TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Proteomik und Bioanalytik

Studies towards the proteome-wide detection, identification and quantification of protein glycosylation

Hannes Hahne

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. D. Haller
Prüfer der Dissertation:	1. Univ.-Prof. Dr. B. Küster
	2. Hon.-Prof. Dr. M. Mann
	(Ludwigs-Maximilians-Universität München)

Die Dissertation wurde am 01.08.2012 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 29.10.2012 angenommen.

Il faut imaginer Sisyphe heureux.

Albert Camus, Le Mythe de Sisyphe, 1942

Content

Abstract

Protein glycosylation is one of the most abundant post-translational modifications of proteins and is involved in virtually any cellular process. While O- and N-linked glycosylation serves a multitude of functions in lumenal compartments, on the cell surface or on secreted proteins, the modification of nucleocytoplasmic proteins with N-acetylglucosamine (O-GlcNAc) mediates and modulates cellular signaling. Mounting evidence indicates that aberrant glycosylation is involved in a variety of inherited and acquired diseases and that altered glycosylation may serve as disease marker or reveal potential drug targets. Mass spectrometry has evolved as key technology for the proteome-wide analysis of glycosylation.

However, the identification of O-GlcNAc peptides by tandem mass spectrometry is still challenging because of the high lability of the O-glycosidic bond in the gas phase. To improve the identification of O-GlcNAc peptides, first an array of peptide fragmentation methods has been systematically evaluated. This comparison led to the development of a simple scoring scheme, which assesses tandem mass spectra for presence and intensity of O-GlcNAc-specific features. The resulting score, termed Oscore, enables the automated interrogation of tandem mass spectra and ranks spectra according to their probability of representing an O-GlcNAc peptide spectrum. Implemented in a two-stage tandem mass spectrometry approach, the Oscore considerably improves the detection of known O-GlcNAc peptides from complex samples. However, the Oscore is most powerful in combination with high resolution/high mass accuracy data. Merely re-analyzing three publically available large-scale proteomic data sets enabled the Oscore-based identification of more than hundred proteins and their sites, suggesting that this modification exists even more widely than previously anticipated. At the same time, these data indicate that the O-GlcNAc modification is considerably less abundant than phosphorylation. Somewhat surprisingly, the re-analysis of two large-scale phosphoproteomic studies uncovered 23 peptides corresponding to 11 mouse proteins, which are modified with phosphorylated O-GlcNAc. Further analyses underpinned the finding that the modification is actual O-GlcNAc-6-phosphate and identified the first modified human protein. Taken together, O-GlcNAc-6-phosphate appears to be a general post-translation modification of mammalian proteins.

Although O-GlcNAc proteins can be identified from large-scale proteomic data, still biochemical enrichment is required for the comprehensive identification of O-GlcNAc proteins and sites. To this end, a metabolic labelling/Click chemistry-based enrichment procedure has been developed, which in combination with β-elimination allowed the identification of 1536 O-GlcNAc proteins and the mapping of 125 O-GlcNAc sites from a single cell line. This approach has then been employed to profile the O-GlcNAc proteome after pharmacological inhibition of O-GlcNAcase. This study indicates that several key signaling proteins, including AMP-activated protein kinase (AMPK) and Sirtuin-1, are involved in nutrient responsive O-GlcNAc cycling.

Last, but not least, carbonyl-reactive tandem mass tags (glycoTMTs) have been explored for the proteome-wide quantification of N-linked glycans. First, basic analytical parameters of glycoTMT reagents with different functional groups were assessed. Merits and limits of the quantification using isobaric and heavy/light labeled glycans were further explored. The practical utility of the developed quantification procedure, which is based on heavy/light labelling of glycans with aminooxy-functionalized TMT, was demonstrated by profiling the N-linked glycan complement of the isogenic human colon carcinoma cell lines SW480 (primary tumor) and SW620 (metastatic tumor). The data revealed significant down-regulation of high-mannose glycans in the metastatic cell line.

Zusammenfassung

Glykosylierung ist eine der abundantesten posttranslationalen Modifikationen von Proteinen und ist an nahezu allen zellulären Vorgängen beteiligt. Während die extrazelluläre und lumenale O- und N-verknüpfte Glykosylierung eine Vielzahl an Funktionen in lumenalen Kompartments, auf der Zelloberfläche und auf sekretierten Proteinen ausübt, hat die Modifikation von nucleo-cytoplasmatischen Proteinen mit N-Acetylglucosamin (O-GlcNAc) hauptsächlich regulatorische Funktion im zellulären Signalgeschehen. Zunehmend wird offenbar, dass aberrante Glykosylierung mit einer Vielzahl an Erbkrankheiten und erworbenen Krankheiten assoziiert ist und als Krankheitsmarker dienen oder mögliche Ansatzpunkte für Medikamente liefern kann. Massenspektrometrie hat sich zu einer Schlüsseltechnologie für die Proteom-weite Untersuchung von Proteinglykosylierungen entwickelt.

Allerdings wird gerade die Identifizierung von O-GlcNAc-modifizierten Peptiden durch Tandemmassenspektrometrie dadurch erschwert, dass die O-glykosidische Bindung in der Gasphase ausgesprochen labil ist. Um die Identifizierung dieser Peptide zu verbessern, wurden systematisch neun verschiedene Fragmentierungsmethoden verglichen und darauf aufbauend einen Scoring-Algorithmus entwickelt, der Tandemmassenspektren auf O-GlcNAc-spezische Signaturen überprüft. Der berechnete Score – Oscore genannt – ermöglicht eine automatische Abschätzung der Wahrscheinlichkeit, ob ein Spektrum zu einem O-GlcNAc-Peptid gehört. Im Rahmen eines zweischrittigen Oscore-basierten Tandemmassenspektrometrie-Experiments konnte die Detektion und Identifizierung von bekannten O-GlcNAc-Peptiden in einer komplexen Probe signifikant verbessert werden. Die Oscore-basierte erneute Analyse von drei öffentlich verfügbaren Zelllinienproteom-Datensätzen ergab die Identifizierung von mehr als Hundert O-GlcNAc-Proteinen, was darauf hindeutet, dass die O-GlcNAc-Modifikation noch abundanter als bislang angenommen ist. Allerdings zeigen dieselben Daten auch, dass Proteinphosphorylierungen deutlich häufiger in der Zelle vorkommen. Die Oscore-basierte erneute Analyse von Phosphoproteom-Datensätzen führte zu der Identifizierung von 23 Peptiden und elf Proteinen aus einem Maus-Datensatz, die mit phosphoryliertem O-GlcNAc modifiziert sind. Weitere Experimente haben zum einen gezeigt, dass die Modifikation mit hoher Wahrscheinlichkeit O-GlcNAc-6-phosphat ist und zum andern das erste menschliche O-GlcNAc-Protein identifiziert. O-GlcNAc-6-phosphat muss vermutlich als eine neue posttranslationale Modifikation von Säugetierproteinen betrachtet werden.

Obwohl O-GlcNAc-modifizierte Proteine – ganz offensichtlich – auch ohne biochemische Anreicherung identifiziert werden können, wird diese doch für eine umfassende Charakterisierung benötigt. Zu diesem Zweck wurde eine Methode entwickelt, die auf der metabolischen Azid-Markierung von O-GlcNAc-Proteinen und Click-Chemie beruht und in Kombination mit β-Eliminierung die Identifizierung von 1536 O-GlcNAc-Proteinen und 125 modifizierten Stellen in einer einzelnen Zelllinie ermöglichte. Eine weitere Studie führte zu der Entdeckung, dass eine Reihe von Schlüsselproteinen im zellulären Signalgeschehen O-GlcNAc-modifiziert und an der O-GlcNAc-abhängingen Zellantwort auf sich ändernde Nahrungsverhältnisse beteiligt sind.

Und schliesslich wurden carbonylreaktive Tandem Mass Tags – glycoTMTs – auf ihr Potential für die Proteom-weite Quantifizierung von N-verknüpften Glykanen untersucht. Um den Nutzen der entwickelten Quantifizierungs-Strategie zu demonstrieren, wurden N-verknüpften Glykane von zwei isogenen Darmkrebszelllinien quantitativ verglichen. Dabei zeigten sich insbesondere signifikante Unterschiede hinsichtlich der Menge an sogenannten ‚high mannose' Glykanen zwischen der vom Primärtumor abstammenden und der metastatischen Zelllinie.

Chapter 1
General Introduction

Proteomics and protein glycosylation

Proteomics

The completion of the human genome project revealed that the number of human genes (20-25,000) was considerably smaller than expected [1]. However, the complexity of an organism is, to a large extent, determined by the dynamic and versatile nature of proteins. Although in principle encoded in the genome, the complexity on protein level arises from combinatorial splicing, processing and modification and is further complicated by the dynamic nature of gene expression, modifications, processing, protein stability, protein-protein interactions, and localization.

Proteomics [2] provides a complementary approach to genomic and transcriptomic technologies and enables the global investigation of biological processes on protein level. Proteomics has undergone an impressive evolution in the past 20 years from two-dimensional gels purely visualizing proteins [3, 4] to the identification and quantification of more than 10,000 proteins from a single cell line [5, 6]. Also the analysis of post-translational modifications (PTMs) such as phosphorylation or ubiquitination can nowadays be achieved to a depth of thousands of modified proteins and ten thousands of sites [7, 8]. Despite recent advances in analytical and computational technologies and instrumentation many areas of proteome research still bear significant future potential.

One of these areas is the proteome-wide analysis of glycosylation. Although protein glycosylation represents one of the most abundant post-translational modifications [9, 10], its large-scale analysis lags far behind simpler and, hence, biochemically more tractable PTMs. Glycoproteomics integrates glycoprotein and proteomic technologies for the system-wide investigation of protein glycosylation. And glycomics, as a closely related field of research, aims at the system-wide interrogation of the glycan complement.

Glycobiology

Glycobiology according to the *Oxford English Dictionary* is “the branch of science concerned with the role of sugars in biological systems”. Although the metabolism and chemistry of sugars and carbohydrates were extensively studied in the first half of the 20th century, they were primarily considered as a source of energy or as structural components. The development of molecular biology since the 1960s has greatly advanced our understanding of biology and facilitated the study of genes and proteins. The study of glycans, in contrast, lagged far behind for many years. This was, by and large, due to their inherent structural complexity, the analytical difficulty in determining their sequences and the difficulties in predicting glycan biosynthesis from DNA templates. However, since then the development of novel technologies, including mass spectrometry and soft ionization techniques, enabled the comprehensive investigation of glycan structures and functions, and established the foundation for an integrated understanding of glycans in cell and molecular biology.

Glycans either as free entities or attached to lipids or proteins are found on all cells and on numerous secreted macromolecules. The most common classes of glycans are depicted in Figure 1. They mediate or modulate cell-cell, cell-matrix and cell-molecule interactions in a wide variety of recognition events, and as such, are considerably involved in the development or functions of complex multicellular organisms. Moreover, the dynamic and reversible attachment of a single sugar, N-acetylglucosamine, to serine or threonine residues of nuclear and cytoplasmic proteins (O-GlcNAc) functions as important modulator of cellular signaling events. Given the important role of

glycans in critical biological processes, it is not surprising that alterations in cellular glycobiology can have severe consequences, and that aberrant glycosylation is involved in a wide variety of diseases.

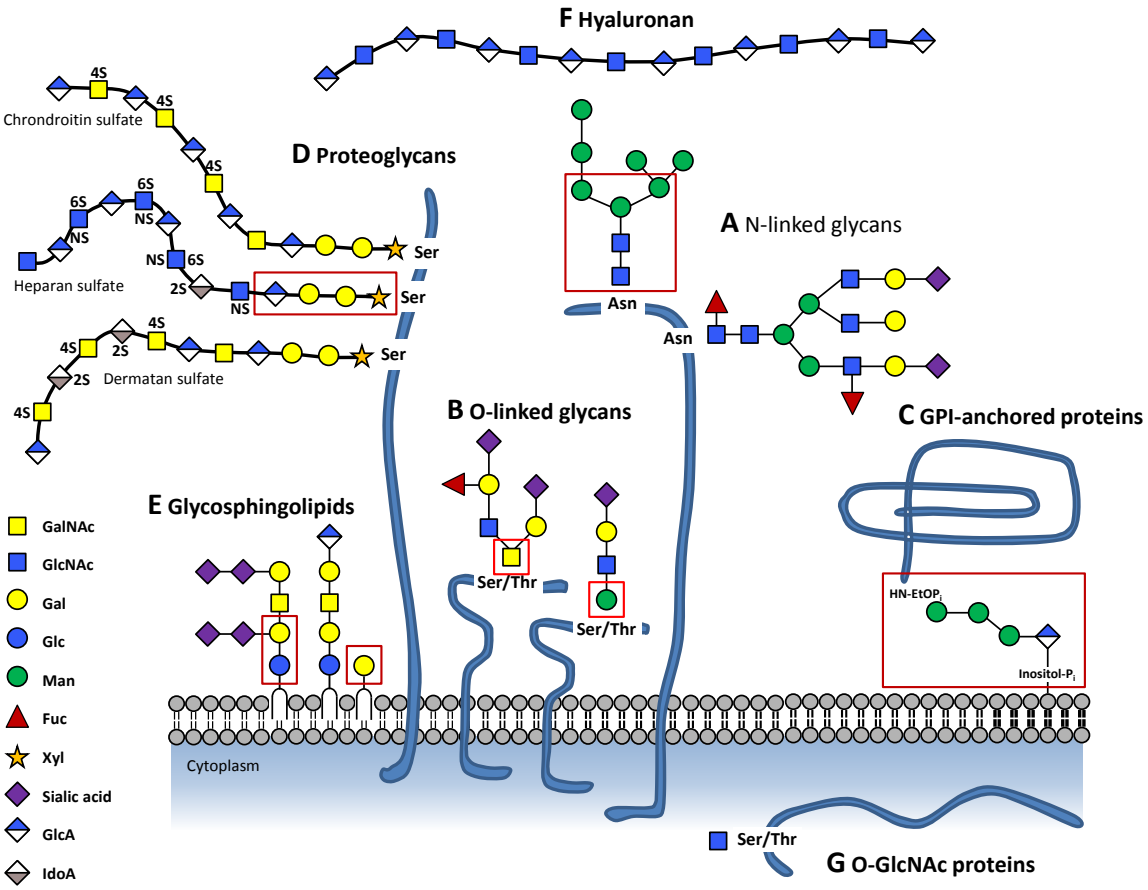

Figure 1 | Overview of common classes of glycans in or on eukaryotic cells; adapted from [11]
Common classes of glycan conjugates in and on eukaryotic cells are **A** N-linked glycans, **B** O-linked glycans, **C** glycosylphosphatidylinositol (GPI)-anchored proteins, **D** proteoglycans, **E** glycosphingolipids, **F** hyaluronan and **G** O-GlcAc modified proteins. Monosaccharide residues are depicted in the recommended nomenclature.

Although, clearly, all classes of glycosylation are involved in key biological functions, it would be beyond the scope of this introduction to cover all aspects of glycosylation. Therefore, the following paragraphs summarize only general features of protein glycosylation with a particular emphasis on N-linked protein glycosylation and O-GlcNAc-modified proteins, which are the underlying subject of this thesis. A comprehensive introduction to glycobiology can be found in ref. 12.

N-linked protein glycosylation

Structure and biosynthesis

N-linked glycans occur on many secreted and extracellular glycoproteins. The reducing end of the glycan is attached to the amide group of an asparagine residue via an N-glycosidic linkage. The asparagine residue is located in an Asn-X-Ser/Thr consensus sequence with X being any amino acid except proline [13]. To a much lesser extent N-glycans are attached to an Asn-X-Cys consensus sequence [14]. N-linked glycans have a defined $Man_3GlcNAc_2$ core structure and can be classified into high mannose-, complex-, and hybrid-type glycans (Figure 2).

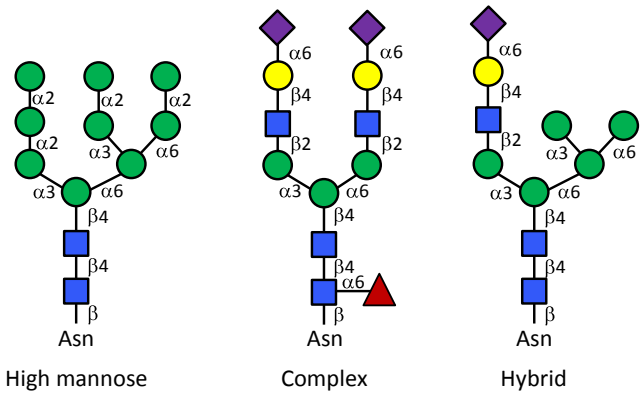

Figure 2 | Typical N-linked glycans; adapted from [15]
Structures of mature N-linked glycans are attached to Asn-X-Ser/Thr acceptor sites on the lumenal or extracellular region of proteins and comprise a common core structure ($Man_3GlcNAc_2$-Asn) which is further processed giving rise to high mannose-, complex- and hybrid-type N-glycans.

Glycoproteins are being synthesized on ribosomes bound to the endoplasmatic reticulum (ER) membrane, and the nascent polypeptides are translocated through the ER membrane. The transfer of a glycan to the acceptor site occurs co-translationally on the lumenal side of the ER membrane.

The biosynthesis of N-linked glycans begins with the step-wise synthesis of a $Man_5GlcNAc_2$-dolichol-diphosphate precursor on the cytoplasmic side of the ER membrane, which is then flipped into the ER lumen and further processed. Upon 'flipping', four additional mannose residues are appended to the Manα1-6 branch, and the Manα1-3 branch is capped with three glucose moieties forming the preassembled $Glc_3Man_9GlcNAc_2$-dolichol-diphosphate precursor.

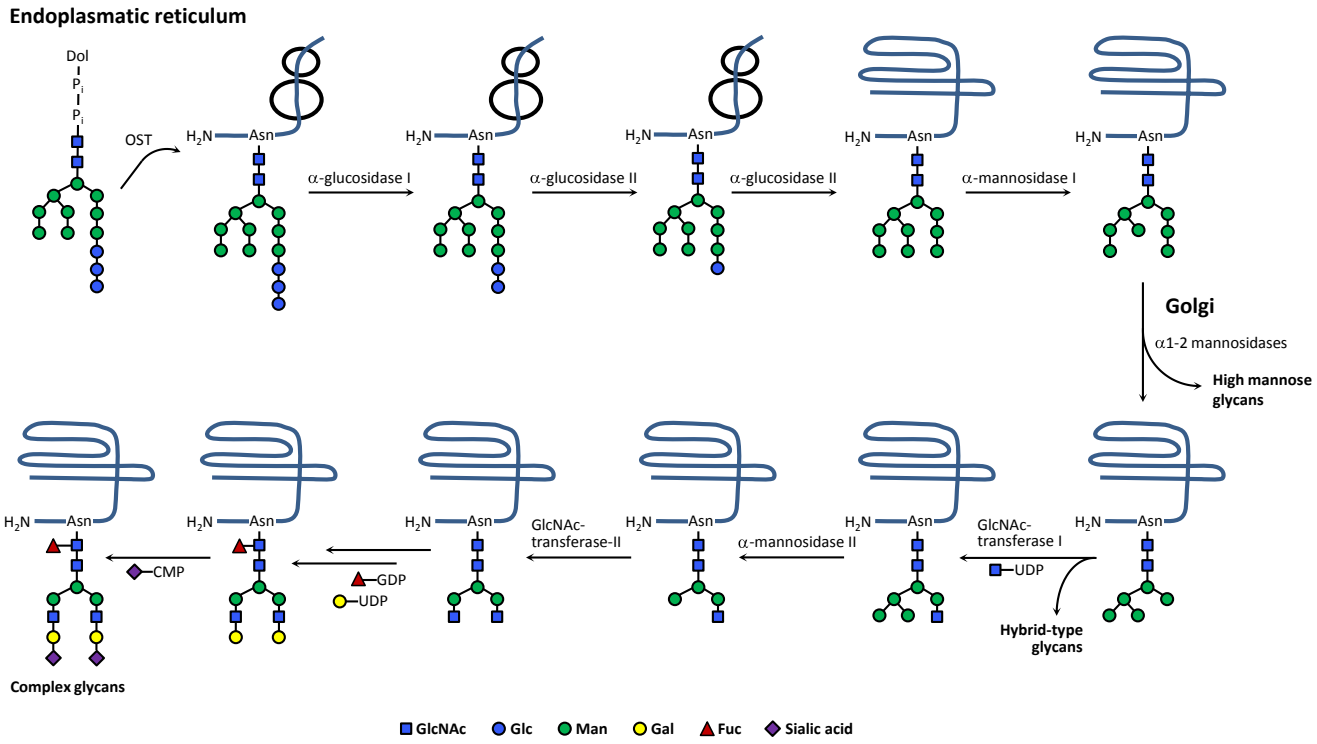

Figure 3 | Schematic representation of some steps in the biosynthesis of N-linked glycans in ER and Golgi
Dol: Dolichol; OST: oligosaccharyl transferase complex

After transfer of the preassembled glycan precursor to the nascent polypeptide by a membrane-bound oligosaccharyl transferase complex, a series of processing reactions trims the N-linked glycan in the ER. Glycans, which are not completely processed to $Man_5GlcNAc_2$-Asn cannot undergo remodeling in the Golgi and, hence, escape further modification as high mannose-type glycans with the composition $Man_{5-9}GlcNAc_2$-Asn. In the Golgi, glycans are further trimmed back by α1-2 mannosidases up to $Man_5GlcNAc_2$-Asn, which represents a key intermediate in the biosynthesis of complex and hybrid N-glycans. Further processing steps involve the action of a variety of glycosidases and glycotransferases resulting in trimming and branching as well as fucosylation of the core, elongation of branches with one or multiple Galβ1–4GlcNAc (LacNAc) repeats and numerous 'capping' reactions, such as the addition of sialic acid, fucose, galactose, N-acetylglucosamine and sulfate. Finally, the biosynthesis of N-glycans result an extensive array of mature, complex N-linked glycans that differ in composition, sequence, branching, capping, linkage, and core modifications.

Glycoproteins exhibit a high degree of heterogeneity with respect to the diversity of attached glycans and their site-specific glycosylation profiles. A homogeneous subpopulation of a given glycoprotein is often referred to as glycoform [16]. The microheterogeneity of membrane-bound and secreted glycoproteins is controlled on multiple levels. A major determinant is clearly the repertoire of processing enzymes present in ER and Golgi, which may differ significantly in a cell-type and species-specific manner [17]. Glycosidases and glycosyltransferases often compete for the same substrate and most require the prior action of other processing enzymes. Thus, even the localization of enzymes within sub-compartments of the Golgi may influence the outcome of glycan processing. The protein itself is another important determinant of glycan diversity. The transfer of the glycan precursor to the nascent polypeptide is mainly determined by the presence of the sequence motif and the absence of proline residues within and adjacent to the motif. N-linked glycosylation is further enhanced adjacent to aromatic residues and in local turns that expose the motifs. In contrast, N-linked glycosylation is less efficient in the proximity to disulfide bonds, transmembrane domains and at the protein termini [18]. Glycoproteins enter the Golgi in their folded state, and secondary and tertiary protein structure then determines further processing reactions by sterically limiting the access of enzymes to the glycan. Also the addition or processing of glycans in close proximity may have an influence on the accessibility. Last, but not least, the attachment and processing of N-linked glycans may as well vary depending on metabolic state and the availability of nucleotide sugar precursors [19].

Functions of N-linked glycans and their role in disease

N-linked glycosylation has been associated with broad range of important physiological functions and influence significantly the life cycle of glycoproteins [12]. The addition of large hydrophilic N-linked glycans may induce changes in the secondary structure and are major determinants of the thermodynamic stability and solubility of proteins [20-22]. In contrast, the distant glucose and mannose residues attached during ER-processing do not interact with the polypeptide backbone, but represent recognition structures for lectins, which assist in folding of nascent polypeptides and direct misfolded polypeptides to degradation [23]. Once correctly folded, the mature glycoproteins can exert their destined function on the cell surface and as secreted glycoproteins. N-linked glycosylation then serves as localization signal targeting proteins to their subcellular destination, regulates the mobility of glycoproteins on or within the cellular membrane through the interaction with lattice forming lectins [24] or protects glycoproteins from proteolysis [25]. N-linked

glycosylation also provides a means to control the half-life of serum glycoproteins [26], and, last but not least, mediate cell-cell interactions [27-31].

As N-linked glycosylation serves a variety of important cellular and physiological functions, it is not surprising that aberrant N-linked glycosylation has been reported in the context of numerous inherited and acquired diseases. Congenital disorders of glycosylation (CDG) is a collection of more than 20 inherited diseases that are associated with impaired N-glycosylation such as unoccupied glycosylation sites (CDG type I) or abnormal glycan profiles (CDG type II) [32]. Aberrant N-linked glycosylation has also been reported in numerous acquired diseases such as cancer or inflammation [33]. The first report of altered glycosylation in cancer dates back to 1969 [34] and it has long been recognized that cell surface glycans contribute to the metastatic and neoplastic properties of tumor cells [35]. Clinical cancer diagnostic markers are often glycoproteins, although most current tests only make use of the protein and not the glycan part [36].

Nuclear and cytoplasmic O-linked *N*-acetylglucosamine

Biosynthesis of O-linked N-acetylglucosamine

The modification of proteins on serine and threonine residues with a single β-N-acetylglucosamine (O-GlcNAc) moiety is an emerging dynamic PTM. It was first discovered by Torres and Hart in 1984 and is distinct from other classes of glycosylation as it occurs solely on cytoplasmic and nuclear proteins [37]. The evolutionary conserved enzyme O-GlcNAc transferase (OGT) catalyzes the attachment of GlcNAc from UDP-GlcNAc to specific protein residues and this reaction can be reversed by O-GlcNAcase. OGT consists of a C-terminal multi-domain catalytic subunit and a N-terminal region containing several tetratricopeptide repeats (TPR), which are proposed to mediate interactions with other proteins, which in turn may confer specificity towards the hundreds of substrates of this enzyme [38]. The nuclear and cytoplasmic O-GlcNAcase (OGA), in contrast, is a bifunctional protein with an N-terminal glycosidase domain, a carboxy-terminal histone acetyl-transferase domain [39] and an interjacent caspase-3 cleavage site [40]. If or how these two enzymatic functions of OGA are interlinked is currently not known.

The sugar precursor for O-GlcNAc is synthesized via the hexosamine biosynthetic pathway (HBP, Figure 4), which plays a central role in carbohydrate metabolism. The end-product of the HBP, UDP-GlcNAc, is not only utilized by OGT, but is required as precursor in the biosynthesis of most glycans. When glucose enters the cell, about 3% are funneled into the HBP [41], which shares the first two enzymes with glycolysis (hexokinase, glucose-6-phsophat-isomerase). The HBP diverges and fructose-6-phosphate is irreversibly converted into glucosamine-6-phosphate (GlcN-6-P) by the glutamine-fructose 6-phosphate transaminase (GFPT). The HBP flux is highly regulated at various levels. The rate-limiting step of the HBP is catalyzed by GFPT and subject to allosteric feedback inhibition by GlcN-6P and UDP-GlcNAc [42]. In addition, GFPT is also regulated at the translational and post-translational level [43]. OGT purifies as multimer and has a remarkably low apparent K_m for UDP-GlcNAc (545 nM). Hence, OGT can still be active at low UDP-GlcNAc concentrations and outcompetes nucleotide transporters of ER and Golgi [44]. Interestingly, the apparent K_m of OGT towards peptide substrates changes with increasing UDP-GlcNAc concentrations, with most peptides becoming better substrates [45]. The nutrient-responsive regulation of the HBP and the unique ability of OGT to respond to changing UDP-GlcNAc levels have prompted researchers to propose that the O-GlcNAc modification may be the terminal step of the nutrient-responsive “hexosamine

signaling pathway", in which UDP-GlcNAc might function as a cellular nutrient and energy sensor [46].

