integrate stimuli [118]. Therefore to understand such widespread networks, it is highly beneficial to analyze PTMs in a global and unbiased manner rather than looking at individual modified proteins.
PTMs can have various effects on the protein carrying the modification, i.e. change its structure, stability, activity, localization and interaction partners. The functional importance of PTMs has become evident in many cases where their deregulation has been linked to a disease [119]. Currently, about 300 different PTMs have been described to physiologically occur on proteins [120]. However, of this large number only very few PTMs are studied routinely and thoroughly. To date, the PTMs that have been targeted in most studies are phosphorylation, acetylation, glycosylation, ubiquitinylation, and

methylation (e.g. [121–129]).

Although many techniques exist to identify PTMs in small focused studies, high-resolution mass spectrometry is currently the main technique for detecting and quantifying protein modifications on a proteome-wide scale. Usually, standard bottom-up approaches are used for this purpose. The introduction of MS-based methods to the PTM field has in some cases multiplied the number of known sites 10 to 100-fold compared to traditional methods [130].

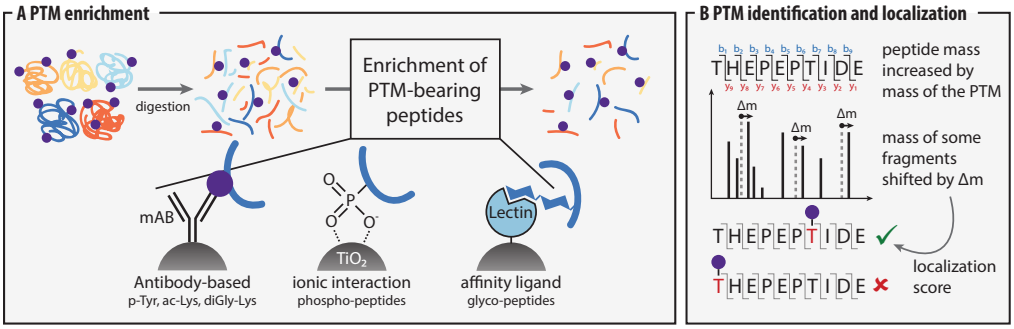


Figure 13: Investigating PTMs by LC-MS/MS. Adapted from [19]. **A** After digesting the proteins into peptides, the peptides carrying the modification of interest can be enriched using specific enrichment strategies. This in turn increases the chance for their comprehensive identification. **B** After adding the mass of the modification of interest and the potentially modified amino acid(s) to the database search, modified peptides can be easily identified. Both the mass of the entire peptide as determined in the full scan, as well as the mass of some fragments in the MS2 scan will be shifted by a certain mass difference ΔM. Because only some of the peptide fragments are shifted by the corresponding ΔM, the site of the modification can also be determined.

Despite numerous success stories, the bottom-up MS-based investigation of PTMs remains challenging for several reasons. (1) Usually only a fraction of a protein is modified, hence PTMs are often present in substoichiometric amounts making them hard to detect in complex mixtures. (2) Depending on the modification, the fragmentation spectra of modified peptides can be complicated and difficult to interpret. (3) If a modified peptide contains several potentially modified residues, it is often challenging or even impossible to determine the exact site. (4) Although *PTM crosstalk* is an important regulatory mechanism [131], multiple PTMs simultaneously occurring on the same protein can often not be detected as such in bottom-up approaches based on short tryptic peptides. (5) The database search space for peptide identification is drastically increased, especially if a PTM can be located on several amino acids and/or the search includes several PTMs at

a time. (6) Finally, quantification is based on single peptides, potentially leading to missing values and some inaccuracies. Many of the aforementioned issues can successfully be solved. To overcome the substoichiometric nature of many PTMs, prefractionation techniques and specific enrichment procedures, for example based on antibodies, ionic interactions and chromatography (see Figure 13A, page 29), have been introduced [132]. High-resolution MS combined with efficient fragmentation and solid data analysis software can in most cases identify PTMs with high confidence and even directly locate the site of the modification (see Figure 13B, page 29). PTM crosstalk can be more closely investigated using alternative proteases creating other and/or longer peptides than trypsin, or using top-down approaches. Quantification can strongly be improved by introducing several replicates per sample, normalizing for changes in the abundance of the corresponding proteins, and applying robust data analysis strategies.

Like for unmodified peptides, quantification of PTM sites can be performed either in a label-free format or using the standard metabolic and chemical labeling techniques. Regarding chemical labeling, however, it needs to be considered that many chemical labeling agents target lysines, which might interfere both with tryptic cleavage and the analysis of PTMs localized on lysine, like acetylation, ubiquitination and methylation. Following the general trend in quantitative proteomics, label-free techniques are gaining momentum also for the analysis of PTMs. Recently a label-free ultra-deep phosphoproteome of a human cancer cell line covering more than 50.000 distinct phosphopeptides was published [133].

After the detection of PTMs on a proteome-wide level has become feasible, the next logical step is to monitor the dynamics of signaling networks upon stimulation, perturbation or under various growth or stress conditions. Highly interesting insights have for example already been obtained by performing time-course experiments in response to certain stimuli like epidermal growth factor (EGF) and transforming growth factor β (TGF-β) [134, 135].

Another intriguing possibility is to determine PTM *site occupancies*, i.e. to determine the fraction of a protein that is modified. If the dectected occupany is high, this can hint at a functional site, whereas low occupancy can hint at non-specific and hence less functionally important origin [136]. Site occupancies can be determined whenever a stimulus is applied, then only three pieces of quantitative information are required: the change in abundance of the modified peptide, the change in abundance of the corresponding unmodified counterpart, and the change in protein abundance [137]. Phosphorylation site

occupancies have been determined on a proteome-wide scale using SILAC and recently also LFQ technologies [133, 137].
High quality quantitative data with a high coverage of sites can be the basis for analyzing further characteristics of a certain PTM. From the sequence around the modification site, specific motifs can be determined, that in turn can point to the modifying enzyme. By integrating other orthogonal data, like interaction, localization, or structural data, a deeper understanding of the PTM under investigation can be obtained.
In summary, MS-based PTM analysis is increasingly revealing an unexpectedly large number of naturally occurring protein modifications. With the development of more enrichment techniques, additional PTMs will become accessible for proteome-wide investigation. By providing quantitative information about PTM sites, MS-based proteomics can now be used to decipher signaling processes both in the healthy and in the diseased state.

1.3.3 Mass spectrometry-based clinical proteomics

MS-based proteomics technologies have improved tremendously over the last years, and increasingly powerful sample preparation techniques, instrumentation and data analysis software are available today. Hence it is now becoming feasible to apply mass spectrometry to adress questions of clinical and medical relevance. This field of research called *clinical proteomics* has a variety of goals, ranging from better characterization of pathological processes on a molecular level to diagnosis, monitoring and optimized treatment of diseases.
The most popular application of MS-based clinical proteomics is the search for protein biomarkers that can pinpoint the presence or reflect the stage of a particular disease, or can be used to classify patients into treatment-relevant subgroups. Biomarkers can be specific cells, (mutated) genes, proteins, lipids, metabolites, or other small molecules, many of which are routinely monitored in standard blood tests. Nevertheless, it is the protein domain that is ultimately affected in a disease, therefore finding protein biomarkers is particularly promising. Several protein biomarkers are already routinely used in the clinic, like C-reactive protein that pinpoints the presence of general inflammation [138], troponin I that indicates a myocardial infarction [139], and prostate-specific antigen that is a marker for prostate cancer [140]. The search for protein biomarkers is facing three major challenges. The first one is the extreme complexity and dynamic range of

the biological material that is the source for biomarker search. The second is the low abundance of many disease-relevant biomarkers, making their discovery reminiscent of finding a needle in the haystack. Finally, the large variability between human individuals and also individual disease characteristics further complicates the situation [141].
Blood is the ultimate source for biomarker discovery. Human blood plasma does not only contain the classical plasma proteins, but also so-called tissue leakage proteins. As blood is in contact with every single tissue in the body, it contains small amounts of proteins from all of these tissues, representing both physiological and pathological processes. Therefore blood plasma most likely represents the most comprehensive human proteome [142]. Blood is also the most sampled biofluid, taken from patients at almost every routine check-up, and a vast infrastructure exists for its storage and analysis. Although the collection of blood in an invasive procedure, it is still very easily accessible compared to e.g. tissue biopsies. Unfortunately, although being such a promising source for protein biomarkers, blood plasma is also the most challenging material for proteomic analysis. The dynamic range in plasma spans an enormous ten to eleven orders of magnitude from the highest to the lowest detected protein so far [143]. It is dominated by very few proteins like albumin present at extremely high concentrations, covering up the low abundant tissue leakage proteins of interest and the even lower abundant cytokines (see Figure 14 A, page 33). This dynamic range is far higher than the dynamic range proteomic technologies can capture; even up-to-date mass spectrometers only reach around six orders of magnitude at best [144].
The pipeline for the development of a new protein biomarker can be divided in several phases (see Figure 14 B, page 33). Particularly in the later phases of biomarker verification, validation and assay development, large numbers of patient-derived samples need to be processed to deal with the natural human and disease variability. At these later stages, so-called immunoaffinity-based MS (IA-MS) approaches are and routinely used. IA-MS methods enrich the protein of interest using antibodies, followed by targeted MS analysis, and have already been successfully applied to quantify various protein biomarkers [145–150]. However the focus of IA-MS and other classic ligand binding assays like the enzyme-linked immunosorbent assay (ELISA) onto one or several candidates makes them unsuitable for phase one, the unbiased discovery of new protein biomarkers. This currently leaves classic data-dependent LC-MS/MS approaches as the main technique for this purpose [141].

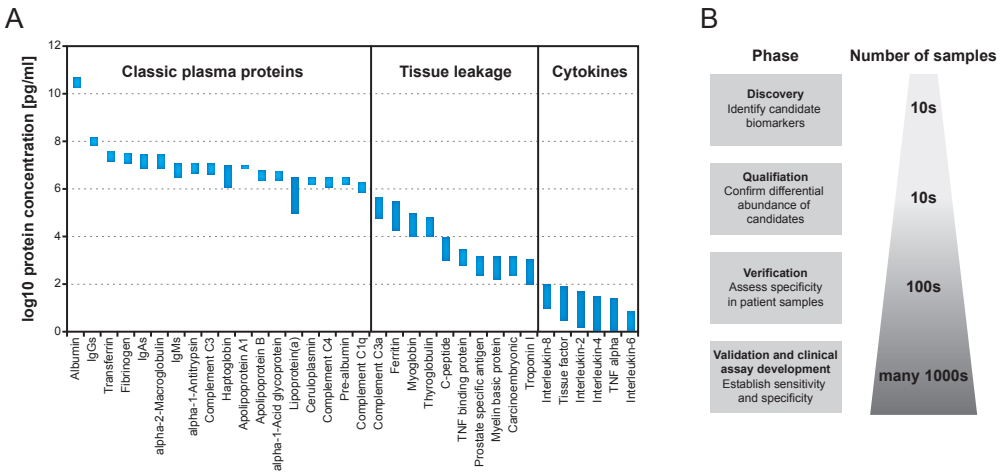


Figure 14: Clinical plasma proteomics. (A) The dynamic range of plasma demonstrated on 34 exemplary plasma proteins. Adapted from [143]. (B) The phases of biomarker discovery and the number of samples required. Adapted from [141].

The search for protein biomarkers still faces many problems, some of them already mentioned, and currently only about one new protein biomarker is introduced per year [141]. Several ideas how to improve this fact have been proposed. Of course methods to more comprehensively cover the blood plasma proteome are highly desirable in this respect. Higher proteome coverage can be achieved by extensive fractionation, however, this in turn multiplies the number of samples, and hence reduces the throughput drastically. Another method to achieve better coverage of lower abundant plasma proteins is to deplete the top abundant ones, reducing the dynamic range by one to two orders of magnitude.

A different approach is to actually move away from the plasma at least for the discovery phase, and use tissues or other biofluids. Often 'proximal fluids', i.e. body fluids that are located more closely to the actual site of the disease, are highly attractive for biomarker discovery [141]. Examples for such proximal fluids are urine for diabetes, kidney disease, bladder cancer etc., cerebrospinal fluid for diseases affecting the brain, bronchoalveolar lavage fluid for lung diseases, and so on. In many cases, the fold difference between the diseased state and the healthy state is higher in proximal fluids, making them a good source for protein biomarker discovery [151, 152]. After identifying a biomarker in a proximal fluid, the final clinical test can in many cases still be developed for the more easily accessible blood.

Genetic and environmental variations introduce noise between samples, which com-

plicates the discovery of new biomarkers. One way to reduce this noise is to use model systems for the discovery phase. Genetically homogenous animals, but also cell lines, can be a good choice for a disease model in this regard.
Finally, ultra high performance mass spectrometers with high resolution, mass accuracy and scan speed can greatly enhance the results in clinical experiments [141]. Therefore the latest generation of Orbitrap instruments (see chapter 1.2.2) is particularly suited for clinical proteomics.
So far, the success of discovery proteomics for finding new biomarker candidates has fallen short of expectations. However with the improved technologies and instruments that are now available, this will likely change in the future. Furthermore, MS-based proteomics is already successfully applied in many functional clinical studies evaluating altered protein-protein interactions or posttranslational modifications in various disease contexts. In many cases, the protein level itself is not affected in a disease, but rather the level of a particular PTM on that protein. The role of PTMs in the pathogenesis of many diseases has recently become more and more acknowledged [136]. In some types of cancer, for example, mutated kinases with higher/lower activity lead to altered phosphorylation levels on their target proteins, which in turn alters their activity and ultimately causes the disease. A well-known example is the constitutively active tyrosine kinase Bcr-Abl, which is created by a translocation between chromosomes 9 and 22 and leads to leukemia [153]. Other PTMs like acetylation, ubiquitination, SUMOylation, glycosylation, glycation and many more can also have a strong impact on the development of diseases [154]. Protein glycation, a modification relevant in diabetes, will be the topic of Results section 2.2.2 in this thesis. MS-based proteomics is currently the only available tool for investigating large-scale variations at the PTM level, both in the physiological and pathological state.
In summary, mass spectrometry holds great promise for clinical applications and will in the future contribute to better diagnose, classify and monitor patients, and to provide optimized treatment strategies for individual patient needs in the context of *personalized medicine*.

1.4 Aims of the thesis

In this thesis, I have developed and applied state-of-the-art mass spectrometry-based proteomics technologies for investigating protein-protein interactions and posttranslational modifications.

I started out with developing a method for investigating protein-protein interactions in budding yeast. The basis for this first and main project was the quantitative BAC-GFP interactomics (QUBIC) [101] methodology from our group, which I transferred from the human to the yeast system. Budding yeast is an attractive model organism for human biology and offers several appealing advantages for investigating interactions. Being a relatively simple organism, yeast can easily be genetically modified, hence endogenous bait expression is possible by tagging the bait proteins directly in their genetic loci. Other genetic alterations are also possible, like knocking out individual complex members to determine complex topologies.
Developing the interaction pipeline for yeast required establishment of the culture conditions and the input amounts, as well as the sample preparation methodology. The immunoprecipitation step had already been optimized for GFP-tagged human proteins in the QUBIC project, and was found to be equally suited to enrich GFP-tagged yeast proteins and their interaction partners. After the wetlab workflow was established, I found that every single pulldown contained almost 2000 proteins, representing about half of the entire yeast proteome. This was a striking discovery, as such a large number of background binders had not even been observed in the human pulldowns. Therefore, the next goal was to develop dedicated data analysis techniques, to detect the few true interactors among the majority of unspecific binders. Since the data was acquired with label-free quantification, distinguishing enriched interactors from unchanging background proved to be relatively straightforward. However, we found that we can make additional use of the large background to improve data quality and obtain high confidence interaction partners. Together with Marco Hein, whose main project was the application of the QUBIC pipeline for mapping the human interactome, I developed several strategies to extract the maximum amount of information contained within the background.
The resulting yeast interaction pipeline stands in stark contrast to classical pulldown experiments where unspecific binders are removed by stringent washing and multiple purification steps, which however leads to the loss of weaker or more transient interaction partners. Since our pulldowns are hardly 'purifications' anymore, I termed the new

methodology 'affinity enrichment mass spectrometry' (AE-MS) to distinguish it from the classical 'affinity purification mass spectrometry' (AP-MS) approaches.

After I had established the pipeline for investigating protein-protein interactions in yeast, my work was continued within the group. Since the yeast interactome in exponentially growing yeast is already quite well understood, the next objective is to investigate this interactome under various conditions and perturbations. To do so, new methods to detect true interactions under close to physiological conditions in a fast and reproducible manner are required. Hence, the aim of this second project was to transform the AE-MS interaction pipeline into a high-throughput format. We first wanted to decrease the time spent on sample preparation and adapted most steps accordingly to a 96-well format. With the drastically reduced sample preparation time, the measurement time of the samples in the mass spectrometer (two hours in the first yeast interaction pipeline) became the major obstacle to increase sample throughput. We hence explored much shorter gradients, and even further reduced the analysis time by implementing a double-barrel column setup driving two analytical columns in parallel. Finally, we were able to measure 96 pulldowns in only about one day with this new high-throughput methodology, and still achieved remarkable coverage for the targeted complexes.

On the side, I applied my expertise of protein-protein interactions in a fruitful collaboration project on human histone variants. Canonical human histones can be replaced by several variants, and this occurs at very specific places in the genome and leads to various functional differences. One way to determine how these variants are targeted to specific chromatin locations and how they exert their differential function, is to investigate which proteins they interact with. Hence the aim of this third project was to identify differential interaction partners of specific H2A variants as compared to the canonical histone H2A. We first performed pulldown experiments of mononucleosomes containing GFP-tagged histone variants in HeLa cells, and identified some highly interesting interaction partners that are now being followed-up by my collaboration partners (ongoing project). Additionally, we also performed similar pulldowns of H2A variants in melanoma cells in collaboration with a third group (presented in this thesis). In this project, our interaction analysis was able to identify the protein Brd2 as a specific interaction partner of H2A.Z containing nucleosomes. After various additional experiments to validate and characterize this protein, Brd2 emerged as a potential target for the therapy of malignant melanoma.

Next to investigating protein-protein interactions, I started to become interested in PTMs, with this interest being triggered by another intriguing collaboration project. The question in this fourth project was how elongation factor P (EF-P), a protein required for efficient translation, is activated in a certain branch of bacteria. While for many bacteria the activation of EF-P is known to occur by modification of a specific lysine residue, other bacteria display an arginine residue in the homologous position. Since arginine is not one of the commonly modified amino acids, the question whether it is modified at all, and if so by what entity, was intriguing. Using phylogenetic tree analysis, my collaboration partner in this project had already identified the potential modifying enzyme. However, what kind of modification this protein might transfer to EF-P in order to activate it, remained completely elusive. Hence, the aim of this project was to investigate whether the potential modifying enzyme actually modifies EF-P, whether modification really occurs on the arginine in question or elsewhere, and finally what the activating modification actually is. Using MS-based proteomics and the 'dependent peptide' search technique, I successfully identified the previously unknown modification that activates EF-P. This project was the first high profile application of the dependent peptide search in our group.

In the fifth and last project, I focused on a clinical question and investigated protein glycation, a posttranslational modification highly relevant in the diagnosis, monitoring and pathology of diabetes. While this modification is extensively studied on a few specific proteins like hemoglobin, no real comprehensive dataset on glycated proteins in blood plasma, the most affected biofluid, currently exists. Hence, the aim of this project was to develop a method to identify glycated proteins from plasma with high confidence using mass spectrometry. In a first step, I wanted to evaluate the specific behavior of this particular modification during HCD fragmentation and the feasibility to study protein glycation on our specific MS instrumentation both *in vitro* on model proteins and *in vivo* in actual plasma. In the future, we will apply the acquired knowledge to investigate glycation directly in patient samples.

2 Results

2.1 Development and application of mass spectrometry-based methods for investigating protein-protein interactions in yeast and human

A deeper understanding of protein-protein interactions can help to answer key questions in biology. Although methods to identify interaction partners of proteins have been available for some time, mapping the interactome is by far not completed. Especially weaker or transient interaction partners often escape detection, hence methods to preserve such proteins are urgently required. In my first project focusing on interactions, I took one step in this direction by developing an efficient yeast pulldown pipeline with very low purification stringency. This led to a background of unprecedented size, and necessitated the development of specialized data analysis techniques described in the first publication of this section.

In the second publication, we took the now established yeast interaction pipeline to the next level, by drastically increasing the sample throughput. To do so, we implemented improvements in the sample preparation workflow as well as in the LC-MS/MS measurement techniques and the data analysis.

While large-scale quantitative interaction networks and methods to acquire such interactomes are highly valuable resources to the scientific community, biologists often focus on very few proteins of their interest. In the third project, I showed that our methodologies are equally well suited for small-scale studies. In this publication, I successfully applied our label-free pulldown technology to find interaction partners of human histone H2A variants in the context of malignant melanoma.

2.1.1 Affinity enrichment mass spectrometry (AE-MS) as a novel concept for investigating protein-protein interactions

Keilhauer, E. C., Hein, M. Y. & Mann, M.

Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS).
Molecular and Cellular Proteomics 2015 Jan; 14(1); 120-135.

The quantitative BAC-GFP interactomics (QUBIC) pipeline developed in our group already represented a powerful workflow for investigating protein-protein interactions in a mammalian cell culture system under near physiological conditions [101]. In the first project of my PhD, I wanted to transfer this pipeline to the yeast system. In contrast to the BAC cell lines, that provide very close to endogenous expression of bait proteins, in yeast genes can be directly tagged in their genetic locus, thereby providing true endogenous expression. Conveniently, an endogenously tagged yeast library employing the GFP-tag had already been created by the Weissman group, initially for protein localization studies, and could be used for my experiments [155]. Since the QUBIC pipeline is based on the GFP-tag, parts of the workflow could be reused for the yeast pulldowns. However, a different upfront lysis protocol based on mechanical beadbeating had to be developed and optimized. Furthermore, I constructed a dedicated control strain with the same genetic background as the strains of the GFP-library, thereby obtaining an optimal control for my interaction experiments.
Surprisingly, I found that the yeast pulldowns behaved differently from the human pulldowns, in that they produced an even larger background (ca. 1.800 protein as opposed to ca. 800 proteins). Hence, dedicated data analysis methods were required to extract true interactors. Together with Marco Hein, who was at the same time applying the QUBIC pipeline for mapping the human interactome, I developed a powerful data analysis pipeline for this purpose. Although the background in the yeast and human system was somewhat different, we found the basic strategies to pull out the proteins of interest to be universally applicable. Finally, I showed that the large number of unspecific binders detected in our pulldowns does not present a hindrance to data analysis, but on the contrary can be leveraged in a very efficient way. Since the pulldowns are highly similar to each other, we found that a dedicated control strain is actually not necessary to interpret the results, but that pulldowns can simply be compared against each other instead. We

also propose a way to efficiently group candidate proteins in 'weak' and 'strong' interactors based on their reproducible enrichment in the pulldown and their intensity profile across all runs.
Since such low-stringency single-step pulldowns do not present real 'purifications' anymore, I termed the novel method *affinity enrichment mass spectrometry (AE-MS)*, to clearly distinguish it from the classic AP-MS protocols mostly based on stringent TAP-tag purification technology. I successfully evaluated this novel concept on a variety of well-known yeast complexes from various cellular compartments, and achieved unprecedented coverage from single pulldowns for many of them.

Research

⌘ *Author's Choice*

Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS)*[S]

Eva C. Keilhauer‡, Marco Y. Hein‡, and Matthias Mann‡§

Protein–protein interactions are fundamental to the understanding of biological processes. Affinity purification coupled to mass spectrometry (AP-MS) is one of the most promising methods for their investigation. Previously, complexes were purified as much as possible, frequently followed by identification of individual gel bands. However, todays mass spectrometers are highly sensitive, and powerful quantitative proteomics strategies are available to distinguish true interactors from background binders. Here we describe a high performance affinity enrichment-mass spectrometry method for investigating protein–protein interactions, in which no attempt at purifying complexes to homogeneity is made. Instead, we developed analysis methods that take advantage of specific enrichment of interactors in the context of a large amount of unspecific background binders. We perform single-step affinity enrichment of endogenously expressed GFP-tagged proteins and their interactors in budding yeast, followed by single-run, intensity-based label-free quantitative LC-MS/MS analysis. Each pull-down contains around 2000 background binders, which are reinterpreted from troubling contaminants to crucial elements in a novel data analysis strategy. First the background serves for accurate normalization. Second, interacting proteins are not identified by comparison to a single untagged control strain, but instead to the other tagged strains. Third, potential interactors are further validated by their intensity profiles across all samples. We demonstrate the power of our AE-MS method using several well-known and challenging yeast complexes of various abundances. AE-MS is not only highly efficient and robust, but also cost effective, broadly applicable, and can be performed in any laboratory with access to high-resolution mass spectrometers. ***Molecular & Cellular Proteomics 14: 10.1074/mcp.M114.041012, 1–16, 2015.***

Protein–protein interactions are key to protein-mediated biological processes and influence all aspects of life. Therefore, considerable efforts have been dedicated to the mapping of protein–protein interactions. A classical experimental approach consists of co-immunoprecipitation of protein complexes combined with SDS-PAGE followed by Western blotting to identify complex members. More recently, high-throughput techniques have been introduced; among these affinity purification-mass spectrometry (AP-MS)[1] (1–3) and the yeast two-hybrid (Y2H) approach (4–6) are the most prominent. AP-MS, in particular, has great potential for detecting functional interactions under near-physiological conditions, and has already been employed for interactome mapping in several organisms (7–15). Various AP-MS approaches have evolved over time, that differ in expression, tagging, and affinity purification of the bait protein; fractionation, LC-MS measurement, and quantification of the sample; and in data analysis. Recent progress in the AP-MS field has been driven by two factors: A new generation of mass spectrometers (16) providing higher sequencing speed, sensitivity, and mass accuracy, and the development of quantitative MS strategies.

In the early days of AP-MS, tagged bait proteins were mostly overexpressed, enhancing their recovery in the pull-down. However, overexpression comes at the cost of obscuring the true situation in the cell, potentially leading to the detection of false interactions (17). Today, increased MS instrument power helps in the detection of bait proteins and interactors expressed at endogenous levels, augmenting the chances to detect functional interactions. In some simple organisms like yeast, genes of interest can directly be tagged in their genetic loci and expressed under their native pro-

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

Received, May 7, 2014 and in revised form, October 17, 2014
Published, MCP Papers in Press, November 2, 2014, DOI 10.1074/mcp.M114.041012
⌘ *Author's Choice*—Final version full access.
Author contributions: E.C.K. and M.M. designed research; E.C.K. performed research; M.Y.H. contributed new reagents or analytic tools; E.C.K., M.Y.H., and M.M. analyzed data; E.C.K. and M.M. wrote the paper.

[1] The abbreviations used are: AP-MS, Affinity purification mass spectrometry; AE-MS, Affinity enrichment mass spectrometry; GFP, Green fluorescent protein; (Co-)IP, (Co-) Immunoprecipitation; Y2H, Yeast two-hybrid; BAC, Bacterial artificial chromosome; QUBIC, Quantitative BAC green fluorescent protein interactomics; TAP, Tandem affinity purification; LFQ, Label-free quantification; MaxLFQ, MaxQuant Label-free quantification; CAA, Chloroacetamide; ES, Experimental series; FDR, False discovery rate; SC, Synthetic complete; YPD, Yeast extract peptone dextrose; BSCG, Bait specific control group; NPC, Nuclear pore complex; SPB, Spindle pole body.

moter. In higher organisms, tagging proteins in their endogenous locus is more challenging, but also for mammalian cells, methods for close to endogenous expression are available. For instance, in controlled inducible expression systems, the concentration of the tagged bait protein can be titrated to close to endogenous levels (18). A very powerful approach is BAC transgenomics (19), as used in our QUBIC protocol (20), where a bacterial artificial chromosome (BAC) containing a tagged version of the gene of interest including all regulatory sequences and the natural promoter is stably transfected into a host cell line.

The affinity purification step has also been subject to substantial changes over time. Previously, AP has been combined with nonquantitative MS as the readout, meaning all proteins identified by MS were considered potential interactors. Therefore, to reduce co-purifying "contaminants," stringent two-step AP protocols using dual affinity tags like the TAP-tag (21) had to be employed. However, such stringent and multistep protocols can result in the loss of weak or transient interactors (3), whereas laborious and partially subjective filtering still has to be applied to clean up the list of identified proteins. The introduction of quantitative mass spectrometry (22–25) to the interactomics field about ten years ago was a paradigm shift, as it offered a proper way of dealing with unspecific binding and true interactors could be directly distinguished from background binders (26, 27). Importantly, quantification enables the detection of true interactors even under low-stringent conditions (28). In turn, this allowed the return to single-step AP protocols, which are milder and faster, and hence more suitable for detecting weak and transient interactors.

Despite these advances, nonquantitative methods—often in combination with the TAP-tagging approach—are still popular and widely used, presumably because of reagent expenses and labeling protocols used in label-based approaches. However, there are ways to determine relative protein abundances in a label-free format. A simple, semi-quantitative label-free way to estimate protein abundance is spectral counting (29). Another relative label-free quantification strategy is based on peptide intensities (30). In recent years high resolution MS has become much more widely accessible and there has been great progress in intensity-based label-free quantification (LFQ) approaches. Together with development of sophisticated LFQ algorithms, this has boosted obtainable accuracy. Intensity-based LFQ now offers a viable and cost-effective alternative to label-based methods in most applications (31). The potential of intensity-based LFQ approaches as tools for investigating protein–protein interactions has already been demonstrated by us (20, 32, 33) and others (34, 35). We have further refined intensity-based LFQ in the context of the MaxQuant framework (36) using sophisticated normalization algorithms, achieving excellent accuracy and robustness of the measured "MaxLFQ" intensities (37).

Another important advance in AP-MS, again enabled by increased MS instrument power, was the development of single-shot LC-MS methods with comprehensive coverage. Instead of extensive fractionation, which was previously needed to reduce sample complexity, nowadays even entire model proteomes can be measured in single LC-MS runs (38). The protein mixture resulting from pull-downs is naturally of lower complexity compared with the entire proteome. Therefore, modern MS obviates the need for gel-based (or other) fractionation and samples can be analyzed in single runs. Apart from avoiding selection of gel bands by visual examination, this has many advantages, including decreased sample preparation and measurement time, increased sensitivity, and higher quantitative accuracy in a label-free format.

In this work, we build on many of the recent advances in the field to establish a state of the art LFQ AE-MS method. Based on our previous QUBIC pipeline (20), we developed an approach for investigating protein–protein interactions, which we exemplify in *Saccharomyces cerevisiae.* We extended the data analysis pipeline to extract the wealth of information contained in the LFQ data, by establishing a novel concept that specifically makes use of the signature of background binders instead of eliminating them from the data set. The large amount of unspecific binders detected in our experiments rendered the use of a classic untagged control strain unnecessary and enabled comparing to a control group consisting of many unrelated pull-downs instead. Our protocol is generic, practical, and fast, uses low input amounts, and identifies interactors with high confidence. We propose that single-step pull-down experiments, especially when coupled to high-sensitivity MS, should now be regarded as *affinity enrichment* rather than affinity purification methods.

EXPERIMENTAL PROCEDURES

Yeast Strains—For all experiments GFP-tagged yeast strains originating from the Yeast-GFP Clone Collection were used, a library with 4156 GFP-tagged proteins representing about 63% of *S. cerevisiae* open reading frames (39). The haploid parental strain of this library, BY4741 (ATCC 201388), served as an initial control strain and to construct the strain pHis3-GFP-HIS3$_k$MX6 (short name pHis3-GFP). To do so, we used the His3 locus in BY4741, which is nonfunctional because of a deletion of several amino acids in the middle of the coding sequence. We amplified a cassette containing a GFP gene without start codon and a His3 gene of *Saccharomyces kluyveri* under control of the TEF promoter and terminator out of the vector pFA6a-GFP(S65T)-HIS3$_k$MX6. This cassette was integrated into the His3 locus of BY4741 directly after the original His3 promoter and start codon by homologous recombination, replacing the rest of the nonfunctional His3 sequence. As a result, our pHis3-GFP strain is able to synthesize histidine and expresses moderate amounts of cytosolic GFP just as the tagged library strains.

Culture of Yeast Strains and anti-GFP Immunoprecipitation—Tagged yeast strains, the parental strain BY4741 and the control strain pHis3-GFP were first grown on plates (YDP plates for BY4741, SC-His plates for all other strains) and then in YPD liquid medium at standard culture conditions. Cell growth was regularly examined by measuring $OD_{600\ nm}$. Yeast cells were grown until they reached an $OD_{600\ nm}$ of around 1, followed by harvesting culture volumes

equaling 50 ODs. For biochemical triplicates (experimental series 1 (ES1)), three times 50 ODs were harvested out of the same culture and from then on processed separately. For biological quadruplicates (experimental series 2 (ES2)), four different colonies were picked on different days and processed separately from the beginning. Yeast cell pellets were dissolved in 1.5 ml lysis buffer (150 mM NaCl, 50 mM Tris HCl pH 7.5, 1 mM $MgCl_2$, 5% glycerol, 1% IGEPAL CA-630 (SIGMA-ALDRICH GmbH, Taufkirchen, Germany), Complete® protease inhibitors (Roche Diagnostics Deutschland GmbH, Mannheim, Germany), and 1% benzonase (Merck KGgA, Darmstadt, Germany)), transferred into FastPrep® tubes (MP Biomedicals GmbH, Eschwege, Germany) containing 1 mm silica spheres (lysing matrix C, MP Biomedicals), frozen in liquid nitrogen and stored at −80 °C until lysis. The frozen samples were thawed and then lysed in a FastPrep24® instrument (MP Biomedicals) for 6 × 1 min at maximum speed. Lysates were cleared by a 10 min centrifugation step at 4 °C and 4000 × *g*; and 800 μl of the clear lysates were transferred into a deep-well plate for immunoprecipitation. IP of yeast protein complexes was essentially performed as described before for a mammalian cell culture system (20). IPs were performed on a Freedom EVO® 200 robot (Tecan Deutschland GmbH, Crailsheim, Germany) equipped with a MultiMACS™ M96 separation unit (Miltenyi Biotec GmbH, Bergisch Gladbach, Germany) that contains a strong permanent magnet. (Miltenyi Biotec also supplies equipment for performing the same pull-downs in a manual fashion.) The basic steps of the IP protocol are as follows: First the lysates are mixed with 50 μl magnetic μMACS Anti-GFP MicroBeads (Miltenyi Biotec) and incubated for 15 min at 4 °C. Because of the favorable kinetics of the microbeads, tagged proteins are efficiently captured in only 15 min (40). Then the Multi-96 separation columns are equilibrated with 250 μl equilibration buffer (same as lysis buffer). After that, the lysates are added to the columns with the magnet turned on, retaining the magnetic MicroBeads on the column. Once all the liquid has passed through the columns, they are first washed with 3 × 800 μl ice cold wash buffer I (0.05% IGEPAL CA-630, 150 mM NaCl, 50 mM Tris HCl pH 7.5, and 5% glycerol), then with 2 × 500 μl of wash buffer II (150 mM NaCl, 50 mM Tris HCl pH 7.5, and 5% glycerol). Afterward 25 μl of elution buffer I (5 ng/μl trypsin, 2 M Urea, 50 mM Tris HCl pH 7.5, and 1 mM DTT) are added and the columns are incubated for 30 min at room temperature. In this "in-column digest," the proteins are partially digested to allow elution from the columns, and reduced by DTT. Subsequently the resulting peptides are eluted and alkylated with 2 × 50 μl elution buffer II (2 M Urea, 50 mM Tris HCl pH 7.5, and 5 mM CAA), and collected in a 96-well plate.

The plate was incubated at room temperature overnight to ensure a complete tryptic digest. The next morning the digest was stopped by addition of 1 μl Trifluoroacetic acid (TFA) per well. The acidified peptides were loaded on StageTips (self-made pipette tips containing two layers of C_{18}) to desalt and purify them according to the standard protocol (41). Every sample was divided onto two StageTips to give one "working" StageTip and one "backup" StageTip. The StageTips were stored at 4 °C until the day of LC-MS/MS measurement.

LC-MS/MS Measurement—Samples were eluted from StageTips with 2 × 20 μl buffer B (80% ACN and 0.5% acetic acid). The organic solvent was removed in a SpeedVac concentrator for 20 min, then the remaining 4 μl of peptide mixture were acidified with 1 μl of buffer A*(2% ACN and 0.1% TFA) resulting in 5 μl final sample size. 2 μl of each sample were analyzed by nanoflow liquid chromatography on an EASY-nLC system (Thermo Fisher Scientific, Bremen, Germany) that was on-line coupled to an LTQ Orbitrap classic (Thermo Fisher Scientific) through a nanoelectrospray ion source (Thermo Fisher Scientific). A 15 cm column with 75 μm inner diameter was used for the chromatography, in-house packed with 3 μm reversed-phase silica beads (ReproSil-Pur C_{18}-AQ, Dr. Maisch GmbH, Germany). Peptides were separated and directly electrosprayed into the mass spectrometer using a linear gradient from 5.6% to 25.6% acetonitrile in 0.5% acetic acid over 100 min at a constant flow of 250 nl/min. The linear gradient was followed by a washout with up to 76% ACN to clean the column for the next run. The overall gradient length was 134 min. The LTQ Orbitrap was operated in a data-dependent mode, switching automatically between one full-scan and subsequent MS/MS scans of the five most abundant peaks (Top5 method). The instrument was controlled using Tune Plus 2.0 and Xcalibur 2.0. Full-scans (*m/z* 300–1650) were acquired in the Orbitrap analyzer with a resolution of 60,000 at 400 *m/z*. The five most intense ions were sequentially isolated with a target value of 1000 ions and an isolation width of 2 m/z and fragmented using CID in the linear ion trap with a normalized collision energy of 40. The activation Q was set to 0.25, the activation time to 30 ms. Maximum ion accumulation times were set to 500 ms for full scans and 1000 ms for MS/MS scans. Dynamic exclusion was enabled; with an exclusion list size of 500 and an exclusion duration of 180 s. Standard MS parameters were set as follows: 2.2 kV spray voltage; no sheath and auxiliary gas; 200 °C heated capillary temperature and 110 V tube lens voltage.

Raw Data Processing—All raw files were analyzed together using the in-house built software MaxQuant (36) (version 1.4.0.6). The derived peak list was searched with the built-in Andromeda search engine (42) against the reference yeast proteome downloaded from Uniprot (http://www.uniprot.org/) on 03–20-2013 (6651 sequences) and a file containing 247 frequently observed contaminants such as human keratins, bovine serum proteins, and proteases. Strict trypsin specificity was required with cleavage C-terminal after K or R, allowing up to two missed cleavages. The minimum required peptide length was set to seven amino acids. Carbamidomethylation of cysteine was set as a fixed modification (57.021464 Da) and N-acetylation of proteins N termini (42.010565 Da) and oxidation of methionine (15.994915 Da) were set as variable modifications. As no labeling was performed, multiplicity was set to 1. During the main search, parent masses were allowed an initial mass deviation of 4.5 ppm and fragment ions were allowed a mass deviation of 0.5 Da. PSM and protein identifications were filtered using a target-decoy approach at a false discovery rate (FDR) of 1%. The second peptide feature was enabled. The match between runs option was also enabled with a match time window of 0.5 min and an alignment time window of 20 min. Relative, label-free quantification of proteins was done using the MaxLFQ algorithm (37) integrated into MaxQuant. The parameters were as follows: Minimum ratio count was set to 1, the FastLFQ option was enabled, LFQ minimum number of neighbors was set to 3, and the LFQ average number of neighbors to 6, as per default. The "proteinGroups" output file from MaxQuant is available in the supplement (supplemental Table S1), as well as all spectra for single-peptide-based protein identifications (supplemental Spectra).

Data Analysis—Further analysis of the MaxQuant-processed data was performed using the in-house developed Perseus software (version 1.4.2.30). The "proteingroups.txt" file produced by MaxQuant was loaded into Perseus. First, hits to the reverse database, contaminants and proteins only identified with modified peptides were eliminated. Then the LFQ intensities were logarithmized, and the pull-downs were divided into ES1 and ES2 and from then on analyzed separately. Samples were first grouped in triplicates or quadruplicates and identifications were filtered for proteins having at least three or four valid values in at least one replicate group, respectively. For every bait a separate grouping was defined, and the data was individually filtered for proteins containing at least two (ES1) or three (ES2) valid values in the specific bait pull-downs. After this, missing values were imputed with values representing a normal distribution around the detection limit of the mass spectrometer. To that end, mean and standard deviation of the distribution of the real intensities were

determined, then a new distribution with a downshift of 1.8 standard deviations and a width of 0.25 standard deviations was created. The total matrix was imputed using these values, enabling statistical analysis. Now a student's *t*-tests was performed comparing the bait pull-down (in replicates) to its individual bait specific control group (BSCG). This BSCG contained all other pull-downs in the data set except those of known complex members. This whole procedure of individual filtering, imputation and *t* test was repeated for every bait. The resulting differences between the logarithmized means of the two groups ("log2(bait/background") and the negative logarithmized *p* values were plotted against each other using R (version 2.15.3) in "volcano plots." We introduced two different cutoff lines with the function $y = c/(x - x_0)$, dividing enriched proteins into mildly and strongly enriched proteins (c = curvature, x_0 = minimum fold change). The positions of the cutoff lines were defined for each experimental series separately by first plotting the distribution of all observed enrichment factors and deriving the standard deviation of this distribution. The x_0 parameter for the inner curve and outer curve was then set to one and two standard deviations (rounded to one significant digit), respectively (supplemental Fig. S6*B* and S6*F*). The curvature parameters were obtained by overlaying all plots within one series, using only pull-downs of functional baits and rather small defined complexes (ES1: all but CDC73, PUP1, and PUP2; ES2: all but NUP84 and NUP145). The c parameter of the outer line was then adjusted to optimally separate true interactors from false positives (for more details see supplemental Fig.S6*C*, 6*D*, 6*G*, and 6*H*). The curvature of the inner line was then set to half of the curvature of the outer line. Cut-off parameters for ES1 were $x_0 = 0.9$ and $c = 4$ for the inner curve, and $x_0 = 1.8$ and $c = 8$ for the outer curve. Cutoff parameters for ES2 were $x_0 = 0.5$ and $c = 4$ for the inner curve, and $x_0 = 1$ and $c = 8$ for the outer curve. For all enriched proteins outside the inner cutoff line, we calculated the Pearson correlation of their LFQ intensity profile across all runs to the LFQ intensity profile of the corresponding bait. Enriched proteins were assigned to interactor confidence classes A, B, or C according to their position in the volcano plot and their correlation value. Cutoffs for the correlation scores were defined for both series individually by analyzing all correlations within one series using a quantile–quantile plot (Q–Q plot), which compares the real distribution of all correlation values to a theoretical normal distribution (supplemental Fig. S6*E* and 6*F*). The correlation cutoff was 0.55 for experimental series 1and 0.35 for experimental series 2. Note that these cutoff criteria do not represent absolute fixed values, but rather help to interpret the individual pull-down result.

RESULTS

Establishing a High Performance AE-MS Method for Detecting Interactions in Yeast—First, we set out to develop a generic and robust, yet high performance affinity enrichment-mass spectrometry (AE-MS) method for investigating protein–protein interactions in yeast. This organism is amenable to genetic and biochemical approaches and has already served as a model in many of the classical interactome studies. We chose to work with a GFP-tag system, because this tag is well tolerated and highly specific antibodies have been generated. Furthermore, a library of GFP-tagged yeast strains is commercially available, covering about 4000 open reading frames, and also offering localization data (34). The GFP-tagged bait proteins in this library are expressed at endogenous levels, a great advantage for detecting functional interactions. We chose a subset of 36 strains from this library, containing tagged bait proteins that are members of characterized complexes from various cellular compartments and cover the entire abundance range of the yeast proteome (supplemental Fig. S1).

Next, we wished to construct a control strain that was as genetically similar to the strains of the library as possible. Because the parental strain of the GFP-library, BY4741, is histidine auxotroph and does not express GFP, we reintroduced the HIS3 selection marker gene and a GFP gene into the dysfunctional HIS3 locus of BY4741 (Experimental Procedures). The resulting control strain can be grown under the same conditions as the strains of the GFP library, expresses moderate amounts of cytosolic GFP and was termed pHIS3-GFP.

An overview of our AE-MS workflow is depicted in Fig. 1. We combined a mild detergent-based lysis buffer with extensive bead beating to efficiently extract yeast proteins without disrupting interactions. We investigated the needed input amounts, and found that a 50 ml yeast culture volume with an $OD_{600\ nm}$ of 1.0 provided ample material for an IP experiment even with very low expressed baits. Starting from these initial 50 ODs of yeast cells allowed us to save material as backup at various stages of the sample preparation. The final amount injected into the mass spectrometer corresponded to only about 5.3 ODs; a very low amount of starting material, especially considering that baits were not overexpressed. The single-step affinity enrichment was performed with highly specific monoclonal anti-GFP antibodies coupled to magnetic microbeads in a flow-through column format using mild washing conditions to preserve weak or transient interactions (Experimental Procedures). The whole pull-down procedure was rather short, taking only about 2.5 h from lysis to elution. Proteins were eluted by in-column predigestion with trypsin, then digested to completion overnight. For all complexes tested, we found that the resulting peptides could be analyzed without any prefractionation in single-shot LC-MS/MS runs on Orbitrap instrumentation, which considerably shortens overall experiment time, provides greater reproducibility especially in a label-free format and higher sensitivity. All experiments were performed in several replicates; either biochemical triplicates (experimental series 1, ES1) or biological quadruplicates (experimental series 2, ES2).

Raw data were analyzed using MaxQuant (36), providing ppm level mass accuracy, confident identification of proteins (False Discovery Rate of less than 1%), and accurate intensity-based label-free quantification, thanks to recently developed sophisticated normalization and matching algorithms (37). Remarkably, all our pull-downs resulted in the identification of thousands of unspecific binders in addition to the specific interactors, leading to quantification of about half of the yeast proteome in every single sample. On the one hand, this was because of the low stringent single-step protocol in which we attempt enrichment instead of proper purification of protein complexes. On the other hand, it resulted from the high instrument sensitivity of the LTQ Orbitrap instrument,

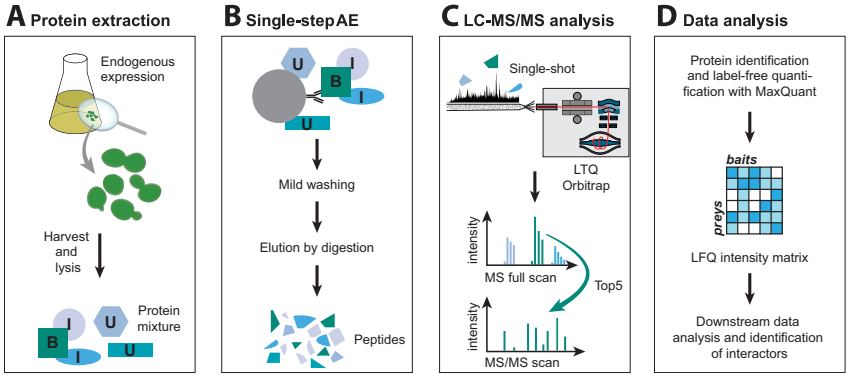


FIG. 1. **Schematic representation of the AE-MS workflow.** *A*, Endogenously expressed GFP-tagged proteins are extracted from yeast cells using mild, nondenaturing conditions. B = Bait, I = Interactor, U = Unspecific binder. *B*, Bait protein and specific interactors are enriched in a single-step immunoprecipitation using anti-GFP antibodies. Subsequently, bound proteins are digested into peptides. *C*, The peptide mixture is analyzed by single-shot liquid chromatography tandem mass spectrometry (LC-MS/MS) on an Orbitrap instrument. *D*, Raw data are processed with MaxQuant to identify and quantify proteins. The resulting label-free quantification (LFQ) intensity matrix is the basis for all downstream data analysis aimed at identifying interactors of the tagged bait proteins.

and was also promoted by the "match between runs" algorithm in MaxQuant. Matching between runs transfers identifications from one MS run to another run, where the same peptide feature was present, but not selected for fragmentation and hence not identified. High confidence matching is enabled by the high mass precision of the Orbitrap and achieved using unique *m/z* and retention time information of the features, after the retention times of all runs have been aligned (43). Processing with matching between runs increased the number of available quantifications in the combined (ES1+ES2) unfiltered LFQ matrix of 196 samples times 2304 proteins from 45 to 80%. The very large number of proteins quantified per IP prompted us to establish novel data analysis strategies, exploiting the information-rich intensity-based LFQ data, as described in the following sections.

AE-MS Produces Internal Beadomes for Every Pull-down—Together, our pull-downs identified a large set of background binders specific for the affinity matrix and conditions used in our experiments. As these proteins are usually detected because they bind to the beads used in the purification, the totality of them has been called the "bead proteome" or "beadome" (44, 45). Instead of having to determine this beadome from separate control experiments, here we detect it as a byproduct in the specific pull-downs ("internal beadome"). In total, after standard filtering (Experimental Procedures) of the data we quantified 2245 different protein groups in the combined ES1 and ES2 experimental series (Fig. 2*A*). Per pull-down, we quantified on average 1860 proteins in ES1 and 1825 proteins in ES2. Only a tiny fraction of the detected proteins in each pull-down were actual interactors of the corresponding tagged protein. For example, using MCM2 as bait recovered the six MCM complex members along with 1891 unspecific background proteins on average. These six proteins constituted only 0.3% of all identified proteins and only 1.3% of the summed LFQ intensity in the corresponding pull-downs, although the bait was among the highest intense proteins.