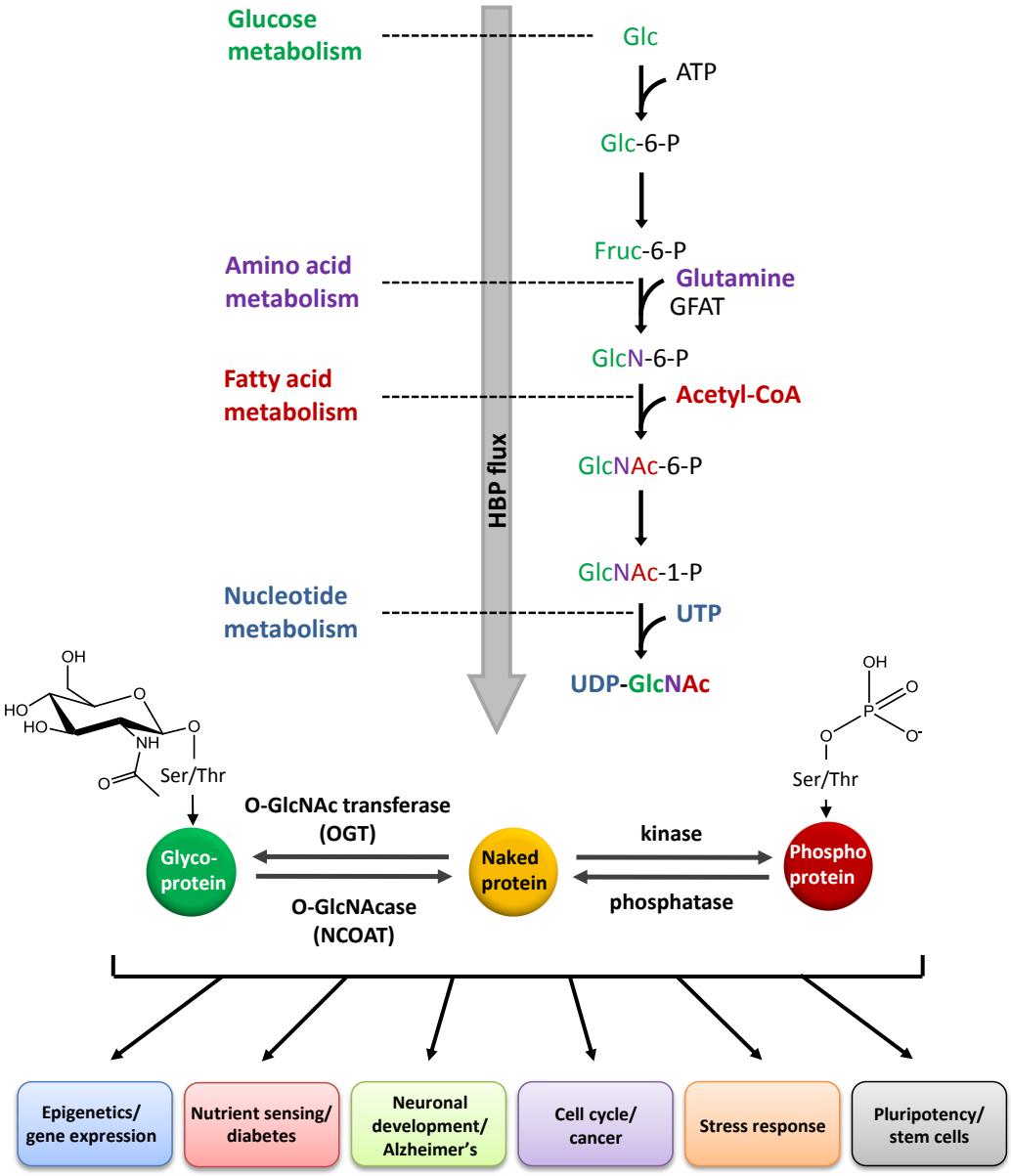

Figure 4 | O-GlcNAc cycling and biological functions; adapted from [47]
The sugar substrate for the O-GlcNAc modification is synthesized via the hexosamine biosynthetic pathway (HBP). A small percentage of intracellular glucose is funneled into the HBP and converted into UDP-GlcNAc. The addition of O-GlcNAc is catalyzed by OGT, and the bifunctional O-GlcNAcase/N-acteyltransferase removes the sugar from nuclear and cytoplasmic proteins. O-GlcNAc is involved in a wide range of cellular processes. GFPT, glutamine-fructose 6-phosphate transaminase.

Cellular functions of O-GlcNAc and role in disease

O-GlcNAc must have many functions beyond nutrient sensing as it is found on a wide range of proteins involved in almost all cellular processes including signaling, cell cycle regulation, transcription and translation regulation, protein trafficking and protein quality control, as well as stress and survival [46-48] (Figure 4). Increasing evidence suggests an interplay between O-GlcNAc and phosphorylation as almost all O-GlcNAc proteins identified so far are also phosphorylated.

Interestingly, many O-GlcNAc sites are identical to or in close proximity to known phosphorylation sites [49] and appear to be mutually exclusive. This has given rise to the concept of the so-called yin yang modification [50]. The links between protein phosphorylation and glycosylation are further strengthened by a recent report showing that kinases are more frequently O-GlcNAc modified than other proteins [51] suggesting that O-GlcNAc may provide an additional level of functional modulation of phosphorylation-regulated signaling cascades.

Regulation by O-GlcNAcylation and phosphorylation has been reported for numerous prominent oncogenes and tumor suppressors, as for instance c-Myc [52], p53 [53] and the estrogen receptor β [54]. The RNA polymerase II subunit A [55] is also a known yin yang protein. The oncogene c-Myc is O-GlcNAc modified on threonine 58, a site which is also phosphorylated by GSK3β, and which is known to be a mutational hot spot in human lymphomas [52]. Another prominent example is the protein tau, the major component of neurofibrillary tangles associated with Alzheimer's disease (AD) which shows altered O-GlcNAcylation/phosphorylation patterns in normal vs. diseased brains [56]. Interestingly, increasing O-GlcNAc levels on tau slows neurodegeneration and stabilizes tau against aggregation, suggesting O-GlcNAcase as a potential target to inhibit progression of AD [57]. Besides the regulation of specific proteins, evidence for globally altered O-GlcNAc levels in human chronic diseases, such as AD, diabetes and cancer is emerging [47, 58].

Mass spectrometry in proteomics and glycomics

Overview

Mass spectrometry is the method of choice for the system-wide analysis of proteins [59] and glycans [60]. The recent, rapid progress of proteomics and glycomics has been substantially enabled by new technologies for peptide sequencing based on mass spectrometry (MS), including soft ionization techniques, such as electrospray ionization and matrix-assisted laser desorption/ionization, and the miniaturization of liquid chromatography (LC) systems. The following paragraphs provide a brief overview of fundamental techniques commonly used in proteomics and glycomics with particular emphasis on protein identification by liquid chromatography-tandem mass spectrometry.

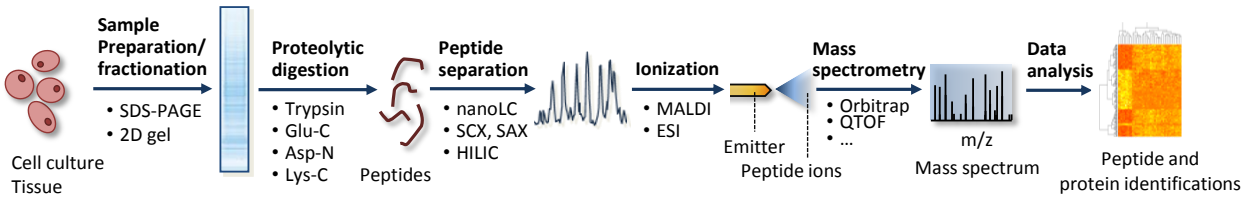

Figure 5 | Generic bottom-up proteomic workflow; adapted from [61]
Proteomic experiments comprise a variety of experimental steps, and each step can be accomplished by various means.

A generic bottom-up (or shotgun) proteomic workflow starting from a certain protein sample, such as a cell lysate or tissue homogenate, may comprise a variety of steps before mass spectrometric analysis (Figure 5). A common feature of every shotgun proteomic workflow is the digestion of proteins into peptides using specific proteases. To reduce the sample complexity, some form of chromatographic or electrophoretic separation is employed prior or subsequent to protein digestion. On protein level, gel electrophoresis followed by in-gel digestion is still widely used.

Common approaches for the separation of peptides are isoelectric focusing (IEF), strong cation or anion exchange chromatography (SCX and SAX, respectively) or hydrophilic interaction chromatography (HILIC). Some form of (affinity) enrichment may be required for the analysis of a (low abundant) sub-proteome such as membrane proteins, post-translationally modified proteins or protein complexes. Following protein digestion, separation and/or enrichment, proteins are identified by tandem mass spectrometry and database search.

A mass spectrometer consists of three parts: an ionization source to generate gas-phase ions, a mass analyzer to separate ions based on their mass-to-charge ratio (m/z) and an ion detector. Ionization of labile biomolecules, such as peptides or glycans, is commonly achieved using so-called soft ionization techniques, such as electrospray ionization (ESI) and matrix-assisted laser desorption/ ionization (MALDI). Upon ionization, the ions are transferred into the mass spectrometer following an electrostatic potential between the source and the mass analyzer, where they are separated according to their m/z ratio. Upon separation, the ions reach the detector and generate a signal. The detector signal and the m/z ratio of the detected ions are then translated into a mass spectrum covering the full range of detected m/z ratios and their intensities.

Soft ionization techniques

Matrix-assisted laser desorption/ionization

MALDI enables the 'soft' transfer of large biomolecules from a solid matrix into the gas phase [62, 63] (Figure 6A). To this end, the analyte molecules are dissolved in a volatile solvent, deposited on a target plate and co-crystallized with an excess of matrix molecules. Matrix molecules are usually small aromatic acids with absorption maxima in the range of the wavelength of the employed laser. The incorporation of analyte molecules into the crystal structure of the matrix molecules is crucial for the following desorption and ionization process. In the ultra-high vacuum of the mass spectrometer the surface of the analyte/matrix crystals is exposed to short, intense laser pulses. The energy transfer occurs via resonant electronic excitation of the π-electrons of the matrix molecules. The first steps of the MALDI process are not yet fully understood [64], but it appears that the energy of the electronically excited matrix molecules relaxes into the crystal lattice and leads almost immediately to the desorption of the analyte and matrix molecules into the gas phase. This expansion process provides rapid cooling of the evaporated analytes and prevents thermal decomposition. Consequently, even very large biomolecules, such as proteins remain intact during desorption. Two models have been postulated for the ion formation in the expanding plume, the so-called 'lucky survivor' and the 'gas phase protonation' model. Depending on experimental conditions, such as laser fluency or matrix, either mechanism seems possible [65]. The MALDI process typically produces singly charged ions, resulting in readily interpretable mass spectra. Larger biomolecules such as proteins can also occur as doubly or even triply charged species. Most commonly utilized lasers are nitrogen lasers with a wavelength of 337 nm and impulse duration of 3-5ns or Nd-YAG laser with a wavelength of 355 nm and impulse duration of 5-15ns.

Electrospray ionization

ESI enables the 'soft' transfer of large and labile biomolecules from solution into gas phase [66]. To this end, analytes are dissolved in a volatile solvent and sprayed from a fine capillary needle, on which a high voltage is applied. The ESI process (Figure 6B) begins with the electrophoretic separation of anions and cations at the tip of the capillary and the formation a Taylor cone. Once the

electric field is stronger than the surface tension, the Taylor cone becomes stable and emits a continuous solvent stream from the tip. With increasing distance, this jet is becoming instable and highly charged droplets are formed. These droplets are stable; however, the continuing evaporation of solvent increases the charge density of the droplet, until the Coulomb repulsion overcomes the surface tension (Raleigh limit). The resulting Coulomb explosion leads to the formation of many smaller droplets. The eventual formation of desolvated ions either occurs by repetitive fission of the droplets until only a single analyte ion remains (charge residue model, CRM) [67] and/or via the emission of analyte ions from the droplet surface (ion emission model, IEM) [68]. ESI typically generates multiply charged analyte ions, which enables the analysis of large biomolecules with mass spectrometers of limited mass range [66]. The development of nanoelectrospray (nanoESI) represented a significant improvement over conventional ESI, as the miniaturization in terms of flow rate and emitter results in a significant increase in ionization efficiency and, ultimately, analytical sensitivity [69, 70]. Notably, this increase is due to the 100-1000x smaller volume of the droplets emitted from the Taylor cone, which comes along with a smaller number of analyte molecules per droplet and, conversely, increased ionization efficiency [69].

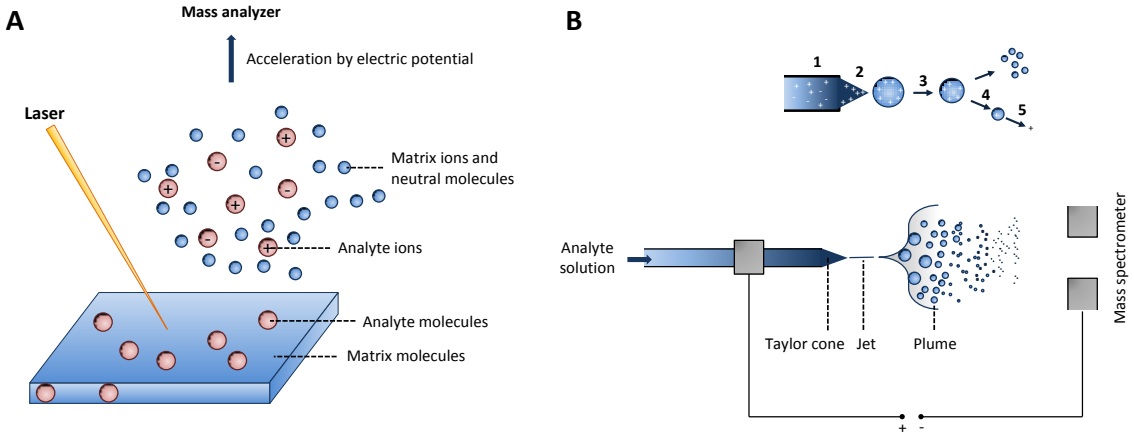

Figure 6 | Soft ionization techniques for labile biomolecules; adapted from [61]
A Formation of analyte ions with MALDI. **B** Formation of analyte ions with ESI. (1) Electrophoretic separation of anions and cations; (2) Taylor cone formation; (3) solvent evaporation from droplets; (4) Coulomb explosion and formation of single solvated analyte ions via CRM or IEM; (5) declustering of solvated ions.

Mass analysis

Many different mass analyzers have been devised and the most common types incorporated into commercial instruments for biomolecular mass spectrometry are briefly presented. A summary of their performance characteristics is given in Table 1.

Table 1 | Performance characteristics of mass analyzers

	TOF	Quadrupole	2D/3D ion trap	Orbitrap
Mass accuracy	< 5 ppm	0.2 – 0.5 Da	0.1 – 0.5 Da	< 2 ppm
Resolving power	> 30,000	< 5000	< 8000	> 200,000
m/z range	> 500,000	< 4000	< 4000	< 2000
Scan velocity	µsec	20 – 200 sec	20 – 200 msec	20 – 200 msec
Dynamic range	$1:10^3$	$1:10^4$	$1:10^3$	$1:10^4$

Time-of-flight

Determining m/z ratios in time-of-flight (TOF) mass analyzers relies on the precisely measured time an ion needs to travel a fixed, field-free distance. All ions acquire the same kinetic energy through acceleration in an electrostatic field, but their velocities differ according to their m/z ratio. The time-of-flight is directly proportional to the square root of their m/z value. An efficient means to improve the resolving power and mass accuracy over linear TOF geometries is the use of a reflectron. A reflectron uses a constant electrostatic field to reflect ions, thereby compensating small differences in the initial kinetic energy of ions with identical m/z value. In addition, also delayed extraction, which is used in combination with MALDI, provides an energy correction required to simultaneously detect all ions of the same m/z ratio irrespective of the initial kinetic energy distribution [71].

Quadrupole

A quadrupole mass analyzer acts as mass filter and is based on the principle that ions of different m/z ratio have different stable trajectories in a quadrupole field. The quadrupole field is generated by two opposing pairs of rods, to which a radio frequency (RF) voltage and a superimposed direct current (DC) are applied. As a consequence of the oscillating electric fields, only certain m/z species have stable trajectories within the quadrupole and can be recorded at the detector. The quadrupole can be used to either select a single m/z value or to scan a range of m/z values by varying the amplitude of the applied fields.

Two- and three-dimensional ion traps

Two- and three-dimensional ion traps enable the storage and manipulation of ions in an adequate electrical field similarly to a quadrupole. Ions can be kept on stable trajectories for a variable duration of time in a (quadrupolar) electric field. A two-dimensional or linear ion trap consists of a quadrupole and the confinement of ions is achieved by a potential barrier on endcap electrodes [72, 73]. A three-dimensional ion trap also consists of two endcap electrodes with hyperbolic geometry and a central ring electrode. The oscillating electric field is applied to the ring electrode and the endcap electrodes are grounded, resulting in a parabolic potential well for the confinement of ions. By systematically varying the field parameters, ions can be ejected from the trap and recorded at a detector in an m/z dependent manner. Depending on the field parameters, ion traps can be utilized to store a broad range of ions or only selected m/z ratios. Linear ion traps have a higher ion capacity, which results in increased sensitivity and dynamic range [74, 75].

Fourier transform ion cyclotron

Mass analysis with a Fourier transform ion cyclotron mass spectrometer (FT-ICR) is based on the cyclotron motion an ion experiences in a strong homogeneous magnetic field. The cyclotron frequency is inversely proportional to the m/z ratio and directly proportional to the magnetic field strength. The ion motion in the magnetic field can be measured as image current, which is induced by the ions passing detector plates. When multiple ions circulate in the ICR cell, the resulting image current (time domain) can be deconvoluted into a mass spectrum (frequency domain) by Fourier transformation. The resulting image current is known as free-induction decay or ‘transient’. An ICR cell consists of two detector plates and two excitation plates. The ion motion in the ICR cell is limited by the magnetic field and an orthogonal electrostatic trapping potential. Once the ions enter the ICR cell, they oscillate closely to the center of the cell. Upon excitation with an RF pulse, ions gain kinetic energy and move to larger orbits, closer to the detections plates. Ions can only be excited when the frequency of the applied RF pulse and the cyclotron frequency are equal (‘resonance excitation’). It

is, therefore, possible to either excite a small selection of ions or a broad range depending on the excitation waveform consisting of the required RF frequencies. The resolving power of FT-ICR mass spectrometers depends on the number of acquired oscillations and is, hence, directly proportional to the strength of the magnetic field (up to 15 Tesla) and the length of the transient. The length of a transient is limited by the mean free path of the ions, which, in turn, depends on the vacuum quality in the ICR cell.

Orbitrap

In an orbitrap mass analyzer ions are stably oscillating in a static electric field generated between an inner spindle electrode and an out electrode. The electrostatic attraction of the inner electrode is compensated by a centrifugal force arising from the initial tangential velocity. Potential barriers generated by end-electrodes prevent loss of ions along the spindle axis. The oscillating ions experience three characteristic frequencies, of which only the frequency of axial oscillations ω_z does not depend on energy, angle, velocity etc. and is, thus, used for mass analysis. The axial oscillations of ions are acquired as image current at the split outer electrodes and deconvoluted into a mass spectrum using Fourier transformation. The ions are injected into the orbitrap mass analyzer in packages collected in an upstream C-shaped ion trap, the so-called C-trap.

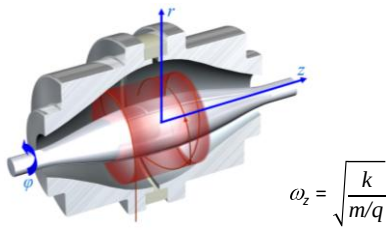

Figure 7 | Model of the orbitrap mass analyzer; adapted from [76]
Upon injection, ions rotate around the inner spindle electrode (ω_φ), and show radial oscillations (ω_r) as well as axial oscillations (ω_z). The frequency of axial oscillations ω_z is directly proportional to the square root of the mass-to-charge ratio and, therefore, used for mass analysis.

The orbitrap mass analyzer features excellent mass accuracy and high resolving power (Table 1). Recent developments in instrumentation, namely a compacter high-field orbitrap, and an enhanced Fourier transform algorithm, which incorporates phase information, significantly improved the resolving power and, thereby, also acquisition speed [77, 78].

Tandem mass spectrometry

Principle

Tandem mass spectrometry enables the structural characterization of peptides (and glycans). The principle idea behind tandem mass spectrometry is that, first, the m/z value of intact analytes are measured, before one analyte is isolated and fragmented within the mass spectrometer. The m/z values of the fragments are then recorded in a so-called tandem mass spectrum. There are two different concepts for tandem mass spectrometry experiments. In the 'tandem in space' approach is the experiment performed in two different mass analyzers within the same instrument. In contrast, in the 'tandem in time' approach, the two steps of the tandem mass spectrometry experiment are performed consecutively within the same mass analyzer. Each approach has certain advantages and disadvantages, but both concepts are widely used. Examples for 'tandem in space' are triple quadrupole (QQQ), quadrupole time-of-flight (QTOF) or quadrupole-orbitrap mass spectrometers, in

which the first quadrupole is used to isolate an peptide and the second mass analyzer to record the fragment mass spectrum. 'Tandem in time' examples are ion traps or FT-ICR instruments. Also combinations of both types exist, such as the linear ion trap-orbitrap instrument.

Fragmentation techniques

The structural characterization of peptides (and glycans) requires the fragmentation of isolated analytes. The most common fragmentation techniques are collision induced dissociation (CID) and electron transfer dissociation (ETD) followed by electron capture dissociation (ECD) and post-source decay (PSD). These techniques are particularly useful for the determination of peptide sequences and glycan structures.

Collision induced dissociation. In CID, peptides are fragmented via multiple collisions with inert gas molecules (He, Ar, N_2). During these collisions the peptides acquire vibrational energy until the breakage of chemical bonds. Most of the bond cleavages observed during CID involve the peptide bond, which renders this technique highly useful for peptide sequence determination. According to the standard nomenclature, the resulting peptide fragments are so-called b- and y-fragment ions (Figure 8). CID techniques can be distinguished into 'low-energy CID' and 'high-energy CID' according to the amount of kinetic energy transferred during the collision process. During high-energy CID (also referred to as beam-type CID), the chemical bonds break almost instantaneous upon collision, leading to information-rich spectra including peptide bond fragments as well as internal and immonium fragments due to multiple bond breakage.

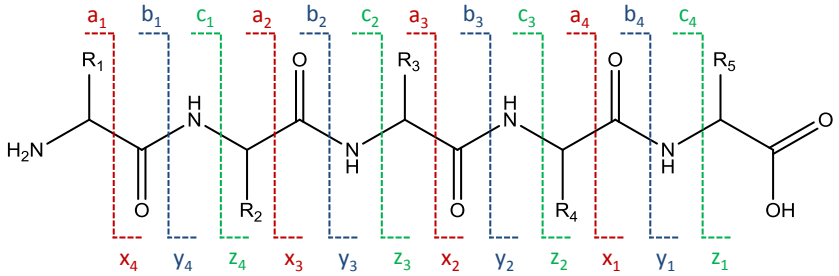

Figure 8 | Peptide fragmentation nomenclature according to Roepsdorf and Fohlmann [79] modified by Johnson *et al.* [80]

Fragment ions which contain the amino-terminus are a, b and c ions, while x, y, and z ions accordingly contain the carboxy-terminus. The fragment ions are named according to the cleaved peptide backbone bond. The mass difference between two adjacent fragments corresponds to the residue mass of an amino acid.

Low-energy CID (or trap-type CID) is often performed in ion traps. The isolated peptides are confined in the ion trap and are excited to larger orbits using resonance excitation. During this process, ions gain kinetic energy and undergo multiple collisions with the gas molecules. The acquired vibrational energy dissipates over the complete molecule leading to a relatively even distribution of fragmented chemical bonds. The fragmentation can be efficiently controlled by the duration and the amplitude of the RF pulse. However, all CID processes are thermodynamically driven resulting in the cleavage of the weakest bond, which poses problems for the analysis of large peptides as well as labile post-translational modifications such as phosphorylation, O-GlcNAc or N-linked glycans. A particular limitation of ion trap CID is the low recovery of fragment ions below ~30% of the precursor mass. This limitation can, in principle, be overcome with alternative ion trap fragmentation techniques

such as pulsed Q dissociation (PQD) [81], high amplitude short time excitation (HASTE technique) [82].

Also glycans can be efficiently fragmented using the described CID fragmentation techniques, and most of the cleavages occur along the glycosidic bonds resulting in B, C, Y or Z type ions or across rings resulting in A and X type ions (Figure 9).

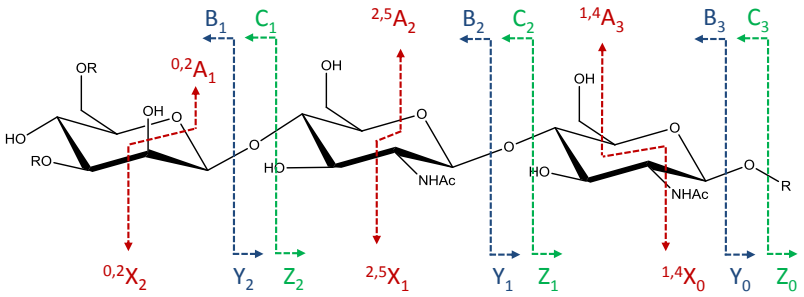

Figure 9 | Glycan fragmentation nomenclature according to Domon and Costello [83]
Fragmentation of the glycosidic bond results in B, C, Y and Z-type ions, while cross-ring cleavages give rise to A and X type ions. X, Y, and Z ions contain the (modified) reducing end of the glycan.

Electron-transfer dissociation. In ETD, the positively charged peptide ions are reacted in an ion trap with an electron donor such as a fluoranthene radical anion [84], which is generated by chemical ionization. The transfer of an additional electron to the peptide leads to a charge-reduced species with an unpaired electron, which is highly instable and undergoes rapid bond cleavage. ETD fragmentation occurs along the peptide backbone, but unlike CID, most ETD fragments result from the cleavage of the N – C_α bond leading to c- and z-ions (Figure 8). Electron-capture dissociation (ECD) is similar to ETD, but uses low energy electrons generated by a heated filament and is primarily implemented on FT-ICR instruments [85]. Unlike CID, the fragmentation mechanisms of ETD and ECD are nonergodic ('kinetically controlled') processes [85] and render peptides with labile modifications amenable to unbiased fragmentation along the backbone [86]. ETD and ECD fragmentation are most efficient with peptides of a high charge density (i. e. high charge, low mass). However, the predominant effect of ETD and ECD for doubly protonated precursors is the charge-reduction of the precursor resulting in low fragmentation efficiency and, eventually, low sensitivity. The fragmentation can be considerably increased by supplemental collisional activation of the non-dissociated electron-transfer products [87].

MS^n fragmentation techniques. The ability to store peptide fragments in ion traps (and FT-ICRs) made the consecutive fragmentation of selected ions possible. In these so-called MS^n approaches (n refers to the tandem MS level), ions are consecutively fragmented and isolated until an MS^n spectrum is recorded. MS^n approaches allow the fragmentation of otherwise un-dissociated fragments, such as neutral loss species from modified peptides, thereby, increasing the coverage of the sequenced peptide. During an MS^3 fragmentation scheme, a peptide is isolated and fragmented before a certain fragment ion is further isolated, fragmented and an MS^3 spectrum recorded. An alternative approach is multistage activation (MSA) [88], in which the fragment ion selected for MS^3 is not isolated, but instead selectively excited and fragmented resulting in the acquisition of a mixed spectrum of tandem MS and MS^3 fragments. In proteomics, MS^3 and MSA are commonly used techniques for the sequencing and site localization of phospho-peptides or other labile

modifications. MS^3 approaches are also gaining popularity in isobaric labeling quantification experiments [89].

Liquid chromatography – tandem mass spectrometry

Even though tandem mass spectrometers are able to analyze mixtures of peptides (or glycans), the complexity of samples generated in typical proteomic experiments easily exceed the peak capacity of tandem mass spectrometers, necessitating additional levels of chromatographic separation prior tandem mass spectrometry-based analysis. To this end, liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) has been developed. While a variety of liquid chromatography techniques is employed in proteomic research, almost exclusively ion-pair reversed phase chromatography is coupled to mass spectrometers. Ion-pair reversed phase chromatography is particularly well suited for the combination with mass spectrometry, since only volatile solvent components are utilized (e. g., water, acetonitrile, organic acids). The enhanced sensitivity in nanoflow LC-MS/MS approaches represented a dramatic improvement over conventional LC-MS/MS for proteomic experiments, where sample quantity is often limiting. Typical nanoflow rates are in the range of 100 to 500 nL/min. Nano-LC columns have an inner diameter of 50 to 100 µm and are packed with fused silica particles of 1 to 5 µm diameter functionalized with C_{18} alkyl chains. In a common setup, peptides are loaded under aqueous conditions onto a trap column for desalting and concentration, and are subsequently eluted and separated onto an analytical (nano-) column with increasing percentage of organic solvent in the mobile phase. Peptides elute off the analytical column in the order of their hydrophobicity.

Liquid chromatography can be coupled to a mass spectrometer in an 'on-line' or 'off-line' configuration. The commonly used 'on-line' coupling is realized with an (nano-) ESI interface. During a LC-ESI-MS/MS experiment peptides elute directly into the (nano-) ESI source and can be immediately analyzed by tandem mass spectrometry, enabling high throughput peptide sequencing. The 'off-line' configuration is realized via MALDI, which involves the collection of eluting peptides on a MALDI target, their co-crystallization with matrix, and tandem mass spectrometry analysis upon completion of the LC experiment.

Data-directed acquisition

A crucial part in every on-line LC-MS/MS experiment is the process of how intact peptide ions are selected for subsequent fragmentation. The most common strategy is the data-dependent acquisition of tandem mass spectra, which is based on the consecutive fragmentation of the most abundant peptides. To avoid repeatedly re-sequencing the same (abundant) peptides, dynamic exclusion lists populated with sequenced precursor ions are used. Moreover, the acquisition process can be directed with global 'inclusion lists' or 'exclusion lists', to support the targeted sequencing of (low abundant) peptides or to avoid the sequencing of undesired peptides, such as those originating from contaminants. Moreover, algorithms have been developed which support the recognition of the LC elution profile of peptides in order to determine the best time point for the sequencing of a peptide. Data-dependent acquisition approaches can further comprise multiple types of fragmentation resulting in complementary tandem mass spectra for the same peptide or MS^n fragmentation techniques supporting the identification of modified peptides, as well as data-dependent decision-tree-based approaches [90] for the optimal choice of fragmentation types for particular peptides. Despite these efforts, data-dependent decision making is still a stochastic

process and replicate experiments of the same sample are highly likely to identify different peptides and proteins.

Protein identification by mass spectrometry

When using mass spectrometry data, proteins can be identified from primary sequence databases by their peptide mass fingerprint (PMF) or by peptide sequence information from tandem mass spectrometry. Briefly, a PMF is a mass spectrum of a peptide mixture resulting from the digestion of a protein by an enzyme [91-93]. Such a mass spectrum provides a fingerprint of great specificity. However, the approach requires that the protein sequence is contained in a database, and can (almost) only deal with proteins purified to homogeneity. Also small proteins are difficult to identify, because they only generate few peptides. The very high mass accuracy afforded by modern mass spectrometers considerably increases the specificity of a PMF.

Proteins can also be identified from peptide tandem mass spectra and primary sequence databases. This approach is commonly referred to as 'database searching' and has been recently comprehensively reviewed [94]. In the database search approach, peptide sequences are identified by correlating acquired peptide tandem mass spectra with theoretical tandem mass spectra predicted for each peptide present in a certain protein sequence database (Figure 10A). For each query, the theoretical peptides are obtained from the sequence database according to a variety of parameters, such as intact peptide mass, enzyme specificity, allowed number of missed cleavage sites, or modifications.

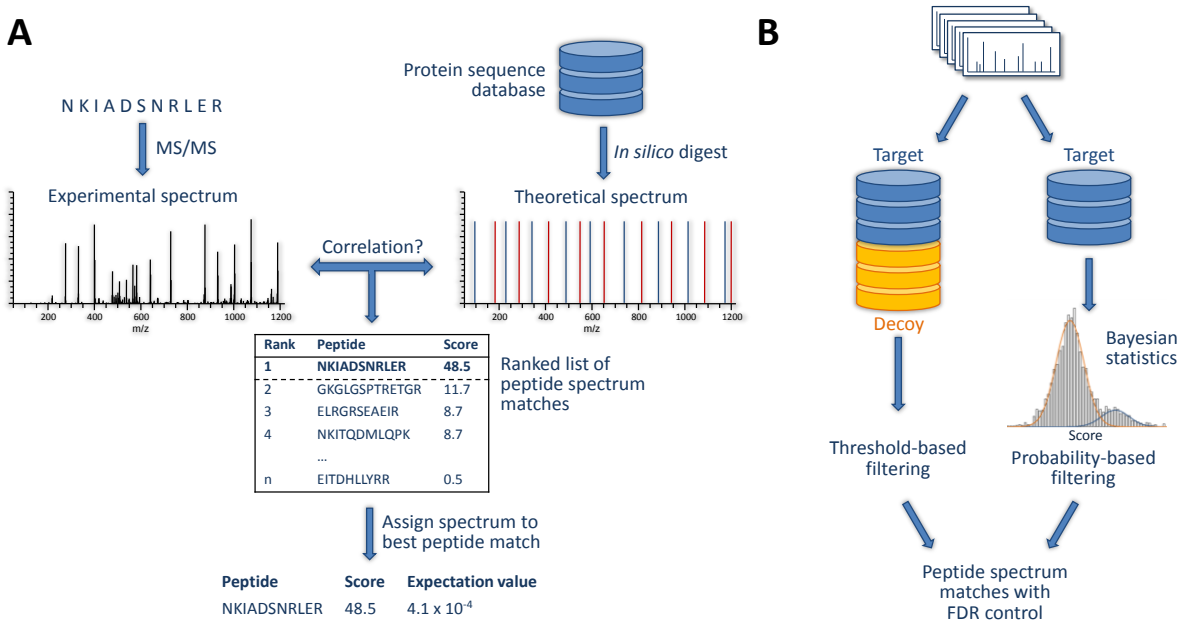

Figure 10 | Peptide identification from tandem mass spectra and database search
A Peptides, and eventually proteins, can be identified from peptide tandem mass spectra by comparison of acquired and theoretical spectra. The higher the degree of similarity, the better is the resulting score for the peptide spectrum match. Scores are often converted into an expectation value representing the probability that the observed match is a random event. **B** False-discovery rate control via target-decoy search and score threshold filtering or (Bayesian statistics) probability-based filtering.

The acquired spectrum and all possible theoretical spectra are then scored according to their degree of similarity taking a variety of information into account (e. g. length of an ion series, gaps, overlap of

ion series, number of fragments matched, percentage of total intensity matched, precursor mass error, fragment mass error, etc.). There are a number of different scoring schemes and an overview is provided in the aforementioned review. Once peptides have been identified by database search, they can be assigned to proteins (or groups of proteins). This is not a trivial procedure, as the same peptide can occur in multiple different protein sequences, and often proteins (or protein isoforms) cannot be distinguished based on the identified peptides [95].

False-positive peptide spectrum assignments (and, hence, false protein identifications) are a common problem of all scoring algorithms, and even a high score or highest rank do not rule out this possibility. Some common sources of false peptide spectrum matches are low quality or chimeric (i. e. mixed) tandem mass spectra, sequenced peptide not present in queried database, or an incorrectly determined charge state or peptide mass [94]. Owing to the large number of tandem mass spectra acquired in a typical proteomics experiment it has become impossible to manually validate all peptide spectrum matches. Two different approaches have been devised to control the number of false positive peptide identification (Figure 10B). Briefly, the so-called target-decoy search strategy allows a global estimation of the false-discovery rate and makes use of a normal 'target' sequence database and a randomized 'decoy' protein sequence database. Peptide spectrum matches in the decoy database are per definition false-positive hits and it is assumed to obtain a similar number of 'invisible' false-positive hits in the target database. The Bayesian statistics approach, in contrast, models the score distribution of false and true hits and estimates a local peptide probability.

Structural analysis of glycans by mass spectrometry

Information obtained by tandem mass spectrometry of intact and fragmented glycans can be utilized to determine composition, sequence, branching and linkage. While protein identification from tandem mass spectrometry data can be performed in a highly automated fashion, structure determination of glycans still requires extensive manual spectrum interpretation.

Glycan compositions can be readily determined from exact mass measurements of intact glycans, which are afforded by high resolution/high ass accuracy mass spectrometers such as orbitrap, FT-ICR or TOF instruments. However, the differentiation of structural isomers requires information from tandem mass spectrometry. Branching information can be deduced from tandem mass spectra of permethylated glycans, because all hydroxy and amino groups are derivatized. The fragmentation of permethylated glycans generates unmodified sites (so-called 'scars') indicating the position of the cleaved bond, which in turn can be read out from cross-ring cleavage ions. Likewise, cross-ring cleavages are highly useful for determining the linkages and substituents of branching monosaccharides. Multiple levels of MS^n fragmentation can be used to produce tandem mass spectra with readily interpretable product ion pattern [96]. In contrast, tandem mass spectra of native or reduced glycans produce less informative spectra because the linkage and branching sites are not tagged by 'scars'. In this case, branching and linkage information is only available, when cross-ring cleavages occur, in which the substituents of the monosaccharide remain intact.

Further detailed structural analyses, such as the determination of LacNAc linkage or sialic acid linkages (α2,3 and α2,6), the differentiation of branching isomers and the determination of antenna composition often require additional tandem mass spectrometry techniques including, but not limited to negative ion mode acquisition of spectra and alternative fragmentation techniques [60, 97].

Mass spectrometry-based approaches to the proteome-wide study of N-linked and O-GlcNAc glycosylation

Proteome-wide study of N-linked protein glycosylation

Mass spectrometry currently is the method of choice for the compositional and structural analysis of protein glycosylation [98], and has developed into an indispensable tool for the investigation of protein glycosylation in physiological and pathological processes [60, 99]. While the proteome-wide analysis of N-linked glycosylation sites can be addressed with an array of glycoproteomic methods after removal of the glycan moiety, the large-scale mass spectrometry-based analysis of intact glycopeptides still represents a considerable challenge [100]. The major difficulty during MS-based analysis of intact glycopeptides is the vastly different stability of O-glycosidic and peptide bonds under typical CID conditions, which impairs the concomitant fragmentation and sequencing of the glycan moiety and the peptide backbone [101].

In glycomic experiments, the proteome-wide analysis of released N-linked glycans can be achieved with a variety of approaches and commonly involves the separation of glycans from the glycoprotein samples, derivatization, LC-based separation of glycans and MALDI or ESI (tandem) mass spectrometry. N-linked glycans are commonly released from peptides or proteins by specific enzymes, such as PNGase F, which cleaves the N-glycosidic Asn – glycan bond, or by chemical means [102]. Upon release, glycans are often derivatized to overcome their lack of chromophores, limited retentivity on many chromatographic supports, poor ionization efficiency during MALDI and ESI, and undesired fragmentation of acidic glycans during MALDI mass spectrometry. The most widely used derivatization strategies include reductive amination, Michael addition, hydrazide or aminooxy labeling, permethylation as well as selective methylation of acidic glycans [102].

According to a recent international inter-laboratory study organized by the Human Disease Glycomics/ Proteome Initiative (HGPI), the most common approaches for the analysis of N-linked glycans can be broadly divided into three groups: *i)* LC-based separation and detection of fluorescently labeled glycans, *ii)* MALDI-MS-based analysis of (per-)methylated glycans, and *iii)* LC-ESI-MS-based analysis of reduced glycans [99]. Quantification by mass spectrometry of the same glycan species across two or more biological states is either achieved by stable isotope labelling or label-free intensity-based quantification akin to common current proteomic methods. Stable isotope labelling typically follows one of two general ideas: Either, glycan stable isotope labelling is performed by permethylation [103-105] or at the reducing end, e. g. using reductive amination [106-109], hydrazone [110] or oxime [111] formation. A detailed introduction into glycan quantification based on stable isotope labelling is provided in Chapter 6.

Fluorescent tagging by reductive amination and HPLC-based separation and detection of glycans has been one of the first approaches for HPLC-based glycan structure determination [112-114]. The derivatization requires a free reducing end on the released glycans, which is usually obtained by enzymatic deglycosylation. Although fluorescent tagging is currently probably the best approach for the quantification of N-linked glycans, it has a couple of drawbacks, namely the resolution of glycan structure and/or compositional isomers and the co-elution of glycans from complex glycomic samples, which compromises glycan quantification. The utility of fluorescence- and HPLC-based quantification is demonstrated in a recent study, which enabled the detection of N-glycosylation features associated with colorectal cancer tissues [115]. In conjunction with mass spectrometry-

based structure elucidation, this study revealed glycan patterns that might be used as candidate biomarkers for colorectal cancer.

Alternatively, (per-)methylated glycans can be analyzed by MALDI-TOF, enabling a rapid screening of glycan compositions. More detailed structural information can be obtained by tandem mass spectrometry. Permethylation stabilizes sialic acid glycans, which otherwise would undergo loss of sialic acid in MALDI-TOF [116]. The directional fragmentation of permethylated glycans facilitates the interpretation of tandem mass spectra and, hence, sequence and branching determination. However, isobaric glycan isomers cannot be readily distinguished. In principle, this problem could be overcome with MS^n approaches (n = 3 … 9) [117]; but often the available sample quantity is limiting, and the unambiguous differentiation of all compositional isomers therefore not possible. Another issue with permethylation is that an incomplete derivatization significantly increases sample and spectrum complexity. A valuable alternative to permethylation is the selective methylesterification of glycans [118]. The potential of the MALDI-TOF analysis of (per-)methylated glycans is nicely exemplified in a recent study of Alley and co-workers, who analyzed the altered glycosylation profile in serum of ovarian cancer patients and found significant differences for various glycan groups [119].

The third alternative for large-scale N-glycan analysis is porous graphitized carbon chromatography (PGC) followed by ESI tandem mass spectrometry. This approach enables the reproducible and sensitive isomer separation of released, reduced glycans. The extraordinary separation capacity of PGC requires that the reducing end is present in the reduced form otherwise separation of anomers can be observed. Although this method enables an excellent separation and the acquisition of tandem mass spectra of glycan isomers, linkage determination requires extensive manual spectrum interpretation. In addition, although chromatographic separation of isomers is indicative of different linkages, which may, in some cases, be deduced from diagnostic fragment ions or retention time, often additional approaches such as exoglycosidase digestion and rechromatography. Using such an approach, Packer and co-workers recently uncovered glycan structure alterations on cell membrane proteins of desoxyepothilone B-resistant leukemia cells [120]. They observed a lower α2-6 sialylation of the core fucosylated α2-6 monosialo-biantennary glycan and could correlate this to a decrease in activity of the corresponding enzyme, β-galactoside α2-6 sialyltransferase, as well as decreased expression of the mRNA.

O-GlcNAc proteomics

The identification of O-GlcNAc proteins and modification sites has seen some progress in the past few years mainly by developments in tandem mass spectrometry [49] and biochemical enrichment techniques. However, large-scale analysis of O-GlcNAc proteins in cells or tissues has not been achieved yet mainly because of the substoichiometric occupancy of O-GlcNAc sites [48, 49, 121] which requires efficient biochemical enrichment. In addition, the chemical lability of the O-glycosidic bond in the gas phase hampers mass spectrometric detection of modified peptides [122, 123].

Under typical CID conditions, O-GlcNAc modified peptides readily loose the GlcNAc moiety and spectra are typically dominated by intense neutral loss species as well as the GlcNAc oxonium ion (*m/z* 204.0866) and its further fragments [124]. Peptide sequence identification is often still possible but site information is irretrievably lost upon dissociation of the O-glycosidic bond. The oxonium ion has been known for a long time as a diagnostically useful reporter ion for O-linked GlcNAc and GalNAc as well as N-linked glycosylation [122] and the initial detection of O-GlcNAc peptides is strongly facilitated in CID-type experiments [121, 124]. Particularly when high resolution and high

mass accuracy instruments are used, the GlcNAc losses along with the GlcNAc oxonium ion and its fragments define a characteristic diagnostic fragment ion pattern, which aids O-GlcNAc peptide identification even in very complex proteomics samples [125-127]. Unlike fragmentation by CID and related approaches, ECD and ETD preserve labile post-translational modifications, thereby enabling the identification of O-GlcNAc modified peptides and direct localization of the PTM site [86, 128-130].

Owing to the low abundance of the O-GlcNAc modification, enrichment of the modified proteins or peptides is an indispensable prerequisite for successful identification by mass spectrometry analysis. O-GlcNAc proteins have been enriched using immunaffinity approaches by monoclonal anti-O-GlcNAc peptide antibodies in combination with shotgun proteomics, which recently enabled the identification of 200 mammalian O-GlcNAc proteins [131]. A similar level of success has been achieved following a tagging-via-substrate approach, which utilizes metabolic labelling of O-GlcNAc proteins with an azide-analogue of the sugar (GlcNAz). The GlcNAz moiety allows the conjugation of proteins to biotin (*via* Staudinger ligation using a biotinylated phosphine), which in turn enables the selective enrichment of O-GlcNAc proteins using avidin/strepatavidin beads. This approach has allowed for the identification of 199 O-GlcNAc proteins, of which 23 were confirmed using reciprocal immunoprecipitation [132]. More recently, Zaro *et al.* explored the possibility of metabolic labeling of O-GlcNAc proteins with alkyne-tagged GlcNAc (GlcNAlk) followed by enrichment based on Cu(I)-catalyzed [3 + 2] azide-alkyne cycloaddition (CuAAC) [133]. The azide/alkyne configuration using an azide-based biotin affinity tag instead if an alkyne probe considerably reduced the copper-mediated protein background often seen during CuAAC [134] and, finally, enabled the identification of 374 candidate O-GlcNAc proteins, 279 of which were not previously reported.

The aforementioned protein-level enrichment approaches are only an indirect means towards the identification of O-GlcNAc proteins and did not yet enable the identification of a single O-GlcNAc site. In contrast, peptide-level enrichment approaches along with ETD mass spectrometry have been shown to be more successful. Notably, O-GlcNAc peptides have been enriched using lectin-affinity chromatography with succinylated wheat germ agglutinin (WGA) allowing for the direct identification of numerous O-GlcNAc sites. This approach, termed lectin weak affinity chromatography (LWAC) [125], recently enabled the identification of 58 O-GlcNAc sites from a mouse post-synaptic density preparation [129] and 142 O-GlcNAc sites on 62 nuclear proteins from mouse embryonic stem cells [135]. An improved version of the LWAC approach, which makes use of WGA immobilized on silica beads (POROS®-immobilized WGA) [51] has been reported by Trinidad and co-workers. WGA immobilized on silica beads shows superior chromatographic performance (smaller bead size, better pressure resistance) over the traditional WGA-agarose columns and allowed them to identify around 1750 O-GlcNAc sites from mouse brain synaptosomes. An interesting alternative for the enrichment of native O-GlcNAc peptides are O-GlcNAc specific monoclonal IgG antibodies, which allowed the identification of 83 O-GlcNAc sites from HEK293T cells [126]. Alternatively, O-GlcNAc peptides can be affinity tagged following a chemoenzymatic approach. Here, an engineered β-1,4-galactosyltransferase attaches a galactose moiety to the GlcNAc and the enzyme accepts various affinity-tagged UDP-galactose analogues as substrates [136, 137]. This chemo-enzymatic approach using a keton-biotin tag led to the identification of numerous O-GlcNAc proteins and their sites from mouse brain tissue [128, 138, 139]. O-GlcNAc peptides can also be labeled chemoenzymatically with an azide-bearing galactose analogue (UDP-GalNAz), which is further conjugated to a photocleavable alkyne-biotin reagent using CuAAC, and released from the avidin

affinity support by photochemical cleavage [140]. This method has been termed chemical/enzymatic photochemical cleavage (CEPC) method. After cleavage, the tagged O-GlcNAc peptides are modified with a basic aminomethyltriazolylacetylgalactosamine (AMT-GalNAc) moiety that facilitates identification and site localization by ETD mass spectrometry. The CEPC approach has enabled the identification of 141 O-GlcNAc sites from less than 15 µg of a spindle and midbody preparation of HeLa cells [130], and, more recently, 458 O-GlcNAc sites in 195 proteins from mouse cerebrocortical brain tissue [141].

Still, despite all enrichment efforts reported to date, the analytical difficulties associated with detecting the O-GlcNAc modification leaves a remarkable discrepancy between the number of reported O-GlcNAc proteins and the number of proteins, for which an O-GlcNAc site has been unambiguously identified. The database dbOGAP [142] which is the most comprehensive and best curated collection of O-GlcNAc proteins and sites currently lists approximately 800 O-GlcNAc protein entries. However, O-GlcNAc site evidence exists only for 172 of these proteins. Hence, direct O-GlcNAc modification and site localization evidence is lacking for almost 80% of all reported putative O-GlcNAc modified proteins. This very large gap needs to be closed by new developments in biochemical and analytical methodology.

Objective and outline of this thesis

O-GlcNAc is an emerging, intracellular PTM with broad biological implications. But the systematic or large-scale analysis of this PTM is hampered by several factors including low stoichiometry and the lability of the O-glycosidic bond during tandem mass spectrometry. Similarly, N-linked glycosylation is involved in a wide variety of physiological and pathophysiological processes. The development of novel biochemical and analytical approaches to study N-linked and O-GlcNAc glycosylation in a quantitative and system-wide fashion is prerequisite for our understanding of important cellular processes and will bear novel insights into the glycobiology of health and disease. The prime objective of this thesis was to advance current proteomic and glycomic technologies for the system-wide analysis of O-GlcNAc proteins and N-linked glycosylation and demonstrate their utility using relevant biological models.

A novel approach for the mass spectrometry-based detection of O-GlcNAc peptides is presented in Chapter 2. Using a synthetic O-GlcNAc peptide library comprising of 72 members with precisely known modification sites, nine different tandem MS acquisition schemes were assessed for their ability to identify O-GlcNAc peptides and to localize their PTM sites. Based on these results, a scoring scheme was developed, termed Oscore, which automatically assesses tandem MS spectra for the presence and intensity of characteristic O-GlcNAc fragment ion patterns. The Oscore discriminates O-GlcNAc peptide spectra from spectra of unmodified peptides with excellent sensitivity and specificity, and the practical utility of the Oscore for the detection and identification of O-GlcNAc peptides using a combination of PQD and ETD fragmentation is illustrated.

The excellent discriminating power of the Oscore in combination with high mass accuracy/high resolution mass spectrometry was exploited in Chapter 3. Data from three recent large-scale proteomic studies of cancer and stem cell lines [5, 143, 144] were re-analyzed using the Oscore algorithm in combination with standard database searching and enabled the identification of hundreds of O-GlcNAc peptides and proteins not identified in the original studies. These results

allowed estimating the relative abundance of O-GlcNAc- and phosphoproteins and afforded further insights into the interplay of O-GlcNAc and phosphorylation.

Chapter 4 describes the discovery of proteins modified by phosphorylated O-GlcNAc. Employing the combination of Oscore and standard Mascot database search led to the identification of numerous proteins modified with a phosphorylated O-GlcNAc moiety. The systematic analysis of tandem mass spectra from phosphorylated O-GlcNAc-modified peptides revealed major fragmentation pathways of this PTM and enabled its structural dissection. Further analyses underpin the finding that this modification is indeed O-GlcNAc-6-phosphate and identified the first human O-GlcNAc-6-phosphate modified protein.

Chapter 5 describes how metabolic labeling of O-GlcNAc proteins with azide-tagged GlcNAc can be exploited in combination with CuAAC and β-elimination for the large-scale identification of O-GlcNAc proteins and their sites. Further insight into the nutrient responsive O-GlcNAc signaling has been obtained by globally analyzing the response of the O-GlcNAc proteome to the inhibition of the nuclear/cytoplasmic O-GlcNAcase, thereby identifying key signaling proteins as O-GlcNAc modified and responsive to O-GlcNAcase inhibition.

A novel approach for the quantification of N-linked glycans based on stable isotope labeled carbonyl-reactive tandem mass tags (glycoTMTs) is described in Chapter 6. The direct comparison of isobaric quantification and quantification of heavy/light TMT-labeled glycans using MALDI-TOF mass spectrometry illustrates advantages and disadvantages of either approach. The quantitative comparison of N-linked glycosylation profiles of primary and metastatic cancer cell lines revealed significant down-regulation of high-mannose glycans in the metastatic tumor cell line, indicating glycoTMT as a valuable tool for the analysis of neutral and acidic N-linked glycans from highly complex samples such as cancer cell lines.