The unspecific binders identified in our internal beadome cover the entire abundance range, with only a small bias toward more highly abundant proteins when compared with the yeast proteome as a whole (46) (Fig. 2*B*). GOBP and GOCC term analysis by category counting of the identified proteins did not indicate cellular functions or compartments that are strongly over-or underrepresented (supplemental Fig. S2*A*). However, the intensity at which we detect proteins in the beadome is dependent on two factors: their abundance in the proteome and their affinity to the beads. Whereas low abundant proteins are generally not found at high intensities in the beadome, the intensities of high abundant proteins can vary from high to low signals (supplemental Fig. S2*B* and 2*C*). Pearson correlation between beadome intensity and proteome copy numbers was 0.53 for both ES1 and ES2. Next, we performed 2D enrichment analysis (47), in which we compared protein annotations between beadome and proteome in an intensity-dependent fashion. The major protein classes that showed higher intensities in the beadome than what would be expected from their cellular abundance were RNA or DNA related (*e.g.* ribosome, spliceosome, nucleolus, and DNA recombination). This confirms former findings that ribosomal proteins have a high affinity to the beads. Interestingly, proteins in metabolic categories, which are ubiquitously present in pull-downs because of their high abundance, tended to be de-enriched (supplemental Fig. S2*D* and 2*E*). We conclude that the beadome is in essence a scaled down version of the

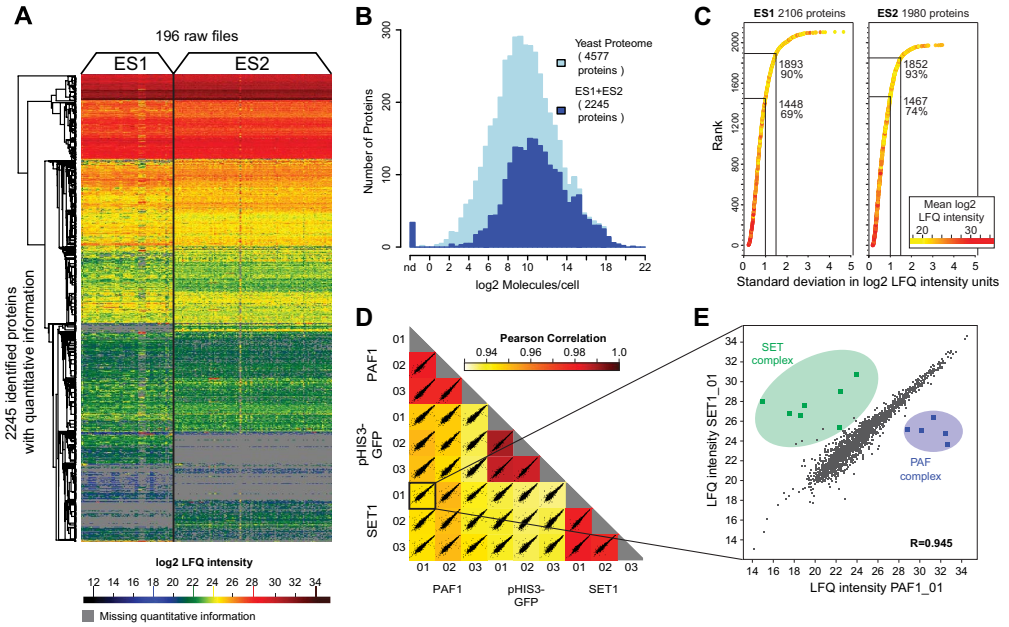


FIG. 2. **The proteomic nature of the background in AE-MS.** *A*, Heatmap of the LFQ intensities of all proteins identified in two experimental series (ES1 and ES2). Hierarchical row clustering was performed on the logarithmized LFQ intensities of more than 2000 quantified prey proteins in the 196 pull-downs, without data imputation. *B*, Histogram of the copy numbers of all proteins quantified in our pull-downs compared with the entire yeast proteome as in Kulak *et al.* *C*, The standard deviation of the LFQ intensity profile for each identified protein was calculated after imputing missing values. Proteins were then ranked according to the standard deviation of their profile. About 70% of detected proteins show a profile varying less than 1 log2 LFQ intensity unit and about 90% vary less than 1.5 log2 LFQ intensity units. *D*, Comparison of the control strain pHIS3-GFP with the two tagged strains SET1-GFP and PAF1-GFP; all measured in triplicates. The matrix of 36 correlation plots reveals very high correlations between LFQ intensities within triplicates (Pearson correlation coefficient $>$ 0.977 for all strains). The correlation between different strains is always higher than 0.935. Average correlation of the corresponding nine comparisons were: SET1-GFP to PAF1-GFP 0.946, SET1-GFP to control strain 0.938, and PAF1-GFP to control strain 0.945. *E*, Zoom into the SET1-GFP_01 *versus* PAF1-GFP_01 correlation plot. The majority of proteins are detected at very similar LFQ intensities in both pull downs. The proteins that differ the most between the two strains are the members of the two targeted complexes highlighted in color.

proteome, albeit with some preferences related to general protein binding properties.

The reproducible identification of unspecific binders across all runs is of course correlated with their intensity; higher intense background binders are more likely to always be detected, whereas background binders that are close to the level of detection may only be identified in some of the runs. Therefore, the LFQ intensity matrix contains missing values among the lower intense proteins (marked gray in Fig. 2*A*). To enable statistical analysis, such missing values can be "imputed." Therefore, after discarding proteins that are not reproducibly detected in at least one replicate group, we imputed the remaining missing values using a normal distribution around the detection limit of the mass spectrometer. These simulated low intensity values fit well into the profiles of the low abundant proteins, and because of its randomness, imputation does not create artifacts in *t*-tests or in intensity profile analyses. A comparison of the data set processed with and without matching identifications between runs, and the result of imputation are illustrated in supplemental Fig. S3.

Most of the background proteins are characterized by highly similar intensities in nearly all of the pull-downs within an experimental series, and we denote these as *typical background binders*. Both in ES1 and ES2 for about 90% of all detected proteins the standard deviation of their intensity profile was lower than 1.5 log2 LFQ intensity units; and for about 70% even lower than 1 (Fig. 2*C*). As expected, this analysis also confirms that proteins with higher intensity tend to have more stable background profiles. Next to the typical background binders, we also found a small number of proteins with irregular profiles. Those *atypical background binders* are usually among the lower intense proteins. Both types of unspecific binders can readily be distinguished from a specific interactor, whose profile ideally fluctuates mildly

around an average background intensity and only deviates from that behavior in specific pull-downs, where it is detected reproducibly and at higher intensities. The relationship of mean LFQ intensity and standard deviation of the intensity profile as well as the profiles of some typical and atypical unspecific binders are further documented in supplemental Fig. S4. Again, there is a clear trend that the intensity profiles of higher intense proteins have a smaller standard deviation. Among the proteins with the highest standard deviation (>1.5 log2 LFQ intensity units) many bait proteins and interactors are found.

A closer look at the heat map in Fig. 2*A* reveals the background in ES1 and ES2 to be slightly different. Sample preparation was similar in both experiments; however, ES1 and ES2 were measured on two different LC-MS systems of the same type but at different time periods, which introduces noticeable variation of the corresponding background. The variation between pull-downs is lower in ES2 because samples were measured directly after each other in contrast to ES1 where samples were measured in blocks according to baits. Because of the slight variations in the background signature between ES1 and ES2, data analysis was performed separately for each experimental series. The differences between ES1 and ES2 allowed us to study the influence of these workflow parameters.

Exploiting the High Coverage Background for Identifying Protein Complexes—Evidently, the extremely large number of unspecific binders detected in addition to the specific interactors in AE-MS represents a completely different experiment readout than that of classic AP-MS protocols. This large background needs specialized data analysis, which is; however, not aimed at removing the unspecific binders, but instead exploits them for high confidence detection of interactors. We recognized four different ways in which the unspecific binders detected in our pull-downs can be used beneficially.

First, they form the basis for intensity-based LFQ in MaxQuant. To produce reliable and accurate quantification results, the normalization procedure performed in MaxQuant requires a background proteome that is assumed to be unchanging. This function is provided here by a large number of unspecific binders identified in all samples. Normalization can then correct for differences in sample loading and sample concentration, which is a prerequisite to making the pull-downs comparable at all and constitutes the basis for further data analysis.

Second, the unspecific binders can serve as a quality control. We observed that deviation of the detected background binders from the standard behavior can indicate insufficient quality of a specific pull-down, which easily became apparent by hierarchical clustering of the data matrix. As an example, see the vertical stripe close to the middle of ES2 in Fig. 2*A*, which is a replicate of a pHIS3-GFP pull-down. Close inspection of the raw data revealed generally low peptide intensities and polymer contamination in this sample. In another case, a difference in background signature was not because of sample quality, but seemed to be because of the nature of the tagged complex: All six proteasome pull-downs reproducibly featured a slightly but clearly different background than the other pull-downs. This can be explained by the fact that proteasome subunits have high cellular copy numbers and are part of a very large complex; together this alters conditions on the beads, "crowding out" some of the normally observed background binders.

Third, the high number of unspecific binders reproducibly quantified in all samples resulted in very high correlations between different pull-downs. In Fig. 2*D*, these correlations are plotted for two tagged strains, SET1-GFP and PAF1-GFP, and the control strain pHIS3-GFP. Within triplicate pull-downs, the average Pearson correlation coefficients were always greater than 0.977. Between the different strains, correlation was always higher than 0.935, indicating that the intensities of the background proteins in the three yeast strains are highly similar. In fact, the correlation of SET1-GFP to PAF1-GFP was even higher than the correlation of SET1-GFP to the control strain pHIS3-GFP (0.945 *versus* 0.937). The proteins most changing in intensity between the two pull-downs were the expected SET1 and PAF1 interactors (Fig. 2*E*). These findings led us to investigate the possibility of comparing pull-downs not to an untagged control strain as it is usually done, but instead to compare them to each other, which will be further explored in the next section.

Finally, we reasoned that next to the pair-wise correlation of samples across all protein intensities, pair-wise correlation of intensity profiles across all samples should contain meaningful information. Specifically, intensity profiles of true interactors across all pull-downs, when compared with the intensity profile of the corresponding bait, should be correlated. The characteristic profile of interactors compared with the unchanging profile of typical background binders or the random profile of atypical background binders could therefore be useful in verifying interactor candidates, as we will demonstrate later on.

Defining Interactors by Comparing Against Other Tagged Strains—To identify interactors of a specific bait protein in the presence of the large amount of background binders, we performed a student's *t* test comparing the LFQ intensities of all proteins identified in replicates of that bait with the LFQ intensities of all proteins identified in the control (Experimental Procedures). When the resulting differences between the log2 mean protein intensities between bait and control are plotted against the negative logarithmized *p* values in volcano plots, the unspecific background binders center around zero. The enriched interactors appear on the right side of the plot, whereas ideally no proteins should appear on the left side when comparing to an empty control, as these would represent proteins depleted by the bait, which is not expected to happen. The higher the difference between the group means (*i.e.* the enrichment) and the *p* value (*i.e.* the reproducibility), the more

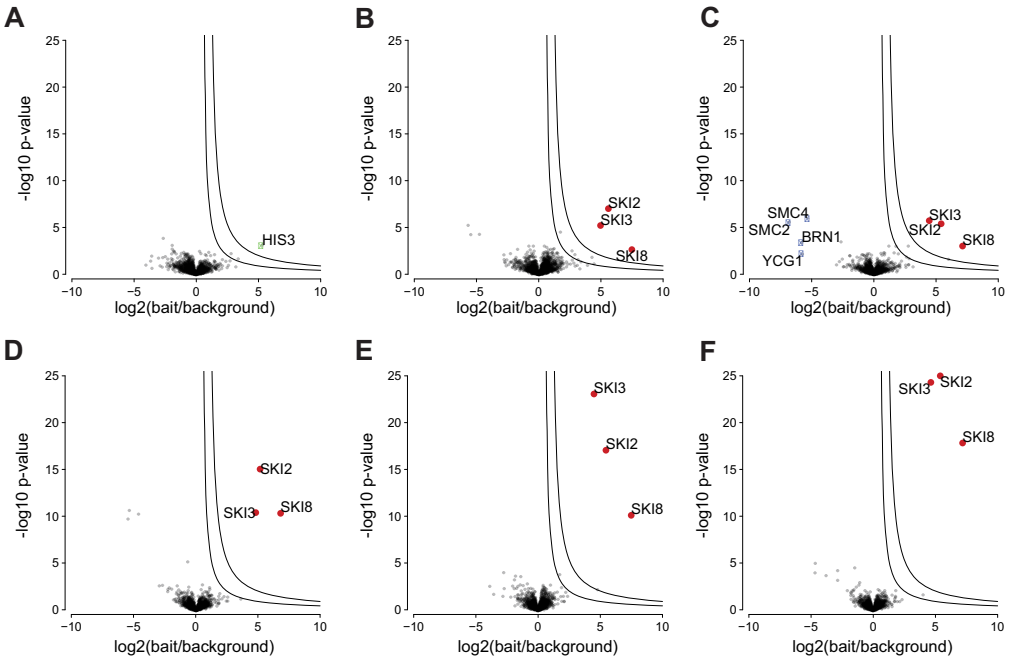


FIG. 3. **Comparing to unrelated tagged strains.** All pull-downs in this figure were measured in quadruplicates. Cut-off lines were those of ES2 (see Experimental Procedures). Red dots represent members of the SKI complex and blue dots represent members of the condensin complex. *A*, Comparison of the control strain pHIS3-GFP against its parental strain BY4741. *B*, Classic comparison of a tagged strain against an untagged control strain, in this case SKI2-GFP against pHIS3-GFP. *C*, SKI2-GFP compared with an unrelated tagged strain, SMC2-GFP. *D*, SKI2-GFP compared with 8 × pHIS3-GFP in quadruplicate (= 32 control pull-downs). *E*, SKI2-GFP compared with eight unrelated tagged strains in quadruplicate (APC1-GFP, CAF1-GFP, CCR4-GFP, PAF1-GFP, PEP5-GFP, SMC1-GFP, SMC2-GFP, and SNF4-GFP = 32 control pull-downs). *F*, SKI2-GFP compared with its bait specific control group (BSCG) consisting of all other pull-downs in the data set except for the SKI3 quadruplicate (= 116 control pull-downs).

the interactors move to the top right corner of the plot, which is the area of highest confidence for a true interaction.

We started by comparing a specific pull-down to an empty control strain as it is usually done in AP-MS experiments. First we used BY4741, the parental strain of the GFP library, as control; however, cross-reactivity of the anti-GFP antibody could occur in the complete absence of GFP. Therefore, we had constructed pHIS3-GFP, a control strain highly similar to the strains of the GFP library, as it could be grown under the same selective conditions and expressed moderate amounts of cytosolic GFP (see above). When we compared the pHIS3-GFP control strain to its parental strain BY4741, we detected only one yeast protein to be enriched, which was imidazoleglycerol-phosphate dehydratase, the protein the HIS3 gene encodes for (Fig. 3*A*). This illustrates that GFP does not interact with any yeast protein, and furthermore demonstrates that our AE-MS workflow is sensitive to an extent that it picks up genetic differences between strains. This confirms the benefits of using a control strain as similar as possible to the actual bait strain, and supports our hypothesis that other tagged strains of the GFP-library could present an excellent control, as they are genetically identical except for the different tagged protein. When we tested this idea on the example of the SKI complex we indeed did not observe any differences in the identified interactors of the bait SKI2, whether we compared with pHis3-GFP or a tagged strain, *e.g.* SMC2-GFP (Fig. 3*B* and 3*C*). As the only side-effect the specific interactors of the other strain now appeared as de-enriched proteins. (We note that even this could be put to good use in certain cases, as it in principle enables detection of the interactors of two different bait proteins in only one comparison and without employing a control.)

A larger control group consisting of many control pull-downs should help to better identify interactors; and we next tested whether this holds true for our pull-downs. Comparing a specific pull-down to eight pHis3-GFP pull-downs, consist-

ing of four biological replicates each, clearly led to better separation of interactors from the background cloud than just comparing to one pHIS3-GFP pull-down (compare Fig. 3*D* to Fig. 3*B*). The larger control group provided a less error-prone average background intensity of every protein, which in turn resulted in higher *p* values of the enriched true interacting proteins. This is particularly beneficial to separate weaker or transient interactors, which by their nature tend to only be mildly enriched, from the background cloud, as long as their low enrichment is highly reproducible. The more control pull-downs are included into the control group, the better the results should become. However, performing a large number of empty control experiments consumes considerable resources. In a human interactome study in 2007 for example, the authors conducted 202 control experiments (12). We reasoned that if we are able to compare tagged strains to each other, we would naturally obtain a large control group without any additional efforts. To test this concept, we first compared the SKI complex pull-downs to eight unrelated tagged strains. This resulted in the same or better statistical improvement of the interactors as we had obtained when using the same number of control strains (Fig. 3*E* and 3*D*). We chose the tagged strains serving as the control group to be unrelated to the specific bait of interest, in the sense that their tagged proteins do not reside in a known complex with this bait. To obtain the largest possible control group, we selected all unrelated pull-downs in the data set and termed this the "bait specific control group" (BSCG). If interacting proteins are included in the BSCG, they can increase the calculated average background intensity of interactors and therefore artificially decrease the *t test* result. For large control groups; however, wrong assignment would generally not dramatically change results, as demonstrated by comparing the SKI2 pull-downs against all other pull-downs in the data set (supplemental Fig. S5). Although we here constructed the BSCG from prior knowledge, it could also be constructed in an iterative way. In the case of SKI2, the BSCG included all pull-downs except the replicates of SKI3, resulting in 116 controls. This led by far to the best separation, placing the SKI complex into the far upper right corner of the volcano plot (Fig. 3*F*). Therefore, we concluded that other pull-downs can serve as excellent controls and in the following determined interactors by comparing each specific pull-down to its BSCG.

Combining Enrichment Over Background with Intensity Profile Analysis Leads to High Quality Interaction Data—To classify a protein as an interactor, we needed to introduce a cutoff that separates enriched proteins from the unchanged cloud of background binders centered around zero in the volcano plots. The position of this cutoff is crucial: A stringent cutoff leads to a low false positive rate, but may miss weaker or more transient interactors, whereas a permissive cutoff would include these, but at the cost of increasing false positives. To preserve information about weak or transient interactors, we decided to use a two cutoff strategy, which divides interactor candidates into mildly and strongly enriched proteins (Fig. 4*A*). To define the position of the two cutoff lines, we plotted the distribution of all enrichment factors within one series and placed two minimum fold change cutoffs at one and two standard deviations, respectively. Interestingly, in the case of ES2, the series with biological quadruplicates that had been measured in one block, the standard deviation was much lower than for ES1. The cutoff lines were placed once for all pull-downs within an experimental series with curvature parameters that best separate the outliers from the cumulative background cloud (for more details see Experimental Procedures and supplemental Fig. S6*A*–6*H*).

We then introduced a new criterion to deal with the false positives among the mildly enriched interactors close to the cutoff lines. This criterion makes use of the above mentioned tendency of intensity profiles of true interactors of a bait protein to be correlated, because interacting proteins should be enriched whenever one of the complex members is tagged. Moreover, slight variations across samples because of background binding should be followed by all complex subunits. This concept requires a complete LFQ intensity matrix, produced by imputing missing values from a suitably chosen random distribution, to not artificially increase or decrease the correlation (Experimental Procedures). To evaluate the similarity of a given profile to the profile of the bait, we calculated the Pearson correlation of the two profiles; and this was repeated for every enriched protein (Fig. 4*B*). Although strongly enriched proteins generally show medium to high correlations, mildly enriched proteins generally show lower correlations, but with a much higher variation from high to even negative values (supplemental Fig. S7). This indicates that true interactors exist among those borderline interactors that can be detected with the help of the correlation analysis. For the example of the MCM4 pull-down in Fig. 4, five out of the six complex members were highly enriched, but one (MCM3) only scored a mild enrichment and moderate *p* value, but a high correlation (0.56), which led to its correct identification as an interactor of MCM4. In this exemplary pull-down, the detected true interactors showed an average correlation of 0.68 to the bait, whereas the detected unspecific binders showed an average correlation of 0.42. In ES2, the average correlation of detected unspecific binders was generally even lower. We determined a series specific correlation cutoff for ES1 and ES2 by evaluating the correlation of all proteins detected in all pull-downs in a Q-Q-plot, which visualizes the real distribution of all correlation values compared with a theoretical normal distribution (supplemental Fig. S6*I* and 6*J*). The point, where actual and theoretical distribution sharply deviated was chosen as the correlation cutoff. Correlation analysis worked particularly well with our data set, as it contains at least two entry points for every complex.

We then proceeded to group enriched proteins into interactor confidence classes A–C by their enrichment, *p* value

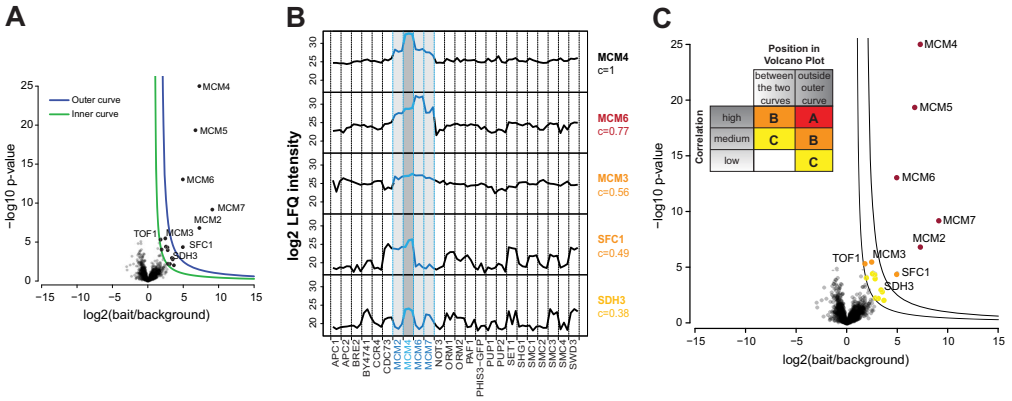


FIG. 4. **Classification of interactors.** Proteins are classified as interactors according to their position in the volcano plot and according to their correlation to the corresponding bait protein. *A*, Volcano Plot. Potential interactors are preclassified according to their position in the volcano plot into "mildly enriched" (between the two curves) and "strongly enriched" (outside the blue curve) proteins *B*, Intensity profile analysis of some enriched proteins from the volcano plot in *A*. From top to bottom: intensity profile of MCM4 (the bait protein), MCM6, and MCM3 (true interactors), and SFC1 and SDH3 (false positives) with the according calculated correlation to the profile of MCM4. *C*, Same volcano plot as in *A*, but with classification of interactors. *Insert:* Enrichment, reproducibility and correlation are combined to score interactors into interactor confidence classes A, B and C. Proteins between the cutoff curves with a low correlation (lower than 0.1) were not considered at all. Both proteins between the cutoff curves with a medium correlation (between 0.1 and the series-specific correlation cutoff) and proteins outside the outer cutoff curve with a low correlation (lower than 0.1) were assigned to class C (noninteractors). Proteins between the cutoff curves with a high correlation (higher than the series-specific correlation cutoff) as well as proteins outside the outer cutoff curve with a medium correlation were assigned to class B (lower confidence interactors). Proteins outside the outer cutoff curve with a high correlation were assigned to class A (high confidence interactors).

and correlation to the bait as summarized in Fig. 4*C*. Class C proteins are proteins between the two cutoff lines with low or medium correlation to the bait and are not regarded as interactors. Class B proteins are proteins between the cutoff lines with high correlation or proteins outside the outer cutoff line with medium correlation, and represent lower confidence interactors. Finally, class A proteins are proteins outside the outer cutoff line with high correlation and are considered high confidence interactors. The result of the classification is shown for the MCM complex in Fig. 4*C*, and the same color scheme is used in all volcano plots throughout the supplemental Material ES1/ES2. Although we found the above classification scheme to be very efficient, it should not be seen as absolute, but rather as a help in interpreting the pull-downs results.

How the intensity profile analysis can recognize false-positives is illustrated by the profiles of SFC1 and SDH3 in Fig. 4*B*. They represent atypical background binders (see above) fluctuating from low to high intensities across pull-downs. Because they appeared by chance in all of the replicates of the specific pull-down they scored both a good enrichment factor and *p* value. However, because of the fluctuations in their profiles, the correlation to the bait intensity profile is poor, which reclassifies SFC1 as lower confidence interactor and SDH3 as noninteractor. Without the correlation analysis, SFC1 would have been considered a high confidence interactor. Conversely, proteins that are only minimally but reproducibly enriched are likely to still be true interactors if they show good correlation (See MCM3 in Fig. 4*B*). Using the data set-dependent cutoff definition, the average complex coverage per pull-down (calculated as true positives/(true positives + false negatives), with true complex members derived from UniProt) was 74% for ES1 and even 83% for ES2. Among the 82 and 79 class A interactors, the false-positive rates (calculated as false positives/(true positives + false positives) were only 6 and 0% for ES1 and ES2 respectively. Among the 32 class B interactors in ES1, the false-positive rate was 53%; however, 15 out of these 17 false positives were downgraded from class A and therefore rightfully classified as lower confidence interactors. Among the 15 class B interactors in ES2, the false positive rate was 20%. False-negative rates in class C (calculated as true complex members falsely classified as class C/all proteins in class C) were very low with 3% (4 out of 133) for ES1 and 6% (2 out of 35) for ES2. For all the aforementioned calculations, the two large complexes (NPC and proteasome) as well as the complexes were no classification could be performed (APC2, CDC73, and TEF1) were excluded.

Defining Complexes of Various Sizes, Abundances, and Cellular Localizations—The bait proteins in our study had been selected to represent a wide range of cellular abundances (supplemental Fig. S1), localizations (*e.g.* cytosolic,

nuclear, and membrane bound), and functions (*e.g.* cell cycle, transcription, translation-elongation, and transport). For each of the pull-downs, the volcano plot containing the results of our analysis is depicted in supplemental Material ES1 and/or supplemental Material ES2. All bait proteins and the page number of the corresponding volcano plot within the supplemental Material ES1/ES2 are summarized in a table on the first page of both files. Given the diversity of these complexes, they serve to illustrate different aspects of our method.

When we used very low abundant proteins as baits, we were still able to identify interactors with a surprisingly high complex coverage, especially considering that our system uses endogenous expression and relatively little input material. For instance the members of the anaphase promoting complex, which has a key regulatory role in the cell cycle, are expressed at an estimated average of about 70 copies per cell in unsynchronized yeast cells (46). Using APC1 (about 30 copies/cell) as the entry point to the APC, our standard pull-down protocol already identified 11 out of 13 APC members. The two missing complex members (APC9 and APC11) are potentially even lower abundant in unsynchronized cells as they were also not detected in a deep yeast proteome (46). Similarly, pull-down of the SET1/COMPASS histone methyltransferase complex by its SET1 (135 copies/cell) and SWD3 (74 copies/cell) subunits revealed all eight complex members as clear outliers in the volcano plots.

Conversely, we were also able to detect interactors of very high abundant proteins. Here the challenge is that these proteins often have very high background intensities – ranging in our workflow to a log2 intensity of up to about 36 – over which they can hardly be further enriched. For the elongation factor CAM1 (49,500 copies/cell, average log2 background intensity 29.9) we identified CAM1 itself and its direct interactor EFB1 with a moderate but clear enrichment but an extremely significant p value ($p < 10^{-25}$). However, TEF1 (630,000 copies/cell, average log2 background intensity of 34.8), another elongation factor 1 complex member, did not register as an interactor as its background intensity is so high that it cannot be significantly further enriched. Even when we tagged TEF1, this bait was not an outlier, although all three interactors CAM1, EFB1, and TEF4 were significantly enriched. We also targeted another very high abundant complex, the ribosome-associated complex (RAC) through its components SSZ1 (59,450 copies/cell, average log2 background intensity of 32.2) and ZUO1 (45,188 copies/cell, average log2 background intensity of 31.4). Although SSZ1 only retrieved itself as outlier, when we tagged ZUO, we could indeed detect SSZ1 with mild enrichment but with a very good p value ($p < 10^{-22}$).

Although the above examples serve as positive controls, illustrating aspects of our affinity enrichment workflow, we also we detected some interactors that are not part of the stable, known core complexes. The MCM complex presents the core of the replicative DNA helicase in yeast and forms a double hexameric ring around the DNA (48). We identified TOF1 (Topoisomerase 1-associated factor 1) which is not part of the core helicase but which has been shown to interact and regulate it (49). TOF1 is an example of an interactor that was promoted to likely interactor status (class B), because of its high correlation with complex members.

The yeast proteasome consists of a 20S core particle composed of 28 α and β-subunits assembled into four rings, and a 19S regulatory particle on both sides of the core composed of 19 proteins. As the proteasome is a highly dynamic holocomplex, its purification is not trivial (50). Using two 20S members, PUP1 (β subunit) and PUP2 (α subunit), retrieved the complete 20S complex and most of the 19S members. Additionally, we found a number of transient interactors, such as the proteasome activator BLM10, the proteasome stabilizing component ECM29, the proteasome chaperone PBA1 and the uncharacterized protein YCR076C. The latter has already been reported to interact with proteasome core particle subunits (51), an association that we now confirm. Other enriched proteins found in the PUP1/PUP2 pull-downs that are not reported to interact with the complex could be proteasome substrates.

The nuclear pore complex (NPC) represents an example of a very large complex (about 30 different proteins in multiple copies) that is embedded in the nuclear membrane (52). Performing pull-downs with two of the subunits (NUP84 and NUP145), we found many components of the NPC (19 and 16 respectively), which, remarkably, is more than what was identified for these two baits in a dedicated membrane interactome (53). Additionally, we identified proteins that are not only components of the NPC but also of the spindle pole body (SPB), namely CDC31 (54, 55) and NDC1 (56). Consequently, other components of the SPB including SPC110 and SPC42 were among the outliers. We also identified the inner nuclear membrane protein HEH2, which has been proposed to be important for a proper distribution of nuclear pores across the nuclear envelope (57).

Two further examples are PAF1 (RNA polymerase II-associated protein 1), pull-down of which resulted in all five core complex members as well as RPO21. This protein is a subunit of the RNA polymerase II. Likewise pull-down of PEP5, a member of the HOPS complex, retrieved all its members, and furthermore VPS8, a component of the CORVET complex sharing four subunits (PEP3,PEP5,VPS16, and VPS33) with the HOPS complex (58).

Apart from core and transient, proteins can also be mutually exclusive complex members. As an example, the SNF1 protein kinase complex is a hetero-trimeric complex consisting of the alpha subunit SNF1, the gamma subunit SNF4, and one of three alternative beta subunits SIP1, SIP2, or GAL83 (Fig. 5*A*) (59). This complex proved to be a good case to investigate the effects of mutually exclusive complex members on the intensity profile analysis. We used SNF4 and GAL83 as baits, hence SIP1 and SIP2 were only identified in the SNF4 pull-down, as expected (Fig. 5*B* and 5*C*). Nevertheless they

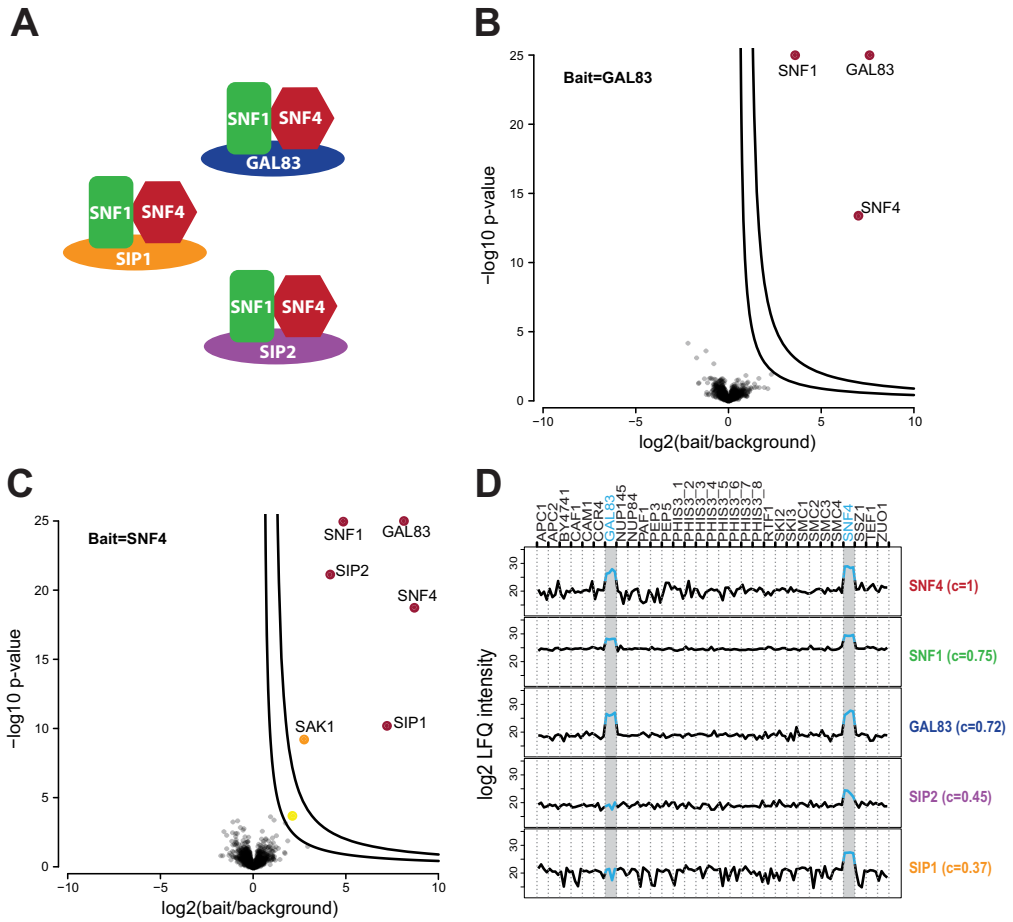


FIG. 5. **Correlation analysis and mutually exclusive binding.** *A*, Schematic representation of the three alternate SNF1 protein kinase complexes. *B*, Volcano plot of GAL83 compared with its bait-specific control group (BSCG). *C*, Volcano plot of SNF4 compared with its BSCG. *D*, Intensity profiles of the gamma subunit SNF4, the alpha subunit SNF1, and the three alternate beta subunits GAL83, SIP1, and SIP2 as well as their correlation to the bait SNF4.

showed a correlation of 0.37 and 0.45, respectively, to the bait SNF4 (Fig. 5*D*), which was higher than the correlation cutoff (0.35 for ES2). This demonstrates the usefulness of correlation analysis for associating even alternative members with the core complex. This complex also illustrates the need for several entry points per complex to recapitulate more complicated complex arrangements such as alternative cellular subcomplexes. Using SNF4 as bait, we additionally identified the protein SAK1, which is an upstream kinase that activates SNF1 (60).

DISCUSSION

For about two decades, AP-MS techniques have been used as tools for investigating protein complexes, and they have been improved greatly during this time. Previously, protein complexes were extensively purified, to reduce the amount of copurifying unspecific binders as much as possible. However, such stringent purification becomes unnecessary as soon as AP is coupled to high resolution, quantitative MS. Quantification can distinguish the true interactors from contaminants. Therefore, protocols can be less stringent, preserving weaker

interactions, while resulting in a higher background. In this work, we have taken this concept to its logical conclusion by employing low stringent single-step enrichment of protein complexes followed by label-free quantitative MS analysis in which we co-purify a very large number of unspecific binders representing about half of the yeast proteome. Complexes can still be confidently identified because of their enrichment in specific bait pull-downs *versus* all other pull-downs. As we do not aim to purify but only to enrich, we suggest terming such methods AE-MS. Our methodology is solely based on intensity-based label-free quantification, which has advanced considerably and for pull-downs is now comparable with label-based quantification approaches like SILAC (20, 33).

Identification of a large number of background binders is unavoidable with modern MS instrumentation. Perhaps counterintuitively, our results demonstrate that these unspecific proteins can actually be beneficial, elevating them from a nuisance to an essential part of the analysis. Apart from their essential use in normalization, they are indicators of the reproducibility within a specific workflow and serving as quality control. As unchanging background binders greatly outnumber changing interactors, pull-downs are highly similar to each other, which in turn obviates the need for a dedicated control strain. Finally, we have shown that reproducible detection of unspecific binders allows further characterization of interactor candidates by correlating their intensity profiles to the profile of the bait. Using our pipeline, we identified interactors of a diverse set of endogenously expressed bait proteins with high confidence, starting from minimal input amounts of unlabeled yeast, and requiring modest measuring times despite replicate analysis. In medium or large-scale projects, our workflow automatically provides a large control group, without actually performing any control pull-downs. However, as illustrated with the SKI complex, using only one tagged strain as control (or an empty stain) already correctly identified all complex members, demonstrating the feasibility of AE-MS also for small scale projects.

Although a large improvement, our AE-MS workflow does not solve all issues in MS-based interaction studies. Membrane complexes always present a challenge because of their hydrophobic nature. However, our protocol yielded excellent results for the HOPS vacuolar membrane complex and the nuclear pore complex without adapting it in any way. For the SPOTS complex, we only retrieved two out of the six complex members. Adapting the type of detergent or the detergent concentration in the lysis buffer may help to better identify membrane complexes (53). To further verify interactors, we have introduced intensity profile analysis, which proved to be very helpful for upgrading weaker interactors and uncovering false positives. As this method relies on correlation to the bait profile, it could; however, not be used in three cases where we did not detect the bait as an outlier (in ES1: APC2 and CDC73; in ES2: TEF1). In the case of CDC73, the bait was incorrectly tagged in the strain we used, as we subsequently found by a control PCR. For APC2 the very low copy number was presumably the reason, as even in ES2 where we found APC2, it was only identified with two peptides. Finally, as already mentioned, for TEF1 the background intensity was so high that it did not form a useful profile. However, the intensity profiling only serves as additional information, and in all these cases the correct interacting proteins were still identified through their enrichment. A final potential caveat for the intensity profile analysis are newly identified proteins interacting with several baits, which decreases their correlation score. However, provided their enrichment is high, they would still be considered (class B) interactors. Examination of the actual intensity profile of such promiscuous interactors could also help in judging whether weak correlation to the bait is caused by strong fluctuation between all samples, making the protein a false positive, or caused by strong fluctuation between several replicate groups, making it a potential link between several complexes.

The two largest yeast interactomes published in 2006 by Gavin *et al.* and Krogan *et al.* both employed TAP-tagging coupled to nonquantitative MS and among other frequency filtering of detected proteins to remove unspecific binders (9, 10). This can be problematic in the case of atypical background binders that appear spontaneously at high intensity in only some pull-downs. In our AE-MS approach, pull-downs are performed in replicates, hence such proteins are rarely scored as interactors. Even if an atypical background binder is by chance detected in all replicates, the intensity profile analysis can still uncover it. With very few exceptions, all of the proteins listed as contaminant in the above studies were also found in our data set. However, they did not appear as interactors in any of our pull-downs other than where expected. The data sets of Gavin *et al.* and Krogan *et al.* only share about one quarter of detected interactions (61) and did not contain 1/3 or 1/2 of the baits that we had tagged here, respectively. For each of the pull-downs that we could compare between all three studies (APC2, BRE2, CCR4, NUP84, NUP145, POP2, RTF1, SET1, SKI2, SMC1, SSZ1, and SWD3) the complex coverage was equal or better using the AE-MS method. In one case, we only retrieved EFB1 as interactors of CAM1 whereas Gavin *et al.* also found TEF1 and TEF2. Although these proteins were also found in a mock TAP-tag purification and therefore included in the contaminant list, we reason that more stringent purification could be helpful for detecting interactors of extremely high expressed proteins such as CAM1.

Recent interaction proteomics efforts typically at least employ semiquantitative approaches; however, removal of contaminants can still be problematic. There is an ongoing collaborative effort to establish a “contaminant repository for affinity purification,” the “CRAPome,” containing control pull-downs from various laboratories performed under various experimental conditions (62). In the case of yeast 17 control pull-downs are currently available, of which 12 have been

performed using GFP-tagged proteins and nano-magnetic beads. However, a larger number of controls may be necessary to comprehensively cover all nonspecific binders and thereby avoid incorrectly classifying a nonspecific binder as an interactor. Our AE-MS method sidesteps this problem, as the samples themselves are the controls. The minor but clear differences between our two experimental series (Fig. 2*A*) demonstrate that minor changes in the workflow like using a different machine of the same type can already alter the detected low abundant background binders, making the notion of a universal CRAPome problematic.

From the differences between the two experimental series we also conclude that for the most optimal output, AE-MS experiments should be executed in a reproducible manner from sample preparation to MS measurement, which should ideally be conducted on one machine and in one batch as in ES2. However, the MaxLFQ normalization algorithm successfully corrected for most of the variability in the ES1 series in general and in the proteasome pull-downs in particular, resulting in excellent results even for ES1.

To perform the AE-MS workflow described here, only three elements were needed: tagged proteins of interest, a high resolution LC-MS system, and sophisticated software to quantify proteins and analyze the data. Here we used the LTQ Orbitrap classic, which—although not being the latest Orbitrap technology—proved to be sufficient for identifying even very low abundant protein complexes. Such technology is now widely accessible, as is the MaxQuant software for performing accurate intensity-based label-free quantification and the Perseus program for statistical analysis of the data. Our AE-MS protocol is equally suited to investigate a small, medium or large number of samples. For a smaller set of samples, SILAC labeling could easily be implemented, which might provide even more accurate ratios in the case of borderline enrichment. More and more AP-MS workflows already use single-step protocols and employ high resolution MS, and therefore rather represent AE-MS methods. The shift in the conceptual framework from AP-MS to AE-MS and the development of sophisticated analysis tools for AE-MS experiments should contribute to higher quality interaction data, thereby making studies more comparable, and helping to solve open challenges in the interactomics field.

Acknowledgments—We thank Nils A. Kulak for input regarding yeast culture, Jürgen Cox for advice regarding data analysis and Roland Wedlich-Söldner for providing the strains of the yeast-GFP clone collection.

* This work was supported by the Bundesministerium für Bildung und Forschung (grant number FKZ01GS0861, DiGtoP consortium) and the European Commission's 7th Framework Program PROSPECTS (HEALTH-F4-2008-201648).

S This article contains supplemental Figs. S1 to S7, Table S1, Spectra, and Experimental Series S1 and S2.

§ To whom correspondence should be addressed: Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, Martinsried (near Munich) D-82152 Germany. Tel.: 49-89-8578 2557; Fax: 49-89-8578 2219; E-mail: mmann@biochem.mpg.de.

DATA AVAILABILITY: The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (63). (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (64) with the data set identifier PXD000955.

REFERENCES

1. Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8,** 645–654
2. Oeffinger, M. (2012) Two steps forward–one step back: advances in affinity purification mass spectrometry of macromolecular complexes. *Proteomics* **12,** 1591–1608
3. Gavin, A. C., Maeda, K., and Kuhner, S. (2011) Recent advances in charting protein–protein interaction: mass spectrometry-based approaches. *Curr. Opin. Biotechnol.* **22,** 42–49
4. Fields, S., and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246
5. Rajagopala, S. V., Sikorski, P., Caufield, J. H., Tovchigrechko, A., and Uetz, P. (2012) Studying protein complexes by the yeast two-hybrid system. *Methods* **58,** 392–399
6. Parrish, J. R., Gulyas, K. D., and Finley, R. L., Jr. (2006) Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **17,** 387–393
7. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147
8. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415,** 180–183
9. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636
10. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440,** 637–643
11. Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433,** 531–537

12. Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J. P., Duewel, H. S., Stewart, II, Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S. L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007) Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3,** 89
13. Kuhner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Castano-Diez, D., Chen, W. H., Devos, D., Guell, M., Norambuena, T., Racke, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Bottcher, B., Frangakis, A. S., Russell, R. B., Serrano, L., Bork, P., and Gavin, A. C. (2009) Proteome organization in a genome-reduced bacterium. *Science* **326,** 1235–1240
14. Guruharsha, K. G., Rual, J. F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D. Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K. H., Yu, C., Parsa, B., Carlson, J. W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S. P., Celniker, S. E., Obar, R. A., and Artavanis-Tsakonas, S. (2011) A protein complex network of Drosophila melanogaster. *Cell* **147,** 690–703
15. Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., Kim, B. J., Li, C., Chen, R., Li, W., Wang, Y., O'Malley, B. W., and Qin, J. (2011) Analysis of the human endogenous coregulator complexome. *Cell* **145,** 787–799
16. Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72,** 1156–1162
17. Gibson, T. J., Seiler, M., and Veitia, R. A. (2013) The transience of transient overexpression. *Nat. Methods* **10,** 715–721
18. Glatter, T., Wepf, A., Aebersold, R., and Gstaiger, M. (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.* **5,** 237
19. Poser, I., Sarov, M., Hutchins, J. R., Heriche, J. K., Toyoda, Y., Pozniakovsky, A., Weigl, D., Nitzsche, A., Hegemann, B., Bird, A. W., Pelletier, L., Kittler, R., Hua, S., Naumann, R., Augsburg, M., Sykora, M. M., Hofemeister, H., Zhang, Y., Nasmyth, K., White, K. P., Dietzel, S., Mechtler, K., Durbin, R., Stewart, A. F., Peters, J. M., Buchholz, F., and Hyman, A. A. (2008) BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5,** 409–415
20. Hubner, N. C., Bird, A. W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., Hyman, A., and Mann, M. (2010) Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189,** 739–754
21. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17,** 1030–1032
22. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
23. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
24. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169
25. Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K., and Hamon, C. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75,** 1895–1904
26. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21,** 315–318
27. Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **33,** 349–355
28. Paul, F. E., Hosp, F., and Selbach, M. (2011) Analyzing protein–protein interactions by quantitative mass spectrometry. *Methods* **54,** 387–395
29. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201
30. Bondarenko, P. V., Chelius, D., and Shaler, T. A. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem.* **74,** 4741–4749
31. Nahnsen, S., Bielow, C., Reinert, K., and Kohlbacher, O. (2013) Tools for label-free peptide quantification. *Mol. Cell Proteomics* **12,** 549–556
32. Luber, C. A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M., Tschopp, J., Akira, S., Wiegand, M., Hochrein, H., O'Keeffe, M., and Mann, M. (2010) Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **32,** 279–289
33. Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M., and Mann, M. (2013) A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. *Mol. Cell* **49,** 368–378
34. Choi, H., Glatter, T., Gstaiger, M., and Nesvizhskii, A. I. (2012) SAINT-MS1: protein–protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J. Proteome Res.* **11,** 2619–2624
35. Poulsen, J. W., Madsen, C. T., Young, C., Poulsen, F. M., and Nielsen, M. L. (2013) Using guanidine-hydrochloride for fast and efficient protein digestion and single-step affinity-purification mass spectrometry. *J. Proteome Res.* **12,** 1020–1030
36. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
37. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13,** 2513–2526
38. Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111 013722
39. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691
40. Hubner, N. C., and Mann, M. (2011) Extracting gene function from protein–protein interactions using Quantitative BAC InteraCtomics (QUBIC). *Methods* **53,** 453–459
41. Rappsilber, J., Mann, M., and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, prefractionation, and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2,** 1896–1906
42. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
43. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11,** M111 014050
44. Trinkle-Mulcahy, L., Boulon, S., Lam, Y. W., Urcia, R., Boisvert, F. M., Vandermoere, F., Morrice, N. A., Swift, S., Rothbauer, U., Leonhardt, H., and Lamond, A. (2008) Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J. Cell Biol.* **183,** 223–239
45. Rees, J. S., Lowe, N., Armean, I. M., Roote, J., Johnson, G., Drummond, E., Spriggs, H., Ryder, E., Russell, S., St Johnston, D., and Lilley, K. S. (2011) In vivo analysis of proteomes and interactomes using Parallel Affinity Capture (iPAC) coupled to mass spectrometry. *Mol. Cell. Proteomics* **10,** M110 002386
46. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11,** 319–324
47. Cox, J., and Mann, M. (2012) 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13,** S12
48. Remus, D., Beuron, F., Tolun, G., Griffith, J. D., Morris, E. P., and Diffley,

J. F. (2009) Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139,** 719–730
49. Nedelcheva, M. N., Roguev, A., Dolapchiev, L. B., Shevchenko, A., Taskov, H. B., Shevchenko, A., Stewart, A. F., and Stoynov, S. S. (2005) Uncoupling of unwinding from DNA synthesis implies regulation of MCM helicase by Tof1/Mrc1/Csm3 checkpoint complex. *J. Mol. Biol.* **347,** 509–521
50. Forster, F., Unverdorben, P., Sledz, P., and Baumeister, W. (2013) Unveiling the long-held secrets of the 26S proteasome. *Structure* **21,** 1551–1562
51. Hatanaka, A., Chen, B., Sun, J. Q., Mano, Y., Funakoshi, M., Kobayashi, H., Ju, Y., Mizutani, T., Shinmyozu, K., Nakayama, J., Miyamoto, K., Uchida, H., and Oki, M. (2011) Fub1p, a novel protein isolated by boundary screening, binds the proteasome complex. *Genes Genet. Syst.* **86,** 305–314
52. Fernandez-Martinez, J., and Rout, M. P. (2012) A jumbo problem: mapping the structure and functions of the nuclear pore complex. *Curr. Opin. Cell Biol.* **24,** 92–99
53. Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B. D., Burston, H. E., Vizeacoumar, F. J., Snider, J., Phanse, S., Fong, V., Tam, Y. Y., Davey, M., Hnatshak, O., Bajaj, N., Chandran, S., Punna, T., Christopolous, C., Wong, V., Yu, A., Zhong, G., Li, J., Stagljar, I., Conibear, E., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2012) Interaction landscape of membrane–protein complexes in Saccharomyces cerevisiae. *Nature* **489,** 585–589
54. Spang, A., Courtney, I., Fackler, U., Matzner, M., and Schiebel, E. (1993) The calcium-binding protein cell division cycle 31 of Saccharomyces cerevisiae is a component of the half bridge of the spindle pole body. *J. Cell Biol.* **123,** 405–416
55. Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B. T. (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651
56. Chial, H. J., Rout, M. P., Giddings, T. H., and Winey, M. (1998) Saccharomyces cerevisiae Ndc1p is a shared component of nuclear pore complexes and spindle pole bodies. *J. Cell Biol.* **143,** 1789–1800
57. Yewdell, W. T., Colombi, P., Makhnevych, T., and Lusk, C. P. (2011) Lumenal interactions in nuclear pore complex assembly and stability. *Mol. Biol. Cell* **22,** 1375–1388
58. Balderhaar, H. J., and Ungermann, C. (2013) CORVET and HOPS tethering complexes - coordinators of endosome and lysosome fusion. *J. Cell Sci.* **126,** 1307–1316
59. Nath, N., McCartney, R. R., and Schmidt, M. C. (2002) Purification and characterization of Snf1 kinase complexes containing a defined Beta subunit composition. *J. Biol. Chem.* **277,** 50403–50408
60. Hedbacker, K., and Carlson, M. (2008) SNF1/AMPK pathways in yeast. *Front. Biosci.* **13,** 2408–2420
61. Kiemer, L., Costa, S., Ueffing, M., and Cesareni, G. (2007) WI-PHI: a weighted yeast interactome enriched for direct physical interactions. *Proteomics* **7,** 932–943
62. Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J. P., St-Denis, N. A., Li, T., Miteva, Y. V., Hauri, S., Sardiu, M. E., Low, T. Y., Halim, V. A., Bagshaw, R. D., Hubner, N. C., Al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W. H., Goudreault, M., Lin, Z. Y., Badillo, B. G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A. J., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I. M., Bennett, K. L., Washburn, M. P., Raught, B., Ewing, R. M., Gingras, A. C., and Nesvizhskii, A. I. (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10,** 730–736
63. Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P. A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H. J., Albar, J. P., Martinez-Bartolome, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223–226
64. Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–D1069

2.1.2 A high-throughput pipeline to measure 96 yeast pulldowns in one day

Hosp F., Scheltema R.A., Eberl H.C., Kulak N.A., **Keilhauer, E. C.**, Mayr, K. & Mann, M.