Abbreviations

AD	Alzheimer's disease
AMT-GalNAc	amino-methyl-triazolyl-acetyl-galactosamine
CDG	congenital disorders of glycosylation
CEPC	chemical/enzymatic photochemical cleavage method
CID	collision-induced dissociation
CRM	charge residue model
CuAAC	Cu(I)-catalyzed [3 + 2] azide-alkyne cyclo-addition
DC	direct current
ECD	electron capture dissociation
ER	endoplasmatic reticulum
ESI	electro-spray ionization
ETD	electron transfer dissociation
FT-ICR	Fourier transform ion cyclotron mass spectrometer
GFPT	glutamine-fructose 6-phosphate transaminase
GlcNAc-6-P	glucosamine-6-phosphate
GlcNAlk	alkyne-tagged GlcNAc
GlcNAz	azide-tagged GlcNAc
glycoTMT	carbonyl-reactive tandem mass tag
GPI	glycosyl-phosphatidyl-inositol
HBP	hexosamine biosynthetic pathway
HILIC	hydrophilic interaction chromatography
IEF	isoelectric focusing
IEM	ion emission model
LacNAc	Galβ1 – 4GlcNAc
LC	liquid chromatography
LC-MS/MS	liquid chromatography coupled to tandem mass spectrometry
LWAC	lectin weak affinity chromato-graphy
MALDI	matrix-assisted laser desorption/ionization
MS	mass spectrometry
MSA	multistage activation
MSn	multiple stage mass spectrometry
OGA	O-GlcNAcase
O-GlcNAc	O-linked β-N-acetylglucosamine
OGT	O-GlcNAc transferase
PGC	porous graphitized carbon chromatography

PMF	peptide mass fingerprint
PQD	pulsed Q dissociation
PSD	post-source decay
PTM	post-translational modification
QQQ	triple quadrupole mass spectrometer
QTOF	quadrupole time-of-flight mass spectrometer
RF	radio frequency
SAX	strong anion exchange chromatography
SCX	strong cation exchange chromatography
TMT	tandem mass tag
TOF	time-of-flight
TPR	tetratricopeptide repeats
WGA	wheat germ agglutinin

References

1. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945.
2. Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., and Humphery-Smith, I. (1995) Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* 16, 1090-1094.
3. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 250, 4007-4021.
4. Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* 26, 231-243.
5. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2012) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7, 548.
6. 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.
7. Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., Goswami, T., Rad, R., Beausoleil, S. A., Villen, J., Haas, W., Sowa, M. E., and Gygi, S. P. (2010) A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* 143, 1174-1189.
8. Wagner, S. A., Beli, P., Weinert, B. T., Scholz, C., Kelstrup, C. D., Young, C., Nielsen, M. L., Olsen, J. V., Brakebusch, C., and Choudhary, C. (2012) Proteomic analyses reveal divergent ubiquitylation site patterns in murine tissues. *Mol Cell Proteomics* [epub ahead of print 2012/07/14].
9. Apweiler, R., Hermjakob, H., and Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1473, 4-8.
10. Van den Steen, P., Rudd, P. M., Dwek, R. A., and Opdenakker, G. (1998) Concepts and principles of O-linked glycosylation. *Crit Rev Biochem Mol Biol* 33, 151-208.
11. Varki, A., and Sharon, N. (2009) Historical Background and Overview. In: Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E., eds. *Essentials of Glycobiology*, Second Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
12. Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E., eds. (2009) *Essentials of Glycobiology*, Second Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
13. Gavel, Y., and von Heijne, G. (1990) Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng* 3, 433-442.
14. Miletich, J. P., and Broze, G. J., Jr. (1990) Beta protein C is not glycosylated at asparagine 329. The rate of translation may influence the frequency of usage at asparagine-X-cysteine sites. *J Biol Chem* 265, 11397-11404.
15. Stanley, P., Schachter, H., and Taniguchi, N. (2009) N-Glycans. In: Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E., eds. *Essentials of Glycobiology*, Second Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
16. Rademacher, T. W., Parekh, R. B., Dwek, R. A., Isenberg, D., Rook, G., Axford, J. S., and Roitt, I. (1988) The role of IgG glycoforms in the pathogenesis of rheumatoid arthritis. *Springer Semin Immunopathol* 10, 231-249.
17. Rudd, P. M., and Dwek, R. A. (1997) Glycosylation: heterogeneity and the 3D structure of proteins. *Crit Rev Biochem Mol Biol* 32, 1-100.
18. Jones, J., Krag, S. S., and Betenbaugh, M. J. (2005) Controlling N-linked glycan site occupancy. *Biochim Biophys Acta* 1726, 121-137.

19. Dennis, J. W., Nabi, I. R., and Demetriou, M. (2009) Metabolism, cell surface organization, and disease. *Cell* 139, 1229-1241.
20. Culyba, E. K., Price, J. L., Hanson, S. R., Dhar, A., Wong, C. H., Gruebele, M., Powers, E. T., and Kelly, J. W. (2011) Protein native-state stabilization by placing aromatic side chains in N-glycosylated reverse turns. *Science* 331, 571-575.
21. Hanson, S. R., Culyba, E. K., Hsu, T. L., Wong, C. H., Kelly, J. W., and Powers, E. T. (2009) The core trisaccharide of an N-linked glycoprotein intrinsically accelerates folding and enhances stability. *Proc Natl Acad Sci U S A* 106, 3131-3136.
22. Shental-Bechor, D., and Levy, Y. (2008) Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc Natl Acad Sci U S A* 105, 8256-8261.
23. Aebi, M., Bernasconi, R., Clerc, S., and Molinari, M. (2010) N-glycan structures: recognition and processing in the ER. *Trends Biochem Sci* 35, 74-82.
24. Dennis, J. W., Lau, K. S., Demetriou, M., and Nabi, I. R. (2009) Adaptive regulation at the cell surface by N-glycosylation. *Traffic* 10, 1569-1578.
25. Wyss, D. F., and Wagner, G. (1996) The structural role of sugars in glycoproteins. *Curr Opin Biotechnol* 7, 409-416.
26. Drickamer, K. (1991) Clearing up glycoprotein hormones. *Cell* 67, 1029-1032.
27. Lowe, J. B., Stoolman, L. M., Nair, R. P., Larsen, R. D., Berhend, T. L., and Marks, R. M. (1990) ELAM-1--dependent cell adhesion to vascular endothelium determined by a transfected human fucosyltransferase cDNA. *Cell* 63, 475-484.
28. Walz, G., Aruffo, A., Kolanus, W., Bevilacqua, M., and Seed, B. (1990) Recognition by ELAM-1 of the sialyl-Lex determinant on myeloid and tumor cells. *Science* 250, 1132-1135.
29. Phillips, M. L., Nudelman, E., Gaeta, F. C., Perez, M., Singhal, A. K., Hakomori, S., and Paulson, J. C. (1990) ELAM-1 mediates cell adhesion by recognition of a carbohydrate ligand, sialyl-Lex. *Science* 250, 1130-1132.
30. Stoolman, L. M., and Rosen, S. D. (1983) Possible role for cell-surface carbohydrate-binding molecules in lymphocyte recirculation. *J Cell Biol* 96, 722-729.
31. Landmesser, L., Dahm, L., Tang, J. C., and Rutishauser, U. (1990) Polysialic acid as a regulator of intramuscular nerve branching during embryonic development. *Neuron* 4, 655-667.
32. Freeze, H. H. (2007) Congenital Disorders of Glycosylation: CDG-I, CDG-II, and beyond. *Curr Mol Med* 7, 389-396.
33. Dube, D. H., and Bertozzi, C. R. (2005) Glycans in cancer and inflammation--potential for therapeutics and diagnostics. *Nat Rev Drug Discov* 4, 477-488.
34. Meezan, E., Wu, H. C., Black, P. H., and Robbins, P. W. (1969) Comparative studies on the carbohydrate-containing membrane components of normal and virus-transformed mouse fibroblasts. II. Separation of glycoproteins and glycopeptides by sephadex chromatography. *Biochemistry* 8, 2518-2524.
35. Taniguchi, N. (2008) Human disease glycomics/proteome initiative (HGPI). *Mol Cell Proteomics* 7, 626-627.
36. Packer, N. H., von der Lieth, C. W., Aoki-Kinoshita, K. F., Lebrilla, C. B., Paulson, J. C., Raman, R., Rudd, P., Sasisekharan, R., Taniguchi, N., and York, W. S. (2008) Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics* 8, 8-20.
37. Torres, C. R., and Hart, G. W. (1984) Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *J Biol Chem* 259, 3308-3317.
38. Jinek, M., Rehwinkel, J., Lazarus, B. D., Izaurralde, E., Hanover, J. A., and Conti, E. (2004) The superhelical TPR-repeat domain of O-linked GlcNAc transferase exhibits structural similarities to importin alpha. *Nat Struct Mol Biol* 11, 1001-1007.

39. Toleman, C., Paterson, A. J., Whisenhunt, T. R., and Kudlow, J. E. (2004) Characterization of the histone acetyltransferase (HAT) domain of a bifunctional protein with activable O-GlcNAcase and HAT activities. *J Biol Chem* 279, 53665-53673.
40. Wells, L., Gao, Y., Mahoney, J. A., Vosseller, K., Chen, C., Rosen, A., and Hart, G. W. (2002) Dynamic O-glycosylation of nuclear and cytosolic proteins: further characterization of the nucleocytoplasmic beta-N-acetylglucosaminidase, O-GlcNAcase. *J Biol Chem* 277, 1755-1761.
41. Marshall, S., Bacote, V., and Traxinger, R. R. (1991) Discovery of a metabolic pathway mediating glucose-induced desensitization of the glucose transport system. Role of hexosamine biosynthesis in the induction of insulin resistance. *J Biol Chem* 266, 4706-4712.
42. Broschat, K. O., Gorka, C., Page, J. D., Martin-Berger, C. L., Davies, M. S., Huang Hc, H. C., Gulve, E. A., Salsgiver, W. J., and Kasten, T. P. (2002) Kinetic characterization of human glutamine-fructose-6-phosphate amidotransferase I: potent feedback inhibition by glucosamine 6-phosphate. *J Biol Chem* 277, 14764-14770.
43. Slawson, C., Copeland, R. J., and Hart, G. W. (2010) O-GlcNAc signaling: a metabolic link between diabetes and cancer? *Trends Biochem Sci* 35, 547-555.
44. Haltiwanger, R. S., Blomberg, M. A., and Hart, G. W. (1992) Glycosylation of nuclear and cytoplasmic proteins. Purification and characterization of a uridine diphospho-N-acetylglucosamine:polypeptide beta-N-acetylglucosaminyltransferase. *J Biol Chem* 267, 9005-9013.
45. Kreppel, L. K., and Hart, G. W. (1999) Regulation of a cytosolic and nuclear O-GlcNAc transferase. Role of the tetratricopeptide repeats. *J Biol Chem* 274, 32015-32022.
46. Love, D. C., and Hanover, J. A. (2005) The hexosamine signaling pathway: deciphering the "O-GlcNAc code". *Sci STKE* 2005, re13.
47. Hart, G. W., Slawson, C., Ramirez-Correa, G., and Lagerlof, O. (2011) Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annu Rev Biochem* 80, 825-858.
48. Hart, G. W., Housley, M. P., and Slawson, C. (2007) Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* 446, 1017-1022.
49. Hu, P., Shimoji, S., and Hart, G. W. (2010) Site-specific interplay between O-GlcNAcylation and phosphorylation in cellular regulation. *FEBS Lett* 584, 2526-2538.
50. Hart, G. W., Greis, K. D., Dong, L. Y., Blomberg, M. A., Chou, T. Y., Jiang, M. S., Roquemore, E. P., Snow, D. M., Kreppel, L. K., Cole, R. N., and et al. (1995) O-linked N-acetylglucosamine: the "yin-yang" of Ser/Thr phosphorylation? Nuclear and cytoplasmic glycosylation. *Adv Exp Med Biol* 376, 115-123.
51. Trinidad, J. C., Barkan, D. T., Gulledge, B. F., Thalhammer, A., Sali, A., Schoepfer, R., and Burlingame, A. L. (2012) Global identification and characterization of both O-GlcNAcylation and phosphorylation at the murine synapse. *Mol Cell Proteomics* [epub ahead of print 2012/05/31].
52. Chou, T. Y., Hart, G. W., and Dang, C. V. (1995) c-Myc is glycosylated at threonine 58, a known phosphorylation site and a mutational hot spot in lymphomas. *J Biol Chem* 270, 18961-18965.
53. Yang, W. H., Kim, J. E., Nam, H. W., Ju, J. W., Kim, H. S., Kim, Y. S., and Cho, J. W. (2006) Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. *Nat Cell Biol* 8, 1074-1083.
54. Cheng, X., and Hart, G. W. (2001) Alternative O-glycosylation/O-phosphorylation of serine-16 in murine estrogen receptor beta: post-translational regulation of turnover and transactivation activity. *J Biol Chem* 276, 10570-10575.
55. Comer, F. I., and Hart, G. W. (2001) Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II. *Biochemistry* 40, 7845-7852.
56. Liu, F., Iqbal, K., Grundke-Iqbal, I., Hart, G. W., and Gong, C. X. (2004) O-GlcNAcylation regulates phosphorylation of tau: a mechanism involved in Alzheimer's disease. *Proc Natl Acad Sci U S A* 101, 10804-10809.

57. Yuzwa, S. A., Shan, X., Macauley, M. S., Clark, T., Skorobogatko, Y., Vosseller, K., and Vocadlo, D. J. (2012) Increasing O-GlcNAc slows neurodegeneration and stabilizes tau against aggregation. *Nat Chem Biol* 8, 393-399.
58. Slawson, C., and Hart, G. W. (2011) O-GlcNAc signalling: implications for cancer cell biology. *Nat Rev Cancer* 11, 678-684.
59. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198-207.
60. Zaia, J. (2010) Mass spectrometry and glycomics. *OMICS* 14, 401-418.
61. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 5, 699-711.
62. Karas, M., and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 60, 2299-2301.
63. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988) Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* 2, 151-153.
64. Knochenmuss, R. (2006) Ion formation mechanisms in UV-MALDI. *Analyst* 131, 966-986.
65. Jaskolla, T. W., and Karas, M. (2011) Compelling evidence for Lucky Survivor and gas phase protonation: the unified MALDI analyte protonation mechanism. *J Am Soc Mass Spectrom* 22, 976-988.
66. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246, 64-71.
67. Wilm, M., and Mann, M. (1994) Electrospray and taylor-cone theory, Dole's beam of macromolecules at last? *Int J Mass Spectrom Ion Process* [epub ahead of print, 167-180.
68. Iribarne, J. V., and Thompson, B. A. (1976) On the evaporation of small ions from charged droplets. *J Chem Phys* 64, 2287-2294.
69. Wilm, M., and Mann, M. (1996) Analytical properties of the nanoelectrospray ion source. *Anal Chem* 68, 1-8.
70. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* 379, 466-469.
71. Whittal, R. M., and Li, L. (1995) High-resolution matrix-assisted laser desorption/ionization in a linear time-of-flight mass spectrometer. *Anal Chem* 67, 1950-1954.
72. Jonscher, K. R., and Yates, J. R., 3rd (1997) The quadrupole ion trap mass spectrometer--a small solution to a big challenge. *Anal Biochem* 244, 1-15.
73. March, R. E. (2009) Quadrupole ion traps. *Mass Spectrom Rev* 28, 961-989.
74. Schwartz, J. C., Senko, M. W., and Syka, J. E. (2002) A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* 13, 659-669.
75. Douglas, D. J., Frank, A. J., and Mao, D. (2005) Linear ion traps in mass spectrometry. *Mass Spectrom Rev* 24, 1-29.
76. Scigelova, M., and Makarov, A. (2006) Orbitrap mass analyzer--overview and applications in proteomics. *Proteomics* 6 Suppl 2, 16-21.
77. Denisov, E., Damoc, E., Lange, O., and Makarov, A. (2012) Orbitrap mass spectrometry with resolving powers above 1,000,000. *International Journal of Mass Spectrometry* [epub ahead of print 2012/06/21].
78. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Mueller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J. J., Cox, J., Horning, S., Mann, M., and Makarov, A. (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol Cell Proteomics* 11, O111 013698.
79. Roepstorff, P., and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* 11, 601.

80. Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T., and Watson, J. T. (1987) Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal Chem* 59, 2621-2625.
81. Schwartz, J. C., Syka, J. E., and Quarmby, S. T. (2005) Improving the Fundamentals of MSn on 2D Ion Traps: New Ion Activation and Isolation Techniques. *53rd ASMS Conference on Mass Spectrommetry*, San Antonio, TX, USA.
82. Cunningham, C., Jr., Glish, G. L., and Burinsky, D. J. (2006) High amplitude short time excitation: a method to form and detect low mass product ions in a quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* 17, 81-84.
83. Domon, B., and Costello, C. E. (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal* 5, 397-409.
84. McAlister, G. C., Phanstiel, D., Good, D. M., Berggren, W. T., and Coon, J. J. (2007) Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Anal Chem* 79, 3525-3534.
85. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J Am Chem Soc* 120, 3265-3266.
86. Mirgorodskaya, E., Roepstorff, P., and Zubarev, R. A. (1999) Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal Chem* 71, 4431-4436.
87. Swaney, D. L., McAlister, G. C., Wirtala, M., Schwartz, J. C., Syka, J. E., and Coon, J. J. (2007) Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal Chem* 79, 477-485.
88. Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F., and Coon, J. J. (2004) A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem* 76, 3590-3598.
89. Ting, L., Rad, R., Gygi, S. P., and Haas, W. (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* 8, 937-940.
90. Swaney, D. L., McAlister, G. C., and Coon, J. J. (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat Methods* 5, 959-964.
91. James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993) Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun* 195, 58-64.
92. Mann, M., Hojrup, P., and Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* 22, 338-345.
93. Pappin, D. J., Hojrup, P., and Bleasby, A. J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 3, 327-332.
94. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 4, 787-797.
95. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 4, 1419-1440.
96. Ashline, D., Singh, S., Hanneman, A., and Reinhold, V. (2005) Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MSn. *Anal Chem* 77, 6250-6262.
97. Zaia, J. (2004) Mass spectrometry of oligosaccharides. *Mass Spectrom Rev* 23, 161-227.
98. Marino, K., Bones, J., Kattla, J. J., and Rudd, P. M. (2010) A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol* 6, 713-723.
99. Wada, Y., Azadi, P., Costello, C. E., Dell, A., Dwek, R. A., Geyer, H., Geyer, R., Kakehi, K., Karlsson, N. G., Kato, K., Kawasaki, N., Khoo, K. H., Kim, S., Kondo, A., Lattova, E., Mechref, Y., Miyoshi, E., Nakamura, K., Narimatsu, H., Novotny, M. V., Packer, N. H., Perreault, H., Peter-Katalinic, J., Pohlentz, G., Reinhold, V. N., Rudd, P. M., Suzuki, A., and Taniguchi, N. (2007) Comparison of the methods for profiling glycoprotein glycans--HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study. *Glycobiology* 17, 411-422.

100. Pan, S., Chen, R., Aebersold, R., and Brentnall, T. A. (2011) Mass spectrometry based glycoproteomics--from a proteomics perspective. *Mol Cell Proteomics* 10, R110 003251.
101. Seipert, R. R., Dodds, E. D., Clowers, B. H., Beecroft, S. M., German, J. B., and Lebrilla, C. B. (2008) Factors that influence fragmentation behavior of N-linked glycopeptide ions. *Anal Chem* 80, 3684-3692.
102. Ruhaak, L. R., Zauner, G., Huhn, C., Bruggink, C., Deelder, A. M., and Wuhrer, M. (2010) Glycan labeling strategies and their use in identification and quantification. *Anal Bioanal Chem* 397, 3457-3481.
103. Alvarez-Manilla, G., Warren, N. L., Abney, T., Atwood, J., 3rd, Azadi, P., York, W. S., Pierce, M., and Orlando, R. (2007) Tools for glycomics: relative quantitation of glycans by isotopic permethylation using 13CH3I. *Glycobiology* 17, 677-687.
104. Kang, P., Mechref, Y., Kyselova, Z., Goetz, J. A., and Novotny, M. V. (2007) Comparative glycomic mapping through quantitative permethylation and stable-isotope labeling. *Anal Chem* 79, 6064-6073.
105. Atwood, J. A., 3rd, Cheng, L., Alvarez-Manilla, G., Warren, N. L., York, W. S., and Orlando, R. (2008) Quantitation by isobaric labeling: applications to glycomics. *J Proteome Res* 7, 367-374.
106. Yuan, J., Hashii, N., Kawasaki, N., Itoh, S., Kawanishi, T., and Hayakawa, T. (2005) Isotope tag method for quantitative analysis of carbohydrates by liquid chromatography-mass spectrometry. *J Chromatogr A* 1067, 145-152.
107. Bowman, M. J., and Zaia, J. (2007) Tags for the stable isotopic labeling of carbohydrates and quantitative analysis by mass spectrometry. *Anal Chem* 79, 5777-5784.
108. Bowman, M. J., and Zaia, J. (2010) Comparative glycomics using a tetraplex stable-isotope coded tag. *Anal Chem* 82, 3023-3031.
109. Xia, B., Feasley, C. L., Sachdev, G. P., Smith, D. F., and Cummings, R. D. (2009) Glycan reductive isotope labeling for quantitative glycomics. *Anal Biochem* 387, 162-170.
110. Walker, S. H., Budhathoki-Uprety, J., Novak, B. M., and Muddiman, D. C. (2011) Stable-isotope labeled hydrophobic hydrazide reagents for the relative quantification of N-linked glycans by electrospray ionization mass spectrometry. *Anal Chem* 83, 6738-6745.
111. Uematsu, R., Furukawa, J., Nakagawa, H., Shinohara, Y., Deguchi, K., Monde, K., and Nishimura, S. (2005) High throughput quantitative glycomics and glycoform-focused proteomics of murine dermis and epidermis. *Mol Cell Proteomics* 4, 1977-1989.
112. Natsuka, S., and Hase, S. (1998) Analysis of N- and O-glycans by pyridylamination. *Methods Mol Biol* 76, 101-113.
113. Guile, G. R., Rudd, P. M., Wing, D. R., Prime, S. B., and Dwek, R. A. (1996) A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles. *Anal Biochem* 240, 210-226.
114. Rudd, P. M., Guile, G. R., Kuster, B., Harvey, D. J., Opdenakker, G., and Dwek, R. A. (1997) Oligosaccharide sequencing technology. *Nature* 388, 205-207.
115. Balog, C. I., Stavenhagen, K., Fung, W. L., Koeleman, C. A., McDonnell, L. M., Verhoeven, A., Mesker, W. E., Tollenaar, R. A., Deelder, A. M., and Wuhrer, M. (2012) N-glycosylation of colorectal cancer tissues: a liquid chromatography and mass spectrometry-based investigation. *Mol Cell Proteomics* [epub ahead of print 2012/05/11].
116. Harvey, D. J. (1999) Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrom Rev* 18, 349-450.
117. Stumpo, K. A., and Reinhold, V. N. (2010) The N-glycome of human plasma. *J Proteome Res* 9, 4823-4830.
118. Powell, A. K., and Harvey, D. J. (1996) Stabilization of sialic acids in N-linked oligosaccharides and gangliosides for analysis by positive ion matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* 10, 1027-1032.

119. Alley, W. R., Jr., Vasseur, J. A., Goetz, J. A., Svoboda, M., Mann, B. F., Matei, D. E., Menning, N., Hussein, A., Mechref, Y., and Novotny, M. V. (2012) N-linked glycan structures and their expressions change in the blood sera of ovarian cancer patients. *J Proteome Res* 11, 2282-2300.
120. Nakano, M., Saldanha, R., Gobel, A., Kavallaris, M., and Packer, N. H. (2011) Identification of glycan structure alterations on cell membrane proteins in desoxyepothilone B resistant leukemia cells. *Mol Cell Proteomics* 10, M111 009001.
121. Haynes, P. A., and Aebersold, R. (2000) Simultaneous detection and identification of O-GlcNAc-modified glycoproteins using liquid chromatography-tandem mass spectrometry. *Anal Chem* 72, 5402-5410.
122. Huddleston, M. J., Bean, M. F., and Carr, S. A. (1993) Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests. *Anal Chem* 65, 877-884.
123. Chalkley, R. J., and Burlingame, A. L. (2001) Identification of GlcNAcylation sites of peptides and alpha-crystallin using Q-TOF mass spectrometry. *J Am Soc Mass Spectrom* 12, 1106-1113.
124. Chalkley, R. J., and Burlingame, A. L. (2003) Identification of novel sites of O-N-acetylglucosamine modification of serum response factor using quadrupole time-of-flight mass spectrometry. *Mol Cell Proteomics* 2, 182-190.
125. Vosseller, K., Trinidad, J. C., Chalkley, R. J., Specht, C. G., Thalhammer, A., Lynn, A. J., Snedecor, J. O., Guan, S., Medzihradszky, K. F., Maltby, D. A., Schoepfer, R., and Burlingame, A. L. (2006) O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. *Mol Cell Proteomics* 5, 923-934.
126. Zhao, P., Viner, R., Teo, C. F., Boons, G. J., Horn, D., and Wells, L. (2011) Combining high-energy C-trap dissociation and electron transfer dissociation for protein O-GlcNAc modification site assignment. *J Proteome Res* 10, 4088-4104.
127. Carapito, C., Klemm, C., Aebersold, R., and Domon, B. (2009) Systematic LC-MS analysis of labile post-translational modifications in complex mixtures. *J Proteome Res* 8, 2608-2614.
128. Khidekel, N., Ficarro, S. B., Clark, P. M., Bryan, M. C., Swaney, D. L., Rexach, J. E., Sun, Y. E., Coon, J. J., Peters, E. C., and Hsieh-Wilson, L. C. (2007) Probing the dynamics of O-GlcNAc glycosylation in the brain using quantitative proteomics. *Nat Chem Biol* 3, 339-348.
129. Chalkley, R. J., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc Natl Acad Sci USA* 106, 8894-8899.
130. Wang, Z., Udeshi, N. D., Slawson, C., Compton, P. D., Sakabe, K., Cheung, W. D., Shabanowitz, J., Hunt, D. F., and Hart, G. W. (2010) Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Sci Signal* 3, ra2.
131. Teo, C. F., Ingale, S., Wolfert, M. A., Elsayed, G. A., Not, L. G., Chatham, J. C., Wells, L., and Boons, G. J. (2010) Glycopeptide-specific monoclonal antibodies suggest new roles for O-GlcNAc. *Nat Chem Biol* 6, 338-343.
132. Nandi, A., Sprung, R., Barma, D. K., Zhao, Y., Kim, S. C., and Falck, J. R. (2006) Global identification of O-GlcNAc-modified proteins. *Anal Chem* 78, 452-458.
133. Zaro, B. W., Yang, Y. Y., Hang, H. C., and Pratt, M. R. (2011) Chemical reporters for fluorescent detection and identification of O-GlcNAc-modified proteins reveal glycosylation of the ubiquitin ligase NEDD4-1. *Proc Natl Acad Sci U S A* 108, 8146-8151.
134. Speers, A. E., and Cravatt, B. F. (2004) Profiling enzyme activities in vivo using click chemistry methods. *Chem Biol* 11, 535-546.
135. Myers, S. A., Panning, B., and Burlingame, A. L. (2011) Polycomb repressive complex 2 is necessary for the normal site-specific O-GlcNAc distribution in mouse embryonic stem cells. *Proc Natl Acad Sci USA* 108, 9490-9495.
136. Khidekel, N., Arndt, S., Lamarre-Vincent, N., Lippert, A., Poulin-Kerstien, K. G., Ramakrishnan, B., Qasba, P. K., and Hsieh-Wilson, L. C. (2003) A chemoenzymatic approach toward the rapid and

sensitive detection of O-GlcNAc posttranslational modifications. *J Am Chem Soc* 125, 16162-16163.
137. Rexach, J. E., Clark, P. M., and Hsieh-Wilson, L. C. (2008) Chemical approaches to understanding O-GlcNAc glycosylation in the brain. *Nat Chem Biol* 4, 97-106.
138. Khidekel, N., Ficarro, S. B., Peters, E. C., and Hsieh-Wilson, L. C. (2004) Exploring the O-GlcNAc proteome: direct identification of O-GlcNAc-modified proteins from the brain. *Proc Natl Acad Sci U S A* 101, 13132-13137.
139. Tai, H. C., Khidekel, N., Ficarro, S. B., Peters, E. C., and Hsieh-Wilson, L. C. (2004) Parallel identification of O-GlcNAc-modified proteins from cell lysates. *J Am Chem Soc* 126, 10500-10501.
140. Wang, Z., Udeshi, N. D., O'Malley, M., Shabanowitz, J., Hunt, D. F., and Hart, G. W. (2010) Enrichment and site mapping of O-linked N-acetylglucosamine by a combination of chemical/enzymatic tagging, photochemical cleavage, and electron transfer dissociation mass spectrometry. *Mol Cell Proteomics* 9, 153-160.
141. Alfaro, J. F., Gong, C. X., Monroe, M. E., Aldrich, J. T., Clauss, T. R., Purvine, S. O., Wang, Z., Camp, D. G., 2nd, Shabanowitz, J., Stanley, P., Hart, G. W., Hunt, D. F., Yang, F., and Smith, R. D. (2012) Tandem mass spectrometry identifies many mouse brain O-GlcNAcylated proteins including EGF domain-specific O-GlcNAc transferase targets. *Proc Natl Acad Sci U S A* 109, 7280-7285.
142. Wang, J., Torii, M., Liu, H., Hart, G. W., and Hu, Z. Z. (2011) dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics* 12, 91.
143. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* 11, M111 014050.
144. Phanstiel, D. H., Brumbaugh, J., Wenger, C. D., Tian, S., Probasco, M. D., Bailey, D. J., Swaney, D. L., Tervo, M. A., Bolin, J. M., Ruotti, V., Stewart, R., Thomson, J. A., and Coon, J. J. (2011) Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nat Methods* 8, 821-827.