A double-barrel LC-MS/MS system to quantify 96 interactomes per day
Molecular and Cellular Proteomics 2015 Apr 17 [epub ahead of print]

Since the new yeast AE-MS method had proven very successful, we next wished to further develop it into a high-throughput pipeline. Having spent quite some time already on a purely technological project, I decided not to take the lead in this but rather a supportive role. Hence, the project was first started by Dr. Christian Eberl, a PostDoc in the group. The initial idea was to transfer all sample preparation steps to a 96-well format to be able to process many more pulldowns in parallel. With my help, Chris started to adapt the yeast pulldowns to a 96-well format using filter plates. However he soon left the group, and the project was taken over by a new PostDoc, Dr. Fabian Hosp. After I introduced him to the yeast pulldown pipeline, he successfully optimized the yeast culture by transferring it to 96-well plates, enabling the growth of 96 GFP-strains in parallel. The challenge here is the maximum culture volume of only 2 ml, however, already in the initial yeast pulldown pipeline I had used relatively low input amounts. Yeast lysis could unfortunately not efficiently be performed in the 96-well plates, hence was performed as optimized by me before by beadbeating. For the pulldowns, we switched from the established robotic platform to 96-well plates coated with anti-GFP antibodies and manual pipetting, due to the low input amounts. After a high-throughput sample preparation pipeline was established, the MS measurement time with gradient lengths of over two hours became the major bottleneck in regard of throughput. The gradient length had to be reduced, which proved to be feasible because of the low complexity pulldown samples and the high performance of the Q Exactive HF. To even further increase sample throughput, Dr. Richard Scheltema, another PostDoc in the group, implemented a double-barrel system driving two chromatography columns in parallel. Whenever a gradient was running on one column, the second column could already be loaded with the next sample which drastically shortened machine idling time between runs. Finally, we adapted the data analysis pipeline to the low complexity pulldowns. Due to the double-barrel system, the new high-throughput pipeline now allows to measure up to 96 pulldowns in about one day. We applied the method to investigate the yeast chromatin

remodeling landscape, and obtained high complex coverage for the 21 targeted complexes using our high-throughput format. The described method provides the means for future dynamic interactome analyses, but should also be universally applicable to various kinds of low complexity proteomes.

Technological Innovations and Resources

A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day*[S]

Fabian Hosp‡‖, Richard A. Scheltema‡‖, H. Christian Eberl‡‖¶, Nils A. Kulak‡‖, Eva C. Keilhauer‡, Korbinian Mayr‡, and Matthias Mann‡§

The field of proteomics has evolved hand-in-hand with technological advances in LC-MS/MS systems, now enabling the analysis of very deep proteomes in a reasonable time. However, most applications do not deal with full cell or tissue proteomes but rather with restricted subproteomes relevant for the research context at hand or resulting from extensive fractionation. At the same time, investigation of many conditions or perturbations puts a strain on measurement capacity. Here, we develop a high-throughput workflow capable of dealing with large numbers of low or medium complexity samples and specifically aim at the analysis of 96-well plates in a single day (15 min per sample). We combine parallel sample processing with a modified liquid chromatography platform driving two analytical columns in tandem, which are coupled to a quadrupole Orbitrap mass spectrometer (Q Exactive HF). The modified LC platform eliminates idle time between measurements, and the high sequencing speed of the Q Exactive HF reduces required measurement time. We apply the pipeline to the yeast chromatin remodeling landscape and demonstrate quantification of 96 pulldowns of chromatin complexes in about 1 day. This is achieved with only 500 μg input material, enabling yeast cultivation in a 96-well format. Our system retrieved known complex-members and the high throughput allowed probing with many bait proteins. Even alternative complex compositions were detectable in these very short gradients. Thus, sample throughput, sensitivity and LC/MS-MS duty cycle are improved severalfold compared with established workflows. The pipeline can be extended to different types of interaction studies and to other medium complexity proteomes. ***Molecular & Cellular Proteomics 14: 10.1074/mcp.O115.049460, 2030–2041, 2015.***

Shotgun proteomics is concerned with the identification and quantification of proteins (1–3). Prior to analysis, the proteins are digested into peptides, resulting in highly complex mixtures. To deal with this complexity, the peptides are separated by liquid chromatography followed by online analysis with mass spectrometry (MS), today facilitating the characterization of almost complete cell line proteomes in a short time (3–5). In addition to the characterization of entire proteomes, there is also a great demand for analyzing low or medium complexity samples. Given the trend toward a systems biology view, relatively large sets of samples often have to be measured. One such category of lower complexity protein mixtures occurs in the determination of physical interaction partners of a protein of interest, which requires the identification and quantification of the proteins "pulled-down" or immunoprecipitated via a bait protein. Protein interactions are essential for almost all biological processes and orchestrate a cell's behavior by regulating enzymes, forming macromolecular assemblies and functionalizing multiprotein complexes that are capable of more complex behavior than the sum of their parts. The human genome has almost 20,000 protein encoding genes, and it has been estimated that 80% of the proteins engage in complex interactions and that 130,000 to 650,000 protein interactions can take place in a human cell (6, 7). These numbers demonstrate a clear need for systematic and high-throughput mapping of protein–protein interactions (PPIs) to understand these complexes.

The introduction of generic methods to detect PPIs, such as the yeast two-hybrid screen (Y2H) (8) or affinity purification combined with mass spectrometry (AP-MS)[1] (9), have revolutionized the protein interactomics field. AP-MS in particular has emerged as an important tool to catalogue interactions with the aim of better understanding basic biochemical mech-

From the ‡Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

Received February 27, 2015, and in revised form, February 27, 2015

Published, MCP Papers in Press April 17, 2015, DOI 10.1074/mcp.O115.049460

Author contributions: F.H., R.A.S., C.E., and M.M. designed the research; F.H., R.A.S., and C.E. performed the research; F.H., R.A.S., N.A.K., E.C.K., and K.M. contributed new reagents or analytic tools; F.H., R.A.S., C.E., N.A.K., E.C.K., and M.M. analyzed the data; and F.H., R.A.S., and M.M. wrote the paper.

[1] The abbreviations used are: AP-MS, affinity purification mass spectrometry; FDR, false discovery rate; GFP, green fluorescent protein; MaxLFQ, MaxQuant label-free quantification; PPI, protein–protein interaction; TMT, tandem mass tag; Y2H, yeast two-hybrid.

anisms in many different organisms (10–17). It can be performed under near-physiological conditions and is capable of identifying functional protein complexes (18). In addition, the combination of affinity purification with quantitative mass spectrometry has greatly improved the discrimination of true interactors from unspecific background binders, a long-standing challenge in the AP-MS field (19–21). Nowadays, quantitative AP-MS is employed to address many different biological questions, such as detection of dynamic changes in PPIs upon perturbation (22–25) or the impact of posttranslational signaling on PPIs (26, 27). Recent developments even make it possible to provide abundances and stoichiometry information of the bait and prey proteins under study, combined with quantitative data from very deep cellular proteomes. Furthermore, sample preparation in AP-MS can now be performed in high-throughput formats capable of producing hundreds of samples per day. With such throughput in sample generation, the LC-MS/MS part of the AP-MS pipeline has become a major bottleneck for large studies, limiting throughput to a small fraction of the available samples. In principle, this limitation could be circumvented by multiplexing analysis via isotope-labeling strategies (28, 29) or by drastically reducing the measurement time per sample (30–32). The former strategy requires exquisite control of the processing steps and has not been widely implemented yet. The latter strategy depends on mass spectrometers with sufficiently high sequencing speed to deal with the pull-down in a very short time. Since its introduction about 10 years ago (33), the Orbitrap mass spectrometer has featured ever-faster sequencing capabilities, with the Q Exactive HF now reaching a peptide sequencing speed of up to 17 Hz (34). This should now make it feasible to substantially lower the amount of time spent per measurement.

Although very short LC-MS/MS runs can in principle be used for high-throughput analyses, they usually lead to a drop in LC-MS duty cycle. This is because each sample needs initial washing, loading, and equilibration steps, independent of gradient time, which takes a substantial percentage for most LC setups - typically at least 15–20 min. To achieve a more efficient LC-MS duty cycle, while maintaining high sensitivity, a second analytical column can be introduced. This enables the parallelization of several steps related to sample loading and to the LC operating steps, including valve switching. Such dual analytical column or "double-barrel: setups have been described for various applications and platforms (30, 35–39).

Starting from the reported performance and throughput of workflows that are standard today (16, 21, 40–42), we asked if it would be possible to obtain a severalfold increase in both sample throughput and sensitivity, as well as a considerable reduction in overall wet lab costs and working time. Specifically, our goal was to quantify 96 medium complexity samples in a single day. Such a number of samples can be processed with a 96-well plate, which currently is the format of choice for highly parallelized sample preparation workflows, often with a high degree of automation. We investigated which advances were needed in sample preparation, liquid chromatography, and mass spectrometry. Based on our findings, we developed a parallelized platform for high-throughput sample preparation and LC-MS/MS analysis, which we applied to pull-down samples from the yeast chromatin remodeling landscape. The extent of retrieval of known complex members served as a quality control of the developed pipeline.

EXPERIMENTAL PROCEDURES

Preparation of Yeast Lysates—GFP-tagged yeast strains from the *Saccharomyces cerevisiae* GFP Clone Collection (43), the parental strain BY4741 and the control strain pHis3-GFP (21) were cultured in YPD liquid medium in 96-deep well plates (Sarstedt, Nümbrecht, Germany) at standard conditions. We used 32 distinct yeast strains in biological triplicates, resulting in 96 experimental samples. Yeast cells were grown until they reached an Optical Density$_{600\ nm}$ of around 1, followed by harvesting culture volumes equaling 2 ODs per well. Yeast cell pellets were dissolved in 300 μl lysis buffer (150 mM NaCl, 50 mM Tris-HCl (pH 8.0), 1 mM $MgCl_2$, 5% glycerol, 1% IGEPAL CA-630 (Sigma-Aldrich, Schnelldorf, Germany), complete protease inhibitors (Roche, Mannheim, Germany), 1% benzonase (Merck, Darmstadt, Germany)), transferred into FastPrep tubes (MP Biomedicals, Eschwege, Germany) containing 1 mm silica spheres (lysing matrix C, MP Biomedicals), and lysed in a FastPrep24 instrument (MP Biomedicals) for 6 $\times$ 1 min at maximum speed. Lysates were cleared by centrifugation at 16,100 $\times$ *g* for 10 min at 4 °C.

Affinity Purification—Each well of a GFP-multiTrap plate (ChromoTek, Martinsried, Germany) was washed three times with 200 μl buffer 1 (150 mM NaCl, 50 mM Tris-HCl, pH 8.0) and then incubated with the cleared yeast cell lysate (500 μg total protein extract) with gentle shaking at 100 rpm for 60 min at 4 °C. Next, each well was washed twice with 200 μl buffer 2 (150 mM NaCl, 50 mM Tris-HCl (pH 8.0), 0.25% IGEPAL CA-630) and four times with 200 μl buffer 1 before incubation with 25 μl elution buffer (2 M urea, 20 mM Tris-HCl (pH 8.0), 1 mM DTT, 100 ng sequence-grade modified trypsin (Promega, Madison, WI, USA) at room temperature for 90 min. Subsequently, the resulting peptides were alkylated with 25 μl alkylation buffer (2 M urea, 20 mM Tris HCl (pH 8.0), 5 mM iodoacetamide) and finally washed once with 50 μl urea buffer (2 M urea, 20 mM Tris HCl, pH 8.0) for 10 min, respectively. The supernatants from the elution, alkylation and washing step were collected after each step and combined in a clean 96-well plate. This plate was incubated overnight at room temperature to ensure a complete digest. The next morning, the digest was stopped by addition of 10 μl 10% TFA per well. The acidified peptides were purified on StageTips (44) containing two layers of Poly(styrenedivinylbenzene)-Reversed-Phase Sulfonate (Empore 2241, 3 M, Neuss, Germany) material to desalt and purify the peptides. Samples were eluted from the StageTips with 60 μl elution buffer (80% acetonitrile, 1% ammonium hydroxide) and evaporated in a SpeedVac concentrator for 30 min. The remaining peptide solution volume was adjusted to 4 μl with buffer A* (2% ACN, 0.1% formic acid).

LC-MS/MS Analysis—Online chromatography was performed with a modified Thermo EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific, Bremen, Germany) coupled online to the Q Exactive HF instrument with a nano-electrospray ion source (Thermo Fisher Scientific). Two analytical columns (15 cm long, 75 μm inner diameter) were packed in-house with ReproSil-Pur C_{18} AQ 1.9 μm reversed phase resin (Dr. Maisch GmbH, Ammerbuch, Germany) in buffer A (0.5% formic acid) and matched with regard to back-pressure to

ensure intercolumn reproducibility. During online analysis, the analytical columns were placed in a modified column heater (Sonation GmbH, Biberach, Germany) regulated to a temperature of 55 °C. Modifications to both systems are described in RESULTS. Peptides were loaded onto the analytical columns with buffer A at a back pressure of 650 bar (generally resulting in a flow rate of 500 nL/min) and separated with two distinct linear gradients of 8–30% buffer B (80% ACN and 0.5% formic acid) at a flow rate of 450 nL/min controlled by IntelliFlow technology over 10 min and 22 min, respectively (generally at a back pressure of around 500 bar). Online quality control was performed with SprayQc (45), which was extended with an additional plugin to support a high-voltage switch controlling the spray voltage for the analytical columns (RESULTS). MS data were acquired with a Q Exactive Plus (27 min gradients) and a Q Exactive HF (14 min gradients) instrument, as the latter has been found to be up to twice as fast (34) and thus capable of dealing with the fast chromatography of the 14 min gradient. The instruments were programmed with a data-dependent top 5 and top 10 method, respectively, dynamically choosing the most abundant not yet sequenced precursor ions from the survey scans (300–1,650 Th). Instruments were controlled using Tune 2.5 and Xcalibur 3.0.63. At a maximum ion inject time of 45 ms for both instruments, the cycle time was ~800 ms, sufficient for generating a median of 16 data points (14 min) or 25 data points (27 min) over the observed elution peaks (RESULTS). Further settings were chosen according to their previously determined optimal values (34). Sequencing was done with higher-energy collisional dissociation fragmentation with a target value of 1e5 ions determined with predictive automatic gain control, for which the isolation of precursors was performed with a window of 1.4 Th. Survey scans were acquired at a resolution of 70,000 and 60,000, respectively, at *m/z* 200 and the resolution for HCD spectra was set to 17,500 and 15,000, respectively, at *m/z* 200. Normalized collision energy was set to 27 and the "underfill ratio," specifying the minimum percentage of the target ion value likely to be reached at maximum fill time was defined as 10% (27 min) and 40% (14 min). The elevated sequencing threshold ensured that, with the reduced complexity of samples, the fragmentation scans are of higher quality. Furthermore, the S-lens radio frequency level was set to 60, which gave optimal transmission of the *m/z* region occupied by the peptides from our digest (34). We excluded precursor ions with unassigned, single, or five and higher charge states from fragmentation selection.

Data Analysis—All data were analyzed with the MaxQuant proteomics data analysis workflow version 1.4.3.14 (46). The false discovery rate (FDR) cut off was set to 1% for protein and peptide spectrum matches. Peptides were required to have a minimum length of seven amino acids and a maximum mass of 4,600 Da. MaxQuant was used to score fragmentation scans for identification based on a search with an initial allowed mass deviation of the precursor ion of a maximum of 4.5 ppm after time-dependent mass calibration. The allowed fragment mass deviation was 20 ppm. Fragmentation spectra were identified using the UniprotKB *S. cerevisiae* database (based on 2014–07 release; 6,643 entries) combined with 262 common contaminants by the integrated Andromeda search engine (47). Enzyme specificity was set as C-terminal to arginine and lysine, also allowing cleavage before proline, and a maximum of two missed cleavages. Carbamidomethylation of cysteine was set as fixed modification and N-terminal protein acetylation and methionine oxidation as variable modifications. Both "match between runs," with a maximum time difference of 30 s, and label-free quantification (LFQ) with standard settings, were enabled (48). Additional metadata stored in the RAW files (*e.g.* ion inject time, noise level, etc.) were extracted using MSFileReader (Thermo Scientific) with in-house-developed tools.

Further data analysis with the goal of assigning the interactors was performed with the R scripting and statistical environment (49) using ggplot (50) for data visualization. Briefly, LFQ intensity values were base10 logarithmized, resulting in a normal distribution. Missing values were imputed by randomly selecting from a normal distribution centered on the lower edge of the intensity values (for this normal distribution the shift was set to 1.8 standard deviations from the mean and the width to 0.3 standard deviations; see histograms describing placement in Figs. S8 and S9). Proteins were excluded in subsequent steps for baits with less than two valid values in the triplicate for the bait (mostly presented as significantly depleted proteins due to the imputed character of the intensity values). The fold enrichment was calculated as the mean ratio between the bait measurements and the proteome measurements of the parental strain (conforming to the mean used in the consequent *t* test). For the fold enrichment, the standard error of the mean was additionally determined. Permutation-based FDR-controlled *t* test *p* values were calculated for each protein between the bait triplicate and the parental strain triplicate (employing 250 permutations). The *p* value was adjusted using a scaling factor s0 with a value of 1 prior to FDR control, which magnifies the importance of the difference of the mean (51). Furthermore, the correlation of each protein's LFQ intensity profile (consisting of all the measured intensity values for that protein) to the LFQ intensity profile of the bait was calculated (21), and the resulting correlation *p* values were adjusted to 1% FDR using the Benjamini and Hochberg procedure. Interactor classes were assigned based on the following rules: (A) only <1% FDR *t* test significance, (A+) both <1% FDR *t* test significance, and <1% FDR correlation significance, (B+) both <5% FDR *t* test significance and <1% FDR correlation significance and (B) only <5% FDR *t* test significance. Known interactors from the *Saccharomyces* Genome Database (www.yeastgenome.org) mainly fell in classes A+, A, and B+. Therefore, we conducted follow-up analyses solely on these classes. For each significant outlier, we also introduced a single significance value, based on the s0 scaling introduced in the *t* test, which combines the enrichment value and the *t* test statistic. This is calculated as the distance in log-space from the origin. The higher this value, the better the data quality and experimental success of that particular interactor. Stoichiometry information was determined in two ways. The first, termed interaction stoichiometry, is the ratio between the calculated intensity-based absolute quantification values (determining the copy numbers from the acquired mass spectrometry data) of the interactors to the bait (52). The second, termed abundance stoichiometry, is the ratio between the normal cellular copy numbers of the interactors to the bait.

RESULTS

Reducing the LC-MS/MS Analysis Time—First, we aimed to establish optimal conditions for reducing the LC gradient length. Both the flow rate and gradient starting percentage require adaptations to ensure that the signal of each peptide does not degrade and to maximize the spread of peptides over the gradient. To achieve this, we tested the effect of flow rate (ranging from 200 to 500 nl/min) and gradient length (from 15 to 120 min) on the chromatographic peak-width with a standard HeLa digest on the Q Exactive HF (34). By far, the largest effect on peak width was shortening the gradient length as this provided a reduction of ~75% on the width, while the flow rate reduced it only by ~4% (Fig. 1*A*). With regard to overall proteome depth, we were able to identify about 740 proteins with a standard HeLa digest using the shortest gradient length of 15 min with the Q Exactive HF (Fig. 1*B*). Hence, the complexity of protein samples should not exceed such a number when high sample throughput is en-

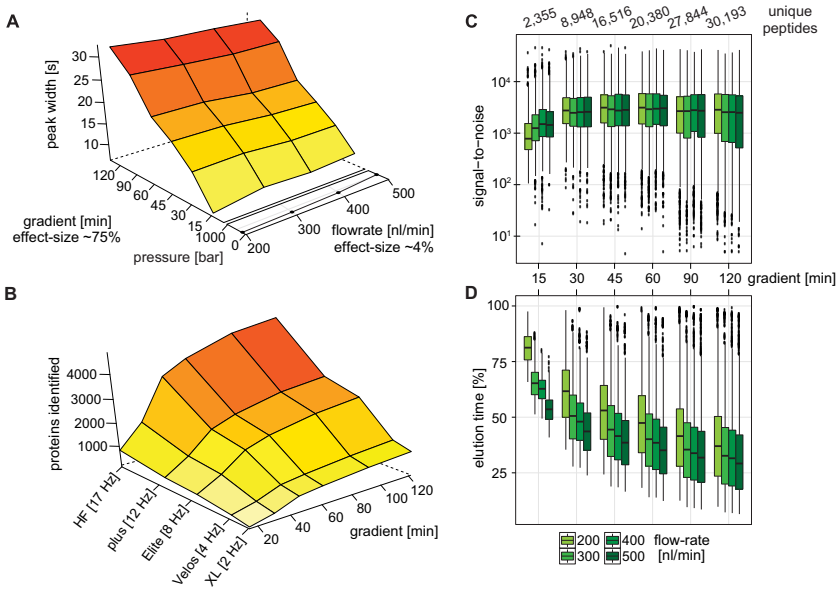


FIG. 1. **Chromatography optimization for very short gradients**. (*A*) Peak-width as a function of gradient length and flow rate. Effect size is the calculation of the reduction compared with the largest change in peak-width. (*B*) Extrapolation of protein identifications as a function of gradient length and scan speed of various MS platforms (Q Exactive HF and plus, Orbitrap Elite, and Velos, LTQ Orbitrap XL, respectively). (*C*) Effect of flow rate on the signal-to-noise for a set of 750 unique isotope patterns identified in all measurements and spread out over the entire gradient. (*D*) Elution time shift induced by higher flow rates, normalized to the gradient length.

visioned. We also determined protein identifications for lower sequencing speed (Fig. 1*B*). Notably, even platforms with lower sequencing speed like the Orbitrap XL identified about 1,000 proteins with a 120 min gradient, suggesting that already this machine generation had the potential to identify all proteins of a lower complexity sample given sufficiently long gradients.

Higher flow rates could have a detrimental effect on the signal-to-noise due to the higher dilution of peptides in the buffer, which we investigated by extracting the signal-to-noise values for a set of 750 isotope patterns identified in all the runs and spread out over the full retention time range. For the longest gradient length of 120 min, we observe a slight decrease in signal-to-noise for the higher flow rates, whereas unexpectedly higher flow rates partially improve the signal-to-noise for the shortest gradient. For the intermediate gradient lengths, the flow rate does not appreciably affect the signal-to-noise ratio. Between the two shortest gradients of 30 and 15 min, we observe a drop in signal-to-noise, which we attribute to imprecision of the buffer delivery by the LC (Fig. 1*C*). Given that it takes time for the buffer mixture to arrive from the mixing T connection to the tip of the analytical column, and therefore for the peptides to elute, the shorter gradients suffer in terms of gradient occupancy (percentage of the gradient occupied by peptides) when using lower flow rates. This is mostly improved by forcing the peptides to elute earlier with higher flow rates. For the shortest gradient lengths, we were able to move the start of peptide elution from 60% in the gradient (at 9 min) to 40% in the gradient (at 6 min), improving the spread of the peptides over the complete gradient and providing better chromatographic resolution. For the 30 min gradient, the first elution was moved from 10 min (35% of the gradient time) to 7 min (25%) (Fig. 1*D*).

Based on these findings, we determined the optimal gradient time to be 27 min with a flow rate of 450 nl/min, which kept the backpressure of the LC pumps at an acceptable level of around 500 bar. This, however, still results in 2 days of measurements for 96 samples. The 12 min gradient at the same flow rate necessary for exactly 24 h of measurement for the same number of samples is expected to have reduced chromatographic performance compared with the 27 min gradient. This period is also too short to transfer the peptides onto the analytical column in parallel. We therefore increased the gradient time to 14 min and activated the loading pump during the intersample preparation time, which reliably loaded all the

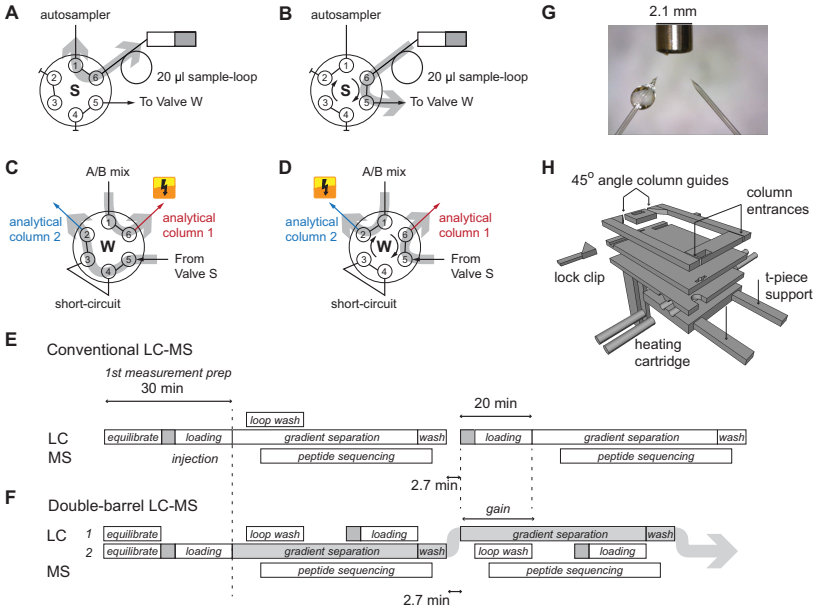


FIG. 2. **Parallel UHPLC operation with two analytical columns.** (*A*) In this position of valve S, the sample pump can fill the sample loop. (*B*) By switching valve S, the contents of the sample loop can be loaded onto one of the analytical columns. (*C*) In this position of valve W, the analytical column 1 can be eluted with the mobile phase, while analytical column 2 is loaded. (*D*) By switching the position of valve W, this behavior is inverted. (*E*) In the conventional setup, the mass spectrometer is not sequencing while the HPLC is loading a new sample. The light gray arrow indicates where the mobile phase is active. (*F*) With the double-barrel setup, this idle-time is circumvented, enabling almost continuous operation. (*G*) Positioning of the analytical columns in reference to the inlet of the mass spectrometer. (*H*) Redesign of the column oven for two analytical columns.

peptides onto the analytical column. Additionally, we increased the starting acetonitrile percentage of the gradient from 2% to 8% (EXPERIMENTAL PROCEDURES) to start the peptide elution at an earlier point of the gradient. Collectively, this resulted in a time frame for peptide elution of 8 min and 18 min, representing 60 and 75%, respectively, of the total measurement time for the 14 and 27 min gradients. At these conditions, the median peak-width (base-to-base) was 6 s (14 min) and 11 s (27 min), respectively.

Double-Barrel Chromatography on the EASY-nLC—Next, we set out to develop a double-barrel chromatography system in order to reduce the idling time of the mass spectrometer during loading of the peptides to the LC column. Unfortunately, no such setup has been described for the Thermo EASY-nLC 1000 UHPLC systems (Thermo Fisher Scientific) that we employ and that are widely used with the Orbitrap-family of mass spectrometers. To address this, we modified the liquid pathway of the EASY-nLC 1000 UHPLC system (Figs. 2*A*-2*D*). In brief, we placed the sample loop directly between the pump S and valve S, allowing the system to utilize pump S as both the sample pickup as well as the sample-loading pump (in the original setup, pump A is used as sample-loading pump). The valve S is connected to valve W (in the original setup this valve is connected to a waste line used for rapid evacuation of the buffers from the lines), which connects to the buffer A and B mixing-T connection and the two analytical columns through standard sample lines. This setup allows loading of one sample onto one of the analytical columns while the other is eluted.

To make use of this new liquid pathway and to drive two analytical columns in parallel, we also modified the "business logic" controlling the UHPLC system. The normally sequential steps in the analysis process (Fig. 2*E*) were altered to work in parallel with each other (Fig. 2*F*). As soon as the preparation for the currently active analytical column has finished, the initiation phase and the valve W has switched to elute the loaded peptides, the inactive analytical column is prepared in parallel for the next sample. This is done in three consecutive steps: First, the sample loop is washed, then the new sample is loaded into the sample loop, and finally the sample is loaded from the sample loop onto the analytical column. With the above described arrangement of the pumps and valves,

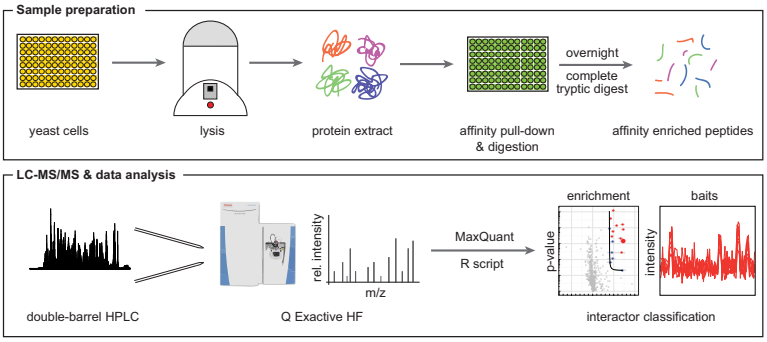


FIG. 3. **Workflow of the high-throughput LC-MS/MS protein interaction analysis pipeline.** Both culturing of yeast cells and affinity purification are performed in 96-well plate format, thus parallelizing sample preparation and minimizing handling errors. LC-MS/MS analysis of 96 pull-down samples in 1 day is achieved through a double-barrel chromatography setup and the increased sequencing speed of the Q Exactive HF mass spectrometer.

these operations can be performed independently for each of the two analytical columns. The intermeasurement time for the double-barrel system was clocked at a maximum of 160 s (Figs. 2*E* and 2*F*), which cannot be further reduced on this particular system due to the necessity of refilling the syringe-based pumps and bringing them back up to pressure (Supplemental Fig. S1).

Finally, we modified our standard analytical column heater (33) to accommodate the two analytical columns. The two columns are now pointing sideways toward the mass spectrometer inlet at a fixed angle of 45 degrees at a distance of roughly 2 mm from each other at the tip ends (equaling the width of the heated capillary mounted on Orbitrap platforms; Fig. 2*G*). As we utilize a fixed setup for the analytical columns, we cannot supply the spray voltage in parallel (Fig. 2*H*). To shift the voltage between the analytical columns, we additionally developed a high-voltage switch capable of supplying electricity to a single analytical column, controllable through a universal serial bus connection (Supplemental Fig. S2). A plugin module that we developed for the SprayQc environment (45) monitors the current position of the valve W and switches the spray voltage to the eluting analytical column according to a user-definable setting.

A Parallel Workflow for Analyzing 96 Pull-Down Samples within a Single Day—A high-throughput platform should be able to prepare samples in a parallelized format and subsequently measure all of them within a very short time period. Here, we developed an analysis pipeline for pull-down samples that is capable of achieving this goal on pull-down samples (Fig. 3). To facilitate a streamlined workflow necessary for achieving high-throughput processing of pull-down samples, we used GFP-tagged yeast strains originating from the yeast GFP clone collection (43). Further improvements were gained by combining both the cultivation of the yeast and the pull-downs in a 96-well format. Each well yields ~50 million yeast cells, equal to 500 μg of protein lysate, which turned out to be sufficient for the pull-down experiments.

Mass Spectrometry Platform Performance on Pull-Down Samples—Using the transcriptional adapter protein ADA2 as a bait, we compared the performance of the Q Exactive HF to that of the LTQ-Orbitrap XL, an instrument introduced about 9 years ago with a sequencing speed of 2 Hz that is frequently used for pull-down analyses. Notably, both instruments were able to identify all known members of the reconstituted ADA2 complex within the commonly used measurement time of 2 h (Supplemental Fig. S3*A*). This suggests that protein interaction data acquired with older Orbitrap generations over the last 10 years would generally gain little by remeasurement as long as extended LC-MS/MS gradients have been used. However, we note that the protein sequence coverage and, consequently, enrichment of the preys (calculated by dividing the MaxLFQ intensity of the interactors by the median of all MaxLFQ intensities) was somewhat improved with the Q Exactive HF, making the setup slightly more sensitive in detecting interactors (Supplemental Fig. S3*B*). Clearly, these gradient times are not making effective use of the superior sequencing speed of the Q Exactive HF. By lowering the measurement time to as low as 15 min, the identification performance of the older platform started to suffer while that of the Q Exactive HF still allowed capturing all the expected interactors (Supplemental Fig. S3*A*). The major difference between the systems was in the sequence coverage per protein, which for the Q Exactive HF remains constant up to 30 min and slightly degrades at 15 min, while it degrades dramatically for the Orbitrap XL (Supplemental Fig. S3*C*). The decreased sequence coverage negatively impacts the ability to accurately quantify proteins as label-free quantification improves with the number of peptides associated to a given protein (48). This is reflected in the measured enrichment ratios, which for the Orbitrap XL made the bait interactors nearly

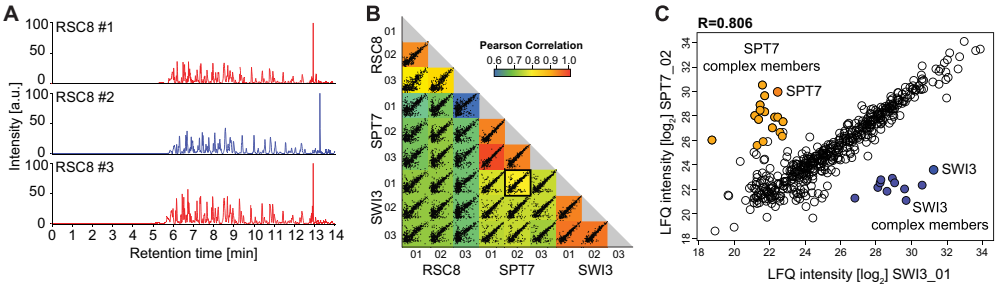


FIG. 4. **Double-barrel chromatography with 14 min gradients on three pull-downs.** (*A*) Base-peak chromatogram of a biological triplicate RSC8 pull-down run on the double-barrel LC-MS/MS setup. Chromatography in all cases is very reproducible. (*B*) Comparison of RSC8, SPT7 and SWI3 pull-downs; all measured in triplicates. The matrix of 36 correlation plots reveals high correlations between MaxLFQ intensities within triplicates. (*C*) Zoom into SPT7_02 *versus* the SWI3_01 correlation plot. While most proteins were detected with very similar MaxLFQ intensities, the two outlier populations marked in orange (SPT7) and blue (SWI3) represent the different complex members of the distinct protein complexes.

indistinguishable from the background, while for the Q Exactive HF it remained superior even at 30 min when comparing to 2 h (Supplemental Fig. S3*B*). Overall, as expected, the Q Exactive HF outperformed the Orbitrap XL for all measurement times tested in terms of prey enrichment, sequence coverage, and isotopic features (Supplemental Fig. S3*B*-S3*D*). While we observed a decrease in obtained sequence information in the 15 min Q Exactive HF methods, these very short runs still yielded sufficiently high sequence coverage to identify the members of the complex under investigation. In conclusion, these results show that mass spectrometers with relatively low sequencing speed can perform equivalently at long gradients for protein interaction studies, whereas very high sequencing speeds are required for high-throughput identification.

Reproducibility of the Data Acquisition System—To investigate the reproducibility of protein quantification between different measurements, we acquired PPI data for the yeast chromatin remodelers RSC8, SPT7, and SWI3 with our workflow. Visual inspection of the chromatograms for the RSC8 pull-down, measured in triplicates, already shows a high degree of technical reproducibility for the double barrel system with back pressure matched analytical columns (Fig. 4*A*). In modern PPI experiments, the number of background binders can be in the thousands as opposed to only a few true interactors. We take advantage of these unspecific binders to estimate reproducibility by calculating the correlation between each pair of the measurements where only the generally small number of true interactors degrade the correlation (21). Most of the detected unspecific binders were indeed reproducibly quantified in all three samples. There was one exception with a slightly reduced Pearson correlation coefficient for the RSC8 pull-down (Fig. 4*B*), for which we concluded based on the large number of imputed values that the enrichment was not completely successful. A small outlier population observed for each bait protein indeed represented the expected interaction partners (Fig. 4*C* and Supplemental Fig. S4). Collectively, these results indicate that our double-barrel setup can be operated with very low MS idling time between two independent measurements and achieves high reproducibility at the same time.

PPI Data Quality from Very Short Gradients—To identify preys of a given bait protein, we classified all interactors into four distinct classes essentially as described (21) and improved on that concept by making it completely data driven (EXPERIMENTAL PROCEDURES). Distinction of specific from unspecific binders was achieved by a permutation-based false-discovery rate approach operating on a *t* test and enrichment with two distinct stringencies (EXPERIMENTAL PROCEDURES; Supplemental Fig. S5*A*). Proteins passing the stringent cutoff represent highly enriched interactors, whereas proteins only passing the less stringent cutoff are characterized as mildly enriched interactors. All other proteins were considered to be unspecific binders. In addition, we used Benjamini–Hochberg-corrected intensity profile correlation of potential interactors compared with the bait protein to minimize false-positive identifications of mildly enriched interactors (EXPERIMENTAL PROCEDURES; Supplemental Figs. S5*E* and S5*F*) (21). With these criteria, interactors were grouped into confidence classes A+, A, B+, and B (Supplemental Figs. S4*C* and S4*G*). Absolute quantification data from whole yeast proteome experiments (53) allowed us to also estimate interaction and abundance stoichiometries for every protein complex under investigation (Supplemental Fig. S5*D*).

To assess the quality control of both the LC-MS/MS measurements and the subsequent interactor classification given our large throughput, we employed three distinct layers. The first layer consists of the real-time validation provided by SprayQc (45). Besides the logic for the voltage switch, this software implements automatic warnings via E-mail to the

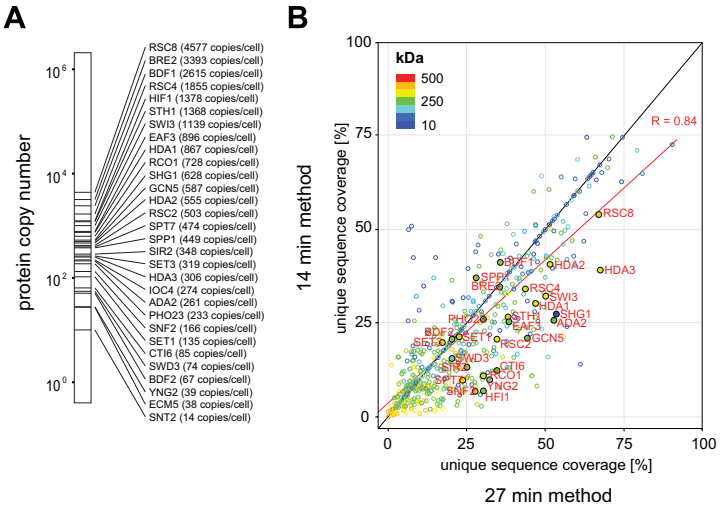


FIG. 5. **Bait and prey characteristics comparing 14 *versus* 27 min gradient methods.** (*A*) The 30 bait proteins selected for the pull-down experiments span several orders of protein expression abundance in *S. cerevisiae*, including several very low abundant proteins (<100 copies per cell). (*B*) Unique sequence coverage for all identified proteins decreases for the 14 min method compared with the 27 min method. Bait proteins are labeled in red.

operator for a large number of components involved in the measurement and reports meta-data for these components (EXPERIMENTAL PROCEDURES). The second layer consists of verification of the sample preparation and LC-MS/MS measurement success by the number of identified proteins per measurement. Given the preponderance of background proteins, this value should be roughly equal for all pull-downs. The histograms displaying the imputed values provide a simple visual guide in the form of the peaks for the imputed proteins (EXPERIMENTAL PROCEDURES). The third layer is the data-driven determination of what constitutes a successful pull-down experiment. For this, we used the information from the volcano plots, specifically the significance value as described (EXPERIMENTAL PROCEDURES). For all the pull-downs, we combine this value for all the baits to determine a valid range for the baits. Anything falling outside this range is flagged as potentially unreliable.

A Snapshot of the S. cerevisiae Chromatin Remodeling Landscape—The data obtained from our very short LC-MS/MS measurements operated with double barrel chromatography demonstrated that AP-MS screens of sufficient quality can be performed in a high-throughput format (Fig. 4). To investigate our workflow on a set of protein complexes involved in a particular biological pathway, we selected 30 distinct bait proteins that are part of the yeast chromatin remodeling landscape. In addition, we also used a GFP-expressing control and the haploid parental strain (EXPERIMENTAL PROCEDURES). Our bait selection spans three orders of expression abundance over the whole yeast proteome (Fig. 5*A*) and includes several baits with very low abundance (<100 copies per cell). We found that the protein input amount of 500 μg, which is much lower than that traditionally used, was sufficient to identify the bait proteins and to retrieve known interactors, even for lowest expressed bait proteins (Supplementary Material_14min and Supplementary Material_27min). Additionally, where possible, we selected multiple baits per protein complex in an attempt to characterize the complex as thoroughly as possible. This collection covers 21 distinct protein complexes subdivided into four enzyme classes: histone acetyltranferase, chromatin remodeling, histone methyltransferase, and histone deacetylase complexes. For the 32 distinct yeast strains, we performed pull-down experiments in biological triplicates, resulting in 96 samples. Each of these pull-down samples was measured with both the 14 and the 27 min LC-MS/MS methods, respectively. Together, the interactomes of 96 pull-down samples were measured in either 47.5 h (27 min method) or 26.7 h (14 min method) of start-to-end complete measurement time, including all overhead. As expected, we found that the sequence coverage of bait proteins and specific interactors was reduced for almost every protein in the 14 compared with the 27 min method (Fig. 5*B*). Nevertheless, the sequence information acquired in the 14 min runs was still sufficient to identify enriched baits and their corresponding preys. We did not experience problems with regard to the bioinformatic enrichment value based on the LFQ intensities, as had been the case for the short gradi-

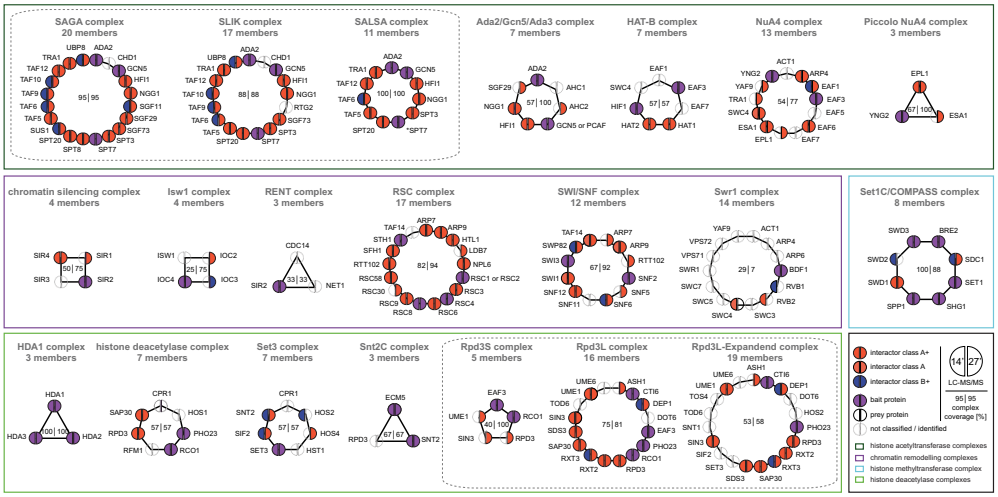


FIG. 6. **Topology network of all interactors.** Global overview of the measured complexes and the success-rate achieved with the 14 *versus* 27 min gradients. Each protein is depicted as a circle, where the left half corresponds to the 14 min and the right half to the 27 min run results. Color coding refers to the different interactor classes and selected bait proteins. Numbers in the center of each complex represent the percentile coverage of the total complex composition as identified by 14 min (left) or 27 min (right) runs. Colored rectangles group the complexes into their distinct biological functionalities.

ents on the older platforms (see Supplemental Figs. S6 and S7, Supplementary Material_14min and Supplementary Material_27min).

To provide an overview of our identified PPIs, we created a topology network of all interactors assigned to one of the defined prey classifications. While the overall interactor class ranking was slightly reduced, we found only small variations in the final complex coverage even though the LC-MS/MS gradient was nearly halved when comparing the 14 to the 27 min method (Fig. 6). Out of 21 protein complexes analyzed, both run times performed equally well in nine cases, whereas the 27 min outperforms the 14 min in ten cases. Conversely, the 14 min runs were better in two cases. The 27 min method allowed a high retrieval of known interactors even for several very low abundant baits with less than 100 copies per cells. While the 14 min method identified less preys of baits with very low abundance, its superior speed allowed throughput of the same sample set in almost half the time.

Remarkably, we could even validate the presence of two different RSC nucleosome-remodeling complexes. The RSC complex is present in two distinct isoforms with distinct roles in the DNA damage response, as defined by the presence of either RSC1 or RSC2 (54, 55). While performing pull-downs on either RSC4 or RSC8, we identified both RSC1 and RSC2 as interactors, demonstrating that RSC4 and RSC8 are part of both RSC complex isoforms (Supplementary Material_27min). In contrast, pulling down RSC2 only resulted in RSC2 but not RSC1 as complex members. These results demonstrate that our workflow is capable of identifying distinct complex compositions in a rapid manner.

Discussion and Outlook—In this study, we have described advances for analyzing up to 96 proteomes with lower complexity in about 1 day of LC-MS/MS data acquisition, including all overhead. Our interaction workflow employs parallelized sample generation in a 96-well format together with a modified LC setup and mass spectrometers with very high sequencing speed. With this combination, we demonstrated a severalfold increase in sample processing throughput and sensitivity, as well as in the LC-MS duty cycle.

Including the preceding yeast cultivation and sample preparation steps, processing of 96 pull-down experiments can be achieved within 48 h. However, several 96 samples could be handled in parallel, allowing nesting upstream sample preparation and downstream LC-MS/MS analysis. This in principle would allow a sustained workflow with a capacity of 96 distinct samples per day. The data presented here were acquired following manual sample preparation. However, the majority of sample preparation steps in our workflow only require liquid handling and are thus easily automated using robotic sample preparation systems.

LC-MS/MS data acquisition within 14 min per sample pushes both the LC and MS systems to their current limits. Consequently, the 14 min runs yielded reduced chromatographic quality compared with the 27 min runs. Although this was still sufficient to yield almost the same complex coverage, the 14 min runs did result in lower sequence coverage for both

bait and prey proteins (Fig 5*B*). This adversely affects analyses and more importantly reduces the enrichment values, making it harder to pinpoint interactors (Fig 2B). Potential optimization could be obtained in an improved experimental design. In this study, we focused on the reproducibility of the complete workflow and chose to perform all steps and measurements in a consecutive series of steps. However, randomizing the measurements, while ensuring that all the replicates of one particular pull-down are always run on the same column, should further improve higher data quality and statistical significance for the interaction determination.

The implementation of double-barrel systems opens up interesting possibilities. On the technological side, it enables automatic detection of a break down in one of the columns due to clogging and reacting to this by using the other column, instead of stopping further analysis. To detect this situation, the software tracks the amount of pressure during the gradient and the flow rate achieved during loading. When the pressure during the gradient or the flow rate during loading exceed critical parameters the system automatically stops operations on this particular column. Further operation is then continued as a single-barrel system. This simple mechanism has the potential to drastically extend the effective up-time and enable almost 24/7 operation of the mass spectrometer. A second technological possibility is the automatic determination of the optimal time for sample loading. The flow rate achieved during loading of the previous sample on the particular analytical column can be used to estimate the required loading time for the current sample. The software then automatically determines the delay required before loading the sample, for instance with a 10 min overhead to ensure that the sample is completely loaded irrespective of fluctuations in the flow rate. This is particularly important for double-barrel-based LC setups as during long gradients it is conceivable that it would be detrimental for the sample to be loaded at the start of the gradient of the other analytical column and then remain at the elevated temperature conditions of the analytical column heater. Third, the described setup could be further extended by using two completely independent UHPLC systems. Even though such a concept is not straightforward to implement on our current system due to software-related issues, the extra redundancy of hardware components would enable troubleshooting of an erroneous UHPLC while the other system maintains measuring. In this way, genuine 24/7 operation of LC-MS/MS data acquisition would be feasible.

Recently, we have reported a high-performance affinity enrichment-mass spectrometry method (21) that uses accurate quantitation of background and unspecific binders for retrieval of true protein complexes. We propose to combine both strategies to allow both the confident retrieval of binding partners and a high throughput. This should be a powerful strategy, especially when a high sequence coverage is not essential (56). Moreover, our results also show that AP-MS can be performed with protein input amounts as low as 500 μg per pull-down and probably much lower in the future, which is considerably less than previously described (21, 42). This increase in sensitivity strongly promotes parallelization and thus throughput efforts. Currently, our pipeline permits a maximum throughput of 96 samples in about 1 day. Employment of other quantification strategies with higher multiplexing, such as TMT labeling for instance, would drastically increase throughput even further.

While we have demonstrated the workflow for protein–protein interactions, our pipeline is generic and can be extended to any kind of protein-based interaction studies in which there is an effective immobilization of the bait material as affinity matrix. We envision other baits such as peptides, DNA, RNA, lipids, or small molecules will greatly facilitate large-scale screening and elucidate drug targets, changes in protein complex formation upon perturbation, and the intertwined relationship between proteins and DNA or RNA.

Finally, the advances described here for the LC-MS/MS part of the workflow can also be extended to the analysis of whole proteomes. For example, biochemical fractionation of whole cell lysates is a routine procedure in mass-spectrometry-based proteomics as it enables much deeper characterization (57, 58). The concomitant increase in LC-MS/MS measurement time caused by the larger number of fractions could be mitigated by using our optimized LC-MS/MS setup. Here, we demonstrated that our very short gradients of 15 min are still able to identify about 700 proteins in a standard HeLa digest (Fig. 1*B*). If such a complexity is not exceeded, high-throughput analysis can be performed even for fractionated whole proteomes of cell lines, small model organisms, or clinical samples. Finally, given the exponential progress in proteomics related technology, it should only be a matter of time until entire proteomes can be measured in minutes.

Acknowledgments—We thank our colleagues at the Max Planck Institute, especially Marco Hein for help and fruitful discussions; Christian Schmid, Martin Wied, Daniel Vik, and Gabriele Sowa for technical assistance; Jochen Rech for providing the GFP yeast strains; scientists at Thermo Scientific, especially Ole Vorm and Ole Hoerning; and Georg Völkle and Wolfgang Schrader at Sonation GmbH. Last, but not least, we thank Skunkworks for inspiring a highly effective collaborative working model.

* The research leading to these results has received funding from the European Commission's 7th Framework Programme (grant agreement HEALTH-F4-2008-201648/PROSPECTS).

§ To whom correspondence should be addressed: mmann@biochem.mpg.de.

[S] This article contains supplemental material Figs. S1 to S9.

‖ These authors contributed equally.

¶ Current address: Cellzome GmbH, Molecular Discovery Research, GlaxoSmithKline, Meyerhofstraβe 1, D-69117 Heidelberg, Germany.

Data availability: Supplementary data is available with this publication at the MCP web site. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (59) (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (60) with the dataset identifier PXD001695.