Chapter 2

A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides

Summary

The modification of serine and threonine residues in proteins by a single N-acetylglucosamine (O-GlcNAc) residue is an emerging post-translational modification (PTM) with broad biological implications. However, the systematic or large-scale analysis of this PTM is hampered by several factors including low stoichiometry and the lability of the O-glycosidic bond during tandem mass spectrometry. Using a library of 72 synthetic glycopeptides, a two-stage tandem MS approach was developed consisting of pulsed Q dissociation (PQD) for O-GlcNAc peptide detection and electron transfer dissociation (ETD) for identification and site localization. The developed approach employs a novel scoring scheme (Oscore), which is based on a set of O-GlcNAc specific fragment ions and discriminates O-GlcNAc peptide spectra from spectra of unmodified peptides with 95% sensitivity and >99% specificity. Integrating the Oscore into the two-stage LC-MS/MS approach detected O-GlcNAc peptides in the low fmol range and at ten-fold better sensitivity than a single data-dependent ETD tandem MS experiment.

Introduction

The modification of proteins on serine and threonine residues with β-N-acetylglucosamine (O-GlcNAc) is an emerging and dynamic post-translational modification (PTM) ubiquitously found on metazoan proteins. It was first discovered by Torres and Hart in 1984 [1] and is found on a wide range of cytoplasmic and nuclear proteins [2]. Further, it is known to be associated with several human diseases [3, 4] including neurodegenerative pathologies [3], type II diabetes [3] as well as cancer [4]. Recent technological progress in O-GlcNAc analytics has, by and large, focussed on biochemical enrichment approaches. Notably, this may be achieved by lectin affinity chromatography [5, 6] or using a chemoenzymatic method in which a β-1,4-galactosyltransferase is used to attach a biotinylated galactose to the endogenous O-GlcNAc moiety [7, 8].

Following some form of enrichment, the discovery of O-GlcNAc modified peptides and proteins is greatly aided by tandem mass spectrometry [6, 9] and reports on the discovery of this modification on individual proteins is increasing at a rapid rate. However, despite recent advances in instrumentation, the mass spectrometric analysis of O-GlcNAc peptides is still difficult and mainly hampered by the substoichiometric occupancy of O-GlcNAc sites [10-12] and by the chemical lability of the O-glycosidic bond in the gas phase [13-16]. Under typical collision-induced dissociation (CID) conditions, O-GlcNAc modified peptides readily lose the GlcNAc moiety and spectra are typically dominated by intense neutral loss species as well as the GlcNAc oxonium ion (*m/z* 204.0866) and further fragments thereof [14]. The GlcNAc oxonium ion is isobaric to that of other GlcNAc epimers (e. g. GalNAc) and, therefore, commonly referred to as HexNAc oxonium ions. The intense HexNAc oxonium ion has been known for a long time as a diagnostically useful reporter ion [10, 13, 17] and used for e. g. precursor scanning on triple quadrupole [10] and quadrupole-time-of-flight mass spectrometers [18]. Unfortunately, the reporter ion may only be occasionally observed in ion trap CID spectra because of the poor recovery of fragment ions in the low *m/z* range [19]. This can be overcome by pulsed Q dissociation (PQD) in the ion trap [20] or so-called higher energy collisional dissociation (HCD) in a conventional multipole collision cell [21] on a LTQ Orbitrap XL mass spectrometer. Still, the dominant break of the O-glycosidic bond strongly reduces the occurrence of sequence-informative peptide fragment ions, which in turn impedes peptide identification and O-GlcNAc site localization. Alternative activation methods that enable sequencing the underlying peptide are neutral loss-triggered MS3 (NL-MS3) [5], multistage activation (MSA) [22], electron-capture [23] and electron-transfer dissociation [24] (ECD and ETD, respectively). The latter two preserve labile post-translational modifications, thereby facilitating both the identification of O-GlcNAc modified peptides and localization of the PTM site [5-8, 25]. One published report used a combination of fragmentation methods in which a CID step is followed by ETD fragmentation of the same precursor if the neutral loss of the HexNAc moiety is present in the CID spectrum (NL-ETD) [9]. Similarly, it would be possible to combine CID and HCD (NL-HCD) but this has not yet been published for O-GlcNAc peptides.

In light of the wide range of fragmentation techniques available on a single mass spectrometric platform (i. e. the LTQ Orbitrap XL ETD), it is timely to revisit which fragmentation technique or combination thereof offers particular advantages for the identification and site localization of O-GlcNAc modified peptides. In fact, no systematic study on O-GlcNAc peptides has yet been published using fragmentation techniques available on a hybrid ion trap - Orbitrap instrument. To this end, nine different tandem MS acquisition schemes were evaluated for their ability to identify O-GlcNAc peptides and to localize their PTM sites using a library containing 72 synthetic glycopeptides. As a

result of this comparison, a two-stage approach for the analysis of O-GlcNAc peptides was developed, facilitating the detection of such peptides by PQD at low collision energy and the identification and site localization by ETD. The two-stage approach makes use of on a novel scoring scheme, which is based on a set of O-GlcNAc specific fragment ions and is able to discriminate O-GlcNAc peptide spectra from unmodified ones with 95% sensitivity and >99% specificity. The two-stage approach allows detection and identification of O-GlcNAc peptides at the low fmol level in a complex proteomic background and is ten-fold more sensitive than a typical data-dependent ETD experiment.

Experimental procedures

O-GlcNAc standard peptides and proteins

Peptides were synthesized on a MultiPep peptide synthesizer (Intavis, Germany) using standard N^{α}-Fmoc solid-phase peptide chemistry. A cytosolic protein extract from exponentially growing *E. coli* was spiked with different amounts (1:10, 1:100, 1:500, 1:1000 w/w) of bovine α-crystallin, followed by trypsin digestion and C_{18} purification prior to LC-MS/MS analysis. (For details, see supplemental methods.)

Nano-liquid chromatography – tandem mass spectrometry

Mass spectrometry was performed on an LTQ Orbitrap XL ETD mass spectrometer (Thermo Fisher Scientific, Germany) connected to a nanoLC Ultra 1D+ liquid chromatography system (Eksigent, CA) using in-house packed precolumn (20 mm x 75 μm ReproSil-Pur C18, Dr. Maisch, Germany) and nanocolumn (200 mm x 50 μm ReproSil-Pur C18, Dr. Maisch, Germany). The mass spectrometer was equipped with a nano-electrospray ion source (Proxeon Biosystems, DK) and the electrospray voltage was applied *via* a liquid junction. (For details, see supplemental methods.) All measurements were performed in positive ion mode. Intact peptide mass spectra were acquired at a resolution of 30,000 (at m/z 400) and an automatic gain control (AGC) target value of 106, followed by fragmentation of the most intense ions, a dynamic exclusion of fragmented precursor ions for 20 seconds, exclusion of singly charged ions and ions without assigned charge state for fragmentation (unless otherwise stated) and internal on-the-fly recalibration using the “lock mass” option. Full scans were acquired in profile mode, whereas all tandem mass spectra were acquired in centroid mode. A complete description of all tandem MS experiments employed in this study can be found in Table S1 in the supplemental methods.

Peaklist generation and database search

Peak processing and peak picking of MS data was performed using Mascot Distiller version 2.2.1 (Matrix Science, UK). Briefly, (a) un-centroiding of tandem MS spectra and (b) precursor charge state re-calculation were enabled, (c) tandem MS spectra of singly charged precursors were discarded, (d) the minimum number of peaks per tandem MS spectrum was set to three, and (e) isotope fitting was disabled for the mass range below *m/z* 205. A brief description of the data processing for NL-MS3, NL-HCD and NL-ETD experiments is available in the supplemental methods. Resulting peaklists were searched using the Mascot search engine version 2.2.04 (Matrix Science, UK) against the complete NCBInr database (02/16/2007, 4,626,804 entries) with sequences of synthetic peptides appended. Search parameters included a precursor tolerance of 10 ppm and a fragment tolerance of 0.5 Da for

linear ion trap spectra. HCD spectra were searched with a fragment tolerance of 0.05 Da. Enzyme specificity was set to trypsin, and up to two missed cleavage sites were allowed. Further parameters accounted for the misassignment of the monoisotopic peak (up to the second isotopic peak), for variable modifications by O-GlcNAc (203.0794 Da at serine or threonine), methylation (14.0157 Da at the C-terminus) and in case of the α-crystallin spiked E. coli samples by carbamido-methylation (57.0215 Da at cystein). Except for ETD experiments, the O-GlcNAc modification definition is crucial for the successful Mascot database search of O-GlcNAc peptide spectra. The neutral losses of 203.0794 Da and 221.0899 Da were defined as both, fragment and precursor neutral loss. Moreover, the HexNAc oxonium ion and its fragments (m/z 126.0550, 138.0550, 144.0655, 168.0655, 186.0761, 204.0866) were ignored for Mascot scoring. The database search results were imported into Scaffold version 2.6.02 (Proteome Software, OR).

Scoring MS spectra for the selective extraction of candidate O-GlcNAc precursors

The raw mass spectrometric data of PQD and HCD experiments were processed using the Mascot Distiller software and parameters exactly as described above. For the extraction of potential O-GlcNAc precursors from the resulting mgf file, an in-house written Perl script was utilized. Briefly, the Perl script parses the mgf file and inserts the rank and normalized intensity of every peak in a spectrum. Based on the precursor *m/z* value and the precursor charge state, the script further calculates the expected *m/z* values for the neutral loss of the HexNAc moiety (*Δm* 203.0794 Da) and the loss of the HexNAc oxonium ion (*Δm* 204.0866 Da, and charge *z*-1). For the computation of the Oscore according to (1), the normalized intensities of the reporter ions (*m/z* 126.0550, 138.0550, 144.0655, 168.0655, 186.0761, 204.0866) and sugar loss ions are first divided by their intensity rank and then summed up if they are within a user-specified *m/z* tolerance (e. g. 10 ppm for HCD, 0.3 Da for PQD). The Perl script exports the list of candidate O-GlcNAc precursors along with the Oscore including the accurate precursor mass, charge state, as well as retention time which was further used to assemble an inclusion list for targeted experiments. Precursors with an Oscore better than 2.0 were included in the inclusion list. Probability computations were performed separately using Microsoft Office Excel 2007 (Microsoft Corporation, WA).

Results

Systematic investigation of tandem MS methods

The two-stage LC-MS/MS strategy for the analysis of O-GlcNAc peptides developed in this work is shown in Figure 1. It consists of a discovery LC-MS/MS experiment for the detection of potential O-GlcNAc peptides using low collision energy PQD, the selective extraction of O-GlcNAc candidates from tandem MS spectra based on a novel spectrum scoring scheme and a targeted ETD experiment for O-GlcNAc peptide identification and site localization. The strategy was inspired by results from a systematic evaluation of nine different tandem MS methods available on an LTQ-Orbitrap XL ETD instrument. For this investigation, a library of glycopeptides with precisely known O-GlcNAc sites was synthesized using a simple randomization approach (Figure 2). Of the 72 possible permutations, 65 glycopeptides could be identified using PQD and ETD tandem MS methods followed by Mascot database search and manual inspection of spectra.

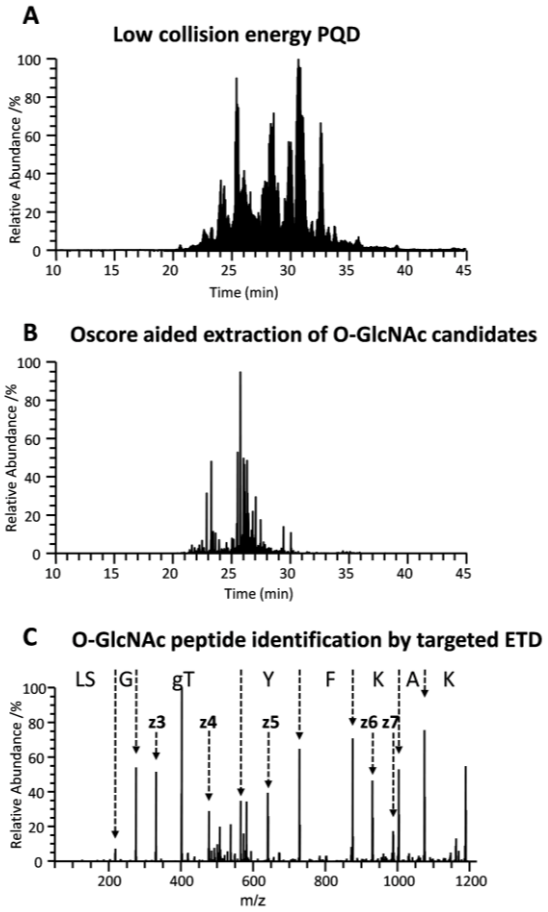

Figure 1 | Analytical strategy for detecting and identifying O-GlcNAc-modified peptides
A key feature of this two-stage approach is the selective extraction of O-GlcNAc candidates from tandem MS spectra based on a novel spectrum scoring scheme.

Figure 2 | O-GlcNAc peptide library synthesized using a sequence randomization approach
The O-GlcNAc peptide library covers a mass range from 1115 to 1286 Da, and represents a heterogeneous set of doubly, triply and quadruply charged peptides ranging from very hydrophilic to hydrophobic and from highly basic to acidic peptides. The dynamic intensity range of the O-GlcNAc library spans almost three orders of magnitude.

Example tandem mass spectra of the peptide LSGgTYFKAK are depicted in Figure 3 and illustrate the merits of each technique. Owing to the chemical lability of the O-glycosidic bond in the gas phase, CID spectra (Figure 3A) are dominated by three signals; the HexNAc oxonium ion at *m/z* 203.98, the neutral loss of the HexNAc moiety from the precursor (*m/z* 493.29) the charge reduced precursor ion that lost the HexNAc oxonium ion.

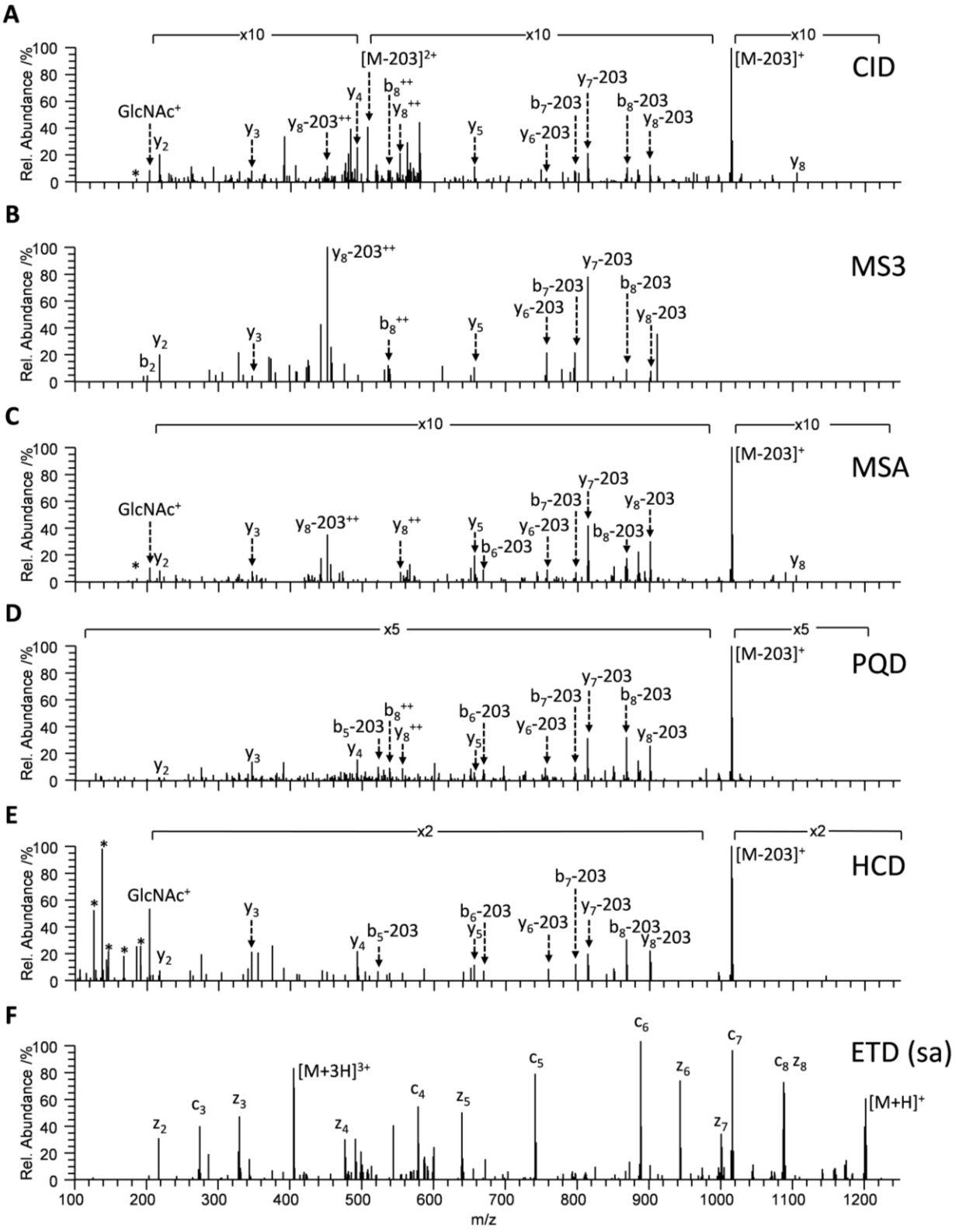

Figure 3 | Example tandem MS spectra of the O-GlcNAc peptide LSGgTYFKAK
Fragments of the HexNAc oxonium ion are marked with an asterisk.

Although the relative intensities of these three signals may vary substantially between different peptides, they usually represent the most intense fragment ions and typically constitute more than 70% of the entire signal in the tandem mass spectrum. Consequently, little if any of the available

signal corresponds to sequence-specific peptide fragment ions which in turn render peptide identification from these spectra difficult. Furthermore, these peptide fragments do generally not retain the O-GlcNAc moiety, thus eliminating direct evidence for the O-GlcNAc modification and site localization from CID spectra. Further fragmentation of the HexNAc neutral loss by NL-MS3 (Figure 3B) or MSA (Figure 3C) significantly increases the yield of peptide fragment ions. Because the neutral loss of the HexNAc group leaves a plain serine or threonine residue at the previously modified O-GlcNAc site, it is impossible to deduce the modification site from NL-MS3 or MSA fragments should more than one possible acceptor site exist within the sequence. PQD and HCD provide access to the full fragment mass range on an LTQ Orbitrap instrument, and hence enable detection of the HexNAc oxonium ion (Figure 3D and E, respectively). Overall, PQD and HCD spectra of O-GlcNAc peptides are quite comparable to CID spectra and hence suffer from similar shortcomings with respect to peptide identification and O-GlcNAc site localisation. However, HCD and, occasionally, PQD fragmentation give rise to further intense peaks below *m/z* 204 which are fragments of the HexNAc oxonium ion [26] (Figure 4). In contrast to the aforementioned activation types, ETD preserves the O-GlcNAc-modification on every peptide fragment ion, thus allowing direct O-GlcNAc site localisation (Figure 3F). However, ETD spectra often exhibit intense non-dissociated electron-transfer products. This can be overcome by supplemental activation of the charge-reduced species [27] resulting in richer spectra than ETD alone. Since the additional radiofrequency pulse does not adversely affect the O-GlcNAc modification, but significantly increases the intensity of peptide fragment ions supporting peptide identification and site localisation, supplemental activation was used for all further experiments involving ETD.

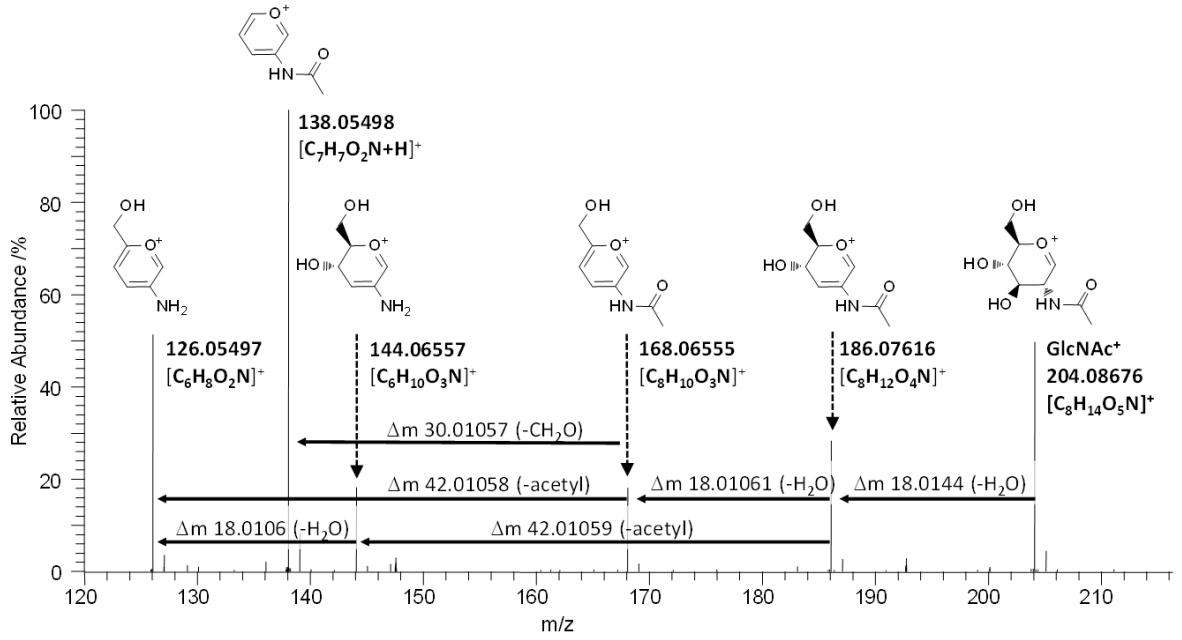

Figure 4 | Fragmentation pathway of the GlcNAc oxonium ion
The GlcNAc oxonium ion gives rise to further fragments during HCD fragmentation.

Searching triplicate LC-MS/MS data from all nine activation types by Mascot identified 48 O-GlcNAc peptides from the library with Mascot ion scores greater than 25 (Table 1). However, the success of identification between individual approaches varied significantly. As shown in Table 1, the results indicate superior performance of PQD with 39 O-GlcNAc peptide identifications followed by ETD with 33 identifications, whereas the least successful approaches were those involving Orbitrap

detection of fragment ions (HCD and ETD [FT]), which only identified 19 and 10 O-GlcNAc peptides, respectively. In total, 1371 tandem mass spectra were matched to the O-GlcNAc peptide library. Of these, 304 spectra point to peptides with multiple serine or threonine residues which carry the risk of false O-GlcNAc-site assignments. Striking differences of the O-GlcNAc site localisation accuracy exist between techniques involving collisional fragmentation and ETD. Both, ETD and ETD (FT) achieve the highest accuracy in O-GlcNAc site assignments (90-100% correct site localisation), while the non-ETD approaches lead to a fairly random assignment of O-GlcNAc sites to serine or threonine residues using Mascot (20-50% correct site localisation).

Table 1 | Comparison of different acquisition modes for O-GlcNAc peptide identification and site localisation

	PQD	ETD (sa)	NL-HCD	NL-ETD	NL-MS3	MSA	CID	HCD	ETD (FT)
Number of identified O-GlcNAc peptides	39	33	32	32	31	28	26	19	10
Correctly localized O-GlcNAc sites	25%	94%	30%	36%	31%	22%	41%	50%	100%
Identified fraction of all O-GlcNAc peptide spectra	227/444	147/512	195/854	229/902	168/633	139/560	126/702	96/428	44/294
Average Mascot ion score	44.8	38.2	35.8	36.8	33.5	31	33.1	40.3	36.2

Recently, the Mascot Delta Score (MD-score) has been introduced as a simple method for confident phosphorylation site assignment [28]. Likewise, the MD-score can be applied to O-GlcNAc spectra to increase confidence in O-GlcNAc site assignments (i. e., high MD-score) and to identify O-GlcNAc site assignments for which evidence from tandem mass spectra is lacking (i. e., low MD-score). While the average MD-score for ETD is 21.0, it is only 0.8 for non-ETD approaches. In addition, the MD-score for 44% of all non-ETD spectra is 0 indicating that no decision at all about site localisation could be made in these cases. It became clear from this systematic investigation, that none of the compared tandem MS approaches was particularly successful in both, O-GlcNAc peptide identification and site localization. It was, therefore, concluded that decoupling O-GlcNAc peptide detection from identification and site localization might improve the analysis of O-GlcNAc peptides, because it would allow combining the best features of each acquisition mode.

Detecting O-GlcNAc peptides with an optimized discovery experiment

Even though CID-type experiments on a LTQ Orbitrap instrument were inappropriate for site localization, the fragment ions involving the sugar moiety can be highly diagnostic. In particular, PQD and HCD provide access to the full fragment mass range enabling the detection of the HexNAc oxonium ion and its fragments. However, the selectivity of the HexNAc oxonium ion may may be compromised by numerous possible interfering peptide fragment ions of very similar mass. The peptide QCPSYFQAK was synthesized with or without O-GlcNAc on the serine residue. In addition to the oxonium ion (*m/z* 204.0866), this peptide can give rise to an a_2(QC) fragment ion (*m/z* 204.0801). This allowed us to investigate how resolution, mass accuracy and collision energy affect the specificity of detection of the HexNAc oxonium ion in the presence of potential interfering ions. Although sulfhydryl groups are typically blocked by alkylation in proteomics experiments, the a_2(QC) fragment was chosen for investigation because it is (along with the isobaric a_3 fragment of the amino acid combination AGC) the only regular peptide fragment within 50 ppm of the mass of the oxonium ion.

Owing to the low resolution of ion trap tandem MS spectra, the a_2(QC) and HexNAc oxonium ions cannot be distinguished by PQD (mass difference 32 ppm). In contrast, this is possible by HCD provided that resolution and/or mass accuracy are sufficiently high. It turns out that 15,000 resolution (FWHM) is insufficient but 30,000 resolution separates both ions to near baseline (Figure 5A). Because HCD spectra acquired at the lowest possible resolution of an Orbitrap (7,500 FWHM) still feature mass accuracy of <10 ppm, unambiguous identification of the oxonium ion is possible by mass accuracy alone provided that one of the two fragment ions has a significantly higher intensity than the other. Yet another alternative for the selective detection of O-GlcNAc peptides is to control the generation of the HexNAc fragment ions by way if the used collision energy. As depicted in Figure 5B, the HexNAc oxonium ion is already generated at low normalized collision energy (NCE) with maxima at 23% NCE (PQD) and 18% NCE (HCD). Both values are considerably lower than the typical NCE values (35-40%) used for CID-type experiments. Concomitantly, peptide backbone fragmentation is much reduced at low collision energy, generating spectra which are almost completely devoid of peptide fragments (Figure 5C).