REFERENCES

1. Wolters, D. A., Washburn, M. P., and Yates, J. R., 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690
2. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
3. Altelaar, A. F., Munoz, J., and Heck, A. J. (2013) Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nature Rev. Genetics* **14,** 35–48
4. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 548
5. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011) The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549
6. Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008) Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 6959–6964
7. Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A. S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabási, A. L., and Vidal, M. (2009) An empirical framework for binary interactome mapping. *Nature Meth.* **6,** 83–90
8. Fields, S., and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340,** 245–246
9. Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. *Nature Rev. Mol. Cell Biol.* **8,** 645–654
10. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183
11. Gavin, A. C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147
12. Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433,** 531–537
13. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636
14. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440,** 637–643
15. Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J. P., Duewel, H. S., Stewart, II, Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S. L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3,** 89
16. Guruharsha, K. G., Rual, J. F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D. Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K. H., Yu, C., Parsa, B., Carlson, J. W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S. P., Celniker, S. E., Obar, R. A., and Artavanis-Tsakonas, S. (2011) A protein complex network of *Drosophila melanogaster*. *Cell* **147,** 690–703
17. Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., Kim, B. J., Li, C., Chen, R., Li, W., Wang, Y., O'Malley, B. W., and Qin, J. (2011) Analysis of the human endogenous coregulator complexome. *Cell* **145,** 787–799
18. Gavin, A. C., Maeda, K., and Kühner, S. (2011) Recent advances in charting protein-protein interaction: Mass spectrometry-based approaches. *Curr. Opin. Biotechnol.* **22,** 42–49
19. Vermeulen, M., Hubner, N. C., and Mann, M. (2008) High confidence determination of specific protein–protein interactions using quantitative mass spectrometry. *Curr. Opin. Biotechnol.* **19,** 331–337
20. Paul, F. E., Hosp, F., and Selbach, M. (2011) Analyzing protein–protein interactions by quantitative mass spectrometry. *Methods* **54,** 387–395
21. Keilhauer, E. C., Hein, M. Y., and Mann, M. (2014) Accurate protein complex retrieval by affinity enrichment MS rather than affinity purification MS. *Mol. Cell. Proteomics*
22. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nature Biotechnol.* **21,** 315–318
23. Bantscheff, M., Eberhard, D., Abraham, Y., Bastuck, S., Boesche, M., Hobson, S., Mathieson, T., Perrin, J., Raida, M., Rau, C., Reader, V., Sweetman, G., Bauer, A., Bouwmeester, T., Hopf, C., Kruse, U., Neubauer, G., Ramsden, N., Rick, J., Kuster, B., and Drewes, G. (2007) Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotechnol.* **25,** 1035–1044
24. Rinner, O., Mueller, L. N., Hubálek, M., Muller, M., Gstaiger, M., and Aebersold, R. (2007) An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nature Biotechnol.* **25,** 345–352
25. Mousson, F., Kolkman, A., Pijnappel, W. W., Timmers, H. T., and Heck, A. J. (2008) Quantitative proteomics reveals regulation of dynamic components within TATA-binding protein (TBP) transcription complexes. *Mol. Cell. Proteomics* **7,** 845–852
26. Selbach, M., Paul, F. E., Brandt, S., Guye, P., Daumke, O., Backert, S., Dehio, C., and Mann, M. (2009) Host cell interactome of tyrosine-phosphorylated bacterial proteins. *Cell Host Microbe* **5,** 397–403
27. Vermeulen, M., Eberl, H. C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K. K., Olsen, J. V., Hyman, A. A., Stunnenberg, H. G., and Mann, M. (2010) Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142,** 967–980
28. McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., Kuhn, K., Pike, I., Grothe, R. A., Blethrow, J. D., and Gygi, S. P. (2012) Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* **84,** 7469–7478
29. Werner, T., Becher, I., Sweetman, G., Doce, C., Savitski, M. M., and Bantscheff, M. (2012) High-resolution enabled TMT 8-plexing. *Anal. Chem.* **84,** 7188–7194
30. Baker, E. S., Livesay, E. A., Orton, D. J., Moore, R. J., Danielson, W. F., 3rd, Prior, D. C., Ibrahim, Y. M., LaMarche, B. L., Mayampurath, A. M., Schepmoes, A. A., Hopkins, D. F., Tang, K., Smith, R. D., and Belov, M. E. (2010) An LC-IMS-MS platform providing increased dynamic range

for high-throughput proteomic studies. *J. Proteome Res.* **9,** 997–1006
31. Falkenby, L. G., Such-Sanmartin, G., Larsen, M. R., Vorm, O., Bache, N., and Jensen, O. N. (2014) Integrated solid-phase extraction-capillary liquid chromatography (speLC) interfaced to ESI-MS/MS for fast characterization and quantification of protein and proteomes. *J. Proteome Res.* **13,** 6169–6175
32. Binai, N. A., Marino, F., Soendergaard, P., Bache, N., Mohammed, S., and Heck, A. J. (2015) Rapid analyses of proteomes and interactomes using an integrated solid-phase extraction-liquid chromatography-MS/MS system. *J. Proteome Res.* **14,** 977–985
33. Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J., and Mann, M. (2011) Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics* **10,** M110.003699
34. Scheltema, R. A., Hauschild, J. P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., and Mann, M. (2014) The Q Exactive HF, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell. Proteomics* **13,** 3698–3708
35. Shen, Y., Tolić, N., Zhao, R., Pasa-Tolić, L., Li, L., Berger, S. J., Harkewicz, R., Anderson, G. A., Belov, M. E., and Smith, R. D. (2001) High-throughput proteomics using high-efficiency multiple-capillary liquid chromatography with on-line high-performance ESI FTICR mass spectrometry. *Anal. Chem.* **73,** 3011–3021
36. Belov, M. E., Anderson, G. A., Wingerd, M. A., Udseth, H. R., Tang, K., Prior, D. C., Swanson, K. R., Buschbach, M. A., Strittmatter, E. F., Moore, R. J., and Smith, R. D. (2004) An automated high performance capillary liquid chromatography-Fourier transform ion cyclotron resonance mass spectrometer for high-throughput proteomics. *J. Amer. Soc. Mass Spec.* **15,** 212–232
37. Bonneil, E., Tessier, S., Carrier, A., and Thibault, P. (2005) Multiplex multidimensional nanoLC-MS system for targeted proteomic analyses. *Electrophoresis* **26,** 4575–4589
38. Livesay, E. A., Tang, K., Taylor, B. K., Buschbach, M. A., Hopkins, D. F., LaMarche, B. L., Zhao, R., Shen, Y., Orton, D. J., Moore, R. J., Kelly, R. T., Udseth, H. R., and Smith, R. D. (2008) Fully automated four-column capillary LC-MS system for maximizing throughput in proteomic analyses. *Anal. Chem.* **80,** 294–302
39. Orton, D. J., Wall, M. J., and Doucette, A. A. (2013) Dual LC-MS platform for high-throughput proteome analysis. *J. Proteome Res.* **12,** 5963–5970
40. Behrends, C., Sowa, M. E., Gygi, S. P., and Harper, J. W. (2010) Network organization of the human autophagy system. *Nature* **466,** 68–76
41. Collins, B. C., Gillet, L. C., Rosenberger, G., Röst, H. L., Vichalkovski, A., Gstaiger, M., and Aebersold, R. (2013) Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14–3-3 system. *Nature Meth.* **10,** 1246–1253
42. Poulsen, J. W., Madsen, C. T., Young, C., Poulsen, F. M., and Nielsen, M. L. (2013) Using guanidine-hydrochloride for fast and efficient protein digestion and single-step affinity-purification mass spectrometry. *J. Proteome Res.* **12,** 1020–1030
43. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691
44. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
45. Scheltema, R. A., and Mann, M. (2012) SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11,** 3458–3466
46. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26,** 1367–1372
47. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
48. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13,** 2513–2526
49. Ihaka, R., and Gentleman, R. (1996) R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5,** 299–314
50. Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.
51. Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 5116–5121
52. Schwänhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342
53. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Meth.* **11,** 319–324
54. Cairns, B. R., Schlichter, A., Erdjument-Bromage, H., Tempst, P., Kornberg, R. D., and Winston, F. (1999) Two functionally distinct forms of the RSC nucleosome-remodeling complex, containing essential AT hook, BAH, and bromodomains. *Molecular Cell* **4,** 715–723
55. Chambers, A. L., Brownlee, P. M., Durley, S. C., Beacham, T., Kent, N. A., and Downs, J. A. (2012) The two different isoforms of the RSC chromatin remodeling complex play distinct roles in DNA damage responses. *PloS One* **7,** e32016
56. Ong, S. E., Li, X., Schenone, M., Schreiber, S. L., and Carr, S. A. (2012) Identifying cellular targets of small-molecule probes and drugs with biochemical enrichment and SILAC. *Meth. Mol. Biol.* **803,** 129–140
57. Yang, F., Shen, Y., Camp, D. G., 2nd, and Smith, R. D. (2012) High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev. Proteomics* **9,** 129–134
58. Kelstrup, C. D., Jersie-Christensen, R. R., Batth, T. S., Arrey, T. N., Kuehn, A., Kellmann, M., and Olsen, J. V. (2014) Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J. Proteome Res.* **13,** 6187–6195
59. Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P. A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H. J., Albar, J. P., Martinez-Bartolomé, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnol.* **32,** 223–226
60. Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–1069

2.1.3 Identifying specific interaction partners of humane histone H2A variants

Vardabasso, C., Gaspar-Maia, A.,Pünzeler, S., Valle-Garcia, D., Hasson, D., Straub, T., **Keilhauer, E. C.**, Strub, T., Panda, T., Segura, M., Chung, C., Verma, A., Mann, M., Hernando, E., Hake, S. B. & Bernstein, E.

Histone variant H2A.Z.2 mediates proliferation and drug sensitivity of malignant melanoma
Molecular Cell Accepted May 2015

In a very fruitful collaboration project with Sebastian Pünzeler from the group of Dr. Sandra Hake, we set out to identify interaction partners of human histone H2A variants. For many of the canonical core histone proteins including H2A, certain low abundant variants exist. These variants show specific expression and chromatin localization, modify the properties of the nucleosomes they are incorporated in, and affect transcription. However, often little is known on how their distinct localization and function is achieved. We hypothesized that the H2A variants attract different interactions partners than the canonical histones.
Sebastian Pünzeler was specifically interested in three variants of histone H2A, called H2A.Z.1, H2A.Z.2.1 and H2A.Z.2.2. He had already established cell lines expressing GFP-tagged versions of canonical H2A and the three variants. To investigate the properties of the variants in their 'natural context', he prepared nuclear extracts from these tagged cells and digested the chromatin into mononucleosomes using micrococcal nuclease. Together, we then implemented our MS-compatible immunoprecipitation protocol in the laboratory of Dr. Sandra Hake, to enrich whole mononucleosomes containing either canonical H2A or the three variants, and to identify their interaction partners by label-free quantitative mass spectrometry.
We first performed such pulldowns in HeLa cells, where we could identify many of the known general H2A interactors, but also intriguing new candidates specific for the variants. One particular interesting candidate is now further validated and investigated by Sebastian Pünzeler (manuscript in preparation).
Contributing to a project of Dr. Chiara Vardabasso from the group of Prof. Emily Bernstein, Sebastian Pünzeler and I then also performed pulldowns of H2A variants in melanoma cells, which is the work presented in the following. The aim of Dr. Chiara Vardabasso's project was to elucidate the role of the variants H2A.Z.1 and H2A.Z.2 in the

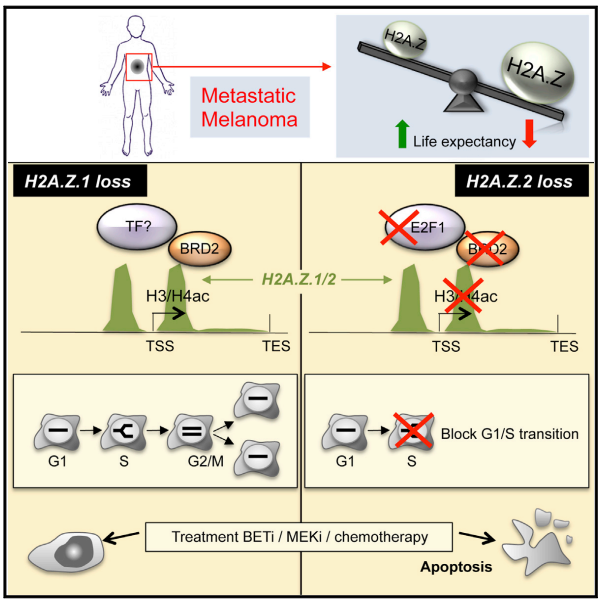
context of metastatic melanoma. Next to many other interesting findings in this publication, our interaction analysis identified several melanoma-specific H2A.Z interactors. The most interesting of these interactors was Brd2, a protein known to interact with the transcription factor E2F1, which in turn controls the expression of a number of genes involved in cell cycle regulation. Our discovery of Brd2 as a specific H2A.Z interactor in melanoma cells was validated using complementary approaches. Although Brd2 was identified as an interactor of both H2A.Z.1 and H2A.Z.2 in our pulldown experiments, only H2A.Z.2 knockdown reduced Brd2 levels in melanoma cells. Therefore, Brd2 was proposed as crucial component of an H2A.Z.2-Brd2-E2F1 axis driving melanoma progression, and as a potential key target for melanoma therapy.

Article

Molecular Cell

Histone Variant H2A.Z.2 Mediates Proliferation and Drug Sensitivity of Malignant Melanoma

Graphical Abstract



Highlights

- High levels of H2A.Z isoforms in metastatic melanoma correlate with poor survival
- H2A.Z.2 promotes expression of E2F targets that display unique H2A.Z occupancy
- BRD2 and E2F1 bind E2F targets in an H2A.Z.2-dependent manner
- H2A.Z.2 silencing sensitizes melanoma cells to chemo- and targeted therapies

Authors

Chiara Vardabasso, Alexandre Gaspar-Maia, Dan Hasson, ..., Eva Hernando, Sandra B. Hake, Emily Bernstein

Correspondence

sandra.hake@med.uni-muenchen.de (S.B.H.), emily.bernstein@mssm.edu (E.B.)

In Brief

Vardabasso et al. establish a role for the histone variant H2A.Z.2 as a driver of malignant melanoma. H2A.Z.2 promotes cell proliferation by regulating expression of E2F targets, which are bound by BRD2 and E2F1 in an H2A.Z.2-dependent manner. High levels of H2A.Z.2 correlate with decreased survival, and its depletion sensitizes cells to therapy.

Accession Numbers

GSE59060

Vardabasso et al., 2015, Molecular Cell 59, 75–88
July 2, 2015
http://dx.doi.org/10.1016/j.molcel.2015.05.009

CellPress

Histone Variant H2A.Z.2 Mediates Proliferation and Drug Sensitivity of Malignant Melanoma

Chiara Vardabasso,[1,2] Alexandre Gaspar-Maia,[1,9] Dan Hasson,[1,2,9] Sebastian Pünzeler,[4,9] David Valle-Garcia,[1,5,9] Tobias Straub,[4] Eva C. Keilhauer,[6] Thomas Strub,[1,2] Joanna Dong,[1,2] Taniya Panda,[1] Chi-Yeh Chung,[1] Jonathan L. Yao,[2,3] Rajendra Singh,[2,3] Miguel F. Segura,[7,10] Barbara Fontanals-Cirera,[7] Amit Verma,[8] Matthias Mann,[6] Eva Hernando,[7] Sandra B. Hake,[4,*] and Emily Bernstein[1,2,*]
[1]Department of Oncological Sciences
[2]Department of Dermatology
[3]Department of Pathology
Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[4]Center for Integrated Protein Science Munich and Department of Molecular Biology, Adolf-Butenandt Institute, Ludwig-Maximilians University, 80336 Munich, Germany
[5]Molecular Genetics Department, Institute for Cellular Physiology, National Autonomous University of Mexico, 04510 Mexico City, Mexico
[6]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[7]Department of Pathology and Interdisciplinary Melanoma Cooperative Group, New York University Langone Medical Center, New York, NY 10016, USA
[8]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[9]These authors contributed equally to this work
[10]Present address: Vall d'Hebron Institut de Recerca (VHIR), 08035 Barcelona, Spain
*Correspondence: sandra.hake@med.uni-muenchen.de (S.B.H.), emily.bernstein@mssm.edu (E.B.)
http://dx.doi.org/10.1016/j.molcel.2015.05.009

SUMMARY

Histone variants are emerging as key regulatory molecules in cancer. We report a unique role for the H2A.Z isoform H2A.Z.2 as a driver of malignant melanoma. H2A.Z.2 is highly expressed in metastatic melanoma, correlates with decreased patient survival, and is required for cellular proliferation. Our integrated genomic analyses reveal that H2A.Z.2 controls the transcriptional output of E2F target genes in melanoma cells. These genes are highly expressed and display a distinct signature of H2A.Z occupancy. We identify BRD2 as an H2A.Z-interacting protein, levels of which are also elevated in melanoma. We further demonstrate that H2A.Z.2-regulated genes are bound by BRD2 and E2F1 in an H2A.Z.2-dependent manner. Importantly, H2A.Z.2 deficiency sensitizes melanoma cells to chemotherapy and targeted therapies. Collectively, our findings implicate H2A.Z.2 as a mediator of cell proliferation and drug sensitivity in malignant melanoma, holding translational potential for novel therapeutic strategies.

INTRODUCTION

Malignant melanoma is the most lethal form of skin cancer, has an increasing incidence, and remains largely incurable. Whereas advances in immune and targeted therapies have made tremendous progress recently (Chapman et al., 2011; Kaufman et al., 2013), they are effective only in distinct subsets of patients or result in the emergence of drug resistance (Lito et al., 2013). Thus, investigation of alternative approaches is essential.

Recent studies have shed light on the importance of epigenetic regulation in melanoma biology. Key roles for BRD4 (Segura et al., 2013), histone methyltransferases SETDB1 (Ceol et al., 2011) and EZH2 (Zingg et al., 2015), and the histone variant macroH2A (Kapoor et al., 2010) have been reported. Relevant to the present study, histone variants and their chaperones are emerging as key regulatory molecules in cancer (Vardabasso et al., 2013).

H2A.Z is a highly conserved H2A variant, with only 60% identity to canonical H2A, and is expressed and incorporated into chromatin throughout the cell cycle (Bönisch and Hake, 2012). Although somewhat confounded by species-specific functions and context-dependencies, the role of H2A.Z in transcriptional regulation is well established (Svotelis et al., 2009). H2A.Z is enriched at gene promoters, as well as other regulatory regions, generally exerting a positive role on transcription (Hu et al., 2013; Obri et al., 2014).

Two distinct H2A.Z isoforms, H2A.Z.1 and H2A.Z.2, have been identified in the vertebrate genome as products of two non-allelic genes, *H2AFZ* and *H2AFV*, respectively (Dryhurst et al., 2009; Horikoshi et al., 2013; Matsuda et al., 2010). While differing by only three amino acids at the protein level, H2A.Z.1 and H2A.Z.2 are encoded by distinct nucleotide sequences. Isoform-specific functions remain unclear, and H2A.Z.1 mouse knockout studies suggest that the two genes are non-redundant (Faast et al., 2001). In the context of tumorigenesis, H2A.Z is overexpressed in breast, prostate, and bladder cancers, where, in some cases, it regulates proliferation (reviewed in Vardabasso et al., 2013). However, these studies either focused solely on H2A.Z.1, or did not clearly distinguish between isoforms.

Here we report a distinct role for H2A.Z.2 in melanoma. H2A.Z.2 is highly expressed in melanoma and drives proliferation by promoting expression of E2F target genes. These cell cycle regulatory genes are highly expressed and acquire a unique signature of H2A.Z occupancy—high promoter enrichment and gene body depletion. We further identified the BET (bromodomain and extraterminal domain) protein BRD2 as an H2A.Z interacting protein, whose levels are also elevated in melanoma specimens. Depletion of H2A.Z.2 results in reduced histone acetylation, BRD2 and E2F1 levels, and impairs recruitment of BRD2 and E2F1 to its target genes. Moreover, H2A.Z.2 deficiency cooperates with BET or MEK inhibition to induce melanoma cell death. Hence, our studies suggest that targeting H2A.Z deposition may be effective therapeutically in combination with existing or emerging therapies for melanoma.

RESULTS

H2A.Z Isoforms Are Overexpressed in Melanoma

By probing a panel of primary and metastatic melanoma cell lines, we detected increased levels of H2A.Z protein in metastatic cells (Figure 1A). Immunoblotting of histones extracted from benign nevi and melanoma specimens revealed increased H2A.Z in melanoma tissues (Figure 1B). We also investigated H2A.Z levels in human primary melanocytes induced to senesce via serial passaging (replicative senescence) and $BRAF^{V600E}$ (oncogene-induced senescence) (Duarte et al., 2014). We observed diminished H2A.Z upon both modes of senescence (Figure S1A). Together, these data link global levels of H2A.Z to cellular proliferation.

To assess whether *H2A.Z* expression is regulated transcriptionally, as well as to examine the individual *H2A.Z* isoforms (which is not possible with currently available antibodies), we performed quantitative RT-PCR (qRT-PCR) using isoform-specific primers. *H2A.Z* isoforms are decreased in human melanocytes induced to senescence (Figure S1A). Conversely, in a panel of benign nevi and melanoma specimens, we observed increased *H2A.Z.1* and *H2A.Z.2* mRNA in melanoma (Figure 1C). *H2A.Z.1* and *H2A.Z.2* mRNA levels were also increased in cell lines derived from metastatic versus primary melanoma (Figure S1B). Analysis of published transcriptional data is consistent with these findings (Talantov et al., 2005; Riker et al., 2008; Xu et al., 2008) (Figure S1C). Finally, in a cohort of patients followed clinically for 3 years after excision of metastatic lesions (Bogunovic et al., 2009), patients with high *H2A.Z.1* and *H2A.Z.2* showed significantly lower survival (Figure 1D). Collectively, these findings suggest that *H2A.Z* isoforms have a functional role in melanoma progression.

We next performed quantitative copy number analysis of *H2A.Z.1* and *H2A.Z.2* in nevi and metastases by qPCR and detected copy gains for both (Figure 1E). The Cancer Genome Atlas (TCGA) reports increased copy number in 13% and 52% of cutaneous melanomas for *H2A.Z.1* and *H2A.Z.2*, respectively, which correlates with increased mRNA levels (Figure S1D). Fluorescence in situ hybridization (FISH) of melanoma cell lines corroborated these findings (Figure S1E).

H2A.Z.2 Depletion Induces G1/S Arrest in Melanoma Cells

Next, we investigated the functional consequences of depleting H2A.Z.1 and H2A.Z.2 in melanoma cell lines. Using sequence-specific shRNAs for H2A.Z isoforms, we established stable SK-mel147 ($NRAS^{Q61R}$), WM266-4 ($BRAF^{V600D}$), and 501mel ($BRAF^{V600E}$) cell lines targeting either H2A.Z.1 or H2A.Z.2. Knockdown was monitored by qRT-PCR and/or immunoblot (Figures S2A–S2C). As H2A.Z.1 is the predominant isoform in melanoma (via RNA sequencing, below), its knockdown can be appreciated at the protein level, whereas H2A.Z.2 knockdown is obscured by H2A.Z.1 (Figures S2A and S2B).

We observed that loss of H2A.Z.2, but not H2A.Z.1, reduced proliferation in all cell lines (Figures 2A, 2B; Figure S2D). To confirm these variant-specific effects, we generated cells stably expressing H2A.Z.1 or an shRNA-resistant H2A.Z.2 that were infected with sh_Z.2 and a control (sh_scr). Only those cells expressing an shRNA-resistant H2A.Z.2 were able to overcome the proliferation defect induced by sh_Z.2 (Figure 2C). Interestingly, HeLa cells depleted of H2A.Z isoforms did not show proliferation defects (Figure S2E). Thus, H2A.Z isoforms exert distinct and non-redundant functions in melanoma cells.

H2A.Z.2 knockdown induced a G1/S cell cycle arrest (Figures 2D, 2E, S2F, and S2G), accompanied by hypophosphorylation of Rb and decreased levels of cyclins E and A (Figure 2F). This phenotype was not consistent with cellular senescence because the expression of cyclin-dependent kinase (CDK) inhibitors (Figure S2H) and β-galactosidase activity (data not shown) were not increased. Moreover, we observed minimal cell death (Figure S2I). Next, SK-mel147 cells were arrested at early S phase via double thymidine block and subsequently released. Both control and H2A.Z.1 knockdown cells progressed through S and G2/M phases, and at 10 hr, ~30% of the cells re-entered G1. However, H2A.Z.2-depleted cells remained largely arrested for the entire duration of the assay (Figure 2G). These findings suggest that H2A.Z.2 loss causes delayed entry into S phase. These data are strikingly similar to *htz1* (H2A.Z) mutant budding yeast, which shows delayed DNA replication and cell cycle progression (Dhillon et al., 2006).

H2A.Z.2 Regulates E2F Target Genes

To further understand the observed proliferation defect, we characterized the transcriptional profile of H2A.Z.2-deficient cells. We used Affymetrix microarrays for SK-mel147 and WM266-4 cells depleted of either H2A.Z.1 or H2A.Z.2 (Figure 3A; Table S1). Interestingly, the majority of genes were downregulated (Figures 3A and 3B), with only 35 overlapping genes between H2A.Z.1 and H2A.Z.2 knockdown in SK-mel147 cells (Figure 3B). Similar expression data were observed for WM266-4 cells (Figures S3A and S3B; Table S1).

Consistent with the observed phenotype, functional annotation revealed that H2A.Z.2-regulated genes are enriched for cell cycle regulators (Figures 3C, 3D and S3C). This is in contrast to H2A.Z.1-regulated genes, which are enriched for immunological pathways (Figure S3D). This is in line with the lack of cell

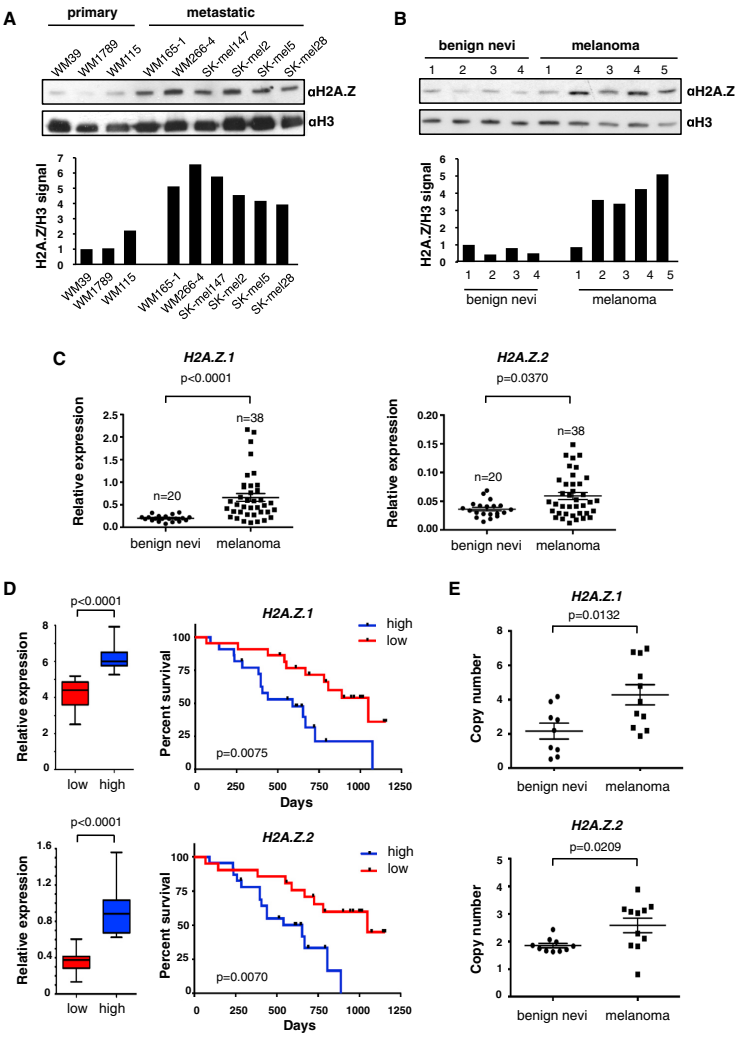


Figure 1. H2A.Z.1 and H2A.Z.2 Are Overexpressed in Melanoma

(A) Chromatin extracted from primary and metastatic cell lines probed with H2A.Z antibody; H3 used for loading. Signals quantified by densitometry. See also Figure S1B for mRNA expression.

(B) H2A.Z immunoblot of acid extracted histones from fresh-frozen human benign nevi and melanoma specimens; H3 used for loading. Signals quantified as in (A).

(C) Expression analysis by qRT-PCR of *H2A.Z.1* and *H2A.Z.2* in benign nevi (n = 20) and melanoma (n = 38). Values normalized to GAPDH; mean ± SEM. Mann-Whitney test (two-tailed).

(D) Survival of melanoma patients with high and low (above or below the median, respectively) mRNA levels of *H2A.Z.1* and *H2A.Z.2*. Gene expression data of 44 metastatic melanoma tissues (Bogunovic et al., 2009) were used to define high and low expressor groups (boxplots, Mann-Whitney test) and to generate Kaplan-Meier curves (log-rank test).

(E) Analysis of *H2A.Z.1* and *H2A.Z.2* gene copy number by qPCR of a subset of benign nevi and melanoma in (C), relative to primary melanocytes. Data are mean ± SEM; unpaired Student's test (two-tailed).

See also Figures S1D and S1E.

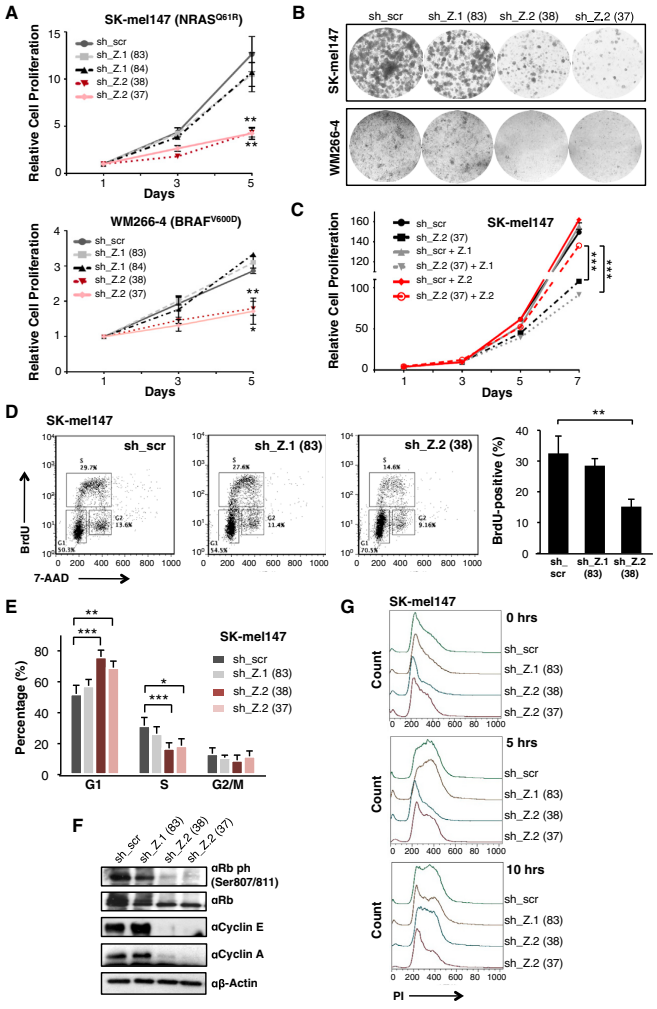


Figure 2. H2A.Z.2 Depletion Induces G1/S Arrest in Melanoma Cells

(A) Proliferation curves of SK-mel147 and WM266-4 cells expressing control and isoform-specific shRNAs as shown. Data are mean ± SEM (n ≥ 3); two-way ANOVA. See also Figure S2D.

(B) Colony formation assays of SK-mel147 and WM266-4 cells expressing shRNAs as in (A).

(C) Proliferation assay of SK-mel147 cells expressing H2A.Z.1, shRNA-resistant H2A.Z.2, or empty vector control. Each line was infected with H2A.Z.2 shRNA (sh_37) and with sh_scr. Data are mean ± SEM (n = 2); two-way ANOVA.

(D) BrdU staining of SK-mel147 cells expressing shRNAs as in (A). Profiles from one representative experiment are displayed. BrdU-positive S phase cells are shown as mean values ± SD (n ≥ 3) (right); unpaired Student's test (two-tailed). See also Figure S2G.

(E) Percentage of SK-mel147 in G1, S, or G2/M phases, as revealed by PI incorporation. Values are mean ± SD (n ≥ 3); unpaired Student's test (two-tailed). Asterisks as follows, in all figures: *p < 0.05, **p < 0.01, ***p < 0.001. See also Figure S2F.

(F) Whole-cell extracts from shRNA-expressing SK-mel147 cells were immunoblotted for unmodified and phosphorylated Rb and cyclins. B-actin used as loading control.

(G) shRNA-expressing SK-mel147 cells were synchronized at early S phase by a double thymidine block and cell synchrony monitored by flow cytometry of PI stained cells at 5-hr intervals. Flow cytometry profiles from a representative experiment are shown.

cycle defects observed upon H2A.Z.1 knockdown and implicates a distinct role for H2A.Z.1 in melanoma.

Gene set enrichment analysis (GSEA) and transcription factor (TF) analysis further demonstrated that the H2A.Z.2-regulated genes are associated with transcriptional hallmarks of advanced melanoma and are targets of the E2F family, including E2F1 and E2F4 (Figures 3E, 3F, and S3E). Furthermore, qRT-PCR analysis revealed that E2F target gene expression correlates with H2A.Z.2 levels in human melanoma (Figure 3G). Given that E2F1 and E2F4 promote melanoma progression and metastasis (Alla et al., 2010; Ma et al., 2008), these results implicate concerted H2A.Z.2-E2F function in melanoma progression.

H2A.Z.2-Regulated Genes Show a Unique Signature of H2A.Z Occupancy

We next performed native chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) of H2A.Z to determine its genomic occupancy in melanoma cells (Figure 4). ChIP-seq of SK-mel147 cells stably expressing N-terminally eGFP-tagged H2A.Z.1 or H2A.Z.2 (along with eGFP-H2A as a control) was also carried out (Figure S4A), and exhibited highly overlapping genome-wide occupancy patterns with endogenous H2A.Z (Figures S4B and S4C) as well as with each other (Figure S4D). Therefore, we used endogenous H2A.Z ChIP-seq for further analyses.

Among H2A.Z-bound sites, 14% lie within promoters and 29% in gene bodies (Figure 4A). By integrating RNA sequencing and ChIP analyses from SK-mel147 cells, we found that H2A.Z promoter levels positively correlate with expression (Figure 4B), as previously reported (Barski et al., 2007; Hu et al., 2013). Intriguingly, gene body occupancy shows a striking negative correlation with gene expression (Figure 4B).

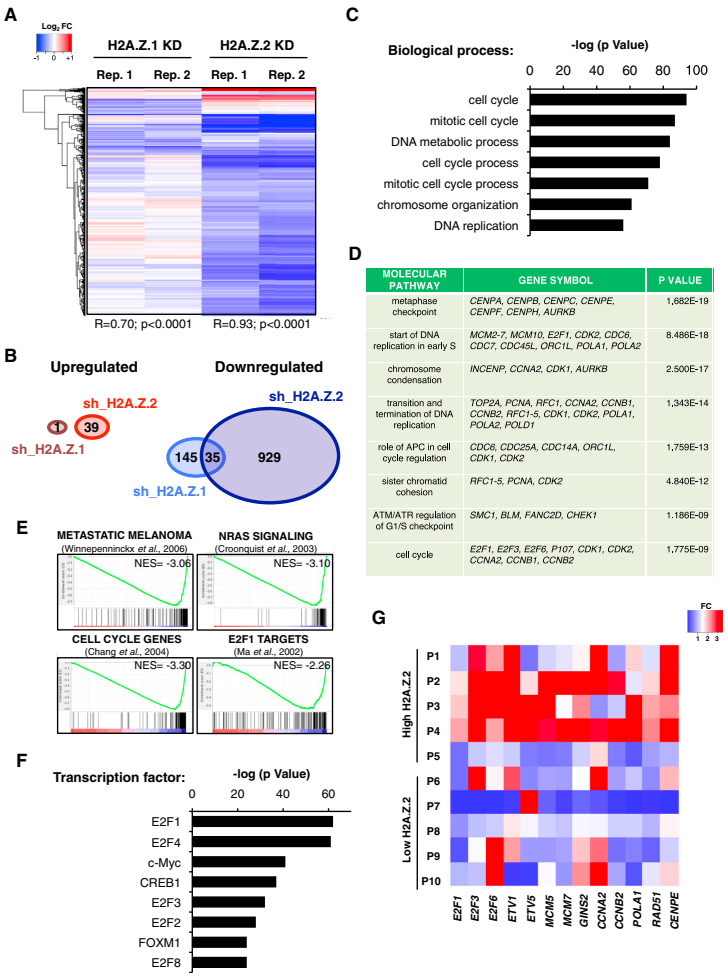


Figure 3. H2A.Z.2 Regulates Cell Cycle-Promoting Genes

(A) Gene expression profiles of SK-mel147 cells upon H2A.Z.1 and H2A.Z.2 knockdown (day 8 post-infection). Two biological replicates (with Pearson correlation), and genes displaying a significant (lfdr < 0.2) change in each replicate are shown.

(B) Venn diagrams exhibiting the numbers of genes that are significantly up- and downregulated upon H2A.Z.1 and H2A.Z.2 knockdown in SK-mel147 cells. See also Figures S3A and S3B.

(C) Functional annotation (biological process) of genes downregulated upon H2A.Z.2 knockdown in SK-mel147 cells. Enriched groups are ranked by the most significant p value.

(D) Functional annotation (molecular pathways) of genes as described in (C). Selected genes belonging to each pathway are shown; p value indicated.

(E) GSEA plots of genes altered upon H2A.Z.2 knockdown in SK-mel147 display negative correlation gene signatures as shown. FDR = 0.0; NES (normalized enrichment score) as indicated.

(F) TF regulation analysis of genes as described in (C). Enriched groups are ranked by the most significant p value. Analyses for (C), (D), and (F) were performed with MetaCore. See also Figures S3C–S3E.

(G) Heatmap generated by qRT-PCR values of the indicated genes in a subset of melanoma specimens (P1-P10) from Figure 1C. P1-P5 = high H2A.Z.2 and P6-P10 = low H2A.Z.2 expression levels (above and below the median, respectively). Expression levels of each gene are shown as fold change (FC) relative to one patient (not shown).

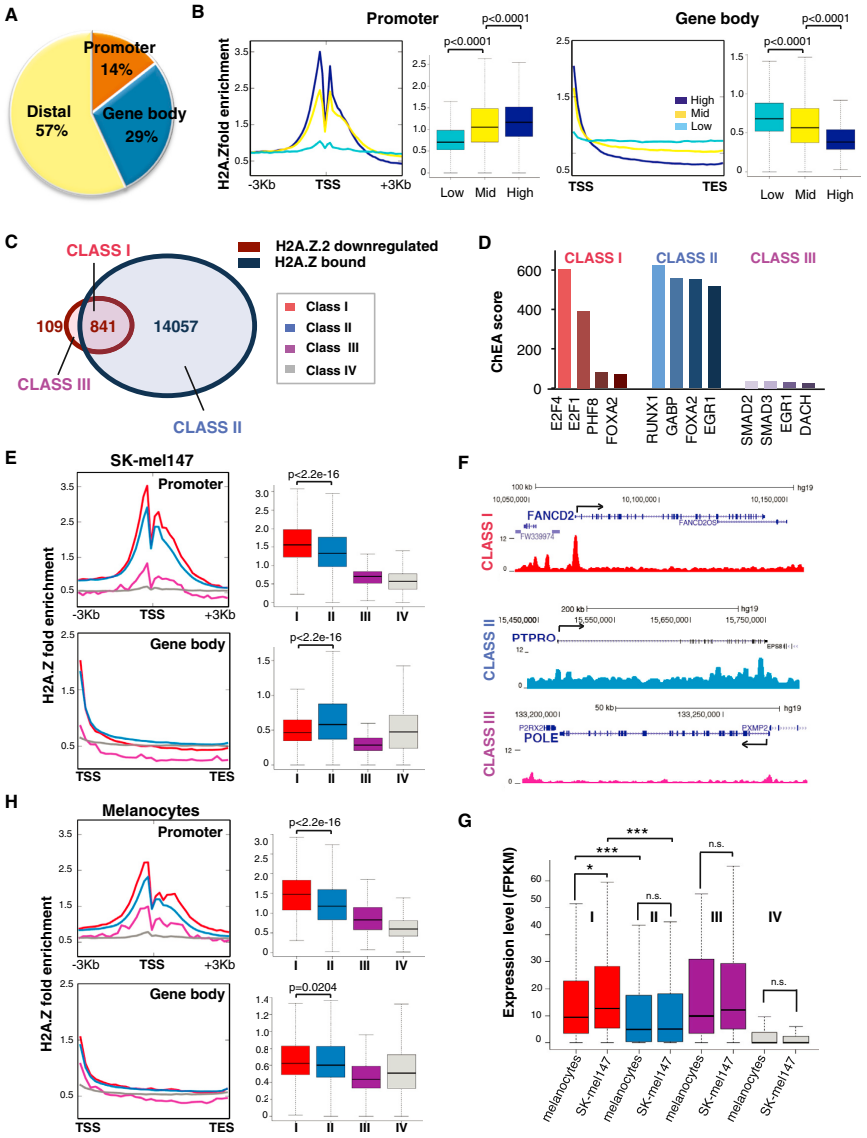


Figure 4. A Unique Signature of H2A.Z Occupancy at H2A.Z.2-Regulated Genes

(A) Pie chart displaying the percentages of H2A.Z peaks occupying promoters, gene bodies and distal regions. Promoters: −3 kb < TSS < +1 kb; gene bodies: from +1 kb > TSS to TES; all other regions defined as distal. TSS, transcription start site; TES, transcription end site.

(B) Correlation of H2A.Z signals at the promoter or gene body with mRNA expression levels. Genes were divided by expression level into high (top 25%), medium (middle 50%), and low (bottom 25%) from RNA sequencing data. Fold enrichment profiles (sliding 100 bp window) and boxplots were calculated around the TSS (−3 kb, +3 kb) and over the gene body (TSS to TES) for each group; Mann-Whitney test (two-tailed).

(legend continued on next page)

Finally, in line with previous reports (Barski et al., 2007; Hu et al., 2013; Obri et al., 2014), our results show that half of the H2A.Z-bound sites lie within intergenic regions (Figure 4A).

Next, we integrated H2A.Z ChIP-seq with H2A.Z.2 downregulated genes. Taking into account both promoter and gene body occupancy, we defined four classes of genes: H2A.Z-bound and H2A.Z.2 downregulated (Class I), H2A.Z-bound but not H2A.Z.2 downregulated (Class II), H2A.Z.2 downregulated but not H2A.Z-bound (Class III), and neither downregulated nor bound (Class IV) (Figure 4C; Table S2). Gene ontology analyses revealed that Class I is enriched for cell cycle genes, while Class II is enriched for metabolic processes (Figure S4E). Accordingly, ChIP Enrichment Analysis (ChEA2) (Kou et al., 2013) uncovered distinct TF binding profiles for each Class, with only Class I showing enrichment for E2Fs (Figure 4D). By examining H2A.Z distribution, we found that Class I genes were significantly enriched at the promoter and depleted within the gene body (Figures 4E and 4F). Conversely, many Class II genes lie within broader H2A.Z domains (Figure 4F). Expression of Class I genes is significantly higher than Class II genes in melanoma (Figure 4G), consistent with our findings in Figure 4B.

H2A.Z ChIP-seq in normal human melanocytes revealed that the Class I signature is not detectable because H2A.Z is not depleted in the gene body (Figure 4H). Consistent with this, Class I genes are expressed at significant lower levels in melanocytes than melanoma cells, whereas expression of all other classes remains largely unchanged (Figure 4G).

Collectively, our analyses revealed that H2A.Z.2 regulated/H2A.Z bound genes (Class I) have unique features in melanoma cells. They are E2F targets, highly expressed, and enriched for H2A.Z at the promoter and depleted in the gene body. These features do not apply to H2A.Z.1-downregulated genes (Figures S4F and S4G; Table S3), suggesting a unique chromatin signature at genes that regulate melanoma cell proliferation.

BRD2 Interacts with H2A.Z-Containing Nucleosomes and Is Overexpressed in Melanoma

To further decipher H2A.Z function, we investigated the factors that interact with H2A.Z-containing nucleosomes in melanoma cells. To this aim, we used unbiased label-free quantitative mass spectrometry (MS) (Eberl et al., 2013). Chromatin isolated from SK-mel147 cells stably expressing eGFP, eGFP-H2A, eGFP-H2A.Z.1, or eGFP-H2A.Z.2 was digested to mononucleosomes (Figure S5A), immunoprecipitated (Figure S5B), and analyzed with liquid chromatography-tandem mass spectrometry (LC-MS/MS). We found significant enrichment of ~45 H2A.Z interactors as compared to H2A-containing nucleosomes (Figures 5A and S5C). The majority of these proteins were found in both H2A.Z variant IPs, including members of the H2A.Z histone chaperone complex, SRCAP (Billon and Côté, 2013).

We identified BRD2 to be enriched in H2A.Z.1- and H2A.Z.2-containing nucleosomes (Figures 5A and 5B). BET proteins (BRD2, BRD3, BRD4, and BRDT) bind to acetylated lysine residues in histones (LeRoy et al., 2008, 2012) and function as scaffolds to recruit chromatin modifying enzymes and TFs, thereby coupling histone acetylation to transcription (reviewed in Belkina and Denis, 2012). Whereas BRD2 and BRD4 are both overexpressed in melanoma (Segura et al., 2013), only BRD2 specifically interacts with H2A.Z-containing nucleosomes (Figure 5A and data not shown). We next tested whether hyperacetylation of histones would enhance the interaction between BRD2 and H2A.Z isoforms in melanoma cells. Treatment with the HDAC inhibitor trichostatin A (TSA) resulted in increased histone H4 and H2A.Z acetylation and increased BRD2 chromatin association (Figure 5C). Furthermore, the BRD2-H2A.Z interaction was enhanced (Figure 5C). By probing primary and metastatic melanoma cell lines (as in Figure 1A), we observed hyperacetylation of H4 and H2A.Z, and high levels of BRD2 in metastatic cells (Figure 5D). Collectively, these results are consistent with the fact that BRD2 is recruited to chromatin by a combination of acetylated H4 (H4ac) and H2A.Z (Draker et al., 2012).

Through immunohistochemistry (IHC) of BRD2 in a cohort of patient samples including benign nevi, thick primary melanoma, and metastatic melanoma, we detected a significant increase of BRD2 in primary and metastatic melanoma specimens as compared to dermal melanocytes of nevi (Figure 5E). Next, we investigated BRD2 knockdown in multiple melanoma cell lines and observed proliferation defects via G1/S arrest (Figures 5F, 5G, and S5D–S5F). BRD2 knockdown altered gene expression of selected E2F targets (Figures 5H and S5G). Collectively, BRD2 knockdown recapitulated the phenotype observed upon H2A.Z.2 loss, suggesting that H2A.Z.2 and BRD2 work together to promote Class I transcription.

(C) Venn diagrams displaying H2A.Z.2 downregulated genes and H2A.Z bound genes by ChIP-seq in SK-mel147. Class I (downregulated in H2A.Z.2 knockdown and bound by H2A.Z, red); Class II (bound by H2A.Z but unaffected by H2A.Z.2 knockdown, light blue); Class III (downregulated by H2A.Z.2 knockdown but not bound by H2A.Z, purple); Class IV (11,003 genes that are not downregulated by H2A.Z.2 knockdown and not bound, gray).

(D) ChIP enrichment analysis tool (ChEA2) analysis of Class I, Class II, and Class III genes (as defined and color coded in A). The ChEA2 database contains ChIP-seq data of 200 transcription factors from 221 publications for a total of 458,471 TF-target interactions (Kou et al., 2013). Transcription factors are ranked by ChEA combined score.

(E) H2A.Z occupancy at the promoter and gene body of the four classes of genes defined in (C), in SK-mel147 cells. Profiles and boxplots represent fold enrichment over input. Mann-Whitney test (two-tailed).

(F) Captures of the UCSC genome browser (GRCh37/hg19) showing the ChIP-seq profiles for H2A.Z for genes belonging to each of the classes defined in (C). RefSeq annotated genes are displayed on top.

(G) Boxplot representing expression levels (FPKM) for each gene class as in (C), in primary melanocytes and in SK-mel147. Mann-Whitney test (two-tailed). Two-way ANOVA.

(H) H2A.Z occupancy at the promoter and gene body of the four classes of genes defined in (C) in primary melanocytes.

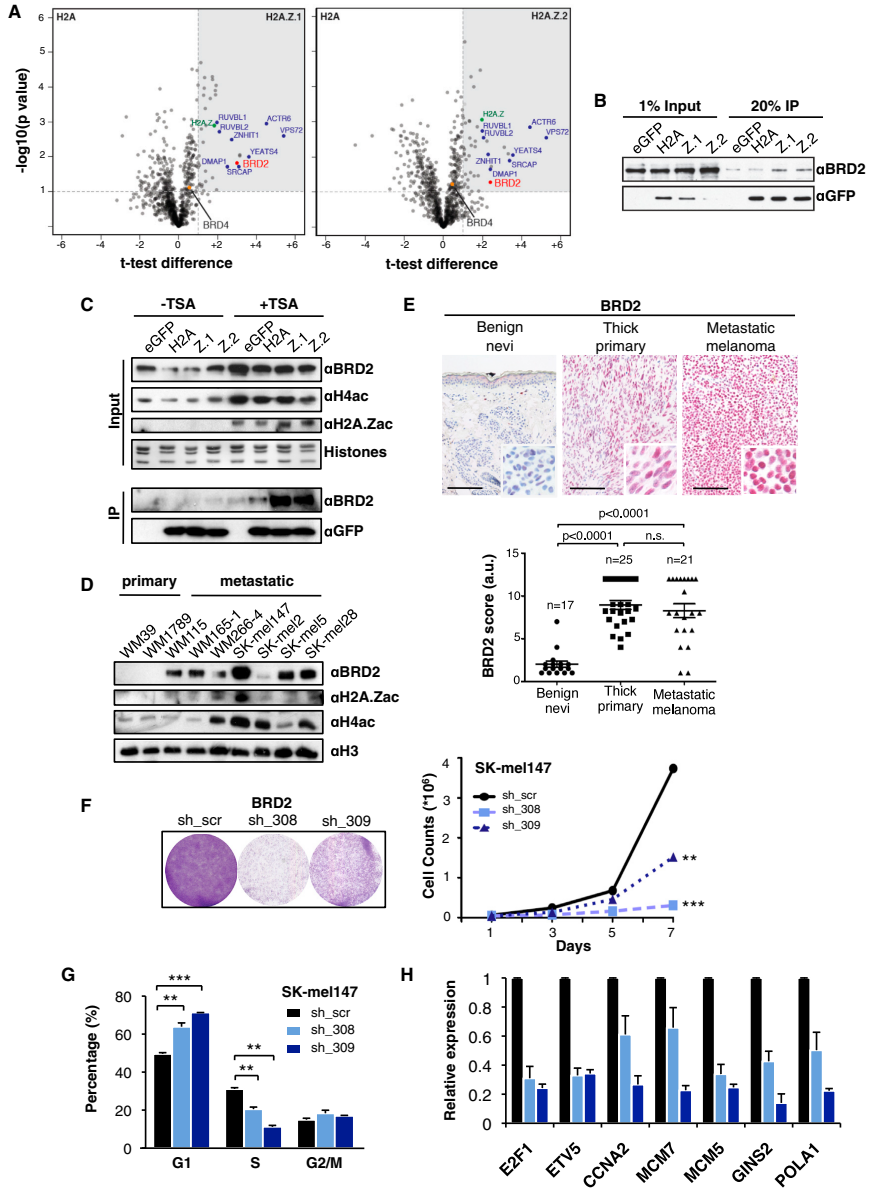


Figure 5. BRD2 Interacts with H2A.Z-Containing Nucleosomes and Is Overexpressed in Melanoma

(A) Volcano plots of label-free interactions of eGFP-H2A.Z.1- or eGFP-H2A.Z.2-containing nucleosomes. Significantly enriched proteins over eGFP-H2A containing nucleosomes are shown in the upper right box (gray shading). Members of the H2A.Z-specific chaperone/remodeling complex SRCAP are highlighted in blue, H2A.Z in green, BRD2 and BRD4 as red and orange dots, respectively. See also Figures S5A–S5C.

(legend continued on next page)

An H2A.Z.2-BRD2-E2F1 Axis in Melanoma

To test this hypothesis, we performed ChIP-seq of BRD2 in SK-mel147 cells, and found a similar genomic distribution as H2A.Z (Figures 4A and S6A). We next determined the extent of co-localization of BRD2 and H2A.Z genome-wide and found that BRD2 peaks overlap with H2A.Z largely at promoters (Figure 6A). Promoters of Classes I and II genes are bound by BRD2 (Figures 6B and S6B), consistent with the fact that BRD2 interacts with both H2A.Z isoforms. However, H2A.Z and BRD2 have the highest enrichment at Class I genes (Figures 4E and 6C).

Because our in silico analyses predicted that Class I genes are E2F targets (Figure 4D), and BRD2 interacts with E2F1 to mediate its recruitment to chromatin (Denis et al., 2006; Sinha et al., 2005), we next queried whether Class I genes are bound by E2F1. ChIP-seq of E2F1 in SK-mel147 cells showed this was indeed the case (Figures 6B, 6C, and S6B). Taken together, these data suggest that H2A.Z.2 works cooperatively with BRD2 and E2F1 (Figure 6D) to promote high levels of transcription at Class I genes in melanoma. Next, we probed a panel of benign nevi and melanoma tissues for H2A.Z, BRD2, E2F1, and H4ac, and found evidence of the H2A.Z-BRD2-E2F axis in melanoma specimens (Figure 6E). This reinforces the relevance of our findings for melanoma disease.

H2A.Z.2 Depletion Impairs BRD2 and E2F1 Function

The findings above prompted us to investigate BRD2 and E2F1 levels upon H2A.Z silencing. We observed marked reduction of BRD2 and E2F1 levels upon H2A.Z.2, but not H2A.Z.1, knockdown across melanoma cell lines (Figures 6F and S6C). BET family members are not transcriptionally regulated by H2A.Z.2 (Figure S6D), suggesting that H2A.Z.2 stabilizes these factors. This was paralleled by a dramatic loss of H4 and H3 acetylation (Figures 6F, S6C, and S6E). These data suggest that BRD2 and E2F1 chromatin recruitment to Class I genes, mediated by histone acetylation, is impaired in H2A.Z.2-deficient cells. Thus, we performed ChIP-qPCR for BRD2 and E2F1 in either control or H2A.Z isoform-depleted cells, and found that BRD2 and E2F1 recruitment to Class I promoters is dependent on H2A.Z.2 (Figure 6G). Overall, H2A.Z.2 deficiency results in dramatic alterations of chromatin structure, thereby clearly distinguishing it from H2A.Z.1. Overall, our ChIP studies demonstrate that H2A.Z.2, BRD2, and E2F1 work cooperatively to promote high levels of transcription at cell cycle-promoting genes in melanoma.

H2A.Z.2 Deficiency Sensitizes Melanoma Cells to Therapy

Because we observed loss of histone acetylation upon H2A.Z.2 knockdown, we queried whether H2A.Z.2 depletion might potentiate the effects of BET inhibitors (BETi). BETi prevent the acetyl-lysine binding of bromodomains with high affinity and are effective agents in a number of tumors (Dawson et al., 2012; Segura et al., 2013). We first assessed the sensitivity of melanoma cells to JQ1 (Filippakopoulos et al., 2010) (Figure 7A) and found a dose-dependent growth inhibitory effect in the majority of cell lines tested (Figures S7A–S7C). Cells treated with JQ1 for 4 days accumulated in G2/M (Figure S7B), with minor induction of apoptosis (data not shown).

Next, we investigated whether H2A.Z.2 knockdown cooperates with JQ1 to enhance the antiproliferative effect of melanoma cells. Whereas JQ1 treatment or H2A.Z.2 knockdown alone induced growth arrest (Figures S7B and 2D–2G), the combination resulted in cell death in both *BRAF* and *NRAS* mutant lines (Figures 7B and S7D). Functional annotation of the SK-mel147 transcriptome upon JQ1 treatment (Table S4) revealed enrichment in developmental processes, distinguishing it from cell cycle annotation associated with H2A.Z.2 silencing (see Figures 3C and S7E). Consistent with this synergy, we observed that the transcriptional profiles of H2A.Z.2 knockdown and BETi show minimal overlap (Figure 7C; Table S4), suggesting that whereas H2A.Z.2-regulated genes are BRD2 and E2F1 targets, the mode of action of JQ1 is largely distinct from H2A.Z.2. Although H2A.Z and BRD2 are enriched on promoters of JQ1 regulated genes (however, significantly less so than on H2A.Z.2-regulated genes), E2F1 binding is absent (Figure 7D). JQ1 regulated genes are instead targets of distinct TFs (Figure S7F).

H2A.Z.2 loss not only enhances sensitivity of melanoma cells to BETi, but also to chemotherapy and targeted therapies (MEKi) used clinically for melanoma (Figure 7E). Collectively, these data suggest that H2A.Z.2 is a critical mediator of melanoma drug sensitivity and regulating its deposition may serve as an important target for novel therapeutic strategies.

DISCUSSION

H2A.Z.2 Is a Driver of Malignant Melanoma

While histone variants and their chaperones have emerged as critical players in cancer biology (Vardabasso et al., 2013), our mechanistic understanding remains limited. Here, we report a unique role for H2A.Z.2 in driving melanoma cell proliferation

(B) Immunoblots for BRD2 and GFP upon immunoprecipitation of mononucleosomes generated from SK-mel147 cells expressing eGFP, eGFP-H2A, -H2A.Z.1, or -H2A.Z.2.
(C) SK-mel147 cells as in (B) were treated with DMSO or TSA (200 nM for 2 hr), and chromatin was probed for BRD2, H4ac, and H2A.Zac (top). Histones used for loading. Bottom: Immunoblots for BRD2 and GFP upon immunoprecipitation are shown.
(D) Chromatin extracted from primary and metastatic cell lines probed with BRD2, H2A.Zac, and H4ac antibodies; H3 used for loading. See also Figure 1A.
(E) IHC for BRD2 in representative intradermal nevi, thick primary, and metastatic melanoma tissue. Images at 20× magnification; insets at 40× magnification. Scale bar represents 100 μm. Scores derived by multiplying the number of positively stained cells (1–4) by intensity of stain (1–3); Mann-Whitney (two-tailed).
(F) Colony formation and proliferation assays of SK-mel147 cells expressing control or BRD2 shRNAs as shown. Data are mean ± SEM ($n \geq 2$); two-way ANOVA.
(G) Percentage of SK-mel147 cells in G1, S, or G2 phases, as shown by PI incorporation. Values are mean ± SD ($n \geq 3$); unpaired Student's test (two-tailed).
(H) Expression of a handful of Class I genes was analyzed by qRT-PCR upon BRD2 knockdown. Expression is shown normalized to GAPDH and relative to scrambled shRNA. Mean ± SD is shown ($n \geq 3$).

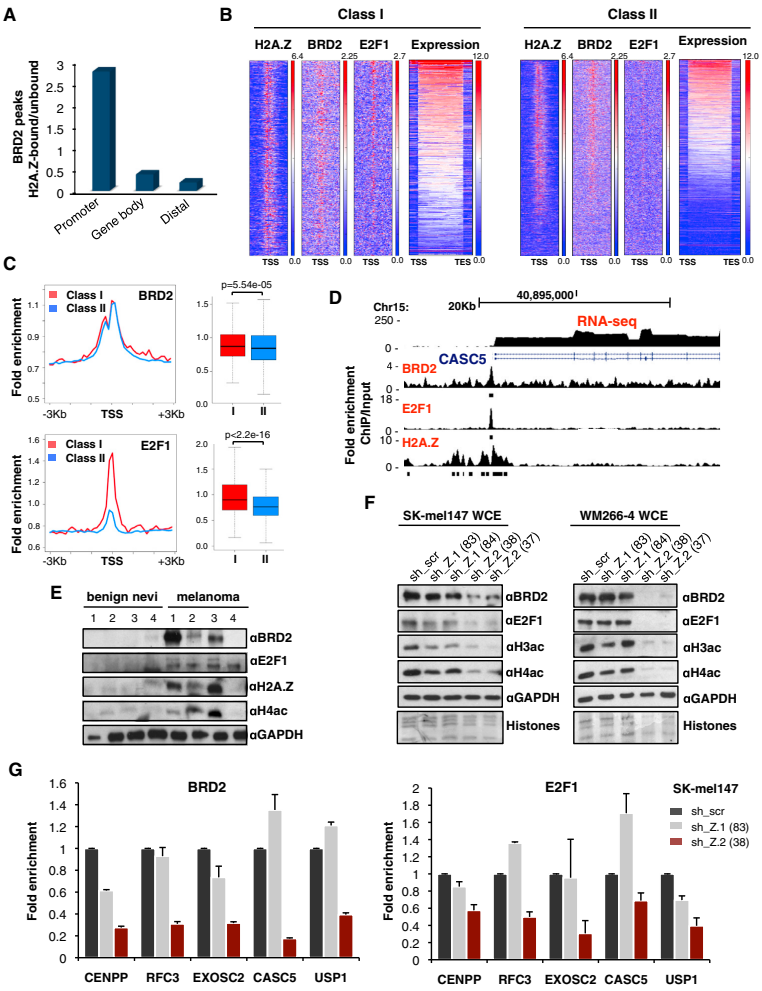


Figure 6. An H2A.Z.2-BRD2-E2F1 Axis in Melanoma

(A) Histograms of the ratio between BRD2 peaks bound by H2A.Z and BRD2 peaks not bound by H2A.Z at promoters, gene bodies, and distal regions as defined in Figure 4A.

(B) Heatmaps of promoters (−3 kb, +3 kb) of Class I and Class II genes based on H2A.Z, BRD2, and E2F1 fold enrichment over input, and ranked by expression level. Expression is indicated as log2 RNA-seq signal. See also Figure S6B.

(C) BRD2 and E2F1 occupancy at the promoter Class I and Class II genes in SK-mel147 cells. Profiles and boxplots represent fold enrichment over input. Mann-Whitney test (two-tailed).

(D) UCSC genome browser (GRCh37/hg19) capture of ~30 kb region of human chromosome 15 depicting a Class I gene. Read counts (normalized fold enrichment of ChIP over input DNA) for BRD2, E2F1, and H2A.Z and FPKM for RNA-seq are shown. RefSeq annotated genes are displayed above.

(E) Whole-cell extracts from fresh-frozen benign nevi and metastatic specimens probed with BRD2, E2F1, H2A.Z, and H4ac antibodies; GAPDH used for loading.

(F) Whole-cell extracts from control and isoform-depleted SK-mel147 and WM266-4 cells were immunoblotted for BRD2, E2F1, H3ac, and H4ac. GAPDH served as loading control. See also Figures S6C and S6E.

(G) ChIP-qPCR for BRD2 (left) and E2F1 (right) at Class I genes in SK-mel147 expressing control or isoform-specific shRNAs as indicated. Fold enrichment ChIP/input is plotted relative to scrambled shRNA. One representative experiment shown; values are mean ± SD (n ≥ 2).

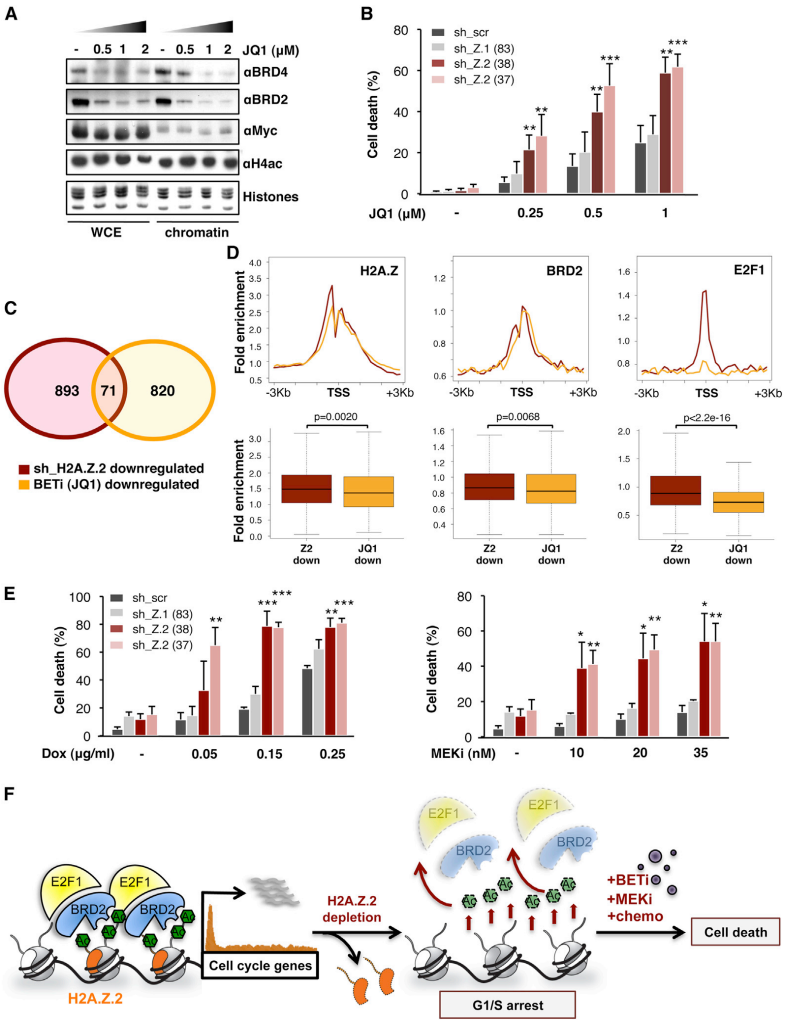


Figure 7. H2A.Z.2 Deficiency Sensitizes Melanoma Cells to Chemotherapy and Targeted Therapies

(A) Chromatin and whole-cell extracts from SK-mel147 cells exposed to DMSO or 0.5–2 μM of JQ1 for 2 days were immunoblotted for BRD2, BRD4, Myc, and H4ac. Amido black staining of histones for loading.

(B) SK-mel147 cells were infected with shRNAs as shown and subsequently treated with JQ1 as indicated for 4 days. Percentage of Annexin V positive cells shown. Values are mean ± SD ($n \geq 3$); unpaired Student's test (two-tailed).

(C) Venn diagram of genes downregulated in SK-mel147 upon H2A.Z.2 knockdown (red) or JQ1 treatment (1 μM for 6 hr; cutoff of FC ≥ -2), (orange).

(D) H2A.Z, BRD2, and E2F1 occupancy at the promoter of H2A.Z.2 downregulated (red) or JQ1 downregulated genes (orange). Profiles and boxplots represent fold enrichment over input. Mann-Whitney test (two-tailed).

(E) SK-mel147 cells were infected as in (B) and treated with doxorubicin as indicated (Dox, left) and MEK inhibitor PD325901 (MEKi, right). Percentage of sub-G1 cells upon 2 days of treatment shown. Values are mean ± SD ($n \geq 3$); unpaired Student's test (two-tailed).

(F) A model for the H2A.Z.2-dependent regulation of cell cycle gene transcription in melanoma. Depletion of H2A.Z.2 results in reduced histone acetylation, BRD2 and E2F1 levels, impairs recruitment of BRD2 and E2F1 to its target genes, and induces G1/S arrest. Combining depletion of H2A.Z.2 loss with targeted therapy or chemotherapy leads to cell death.

and drug sensitivity. Our study suggests a melanoma-specific role for H2A.Z.2 in promoting proliferation, and it will be of interest to learn if H2A.Z.2 plays a similar role in other tumors. Importantly, we do not exclude a role for H2A.Z.1 in melanoma because it is also upregulated and correlates with shorter patient survival.

Because H2A.Z isoforms have distinct roles in melanoma, we hypothesized they may have unique interaction partners and genomic occupancy. Our studies indicate that H2A.Z.1 and H2A.Z.2 share genomic occupancy patterns and interact with similar histone chaperone complexes. However, it is clear that H2A.Z.2 is critical for promoting cell cycle progression in melanoma, and acts distinctly from H2A.Z.1. Our data strongly suggest that a unique property of H2A.Z.2 is to promote and/or maintain BRD2, E2F1, and histone acetylation levels. While the exact mechanism remains unclear, H2A.Z.2 likely acts together with histone acetylation to recruit co-activators and TFs, such as BRD2 and E2F1, respectively, to promote expression of cell cycle regulators (see below).

A Unique Signature of H2A.Z Occupancy at E2F Target Genes

Our analyses revealed that H2A.Z.2 promotes the expression of E2F target genes. In melanoma cells, these genes are characterized by a unique signature of H2A.Z occupancy—highly enriched at the TSS and depleted within the gene body—and this pattern associates with high gene expression levels. Our findings are in line with previous observations in plants and yeast (Coleman-Derr and Zilberman, 2012; Sadeghi et al., 2011; Zilberman et al., 2008); for example, H2A.Z is excluded from the bodies of actively transcribed genes in *Arabidopsis* (Coleman-Derr and Zilberman, 2012; Zilberman et al., 2008). Intriguingly, the DREAM complex was recently reported to promote H2A.Z gene body incorporation to repress cell cycle progression genes in *C. elegans* (Latorre et al., 2015). Together, these studies suggest that H2A.Z is differentially distributed across promoters and gene bodies at distinct subsets of genes to regulate their expression levels.

An H2A.Z.2-BRD2-E2F1 Axis in Melanoma

Our study has identified BRD2 as an H2A.Z-interacting protein in malignant melanoma. Work by Denis and colleagues initially demonstrated that BRD2 has oncogenic potential: BRD2 transforms mouse fibroblasts in the context of oncogenic Ras (Denis et al., 2000), and Eμ-*BRD2* transgenic mice develop B cell lymphoma and leukemia (Greenwald et al., 2004). In fact, BRD2 has a crucial role in cell cycle control, and by interaction with E2F1, it regulates the expression of cyclins and other cell cycle regulatory genes (Denis et al., 2000, 2006; Sinha et al., 2005).

Our loss-of-function approach revealed that the chromatin association and total levels of BRD2, E2F1, and histone acetylation are H2A.Z.2 dependent. This is in line with the fact that BRD2's preference for H2A.Z-containing nucleosomes is mediated by a combination of hyperacetylated H4, and features on H2A.Z itself (Draker et al., 2012), and that histone acetyltransferase (HAT) activity is contained within BRD2 nuclear complexes (Sinha et al., 2005). Furthermore, we find evidence of an H2A.Z-BRD2-E2F axis in melanoma tissues. Accordingly, our ChIP analyses show that BRD2 and E2F1 are enriched at promoters of Class I genes and that H2A.Z.2 is required for recruitment of these factors to these E2F targets. Interestingly, Draker et al. found that recruitment of BRD2 to androgen receptor (AR)-regulated genes in prostate cancer cells is dependent on H2A.Z.1. Thus, BRD2 may associate with distinct TFs and H2A.Z isoforms to achieve oncogenic gene transcription in different tumor types.

Collectively, we envision that H2A.Z.2 recruits BRD2 and E2Fs, along with HAT activity, to E2F target genes in melanoma cells. This in turn results in increased expression of cell cycle genes, and ultimately promotes proliferation (Figure 7F). Our findings implicate the H2A.Z.2-BRD2-E2F1 axis as a driver of melanoma progression. Of these molecules, BRD2 represents a key target for therapy.

Novel Epigenetic Therapeutic Strategies to Treat Melanoma

Metastatic melanoma is notoriously refractory to conventional cancer therapies and remains largely resistant to current targeted therapies (Lito et al., 2013). Here we show that in combination with BET inhibition, H2A.Z.2 depletion is effective in inducing cell death. Because a tool to disrupt H2A.Z deposition is currently lacking, it is plausible that combining BETi with a potent inhibitor of HAT activity will potentiate melanoma cell death (Figure 7F). This combination may not only evict BET proteins from chromatin, but cause additional destabilization of BET proteins and their associated TFs due to loss of acetylation (Figure 7F). It will be of interest to create BRD-specific inhibitors, if achievable, because our study suggests that BRDs function distinctly in disease. Finally, our findings implicate H2A.Z.2 as a mediator of cell proliferation and drug sensitivity in malignant melanoma. Because histone modification and deposition are reversible processes, our study holds therapeutic potential for this highly intractable neoplasm.

EXPERIMENTAL PROCEDURES

Cell Culture, Plasmids, and Infections

Primary (WM115, WM1789, WM39), metastatic (SK-mel147, WM266-4, 501mel, A375, SK-mel2, SK-mel28, SK-mel239, SK-mel5, M14, WM165-1) melanoma cell lines, HeLa cells, and human melanocytes were cultured as described in the Supplemental Experimental Procedures. Lentiviral vectors and shRNAs used for the generation of stable cell lines are described in the Supplemental Experimental Procedures. Infections were performed according to standard procedures (Kapoor et al., 2010).

Chromatin Fractionation, Acid Extraction of Histones, and Immunoblotting

Chromatin fractionation and acid extraction of histones were performed as described (Kapoor et al., 2010) and in the Supplemental Experimental Procedures. Antibodies used in this study are listed in the Supplemental Experimental Procedures.

Clinical Specimens

Approval to collect melanoma specimens was granted by Mount Sinai Biorepository Cooperative and the New York University Interdisciplinary Melanoma Cooperative Group (project number HSD08-00565 and IRB number 10362, respectively). Approval to collect benign nevi was granted by ISMMS (Icahn School of Medicine at Mount Sinai) Division of Dermatopathology (project number 08-0964).

RNA Extraction, qRT-PCR, and Microarray Hybridization

For RNA extraction, qRT-PCR, and primers, see the Supplemental Experimental Procedures. RNA amplification, labeling, and hybridization to Human Gene 1.0 ST Arrays (Affymetrix) were performed as described previously (Wiedemann et al., 2010), and data processed in R/bioconductor (http://www.bioconductor.org). For data analysis, see the Supplemental Experimental Procedures.

Cell Proliferation, Colony Formation, and Flow Cytometry

For proliferation curves, cells were counted up to 7 days and normalized to cell counts at day 1. Colony formation assay was performed by seeding cells at low density and allowing growth for 2 weeks. Cells were washed with phosphate buffered saline, fixed in 10% methanol/acetic acid solution, and stained with 1% crystal violet. Flow cytometry experiments were performed as described in the Supplemental Experimental Procedures.

Native and Crosslinked ChIP and Next-Generation Sequencing

Chromatin from SK-mel147 cells was digested with micrococcal nuclease (MNase) and used for ChIP with H2A.Z (Abcam ab4174) and GFP Trap Beads (Chromotek), essentially as described (Hasson et al., 2013). SK-mel147 cells stably expressing control or isoform-specific shRNAs were crosslinked for 10 min with 1% formaldehyde and immunoprecipitated with BRD2 and E2F1 antibodies (Bethyl Laboratories A302-583A and Santa Cruz sc-193, respectively) as described in the Supplemental Experimental Procedures. Sequencing libraries were generated and barcoded for multiplexing as described (Hasson et al., 2013) and libraries were submitted for 100-bp, single-end Illumina sequencing on a HiSeq 2500. For data processing and analysis, see the Supplemental Experimental Procedures.

RNA Sequencing

Total RNA samples were isolated from human melanocytes and enantiomer- or JQ1-treated SK-mel147 using miRNeasy mini kit (QIAGEN) following manufacturer's protocol. Sequencing libraries were prepared and data analysis performed as described in the Supplemental Experimental Procedures.

Mononucleosome Immunoprecipitation

Mononucleosomes were generated according to (Sansoni et al., 2014) and described in the Supplemental Experimental Procedures.

LC-MS/MS Analysis and MS Data Analysis

See the Supplemental Experimental Procedures for details.

Statistical Methodologies

Statistical tests were applied as indicated in figure legends. Asterisks are as follows: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Boxplots represent Tukey boxplots with outliers omitted.

ACCESSION NUMBERS

The accession number for all microarray, RNA-seq, and ChIP-seq data sets reported in this paper is NCBI GEO: GSE59060.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2015.05.009.

AUTHOR CONTRIBUTIONS

C.V. and E.B. conceived this study with guidance from S.B.H. C.V. performed all qPCR and shRNA studies and generated eGFP melanoma cell lines. C.V. and T.P. performed immunoblots and cell cycle analyses. S.B.H. and T. Straub performed and analyzed microarrays, respectively. C.V., D.H., and S.P. performed ChIP-seq experiments, and data analyses were led by C.V., A.G.-M., D.V.-G., D.H., and T. Straub. S.P. performed IPs for MS, and E.C.K. performed MS with the support of M.M. S.P. analyzed MS results under the guidance of E.C.K. C.V. and T. Strub performed drug treatments, and S.P., C.V., and T. Strub performed BRD2 immunoblots. J.D. performed BRD2 IHC, and J.L.Y. and R.S. scored tissues. C.-Y.C. assisted with network analyses, and A.V. provided copy number analysis support. M.F.S., B.F.-C., and E.H. provided assistance with patient data and with BETi experiments. B.F.-C. prepared samples for RNA-seq, and D.H. analyzed RNA-seq data. C.V., A.G.-M., D.V.-G., D.H., S.P., E.C.K., S.B.H., and E.B. designed experiments and interpreted results. C.V., E.B., and S.B.H. wrote the manuscript with contributions from all other coauthors.

ACKNOWLEDGMENTS

The authors are grateful for assistance and reagents provided by Luis Duarte, Avnish Kapoor, Nicholas Mills, Clemens Bönisch, Pauline Rimmele, Danielle Martinez, and the laboratories of Robert Fisher, Stuart Aaronson, Ramon Parsons, and Jay Bradner. We thank Cristina Montagna and Jidong Shan at Molecular Cytogenetic Core at Albert Einstein College of Medicine; Genomics Core Facility at Mount Sinai; Avi Ma'ayan for statistical support; Robert Phelps, Madelaine Haddican, Shelbi Jim On, and Giselle Singer for assistance with benign nevi collection; and Mark Lebwohl for his support. Funding was supported by a Melanoma Research Development Award (Mount Sinai) to C.V.; DOD BCRP Postdoctoral fellowship (W81XWH-11-1-0018) and NYSCF Druckenmiller Fellowship to A.G.-M.; graduate fellowship from CONACyT (239663) to D.V.-G.; International Max Planck Research School for Life Science (IMPRS-LS) to S.P.; Bundesministerium für Bildung und Forschung (FKZ01GS0861, DiGtoP consortium) to M.M.; DOD Collaborative Awards CA093471 (W81XWH-10-1-080), 1R01CA155234, and 1R01CA163891 to E.H.; Deutsche Forschungsgemeinschaft through the Collaborative Research Center SFB 1064 (project A10 to S.B.H. and project Z04 to T.S.); HA 5437/4-1 and the Center for Integrated Protein Science Munich (CIPSM) to S.B.H.; Worldwide Cancer Research, Hirschl/Weill-CaulierResearch Award; and NCI/NIH R01CA154683 to E.B. We dedicate this study in memory of Estela Medrano.

Received: October 30, 2014
Revised: March 24, 2015
Accepted: April 30, 2015
Published: June 4, 2015

REFERENCES

Alla, V., Engelmann, D., Niemetz, A., Pahnke, J., Schmidt, A., Kunz, M., Emmrich, S., Steder, M., Koczan, D., and Pützer, B.M. (2010). E2F1 in melanoma progression and metastasis. J. Natl. Cancer Inst. *102*, 127–133.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Belkina, A.C., and Denis, G.V. (2012). BET domain co-regulators in obesity, inflammation and cancer. Nat. Rev. Cancer *12*, 465–477.

Billon, P., and Côté, J. (2013). Precise deposition of histone H2A.Z in chromatin for genome expression and maintenance. Biochim. Biophys. Acta *1819*, 290–302.

Bogunovic, D., O'Neill, D.W., Belitskaya-Levy, I., Vacic, V., Yu, Y.L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A.C., et al. (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. Proc. Natl. Acad. Sci. USA *106*, 20429–20434.

Bönisch, C., and Hake, S.B. (2012). Histone H2A variants in nucleosomes and chromatin: more or less stable? Nucleic Acids Res. *40*, 10719–10741.

Ceol, C.J., Houvras, Y., Jane-Valbuena, J., Bilodeau, S., Orlando, D.A., Battisti, V., Fritsch, L., Lin, W.M., Hollmann, T.J., Ferré, F., et al. (2011). The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. Nature *471*, 513–517.

Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., et al.; BRIM-3 Study Group

(2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N. Engl. J. Med. *364*, 2507–2516.

Coleman-Derr, D., and Zilberman, D. (2012). Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. PLoS Genet. *8*, e1002988.

Dawson, M.A., Kouzarides, T., and Huntly, B.J. (2012). Targeting epigenetic readers in cancer. N. Engl. J. Med. *367*, 647–657.

Denis, G.V., Vaziri, C., Guo, N., and Faller, D.V. (2000). RING3 kinase transactivates promoters of cell cycle regulatory genes through E2F. Cell Growth Differ. *11*, 417–424.

Denis, G.V., McComb, M.E., Faller, D.V., Sinha, A., Romesser, P.B., and Costello, C.E. (2006). Identification of transcription complexes that contain the double bromodomain protein Brd2 and chromatin remodeling machines. J. Proteome Res. *5*, 502–511.

Dhillon, N., Oki, M., Szyjka, S.J., Aparicio, O.M., and Kamakaka, R.T. (2006). H2A.Z functions to regulate progression through the cell cycle. Mol. Cell. Biol. *26*, 489–501.

Draker, R., Ng, M.K., Sarcinella, E., Ignatchenko, V., Kislinger, T., and Cheung, P. (2012). A combination of H2A.Z and H4 acetylation recruits Brd2 to chromatin during transcriptional activation. PLoS Genet. *8*, e1003047.

Dryhurst, D., Ishibashi, T., Rose, K.L., Eirín-López, J.M., McDonald, D., Silva-Moreno, B., Veldhoen, N., Helbing, C.C., Hendzel, M.J., Shabanowitz, J., et al. (2009). Characterization of the histone H2A.Z-1 and H2A.Z-2 isoforms in vertebrates. BMC Biol. *7*, 86.

Duarte, L.F., Young, A.R.J., Wang, Z., Wu, H.-A., Panda, T., Kou, Y., Kapoor, A., Hasson, D., Mills, N.R., Ma'ayan, A., et al. (2014). Histone H3.3 and its proteolytically processed form drive a cellular senescence programme. Nat. Commun. *5*, 5210.

Eberl, H.C., Spruijt, C.G., Kelstrup, C.D., Vermeulen, M., and Mann, M. (2013). A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. Mol. Cell *49*, 368–378.

Faast, R., Thonglairoam, V., Schulz, T.C., Beall, J., Wells, J.R., Taylor, H., Matthaei, K., Rathjen, P.D., Tremethick, D.J., and Lyons, I. (2001). Histone variant H2A.Z is required for early mammalian development. Curr. Biol. *11*, 1183–1187.

Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W.B., Fedorov, O., Morse, E.M., Keates, T., Hickman, T.T., Felletar, I., et al. (2010). Selective inhibition of BET bromodomains. Nature *468*, 1067–1073.

Greenwald, R.J., Tumang, J.R., Sinha, A., Currier, N., Cardiff, R.D., Rothstein, T.L., Faller, D.V., and Denis, G.V. (2004). E mu-BRD2 transgenic mice develop B-cell lymphoma and leukemia. Blood *103*, 1475–1484.

Hasson, D., Panchenko, T., Salimian, K.J., Salman, M.U., Sekulic, N., Alonso, A., Warburton, P.E., and Black, B.E. (2013). The octamer is the major form of CENP-A nucleosomes at human centromeres. Nat. Struct. Mol. Biol. *20*, 687–695.

Horikoshi, N., Sato, K., Shimada, K., Arimura, Y., Osakabe, A., Tachiwana, H., Hayashi-Takanaka, Y., Iwasaki, W., Kagawa, W., Harata, M., et al. (2013). Structural polymorphism in the L1 loop regions of human H2A.Z.1 and H2A.Z.2. Acta Crystallogr. D Biol. Crystallogr. *69*, 2431–2439.

Hu, G., Cui, K., Northrup, D., Liu, C., Wang, C., Tang, Q., Ge, K., Levens, D., Crane-Robinson, C., and Zhao, K. (2013). H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. Cell Stem Cell *12*, 180–192.

Kapoor, A., Goldberg, M.S., Cumberland, L.K., Ratnakumar, K., Segura, M.F., Emanuel, P.O., Menendez, S., Vardabasso, C., Leroy, G., Vidal, C.I., et al. (2010). The histone variant macroH2A suppresses melanoma progression through regulation of CDK8. Nature *468*, 1105–1109.

Kaufman, H.L., Kirkwood, J.M., Hodi, F.S., Agarwala, S., Amatruda, T., Bines, S.D., Clark, J.I., Curti, B., Ernstoff, M.S., Gajewski, T., et al. (2013). The Society for Immunotherapy of Cancer consensus statement on tumour immunotherapy for the treatment of cutaneous melanoma. Nat. Rev. Clin. Oncol. *10*, 588–598.

Kou, Y., Chen, E.Y., Clark, N.R., Tan, C.M., and Ma'ayan, A. (2013). ChEA2: gene-set libraries from ChIP-X experiments to decode the transcription regulome. Multidisciplinary research and practice for information systems. CD-ARES 2013. Lect. Notes Comput. Sci. *8127*, 416–430.

Latorre, I., Chesney, M.A., Garrigues, J.M., Stempor, P., Appert, A., Francesconi, M., Strome, S., and Ahringer, J. (2015). The DREAM complex promotes gene body H2A.Z for target repression. Genes Dev. *29*, 495–500.

LeRoy, G., Rickards, B., and Flint, S.J. (2008). The double bromodomain proteins Brd2 and Brd3 couple histone acetylation to transcription. Mol. Cell *30*, 51–60.

LeRoy, G., Chepelev, I., DiMaggio, P.A., Blanco, M.A., Zee, B.M., Zhao, K., and Garcia, B.A. (2012). Proteogenomic characterization and mapping of nucleosomes decoded by Brd and HP1 proteins. Genome Biol. *13*, R68.

Lito, P., Rosen, N., and Solit, D.B. (2013). Tumor adaptation and resistance to RAF inhibitors. Nat. Med. *19*, 1401–1409.

Ma, Y., Kurtyka, C.A., Boyapalle, S., Sung, S.S., Lawrence, H., Guida, W., and Cress, W.D. (2008). A small-molecule E2F inhibitor blocks growth in a melanoma culture model. Cancer Res. *68*, 6292–6299.

Matsuda, R., Hori, T., Kitamura, H., Takeuchi, K., Fukagawa, T., and Harata, M. (2010). Identification and characterization of the two isoforms of the vertebrate H2A.Z histone variant. Nucleic Acids Res. *38*, 4263–4273.

Obri, A., Ouararhni, K., Papin, C., Diebold, M.L., Padmanabhan, K., Marek, M., Stoll, I., Roy, L., Reilly, P.T., Mak, T.W., et al. (2014). ANP32E is a histone chaperone that removes H2A.Z from chromatin. Nature *505*, 648–653.

Riker, A.I., Enkemann, S.A., Fodstad, O., Liu, S., Ren, S., Morris, C., Xi, Y., Howell, P., Metge, B., Samant, R.S., et al. (2008). The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. BMC Med. Genomics *1*, 13.

Sadeghi, L., Bonilla, C., Strålfors, A., Ekwall, K., and Svensson, J.P. (2011). Podbat: a novel genomic tool reveals Swr1-independent H2A.Z incorporation at gene coding sequences through epigenetic meta-analysis. PLoS Comput. Biol. *7*, e1002163.

Sansoni, V., Casas-Delucchi, C.S., Rajan, M., Schmidt, A., Bonisch, C., Thomae, A.W., Staege, M.S., Hake, S.B., Cardoso, M.C., and Imhof, A. (2014). The histone variant H2A.Bbd is enriched at sites of DNA synthesis. Nucleic Acids Res. *42*, 6405–6420.

Segura, M.F., Fontanals-Cirera, B., Gaziel-Sovran, A., Guijarro, M.V., Hanniford, D., Zhang, G., González-Gomez, P., Morante, M., Jubierre, L., Zhang, W., et al. (2013). BRD4 sustains melanoma proliferation and represents a new target for epigenetic therapy. Cancer Res. *73*, 6264–6276.

Sinha, A., Faller, D.V., and Denis, G.V. (2005). Bromodomain analysis of Brd2-dependent transcriptional activation of cyclin A. Biochem. J. *387*, 257–269.

Svotelis, A., Gévry, N., and Gaudreau, L. (2009). Regulation of gene expression and cellular proliferation by histone H2A.Z. Biochem. Cell Biol. *87*, 179–188.

Talantov, D., Mazumder, A., Yu, J.X., Briggs, T., Jiang, Y., Backus, J., Atkins, D., and Wang, Y. (2005). Novel genes associated with malignant melanoma but not benign melanocytic lesions. Clin. Cancer Res. *11*, 7234–7242.

Vardabasso, C., Hasson, D., Ratnakumar, K., Chung, C.Y., Duarte, L.F., and Bernstein, E. (2013). Histone variants: emerging players in cancer biology. Cell. Mol. Life Sci. *71*, 379–404.

Wiedemann, S.M., Mildner, S.N., Bönisch, C., Israel, L., Maiser, A., Matheisl, S., Straub, T., Merkl, R., Leonhardt, H., Kremmer, E., et al. (2010). Identification and characterization of two novel primate-specific histone H3 variants, H3.X and H3.Y. J. Cell Biol. *190*, 777–791.

Xu, L., Shen, S.S., Hoshida, Y., Subramanian, A., Ross, K., Brunet, J.P., Wagner, S.N., Ramaswamy, S., Mesirov, J.P., and Hynes, R.O. (2008). Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases. Mol. Cancer Res. *6*, 760–769.

Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature *456*, 125–129.

Zingg, D., Debbache, J., Schaefer, S.M., Tuncer, E., Frommel, S.C., Cheng, P., Arenas-Ramirez, N., Haeusel, J., Zhang, Y., Bonalli, M., et al. (2015). The epigenetic modifier EZH2 controls melanoma growth and metastasis through silencing of distinct tumour suppressors. Nat. Commun. *6*, 6051.

2.2 Investigating unknown and unusual posttranslational modifications by mass spectrometry-based proteomics

Specific addition and removal of posttranslational modifications is an important cellular mechanism to modulate protein structure, localization and function. For instance, in order to become activated, many proteins are dependent on the addition of a specific PTM at a specific site. In the first publication in the area of PTMs, I set out to identify a previously unknown activating PTM. Modifications of unknown composition can naturally not be detected by conventional database searching, hence in this project I applied the unbiased dependent peptide search approach. With this, I discovered the modification that activates elongation factor P in a specific branch of bacteria.

In contrast to the specific enzyme-mediated processes mentioned above, unspecific (non-enzymatic) addition of modifications to proteins can have detrimental effects. Unsurprisingly, such modifications are often associated with diseases. The second PTM project had a clinical focus: I developed a mass spectrometry-based method for investigating protein glycation, an unspecific PTM relevant in diabetes. Since in this case the modification mass was known, I could use a conventional database search approach. I successfully evaluated the HCD fragmentation behavior of glycated peptides on model proteins, and finally applied the method for detecting protein glycation directly in blood plasma.

2.2.1 Identification of the previously unknown modification that activates elongation factor P

Lassak, J., **Keilhauer, E. C.**, Fürst, M., Wuichet, K., Gödeke, J., Starosta, A.L., Chen, J., Søgaard-Andersen, L.,Rohr, D., Wilson, D.N., Häussler, S. Mann, M. & Jung, K.

Arginine-rhamnosylation as new strategy to activate translation elongation factor P
Nature Chemical Biology 2015 Apr; 11(4); 266-270

This highly interesting collaboration project was initiated by Dr. Jürgen Lassak, then a senior PostDoc in the group of Prof. Kirsten Jung. The subject of this study was elongation factor P (EF-P), a protein required to resolve ribosome stalling caused by certain polyproline motifs. To be functional, EF-P needs to be posttranslationally activated. In eukaryotes and certain bacteria like *Escherichia coli* (*E.coli*), the corresponding activation systems are completely understood, and the corresponding modification is known to occur on a particular lysine. However, for other types of bacteria, both the activating enzyme and associated PTM remained elusive. Dr. Jürgen Lassak identified one particularly interesting branch of bacteria with unknown EF-P activation mechanism, characterized by an arginine at the position homologous to the modified lysine in eukaryotes and *E. coli*. He also bioinformatically identified a protein of unknown function strictly co-occuring with this arginine-type EF-P branch. In further experiments, he determined this protein to be the modifying enzyme necessary for EF-P activation in this group of bacteria. The question what kind of modification this enzyme transfers to EF-P to activate it remained unsolved.
At this point he approached me, and I tried to tackle this problem by MS-based proteomics. To that end, I analyzed EF-P produced both in the presence and the absence of the modifying enzyme by LC-MS/MS. Since I had no potential modification mass to use for standard database searching, I applied the dependent peptide search approach. This search mode outputs complex and long lists of potential modifications, hence discovering the needle in the haystack, i.e. the true modification, proved to be challenging. However, using expert knowledge and spectra exploration I determined a promising candidate: attachment of rhamnose, a 6-deoxy-hexose sugar, to the specific arginine residue of EF-P. Subsequently, we further confirmed this modification by reconstituting the modification reaction *in vitro* and analyzing the resulting modified peptides by MS. Together with other biochemical validation experiments performed by Dr. Jürgen

Lassak and coworkers, we unequivocally showed that EF-P in the chosen model system *Schewanella oneidensis* is activated by arginine-rhamnosylation.

nature chemical biology

ARTICLE

PUBLISHED ONLINE: 16 FEBRUARY 2015 | DOI: 10.1038/NCHEMBIO.1751

Arginine-rhamnosylation as new strategy to activate translation elongation factor P

Jürgen Lassak[1,2]*, Eva Keilhauer[3], Maximilian Fürst[1,2], Kristin Wuichet[4], Julia Gödeke[5], Agata L Starosta[1,6], Jhong-Min Chen[7], Lotte Søgaard-Andersen[4], Jürgen Rohr[7], Daniel N Wilson[1,6], Susanne Häussler[5,8], Matthias Mann[3] & Kirsten Jung[1,2]*

Ribosome stalling at polyproline stretches is common and fundamental. In bacteria, translation elongation factor P (EF-P) rescues such stalled ribosomes, but only when it is post-translationally activated. In *Escherichia coli*, activation of EF-P is achieved by (*R*)-β-lysinylation and hydroxylation of a conserved lysine. Here we have unveiled a markedly different modification strategy in which a conserved arginine of EF-P is rhamnosylated by a glycosyltransferase (EarP) using dTDP-L-rhamnose as a substrate. This is to our knowledge the first report of N-linked protein glycosylation on arginine in bacteria and the first example in which a glycosylated side chain of a translation elongation factor is essential for function. Arginine-rhamnosylation of EF-P also occurs in clinically relevant bacteria such as *Pseudomonas aeruginosa*. We demonstrate that the modification is needed to develop pathogenicity, making EarP and dTDP-L-rhamnose–biosynthesizing enzymes ideal targets for antibiotic development.

Ribosomes translate an mRNA sequence into a polypeptide chain. During this process, specific X-PP-X tripeptide sequence motifs can induce ribosome stalling[1–6]. Eukaryotic/archaeal initiation factor 5A (e/aIF5A) and its bacterial ortholog, EF-P, alleviate the stalled ribosomes by binding and stimulating peptide bond formation[3,4,7–10]. With its three β-barrel domains, the L-shaped EF-P is structurally reminiscent of transfer RNA (tRNA)[11] and binds to the ribosome between the sites of peptidyl-tRNA binding (P-site) and tRNA exiting (E-site)[12]. A positively charged residue at the tip of the loop region in domain I of EF-P protrudes toward the peptidyl-transferase center and can reach into it when elongated by modification[12]. Accordingly, the conserved lysine in e/aIF5A is extended to hypusine by deoxyhypusine synthase and deoxyhypusine hydroxylase[13–15]. Analogously, in bacteria such as *E. coli*, the protruding lysine (K34) of EF-P is (*R*)-β-lysinylated and hydroxylated by the concerted action of EF-P lysyl-transferase (EpmA), lysine aminomutase (EpmB) and EF-P hydroxylase (EpmC; **Fig. 1a**)[16–20]. Here we report the identification of an EF-P subfamily activated by a chemically different modification. Using *Shewanella oneidensis* as a model organism, we found rhamnosylation of a conserved arginine. We also identified the corresponding glycosyltransferase, EarP. This modification is not only crucial for bacterial fitness but also for pathogenicity in *P. aeruginosa*, and thus it might be equally important in other clinically relevant species such as *Neisseria gonorrhoea* or *Bordetella pertussis*.

RESULTS

Identification of the EarP-arginine type EF-P subfamily

Whereas the hypusinylation pathway is highly conserved in archaea and eukaryotes, EpmA (also known as YjeA, PoxA, GenX) and EpmB (YjeK) are only found in about 26% of all bacteria[20] (**Supplementary Results**, **Supplementary Figs. 1–5** and **Supplementary Data Set 1**). The third modification enzyme, EpmC (YfcM), co-occurs with EpmA and EpmB but is restricted almost exclusively to γ-proteobacterial genomes (**Supplementary Figs. 4,5** and **Supplementary Data Set 1**) corroborating its minor role in EF-P function[3,6–8,21]. We hypothesized that the genes encoding EF-P and the associated modification system have coevolved. In a phylogenetic analysis of EF-P sequences, we identified a distinct subfamily, encoded in genomes lacking EpmABC orthologs, that has a strictly conserved arginine (R32) in the position equivalent to K34 in *E. coli* (**Fig. 1b,c**, **Supplementary Fig. 4** and **Supplementary Data Set 1**). The members of this subfamily represent about 9% of all EF-Ps, but the distribution deviates from the currently accepted species phylogeny. As the newly identified EF-P branch encompasses all β-proteobacteria, we hypothesize this subdivision as the phylogenetic origin, with subsequent horizontal transfer into several γ-proteobacterial orders (including *Pseudomonadales*, *Aeromonadales* and *Alteromonadales*) as well as some Fusobacteria, Planctomycetes and Spirochetes (**Supplementary Data Set 1**). We took advantage of the anomalous EF-P phylogeny by searching for putative EF-P modification enzymes associated with this subfamily via gene neighborhood and co-occurrence using STRING[22]. This led us to identify a protein with a conserved domain of unknown function (DUF 2331), which we designated as EarP. Its distribution strictly coincides with the newly identified EF-P subfamily (**Supplementary Fig. 5**). Moreover, the corresponding gene, *earP*, always lies within a four-gene distance to *efp*; in 94% of cases, both genes are directly adjacent (**Supplementary Fig. 6a**).

EF-P and EarP are functionally linked

To investigate whether EF-P and EarP are functionally linked, we used the ubiquitous, facultative anaerobic, alteromonadal γ-proteobacterium *S. oneidensis*. Bacteria of the genus *Shewanella*

[1]Center for Integrated Protein Science Munich, Ludwig-Maximilians-Universität München, Munich, Germany. [2]Department of Biology I, Microbiology, Ludwig-Maximilians-Universität München, Martinsried, Germany. [3]Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried, Germany. [4]Max Planck Institute for Terrestrial Microbiology, Marburg, Germany. [5]Institute for Molecular Bacteriology, Twincore, Centre for Clinical and Experimental Infection Research, a joint venture of the Helmholtz Centre of Infection Research and the Hannover Medical School, Hannover, Germany. [6]Gene Center, Department for Biochemistry, Ludwig-Maximilians-Universität München, Munich, Germany. [7]Department of Pharmaceutical Sciences, College of Pharmacy, University of Kentucky, Lexington, Kentucky, USA. [8]Department of Molecular Bacteriology, Helmholtz Centre for Infection Research, Braunschweig, Germany *e-mail: juergen.lassak@lmu.de or jung@lmu.de

ARTICLE NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.1751

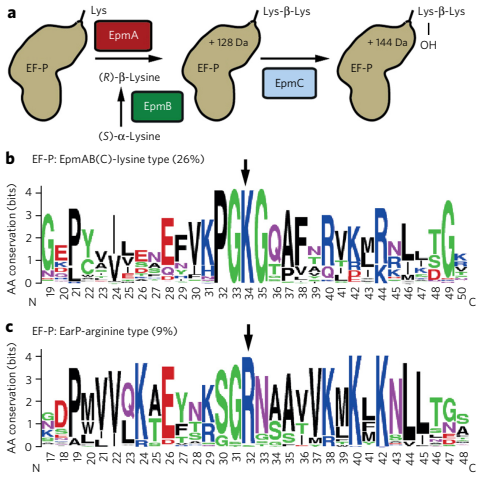


Figure 1 | Bioinformatic identification of the EarP-arginine type EF-P subfamily. (**a**) (*R*)-β-lysinylation and hydroxylation of EF-P in *E. coli*. (**b**,**c**) A 31-amino-acid (aa)-long sequence logo of EF-P encompassing a part of domain I, including the loop region. Arrows point to the positively charged residue at the tip of the loop region. Numbering depicts aa in the polypeptide chain. (**b**) Weblogo generated from EF-P of bacteria encoding EpmA and EpmB. (**c**) Weblogo generated from EF-P of bacteria derived from the newly identified branch that co-occurs with EarP.

are commonly used in microbial fuel cells and have high potential in bioremediation because of their ability to use a wide range of terminal electron acceptors, including heavy metals[23,24].