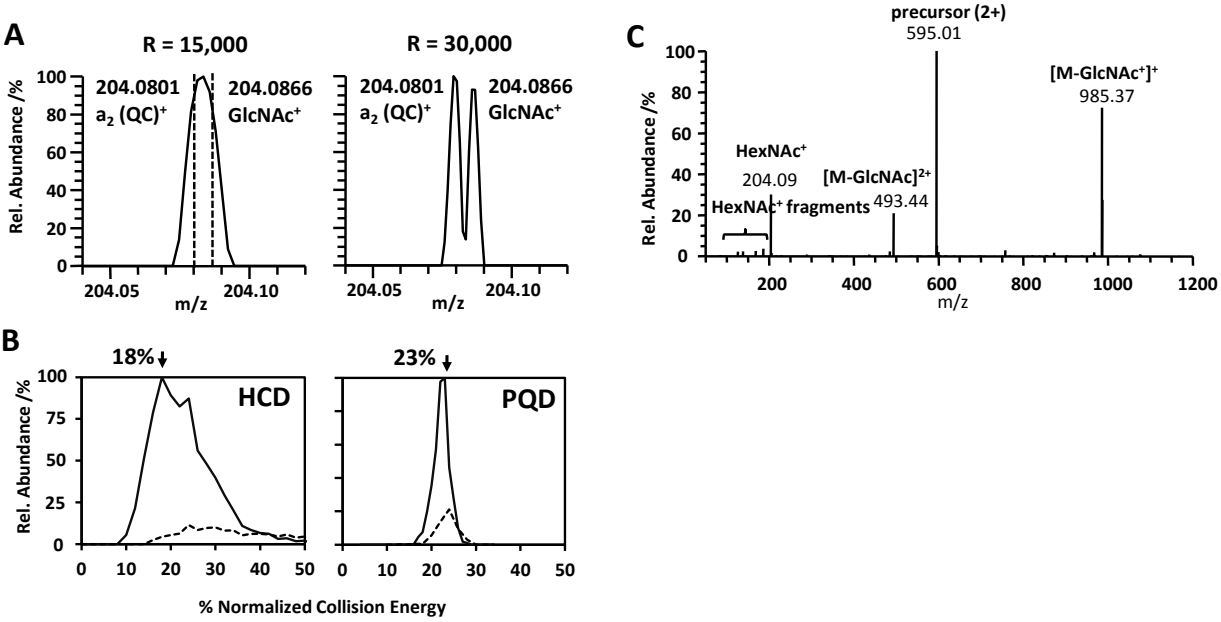

Figure 5 | Resolution, mass accuracy and collision energy affect O-GlcNAc oxonium ion detection
A HCD spectra of QCPgSYFQAK acquired with 15,000 and 30,000 resolution (FWHM at m/z 400). **B** Collision energy characteristic of the HexNAc oxonium ion (solid line) and the interfering a_2(QC) fragment ion (dashed line). **C** Low collision energy PQD spectrum of the peptide LDGgTYFAAK. Note that almost all signals in this spectrum point to an O-GlcNAc modification of the precursor.

HCD detection is inherently less sensitive than PQD, since precursor ions for HCD are isolated in the LTQ, accumulated in the C-trap, before they are injected into the collision octopole, and the fragments are then transferred to the Orbitrap for detection. This process is inevitably accompanied by ion losses, which does not apply for PQD. Second, while the electron multipliers of the LTQ are capable of detecting a single ion, the Orbitrap detector requires a minimum of ~20 charges to detect a signal [29, 30]. This necessitates comparatively high AGC target value settings and consequently longer accumulation times. HCD is also significantly slower than PQD (sequential vs. parallel MS and MS/MS). As expected, a side-by-side comparison of PQD and HCD (in triplicates) using the O-GlcNAc peptide library showed considerable differences in scan speed. While the discovery PQD experiment

generated tandem mass spectra at 2.2 Hz, the speed of HCD data acquisition was only 1.1 Hz. Concomitantly, PQD and HCD generated 320 and 183 O-GlcNAc spectra, respectively, within 45 minutes LC-MS/MS time. All things considered, low energy PQD turned out to be the superior method. It provided sufficient selectivity and was more efficient than HCD for the detection of O-GlcNAc peptides despite its lower mass accuracy and resolution.

Scoring tandem MS spectra for the presence of O-GlcNAc of modified precursors

With an efficient tandem MS method for the generation of diagnostic fragment ions at hand, a simple scoring scheme that differentiates O-GlcNAc from non-O-GlcNAc tandem spectra was devised. This scoring scheme has been termed Oscore because it utilizes and accounts for all spectral features pointing to the O-GlcNAc modification. The Oscore *S* is calculated according to (1)

$$S = -\log_{10} \quad \frac{I_{norm}}{n} \tag{1}$$

where I_{norm} is the normalized intensity (i. e., divided by the sum of all intensities) of up to eight O-GlcNAc-specific spectral features and *n* is the intensity rank within the tandem mass spectrum. For calculation of the Oscore, the fragment intensity is first normalized by the sum of all spectrum features to render the score independent of precursor intensity and, hence, robust against spectra from high abundant precursors. Second, the normalized intensity is further divided by the rank, in order to favour spectra, in which the O-GlcNAc diagnostic fragments are among the most intense peaks. This step concomitantly penalizes spectra, which exhibit intense unspecific signals. The logarithmic transformation is used for convenience to rescale the score. The Oscore is computed using a Perl script which parses the peaklist contained in a mascot generic file (mgf) and calculates the rank and normalized intensities of all peaks in a spectrum before calculating the Oscore. It requires at least one of the O-GlcNAc features to be present in the peaklist within a user-specified mass tolerance. The Oscore script creates a tab-delimited output file containing (among other information) precursor *m/z*, precursor charge state as well as retention time which can be used to build inclusion lists for follow-up targeted experiments.

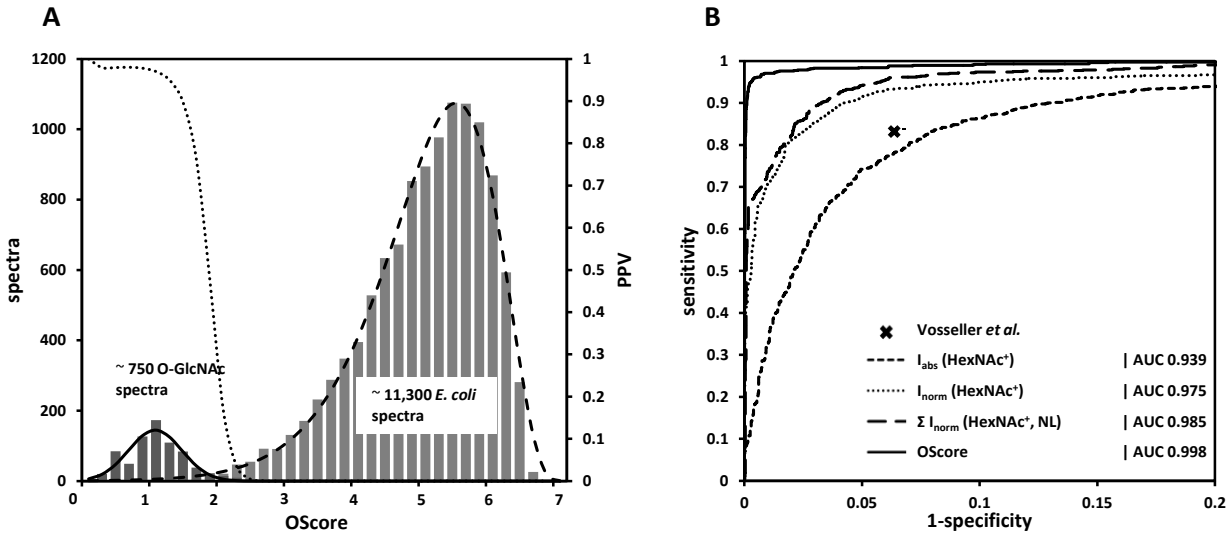

Figure 6 | Oscore-based detection of O-GlcNAc spectra
A Oscore distribution of O-GlcNAc spectra (solid line), non O-GlcNAc spectra (dashed line) and positive predictive value (PPV, dotted line). **B** ROC plots of several alternative O-GlcNAc spectrum classifiers.

In order to assess the discriminating power of the scoring scheme, Oscores were computed for a test set of low collision energy PQD spectra (approximately 750 O-GlcNAc spectra from the O-GlcNAc peptide library and 11,300 non-O-GlcNAc spectra from a tryptic digest of cytosolic *E. coli* proteins). According to Figure 6A, the bimodal Oscore distribution nicely discriminates O-GlcNAc peptides (low Oscores) from unmodified peptides (high Oscores). The Oscore has been compared to other features of O-GlcNAc spectra which could similarly be used as classifier to group O-GlcNAc and non-O-GlcNAc tandem MS spectra, e. g. the approach employed by Vosseller *et al.* [5] which utilizes the combination of the HexNAc oxonium ion and the neutral loss, or the HexNAc oxonium ion intensity, its normalized intensity or the sum of normalized intensities of the oxonium ion and the HexNAc neutral loss. As revealed by a receiver operator characteristic (ROC) analysis, the Oscore outperforms alternative classifiers, and discriminates O-GlcNAc peptide spectra from spectra of unmodified peptides with 95% sensitivity at 99% specificity (Figure 6B). Furthermore, the area under the ROC curve (AUC) of the Oscore is 0.997 indicating very high cumulative accuracy of the classifier.

The bimodal distribution of Oscores allowed the straightforward calculation of the probability that O-GlcNAc spectrum assignments with a given Oscore are correct. Using Bayes' Law and denoting correct and incorrect assignments as "+" and "-", respectively, the positive predictive value (PPV) $p(+|S)$ for an Oscore S can be calculated according to (2)

$$p(+|S) = \frac{p\ S\ +\ p(+)}{p\ S\ +\ p(+)+p(S|-)p(-)} \tag{2}$$

with $p(S|+)$ and $p(S|-)$ being the probabilities of Oscores among O-GlcNAc and non-O-GlcNAc peptides, respectively, and $p(+)$ and $p(-)$ being prior probabilities representing the overall proportion of O-GlcNAc and non-O-GlcNAc spectra in the data set. The calculation of a PPV for a given Oscore from (2) requires accurate models for the Oscore score distributions. The symmetrical distribution of O-GlcNAc spectra was approximated using a Gaussian distribution and the asymmetrically distributed non-O-GlcNAc spectra were modeled on an offset-corrected gamma distribution. Both distributions were fitted to the data using the method of least squares. Thus, with calculated mean μ and standard deviation σ, the probability for a correct O-GlcNAc spectrum assignment with an Oscore S can be calculated according to (3),

$$p\ S\ +\ = \frac{1}{\sqrt{2\pi\sigma}} e^{-\ S-\mu\ ^2/2\sigma^2} \tag{3}$$

while the probability for an incorrect assignment can be calculated according to (4)

$$p\ S\ -\ = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\ (S_m - S)^{\alpha-1} \cdot e^{(S-S_m)/\beta} - S_m^{\alpha-1} \cdot e^{-S_m/\beta} \tag{4}$$

with S_m being the highest observed Oscore and computed parameters α and β. Substitution of $p(S|+)$ and $p(S|-)$ in (2) by the modeled Gaussian and gamma distribution along with computed prior probabilities $p(+)$ and $p(-)$ allowed calculation of PPVs (Figure 6A). It should be noted that low Oscore values correspond to high probabilities and vice versa.

Identification of O-GlcNAc peptides in a complex proteome

To demonstrate the practical utility of the two-stage LC-MS/MS approach, a side-by-side comparison to a conventional data-dependent ETD experiment was performed using a highly complex tryptic digest of cytosolic *E. coli* proteins spiked with decreasing amounts of O-GlcNAc modified bovine α-crystallin (1:10, 1:100, 1:500, 1:1,000 w/w). Bovine α-crystallin is O-GlcNAc-modified at two sites (serine 162 of chain A, threonine 170 of chain B, see supplemental spectra). Serine 162 of chain A is

modified at a stoichiometry of 10% [14] and Thr 170 is barely detectable. Considering that chain A and B are present in 1:1 stoichiometry, the molar content of O-GlcNAc of bovine α-crystallin is in the range of 5%. Hence, the actual spiking ratios in the experiment are in the order of 1:200 to 1:20,000 (w/w). The results are summarized in Table 2 and Figure 7.

Table 2 | Comparison of the two-stage approach and a conventional data-dependent ETD experiment

Spiking ratio	O-GlcNAc peptide /fmol	Discovery PQD			Targeted ETD		Data-dependent ETD	
		intensity	MS/MS	Best Oscore	MS/MS	Best Mascot ion score	MS/MS	Best Mascot ion score
1:200	700	2.1e6	4	1.2	8	55	3	24
1:2,000	70	2.5e5	1	1.8	3	24	1	-
1:10,000	14	5.3e4	-	-	1	-	-	-
1:20,000	7	2.9e4	-	-	-	-	-	-

The data dependent ETD experiment identified the O-GlcNAc peptide AIPVgSREEKPSSAPSS at a spiking ratio of 1:200. The two-stage approach detects and identifies the O-GlcNAc peptide still in the 1:2,000 sample, suggesting an approximately ten-fold increased sensitivity over the conventional data-dependent approach. Both, the limit of detection and the limit of identification are reached at the spiking ratio of 1:2,000, corresponding to 70 fmol O-GlcNAc peptide on column. Notably, the increase in sensitivity comes along with a significant increase in Mascot ion score (55 vs. 24) at the same spiking ratio, thus providing higher confidence for the O-GlcNAc peptide identification.

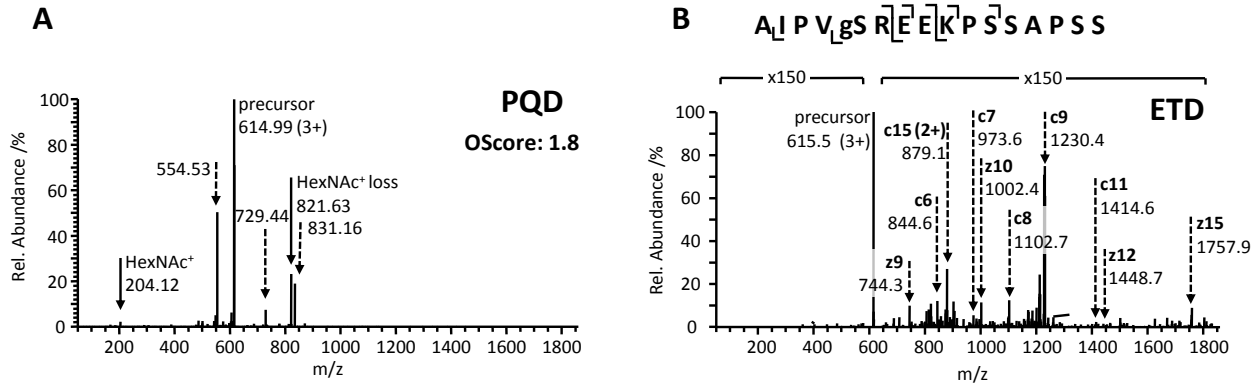

Figure 7 | Limit of O-GlcNAc peptide detection and identification in a complex proteome
A PQD spectrum of the spiked O-GlcNAc peptide (solid arrows indicate O-GlcNAc ions; dashed arrows point to unexplained signals). **B** The targeted ETD spectrum leads to the identification of AIPVgSREEKPSSAPSS.

Discussion

Systematic evaluation of tandem MS techniques

In light of the poor CID fragmentation of O-GlcNAc peptides, the numerous fragmentation approaches available on an LTQ Orbitrap XL ETD instrument have been revisited for their merits in O-GlcNAc peptide identification and site-localization. Surprisingly, the highest number of O-GlcNAc peptides was identified by PQD (Table 1). ETD, NL-ETD and NL-HCD also led to a reasonable number of O-GlcNAc peptide identifications. For the latter two, this is reasonable as both spectra are triggered by the detection of a diagnostic neutral loss in the preceding CID spectrum. However,

along with other CID-type fragmentation techniques, PQD spectra could not be utilized to localize O-GlcNAc sites reliably. Here, ETD fragmentation offers the distinct advantage that it preserves the O-GlcNAc modification and thus enables the direct inference of the accurate site of modification (Table 1). But ETD has its limitations too: increasing mass and decreasing charge density of the precursor diminish the fragmentation efficiency of ETD [31] and may render the O-GlcNAc site determination with ETD impossible for large peptides. Among the CID-like fragmentation approaches, HCD was the most accurate fragmentation technique with 50% correctly identified O-GlcNAc sites by Mascot. This can be reasoned by the high mass accuracy and high dynamic range of HCD spectra [21], which also allow deducing the O-GlcNAc localization from very low intensity signals.

Scoring tandem mass spectra for presence of the O-GlcNAc modification

The Oscore is a conceptually new and straightforward approach to evaluating the presence of an O-GlcNAc modification of potentially modified peptides based on tandem mass spectra. The Oscore does not require the detection of sequence-informative peptide fragments, but instead exclusively relies on the presence and intensity of up to eight different fragments originating from the breakage of the O-glycosidic bond (Figure 3b). Unlike other PTM scores [32-36], the Oscore does not contribute any information about the localization of O-GlcNAc modification or the underlying peptide sequence. Instead, it assesses tandem MS spectra of complex peptide mixtures for the presence of the modification (Figure 6). Using Bayesian statistics, the Oscore can be further transformed into a positive predictive value for a given Oscore. While these probability computations require a sufficient number of O-GlcNAc spectra to model score distributions of O-GlcNAc and non-O-GlcNAc spectra, the Oscore itself is calculated on a single spectrum basis and as such indicates the presence of an O-GlcNAc modification by a low Oscore irrespective of whether a single or hundreds of O-GlcNAc peptides are present in a sample. By design, the Oscore is independent of the precursor signal intensity and as such independent of the amount of peptide on column. However, the quality of the Oscore will decrease with decreasing signal-to-noise ratio of the precursor ion because the contribution of the chemical noise inevitably increases until the resulting tandem MS spectrum no longer primarily reflects the isolated O-GlcNAc peptide.

Nevertheless, the score is quite robust as exemplified in Figure 7. Apart from diagnostic fragments for O-GlcNAc, this PQD spectrum contains three intense signals (554.53 *m/z*, 729.44 *m/z*, 831.16 *m/z*) which cannot be explained by typical fragment ions for the peptide AIPVgSREEKPSSAPSS, but very likely result from co-isolation and co-fragmentation of another peptide. With an Oscore of 1.8, the precursor ion generating this mixed tandem mass spectrum is a reasonable candidate for inclusion in a targeted ETD experiment which confirmed the sequence and modification.

Peptides modified by N- or O-linked glycans will probably also result in fairly low Oscores (i. e. high O-GlcNAc probability), as they may lose multiple HexNAc groups from their non-reducing ends. On the other hand, fragmentation of complex glycans will also result in numerous signals that do not indicate the O-GlcNAc modification, and, hence, increase the Oscore and decrease the probability of a false-positive O-GlcNAc spectrum assignment. The recently reported intracellular single N-linked HexNAc modification presumably resulting from breakdown of glycoproteins [6, 8], will probably not interfere, since the N-glycosidic linkage is more stable than the O-glycosidic bond under CID conditions [37] and the targeted ETD experiment would resolve this particular issue. Future experiments will address if and to which extent this modification may influence the discovery of O-GlcNAc peptides.

In addition to its utility in identifying candidate O-GlcNAc species in complex mixtures, the Oscore may also support O-GlcNAc peptide identification by database searching. In particular in large-scale studies, the poor CID fragmentation of O-GlcNAc peptides, along with low search engine scores, is likely to result in both, an unnecessarily high proportion of overlooked (i. e., false-negative) as well as false-positive O-GlcNAc peptide-spectrum matches (PSMs). The Oscore provides complementary information to that used by search engines, which may be used to 'rescue' genuine O-GlcNAc identifications despite having low search engine scores and to discriminate correct from incorrect O-GlcNAc PSMs. Both ways around, data quality would increase significantly. Another application of the Oscore may be the retrospective analysis of existing data sets. This, however, would only work for data sets that were created by tandem MS including the full mass range to ensure the detection of the HexNAc oxonium ion and its fragments.

Sensitivity of detection and identification

Compared to a conventional data-dependent ETD experiment, the two-stage approach resulted in a ten-fold increased sensitivity and significantly improved Mascot ion scores for the analyzed O-GlcNAc peptide of bovine α-crystallin spiked into a tryptic digest of *E. coli* proteins (Table 2 and Figure 7). These improvements were achieved by decoupling the detection of potential O-GlcNAc precursors (by PQD) from their actual identification and site localization (by ETD), as well as by scoring tandem mass spectra for the presence of the O-GlcNAc moiety. This ten-fold gain in sensitivity comes, however, at the expense of requiring twice the amount of sample and measurement time. The limit of detection and identification determined here for a single peptide spiked into a highly complex background (low fmol range) probably represents a very conservative estimate. For less complex mixtures such as O-GlcNAc enriched proteomes or single O-GlcNAc proteins, one might expect limits of detection and identification in the mid amol range.

The PQD discovery experiment and the classical data-dependent ETD experiment acquired a similar number of tandem MS spectra at each spiking level (Table 2). They, therefore, had the same chance of detecting and identifying an O-GlcNAc modified peptide. However, at the spiking level of 1:2,000 (w/w), only the PQD discovery experiment in conjunction with the Oscore, allowed the identification of the precursor 615.6461 *m/z* (3+) as a potentially modified peptide. In contrast, the conventional data dependent ETD experiment did not lead to a successful O-GlcNAc peptide identification at spiking ratios of less than 1:200 (w/w). Although the respective precursor ion can still be detected in the full scan spectra at higher spiking ratios, it was no longer among the species selected for fragmentation by PQD in a discovery experiment or a conventional data-dependent ETD experiment (Table 2).

While the PQD discovery experiment as well as the Oscore was developed to maximize selectivity and sensitivity, the second-stage experiment focused on the targeted peptide identification and O-GlcNAc site localization. ETD was selected for this purpose because of its outstanding accuracy in O-GlcNAc site identification and its sound peptide identification performance (Table 1). Employing ETD in a targeted fashion enabled the acquisition of multiple ETD spectra across the chromatographic peak which was accomplished by disabling the monoisotopic precursor selection as well as the rejection of unassigned charge states in the MS acquisition software. Owing to an enhanced signal-to-noise ratio and reliable ion statistics in tandem MS spectra, the acquisition of multiple ETD spectra per chromatographic peak resulted in both, an increased chance to identify the O-GlcNAc

peptide, as well as a higher confidence that the O-GlcNAc peptide-spectrum match is correct (Table 2).

Translation of the two-stage approach to other MS instruments

The two-stage approach has been developed using an LTQ Orbitrap XL ETD mass spectrometer. However, the approach can be easily translated to any other type of mass spectrometer which is capable of detecting low-mass ions in tandem mass spectra and is ETD-enabled. When doing so, three aspects have to be considered. First, for the discovery experiment, it is important to adjust the fragmentation amplitude to generate O-GlcNAc spectra showing O-GlcNAc diagnostic fragments while suppressing (interfering) peptide fragment ions. Second, the peak list generated as input for the Oscore script has to be converted into the Mascot generic format, e. g. using one of the free available peak list conversion tools. Third, the resulting Oscore distribution of O-GlcNAc and non-O-GlcNAc spectra will likely be different from instrument to instrument. Figure 8 shows the Oscore distribution of O-GlcNAc and non-O-GlcNAc spectra acquired on an amaZon ETD ion trap mass spectrometer. As expected, the Oscore distributions of PQD (Figure 6) and the corresponding PAN experiment on the amaZon instrument (Figure 8) are alike, but span a different scale. Consequently, the Oscore threshold for O-GlcNAc candidates has to be adjusted. The approach lends itself to real-time decision making akin to what has been proposed for the analysis of phosphopeptides [31] and further improvements should arise from increasing scan speed and sensitivity of ion trap - Orbitrap [38] and quadrupole TOF mass spectrometers [39, 40].

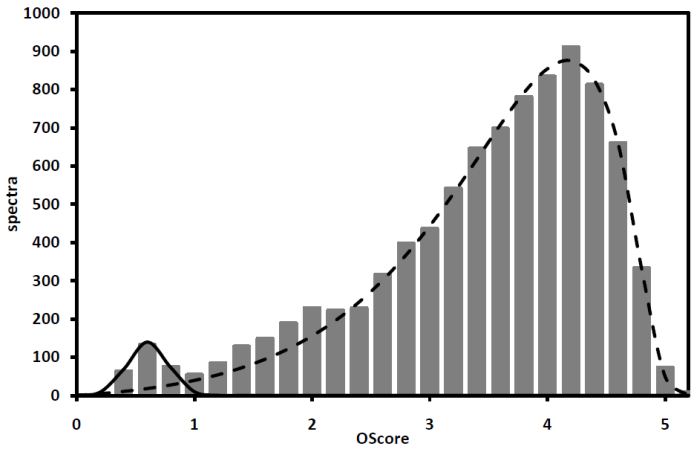

Figure 8 | Oscore distribution of O-GlcNAc (O-GlcNAc peptide library) and non-O-GlcNAc spectra (*E. coli* digest) acquired on an amaZon ETD ion trap mass spectrometer (Bruker Daltonics, Bremen) using the PAN scan mode

Like PQD, the PAN fragmentation technique enables the detection of low m/z fragments in ion trap tandem mass spectra.

Conclusion

The developed analytical strategy has great potential for the broad-scale discovery of O-GlcNAc-containing proteins, particularly if combined with O-GlcNAc-specific enrichment tools. Further, the Oscore can be expected to become a valuable tool to improving the quality of O-GlcNAc peptide spectrum assignments. Finally, it can be anticipated that the increase in the number of documented O-GlcNAc proteins discovered in this or similar ways will shed further light on the functional significance of this emerging intracellular protein modification.

Acknowledgments

The author is indebted to Simone Lemeer for valuable support in aspects of mass spectrometry, to Kurt Fellenberg for an introduction into Perl programming and to Andrea Hubauer for an introduction into solid-phase peptide synthesis.

Abbreviations

AGC	automatic gain control
ETD (sa)	ETD with supplemental activation
gS	β-O-GlcNAc modified serine residue
gT	β-O-GlcNAc modified threonine residue
HCD	higher energy C-trap dissociation
NCE	normalized collision energy
NL-ETD	neutral loss-triggered ETD
NL-HCD	neutral loss-triggered HCD
NL-MS3	neutral loss-triggered MS3
MSA	multistage activation
PQD	pulsed Q dissociation
PSM	peptide spectrum match

References

1. Torres, C. R., and Hart, G. W. (1984) Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *J Biol Chem* 259, 3308-3317.
2. Love, D. C., and Hanover, J. A. (2005) The hexosamine signaling pathway: deciphering the "O-GlcNAc code". *Sci STKE* 2005, re13.
3. Dias, W. B., and Hart, G. W. (2007) O-GlcNAc modification in diabetes and Alzheimer's disease. *Mol Biosyst* 3, 766-772.
4. Chou, T. Y., and Hart, G. W. (2001) O-linked N-acetylglucosamine and cancer: messages from the glycosylation of c-Myc. *Adv Exp Med Biol* 491, 413-418.
5. Vosseller, K., Trinidad, J. C., Chalkley, R. J., Specht, C. G., Thalhammer, A., Lynn, A. J., Snedecor, J. O., Guan, S., Medzihradszky, K. F., Maltby, D. A., Schoepfer, R., and Burlingame, A. L. (2006) O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. *Mol Cell Proteomics* 5, 923-934.
6. Chalkley, R. J., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc Natl Acad Sci USA* 106, 8894-8899.
7. Khidekel, N., Ficarro, S. B., Clark, P. M., Bryan, M. C., Swaney, D. L., Rexach, J. E., Sun, Y. E., Coon, J. J., Peters, E. C., and Hsieh-Wilson, L. C. (2007) Probing the dynamics of O-GlcNAc glycosylation in the brain using quantitative proteomics. *Nat Chem Biol* 3, 339-348.
8. Wang, Z., Udeshi, N. D., Slawson, C., Compton, P. D., Sakabe, K., Cheung, W. D., Shabanowitz, J., Hunt, D. F., and Hart, G. W. (2010) Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Sci Signal* 3, ra2.
9. Teo, C. F., Ingale, S., Wolfert, M. A., Elsayed, G. A., Not, L. G., Chatham, J. C., Wells, L., and Boons, G. J. (2010) Glycopeptide-specific monoclonal antibodies suggest new roles for O-GlcNAc. *Nat Chem Biol* 6, 338-343.
10. Haynes, P. A., and Aebersold, R. (2000) Simultaneous detection and identification of O-GlcNAc-modified glycoproteins using liquid chromatography-tandem mass spectrometry. *Anal Chem* 72, 5402-5410.
11. Hart, G. W., Housley, M. P., and Slawson, C. (2007) Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* 446, 1017-1022.
12. Hu, P., Shimoji, S., and Hart, G. W. (2010) Site-specific interplay between O-GlcNAcylation and phosphorylation in cellular regulation. *FEBS Lett* 584, 2526-2538.
13. Huddleston, M. J., Bean, M. F., and Carr, S. A. (1993) Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests. *Anal Chem* 65, 877-884.
14. Chalkley, R. J., and Burlingame, A. L. (2001) Identification of GlcNAcylation sites of peptides and alpha-crystallin using Q-TOF mass spectrometry. *J Am Soc Mass Spectrom* 12, 1106-1113.
15. Jebanathirajah, J., Steen, H., and Roepstorff, P. (2003) Using optimized collision energies and high resolution, high accuracy fragment ion selection to improve glycopeptide detection by precursor ion scanning. *J Am Soc Mass Spectrom* 14, 777-784.
16. Medzihradszky, K. F., Gillece-Castro, B. L., Townsend, R. R., Burlingame, A. L., and Hardy, M. R. (1996) Structural elucidation of O-linked glycopeptides by high energy collision-induced dissociation. *J Am Soc Mass Spectrom* 7, 319-328.
17. Carr, S. A., Huddleston, M. J., and Bean, M. F. (1993) Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Sci* 2, 183-196.
18. Carapito, C., Klemm, C., Aebersold, R., and Domon, B. (2009) Systematic LC-MS analysis of labile post-translational modifications in complex mixtures. *J Proteome Res* 8, 2608-2614.

19. Cunningham, C., Jr., Glish, G. L., and Burinsky, D. J. (2006) High amplitude short time excitation: a method to form and detect low mass product ions in a quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* 17, 81-84.
20. Schwartz, J. C., Syka, J. E., and Quarmby, S. T. (2005) Improving the Fundamentals of MSn on 2D Ion Traps: New Ion Activation and Isolation Techniques. *53rd ASMS Conference on Mass Spectrommetry*, San Antonio, TX, USA.
21. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* 4, 709-712.
22. Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F., and Coon, J. J. (2004) A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem* 76, 3590-3598.
23. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J Am Chem Soc* 120, 3265-3266.
24. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* 101, 9528-9533.
25. Mirgorodskaya, E., Roepstorff, P., and Zubarev, R. A. (1999) Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal Chem* 71, 4431-4436.
26. Chalkley, R. J., and Burlingame, A. L. (2003) Identification of novel sites of O-N-acetylglucosamine modification of serum response factor using quadrupole time-of-flight mass spectrometry. *Mol Cell Proteomics* 2, 182-190.
27. Swaney, D. L., McAlister, G. C., Wirtala, M., Schwartz, J. C., Syka, J. E., and Coon, J. J. (2007) Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal Chem* 79, 477-485.
28. Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., and Kuster, B. (2011) Confident phosphorylation site localization using the Mascot Delta Score. *Mol Cell Proteomics* 10, M110 003830.
29. Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., and Horning, S. (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem* 78, 2113-2120.
30. Bantscheff, M., Boesche, M., Eberhard, D., Matthieson, T., Sweetman, G., and Kuster, B. (2008) Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol Cell Proteomics* 7, 1702-1713.
31. Swaney, D. L., McAlister, G. C., and Coon, J. J. (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat Methods* 5, 959-964.
32. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 24, 1285-1292.
33. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2006) 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, 935-948.
34. 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.
35. Mortensen, P., Gouw, J. W., Olsen, J. V., Ong, S. E., Rigbolt, K. T., Bunkenborg, J., Cox, J., Foster, L. J., Heck, A. J., Blagoev, B., Andersen, J. S., and Mann, M. (2010) MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J Proteome Res* 9, 393-403.
36. Ruttenberg, B. E., Pisitkun, T., Knepper, M. A., and Hoffert, J. D. (2008) PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. *J Proteome Res* 7, 3054-3059.

37. Medzihradszky, K. F. (2005) Characterization of protein N-glycosylation. *Methods Enzymol* 405, 116-138.
38. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics* 8, 2759-2769.
39. Ow, S. Y., Noirel, J., Salim, M., Evans, C., Watson, R., and Wright, P. C. (2010) Balancing robust quantification and identification for iTRAQ: application of UHR-ToF MS. *Proteomics* 10, 2205-2213.
40. Ibrahim, Y. M., Prior, D. C., Baker, E. S., Smith, R. D., and Belov, M. E. (2010) Characterization of an Ion Mobility-Multiplexed Collision Induced Dissociation-Tandem Time-of-Flight Mass Spectrometry Approach. *Int J Mass Spectrom* 293, 34-44.

Chapter 3

Discovery of O-GlcNAc-modified proteins in published large-scale proteome data

Summary

The attachment of β-N-acetylglucosamine to serine or threonine residues (O-GlcNAc) is a post-translational modification on nuclear and cytoplasmic proteins with emerging roles in numerous cellular processes, such as signal transduction, transcription and translation. It is further presumed that O-GlcNAc can exhibit a site-specific, dynamic and possibly functional interplay with phosphorylation. O-GlcNAc proteins are commonly identified by tandem mass spectrometry following some form of biochemical enrichment. In the present study, it was assessed if, and to which extent, O-GlcNAc-modified proteins can be discovered from existing large-scale proteome data sets. To this end, a straightforward O-GlcNAc identification strategy was conceived based on the recently developed Oscore software that automatically analyzes tandem mass spectra for the presence and intensity of O-GlcNAc diagnostic fragment ions. Using the Oscore, hundreds of O-GlcNAc peptides not initially identified in these studies were discovered, and most of which have not been described before. Merely re-searching this data extended the number of known O-GlcNAc proteins by almost 100 suggesting that this modification exists even more widely than previously anticipated and the modification is often sufficiently abundant to be detected without enrichment. However, a comparison of O-GlcNAc and phospho-identifications from the very same data indicates that the O-GlcNAc modification is considerably less abundant than phosphorylation. The discovery of numerous doubly modified peptides (i. e., peptides with one or multiple O-GlcNAc or phosphate moieties), suggests that O-GlcNAc and phosphorylation are not necessarily mutually exclusive, but can occur simultaneously at adjacent sites.

Introduction

The modification of proteins with β-N-acetylglucosamine (O-GlcNAc) is an emerging dynamic post-translational modification of serine or threonine residues of proteins. O-GlcNAc is found on a wide range of proteins involved in virtually all cellular processes as well as various human diseases [1, 2] including cancer [3]. In addition, O-GlcNAc can interplay with phosphorylation, which, for instance, modulates the stability and activity of p53 [4]. Despite its biological importance, the analysis of O-GlcNAc-modified proteins remains highly challenging. In fact, of the approximately 800 reported O-GlcNAc proteins, direct and unambiguous evidence for the site of O-glycosylation is available for less than 25% of these [5].

The identification of O-GlcNAc proteins is typically achieved by combining selective enrichment and liquid chromatography tandem mass spectrometry (LC-MS/MS). Albeit powerful, the identification of modified peptides and sites is hindered by the substoichiometric occupancy of O-GlcNAc sites [2] and the lability of the O-glycosidic bond in the gas phase [6]. In mass spectrometry-based proteomics, peptides are usually sequenced via collision-induced dissociation (CID). However, under typical CID conditions, the concurrent O-GlcNAc peptide and site identification is difficult, because peptides readily lose the GlcNAc moiety, and spectra are dominated by neutral loss species along with the GlcNAc oxonium ion and fragments thereof [7]. Peptide sequence identification is often still possible from fragments that lost the O-GlcNAc moiety, but site information is irretrievably lost upon dissociation of the O-glycosidic bond. In contrast, the fragmentation of peptides with electron capture dissociation (ECD) or electron transfer dissociation (ETD) typically preserves PTM sites and allows the direct and simultaneous identification of O-GlcNAc peptide sequences and sites [8, 9] but these techniques also have shortcomings notably concerning sensitivity on most current commercial platforms.

Although not ideal for O-GlcNAc site localization, the initial detection of O-GlcNAc peptides is strongly facilitated in CID-type experiments [10, 11] because diagnostic GlcNAc losses along with the GlcNAc oxonium ion and its fragments define a characteristic pattern, which identifies O-GlcNAc peptides even in very complex proteomics samples [9]. The availability of high resolution and high mass accuracy instruments further improves the selectivity of these diagnostic fragment ions [12, 13].

The recently developed a bioinformatics tool, termed Oscore automatically assesses tandem MS spectra for the presence and intensity of O-GlcNAc diagnostic fragment ions and, in turn, allows ranking spectra according their probability of representing an O-GlcNAc peptide [12]. On a test data set of 750 O-GlcNAc spectra and 11,300 spectra from unmodified peptides, the Oscore was able to discriminate O-GlcNAc spectra from spectra of unmodified peptides with 95% sensitivity and >99% specificity and outperformed alternative approaches such as the simple filtering for diagnostic ions. The present study shows that the Oscore can be applied to existing large-scale proteomic data to discover hundreds of O-GlcNAc peptides not initially identified in these studies. Merely re-searching this data extended the number of known O-GlcNAc proteins by almost 100 suggesting that this modification exists even more widely than previously anticipated and is often abundant enough to be detected without specific biochemical enrichment.

Experimental procedures

Publically available data

Publically available raw mass spectrometric data from published proteome-wide studies of eleven different cell lines [14], HeLa cells [15], as well as data from published proteome-wide and phospho-proteome studies of hES and iPS cells [16] were downloaded from respective repositories.

Data analysis

The mass spectrometric data were processed essentially as described [12]. Briefly, peak picking and processing was performed using Mascot Distiller 2.4.2.0 (Matrix Science, London, UK) in which merging of tandem MS spectra from the same precursor as well as isotope fitting of fragments below *m/z* 205 was disabled. The resulting peak list files were processed by the Oscore perl script, which calculates the Oscore for every peptide precursor for which the tandem MS spectrum contains at least one diagnostic O-GlcNAc feature within a tolerance of 10 ppm. The peak list files were searched with Mascot 2.3.0 against the UniProtKB complete human (download date 26.10.2010, 110,550 sequences) combined with sequences of common contaminants. In case of the phospho-proteome dataset of hES and iPS cells [16], the spectra were searched against a subset database generated with Scaffold 3.3.1 (Proteome Software, Portland, OR) including only protein identifications from the respective full proteome data set (11,288 sequences). Carbamido-methylation of cysteine residues, oxidation of methionine, and HexNAc modification of serine, threonine and asparagine residues were taken into account as variable modifications. Where applicable, phosphorylation of serine, threonine and tyrosine residues was set as variable modification. Likewise, 4-plex or 8-plex iTRAQ was set as fixed modification at the peptide amino-terminus and lysine side chain for data sources using these peptide tags. According to the proteases employed in the original studies, enzyme specificity was set to trypsin (lysine, arginine), LysC (lysine), or GluC (aspartic acid, glutamic acid) allowing for up to two missed cleavage sites. The target-decoy option of Mascot was enabled and peptide mass tolerance was set to 10 ppm and fragment mass tolerance to 0.02 Da. Search results were imported into Scaffold 3.3.1. Proteins were required to have at least 99% protein probability and 80% peptide probability. Candidate O-GlcNAc spectra were filtered against false-positive O-GlcNAc peptide-spectrum-matches (PSMs) to retain only O-GlcNAc PSMs with Oscores smaller than 2.3. Candidate O-GlcNAc PSMs were inspected and validated manually.

A list of known human and murine O-GlcNAc proteins and sites was compiled from recent publications [13, 17-19] as well as from the databases dbOGAP [5] and PhosphositePlus [20]. Information on phosphorylated and ubiquitinylated proteins was retrieved from the PhosphositePlus database. Reported N-linked glycosylation sites were extracted from from UniProtKB, and sub-cellular localization information from Ingenuity Pathway Analysis software (Ingenuity Systems, Redwood City, CA).

Results and discussion

Oscore-based O-GlcNAc protein identification strategy

The recently developed Oscore is a reliable means to assess the probability of a tandem MS spectrum to represent an O-GlcNAc modified peptide [12]. The high specificity of the score is further increased by the high mass accuracy provided by modern mass spectrometers. It was therefore reasoned that it may be possible to identify O-GlcNAc modified peptides from large-scale proteomic data and, if so, to assess the overall abundance of the modification. To this end, a number of published data sets were downloaded from public data repositories which were all acquired on dual pressure linear ion trap Orbitrap hybrid mass spectrometers using HCD fragmentation [21]. The first data set comprises the label-free comparison of eleven commonly used cell lines [14]; the second data set comprises a comprehensive characterization of the HeLa cancer cell line proteome employing multiple protease digestion [15] and the third data set represents an iTRAQ-based quantitative comparison of the proteome and the phosho-proteome of four human embryonic stem (hES) cell lines and four induced pluripotent stem (iPS) cell lines [16]. Together, these data sets constitute 13,897,945 tandem MS spectra.

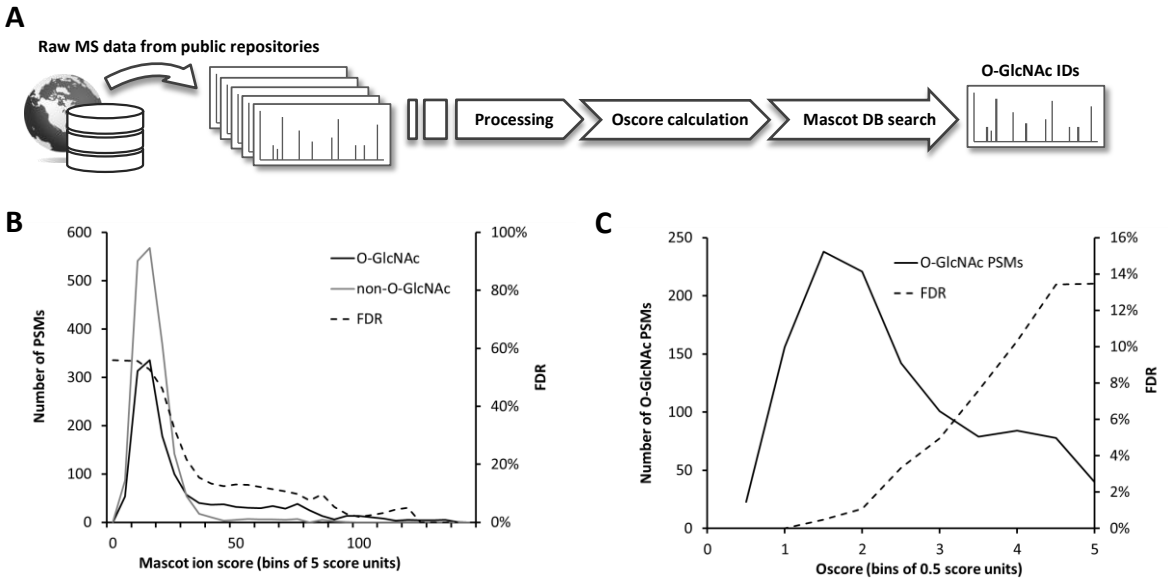

Figure 1 | O-GlcNAc protein identification strategy
A Raw LC-MS/MS data is downloaded from public data repositories, tandem mass spectra are processed into peak lists. These are examined for candidate O-GlcNAc information using the Oscore and identified by database searching using Mascot. **B** Mascot ion score distribution of candidate O-GlcNAc PSMs. For FDR estimation, PSMs which were assigned to O-GlcNAc-modified sequences, but did not contain O-GlcNAc diagnostic features, were considered as false-positive hits and are indicated as "non-O-GlcNAc" PSMs. **C** Oscore distribution of candidate O-GlcNAc PSMs. The FDR is calculated on target and decoy PSMs and spectra with Oscores of <2.3 correspond to an FDR of 2.5%.

The straightforward strategy for data re-analysis combines standard Mascot database searching and Oscoring of tandem mass spectra for the assessment of potential O-GlcNAc spectra (Figure 1A). Both algorithms exploit complementary properties of tandem MS spectra. While the Mascot ion score reflects peptide sequence information, the Oscore assesses tandem MS spectra solely based on the presence of O-GlcNAc diagnostic fragment ions. Given the particular fragmentation behavior of O-

GlcNAc peptides, the Mascot ion score alone is not able to discriminate accurately between O-GlcNAc and non-O-GlcNAc spectra (Figure 1B). However, when O-GlcNAc PSMs assigned by Mascot are re-assessed according to their Oscore, it is easily possible to discriminate between O-GlcNAc and non-O-GlcNAc spectra. Low Oscores represent strong O-GlcNAc spectra, high Oscores represent weak or unlikely O-GlcNAc spectra and no Oscore represent the absence of typical O-GlcNAc features. The Oscore-based ranking of O-GlcNAc PSMs then allows filtering the data at the desired target-decoy FDR while maintaining adequate sensitivity (Figure 1C).

O-GlcNAc sites from HCD spectra

The Oscore-based re-analysis of three comprehensive cell line proteome data sets resulted in the identification of 158 O-GlcNAc peptides containing 194 sites from 628 spectra (Table 1). Manual interpretation of the best PSM for every peptide allowed the unambiguous localisation of 26 O-linked GlcNAc and 12 N-linked GlcNAc sites (see below). The localisation of 13 sites could be narrowed down to three or less residues, and the localisation of 140 sites remained ambiguous.

Table 1 | O-GlcNAc protein and peptide identifications from published large-scale proteome studies

Project	MS/MS	PSM	Peptides	Sites	Proteins
Geiger *et al.*	5,985,620	454	104	125	76
Nagaraj *et al.*	4,829,525	75	36	38	29
Phanstiel *et al.*	1,766,566	99	41	50	32
Total	12,581,711	628	158*	194*	114*
Phanstiel *et al.*(phospho data set)	1,316,234	107	28	34	22
Total + phospho	13,897,945	735	174*	204*	124*

*nonredundant peptides, sites and proteins

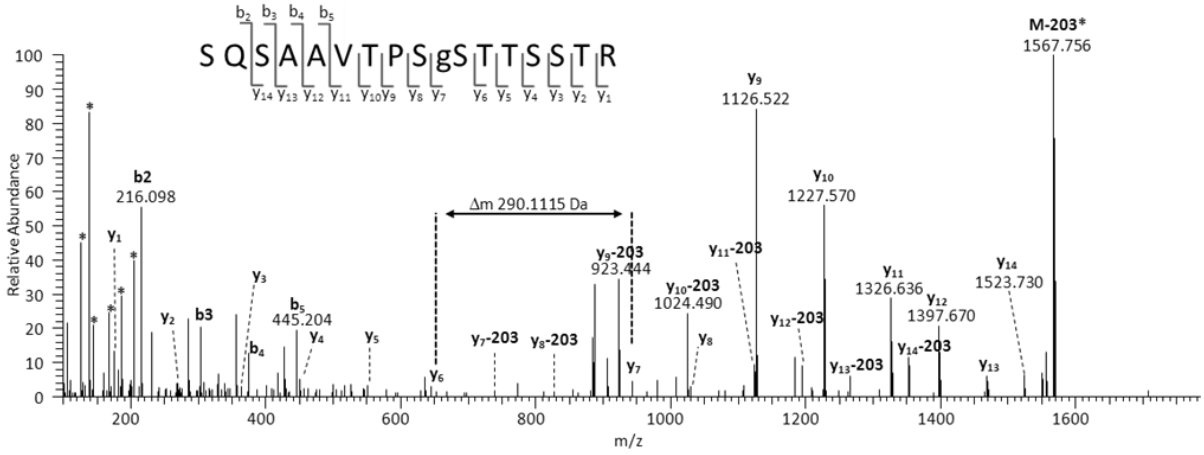

Figure 2 | Example HCD spectrum of a novel O-GlcNAc site corresponding to the sequence SQSAAVTPSgSTTSSTR of the proteasomal ubiquitin receptor ADRM1

The large dynamic range of the HCD spectrum and the high mass accuracy allows determining the peptide sequence and the localisation of the O-GlcNAc site despite the presence of nine alternative modification sites. Diagnostic O-GlcNAc features (i. e. reporter ions and the GlcNAc oxonium loss) are depicted as well (*).

An example O-GlcNAc HCD spectrum is depicted in Figure 2. The high mass accuracy and the large dynamic range of HCD spectra facilitate not only the identification of the SQSAAVTPS**gS**TTSSTR peptide from ADRM1, but also support the detection of the PTM via diagnostic fragments and allows

the unambiguous localisation of the O-GlcNAc site even in the presence of nine alternative sites. Although it has been possible to identify numerous O-GlcNAc sites from HCD spectra, the low stability of the O-glycosidic bond during CID conditions render the localization of O-GlcNAc sites very difficult. Clearly, the fragmentation method of choice for an accurate O-GlcNAC site localization is ETD, which retains the O-GlcNAc site during fragmentation and enables direct site localization. However, stretches of serine and threonine residues around the actual O-GlcNAc site further impede site localization. Only five out of 158 peptides have only a single possible O-GlcNAc site (Ser, Thr), and the average number of potential sites per peptide is 5.6. This is consistent with published O-GlcNAc transferase consensus motifs [5]. Interestingly, non-modified peptides contain only 1.5 possible O-GlcNAc acceptor sites and phospho-peptides (see below) harbour 3.3 possible O-GlcNAc sites, suggesting that O-GlcNAc is more likely to occur on serine/threonine-rich peptides.

Among the 158 GlcNAc peptides are 12 peptides for which the GlcNAc modification could be localized to N-linked asparagine residues within an NX[ST] consensus motif. In addtion, 20 peptides for which the site of modification could not be reliably deduced from tandem mass spectra, harbour N-linked glycosylation sites reported in UniProt (also see Table S4). Although single N-linked GlcNAc residues are not generally expected to be present on proteins, this result is in accordance with previous findings [18]. A possible explanation raised by Chalkley *et al.* is that these N-linked HexNAc peptides are artefacts formed upon cell lysis by the activity of the cytosolic endo-β-N-acetylglucosaminidase. The enzyme cleaves the β-1,4-glycosidic bond in the N,N'-diactylchitobiose core of high mannose glycopeptides and glycoproteins leaving an N-linked GlcNAc residue. However, these N-GlcNAc peptides, as well as peptides from O-glycans, may also arise from in-source fragmentation of the glycan structure in the high pressure region at the front end of the mass spectrometer.

Identified O-GlcNAc proteins

After processing more than 12 million tandem mass spectra, 628 O-GlcNAc spectra corresponding to 158 peptides and 114 candidate O-GlcNAc proteins were identified. The three re-examined studies contribute common and exclusive protein identifications (Figure 3A). The highest number of modified proteins originates from the eleven cell line proteomes profiled by Geiger *et al.* [14].

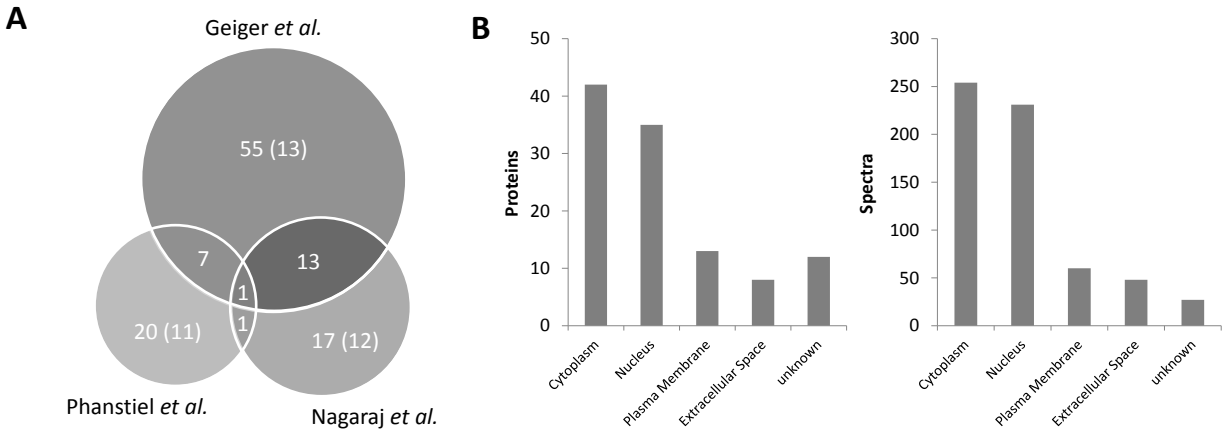

Figure 3 | Number of O-GlcNAc proteins identified in different studies from various cell lines
A Number of O-GlcNAc proteins and known proteins (in parentheses) from three different data sets. **B** Subcellular localization of candidate O-GlcNAc proteins and spectra.

Within that study, the number of identified spectra and proteins varies significantly between cell lines and may reflect cell-type specific differences of protein expression and O-GlcNAcylation. Interestingly, the analysis of the HeLa deep proteome published by Nagaraj *et al.* [15] also contributed a significant number of exclusive and novel O-GlcNAc proteins, even though the HeLa cell line was also part of the panel analyzed by Geiger *et al.* [14]. A closer inspection of the data revealed that 16 out of the 18 exclusive protein identifications originate from GluC (7 proteins) or LysC digests (9 proteins), underscoring the usefulness of multiple protease digestion for proteomics in general and O-GlcNAc and PTM studies in particular. Interestingly, the only O-GlcNAc protein identified in all studies is the Host cell factor 1, a protein known to be highly O-GlcNAcylated.

It is of note that for ten proteins, the GlcNAc site was assigned to an asparagine residue (N-GlcNAc). Moreover, although O-GlcNAc has been reported for proteins of almost all cellular compartments as well as on extracellular proteins [22], it cannot be precluded that several of the identified ER- and Golgi-resident proteins are early synthesis products of O-GalNAc-type glycans. The subcellular localization of candidate O-GlcNAc proteins is depicted in Figure 3B. For 47 of the identified proteins, the O-GlcNAc modification has been previously reported, while 57 represent novel O-GlcNAc proteins. In addition, for nine of the known O-GlcNAc proteins, this study reports direct evidence for the modification for the first time. Collectively, this data shows that O-GlcNAc modified peptides can be identified from large-scale proteomic data, which makes a point in favour of sharing proteomic data with the scientific community.