In a first step, we generated markerless in-frame deletions of *S. oneidensis efp* (locus tag SO_2328) and *earP* (locus tag SO_2329) and phenotypically characterized the resulting mutant strains, $\Delta efp_{S.o.}$ and $\Delta earP_{S.o.}$, respectively. Bacteria lacking *efp*, such as *E. coli* or *Agrobacterium tumefaciens*, have diminished growth rates[16,25]. In line with these results, deleting either $efp_{S.o.}$ or $earP_{S.o.}$ increased the doubling time from 40 min to 110 min and 70 min, respectively, which was reversed by providing the corresponding gene copy in *trans* (**Fig. 2a** and **Supplementary Fig. 6a**). In parallel, we analyzed the growth of the $\Delta efp_{S.o.}$ strain, which encodes an EF-P variant where the strictly conserved R32 (**Fig. 1c**) was substituted by either lysine (R32K) or alanine (R32A). Both strains phenocopied $\Delta efp_{S.o.}$, demonstrating the importance of this conserved arginine for EF-$P_{S.o.}$ function (**Fig. 2a** and **Supplementary Fig. 6a**). We also investigated whether the synthesis of polyproline-containing proteins is affected in the absence of $efp_{S.o.}$ and $earP_{S.o.}$. Therefore, we used the reporter plasmid p3LC-TL30-3P, which encodes a LacZ variant that is preceded by a stretch of three proline residues (**Fig. 2b**)[3,7]. As expected, the β-galactosidase activities of the $\Delta efp_{S.o.}$ or $\Delta earP_{S.o.}$ strains were both reduced by about tenfold, providing clear experimental evidence that EarP is required for EF-P activity.

EarP is sufficient to activate EF-P

To test whether $EarP_{S.o.}$ is sufficient for activation of EF-$P_{S.o.}$, we examined the phenotypes of an *E. coli efp* deletion ($\Delta efp_{E.c.}$) heterologously producing EF-$P_{S.o.}$ and $EarP_{S.o.}$. As a readout, we used a chromosomal P_{cadBA}-dependent *lacZ* reporter in which activation of P_{cadBA} strictly depends on the pH-responsive transcriptional activator CadC[26]. Translation of CadC is impaired in cells lacking active EF-P because of the presence of a polyproline motif (**Supplementary Fig. 6b**)[7]. As expected, the *E. coli* $\Delta efp_{E.c.}$ P_{cadBA}::*lacZ* strain was characterized by low-level β-galactosidase activity, and wild-type–like *lacZ* expression was restored in the presence of a plasmid expressing EF-$P_{E.c.}$. Whereas β-galactosidase activity remained low when EF-$P_{S.o.}$ or $EarP_{S.o.}$ were expressed alone, simultaneous production of both proteins complemented for the lack of EF-$P_{E.c.}$. The diminished growth rate of the $\Delta efp_{E.c.}$ mutant[18] was consistently eliminated when EF-$P_{S.o.}$ and $EarP_{S.o.}$ were produced together, regardless of the presence of the *E. coli* EF-P modification enzymes EpmA or EpmB (**Supplementary Fig. 6c**). Next, we asked whether EpmABC and EarP had adapted to specifically activate their corresponding EF-Ps. Therefore EF-$P_{E.c.}$ K34 and EF-$P_{S.o.}$ R32 substitution variants were produced in $\Delta efp_{E.c.}$, and P_{cadBA} activation, growth rate or both were investigated in the resultant strains (**Supplementary Fig. 6b**,**c**). Neither EF-$P_{S.o.}^{R32K}$ with EpmABC nor the EF-$P_{E.c.}^{K34R}$ with EarP could reverse the mutant phenotype, thus further corroborating our hypothesis for the coevolution of EF-P with its associated modification system.

EarP modifies EF-P at Arg32

Having demonstrated that $EarP_{S.o.}$ is necessary and sufficient for specific activation of EF-$P_{S.o.}$, we addressed whether $EarP_{S.o.}$ post-translationally modifies the conserved R32 of EF-$P_{S.o.}$. To that end, we overproduced His_6-tagged EF-$P_{S.o.}$ in two *S. oneidensis* strains, $\Delta efp_{S.o.}$ and $\Delta efp_{S.o.}/\Delta earP_{S.o.}$. Purified EF-$P_{S.o.}$ from these two strains was proteolytically digested, and the resulting peptides were analyzed by high-resolution LC/MS/MS using an unbiased 'dependent peptide' search[27]. We detected eight high-confidence R32-containing peptides that were 146.058 Da heavier than their unmodified counterparts (**Fig. 3** and **Supplementary Fig. 7**). This mass shift only occurred on R32-containing peptides of EF-$P_{S.o.}$ and never when EF-$P_{S.o.}$ was produced in cells lacking EarP (**Supplementary Fig. 8**). The fragmentation pattern of the modified peptides strongly suggested the modification site to be on R32 (**Fig. 3**). To further confirm this, we performed a standard variable modification search with a potential arginine mass shift of 146.058 Da, corresponding to a molecular composition of $C_6H_{10}O_4$ (146.0579 Da). This second analysis identified the modification to exist exclusively on R32 of EF-P, confirming both the molecular

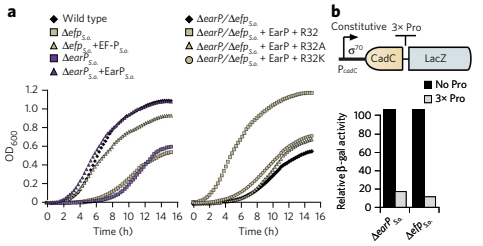


Figure 2 | Phenotypic analysis of *S. oneidensis* MR-1 *earP* and *efp* deletion mutants. (**a**) Growth of *S. oneidensis* MR-1 strains. Left, wild-type strains in comparison to $\Delta efp_{S.o.}$ and $\Delta earP_{S.o.}$ deletion strains and after complementation in *trans* (+$efp_{S.o.}$, +$earP_{S.o.}$). Right, the $\Delta efp_{S.o.}$ deletion strain and after complementation with plasmids encoding His_6 versions of EF-$P_{S.o.}$ (+R32) or the corresponding substitution variants EF-$P_{S.o.}^{R32A}$ (+R32A) and EF-$P_{S.o.}^{R32K}$ (+R32K), respectively. The presented growth curves are average data from three independent data sets, with statistical error below 10%. (**b**) β-galactosidase (β-gal) activity assay of *S. oneidensis* MR-1 $\Delta efp_{S.o.}$, $\Delta earP_{S.o.}$ encoding a constitutively produced LacZ-hybrid without (black bars) or with (gray bars) a polyproline motif (3× Pro). β-galactosidase activity is given in percent and is normalized to the wild-type values. The relative activities are average data from three independent data sets, with statistical error below 10%.

NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.1751 ARTICLE

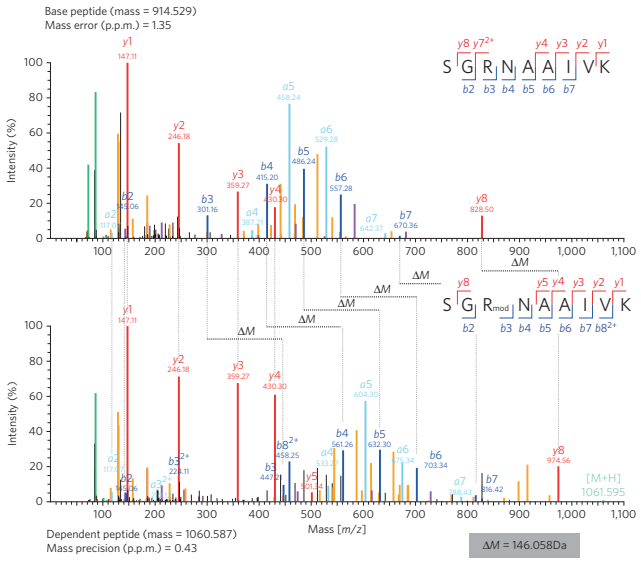


Figure 3 | Dependent peptide MS analysis of the *S. oneidensis* MR-1 EF-P modification. C-terminally His_6-tagged $EF\text{-}P_{S.o.}$ was analyzed upon homologous overproduction. Modified R32-containing 'dependent' peptides (146.058 Da heavier) were identified by a characteristic Δ*M* mass shift of the precursor and several fragment ions when compared to the unmodified 'base peptide'. MS/MS spectra of the best-scoring base peptide–dependent peptide pair are shown. Peak colors: red, blue and light blue refer to *y*, *b* and *a* ions (according to Roepstorff-Fohlmann-Biemann nomenclature), respectively; light green, molecular ion; yellow, neutral losses; purple, internal fragments; green, immonium ions, black peaks, unassigned.

composition and the location of the modification (**Supplementary Fig. 9**). Notably, no peptides with this modification could be found when two arginine substitution EF-P variants were analyzed ($His_6\text{-}EF\text{-}P_{S.o.}^{R32K}$ and $His_6\text{-}EF\text{-}P_{S.o.}^{R32A}$; **Supplementary Fig. 10**). This observation explains the *efp*-null mutant phenotype of strains encoding $His_6\text{-}EF\text{-}P_{S.o.}^{R32K}$ or $His_6\text{-}EF\text{-}P_{S.o.}^{R32A}$ substitution variants (**Fig. 2a** and **Supplementary Fig. 6**) as a consequence of the absence of $EF\text{-}P_{S.o.}$ R32 modification by $EarP_{S.o.}$.

EarP is a rhamnosyltransferase

We reasoned that a modification on R32 with a molecular composition of $C_6H_{10}O_4$ might be the result of N-glycosylation with an activated deoxyhexose sugar ($C_6H_{12}O_5 - H_2O = C_6H_{10}O_4$). *E. coli* synthesizes two nucleotide diphosphate deoxyhexoses: GDP-L-fucose and dTDP-L-rhamnose (**Fig. 4a**)[28]. The biosynthesis genes encoding the latter (represented by *rmlD*) are strictly conserved in bacteria encoding EarP but are less frequently found in bacteria with EpmABC (**Supplementary Fig. 5**). To test whether dTDP-L-rhamnose or GDP-L-fucose might act as a substrate for $EarP_{S.o.}$, we interrupted the corresponding synthesis pathways. GDP-L-fucose and dTDP-L-rhamnose are synthesized from fructose-6-phosphate and glucose-1-phosphate, respectively, with each biosynthetic pathway encompassing four specific enzymes (**Fig. 4a**). To prevent formation of GDP-L-fucose, we deleted *fcl*, which arrested synthesis at the GDP-4-keto-6-deoxy-D-mannose step, whereas dTDP-L-rhamnose formation was blocked at the intermediate dTDP-4-keto-6-deoxy-L-mannose step by deletion of *rmlD* (**Fig. 4a**). The deletions were generated in an *E. coli* P_{cadBA}::*lacZ* strain lacking *efp*. The β-galactosidase activities of both resultant strains deficient either in GDP-L-fucose ($\Delta efp_{E.c.}/\Delta fcl_{E.c.}$) or in dTDP-L-rhamnose ($\Delta efp_{E.c.}/\Delta rmlD_{E.c.}$) were comparable to wild-type activity when a copy of $efp_{E.c.}$ was provided in *trans* (**Fig. 4b**). Similarly, the $\Delta efp_{E.c.}/\Delta fcl_{E.c.}$ strain could be complemented with $EF\text{-}P_{S.o.}/EarP_{S.o.}$, indicating that GDP-L-fucose is not a substrate for EarP. In stark contrast, the coproduction of $EF\text{-}P_{S.o.}/EarP_{S.o.}$ in the context of $\Delta efp_{E.c.}/\Delta rmlD_{E.c.}$ cells phenocopied the $\Delta efp_{E.c.}$ strain, suggesting that dTDP-L-rhamnose is the substrate needed for $EarP_{S.o.}$ to activate $EF\text{-}P_{S.o.}$ To further corroborate this result, we deleted *rmlC* (locus tag SO_3160) in *S. oneidensis* and analyzed translation of a polyproline-containing LacZ reporter hybrid in the resultant dTDP-L-rhamnose–deficient strain $\Delta rmlC_{S.o.}$ (**Fig. 4c**)[3,7]. As observed for $\Delta efp_{S.o.}$ or $\Delta earP_{S.o.}$, β-galactosidase activity was low in the $\Delta rmlC_{S.o.}$ strain harboring the p3LC-TL30-3P reporter. Consistent with our previous results, *in vivo* activation of $EF\text{-}P_{S.o.}$ strictly depends on the biosynthesis of dTDP-L-rhamnose, leading us to conclude that dTDP-L-rhamnose acts as the substrate for $EF\text{-}P_{S.o.}$ modification and that it was recruited for this role upon the development of the EarP–EF-P phylogenetic relationship (**Supplementary Figs. 4**,**5** and **Supplementary Data Set 1**).

To directly demonstrate that $EarP_{S.o.}$ can glycosylate $EF\text{-}P_{S.o.}$ using dTDP-L-rhamnose as a substrate, we performed *in vitro* glycosylation reactions with purified components. LC/MS/MS analysis, performed as described above, revealed the presence of R32 rhamnosylation of wild-type $EF\text{-}P_{S.o.}$ if and only if all three components were provided (**Supplementary Fig. 11**). Collectively, our data demonstrate unambiguously both *in vivo* and *in vitro* that EarP is an EF-P arginine rhamnosyltransferase essential for post-translational activation.

EF-P stimulates peptide bond formation indirectly

Rhamnosyl-arginine differs substantially from (*R*)-β-lysinyl-hydroxylysine and hypusine of $EF\text{-}P_{E.c.}$ and a/eIF5A, respectively, raising the question of how this unusual extension protrudes into the peptidyl-transferase center of the ribosome (**Fig. 5a**–**d**). To investigate this, we generated molecular models for the different modifications of EF-P orthologs based on the crystal structure of unmodified *Thermus thermophilus* EF-P bound to the 70S ribosome[12]. These models suggest that the (*R*)-β-lysinylation found on $EF\text{-}P_{E.c.}$ could reach within 2 Å of the proline attached to the P-site tRNA (**Fig. 5b**), whereas the hypusine and rhamnose-arginine modifications are shorter and cannot reach the P-site proline (**Fig. 5c**,**d**). Therefore, EF-P bearing either hypusine or rhamnose-arginine modifications is not likely to stimulate peptide-bond formation by directly influencing the conformation of the polypeptide chain but rather does so indirectly by interacting with and stabilizing the CCA-end of the P-site peptidyl-Pro-tRNA.

EarP and EF-P are essential for *P. aeruginosa* pathogenicity

Distinct bacterial strategies to functionalize EF-P may provide a basis for development of customized antibiotics. Deleting EF-P or its modifying enzymes has been shown to reduce bacterial fitness[16,25] and lead to a loss of pathogenicity in *Salmonella enterica* and *A. tumefaciens*. To test whether rhamnosylated EF-P is also required

ARTICLE NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.1751

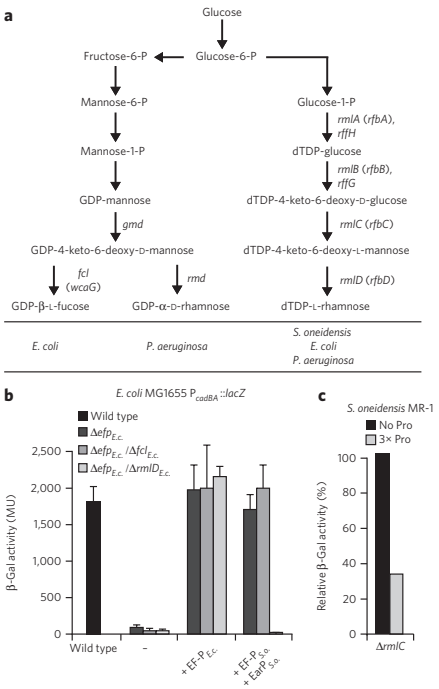


Figure 4 | *In vivo* analysis of *S. oneidensis* MR-1 EF-P functionality depending on NDP-deoxyhexose biosyntheses. (**a**) Biosynthesis pathways for dTDP-L-rhamnose, GDP-D-rhamnose and GDP-L-fucose. Arrows depict sugar conversion steps. Specific conversion steps are associated with the corresponding biosynthesis gene. Paralogous genes are separated by a comma, and alternative names are given in parentheses. (**b**) β-galactosidase (β-gal) activity of *E. coli* P_{cadBA}-*lacZ* reporter wild-type (WT) and *efp* deletion strain ($\Delta efp_{E.c.}$) as well as the $\Delta efp_{E.c.}/\Delta fcl_{E.c.}$ (GDP-L-fucose deficient) and $\Delta efp_{E.c.}/\Delta rmlD_{E.c.}$ (dTDP-L-rhamnose deficient) double deletion mutants and after complementation either with EF-$P_{E.c.}$ or EF-$P_{S.o.}$ in combination with $EarP_{S.o.}$. Cells were incubated under *cadBA*-inducing conditions. Data represent mean values from three independent replicates ± s.d. (**c**) β-galactosidase activity assay of *S. oneidensis* MR-1 $\Delta rmlC_{S.o.}$ encoding a constitutively produced LacZ hybrid without (black bars) or with (gray bars) a polyproline motif (3× Pro). β-galactosidase activity is given in percent and is normalized to the wild-type values. The relative activities are average data from three independent data sets, with statistical error below 10%.

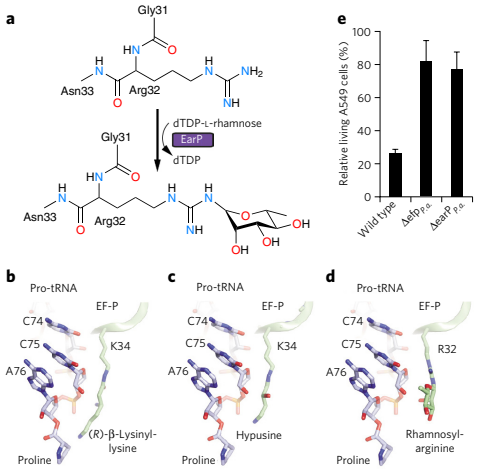


Figure 5 | EF-P rhamnosylation, mode of action and impact on pathogenicity. (**a**) Arginine-rhamnosylation by EarP using dTDP-L-rhamnose as substrate. (**b**–**d**) Models of different modified EF-P proteins bound to the ribosome. The CCA-end of P-site bound Pro-tRNA (blue) is shown for reference. Models based on *T. thermophilus* EF-P–70S structure[12]. (**b**) K34 of *E. coli* EF-P post-translationally modified with (*R*)-β-lysine. (**c**) EF-P bearing the K34 hypusine modification. (**d**) R32 of EF-P modified by L-rhamnose. (**e**) Effects of $\Delta efp_{P.a.}$ and $\Delta earP_{P.a.}$ on *P. aeruginosa* pathogenicity. Cytotoxicity of *P. aeruginosa* strains was assessed by infecting A549-Gluc cells, which secrete Gaussia luciferase, as a measure of cell integrity. Data represent mean values from three independent replicates ± s.d.

to develop pathogenicity in the *P. aeruginosa* strain PAO1, we investigated transposon mutants of $efp_{P.a.}$ (locus tag PA2851) and $earP_{P.a.}$ (locus tag PA2852) in an infection assay using the human cell line A549-Gluc (**Fig. 5e**). Whereas wild-type *P. aeruginosa* decreased the number of living cells by about 80%, infection with $\Delta efp_{P.a.}$ or $\Delta earP_{P.a.}$ mutants had no effect on cell viability. Pathogenicity of *P. aeruginosa* PAO1 is dependent on a large number of cell-associated and extracellular virulence factors, such as rhamnolipids and pyocyanin, that are important for colonization and invasion during infection[29]. Bioinformatic analysis on the *P. aeruginosa* proteome revealed that the synthesis of those virulence factors involves polyproline-containing proteins, suggesting a dependence on EF-P for their translation (**Supplementary Table 4**). Consistently, $efp_{P.a.}$ or $earP_{P.a.}$ disruption mutants showed a substantial decrease in the production of rhamnolipids and pyocyanin, and production was restored by introducing $EarP_{S.o.}$ and EF-$P_{S.o.}$ but not the substitution mutants EF-$P_{S.o.}^{R32A}$ and EF-$P_{S.o.}^{R32K}$ (**Supplementary Fig. 12**). Therefore, both EF-$P_{P.a.}$ and the corresponding rhamnosyltransferase $EarP_{P.a.}$ contribute to pathogenicity in *P. aeruginosa*.

DISCUSSION

Protein glycosylation is a commonly used strategy to alter structural and functional properties of a protein. However, until recently, N-linked glycosylation was almost exclusively associated with asparagine. The only known additions of a sugar to arginine were restricted to two reported examples on eukaryotic proteins. Arginine glycosylation was discovered first in search of a protein primer for starch synthesis in 1995 (ref. 30). Here, the authors identified sweet corn amylogenin to be self-β-glucosylated. Second, in 2013, two independent research groups showed that NleB, an enteropathogenic *E. coli* type III secretion system effector, antagonizes death receptor signaling by modifying conserved arginines in human death receptor domains with N-acetylglucosamine (GlcNAc)[31,32]. With *S. oneidensis* and *P. aeruginosa* EF-P, we now report what is to our knowledge the first arginine-glycosylated bacterial protein, thus demonstrating that this type of post-translational modification is not restricted to eukaryotic proteins but is common to other domains of life.

NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.1751 ARTICLE

Received 8 September 2014; accepted 22 December 2014; published online 16 February 2015

METHODS

Methods and any associated references are available in the online version of the paper.

References

1. Tanner, D.R., Cariello, D.A., Woolstenhulme, C.J., Broadbent, M.A. & Buskirk, A.R. Genetic identification of nascent peptides that induce ribosome stalling. *J. Biol. Chem.* **284**, 34809–34818 (2009).
2. Woolstenhulme, C.J. *et al.* Nascent peptides that block protein synthesis in bacteria. *Proc. Natl. Acad. Sci. USA* **110**, E878–E887 (2013).
3. Peil, L. *et al.* Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. *Proc. Natl. Acad. Sci. USA* **110**, 15265–15270 (2013).
4. Hersch, S.J. *et al.* Divergent protein motifs direct elongation factor P–mediated translational regulation in *Salmonella enterica* and *Escherichia coli*. *mBio* **4**, e00180–e00113 (2013).
5. Elgamal, S. *et al.* EF-P dependent pauses integrate proximal and distal signals during translation. *PLoS Genet.* **10**, e1004553 (2014).
6. Starosta, A.L. *et al.* Translational stalling at polyproline stretches is modulated by the sequence context upstream of the stall site. *Nucleic Acids Res.* **42**, 10711–10719 (2014).
7. Ude, S. *et al.* Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science* **339**, 82–85 (2013).
8. Doerfel, L.K. *et al.* EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. *Science* **339**, 85–88 (2013).
9. Gutierrez, E. *et al.* eIF5A promotes translation of polyproline motifs. *Mol. Cell* **51**, 35–45 (2013).
10. Saini, P., Eyler, D.E., Green, R. & Dever, T.E. Hypusine-containing protein eIF5A promotes translation elongation. *Nature* **459**, 118–121 (2009).
11. Hanawa-Suetsugu, K. *et al.* Crystal structure of elongation factor P from *Thermus thermophilus* HB8. *Proc. Natl. Acad. Sci. USA* **101**, 9595–9600 (2004).
12. Blaha, G., Stanley, R.E. & Steitz, T.A. Formation of the first peptide bond: the structure of EF-P bound to the 70S ribosome. *Science* **325**, 966–970 (2009).
13. Park, M.H., Cooper, H.L. & Folk, J.E. Identification of hypusine, an unusual amino acid, in a protein from human lymphocytes and of spermidine as its biosynthetic precursor. *Proc. Natl. Acad. Sci. USA* **78**, 2869–2873 (1981).
14. Park, M.H., Cooper, H.L. & Folk, J.E. The biosynthesis of protein-bound hypusine (*N*-ε-(4-amino-2-hydroxybutyl)lysine). Lysine as the amino acid precursor and the intermediate role of deoxyhypusine (*N*-ε-(4-aminobutyl) lysine). *J. Biol. Chem.* **257**, 7217–7222 (1982).
15. Cooper, H.L., Park, M.H., Folk, J.E., Safer, B. & Braverman, R. Identification of the hypusine-containing protein hy+ as translation initiation factor eIF-4D. *Proc. Natl. Acad. Sci. USA* **80**, 1854–1857 (1983).
16. Navarre, W.W. *et al.* PoxA, YjeK, and elongation factor P coordinately modulate virulence and drug resistance in *Salmonella enterica*. *Mol. Cell* **39**, 209–221 (2010).
17. Gilreath, M.S. *et al.* β-Lysine discrimination by lysyl-tRNA synthetase. *FEBS Lett.* **585**, 3284–3288 (2011).
18. Yanagisawa, T., Sumida, T., Ishii, R., Takemoto, C. & Yokoyama, S. A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P. *Nat. Struct. Mol. Biol.* **17**, 1136–1143 (2010).
19. Peil, L. *et al.* Lys34 of translation elongation factor EF-P is hydroxylated by YfcM. *Nat. Chem. Biol.* **8**, 695–697 (2012).
20. Bailly, M. & de Crecy-Lagard, V. Predicting the pathway involved in post-translational modification of elongation factor P in a subset of bacterial species. *Biol. Direct* **5**, 3 (2010).
21. Bullwinkle, T.J. *et al.* (*R*)-β-Lysine–modified elongation factor P functions in translation elongation. *J. Biol. Chem.* **288**, 4416–4423 (2013).
22. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
23. Hau, H.H. & Gralnick, J.A. Ecology and biotechnology of the genus *Shewanella*. *Annu. Rev. Microbiol.* **61**, 237–258 (2007).
24. Fredrickson, J.K. *et al.* Towards environmental systems biology of *Shewanella*. *Nat. Rev. Microbiol.* **6**, 592–603 (2008).
25. Peng, W.T., Banta, L.M., Charles, T.C. & Nester, E.W. The *chvH* locus of *Agrobacterium* encodes a homologue of an elongation factor involved in protein synthesis. *J. Bacteriol.* **183**, 36–45 (2001).
26. Haneburger, I., Eichinger, A., Skerra, A. & Jung, K. New insights into the signaling mechanism of the pH-responsive, membrane-integrated transcriptional activator CadC of *Escherichia coli*. *J. Biol. Chem.* **286**, 10681–10689 (2011).
27. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
28. Mäki, M. & Renkonen, R. Biosynthesis of 6-deoxyhexose glycans in bacteria. *Glycobiology* **14**, 1R–15R (2004).
29. Jimenez, P.N. *et al.* The multiple signaling systems regulating virulence in *Pseudomonas aeruginosa*. *Microbiol. Mol. Biol. Rev.* **76**, 46–65 (2012).
30. Singh, D.G. *et al.* β-Glucosylarginine: a new glucose-protein bond in a self-glucosylating protein from sweet corn. *FEBS Lett.* **376**, 61–64 (1995).
31. Li, S. *et al.* Pathogen blocks host death receptor signalling by arginine GlcNAcylation of death domains. *Nature* **501**, 242–246 (2013).
32. Pearson, J.S. *et al.* A type III effector antagonizes death receptor signalling during bacterial gut infection. *Nature* **501**, 247–251 (2013).

Acknowledgments

We would like to thank I. Weitl for excellent technical assistance. This work was supported by the Deutsche Forschungsgemeinschaft: Center for Integrated Protein Science Munich (CiPSM; Exc114/2 to K.J. and WI3285/4-1 to D.N.W.) and the Max Planck Society (to E.K., K.W., L.S.-A. and M.M.). A.L.S. is funded by an AXA Research Fund Postdoctoral Fellowship. The work of S.H. and J.G. was supported by the Helmholtz Association and the Bundesministerium für Bildung und Forschung. J.-M.C. and J.R are funded by the US National Institutes of Health (CA 091901).

Author contributions

J.L. and K.J. designed the experiments and wrote the manuscript. J.L., A.L.S. and D.N.W. discovered EarP and arginine-type EF-P using bioinformatics. K.W., J.L. and L.S.-A. performed phylogenetic analyses. J.L. and M.F. generated *S. oneidensis* and *E. coli* deletion mutants as well as all of the plasmids used in the study. All phenotypic analysis was performed by J.L. and M.F. J.L. purified proteins for MS analysis, which was performed by E.K. and analyzed by E.K. and M.M. TDP-L-rhamnose was synthesized by J.-M.C. and J.R. for *in vitro* glycosylation performed by J.L. Modeling of modified EF-P to the ribosome was done by D.N.W. Phenotypic characterization of *P. aeruginosa* Δ*efp* and Δ*earP* mutants was performed by J.G. and analyzed by J.G. and S.H.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to J.L. or K.J.

ONLINE METHODS

Bioinformatics software. Hidden Markov model (HMM) analyses were carried out using the HMMER3 software package[33]. Multiple sequence alignments were constructed using the l-ins-i algorithm of the MAFFT version 6.864b software package[34]. BLASTP searches were performed using the BLAST+ software package version 2.2.26 (ref. 35). All phylogenetic trees were constructed using FastTree with default settings[36]. Sequences logos were created using the Weblogo server[37].

Genome set and domain architecture. 1,611 completely sequenced prokaryote genomes from a previously defined set available from 4 April 2012 were collected[38]. We used a set of 1,004 genomes with reduced redundancy based on the 16S comparison for all sequence analyses. The Pfam26 HMM library[39] was used to define domain architecture of all sequences with default gathering thresholds. In the event of domain overlaps, the highest-scoring domain model was chosen for the final architecture.

EarP, EF-P, RmlD and EpmC were identified by collecting sequences that contain the DUF2331, EF-P, RmlD_sub_bind and DUF462 domains, respectively. EpmA consists of a class II tRNA synthetase domain (tRNA-synt_2), which is found in a variety of proteins. All 2,684 sequences containing the tRNA-synt_2 domain were collected and aligned. The core region corresponding to the tRNA-synt_2 domain was extracted from the multiple sequence alignment and used to build a phylogenetic tree. Domain architecture and gene neighborhood analyses identified a conserved clade of 239 sequences in the tree that were determined to be EpmA homologs because of their genomic context association with EF-P homologs (**Supplementary Fig. 1**). EpmB is composed of a Radical SAM domain (Radical_SAM), which is found in over 15,000 proteins in our sequence set, making a phylogenetics-based approach to EpmB identification challenging. EpmB is often encoded near *efp* or *epmA*. We collected all of the sequences with Radical_SAM domains that were encoded within a distance of four genes to *efp* or *epmA* homologs (181 sequences) and aligned them. A tree constructed from the alignment revealed two distinct clades, one with many short branches presumed to be true EpmB sequences and another composed of many long branching sequences, which suggest divergence. Furthermore, some sequences in the divergent clade are from genomes represented in the EpmB-associated clade, supporting that they are not true EpmB orthologs (**Supplementary Fig. 2**). These 36 sequences of the divergent clade were removed from the set, and the remaining sequences were used as queries in BLASTP searches against our representative sequence set. All sequences with an *e*-value of 0.0001 or less were collected (515 sequences) and aligned. A phylogenetic tree was built from the core region of the multiple sequence alignment corresponding to the Radical_SAM domain. Conserved clades that were associated with EF-P or EpmA on the basis of genome context and phylogenetic distribution were identified (**Supplementary Fig. 3**). The 236 sequences that are members of these clades were defined as EpmB homologs, and this is consistent with the 239 EpmA homologs we identified that are presumably part of the same pathway. All EarP, EF-P, RmlD, EpmC, EpmA and EpmB homologs in the representative genome set can be found in **Supplementary Data Set 1**. We further identified all of the EarP homologs in the full 1,611 genome set on the basis of the presence of the DUF2331 domain (**Supplementary Data Set 2**). We identified all of the EarP-associated EF-Ps in this set using a HMM built from an alignment of the EarP-associated EF-Ps identified in the EF-P phylogenetic analyses. All sequences with a score greater than or equal to that of the lowest-scoring member of the representative set of EarP-associated EF-Ps were identified as homologs (**Supplementary Data Set 2**).

Phylogenetic analysis of EF-P homologs. A phylogenetic tree was built from the core region of a multiple sequence alignment of EF-P homologs. Conserved clades associated with EarP, EpmA or EpmB on the basis of genome context and phylogenetic distribution data were identified (**Supplementary Fig. 4**). Sequences from these clades were collected, with the exception of those from an EpmA/EpmB-associated subfamily that includes members of the EF-P–like family (the YeiP subfamily), which lack the conserved lysine. The collected sequences were aligned, and the core region of the alignment was used to construct a phylogenetic tree (**Supplementary Fig. 5**).

Oligonucleotides, plasmids and bacterial strain construction. Primers, plasmids and strains used in this study are listed in **Supplementary Tables 1–3**. Of note, transposon mutant *P. aeruginosa* PAO1 ID 30853 from the Washington Genome Center[40] with a transposon insertion in open reading frame (ORF) PA4684 was used as a wild-type control[41]. Strains BW-Δ*efp*/*epmA*::npt, BW-Δ*efp*/*epmB*::*cat*, MG-CR-*efp-fcl* and MG-CR-*efp-rmlD* were constructed by using pRED/ET recombination technology together with *rpsL* counterselection[42] in accordance to the technical protocol of the Quick and Easy *E. coli* Gene Deletion Kit of Gene Bridges (http://www.genebridges.com/). Strains $\Delta efp_{S.o.}$, $\Delta earP_{S.o.}$, $\Delta efp_{S.o.}/\Delta earP_{S.o.}$ and $\Delta rmlC_{S.o.}$ were constructed as essentially described in ref. 43, leaving terminal sections of the target gene.

Molecular biology methods. Enzymes and kits were used according to the manufacturer's directions. Genomic DNA was purified according to standard protocols. DNA fragments were purified from agarose gels using a high-yield PCR cleanup and gel extraction kit (Sued-Laborbedarf). Restriction endonucleases were purchased from New England Biolabs. Sequence amplifications by PCR were performed by using the Phusion high-fidelity DNA polymerase from Finnzymes or the Taq DNA polymerase from New England Biolabs, respectively. All *efp* mutants were constructed by one- or two-step PCR using mismatched primer pairs[44].

Growth conditions. *E. coli*, *P. aeruginosa* PAO1 and *S. oneidensis* MR-1 were routinely grown at 37 °C (for *E. coli* and *P. aeruginosa*) and 30 °C (*S. oneidensis*), unless indicated otherwise. According to the NaCl modification of Miller, lysogeny broth (LB)[45] was used as complex medium. When indicated, LB was buffered with 100 mM sodium-phosphate to pH 5.8. Microaerobic conditions were achieved by growing cells in closed Eppendorf cups with minimal agitation. Antibiotics were used when necessary with the following concentrations: 100 μg/ml ampicillin sodium salt, 50 μg/ml kanamycin sulfate, 34 μg/ml chloramphenicol, 50 μg/ml streptomycin sulfate or 15 μg/ml tetracycline hydrochloride. For blue-white selection, LB agar plates were additionally supplemented with 80 μM 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-Gal; Sigma Aldrich) and 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG; Sigma Aldrich).

β-Galactosidase activity assay. Cells expressing *lacZ* under control of the *cadC* or *cadBA* promoters were grown in buffered LB medium to mid-exponential growth phase or overnight and harvested by centrifugation. β-Galactosidase activities were determined for at least three independent experiments and are given in Miller units (MU). The significance of the results was determined by applying the two-sided Student's *t*-test, and results were considered significantly different if $P < 0.05$.

Overproduction and purification of recombinant proteins. C-terminal His_6-tagged $EF\text{-}P_{S.o.}$ as well as the corresponding R32A and R32K substitution variants were overproduced either in *S. oneidensis* MR-1 wild-type, $\Delta efp_{S.o.}$ and $\Delta efp_{S.o.}/\Delta earP_{S.o.}$ mutant strains or in *E. coli* LMG194 and grown in LB overnight at 16 °C after the addition of 0.2% (w/v) L-arabinose to exponentially grow cells. Similarly, C-terminal His_6-tagged $EarP_{S.o.}$ was produced in *E. coli* LMG194. Cells were lysed and purified using Ni-NTA (Qiagen) and 250 mM imidazole. For further MS analysis and *in vitro* rhamnosylation, proteins were dialyzed against reaction buffer (50 mM HEPES, pH 7.8, 100 mM NaCl, 50 mM KCl, 10 mM $MgCl_2$, 2.5 mM β-mercaptoethanol, 2.0% (v/v) glycerol).

Protein digestion and sample preparation for MS. Purified EF-P was predigested with LysC for 3 h in 8 M urea, 50 mM Tris-HCl, pH 7.5, 1 mM DTT and 5 mM chloroacetamide (CAA). Then, samples were diluted 1:4 with 50 mM Tris HCl, pH 7.5, to decrease the urea concentration to 2 M and digested overnight with trypsin. Peptide mixtures were purified on C18 StageTips[46].

LC/MS/MS analysis. Peptides were eluted from the C18 StageTips according to the standard protocol. They were analyzed by reversed-phase LC on an EASY-nLC 1000 system (Thermo Fisher Scientific) directly coupled to a quadrupole Orbitrap mass spectrometer (Q Exactive, Thermo Fisher Scientific). HPLC columns with a length of 50 cm and an inner diameter of 75 μm were packed in-house with ReproSil-Pur 120 C18-AQ 1.9-μm particles (Dr. Maisch GmbH). Peptide mixtures were separated using gradients of either 60 min or 140 min (total run time plus washout) and a two-buffer system: buffer A++ (0.1% formic acid) and buffer B++ (80% acetonitrile in 0.1% formic acid). The flow rate was set to 250 nl/min, and the column was heated to 50 °C using a column oven (Sonation GmbH). Peptides eluting from the column were directly sprayed into the mass spectrometer; spray voltage was set to 2.3–2.4 kV, and the capillary temperature was set to 250 °C. The mass spectrometer was operated in a data-dependent mode with switching between a survey scan and fragmentation

doi:10.1038/nchembio.1751

scans of the top five most abundant peaks. In the 60-min method, full scans were acquired at a resolution of 140,000 with an AGC target of three E06 ions and a maximum injection time of 20 ms. Precursors were selected with an isolation window of 3 Th, and MS2 scans were acquired at a resolution of 17,500 with an AGC target of one E05 ion and a maximum injection time of 120 ms. In the 140-min method, full scans were acquired at a resolution of 70,000 with an AGC target of three E06 ions and a maximum injection time of 20 ms. Precursors were selected with an isolation window of 2 Th, and MS2 scans were acquired at a resolution of 35,000 with an AGC target of one E05 ion and a maximum injection time of 120 ms. In both cases, peptides were fragmented by higher-energy collisional dissociation (HCD) with a normalized collision energy of 25. To minimize resequencing of peptides, dynamic exclusion was enabled within a time window of 20 s.

MS data analysis. MS raw files were processed using MaxQuant[27] version 1.5.0.0. MS/MS spectra were searched using the Andromeda search engine[47] against FASTA files obtained from Uniprot adapted to the corresponding sample. For EF-P samples from homologous production in *S. oneidensis*, raw data were searched against the *S. oneidensis* reference proteome downloaded from Uniprot on 20 January 2014 and a FASTA file containing the sequence of His_6-tagged EF-$P_{S.o.}$. Depending on the experiment, we added additional FASTA files containing the sequence of His_6-tagged EF-$P_{S.o.}^{R32A}$ or EF-$P_{S.o.}^{R32K}$. Cysteine carbamidomethylation was set as a fixed modification; N-terminal acetylation and methionine oxidation were set as variable modifications. Trypsin was chosen as the specific enzyme, with a maximum of two missed cleavages allowed. Peptide and protein identifications were filtered at a 1% false discovery rate (FDR). The initial mass tolerance was set to 4.5 p.p.m. for the precursor masses and to 20 p.p.m. for the fragment masses. All of the other parameters were left at standard settings. For dependent peptide analysis (essentially as described in ref. 47), the corresponding feature was enabled. For the variable modification searches, L-rhamnose $-H_2O$ ($C_6H_{10}O_4$ = 146.058 Da) was first defined as a variable modification with specificity for arginine in the Andromeda configuration and was subsequently added to the other variable modifications in the MaxQuant search.

Bioinformatic analysis of the MaxQuant processed data was performed using the Perseus software (version 1.4.2.35, available in the MaxQuant environment). In brief, for dependent peptide analysis, the 'all.peptides.txt' table was loaded and filtered for DP decoy ≠ '+', 'DP Protein' = EF-P, 'DP Base sequence' containing 'SGR', 'DP Mass Difference' > 0 and 'DP Score' > 80. Remaining hits were further validated in a manual fashion. Spectra were visualized using the Viewer program (version 1.5.0.0, integrated into MaxQuant) and annotated using the Expert System[48]. For rhamnosylation analysis, the 'Rhamsites.txt' table was loaded, and the site table was expanded, logarithmized and then filtered for 'Protein' = EF-P, 'Localization Probability' = 1 and 'Score' > 80.

Enzymatic total synthesis of TDP-L-rhamnose. The synthesis was carried out in two steps. First, TDP-4-keto-6-deoxy-D-glucose was prepared from glucose-1-phosphate using two purified *E. coli* enzymes (RmlA and RmlB) and TTP. TTP was generated *in situ* from TMP with a mix of TMP kinase and acetate kinase. TDP-4-keto-6-deoxy-D-glucose is a key intermediate of many 6-deoxyhexoses and can be stored at −80 °C. *E. coli* BL21(DE3) (EMD 4 Biosciences) was used for the conversion of TDP-4-keto-6-deoxy-D-glucose to TDP-L-rhamnose. This *E. coli* strain contains naturally TDP-4-ketorhamnose 3,5-epimerase (RmlC) and TDP-4-keto-rhamnose reductase (RmlD). Crude cell lysates of *E. coli* BL21 (DE3) were added to a solution of TDP-4-keto-6-deoxy-D-glucose in phosphate buffer (50 mM, pH 7.5), $MgCl_2$ (4 mM) and NADH (5 mM). The reaction continued for 3 h at 37 °C. The production of TDP-L-rhamnose was monitored by HPLC[49,50].

The resulting solution containing TDP-L-rhamnose was desalted by size-exclusion chromatography (BioGel P2 column) and further purified by HPLC (Waters 600 system consisting of a controller, a Waters 996 photodiode array detector and a Delta 600 pump; a Dionex CarboPac PA1, 4 × 250 mm column was used for 60-min runs at a flow rate of 1.0 ml/min, and UV monitoring absorbance was set at 254 nm). The gradient used was as follows: solvent A, water; solvent B, 0.5 mM ammonium acetate solution. Solvent B was increased from 5% to 20% (0 min to 15 min), from 20% to 60% (15 min to 35 min), and then from 60% to 100% (35 min to 45 min); it was kept at 100% for 5 min before it was decreased back to 5% within 2 min, and it was kept at 5% for the last 8 min[49,51]. Under these conditions, TDP-L-rhamnose elutes at 34.04 min.

***In vitro* glycosylation.** A total of 10 µM of His_6-EF-$P_{S.o.}$ or substitution variants and 10 µM His_6-Ear$P_{S.o.}$ were incubated in reaction buffer (50 mM HEPES, pH 7.8, 100 mM NaCl, 50 mM KCl, 10 mM $MgCl_2$, 2.5 mM β-mercaptoethanol, 2.0% (v/v) glycerol) with 100 µM of dTDP-L-rhamnose for 60 min at 30 °C.

Cytotoxicity assay. The cytotoxicity of the *P. aeruginosa* strains was assayed as described earlier[52]. Briefly, eukaryotic A549-Gluc cells were cultured in completed Dulbecco's modified Eagle's medium at 37 °C with 5% CO_2. A549-Gluc cells were generated from A549 by lentiviral gene transfer as described previously[53,54]. Cytotoxicity of the *P. aeruginosa* strains was assessed by infecting A549-Gluc cells, which secrete Gaussia luciferase, as a measure of cell integrity. A549-Gluc cells were seeded in 96-well plates at a density of 2.5–5 × 10^4 cells per well and grown until ~90% confluence. After washing, cells were inoculated with 6-h-old *P. aeruginosa* LB cultures adjusted to a multiplicity of infection (MOI) of 200 and centrifuged to increase cell-cell contact. 100 µl cell culture supernatants were collected after 3 h of incubation, and Gaussia luciferase activity was measured for 0.1 s using an LB 960 Centro XS3 plate luminometer (Berthold Technologies) after the addition of 60 µl of 10 µM coelenterazine (PJK GmbH). Luciferase activities were determined for at least three independent experiments. The significance of the results was determined by applying the two-sided Student's *t*-test, and results were considered significantly different if $P < 0.05$.

Assay for pyocyanin production. Pyocyanin was extracted from *P. aeruginosa* supernatant and measured according to ref. 55. 5 ml of 24-h-old cultures were extracted with 1 volume of chloroform and then re-extracted into 0.2 N HCl to give a pink solution. The aqueous layer was transferred to a fresh tube, and absorbance was measured at 520 nm. Pyocyanin produced per milliliter of culture supernatant was calculated as described elsewhere[55].

Quantification of rhamnolipid production. Overnight cultures of *P. aeruginosa* were freshly diluted to an OD_{600nm} of 0.05 and incubated in LB at 37 °C for 48 h. The colorimetric analysis of the orcinol reaction was adopted from the method described in ref. 56. Briefly, 300 µl of culture supernatant were extracted twice with diethylether. After evaporation of the pooled fraction, the remaining fraction was dissolved in distilled water and incubated with 100 µl 1.6% (w/v) orcinol and 800 µl 60% sulfuric acid. After heating to 80 °C and shaking at 175 r.p.m. for 30 min, the adsorption at 421 nm was determined. In parallel, L-rhamnose at defined concentrations was assayed as described above and used as a standard for determining the L-rhamnose in the culture samples. Rhamnolipid concentrations were then calculated with the assumption that 1 µg of L-rhamnose corresponds to 2.5 µg of rhamnolipid[57]. Rhamnolipids were quantified from at least three independent experiments. The significance of the results was determined by applying the two-sided Student's *t*-test, and results were considered significantly different if $P < 0.05$.

Molecular modeling. The molecular model for modified EF-Ps on the ribosome was generated using the crystal structure of unmodified *T. thermophilus* EF-P bound to a *T. thermophilus* 70S ribosome programmed with $tRNA^{fMet}$ at the P-site[12]. The proline residue was modeled onto the CCA-end of the P-site tRNA by aligning the structure of the 50S subunit with an aminoacylated tRNA substrate in the P-site (PDB1VQN) and by mutagenesis of the amino acid to proline using Coot[58]. The models for EF-P bearing lysinylation or hypusine modifications were generated by mutagenesis of Arg32 to lysine of ribosome-bound *T. thermophilus* EF-P and then addition of the required modification moieties to the ε-amino group of the lysine, whereas for EF-P bearing the rhamnosylation modification, an L-rhamnose moiety was added to the η^2-amino group of Arg32 of ribosome-bound *T. thermophilus* EF-P. All of the models were generated and refined in Coot[58], and images were produced using the PyMOL Molecular Graphics System (Version 1.5.0.4 Schrödinger, LLC).

33. Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome. Inform.* **23**, 205–211 (2009).
34. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

35. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
36. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
37. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
38. Keilberg, D., Wuichet, K., Drescher, F. & Sogaard-Andersen, L. A response regulator interfaces between the Frz chemosensory system and the MglA/MglB GTPase/GAP module to regulate polarity in *Myxococcus xanthus*. *PLoS Genet.* **8**, e1002951 (2012).
39. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
40. Jacobs, M.A. *et al.* Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. USA* **100**, 14339–14344 (2003).
41. Schmidt, J. *et al.* The *Pseudomonas aeruginosa* chemotaxis methyltransferase CheR1 impacts on bacterial surface sampling. *PLoS ONE* **6**, e18184 (2011).
42. Heermann, R., Zeppenfeld, T. & Jung, K. Simple generation of site-directed point mutations in the *Escherichia coli* chromosome using Red®/ET® Recombination. *Microb. Cell Fact.* **7**, 14 (2008).
43. Lassak, J., Henche, A.L., Binnenkade, L. & Thormann, K.M. ArcS, the cognate sensor kinase in an atypical Arc system of *Shewanella oneidensis* MR-1. *Appl. Environ. Microbiol.* **76**, 3263–3274 (2010).
44. Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K. & Pease, L.R. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* **77**, 51–59 (1989).
45. Bertani, G. Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*. *J. Bacteriol.* **62**, 293–300 (1951).
46. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
47. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
48. Neuhauser, N., Michalski, A., Cox, J. & Mann, M. Expert system for computer-assisted annotation of MS/MS spectra. *Mol. Cell. Proteomics* **11**, 1500–1509 (2012).
49. Kharel, M.K., Lian, H. & Rohr, J. Characterization of the TDP-D-ravidosamine biosynthetic pathway: one-pot enzymatic synthesis of TDP-D-ravidosamine from thymidine-5-phosphate and glucose-1-phosphate. *Org. Biomol. Chem.* **9**, 1799–1808 (2011).
50. Wang, G. *et al.* Cooperation of two bifunctional enzymes in the biosynthesis and attachment of deoxysugars of the antitumor antibiotic mithramycin. *Angew. Chem. Int. Edn. Engl.* **51**, 10638–10642 (2012).
51. Wang, G., Kharel, M.K., Pahari, P. & Rohr, J. Investigating mithramycin deoxysugar biosynthesis: enzymatic total synthesis of TDP-D-olivose. *ChemBioChem* **12**, 2568–2571 (2011).
52. Gödeke, J., Pustelny, C. & Haussler, S. Recycling of peptidyl-tRNAs by peptidyl-tRNA hydrolase counteracts azithromycin-mediated effects on *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **57**, 1617–1624 (2013).
53. Haid, S., Windisch, M.P., Bartenschlager, R. & Pietschmann, T. Mouse-specific residues of claudin-1 limit hepatitis C virus genotype 2a infection in a human hepatocyte cell line. *J. Virol.* **84**, 964–975 (2010).
54. Gentzsch, J. *et al.* Hepatitis C virus complete life cycle screen for identification of small molecules with pro- or antiviral activity. *Antiviral Res.* **89**, 136–148 (2011).
55. Essar, D.W., Eberly, L., Hadero, A. & Crawford, I.P. Identification and characterization of genes for a second anthranilate synthase in *Pseudomonas aeruginosa*: interchangeability of the two anthranilate synthases and evolutionary implications. *J. Bacteriol.* **172**, 884–900 (1990).
56. Wilhelm, S., Gdynia, A., Tielen, P., Rosenau, F. & Jaeger, K.E. The autotransporter esterase EstA of *Pseudomonas aeruginosa* is required for rhamnolipid production, cell motility, and biofilm formation. *J. Bacteriol.* **189**, 6695–6703 (2007).
57. Ochsner, U.A., Koch, A.K., Fiechter, A. & Reiser, J. Isolation and characterization of a regulatory gene affecting rhamnolipid biosurfactant synthesis in *Pseudomonas aeruginosa*. *J. Bacteriol.* **176**, 2044–2054 (1994).
58. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

doi:10.1038/nchembio.1751

2.2.2 Investigating protein glycation in human blood plasma using higher-energy collisional dissociation mass spectrometry

Keilhauer, E. C., Geyer, P. E. & Mann, M.

HCD fragmentation of glycated peptides
Submitted to the *Journal of Proteome Research*

Recent technological improvements have advanced mass spectrometry to a stage where it can be applied to clinical questions. In this fifth and last project, I set out to investigate a PTM highly relevant in diabetes, called protein glycation. Diabetes, a chronic disease characterized by abnormal glucose metabolism, is amongst the top ten causes of death worldwide. Hence new methods for better diagnosis, monitoring and treatment, as well as a deeper understanding of the disease in order to potentially prevent it, are urgently needed.
Protein glycation is formed by the non-enzymatic reaction between glucose (or other reducing sugars) with the amino groups of proteins, a reaction commonly known as Maillard reaction. Since the glycation reaction is concentration-dependent, the increased level of blood glucose in diabetic patients results in an increase of this particular PTM. Although protein glycation is well understood for hemoglobin, and measuring glycated hemoglobin (*HbA1c*) is actually one of the standard procedures to diagnose and monitor diabetes, our knowledge about other glycated proteins is quite limited. Every protein in contact with glucose is a potential target for this unspecific modification, hence increased knowledge about other glycation targets is highly desirable.
In this project I set out to develop an MS-based method for investigating glycated peptides on the benchtop Orbitrap platforms used in our laboratory, that employ HCD fragmentation. I first evaluated the fragmentation behavior of the glycated peptides in HCD using model proteins, and adapted the data acquisition and analysis accordingly. I then performed proof-of-principle experiments for detecting glycated proteins in complex matrices like HeLa lysate and finally human blood plasma. In the future, I plan to further optimize this workflow, particularly by implementing a quantification strategy, and subsequently apply it to investigate protein glycation directly in patient samples.

HCD fragmentation of glycated peptides

Eva C. Keilhauer, Philipp E. Geyer and Matthias Mann*

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

*Corresponding author

ABSTRACT

Protein glycation is a concentration-dependent non-enzymatic reaction of reducing sugars with amine groups of proteins to form early as well as advanced glycation products (AGEs). Glycation is a highly disease-relevant modification, but is typically only studied on few blood proteins. To complement our blood proteomics studies in diabetics, we here investigate protein glycation by higher-energy collisional dissociation (HCD) fragmentation on Orbitrap mass spectrometers. We established parameters to most efficiently fragment and identify early glycation products on *in vitro* glycated model proteins. Retaining standard collision energies does not degrade performance if the most dominant neutral loss of H_6O_3 is included into the database search strategy. Glycation analysis of the entire HeLa proteome revealed an unexpected intracellular preponderance for arginine over lysine modification in early and advanced glycation products. Single-run analysis from 1 µl of undepleted and unenriched blood plasma identified 101 early glycation sites as well as numerous AGE sites on diverse plasma proteins. We conclude that HCD fragmentation is well suited for analyzing glycated peptides, and that the diabetic status of patients can be directly diagnosed from single-run plasma proteomics measurements.

Keywords: protein glycation, higher-energy collisional dissociation, diabetes, blood plasma, AGEs

INTRODUCTION

Protein glycation, in contrast to enzyme-mediated glycosylation, it is produced by the non-enzymatic reaction of glucose molecules or other reducing sugars with amine groups of proteins and is also known as Maillard reaction.[1] Glucose first attaches to form a Schiff base, which then rearranges into the relatively stable Amadori compound[2], to which we refer here as 'early glycation product'. Glycated proteins can further react to form advanced glycation endproducts (AGEs), or proteins can directly react with glucose-derived reactive dicarbonyls like methylglyoxal to form AGEs.[3] Glucose is an essential and omnipresent energy source in humans, and is tightly regulated in a narrow concentration band in healthy individuals. Dysregulation of glucose levels is the principal feature of diabetes, a growing health epidemic, currently affecting an estimated 415 million individuals worldwide according to the International Diabetes Federation (IDF) Diabetes Atlas (7th edition). The extent of protein glycation and AGEs are increased in proportion to the glucose concentration, and the glycation level of one particular blood protein, hemoglobin, is routinely assessed in the diagnosis of diabetes as well as for long-term monitoring of blood glucose levels of diabetes patients. More specifically, glycation of the N-terminal valine of the hemoglobin beta-chain is assessed, a clinical parameter known as HbA1c.[4, 5] Since the lifespan of erythrocytes and hence hemoglobin is around 120 days, the HbA1c value reflects the average blood glucose concentration of the last six to eight weeks[2, 6]. Hence the HbA1c-test is often more robust than oral glucose tolerance tests that can be influenced by various factors such as recent food intake, exercise and blood sampling time. If the HbA1c value can be stabilized close to normal levels, patients have a much better prognosis and less diabetic complications than those with poorly controlled HbA1c values.[7] Glycation and AGEs are central to the development of typical diabetic complications, and also play a role in ageing and neurodegenerative and cardiovascular diseases.[8-13]

The current and strong focus on glycated hemoglobin and few more proteins is presumably due to a lack of appropriate methods to robustly detect, characterize and quantify other glycated proteins. Owing to its extreme complexity and extraordinary dynamic range, blood plasma is the most challenging proteome.[14-16] However,

investigation of other glycated proteins could help to better diagnose, monitor and understand metabolic conditions such as diabetes. For example, measuring several glycated proteins with different lifespans might yield a more detailed picture of blood glucose levels of patients over the last days to weeks.[17-19]

Mass spectrometry (MS) is the method of choice to investigate post-translational protein modifications (PTMs) in an unbiased manner.[20] Analysis of glycation in body fluids has been challenging because of its low-stoichiometry and enrichment strategies such as boronate affinity chromatography (BAC) are typically employed.[21, 22] Sample complexity is often additionally reduced by depleting the most abundant plasma proteins and/or fractionating the plasma on the peptide level. In this way, and by pooling and fractionating a large number of diverse samples, the most comprehensive study to date found evidence of around 1100 glycated proteins from human plasma.[23] Such elaborate protocols are useful for generating glycation site resources, however they are not practical for clinical tests. We have recently reported a method called 'plasma proteome profiling', that allows measuring hundreds of plasma proteins from only 1 µl of plasma in a single-run format without depletion or fractionation.[24] We therefore wondered if we could complement the patient information gained from plasma proteome profiling with the diabetic status by determining glycation of plasma proteins.

Glycated peptides have been studied by MS/MS using various fragmentation techniques.[6] Collision-induced dissociation (CID)[25] in ion traps suffers from dominant neutral losses of the labile Amadori compound, often leading to insufficient fragmentation of the peptide backbone for identification of the peptide sequence and the glycation site.[26, 27] Neutral loss-triggered MS^3 scans partly alleviate this problem, but at the cost of lower throughput and sensitivity.[28] Electron-transfer dissociation (ETD)[29], a technology generally known to be well suited for investigating labile modifications, is very effective for glycated peptides. Using ETD, no neutral losses and almost complete series of c-and z-ions were observed.[30] However, ETD is only implemented on specialized mass spectrometers and not on the benchtop Orbitrap instruments that are routinely used in many laboratories. Initial promising results have also been obtained for higher-energy collisional dissociation (HCD)[31] fragmentation, however, so far always in combination with other techniques.[28]

As the benchtop Orbitrap instruments (Q Exactive) exclusively feature HCD fragmentation, we therefore set out to systematically evaluate how well glycated peptides can be fragmented and analyzed with HCD-MS2 scans alone.

EXPERIMENTAL SECTION

IN VITRO GLYCATION OF BSA AND HSA

Both bovine serum albumin (BSA) and human serum albumin (HSA) (human fraction 5 powder) were purchased from Sigma Aldrich. BSA (100 mg/ml) was incubated with 1 M glucose in 50 mM Tris HCl buffer pH 7.5 at room temperature for the indicated times. HSA (10 mg/ml) was incubated with 1 M glucose in the same buffer for 48 h. Both BSA and HSA were digested with trypsin (Promega) with an enzyme to protein ratio of around 1:20 to 1:50 in digestion buffer (2 M urea and 1 mM dithiothreitol (DTT) in 50 mM TrisHCl pH 7.5). After 20 min, 5 mM chloroacetamide (CAA) was added to the samples, then they were incubated overnight to ensure a complete digest. On the next day, the digest was stopped by addition of 1 µl trifluoroacetic acid (TFA) per sample. The peptides were desalted and purified on StageTips (self-made pipette tips containing two layers of C_{18} material) according to the standard protocol.[32] The StageTips were stored at 4 °C until the sample was measured. BSA and HSA samples were eluted from the C_{18} StageTips with 2×20 µl buffer B (80 % acetonitrile (ACN), 0.5 % acetic acid). The organic solvent was removed in a SpeedVac concentrator for 20 min, then the peptide mixture was acidified with buffer A* (2 % ACN, 0.1 % TFA) to a final sample size of 5 µl.

PREPARATION OF HELA DIGESTS

HeLa cells were cultured in high glucose DMEM with 10 % fetal bovine serum and 1 % penicillin-streptomycin (all from Life Technologies). Around $5x10^7$ cells were harvested and lysed in 6 M urea/2 M thiourea. Proteins were reduced with 1 mM DTT for 30 min at room temperature, then alkylated with 5mM iodoacetamide (IAA) for 20 min in the dark. Proteins were digested overnight with LysC and trypsin. The digest was stopped by adding TFA, then peptides were purified on StageTips as described above.

PREPARATION OF WHOLE BLOOD AND BLOOD PLASMA SAMPLES: PROTEIN DIGESTION AND IN-STAGETIP PURIFICATION.

Sample preparation for plasma was done as described previously.[33] Briefly, 1 µl of plasma was mixed with 24 µl of SDC reduction and alkylation buffer[34]. After protein denaturation by boiling for 10 min, LysC and trypsin were added in a 1:100 ratio (µg enzyme to µg protein) and digestion was performed for 1 h at 37 °C. Peptides were acidified by adding 125 µl ethylacetate/1 % TFA and 20 µg were transferred to StageTips, containing two 14-gauge SDB-RPS plugs. Washing steps included two times 100 µl ethylacetate/1 % TFA and one time 100 µl ddH2O/0.2 % TFA. The purified peptides were eluted with 60 µl of elution buffer (80 % acetonitrile, 19 % ddH2O, 1 % ammonia) into auto sampler vials. The collected material was dried to completion using a SpeedVac centrifuge at 45 °C (Eppendorf, Concentrator plus). Peptides were suspended in 2 % acetonitrile, 0.1 % TFA and sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510) prior to analysis.

The sample preparation procedure for whole blood included an additional sonication step of 15 min by a Diagenode Bioruptor prior to digestion.

LC-MS/MS MEASUREMENT OF HSA AND BSA

Samples were analyzed by nanoflow liquid chromatography (LC-MS/MS) on an EASY-nLC HPLC system (Thermo Fisher Scientific) that was on-line coupled to either an Q Exactive plus or a Q Exactive HF mass spectrometer (all Thermo Fisher Scientific) through a nanoelectrospray ion source (Thermo Fisher Scientific). A 50 cm column with a 75 µm inner diameter in-house packed with 1.9 µm reversed-phase silica beads (ReproSil-Pur C_{18}-AQ, Dr. Maisch GmbH) was used for the chromatography. Peptides were separated using a linear gradient from 5.6 % to 25.6 % ACN in 0.1 % formic acid at a constant flow of 250 nl/min, then directly electrosprayed into the mass spectrometer. Overall gradient length was one hour. The column oven (Sonation GmbH) was heated to 55 °C. The spray voltage was set to 2.4 kV and the heated capillary temperature to 250 °C.

BSA/HSA samples were measured using a data-dependent top10 method, the BSA glycation time course was measured in a top1 method. Instruments were controlled by Tune Plus 2.0 and Xcalibur 2.0. On the Q Exactive plus, full scans (*m*/*z* 300–1,650) were acquired with a resolution of 70,000 at 200 *m*/*z* and an AGC target of 3E06 ions and fragmentation scans with a resolution of 17,500 at 200 *m*/*z* and an AGC target of 1E05 ions. Maximum ion accumulation times were 20 ms for the full scans and 120 ms for the fragmentation scans. On the Q Exactive HF, full scans (*m*/*z* 300–1,650) were acquired with a resolution of 60,000 at 200 *m*/*z* and an AGC target of 3E06 ions and fragmentation scans with a resolution of 16,000 at 200 *m*/*z* and an AGC target of 1E05 ions. Maximum ion accumulation times were 120 ms for both full scans and fragmentation scans. The most intense ions from the full scans were isolated with an isolation width of 1.4 m/z and

fragmented using HCD, with a normalized collision energy (NCE) of 25 % (Q Exactive plus) or 27 % (Q Exactive HF) unless specified otherwise in the text. Dynamic exclusion was enabled for a duration of 20 s.

LC-MS/MS MEASUREMENT OF HELA DIGESTS

Samples were measured on a Q Exactive HF essentially as described for BSA and HSA with the following alterations: Gradient length was 2 h, HeLa samples were measured in top15 mode and the maximum ion accumulation time for fragmentation scans was 25 ms.

LC-MS/MS MEASUREMENT OF BLOOD PLASMA/ WHOLE BLOOD

Samples were measured on a Q Exactive HF essentially as described for BSA and HSA with the following alterations: Column length was 40 cm and the column oven temperature was set to 60 °C. Gradient length was 100 min and samples were measured using a data-dependent top15 method. Full scans (*m*/*z* 300–1,650) were acquired with a resolution of 120,000 at 200 *m*/*z,* an AGC target of 3E06 ions and a maximum injection time of 55 ms. An isolation window of 1.5 m/z and a fixed first mass of 100 m/z was used for MS/MS scans. HCD fragmentation was performed with an NCE of 27. MS/MS scans were acquired with a resolution of 30,000 at 200 *m/z* with an AGC target of 1E05 ions and a maximum injection time of 55 ms. Dynamic exclusion was enabled for a duration of 30 s.

DATA ANALYSIS

All raw data was analyzed using the MaxQuant[35] software environment (version 1.5.3.0). The software searched the derived peak list using the built-in Andromeda search engine[36] against either a bovine reference proteome downloaded from Uniprot (http://www.uniprot.org/) on February 2016 (24481 sequences) or against a human reference proteome downloaded from Uniprot in May 2013 (88847 sequences). In all cases, a file containing 247 frequently observed contaminants such as human keratins and proteases was included in the search. Trypsin was chosen as the protease with strict specificity for cleavage C-terminal to K or R required. Up to two missed cleavages per peptide were allowed. The minimum peptide length was set to 7 amino acids. Due to the sample preparation, carbamidomethylation of cysteine was set as a fixed modification (57.021464 Da). N-acetylation of protein N-termini (42.010565 Da) and oxidation of methionine (15.994915 Da) were set as variable modifications. For glycation/AGE analysis, the corresponding modification with/without different neutral losses was defined in Andromeda configuration and added to the variable modifications as stated in the text.(Glycation: 162.052823 Da, CML: 58.005479 Da, CEL: 72.021129 Da, MG-H1: 54.010565 Da, Argpyr: 80.026215 Da, 3DG-H1: 144.042259 Da) All other parameters were left at standard settings. Peptide and protein identifications were filtered at a false discovery rate (FDR) of 1 %. The ‘match between runs’ option was used where specified in the text with a match time window of 0.7 min and an alignment time window of 20 min.

Further analysis of the MaxQuant output tables was performed using the Perseus software (version 1.5.3.0), which is part of the MaxQuant environment. Plots were produced in R (version 2.15.3).

DATA AVAILABILITY

Raw data and MaxQuant output files are accessible via ProteomeXchange[37] with identifier PXD004182.

RESULTS AND DISCUSSION

HCD FRAGMENTATION OF GLYCATED PEPTIDES

Orbitrap mass spectrometers have proven to be powerful instruments for proteomics in general and clinical proteomics in particular, and today are standard in many laboratories. The widespread benchtop quadrupole Orbitrap instruments (Q Exactive family) only feature HCD as fragmentation method. As previous work on glycated peptides had employed ETD or a combination of other fragmentation methods with HCD, we here set out to investigate whether glycated peptides can be identified solely on the basis of HCD-MS/MS scans. As glycation is typically studied in blood plasma, we chose human serum albumin (HSA) as a model protein. We glycated HSA *in vitro*, digested it with trypsin and measured the resulting peptides on a Q Exactive HF without optimizing the instrument in any way. In the MaxQuant data analysis software[35], we included protein glycation ($C_6H_{10}O_5$; 162.0528 Da) as a variable modification on lysine, which is the major target for glycation by glucose, and on arginine. The 'matching between runs' algorithm was enabled between the three technical replicates, which transfers peptide identifications to LC MS/MS runs where the same peptide was present but was not sequenced. Surprisingly, in view of the complex experimental set-up previously employed in the analysis of glycation, already this first experiment identified 45 unique glycation sites on HSA. Most sites (42) were located on lysine, consistent with the fact that this residue is the primary target for this type of glycation, and only three sites were found on arginine. Thus the large majority of the 59 lysines in mature HSA can be glycated *in vitro* by

incubation with high glucose concentrations. Interestingly, UniProt lists only 20 of the 42 lysine sites as glycated *in vitro* or *in vivo*, while 22 of them were incorrectly annotated as 'not glycated' in UniProt (See **Table 1 A**). The five strongest glycation sites as indicated by the MS intensity of their corresponding glycated peptides were K257, K549, K438, K36 and K223 (**Table 1B**). Consistent with our *in vitro* results, K36, K223, K257 and K549 have also been reported to be HSA glycation sites *in vivo.*[38] Interestingly, K438 has not been reported to be glycated before (only K437), and we found no evidence of two other reportedly strong *in vivo* glycation sites, K305 and K463. The fact that we identified such a high number of sites on this widely used model protein suggests that standard HCD-MS/MS scans are remarkably well suited for the characterization of glycated peptides.

OPTIMIZING THE COLLISION ENERGY FOR GLYCATED PEPTIDES

In addition to backbone fragmentation, glycated peptides can also fragment by losing all or part of the Amadori product during CID and HCD fragmentation.[27, 39] Therefore collision energies for HCD might be different for the identification of glycated peptides compared to unmodified peptides, which was suggested by the relatively low identification scores of the glycated HSA peptides described above. Using *in vitro* glycated BSA as a model protein, we performed LC-MS/MS runs with six different normalized collision energies (NCEs) centered around the standard NCE that we use in our shotgun proteomics experiments. Plotting the number of unmodified BSA peptides identified at each collision energy confirmed that an NCE of 25 % on the instrument employed (Q Exactive Plus) was indeed optimal for these peptides (**Figure 1A**). The same analysis revealed a broad optimum NCE for the number of identified glycated peptides, centered between 20 (43 sites) and 25 % (42 sites) (**Figure 1B**). An NCE of 40 %, in contrast, dramatically reduced identification success. Next we investigated for each identified glycation site in which of the measurements at the different NCEs it was best localized to a particular amino acid (localization probability) and where it obtained the maximum database identification score. By these measures, an NCE of 20 % appeared to be optimal for both localization and identification (**Figure 1C, D**).

When we examined the fragmentation spectra of the glycated peptides more closely, we found that at higher NCEs, there were typically no fragments carrying the full modification of 162.053 Da. Furthermore, *b*-ions were mostly absent from the spectra, and often a number of intense peaks in the higher mass range were unexplained by standard backbone fragmentation (for an example, see spectrum in **Figure 2A**). The Amadori compound can lose several water molecules and formaldehyde during CID and HCD fragmentation[27, 39], resulting in residual modification masses of 144.0423 Da, 126.0317 Da, 108.0211 Da, 96.0211 Da and 78.0106 Da (**Figure 2B**). Additionally, we also observed loss of the entire glucose moiety from the fragments and the intact peptide. After annotating the spectrum in **Figure 2A** with these reduced forms of glycation using the expert system for fragment annotation[40], we were able to explain basically all the peaks in the spectrum (**Figure 2C**). Essentially the complete series of backbone fragments were represented in at least one of the possible modification states, with the exception of cleavage between the N-terminal phenylalanine and the glycated lysine. Generally, while the loss of only one water molecule leading to the 144.0423 Da modification seemed to occur rarely, other pathways appeared to be more dominant: the loss of three water molecules leading to the 108.0211 Da modification, and the loss of three water and one formaldehyde molecules leading to the 78.0106 Da modification.

We reasoned that taking the neutral losses into account, we might be able to use our standard collision energy of 25 % (or 27 % on the Q Exactive HF) to both obtain efficient backbone fragmentation, as well as confidently identify glycation sites. As the MaxQuant software only supports one neutral loss per modification to avoid combinatorial explosion, next determined the most common neutral loss in a systematic way. We defined seven different versions of glycation for the search engine: without any neutral loss, with a neutral loss of H_2O, H_4O_2, H_6O_3, CH_6O_3, CH_8O_4, and finally $C_6H_{10}O_5$ corresponding to the entire modification. Interrogating the data file obtained at the NCE of 25 % with the seven different versions of glycation on lysine, we found that a neutral loss of three water molecules (H_6O_3) leading to a residual mass of 108.0211 Da yielded most glycation sites in total (47 sites, see **Figure 3A**). This search mode also produced most high confidence sites, for example 44 sites with an identification score of over 75. CH_8O_4, with a residual modification mass of 78.0106 Da was the next most common neutral loss, followed by

loss of the entire glucose moiety. With these optimized collision and search settings, we now found an additional 17 glycated lysines on BSA, compared to the search without neutral loss (**Figure 3A**). **Figure 3B** illustrates that the neutral loss of three water molecules explains the majority of peaks in the MS/MS spectra. Having established the dominant neutral loss in HCD fragmentation at the standard (and optimal) collision energy of 25 % to be the loss of three water molecules, we subsequently routinely included this neutral loss in the search for glycated peptides.

Applying the three water neutral loss analysis to our previous analysis of *in vitro*-glycated HSA increased the number of unique glycation sites increased from 45 to 54. Among those are 50 lysine residues, meaning that a remarkable 85 % of all lysine residues in the mature HSA sequence can be glycated *in vitro*. This can be explained by the fact that lysine as a charged amino acid is typically surface exposed. The number of glycated arginines went up from three to four, and the additional site at R184 has been reported before (see **Supplementary Table 1)**. Regarding the previously reported *in vivo* glycation sites, we now additionally identify K305, however we still find no evidence for glycated K463.

Assessing the effect of including the dominant neutral loss on the collision energy evaluation, we found that an NCE of 25 % now resulted in the most BSA glycation sites and the total number of sites increased from 43 to 60 (**Supp. Figure 1A**). The best localization was now obtained with an NCE of 30 %, while the highest score was clearly obtained with an NCE of 25 % (**Supp. Figure 1B, C**). Thus the overall optimal collision energy should be between 25 and 30 %. Considering that 25 % is the optimal setting for unmodified peptides and hence peptide backbone fragmentation, and that localization of the glycation site is generally not problematic, we recommend an NCE of 25 % as also optimal for fragmenting glycated peptides, provided the neutral loss of H_6O_3 is taken into account. (Optimal NCEs depend slightly on the specific model and we find an NCE of 27 % to be optimal for glycated and non-glycated peptides on the Q Exactive HF ref[41]).

TIME- DEPENDENCY OF PROTEIN GLYCATION

To investigate the increase of protein glycation over time *in vitro*, we incubated BSA with 1M glucose for 0-30 days, since after 30 days the equilibrium of the reaction forming the Amadori product should have been reached.[42] Samples were analyzed in triplicates for glycation on K and R allowing for a neutral loss of H_6O_3 and without matching between runs. Interestingly, our results revealed some glycation events already on the purchased BSA before *in vitro* incubation with glucose. These are presumably *in vivo* glycations that have remained stably associated with the protein after purification from bovine blood, processing and storage. We identified 11 such sites in all three replicates: K28, K36, K88, K256, K263, K266, K299, K401, K498, K548 and K561. (Note that if comparing BSA to HSA sites there is a plus one difference in amino acid position starting from position 140.) Two of these (K36 and K256) correspond to known HSA *in vivo* sites. With longer incubation the number of detected glycation sites increased substantially (**Figure 4A**). The figure shows a near doubling of detected sites already after one day. This means we are initially detecting the Schiff base adduct, since several days are needed to convert the Schiff base to the more stable Amadori product.[43] On day 30, almost all sites were still found on lysine. There was also a substantial number of doubly glycated peptides, consistent with the fact that glycation inhibits tryptic cleavage (**Figure 4B**), and a clear quantitative increase in glycation over time (**Figure 4C**).

ANALYZING PROTEIN GLYCATION IN CELL LYSATE AND BLOOD PLASMA

To evaluate the feasibility of detecting glycated peptides in a complex matrix without applying any enrichment step, we chose HeLa lysate as a first test matrix. Because glucose concentrations in standard cell culture conditions are already around five times higher than the physiological concentrations in the body (4.5 mg/mL glucose vs. 0.75-1.15 mg/ml in normal human blood[44]), we chose to not further expose the cells to glucose. HeLa lysates were trypsin digested in four workflow replicates, measured in single-shot 2 h measurements on a Q Exactive HF and analyzed for glycation as described before

with matching between runs. Even in the absence of any enrichment, we identified 155 glycation sites on 94 different proteins, with a mean localization probability of 0.95. Surprisingly, and in stark contrast to our model plasma proteins, the most frequently modified amino acid was arginine (83 sites) and not lysine (72 sites) (**Figure 5A**). This indicates that in an intracellular system, arginine and lysine are about equally reactive as targets for glycation by glucose.

We next investigated possible formation of AGEs in the HeLa proteome. Intracellularly, AGEs may not form by reaction with glucose and via the Amadori product, but instead by direct reactions with glucose metabolites.[45] Therefore, we additionally included some major *in vivo* AGEs derived from glyoxal, methylglyoxal or 3-deoxyglucosone into the analysis: carboxymethyllysine (CML), carboxyethyllysine (CEL), methylglyoxal-derived hydroimidazolone (MG-H, on arginine), argpyrimidine (on arginine) and 3-deoxyglucosone-derived hydroimidazolone (3DG-H, on arginine). We indeed found many sites for all of those AGEs, and interestingly detected about 5 times more arginine AGEs than lysine AGEs (see **Figure 5B**). This is consistent with what we found for early glycation and with the fact that methylglyoxal is more reactive towards arginine than lysine.[46] Unexpectedly, argpyrimidine was the most common AGE, even though its half-life under physiological conditions has been reported to be shorter than that of MG-H1 (2-9 days vs. 2-6 weeks).[47] All Hela glycation and AGE sites are listed in **Supplementary Table 2**.

We next went on to test our method on human blood plasma. Exploiting the high scan speed of the Q Exactive HF, we set out to detect glycation sites directly from less than a single drop of human plasma, without depletion of high abundance proteins, peptide fractionation or enrichment of glycated peptides. We performed the plasma analysis in three technical replicates and analyzed the purified peptides in 100 minute gradients using a Top15 method. This yielded 101 glycation sites located on 53 proteins. Similar numbers were obtained in a 2008 study using immunodepletion and boronate affinity enrichment, however, with 5000 times the input material and substantially longer sample processing times.[48] The protein carrying the most glycation sites was albumin with 16 sites, 11 of which were identified with very high localization scores (>0.99): K36, K44,

K161, K214, K223, K249, K257, K375, K402, K549 and K598. Although identified in a direct and relatively straightforward analysis in normal human blood, three of these sites have not been reported to be glycated before according to UniProt (see **Table 1**). Many other typical plasma proteins were found to be glycated, among them apolipoprotein A1 (8 sites), alpha-1-antitrypsin (4 sites), serotransferrin (3 sites), fibrinogen alpha and beta chain (1 site each), and interestingly, many antibody chains. Overall, the plasma glycation sites had a mean localization probability of 0.95, and a mean absolute mass error of only 0.12 ppm (**Supplementary Table 3)**. The vast majority of glycations in plasma was found to be located on lysine (90 vs. 11 sites; **Figure 5C).** This was similar to what we observed on the model proteins before, but very different from the glycated HeLa proteins (see **Figure 5A**). Furthermore, while in the cell lysate, the majority of the peptides was glycated twice, in plasma the majority of the peptides carried only one glycation. We also searched the plasma samples for the five AGEs mentioned above, and found at least 20 sites for each of them, with CML and 3DG-H1 being the most abundant AGEs at 34 sites each (**see Supplementary Table 4**). In contrast to HeLa cells, lysine and arginine AGEs were similarly abundant in plasma (**Figure 5D**).

In a final experiment, we measured whole human blood with all cellular components. Thus it includes the hemoglobin beta-chain (HBB) and its glycation site on the N-terminal valine, which is clinically used to determine the HbA1c value from which diabetes can be diagnosed. We digested and measured whole blood as described before for plasma and analyzed the resulting samples for glycation on valine as well as on lysine and arginine (always including the neutral loss of H_6O_3). We indeed clearly identified the modified valine in position two of HBB (N-terminal position when considering the loss of the initiating methionine), on the easily detectable peptide V*HLTPEEK. Additionally, we found four of the five known lysine glycation sites on HBB, as well as two additional sites that have not been reported before. We also detected all four known lysine glycation sites on the hemoglobin alpha chain (HBA) plus two additional ones (See **Supplementary Table 5** for all hemoglobin sites). If ordered by site intensity, K133 was the strongest site on HBB and K41 on HBA.

CONCLUSIONS AND OUTLOOK

Blood plasma is one of the most challenging proteomes, spanning more than ten orders of magnitude in abundance from the highest to the lowest known plasma protein. Furthermore, PTMs on plasma proteins add another layer of complexity to the inherently intricate plasma proteome. Previous investigations of glycated plasma proteins had relied on extensive sample fractionation, enrichment of glycated peptides and different peptide fragmentation methods.

In the context of our interest in diabetes, we here asked if modern benchtop Orbitrap platforms are capable of the analysis of glycated peptides in plasma. This would be particularly attractive if it could be incorporated into a routine and robust workflow for plasma proteomics.[33]

We evaluated the fragmentation behavior of glycated peptides, and found that HCD-MS/MS scans with the standard collision energy also used by us in proteome measurements are very well suited for identifying and localizing glycation sites. This requires that the prevailing neutral loss of H_6O_3 is taken into account. In this way, we developed a straightforward workflow to detect glycated peptides directly from blood plasma without applying time-consuming depletion, fractionation or enrichment steps. We additionally screened for several well-known AGEs, and found that they can also be efficiently detected from plasma. Our study demonstrates that straightforward plasma proteome analysis can identify early and advanced protein glycation in this challenging body fluid, as part of the routine plasma proteome profiling workflow. Together, this successfully established HCD fragmentation for the investigation of protein glycation in general and early glycation in particular.

It may be interesting to determine the reasons for the marked differences in the glycation behavior of intracellular proteomes and the plasma proteome – in particular the overwhelming preference for lysine over arginine glycation in plasma in contrast to equal occurrence in the cellular proteome.

In the future, we plan to implement a quantification strategy for glycated peptides from patient material, since this would allow to directly assess the level of blood sugar control in any individual in a proteomic study. Clearly, this would be very challenging with label free methods, because of the required accuracy: normal HbA1c values of below 5.7 % need to be robustly distinguished from the pre-diabetic range (5.7-6.4 %) and diabetic values of >6.5 % (Values according to the World health organization report on the use of HbA1c in the diagnosis of diabetes, 2011). We envision the use of isotopic labels that can be introduced into patient material via chemical labeling strategies, such as iTRAQ or TMT. However, ratio compression, which can occur with these techniques, would not be clinically acceptable and additional challenges connected to the fact that trypsin or LysC do not cleave at glycated lysine residues will have to be overcome.

AUTHOR INFORMATION

CORRESPONDING AUTHOR

*Phone: +49 (0)89 8578 2557. Fax: +49 89 8578 2219. Email: mmann@biochem.mpg.de

NOTES

The authors declare no competing financial interests.

ACKNOWLEDGEMENTS

We thank Gaby Sowa, Igor Paron and Korbinian Mayr for technical assistance. We thank Jürgen Cox and Richard Scheltema for advice regarding data analysis. This work was supported by the Max-Planck Society for the Advancement of Science.

ABBREVIATIONS

3DG-H –	3-deoxyglucosone-derived hydroimidazolone
ACN –	Acetonitrile
AGE –	Advanced glycation end-product
BAC –	Boronate affinity chromatography
BSA –	Bovine serum albumin
CAA –	Chloroacetamide
CEL –	Carboxyethyllysine
CID –	Collision-induced dissociation
CML –	Carboxymethyllysine
DTT –	Dithiothreitol
ETD –	Electron-transfer dissociation
FDR –	False discovery rate
HBA	Hemoglobin alpha chain
HbA1c –	Clinical parameter; Glycation on the N-terminal Valine of the hemoglobin beta-chain
HBB–	Hemoglobin beta chain
HCD –	Higher-energy collisional dissociation
HPLC –	High-pressure liquid chromatography
HSA –	Human serum albumin
IAA –	Iodoacetamide
IDF –	International Diabetes Federation
LC-MS/MS –	Liquid chromatography tandem mass spectrometry
MG-H –	Methylglyoxal-derived hydroimidazolone
MS –	Mass Spectrometry
NCE –	Normalized collision energy
PTM –	Post translational modification
SDB-RPS –	Poly(styrenedivinylbenzene) Reversed-Phase Sulfonate
SDC –	Sodiumdeoxycholate
TFA –	Trifluoroacetic acid

REFERENCES

1. Maillard, L. C., Action of amino acids on sugars. Formation of melanoids in a methodical way. *Compt Rend* **1912,** 154, 66-68.
2. Bunn, H. F.; Gabbay, K. H.; Gallop, P. M., The glycosylation of hemoglobin: relevance to diabetes mellitus. *Science* **1978,** 200, (4337), 21-7.
3. Thornalley, P. J.; Battah, S.; Ahmed, N.; Karachalias, N.; Agalou, S.; Babaei-Jadidi, R.; Dawnay, A., Quantitative screening of advanced glycation endproducts in cellular and extracellular proteins by tandem mass spectrometry. *Biochem J* **2003,** 375, (Pt 3), 581-92.
4. Bunn, H. F.; Haney, D. N.; Kamin, S.; Gabbay, K. H.; Gallop, P. M., The biosynthesis of human hemoglobin A1c. Slow glycosylation of hemoglobin in vivo. *J Clin Invest* **1976,** 57, (6), 1652-9.
5. Koenig, R. J.; Peterson, C. M.; Jones, R. L.; Saudek, C.; Lehrman, M.; Cerami, A., Correlation of glucose regulation and hemoglobin AIc in diabetes mellitus. *N Engl J Med* **1976,** 295, (8), 417-20.
6. Zhang, Q.; Ames, J. M.; Smith, R. D.; Baynes, J. W.; Metz, T. O., A perspective on the Maillard reaction and the analysis of protein glycation by mass spectrometry: probing the pathogenesis of chronic disease. *J Proteome Res* **2009,** 8, (2), 754-69.
7. Wang, J.; Yan, G.; Qiao, Y.; Wang, D.; Ma, G.; Tang, C., Different levels of glycosylated hemoglobin influence severity and long-term prognosis of coronary heart disease patients with stent implantation. *Exp Ther Med* **2015,** 9, (2), 361-366.
8. Ulrich, P.; Cerami, A., Protein glycation, diabetes, and aging. *Recent Prog Horm Res* **2001,** 56, 1-21.
9. Pamplona, R.; Naudi, A.; Gavin, R.; Pastrana, M. A.; Sajnani, G.; Ilieva, E. V.; Del Rio, J. A.; Portero-Otin, M.; Ferrer, I.; Requena, J. R., Increased oxidation, glycoxidation, and lipoxidation of brain proteins in prion disease. *Free Radic Biol Med* **2008,** 45, (8), 1159-66.
10. Stirban, A.; Gawlowski, T.; Roden, M., Vascular effects of advanced glycation endproducts: Clinical effects and molecular mechanisms. *Mol Metab* **2014,** 3, (2), 94-108.
11. Thorpe, S. R.; Baynes, J. W., Role of the Maillard reaction in diabetes mellitus and diseases of aging. *Drugs Aging* **1996,** 9, (2), 69-77.
12. Ahmed, N.; Thornalley, P. J., Advanced glycation endproducts: what is their relevance to diabetic complications? *Diabetes Obes Metab* **2007,** 9, (3), 233-45.
13. Goh, S. Y.; Cooper, M. E., Clinical review: The role of advanced glycation end products in progression and complications of diabetes. *J Clin Endocrinol Metab* **2008,** 93, (4), 1143-52.
14. Anderson, N. L.; Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **2002,** 1, (11), 845-67.
15. Baker, E. S.; Liu, T.; Petyuk, V. A.; Burnum-Johnson, K. E.; Ibrahim, Y. M.; Anderson, G. A.; Smith, R. D., Mass spectrometry for translational proteomics: progress and clinical implications. *Genome Med* **2012,** 4, (8), 63.
16. Omenn, G. S., Exploring the human plasma proteome. *Proteomics* **2005,** 5, (13), 3223, 3225.

17. Misciagna, G.; De Michele, G.; Trevisan, M., Non enzymatic glycated proteins in the blood and cardiovascular disease. *Curr Pharm Des* **2007,** 13, (36), 3688-95.
18. Duncan, B. B.; Heiss, G., Nonenzymatic glycosylation of proteins--a new tool for assessment of cumulative hyperglycemia in epidemiologic studies, past and future. *Am J Epidemiol* **1984,** 120, (2), 169-89.
19. Kim, K. J.; Lee, B. W., The roles of glycated albumin as intermediate glycation index and pathogenic protein. *Diabetes Metab J* **2012,** 36, (2), 98-107.
20. Doll, S.; Burlingame, A. L., Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chem Biol* **2015,** 10, (1), 63-71.
21. Gould, B. J.; Hall, P. M., m-Aminophenylboronate affinity ligands distinguish between nonenzymically glycosylated proteins and glycoproteins. *Clin Chim Acta* **1987,** 163, (2), 225-30.
22. Zhang, Q.; Tang, N.; Brock, J. W.; Mottaz, H. M.; Ames, J. M.; Baynes, J. W.; Smith, R. D.; Metz, T. O., Enrichment and analysis of nonenzymatically glycated peptides: boronate affinity chromatography coupled with electron-transfer dissociation mass spectrometry. *J Proteome Res* **2007,** 6, (6), 2323-30.
23. Zhang, Q.; Monroe, M. E.; Schepmoes, A. A.; Clauss, T. R.; Gritsenko, M. A.; Meng, D.; Petyuk, V. A.; Smith, R. D.; Metz, T. O., Comprehensive identification of glycated peptides and their glycation motifs in plasma and erythrocytes of control and diabetic subjects. *J Proteome Res* **2011,** 10, (7), 3076-88.
24. Geyer, P. E.; Kulak, N. A.; Pichler, G.; Holdt, L. M.; Teupser, D.; Mann, M., Plasma proteome profiling to assess human health and disease. *Cell Systems* **2016**.
25. Wells, J. M.; McLuckey, S. A., Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* **2005,** 402, 148-85.
26. Lapolla, A.; Fedele, D.; Reitano, R.; Arico, N. C.; Seraglia, R.; Traldi, P.; Marotta, E.; Tonani, R., Enzymatic digestion and mass spectrometry in the study of advanced glycation end products/peptides. *J Am Soc Mass Spectrom* **2004,** 15, (4), 496-509.
27. Frolov, A.; Hoffmann, P.; Hoffmann, R., Fragmentation behavior of glycated peptides derived from D-glucose, D-fructose and D-ribose in tandem mass spectrometry. *J Mass Spectrom* **2006,** 41, (11), 1459-69.
28. Priego-Capote, F.; Scherl, A.; Muller, M.; Waridel, P.; Lisacek, F.; Sanchez, J. C., Glycation isotopic labeling with 13C-reducing sugars for quantitative analysis of glycated proteins in human plasma. *Mol Cell Proteomics* **2010,** 9, (3), 579-92.
29. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **2004,** 101, (26), 9528-33.
30. Zhang, Q.; Frolov, A.; Tang, N.; Hoffmann, R.; van de Goor, T.; Metz, T. O.; Smith, R. D., Application of electron transfer dissociation mass spectrometry in analyses of non-enzymatically glycated peptides. *Rapid Commun Mass Spectrom* **2007,** 21, (5), 661-6.
31. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **2007,** 4, (9), 709-12.

32. Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2007,** 2, (8), 1896-906.
33. Geyer, P. E.; Kulak, N. A.; Pichler, G.; Holdt, L. M.; Teupser, D.; Mann, M., Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* **2016,** 2, (3), 185-195.
34. Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M., Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* **2014,** 11, (3), 319-24.
35. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008,** 26, (12), 1367-72.
36. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **2011,** 10, (4), 1794-805.
37. Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P. A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H. J.; Albar, J. P.; Martinez-Bartolome, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H., ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **2014,** 32, (3), 223-6.
38. Rondeau, P.; Bourdon, E., The glycation of albumin: structural and functional impacts. *Biochimie* **2011,** 93, (4), 645-58.
39. Priego-Capote, F.; Ramirez-Boo, M.; Finamore, F.; Gluck, F.; Sanchez, J. C., Quantitative analysis of glycated proteins. *J Proteome Res* **2014,** 13, (2), 336-47.
40. Neuhauser, N.; Michalski, A.; Cox, J.; Mann, M., Expert system for computer-assisted annotation of MS/MS spectra. *Mol Cell Proteomics* **2012,** 11, (11), 1500-9.
41. Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M., The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol Cell Proteomics* **2014,** 13, (12), 3698-708.
42. Mortensen, H. B.; Christophersen, C., Glucosylation of human haemoglobin a in red blood cells studied in vitro. Kinetics of the formation and dissociation of haemoglobin A1c. *Clin Chim Acta* **1983,** 134, (3), 317-26.
43. Higgins, P. J.; Bunn, H. F., Kinetic analysis of the nonenzymatic glycosylation of hemoglobin. *J Biol Chem* **1981,** 256, (10), 5204-8.
44. Kratz, A.; Ferraro, M.; Sluss, P. M.; Lewandrowski, K. B., Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Laboratory reference values. *N Engl J Med* **2004,** 351, (15), 1548-63.
45. Johansen, M. B.; Kiemer, L.; Brunak, S., Analysis and prediction of mammalian protein glycation. *Glycobiology* **2006,** 16, (9), 844-53.
46. Rabbani, N.; Thornalley, P. J., The critical role of methylglyoxal and glyoxalase 1 in diabetic nephropathy. *Diabetes* **2014,** 63, (1), 50-2.
47. Sousa Silva, M.; Gomes, R. A.; Ferreira, A. E.; Ponces Freire, A.; Cordeiro, C., The glyoxalase pathway: the first hundred years... and beyond. *Biochem J* **2013,** 453, (1), 1-15.

48. Zhang, Q.; Tang, N.; Schepmoes, A. A.; Phillips, L. S.; Smith, R. D.; Metz, T. O., Proteomic profiling of nonenzymatically glycated proteins in human plasma and erythrocyte membranes. *J Proteome Res* **2008,** 7, (5), 2025-32.

FIGURES

Table 1. ***Detected glycation sites on HSA*** (A) Sites ordered by position and their status in UniProt and/or in a recent review[38] if marked by an asterisk. (B) All sites with three valid values ordered by their mean log2 transformed intensity.

Figure 1. ***Evaluation of different collision energies.*** (A) Number of unmodified BSA peptides identified with six different normalized collision energies (NCEs) from 15 to 40 %. (B) Glycation sites identified when searching for glycation on K and R. (C) Localization score as a function of the NCE. (D) Andromeda database identification score[36] as a function of the NCE.

Figure 2. ***HCD fragmentation behavior of glycated peptides.*** (A) Spectrum of the glycated BSA peptide FK*DLGEEHFK with an NCE of 25 % (the asterisk or green color denotes the position of glycation). An almost complete y-ion series is apparent, however, not a single b-ion was found and many peaks in the spectrum are unexplained. (B) Scheme of proposed pathways generating different neutral losses during CID/HCD fragmentation (adapted from ref [39]) (C) The same spectrum as in (A) now manually annotated with the different neutral losses, which explains essentially all fragments.

Figure 3. ***Evaluation of the different neutral losses.*** (A) Number of glycation sites identified in seven different MaxQuant runs of the same data file with no neutral loss (no NL), neutral loss of H_2O (-18 Da), two H_2O (-36 Da), three H_2O (-54 Da), CH_6O_3 (-66 Da), CH_8O_4 (-84 Da), and of the entire Amadori compound (-162 Da). (B) Same spectrum as in figure 2A and 2C now annotated with an almost complete b-ion series due to integrating the neutral loss of three water molecules in the database search. Asterisks on the b-ions indicate that they carry the residual modification after neutral loss of H_6O_3 (standard annotation feature in the MaxQuant viewer).

Figure 4. ***The time-dependency of the in vitro glycation reaction.*** (A) Number of identified glycation sites in triplicate analysis of BSA *in vitro* glycated with 1M glucose for 1-30 days. (B) Analysis of residue and multiplicity of all glycation sites identified on day 30. (C) Heatmap of the intensities of those glycated lysine sites with more than 50 % valid values over the course of the experiment.

Figure 5. ***Properties of glycated peptides and AGE analysis*****.** (A) Analysis of glycated peptides identified in HeLa lysate, showing the preferred site of glycation and their multiplicity i.e. whether identified peptides were glycated one, two or three times. (B) Same analysis for glycated peptides identified in blood plasma. (C) Barplot depicting the number of proteins, glycation sites and some major AGE sites identified in the HeLa sample. (D) Barplot depicting the number of proteins, glycation sites and some major AGE sites identified in the blood plasma sample.

Table 1 (Keilhauer E. C. *et al.*, 2016)

A

Amino acid	Position	Status
K	28	not glycated
K	36	glycated
K	44	not glycated
K	75	*in vitro* glycated
K	88	not glycated
K	97	not glycated
K	130	not glycated
K	160	glycated*
K	161	*in vitro* glycated
K	183	not glycated
K	186	*in vitro* glycated
K	198	not glycated
K	205	not glycated
K	214	not glycated
K	223	*in vitro* glycated
K	236	not glycated
K	249	*in vitro* glycated
K	257	glycated
K	264	not glycated
K	286	not glycated
K	300	*in vitro* glycated
K	337	*in vitro* glycated
K	341	glycated
K	347	*in vitro* glycated
K	375	glycated
K	383	not glycated
K	396	not glycated
K	402	*in vitro* glycated
K	413	glycated*
K	426	not glycated
K	437	*in vitro* glycated
K	438	not glycated
K	460	not glycated
K	490	not glycated
K	499	not glycated
K	548	not glycated
K	549	glycated
K	558	glycated
K	565	not glycated
K	569	*in vitro* glycated
K	597	*in vitro* glycated
K	598	not glycated
R	141	-
R	210	glycated*
R	361	-

B

Position	Mean log2 intensity
257	34.4
549	33.5
438	33.2
36	32.5
223	32.4
402	32.2
375	31.6
214	31.2
236	30.4
548	30.4
300	29.9
426	29.8
341	29.7
198	29.4
569	29.4
97	29.2
499	29.2
413	29.0
161	28.8
44	28.5
490	28.1
88	27.7
437	27.6
205	27.4
396	27.3
383	27.0
130	26.9
337	26.8
286	26.7
598	26.7
75	26.5
264	26.1
249	25.7
548	25.3
549	25.3
186	24.9
460	24.1

Figure 1 (Keilhauer E. C. *et al.*, 2016)

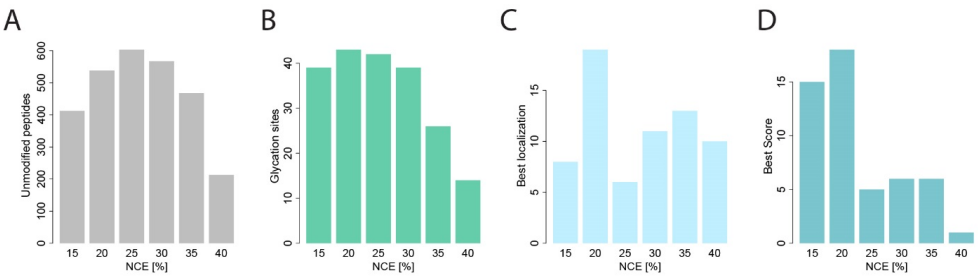


Figure 2 (Keilhauer E. C. *et al.*, 2016)

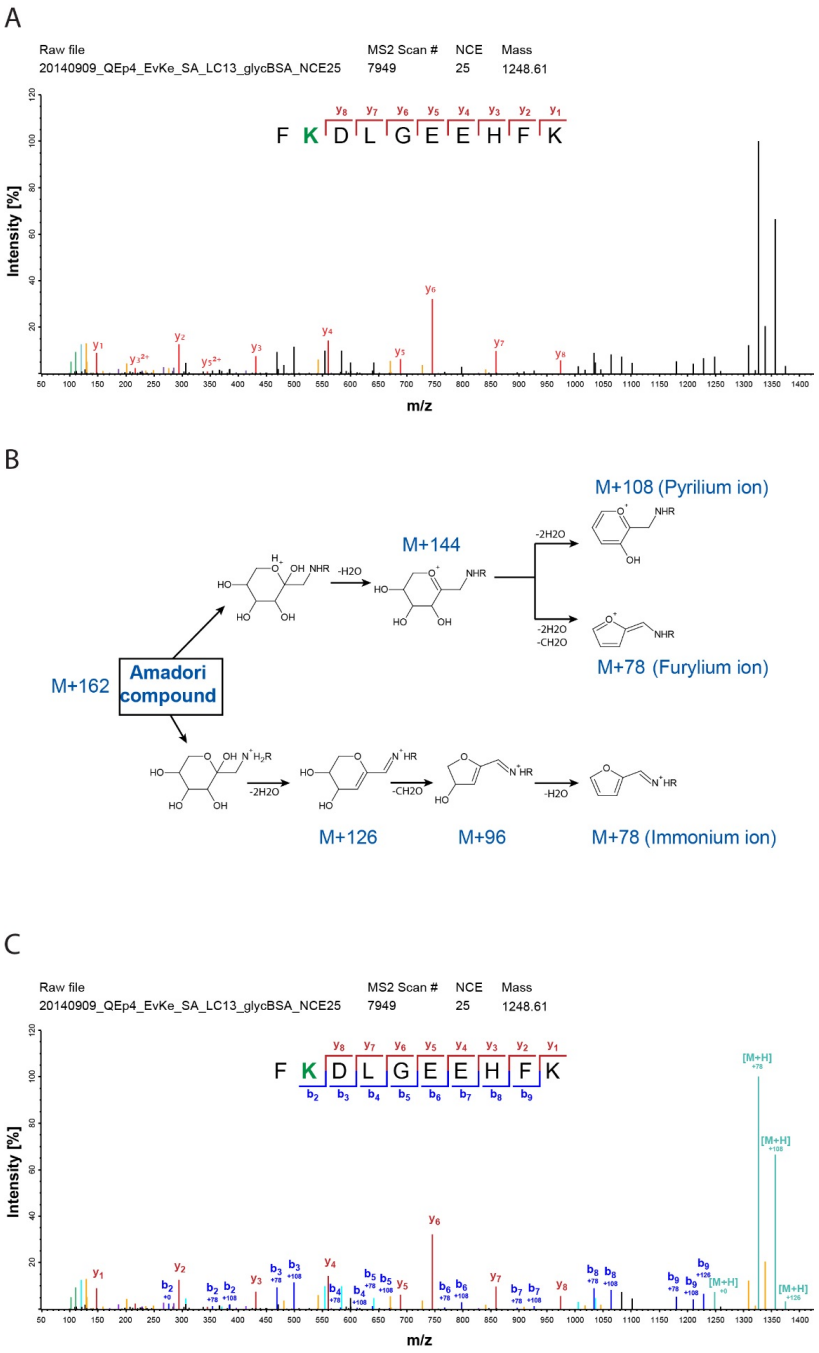


Figure 3 (Keilhauer E. C. *et al.*, 2016)

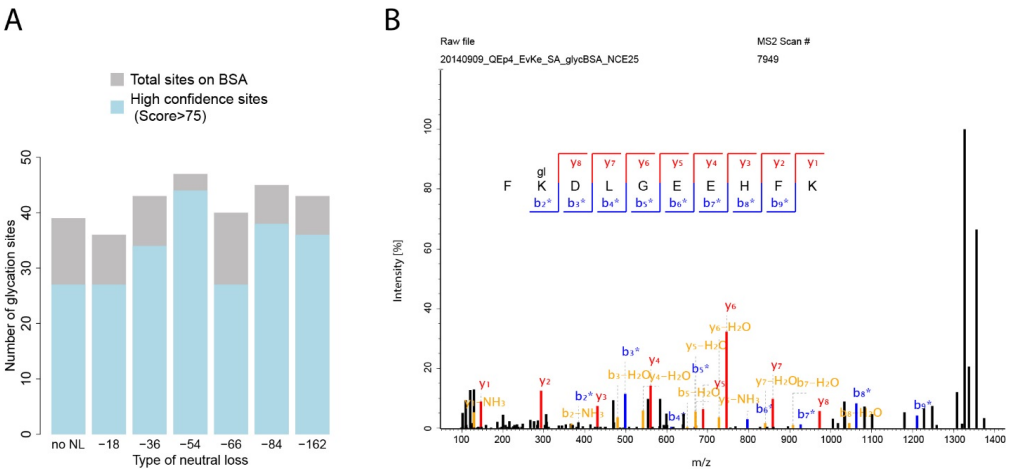


Figure 4 (Keilhauer E. C. *et al.*, 2016)

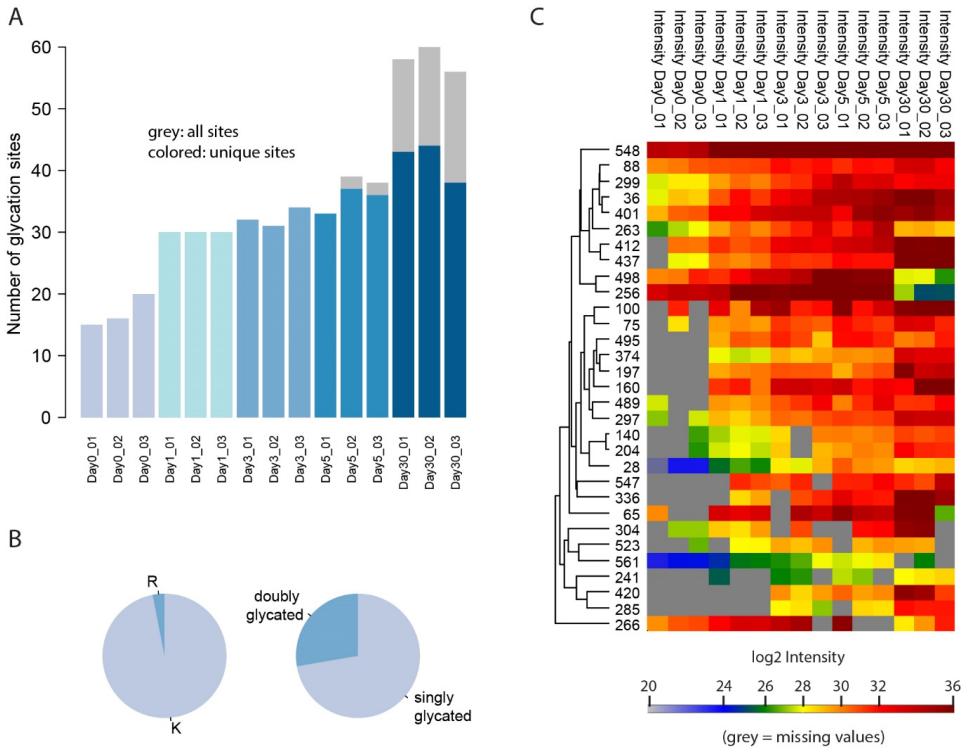
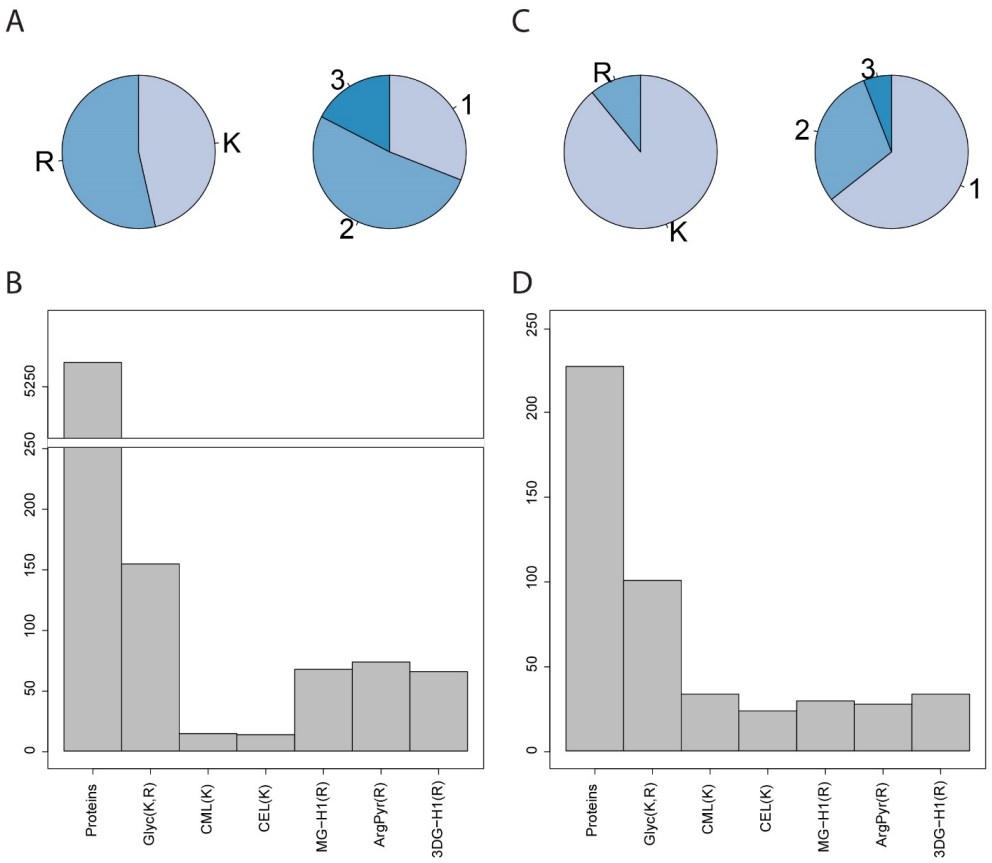
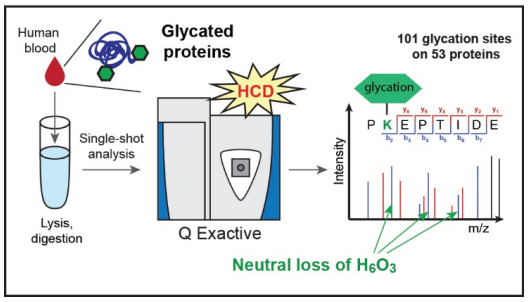


Figure 5 (Keilhauer E. C. *et al.*, 2016)



ABSTRACT GRAPHIC (Keilhauer E. C. *et al.*, 2016)



3 Conclusion and outlook

Mass spectrometry-based proteomics is today established as the most effective technique in proteomic research. It is the only method that can identify proteins in a high-throughput manner from immensely complex samples, and in addition obtain quantitative information about every identified protein. The range of applications in which MS-based proteomics can give new insights is ever growing. In this thesis, I have shown three areas where this technology can make tremendous contributions.

Not long ago, members of protein complexes were identified by developing dedicated multistage biochemical purification schemes for individual stable complexes, a laborious process easily filling a whole doctorate for one complex. Nowadays, large-scale quantitative interaction studies are straightforward to perform, especially with label-free quantification as implemented in the presented AE-MS pipeline. Such studies are now feasible within a relatively short time-frame, due to the development of even faster measurement techniques like the presented double-barrel system. Even though the interaction techniques described in this thesis have been developed for large-scale applications, they are also extremely powerful for answering defined questions, as I have shown in a small-scale project on human histone variants.

The second presented application for MS-based proteomics is the investigation of unusual posttranslational modifications. Next to the tremendous increase in known sites that we have gained from MS-based proteomics experiments for certain well-characterized PTMs like phosphorylation, the possibility of detecting completely unknown modifications, as presented in the EF-P project, is particularly intriguing. I also investigated protein glycation, a non-enzymatic and hence rather untypical PTM relevant in diabetes pathology.

The glycation project is also an example for the third application of MS-based proteomics presented in this thesis, namely clinical proteomics in general and plasma proteomics in particular. The great complexity and dynamic range of blood plasma has so far hampered the successful application of mass spectrometry for investigating plasma proteins. However, now the prerequisites are changing due to various improvements on the technology side, and mass spectrometry will surely soon start to impact on clinical questions.

Interaction proteomics – History, present day and future directions

The global study of protein-protein interactions has only become feasible around the year 2000, when AP-MS techniques became able to create large-scale interaction datasets (see Figure 15). Over the years, the technology has been further refined, with the most relevant step being the implementation of quantitative mass spectrometry into the AP-MS workflow. Since then, it is possible to truly distinguish specific interactors from unspecific background binders. Surprisingly, despite this fact, even today many studies still rely on outdated non-quantitative techniques. Especially since the maturation of label-free approaches into highly accurate quantification strategies, quantitative data can now be acquired in a very straightforward manner. The only requirements for LFQ interaction studies like the AE-MS-pipeline presented in this thesis are replicates and a reproducible sample preparation, two prerequisites that are usually anyway given in biological experiments. When protein quantification finally becomes the standard for interactomics, high confidence large-scale interaction networks will become available for numerous organisms. Such networks will increase our knowledge about which proteins interact with each other, and also allow to obtain additional information like complex topologies and stoichiometries as extremely useful byproducts.

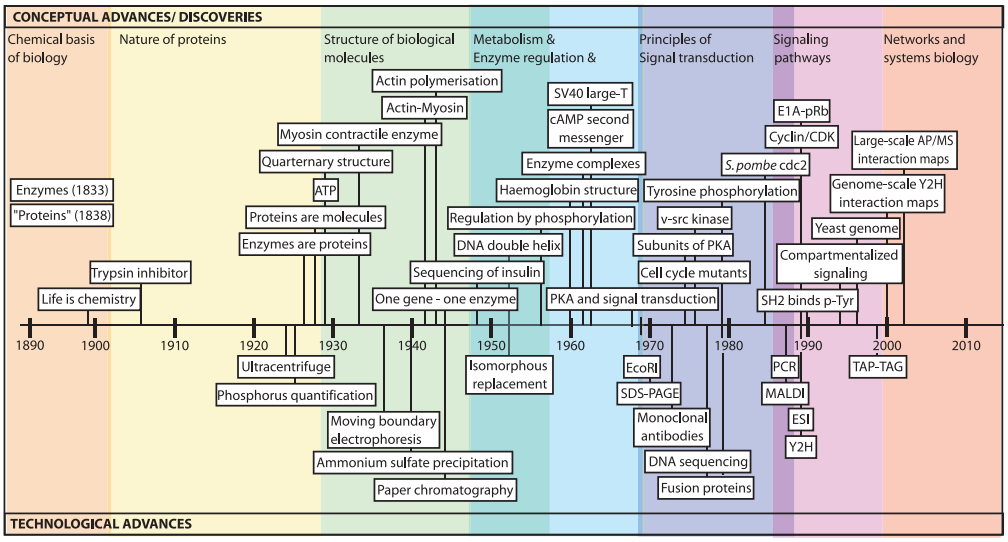


Figure 15: A timeline of some important discoveries and advances in protein research in general and early interactomics in particular. From [156]

The next step forward for interactomics will be to investigate interaction networks under various conditions and perturbations. Right now, most datasets are produced from exponentially growing cells. However, just like the proteome, the interactome is highly dynamic. Investigating cells under various conditions could lead to the discovery of completely new protein complexes, that only form under these conditions and hence could not be detected so far. How interaction networks change in various diseases, and how such changes can impact on the disease biology, will be of particular interest in this area.

Endogenous expression of bait proteins is a prerequisite to obtain meaningful interaction results. So far, true endogenous expression of tagged bait proteins has only been possible in lower complexity organisms, but not in humans. Although technologies such as the BAC strategy used in the QUBIC approach are getting very close to that desired goal, this technique still introduces another copy of the endogenous gene and hence provides close to, but not completely endogenous regulation and expression. However, recently a technique that finally allows tagging in the endogenous locus has been described. The so called CRISPR/Cas system is an acquired immunity mechanism in prokaryotes, that can be be used to edit the human (or any other) genome at any desired location [157–160]. To do so, an appropriate guide RNA that binds the desired DNA locus, and the Cas9 enzyme that recognizes the guide RNA and cleaves the DNA at the targeted position, are transfected into cells. Libraries of human cells endogenously expressing tagged proteins are likely being created right now, and will present excellent resources for mapping the human interactome. With this system, other genetic modifications like the introduction of point mutations, the removal of protein domains, or the knock-out of entire proteins are also possible, and can give further insight into the properties of interactions.

Even though classic AP-MS can address many questions in interaction proteomics, several powerful complementary MS-based approaches exist. Crosslinking techniques can determine complex topologies or retain transient interactors. Their successful large-scale application has been hampered so far by the complex and hard to interpret fragmentation spectra that are produced from crosslinked peptides. However, new developments in this area show great promise, and crosslinking techniques will likely gain momentum in the future. Recently CID/HCD-cleavable crosslinkers have been developed; with this technology crosslinked peptides can now be separated into the two linked peptides by applying modest collision energies in the source region of the mass spectrometer. Subsequently, the now unlinked peptides can be individually isolated and fragmented

leading to MS2 spectra of 'normal' complexity [161]. The recently introduced 'BioID' technology takes a completely different approach to investigate weak and/or transient interactors [162]. In this approach, a biotin ligase is fused to the protein of interest, leading to biotinylations on proteins in close proximity. The proteins modified in this way can be isolated by affinity purification and subsequently analyzed by MS. This approach can also be used to investigate insoluble proteins, and hence complement standard AP-MS datasets. Some progress has also been made in the field of top-down and native proteomics. Native MS approaches can now successfully be used to measure whole protein complexes and even structures as large as virus capsid-antibody conjugates [163].
Collectively, the interaction projects presented in this thesis have demonstrated that MS-based interaction proteomics is a highly powerful technique, and the described developments will hopefully contribute to its further success in the future.

Posttranslational modifications – Investigating less characterized modifications

Mass spectrometry-based proteomics has tremendously increased our knowledge about certain PTMs, however this trend has generally been restricted to modifications with known composition, and modifications where efficient enrichment strategies are available.
Since bottom-up MS relies on database searching to identify peptides and modified peptides, it can inherently not discover modifications with unknown mass. In this thesis, I have applied an interesting search mode that allows for the unbiased detection of completely unknown modifications from standard shotgun experiments. This 'dependent peptide search' compares all unidentified with all identified peptides, based on the assumption that some of the former could not be identified by the standard search because they are modified versions of already identified peptides. So far, the dependent peptide search can only identify dependent peptides when the unmodified counterpart is also present in the sample, however, this can be resolved by not taking all identified peptides as basis for the search but all theoretical peptides. In the EF-P project, I have successfully applied this technique for detecting the modification that activates EF-P, demonstrating the power of the approach. Especially for relatively specific questions, like the one presented here investigating one modification on one particular protein, we think that this search mode can have a big impact in the future.
PTMs on proteins are mostly substoichiometric, hence their successful detection is usually based on specific enrichment to aid identification by MS. Therefore, the analysis

of many other highly interesting PTMs, for which no specific enrichment strategy is available, has lagged far behind. More focus should in the future be put on such under-investigated PTMs, by either developing the required enrichment methods, or by finding ways to investigate them without enrichment. Recently, several new antibody-based enrichment strategies have been developed e.g. for ubiquitination [164], histidine phosphorylation [165] and arginine-methylation [166]. Due to the increased scanning speed and dynamic range of modern MS instrumentation, PTMs can in many cases indeed be detected without enrichment, however naturally not to the same depth.
In this thesis, I have studied protein glycation, an unspecific PTM (i.e. not added by an enzymatic process) where our knowledge so far is quite restricted. In the case of glycation, the thorough investigation has not been prevented by a lack of efficient enrichment strategies, but simply by the fact that is it primarily occurring in blood, the most challenging sample for MS-based proteomics. We propose that the in-depth study of glycated proteins will give new insights in the pathology of diabetes, and potentially allow to better diagnose and/or monitor the disease. Likewise, the investigation of other unusual PTMs should yield interesting new physiological or pathological regulatory mechanisms.

Clinical proteomics

Despite all the knowledge we have gathered in the natural and medical sciences, the number of diseases we completely understand down to the molecular level is relatively small, which hampers the development of new drugs in a targeted manner. Genomic techniques have in many cases been able to identify the mutations underlying certain diseases, however, the effects of those mutations often remain elusive. Well-known examples for such cases include Huntington's disease and hereditary forms of Parkinson's disease. Hence it has become clear, that in order to understand disease pathology, we should concentrate more of our efforts on investigating the proteome.
Initially hampered by technical issues, proteomics is just beginning to move into the clinical field. One challenge was to achieve accurate quantification of proteins from patient samples, which inherently can not be metabolically labeled. However, now patient samples can easily be quantified using chemical labeling techniques (e.g. [167]), using specialized metabolically labeled standards as in the super-SILAC approach [59, 168], or of course using label-free strategies [77]. Another issue was the accessibility of tissue samples from biobanks, a highly valuable source for clinical proteomics research. Such

samples are preserved by fixing with formalin and embedding in paraffin. Some years ago, protocols to efficiently extract peptides from such formalin fixed paraffin embedded (FFPE) samples for MS analysis have been developed, successfully solving this problem [169, 170]. Finally, as already discussed in the introduction, one of the most desirable input materials for clinical studies, namely blood plasma, is also the most challenging one for proteomics research. To some degree, targeted methods successfully circumvent the dynamic range problem in plasma by following only a limited number of analytes, however, they can only be used when the proteins of interest are already known. For the unbiased discovery of protein biomarkers, data-dependent shotgun proteomics is the only way to go. Technical advances on the instrumentation side have expanded the limits of this technology, and together with sophisticated fractionation techniques we hopefully soon can reach sufficient depth to measure down to the highly interesting regulatory plasma proteins and tissue leakage proteins. However, valuable information is already contained in the top abundant proteins, which we can easily measure today. These include for example apolipoproteins, some of which are involved in the development of vascular diseases and hence heart disease and stroke.

Many challenges still remain to be solved before proteomics can be routinely applied to diagnose patients. Most importantly, measurements have to become highly reproducible and robust to allow statistically sound conclusions. Nevertheless, at some point in the future probably not far from now, proteomics will definitely have an enormous impact on the way we diagnose and monitor diseases, and help to provide patients with customized and hence maximal effective therapies.

References

[1] Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**(6788), 837–846 Jun (2000). (↑ p. 1)

[2] Zubarev, R. A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13**(5), 723–726 Mar (2013). (↑ p. 1)

[3] Ong, S.-E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1**(5), 252–262 Oct (2005). (↑ p. 1)

[4] O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**(10), 4007–4021 May (1975). (↑ p. 1)

[5] Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**(20), 2299–2301 Oct (1988). (↑ p. 2)

[6] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**(4926), 64–71 Oct (1989). (↑ p. 2)

[7] Venter, J. C., Adams, M. D., Myers, E. W., *et al.* The sequence of the human genome. *Science* **291**(5507), 1304–1351 Feb (2001). (↑ p. 2)

[8] Lander, E. S., Linton, L. M., Birren, B., *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921 Feb (2001). (↑ p. 2)

[9] Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**(6928), 198–207 Mar (2003). (↑ p. 3)

[10] Heck, A. J. R. Native mass spectrometry: a bridge between interactomics and structural biology. *Nat Methods* **5**(11), 927–933 Nov (2008). (↑ p. 3)

[11] Parks, B. A., Jiang, L., Thomas, P. M., *et al.* Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers. *Anal Chem* **79**(21), 7984–7991 Nov (2007). (↑ p. 3)

[12] Catherman, A. D., Skinner, O. S., & Kelleher, N. L. Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun* **445**(4), 683–693 Mar (2014). (↑ p. 3)

[13] Boeri Erba, E. & Petosa, C. The emerging role of native mass spectrometry in characterizing the structure and dynamics of macromolecular complexes. *Protein Sci* Feb (2015). (↑ p. 3)

[14] Olsen, J. V., Ong, S.-E., & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**(6), 608–614 Jun (2004). (↑ p. 3)

[15] Nagaraj, N., Wisniewski, J. R., Geiger, T., *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548 (2011). (↑ pp. 3 and 22)

[16] Choudhary, G., Wu, S.-L., Shieh, P., & Hancock, W. S. Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *J Proteome Res* **2**(1), 59–67 (2003).

[17] Biringer, R. G., Amato, H., Harrington, M. G., *et al.* Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS. *Brief Funct Genomic Proteomic* **5**(2), 144–153 Jun (2006). (↑ p. 3)

[18] Wu, C., Tran, J. C., Zamdborg, L., *et al.* A protease for 'middle-down' proteomics. *Nat Methods* **9**(8), 822–824 Aug (2012). (↑ p. 3)

[19] Hein, M. Y., Sharma, K., Cox, J., & Mann, M. *Handbook of Systems Biology: Concepts and Insights*, chapter 1: Proteomic Analysis of Cellular Systems, 3–25. Academic Press (2013). (↑ pp. 4, 14, 18, 24, 26, 27, and 29)

[20] Michalski, A., Cox, J., & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10**(4), 1785–1793 Apr (2011). (↑ p. 5)

[21] Bilbao, A., Varesio, E., Luban, J., *et al.* Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **15**(5-6), 964–980 Mar (2015). (↑ p. 6)

[22] Gillet, L. C., Navarro, P., Tate, S., *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11**(6), O111.016717 Jun (2012). (↑ p. 6)

[23] Wells, J. M. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* **402**, 148–185 (2005). (↑ p. 6)

[24] Olsen, J. V., Macek, B., Lange, O., *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**(9), 709–712 Sep (2007). (↑ p. 6)

[25] Zubarev, R. A. Electron-capture dissociation tandem mass spectrometry. *Curr Opin Biotechnol* **15**(1), 12–16 Feb (2004). (↑ p. 7)

[26] Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **101**(26), 9528–9533 Jun (2004). (↑ p. 7)

[27] Mikesh, L. M., Ueberheide, B., Chi, A., *et al.* The utility of ETD mass spectrometry in proteomic analysis. *Biochim Biophys Acta* **1764**(12), 1811–1822 Dec (2006). (↑ p. 7)

[28] Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **11**(11), 601 Nov (1984). (↑ p. 7)

[29] Biemann, K. Mass spectrometry of peptides and proteins. *Annu Rev Biochem* **61**, 977–1010 (1992). (↑ p. 7)

[30] Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**(9), 699–711 Sep (2004). (↑ pp. 7, 8, and 10)

[31] Wysocki, V. H., Tsaprailis, G., Smith, L. L., & Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* **35**(12), 1399–1406 Dec (2000). (↑ p. 8)

[32] Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* **24**(4), 508–548 (2005). (↑ pp. 8 and 9)

[33] Michalski, A., Neuhauser, N., Cox, J., & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *J. Proteome Res.* **11**(11), 5479–5491 Nov (2012). (↑ p. 9)

[34] Neuhauser, N., Michalski, A., Cox, J., & Mann, M. Expert system for computer-assisted annotation of MS/MS spectra. *Mol. Cell. Proteomics* **11**(11), 1500–1509 Nov (2012). (↑ p. 10)

[35] Marshall, A. G., Hendrickson, C. L., & Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* **17**(1), 1–35 (1998). (↑ p. 10)

[36] Hu, Q., Noll, R. J., Li, H., *et al.* The Orbitrap: a new mass spectrometer. *J Mass Spectrom* **40**(4), 430–443 Apr (2005). (↑ pp. 10 and 12)

[37] Schwartz, J. C., Senko, M. W., & Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* **13**(6), 659–669 Jun (2002). (↑ p. 10)

[38] Makarov, A. & Denisov, E. Dynamics of ions of intact proteins in the Orbitrap mass analyzer. *J Am Soc Mass Spectrom* **20**(8), 1486–1495 Aug (2009). (↑ p. 10)

[39] Olsen, J. V., de Godoy, L. M. F., Li, G., *et al.* Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4**(12), 2010–2021 Dec (2005). (↑ p. 11)

[40] Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem* **72**(6), 1156–1162 Mar (2000). (↑ pp. 11 and 12)

[41] Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal Chem* **85**(11), 5288–5296 Jun (2013). (↑ p. 11)

[42] Makarov, A., Denisov, E., & Lange, O. Performance evaluation of a high-field Orbitrap mass analyzer. *J Am Soc Mass Spectrom* **20**(8), 1391–1396 Aug (2009). (↑ p. 11)

[43] Michalski, A., Damoc, E., Lange, O., *et al.* Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **11**(3), O111.013698 Mar (2012). (↑ pp. 12 and 13)

[44] Kelstrup, C. D., Jersie-Christensen, R. R., Batth, T. S., *et al.* Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J Proteome Res* **13**(12), 6187–6195 Dec (2014). (↑ pp. 12 and 13)

[45] Makarov, A., Denisov, E., Kholomeev, A., *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem* **78**(7), 2113–2120 Apr (2006). (↑ pp. 12 and 13)

[46] Michalski, A., Damoc, E., Hauschild, J.-P., *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**(9), M111.011015 Sep (2011). (↑ pp. 12 and 13)

[47] Scheltema, R. A., Hauschild, J.-P., Lange, O., *et al.* The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol Cell Proteomics* **13**(12), 3698–3708 Dec (2014). (↑ pp. 12 and 13)

[48] Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **227**(4693), 1435–1441 Mar (1985). (↑ p. 14)

[49] Consortium, T. U. UniProt: a hub for protein information. *Nucleic Acids Res* **43**(Database issue), D204–D212 Jan (2015). (↑ p. 14)

[50] Eng, J. K., McCormack, A. L., & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**(11), 976–989 Nov (1994). (↑ p. 14)

[51] Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**(18), 3551–3567 Dec (1999). (↑ p. 14)

[52] Cox, J., Neuhauser, N., Michalski, A., *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**(4), 1794–1805 Apr (2011). (↑ p. 14)

[53] Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**(12), 1367–1372 Dec (2008). (↑ pp. 14, 15, and 21)

[54] Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**(3), 207–214 Mar (2007). (↑ p. 15)

[55] Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**(10), 1419–1440 Oct (2005). (↑ p. 15)

[56] Savitski, M. M., Nielsen, M. L., & Zubarev, R. A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics* **5**(5), 935–948 May (2006). (↑ pp. 15 and 16)

[57] Everley, R. A., Kunz, R. C., McAllister, F. E., & Gygi, S. P. Increasing throughput in targeted proteomics assays: 54-plex quantitation in a single mass spectrometry run. *Anal Chem* **85**(11), 5340–5346 Jun (2013). (↑ p. 17)

[58] Ong, S.-E., Blagoev, B., Kratchmarova, I., *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**(5), 376–386 May (2002). (↑ pp. 18 and 19)

[59] Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7**(5), 383–385 May (2010). (↑ pp. 18 and 137)

[60] Oda, Y., Huang, K., Cross, F. R., Cowburn, D., & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96**(12), 6591–6596 Jun (1999). (↑ p. 18)

[61] Gouw, J. W., Krijgsveld, J., & Heck, A. J. R. Quantitative proteomics by metabolic labeling of model organisms. *Mol Cell Proteomics* **9**(1), 11–24 Jan (2010). (↑ p. 18)

[62] Thompson, A., Schäfer, J., Kuhn, K., *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**(8), 1895–1904 Apr (2003). (↑ p. 18)

[63] Ross, P. L., Huang, Y. N., Marchese, J. N., *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**(12), 1154–1169 Dec (2004). (↑ p. 18)

[64] DeSouza, L. V., Taylor, A. M., Li, W., *et al.* Multiple reaction monitoring of mTRAQ-labeled peptides enables absolute quantification of endogenous levels of a potential cancer marker in cancerous and normal endometrial tissues. *J Proteome Res* **7**(8), 3525–3534 Aug (2008). (↑ p. 18)

[65] Hsu, J.-L., Huang, S.-Y., Chow, N.-H., & Chen, S.-H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* **75**(24), 6843–6852 Dec (2003). (↑ p. 18)

[66] Hanke, S., Besir, H., Oesterhelt, D., & Mann, M. Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *J. Proteome Res.* **7**(3), 1118–1130 Mar (2008). (↑ p. 18)

[67] Zeiler, M., Straube, W. L., Lundberg, E., Uhlen, M., & Mann, M. A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol. Cell. Proteomics* **11**(3), O111.009613 Mar (2012). (↑ p. 18)

[68] Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100**(12), 6940–6945 Jun (2003). (↑ p. 18)

[69] Beynon, R. J., Doherty, M. K., Pratt, J. M., & Gaskell, S. J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat Methods* **2**(8), 587–589 Aug (2005). (↑ p. 18)

[70] Brun, V., Dupuis, A., Adrait, A., *et al.* Isotope-labeled protein standards: toward absolute quantitative proteomics. *Mol Cell Proteomics* **6**(12), 2139–2149 Dec (2007). (↑ p. 18)

[71] Singh, S., Springer, M., Steen, J., Kirschner, M. W., & Steen, H. FLEXIQuant: a novel tool for the absolute quantification of proteins, and the simultaneous identification and quantification of potentially modified peptides. *J Proteome Res* **8**(5), 2201–2210 May (2009). (↑ p. 18)

[72] Liu, H., Sadygov, R. G., & Yates, 3rd, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**(14), 4193–4201 Jul (2004). (↑ pp. 18 and 20)

[73] Bondarenko, P. V., Chelius, D., & Shaler, T. A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* **74**(18), 4741–4749 Sep (2002). (↑ pp. 18 and 21)

[74] Cox, J., Hein, M. Y., Luber, C. A., *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**(9), 2513–2526 Sep (2014). (↑ pp. 18 and 21)

[75] Schwanhäusser, B., Busse, D., Li, N., *et al.* Global quantification of mammalian gene expression control. *Nature* **473**(7347), 337–342 May (2011). (↑ pp. 18 and 22)

[76] Bantscheff, M., Lemeer, S., Savitski, M. M., & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* **404**(4), 939–965 Sep (2012). (↑ p. 19)

[77] Megger, D. A., Bracht, T., Meyer, H. E., & Sitek, B. Label-free quantification in clinical proteomics. *Biochim Biophys Acta* **1834**(8), 1581–1590 Aug (2013). (↑ pp. 20 and 137)

[78] Werner, T., Sweetman, G., Savitski, M. F., *et al.* Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal Chem* **86**(7), 3594–3601 Apr (2014). (↑ p. 20)

[79] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **389**(4), 1017–1031 Oct (2007). (↑ pp. 20 and 28)

[80] Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12**(8), 1231–1245 Aug (2002). (↑ p. 20)

[81] Ishihama, Y., Oda, Y., Tabata, T., *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4**(9), 1265–1272 Sep (2005). (↑ p. 20)

[82] Geiger, T., Wehner, A., Schaab, C., Cox, J., & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11**(3), M111.014050 Mar (2012). (↑ p. 21)

[83] Wisniewski, J. R., Hein, M. Y., Cox, J., & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics* **13**(12), 3497–3506 Dec (2014). (↑ p. 22)

[84] de Godoy, L. M. F., Olsen, J. V., Cox, J., *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**(7217), 1251–1254 Oct (2008). (↑ p. 22)

[85] Beck, M., Schmidt, A., Malmstroem, J., *et al.* The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549 (2011). (↑ p. 22)

[86] Kim, M.-S., Pinto, S. M., Getnet, D., *et al.* A draft map of the human proteome. *Nature* **509**(7502), 575–581 May (2014). (↑ p. 22)

[87] Wilhelm, M., Schlegl, J., Hahne, H., *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**(7502), 582–587 May (2014). (↑ p. 22)

[88] Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**(4705), 1315–1317 Jun (1985). (↑ p. 23)

[89] Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**(6230), 245–246 Jul (1989). (↑ p. 23)

[90] Parrish, J. R., Gulyas, K. D., & Finley, Jr, R. L. Yeast two-hybrid contributions to interactome mapping. *Curr Opin Biotechnol* **17**(4), 387–393 Aug (2006).

[91] Rajagopala, S. V., Sikorski, P., Caufield, J. H., Tovchigrechko, A., & Uetz, P. Studying protein complexes by the yeast two-hybrid system. *Methods* **58**(4), 392–399 Dec (2012). (↑ p. 23)

[92] Dunham, W. H., Mullin, M., & Gingras, A.-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **12**(10), 1576–1590 May (2012). (↑ p. 23)

[93] Rigaut, G., Shevchenko, A., Rutz, B., *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**(10), 1030–1032 Oct (1999). (↑ p. 24)

[94] Gavin, A.-C., Bösche, M., Krause, R., *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**(6868), 141–147 Jan (2002). (↑ p. 24)

[95] Gavin, A.-C., Aloy, P., Grandi, P., *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**(7084), 631–636 Mar (2006).

[96] Krogan, N. J., Cagney, G., Yu, H., *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**(7084), 637–643 Mar (2006). (↑ p. 24)

[97] Gingras, A.-C., Gstaiger, M., Raught, B., & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8**(8), 645–654 Aug (2007). (↑ p. 25)

[98] Ho, Y., Gruhler, A., Heilbut, A., *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**(6868), 180–183 Jan (2002). (↑ p. 25)

[99] Ewing, R. M., Chu, P., Elisma, F., *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* **3**, 89 (2007). (↑ p. 25)

[100] Poser, I., Sarov, M., Hutchins, J. R. A., *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**(5), 409–415 May (2008). (↑ p. 25)

[101] Hubner, N. C., Bird, A. W., Cox, J., *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189**(4), 739–754 May (2010). (↑ pp. 25, 26, 35, and 41)

[102] Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M., & Mann, M. A map of general and specialized chromatin readers in mouse tissues generated by label-free interac-

tion proteomics. *Mol. Cell* **49**(2), 368–378 Jan (2013). (↑ p. 25)

[103] Smits, A. H., Jansen, P. W. T. C., Poser, I., Hyman, A. A., & Vermeulen, M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res.* **41**(1), e28 Jan (2013). (↑ p. 26)

[104] Rappsilber, J., Siniossoglou, S., Hurt, E. C., & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal Chem* **72**(2), 267–275 Jan (2000). (↑ p. 27)

[105] Leitner, A., Walzthoeni, T., Kahraman, A., *et al.* Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol Cell Proteomics* **9**(8), 1634–1649 Aug (2010).

[106] Chen, Z. A., Jawhari, A., Fischer, L., *et al.* Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J* **29**(4), 717–726 Feb (2010). (↑ p. 27)

[107] Sutherland, B. W., Toews, J., & Kast, J. Utility of formaldehyde cross-linking and mass spectrometry in the study of protein-protein interactions. *J Mass Spectrom* **43**(6), 699–715 Jun (2008). (↑ p. 27)

[108] Vermeulen, M., Mulder, K. W., Denissov, S., *et al.* Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**(1), 58–69 Oct (2007). (↑ p. 27)

[109] Schulze, W. X. & Mann, M. A novel proteomic screen for peptide-protein interactions. *J. Biol. Chem.* **279**(11), 10756–10764 Mar (2004).

[110] Hanke, S. & Mann, M. The phosphotyrosine interactome of the insulin receptor family and its substrates IRS-1 and IRS-2. *Mol. Cell. Proteomics* **8**(3), 519–534 Mar (2009). (↑ p. 27)

[111] Mittler, G., Butter, F., & Mann, M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res.* **19**(2), 284–293 Feb (2009). (↑ p. 27)

[112] Viturawong, T., Meissner, F., Butter, F., & Mann, M. A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation. *Cell Rep* **5**(2), 531–545 Oct (2013). (↑ p. 27)

[113] Butter, F., Scheibe, M., Mörl, M., & Mann, M. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc. Natl. Acad. Sci. U. S. A.* **106**(26), 10626–10631 Jun (2009). (↑ p. 27)

[114] Scheibe, M., Arnoult, N., Kappei, D., *et al.* Quantitative interaction screen of telomeric repeat-containing RNA reveals novel TERRA regulators. *Genome Res.* **23**(12), 2149–2157 Dec (2013). (↑ p. 27)

[115] Sun, B. & He, Q.-Y. Chemical proteomics to identify molecular targets of small compounds. *Curr Mol Med* **13**(7), 1175–1191 Aug (2013). (↑ p. 28)

[116] Sharma, K., Weber, C., Bairlein, M., *et al.* Proteomics strategy for quantitative protein interaction profiling in cell extracts. *Nat Methods* **6**(10), 741–744 Oct (2009). (↑ p. 28)

[117] Bantscheff, M., Hopf, C., Savitski, M. M., *et al.* Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nat Biotechnol* **29**(3), 255–265 Mar (2011). (↑ p. 28)

[118] Jordan, J. D., Landau, E. M., & Iyengar, R. Signaling networks: the origins of cellular multitasking. *Cell* **103**(2), 193–200 Oct (2000). (↑ p. 28)

[119] Hennrich, M. L. & Gavin, A.-C. Quantitative mass spectrometry of posttranslational modifications: Keys to confidence. *Sci Signal* **8**(371), re5 (2015). (↑ p. 28)

[120] Witze, E. S., Old, W. M., Resing, K. A., & Ahn, N. G. Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* **4**(10), 798–806 Oct (2007). (↑ p. 28)

[121] Olsen, J. V., Blagoev, B., Gnad, F., *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**(3), 635–648 Nov (2006). (↑ p. 29)

[122] Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**(7), 1174–1189 Dec (2010).

[123] Kim, S. C., Sprung, R., Chen, Y., *et al.* Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Mol Cell* **23**(4), 607–618 Aug (2006).

[124] Choudhary, C., Kumar, C., Gnad, F., *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**(5942), 834–840 Aug (2009).

[125] Kaji, H., Kamiie, J.-I., Kawakami, H., *et al.* Proteomics reveals N-linked glycoprotein diversity in Caenorhabditis elegans and suggests an atypical translocation mechanism for integral membrane proteins. *Mol Cell Proteomics* **6**(12), 2100–2109 Dec (2007).

[126] Zielinska, D. F., Gnad, F., Wisniewski, J. R., & Mann, M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* **141**(5), 897–907 May (2010).

[127] Wagner, S. A., Beli, P., Weinert, B. T., *et al.* A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell. Proteomics* **10**(10), M111.013284 Oct (2011).

[128] Kim, W., Bennett, E. J., Huttlin, E. L., *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol Cell* **44**(2), 325–340 Oct (2011).

[129] Ong, S.-E., Mittler, G., & Mann, M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat. Methods* **1**(2), 119–126 Nov (2004). (↑ p. 29)

[130] Choudhary, C. & Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **11**(6), 427–439 Jun (2010). (↑ p. 29)

[131] Hunter, T. The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol Cell* **28**(5), 730–738 Dec (2007). (↑ p. 29)

[132] Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9**(20), 4632–4641 Oct (2009). (↑ p. 30)

[133] Sharma, K., D'Souza, R. C. J., Tyanova, S., *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep* **8**(5), 1583–1594 Sep (2014). (↑ pp. 30 and 31)

[134] Blagoev, B., Ong, S.-E., Kratchmarova, I., & Mann, M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat. Biotechnol.* **22**(9), 1139–1145 Sep (2004). (↑ p. 30)

[135] D'Souza, R. C. J., Knittle, A. M., Nagaraj, N., *et al.* Time-resolved dissection of early phosphoproteome and ensuing proteome changes in response to TGF-β. *Sci Signal* **7**(335), rs5 Jul (2014). (↑ p. 30)

[136] Olsen, J. V. & Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics* **12**(12), 3444–3452 Dec (2013). (↑ pp. 30

and 34)

[137] Olsen, J. V., Vermeulen, M., Santamaria, A., *et al.* Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal* **3**(104), ra3 (2010). (↑ pp. 30 and 31)

[138] Danesh, J., Wheeler, J. G., Hirschfield, G. M., *et al.* C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *N Engl J Med* **350**(14), 1387–1397 Apr (2004). (↑ p. 31)

[139] Antman, E. M., Tanasijevic, M. J., Thompson, B., *et al.* Cardiac-specific troponin I levels to predict the risk of mortality in patients with acute coronary syndromes. *N Engl J Med* **335**(18), 1342–1349 Oct (1996). (↑ p. 31)

[140] Catalona, W. J., Partin, A. W., Slawin, K. M., *et al.* Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. *JAMA* **279**(19), 1542–1547 May (1998). (↑ p. 31)

[141] Rifai, N., Gillette, M. A., & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **24**(8), 971–983 Aug (2006). (↑ pp. 32, 33, and 34)

[142] Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **1**(11), 845–867 Nov (2002). (↑ p. 32)

[143] Baker, E. S., Liu, T., Petyuk, V. A., *et al.* Mass spectrometry for translational proteomics: progress and clinical implications. *Genome Med* **4**(8), 63 (2012). (↑ pp. 32 and 33)

[144] Thakur, S. S., Geiger, T., Chatterjee, B., *et al.* Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics* **10**(8), M110.003699 Aug (2011). (↑ p. 32)

[145] Barnidge, D. R., Goodmanson, M. K., Klee, G. G., & Muddiman, D. C. Absolute quantification of the model biomarker prostate-specific antigen in serum by LC-Ms/MS using protein cleavage and isotope dilution mass spectrometry. *J Proteome Res* **3**(3), 644–652 (2004). (↑ p. 32)

[146] Anderson, N. L., Anderson, N. G., Haines, L. R., *et al.* Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res* **3**(2), 235–244 (2004).

[147] Oe, T., Ackermann, B. L., Inoue, K., *et al.* Quantitative analysis of amyloid beta peptides in cerebrospinal fluid of Alzheimer's disease patients by immunoaffinity purification and stable isotope dilution liquid chromatography/negative electrospray ionization tandem mass spectrometry. *Rapid Commun Mass Spectrom* **20**(24), 3723–3735 (2006).

[148] Kilpatrick, E. L. & Bunk, D. M. Reference measurement procedure development for C-reactive protein in human serum. *Anal Chem* **81**(20), 8610–8616 Oct (2009).

[149] Neubert, H., Muirhead, D., Kabir, M., *et al.* Sequential protein and peptide immunoaffinity capture for mass spectrometry-based quantification of total human β-nerve growth factor. *Anal Chem* **85**(3), 1719–1726 Feb (2013).

[150] Lesur, A., Ancheva, L., Kim, Y. J., *et al.* Screening protein isoforms predictive for cancer using immunoaffinity capture and fast LC-MS in PRM mode. *Proteomics Clin Appl* Feb (2015). (↑ p. 32)

[151] Rosty, C., Christa, L., Kuzdzal, S., *et al.* Identification of hepatocarcinoma-intestine-pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocar-

cinoma by protein biochip technology. *Cancer Res* **62**(6), 1868–1875 Mar (2002). (↑ p. 33)

[152] Sedlaczek, P., Frydecka, I., Gabrys, M., *et al.* Comparative analysis of CA125, tissue polypeptide specific antigen, and soluble interleukin-2 receptor alpha levels in sera, cyst, and ascitic fluids from patients with ovarian carcinoma. *Cancer* **95**(9), 1886–1893 Nov (2002). (↑ p. 33)

[153] NOWELL, P. C. & HUNGERFORD, D. A. Chromosome studies in human leukemia. II. Chronic granulocytic leukemia. *J Natl Cancer Inst* **27**, 1013–1035 Nov (1961). (↑ p. 34)

[154] Karve, T. M. & Cheema, A. K. Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. *J Amino Acids* **2011**, 207691 (2011). (↑ p. 34)

[155] Huh, W.-K., Falvo, J. V., Gerke, L. C., *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**(6959), 686–691 Oct (2003). (↑ p. 41)

[156] Gingras, A.-C. Protein phosphatases, from molecules to networks. *EMBO Rep.* **12**(12), 1211–1213 Dec (2011). (↑ p. 134)

[157] Mali, P., Yang, L., Esvelt, K. M., *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**(6121), 823–826 Feb (2013). (↑ p. 135)

[158] Cong, L., Ran, F. A., Cox, D., *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**(6121), 819–823 Feb (2013).

[159] Jinek, M., East, A., Cheng, A., *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471 (2013).

[160] Cho, S. W., Kim, S., Kim, J. M., & Kim, J.-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol* **31**(3), 230–232 Mar (2013). (↑ p. 135)

[161] Soderblom, E. J. & Goshe, M. B. Collision-induced dissociative chemical cross-linking reagents and methodology: Applications to protein structural characterization using tandem mass spectrometry analysis. *Anal Chem* **78**(23), 8059–8068 Dec (2006). (↑ p. 136)

[162] Roux, K. J., Kim, D. I., & Burke, B. BioID: a screen for protein-protein interactions. *Curr Protoc Protein Sci* **74**, Unit 19.23. (2013). (↑ p. 136)

[163] Bereszczak, J. Z., Havlik, M., Weiss, V. U., *et al.* Sizing up large protein complexes by electrospray ionisation-based electrophoretic mobility and native mass spectrometry: morphology selective binding of Fabs to hepatitis B virus capsids. *Anal Bioanal Chem* **406**(5), 1437–1446 Feb (2014). (↑ p. 136)

[164] Na, C. H., Jones, D. R., Yang, Y., *et al.* Synaptic protein ubiquitination in rat brain revealed by antibody-based ubiquitome analysis. *J Proteome Res* **11**(9), 4722–4732 Sep (2012). (↑ p. 137)

[165] Kee, J.-M., Oslund, R. C., Perlman, D. H., & Muir, T. W. A pan-specific antibody for direct detection of protein histidine phosphorylation. *Nat Chem Biol* **9**(7), 416–421 Jul (2013). (↑ p. 137)

[166] Geoghegan, V., Guo, A., Trudgian, D., Thomas, B., & Acuto, O. Comprehensive identification of arginine methylation in primary T cells reveals regulatory roles in cell signalling. *Nat Commun* **6**, 6758 (2015). (↑ p. 137)

[167] Adav, S. S., Qian, J., Ang, Y. L., *et al.* iTRAQ quantitative clinical proteomics revealed role of Na(+)K(+)-ATPase and its correlation with deamidation in vascular dementia. *J Proteome Res* **13**(11), 4635–4646 Nov (2014). (↑ p. 137)

[168] Deeb, S. J., D'Souza, R. C. J., Cox, J., Schmidt-Supprian, M., & Mann, M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. *Mol. Cell. Proteomics* **11**(5), 77–89 May (2012). (↑ p. 137)

[169] Hood, B. L., Darfler, M. M., Guiel, T. G., *et al.* Proteomic analysis of formalin-fixed prostate cancer tissue. *Mol Cell Proteomics* **4**(11), 1741–1753 Nov (2005). (↑ p. 138)

[170] Ostasiewicz, P., Zielinska, D. F., Mann, M., & Wisniewski, J. R. Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. *J. Proteome Res.* **9**(7), 3688–3700 Jul (2010). (↑ p. 138)

Acknowledgements

I want to express my deep gratitude to everyone who contributed to this work in one or the other way:

First of all thanks to Matthias Mann for getting me into doing a PhD at all, since I initially wanted to leave for industry directly after my master thesis. Now I am deeply grateful for the wonderful experience of my PhD, during which I have learned and experienced countless new things. Thank you for being extremely supportive all along the way, and for the excellent environment you provide.

Thanks to the members of my thesis advisory committee (Prof. Dr. Albert Heck, Dr. Sandra Hake and Dr. Andreas Pichlmair) for great input and a very friendly atmosphere.

Thanks to Jürgen Lassak and Sebastian Pünzeler for the fruitful collaborations, it was inspiring to work with you.

Thanks to all current and former members of the Mann department for the nice atmosphere at work and lots of fun outside the lab e.g. during the retreat, the MaxQuant summer school, all the conferences and so on and so forth.

Thanks to all my office mates from the 'Interaction office' and the 'Blümchen office' for countless good discussions, both scientific and non-scientific.

A tremendous thanks to Marco Hein. I'm very sure that if I wouldn't have ended up sitting next to you I would have learned a lot less in the last four years! For example, without you this thesis would have been written in Word instead of in LaTeX... You have become a good friend and I really miss our everyday talks already... Also thank you for proofreading the introduction of this thesis.

Thanks to Sally Deeb, Marlis Zeiler and Gabi Stöhr for numerous nice lunch/coffee/etc. breaks, other activities and for becoming friends along the way...

Thanks to Scarlet Beck for the pregnancy walks and talks.

Thanks to Richard Scheltema for the fun times in Croatia and at various other occasions, and for your deep knowledge of mass spectrometry.

Thanks to Korbinian Mayr for your incredible technical know-how and your tireless help with the instruments.

Thanks to Igor Paron for being such a positive person and for your help with the instruments.

Thanks to Alison Dalfovo and Theresa Schneider for the great administrative support.

Thanks to Gabi Sowa for excellent chromatography columns and other supplies.

Thanks to Tar Viturawong for introducing me to R.

Thanks to my childhood friend Mona Reineck for the Tuesday lunches at the Bio Mensa, and for sharing the interest in natural sciences since we were sitting next to each other in Mrs. Rabe's Biology LK...

Thanks to all my non-scientific friends for providing me with lots of nice free-time activities that took my mind of science!

Especially thanks to Hannah Stark and Anna Oberländer for always being there for me.

Thanks to my whole family for always being supportive despite not really understanding what I'm doing and what it's good for... :)

Thanks to my father and thanks to my dad, I wish both of you could be here to see this day.

Thanks to my unborn child for being an incredible motivation in general, and particularly to finish this thesis in time.

Thanks to Andreas Lalakos for your support and your love.

Finally, the biggest thanks of all goes to my mum, Andrea Keilhauer, for simply everything.

Nichts in der Geschichte des Lebens
ist beständiger als der Wandel

(Charles Darwin)