O-GlcNAc is less abundant than phosphorylation

The modified and unmodified peptides identified in the present re-analysis of proteomic data enabled us to perform a crude estimation of the frequency and abundance of these modifications on the most abundant modified proteins. The re-analysis of the Geiger *et al*. data (11 cell lines) identified 2,023,960 tandem mass spectra, 6,124 of which correspond to phosphorylated peptides and 454 matched to O-GlcNAc peptides. Hence, the frequency of phospho-spectra is 1 in 334 and the frequency of O-GlcNAc spectra is 1 in 4500 indicating that O-GlcNAc is numerically ~13 fold less frequent than phosphorylation. Clearly, this estimation rests upon the assumption that O-GlcNAcylated peptides are, by and large, identified at the same rate as phosphopeptides from HCD data, which may not necessarily be the case (although probably approximately true). The protein abundance for all eleven cell lines has been expressed as the logarithmic normalized spectral abundance factor [23] (NSAF, Figure 4A). As expected, the detected modified proteins are mostly among the medium to high abundant proteins. Interestingly, but somewhat unexpectedly, the NSAF distributions of O-GlcNAc- and phospho-proteins are quite similar. This clearly indicates that the observed O-GlcNAc- and phospho-proteins are, by and large, equally abundant, but that the O-GlcNAc modification is less frequent. Alternatively, the distribution of peptide precursor intensities (Figure 4B) was used as a proxy for the abundance of the detected (modified) peptides.

The data shows that the distributions of phospho-peptides and ordinary peptides are very similar. In contrast, the distribution of O-GlcNAc peptides is massively skewed towards high intensity proteins indicating that many high abundance proteins are also O-GlcNAc modified and that the site occupancy of the detected peptides is likely significantly higher for O-GlcNAc peptides than for phospho-peptides. To test this hypothesis, the site occupancy of all identified O-GlcNAc and phospho-peptides was estimated via the summed precursor intensities for modified and unmodified peptides.

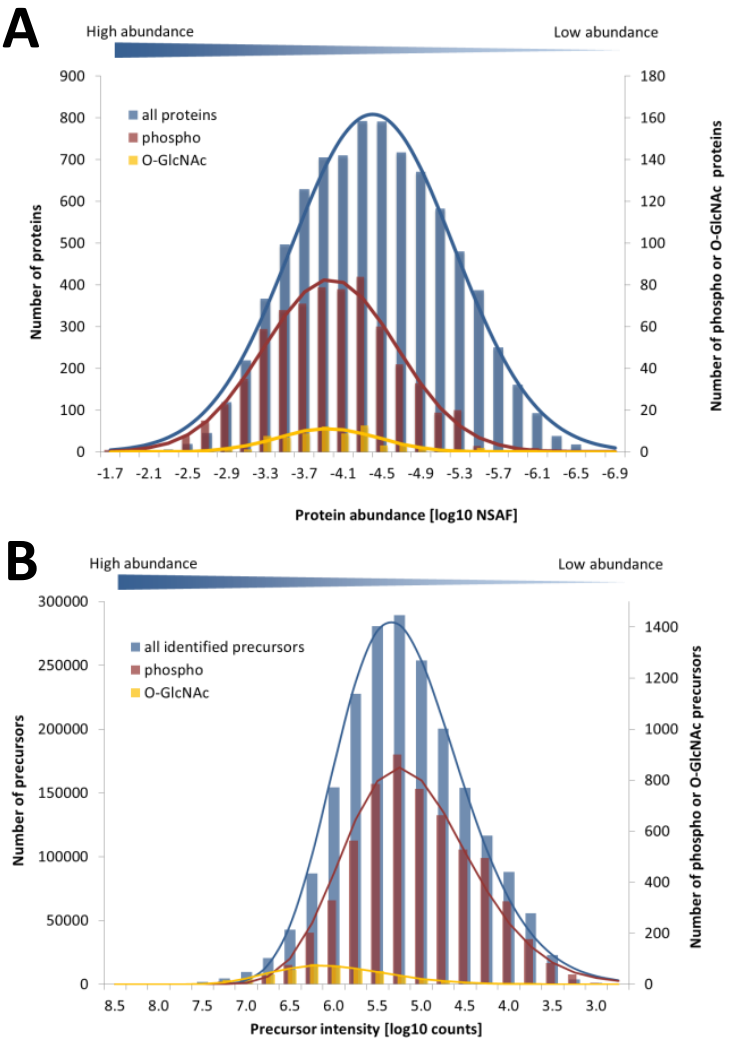

Figure 4 | Protein abundance data for the eleven cell lines analyzed by Geiger *et al.*
A Protein abundance distribution (expressed as logarithmic NSAF) of 76 O-GlcNAc- and 736 phospho-proteins. **B** Protein abundance distribution (expressed as summed peptide precursor intensity) of 454 O-GlcNAc spectra, 6124 phospho-spectra and >2 million unmodified spectra. Note the secondary y-axis for O-GlcNAc- and phospho-identifications.

By this method, an average site occupancy of 0.73 for phospho-peptides and of 0.90 for O-GlcNAc peptides was found. This difference in site occupancy is supported by the fact that the unmodified peptide counterpart could be identified for 46% of all phospho-peptides, but only for 26% of the O-GlcNAc peptides. Clearly, the above estimates are crude because the assumption that the detection efficiencies of modified and unmodified peptides by the employed methods are not grossly different may not be well justified. Still, the data suggests that the O-GlcNAc modification appears to be considerably less frequent than phosphorylation. At the same time, however, the average occupancy of the detected sites appears to be rather high indicating that many of the observed (i. e., abundant) O-GlcNAc proteins are stably modified under physiological conditions. This is consistent with recent *in vitro* data on human O-GlcNAc transferase suggesting that some substrates are constitutively modified [24].

Simultaneous O-GlcNAc/phospho occupancy of proximal sites

Given the potential interplay of O-GlcNAc and phosphorylation [25], it was further investigated whether O-GlcNAc peptide identifications are also possible from large-scale phospho-proteome data. To this end, the Oscore-strategy was employed to identify O-GlcNAc sites from the phospho-proteome of hES and iPS cells [16]. Overall, 107 spectra corresponding to 28 O-GlcNAc-modified peptides and 34 O-GlcNAc sites on 22 proteins (Table 1) could be identified. Of these peptides, 67% were doubly modified with one or multiple O-GlcNAc and phosphate moieties. The identification of O-GlcNAc peptides, which are not phosphorylated, is not surprising given that only around 50% of all identified peptides from the phospho-proteome data harbour phosphorylation sites.

According to common notion, the cross-talk between O-GlcNAc and phosphorylation on identical or proximal sites is extensive and usually referred to as being either antagonistic or synergistic [1]. Most of the reported cases in the literature show competitive occupancy by O-GlcNAc or phosphate of the same or neighbouring residues, and it is argued that the reciprocal exclusion results from either the large size of an O-GlcNAc residue (with an Stokes radius four to five-fold larger than a phosphate moiety) or by the negative charge of the phosphate group or by conformational changes induced by either modification [26]. The observation of 23 doubly modified peptides with a median length of 24 residues suggest that both modifications cannot only occur simultaneously on distal sites of the same protein, but that also proximal residues can be occupied by O-GlcNAc and phosphate simultaneously. A striking example is given by the peptide SEA**pSg(SS)**PPVVTSSSHSR of the SOX2 transcription factor. Here, the tandem mass spectrum localizes the phosphorylation at S4 and the O-GlcNAc modification at either S5 or S6, indicating that both modifications can, at the same time, occur even on (almost) adjacent sites.

Functional roles of novel human O-GlcNAc proteins

Numerous of the novel O-GlcNAc proteins highlight the emerging role of O-GlcNAc as part of the histone code and in the regulation of histone modifications [1 , 27]. Among the novel proteins identified, histone H2B is a particularly interesting case as three O-GlcNAc sites were identified which are in close proximity to (di-)methylation, ubiquitination and phosphorylation sites (Figure 5). O-GlcNAcylation of S113 has, very recently, been reported to facilitate monoubiquitination at K121. Interestingly, here, the O-GlcNAc moiety seems to act as primer for a histone H2B ubiquitin ligase, and monoubiquitination presumably results in transcriptional activation [28]. Although the precise roles of the novel O-GlcNAc sites between T53 and S65 on H2B are unknown, one might speculate about further relationships of O-GlcNAc and ubiquitination.

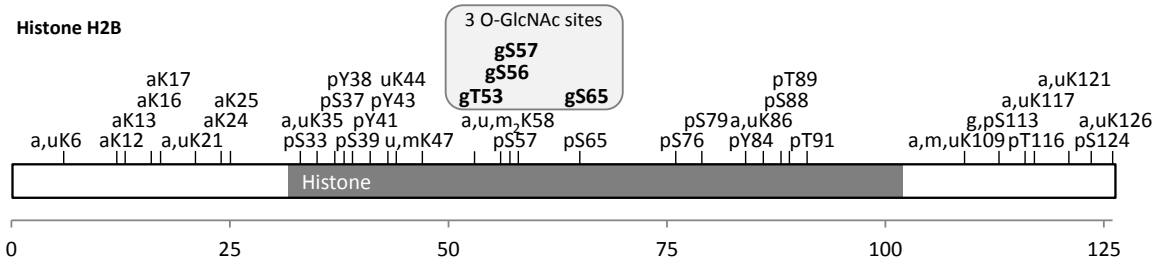

Figure 5 | Graphical representation of post-translational modifications along the sequence of histone H2B
a: acetylation; g: O-GlcNAc; m: methylation; p: phosphorylation; u: ubiquitination

Further noteworthy examples for O-GlcNAc modified proteins include the transcription factors SOX-2 and Sal-like protein 4 (SALL4) as well as STAT3, which have been discovered in the hES and iPS cell proteomes [16]. While SALL4 and SOX-2 have been previously reported to be O-GlcNAc-modified in mouse [19], no site has been determined yet for STAT3 [29]. The STAT3 O-GlcNAc site could be localized between T714 and T721. For SALL4, three novel O-GlcNAc sites have been found: one site between S480 and T501, one site at T608, S609 or S612; and one additional site between T608 and S628. All three proteins are involved in maintaining stem cell identity and governing stem cell-renewal [30, 31] by up-regulating pluripotency genes and down-regulating developmental genes. The discovery of novel O-GlcNAc-modified stem cell transcription factors is in line with the finding that O-GlcNAc transferase might regulate transcription during early development via the modification of proteins required to maintain the embryonic stem cell transcriptional repertoire [19].

Conclusion

The re-analysis of >13 million tandem mass spectra from four large-scale human proteome and phosphoproteome data sets identified several hundred O-GlcNAc modified peptides, most of which have not been reported before. This shows that at least some O-GlcNAc modified proteins are abundant enough so that they can be identified without biochemical enrichment. The current study also makes a point in favor of sharing data between laboratories because one can expect to be able to discover many hundred more modified peptides from the vast quantities of published proteomic data. Interestingly, the number of O-GlcNAc peptides and sites reported in this work is larger than those of most other O-GlcNAc studies which all utilize some form of biochemical enrichment. This may indicate that the development of such enrichment methods is still in its infancy. The fact that the number and abundance of O-GlcNAc peptides identified 'in passing' as it were, is much smaller than those of phosphorylated peptides further highlights the need for the development of better biochemical tools.

Acknowledgments

The author is indebted to Amin Moghaddas Gholami for re-writing the Oscore code and to the originators of the mass spectrometry data used in this study for making this data available to the community.

Abbreviations

O-GlcNAc	O-linked β-N-acetylglucosamine
gS	O-GlcNAc modified serine
gT	O-GlcNAc modified threonine
HCD	higher collision energy dissociation
HexNAc	N-acetylgalactosamine, N-acetylglucosamine
MS	mass spectrometry
NSAF	normalized spectral abundance factor
PSM	peptide spectrum match

References

1. Hart, G. W., Slawson, C., Ramirez-Correa, G., and Lagerlof, O. (2011) Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annu Rev Biochem* 80, 825-858.
2. Hu, P., Shimoji, S., and Hart, G. W. (2010) Site-specific interplay between O-GlcNAcylation and phosphorylation in cellular regulation. *FEBS Lett* 584, 2526-2538.
3. Slawson, C., and Hart, G. W. (2011) O-GlcNAc signalling: implications for cancer cell biology. *Nat Rev Cancer* 11, 678-684.
4. Yang, W. H., Kim, J. E., Nam, H. W., Ju, J. W., Kim, H. S., Kim, Y. S., and Cho, J. W. (2006) Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. *Nat Cell Biol* 8, 1074-1083.
5. Wang, J., Torii, M., Liu, H., Hart, G. W., and Hu, Z. Z. (2011) dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics* 12, 91.
6. Huddleston, M. J., Bean, M. F., and Carr, S. A. (1993) Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests. *Anal Chem* 65, 877-884.
7. Chalkley, R. J., and Burlingame, A. L. (2001) Identification of GlcNAcylation sites of peptides and alpha-crystallin using Q-TOF mass spectrometry. *J Am Soc Mass Spectrom* 12, 1106-1113.
8. Mirgorodskaya, E., Roepstorff, P., and Zubarev, R. A. (1999) Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal Chem* 71, 4431-4436.
9. Vosseller, K., Trinidad, J. C., Chalkley, R. J., Specht, C. G., Thalhammer, A., Lynn, A. J., Snedecor, J. O., Guan, S., Medzihradszky, K. F., Maltby, D. A., Schoepfer, R., and Burlingame, A. L. (2006) O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. *Mol Cell Proteomics* 5, 923-934.
10. Haynes, P. A., and Aebersold, R. (2000) Simultaneous detection and identification of O-GlcNAc-modified glycoproteins using liquid chromatography-tandem mass spectrometry. *Anal Chem* 72, 5402-5410.
11. Chalkley, R. J., and Burlingame, A. L. (2003) Identification of novel sites of O-N-acetylglucosamine modification of serum response factor using quadrupole time-of-flight mass spectrometry. *Mol Cell Proteomics* 2, 182-190.
12. Hahne, H., and Kuster, B. (2011) A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. *J Am Soc Mass Spectrom* 22, 931-942.
13. Zhao, P., Viner, R., Teo, C. F., Boons, G. J., Horn, D., and Wells, L. (2011) Combining high-energy C-trap dissociation and electron transfer dissociation for protein O-GlcNAc modification site assignment. *J Proteome Res* 10, 4088-4104.
14. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* 11, M111 014050.
15. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2012) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7, 548.
16. Phanstiel, D. H., Brumbaugh, J., Wenger, C. D., Tian, S., Probasco, M. D., Bailey, D. J., Swaney, D. L., Tervo, M. A., Bolin, J. M., Ruotti, V., Stewart, R., Thomson, J. A., and Coon, J. J. (2011) Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nat Methods* 8, 821-827.
17. Wang, Z., Udeshi, N. D., Slawson, C., Compton, P. D., Sakabe, K., Cheung, W. D., Shabanowitz, J., Hunt, D. F., and Hart, G. W. (2010) Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Sci Signal* 3, ra2.

18. Chalkley, R. J., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc Natl Acad Sci USA* 106, 8894-8899.
19. Myers, S. A., Panning, B., and Burlingame, A. L. (2011) Polycomb repressive complex 2 is necessary for the normal site-specific O-GlcNAc distribution in mouse embryonic stem cells. *Proc Natl Acad Sci USA* 108, 9490-9495.
20. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40, D261-270.
21. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics* 8, 2759-2769.
22. Matsuura, A., Ito, M., Sakaidani, Y., Kondo, T., Murakami, K., Furukawa, K., Nadano, D., Matsuda, T., and Okajima, T. (2008) O-linked N-acetylglucosamine is present on the extracellular domain of notch receptors. *J Biol Chem* 283, 35486-35495.
23. Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., and Washburn, M. P. (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J Proteome Res* 5, 2339-2347.
24. Shen, D. L., Gloster, T. M., Yuzwa, S. A., and Vocadlo, D. J. (2012) Insights into O-GlcNAc processing and dynamics through kinetic analysis of O-GlcNAc transferase and O-GlcNAcase activity on protein substrates. *J Biol Chem* 287, 15395-15408.
25. Hart, G. W., Housley, M. P., and Slawson, C. (2007) Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* 446, 1017-1022.
26. Chen, Y. X., Du, J. T., Zhou, L. X., Liu, X. H., Zhao, Y. F., Nakanishi, H., and Li, Y. M. (2006) Alternative O-GlcNAcylation/O-phosphorylation of Ser16 induce different conformational disturbances to the N terminus of murine estrogen receptor beta. *Chem Biol* 13, 937-944.
27. Hanover, J. A. (2010) Epigenetics gets sweeter: O-GlcNAc joins the "histone code". *Chem Biol* 17, 1272-1274.
28. Fujiki, R., Hashiba, W., Sekine, H., Yokoyama, A., Chikanishi, T., Ito, S., Imai, Y., Kim, J., He, H. H., Igarashi, K., Kanno, J., Ohtake, F., Kitagawa, H., Roeder, R. G., Brown, M., and Kato, S. (2011) GlcNAcylation of histone H2B facilitates its monoubiquitination. *Nature* 480, 557-560.
29. Whelan, S. A., Lane, M. D., and Hart, G. W. (2008) Regulation of the O-linked beta-N-acetylglucosamine transferase by insulin signaling. *J Biol Chem* 283, 21411-21417.
30. Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947-956.
31. Zhang, J., Tam, W. L., Tong, G. Q., Wu, Q., Chan, H. Y., Soh, B. S., Lou, Y., Yang, J., Ma, Y., Chai, L., Ng, H. H., Lufkin, T., Robson, P., and Lim, B. (2006) Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nat Cell Biol* 8, 1114-1123.

Chapter 4

Discovery of O-GlcNAc-6-phosphate-modified proteins in large-scale phosphoproteomics data

Summary

Phosphorylated O-GlcNAc is a novel posttranslational modification that has so far only been found on the neuronal protein AP180 from the rat (Graham et al., *J. Proteome Res.* 2011, **10,** 2725-33). Upon collision induced dissociation, the modification generates a highly mass deficient fragment ion (m/z 284.0530) that can be used as a reporter for the identification of phosphorylated O-GlcNAc. Using a publically available mouse brain phosphoproteome data set, the recently developed Oscore software was employed to re-evaluate high resolution/high accuracy tandem mass spectra and discovered the modification on 23 peptides corresponding to 11 mouse proteins. The systematic analysis of 220 candidate phosphoGlcNAc tandem mass spectra as well as a synthetic standard enabled the dissection of the major phosphoGlcNAc fragmentation pathways, suggesting that the modification is O-GlcNAc-6-phosphate. Further, the classical O-GlcNAc modification often exists on the same peptides indicating that O-GlcNAc-6-phosphate may biosynthetically arise in two steps involving the O-GlcNAc transferase and a currently unknown kinase. Many of the identified proteins are involved in synaptic transmission and for Ca^{2+}/calmodulin kinase IV, the O-GlcNAc-6-phosphate modification was found in the vicinity of two autophosphorylation sites required for full activation of the kinase suggesting a potential regulatory role for O-GlcNAc-6-phosphate. By re-analyzing mass spectrometric data from human embryonic and induced pluripotent stem cells, this study also identified Zinc finger protein 462 (ZNF462) as the first human O-GlcNAc-6-phosphate modified protein. Collectively, the data suggests that O-GlcNAc-6-phosphate is a general post-translation modification of mammalian proteins with a variety of possible cellular functions.

Introduction

The attachment of N-acetylglucosamine (O-GlcNAc) to serine and threonine residues of nuclear and cytoplasmic proteins is a dynamic post-translational modification with emerging roles in important cellular processes such as transcription, translation, cytokinesis and signaling [1-4]. O-GlcNAcylation has been linked to phosphorylation as both modifications can occupy the same or adjacent sites [2] and a functional relationship of both modifications has been identified in some cases. For instance, the interplay between O-GlcNAcylation and phosphorylation modulates the stability and activity of p53 [5]. However, recent data revealed the frequent co-occurrence of O-GlcNAc and phosphate at proximal sites [6], suggesting the reciprocal regulation by O-GlcNAcylation and phosphorylation may not be a very general mechanism. Moreover, it has also been found that the distribution of O-GlcNAc sites relative to phosphorylation sites is rather random and that the modification rates at sites detected with both modifications are almost equal, indicating that, on a global level, the substrate recognition of both pathways is not interconnected [7].

The identification of O-GlcNAc-modified proteins is typically achieved by combining selective enrichment and liquid chromatography tandem mass spectrometry (LC-MS/MS). In mass spectrometry based proteomics, peptides are usually analysed by some form of collision-induced dissociation (CID). But, owing to the lability of the O-glycosidic bond under typical CID conditions, the direct and simultaneous identification of O-GlcNAc peptides and sites is difficult. Fragment ion spectra of O-GlcNAc peptides are dominated by the sugar fragments and the the GlcNAc oxonium ion cannot be distinguished from other isobaric HexNAc epimers (e. g., GalNAc). Still, the fragment ions generated by the cleavage of the O-glycosidic bond define a highly useful pattern, which significantly facilitates the (automated) discovery of glycopeptides in general and O-GlcNAc peptides in particular even in complex samples [8-14]. The specificity of these diagnostic fragment ions is further increased when identified from high resolution and high mass accuracy tandem MS spectra [14, 15]. To interrogate such data systematically, a simple scoring scheme, termed Oscore, was recently developed which automatically assesses tandem mass spectra for the presence and intensity of O-GlcNAc (HexNAc) diagnostic fragment ions and, in turn, allows ranking spectra according their probability of representing an O-GlcNAc peptide [15]. A combined search strategy using the protein identification software Mascot and the Oscore algorithm enabled the identification of hundreds of O-GlcNAc peptides from large-scale proteome data [6].

Very recently, phosphorylated O-GlcNAc (phosphoGlcNAc) has been identified for the first time on the synapse-specific protein AP180 purified from rat brain [16]. In light of this exciting discovery, this study reports on the extension of the combined Mascot/Oscore approach for the discovery of proteins modified with phosphoGlcNAc. The Oscore has been first adapted for the detection of phosphoGlcNAc and then re-assessed a large-scale phosphoproteomic data set from murine brain [17]. This led to the discovery of 23 phosphoGlcNAc peptides on 11 phosphoGlcNAc proteins. Based on the fragmentation patterns of 220 candidate phosphoGlcNAc spectra and a synthetic standard, O-GlcNAc-6-phosphate could be deduced as the most likely molecular entity. Finally, the re-analysis of a phosphoproteome study of human embryonic (hES) and induced pluripotent stem (iPS) cells [18], revealed evidence for the first time that a human protein may be modified by O-GlcNAc-6-phosphate suggesting that this PTM may exist more generally in mammalian systems.

Experimental procedures

Publically available data

Data from published phosphoproteome-wide studies [17, 18] were downloaded from the Tranche data repository and http://scor.chem.wisc.edu, respectively.

Data analysis

The mass spectrometric data were processed essentially as described [6]. The data processing with Mascot Distiller 2.4.2.0 (Matrix Science, London, UK) was slightly modified to account for the particular fragmentation behaviour of O-GlcNAc-6-phosphate peptides. Briefly, the isotope fitting for fragments below *m/z* 285 was disabled during peak picking, and the Oscore script was adapted to consider diagnostic O-GlcNAc-6-phosphate fragment ion features (Table 1) within a mass tolerance of 10 ppm. For the identification of O-GlcNAc-6-phosphate peptides, the generated peak list files contained only spectra for which an Oscore could be calculated. The peak list files were then searched with Mascot 2.3.0 against the UniProtKB complete mouse proteome (download date 26.10.2010, 73,688 sequences) combined with sequences of common contaminants. In case of the phosphoproteome dataset of hES and iPS cells [18], spectra were searched against a subset database generated with Scaffold 3.3.1 (Proteome Software, Portland, OR) including only protein identifications from the respective full proteome data set (11,288 sequences). Carbamidomethylation of cysteine residues, oxidation of methionine, HexNAc modification of serine, threonine and asparagine residues, phosphoHexNAc modification of serine and threonine residues as well as phosphorylation at serine, threonine and tyrosine residues were taken into account as variable modifications. Where appropriate, 4-plex or 8-plex iTRAQ was set as fixed modification at peptide amino-termini and lysine side chains. Enzyme specificity was set to trypsin with up to two missed cleavage sites. The target-decoy option of Mascot was enabled and peptide mass tolerance was set to 10 ppm and fragment mass tolerance to 0.02 Da. Search results were imported into Scaffold 3.3.1 and filtered for O-GlcNAc and O-GlcNAc-6-phosphate containing peptides. Candidate O-GlcNAc and O-GlcNAc-6-phosphate spectra were inspected and validated manually (see Supplemental Spectra). Ascore-based localization probabilities for phosphosites [19] from the complete mouse brain phosphoproteome data set were calculated with Scaffold PTM 1.1.3 (Proteome Software, Portland, OR).

A list of known human and murine O-GlcNAc proteins and sites was compiled from recent publications [14, 20-22] as well as from the databases dbOGAP [23] and PhosphositePlus [24]. Similarly, phosphosite and other PTM information was retrieved from UniProtKB and PhosphositePlus.

Results and discussion

Identification of O-GlcNAc- and O-GlcNAc-phosphate-modified peptides from mouse brain

A simple O-GlcNAc protein identification strategy which uses the Oscore as a means to re-assess peptide-spectrum-matches (PSMs) from standard database search algorithms led to the identification of hundreds of O-GlcNAc peptides from large-scale proteomic and phosphoproteomic data sets [6]. We hypothesized that it may be possible to identify O-GlcNAc-phosphate modified peptides in a similar way from phosphoproteomic data sets, as these peptides may co-purify with ordinary phosphopeptides during biochemical enrichment and which should exhibit a fragmentation pattern that can be readily discovered using the Oscore algorithm [15]. To this end, a publically available mouse brain phosphoproteomic data set [17] was downloaded, which had been acquired on a dual pressure linear ion trap Orbitrap hybrid mass spectrometer using higher energy collision dissociation (HCD) [25]. In addition, the Oscore configuration was adapted to consider typical O-GlcNAc-phosphate and O-GlcNAc diagnostic fragment ions (see below and Table 1).

Table 1 | Diagnostic HexNAc-phosphate fragments

Fragment	Empirical formula	Prevalence	Calculated m/z	Average observed m/z	R.M.S /ppm
284	$[C_8H_{15}O_8NP]^+$	100%	284.052980	284.0534	3.4
266	$[C_8H_{13}O_7NP]^+$	24%	266.042415	266.0402	4.5
248	$[C_8H_{11}O_6NP]^+$	49%	248.031850	248.0318	2.7
206	$[C_6H_9O_5NP]^+$	49%	206.021285	206.0217	2.6
204	$[C_8H_{14}O_5N]^+$	6%	204.086649	204.0877	2.8
186	$[C_8H_{12}O_4N]^+$	30%	186.076084	186.0764	3.2
168	$[C_8H_{10}O_3N]^+$	77%	168.065519	168.0656	3.4
150	$[C_8H_8O_2N]^+$	45%	150.054955	150.0551	3.6
144	$[C_6H_{10}O_3N]^+$	24%	144.065519	144.0658	3.6
138	$[C_7H_8O_2N]^+$	66%	138.054955	138.0551	2.9
126	$[C_6H_8O_2N]^+$	75%	126.054955	126.0549	2.8

Using the combination of Mascot database searching and Oscore-based evaluation of PSMs, 23 O-GlcNAc-phosphate and 34 O-GlcNAc peptides were identified (Table 2). Figure 1 shows an example for an O-GlcNAc-phosphate modified peptide derived from the SH3 domain containing scaffold protein Shank2. Clearly, O-GlcNAc and O-GlcNAc-phosphate fragments are not readily distinguishable by mass spectrometry from other possible HexNAc(-phosphate) epimers, but represent the most likely explanation for the identified HexNAc-phosphate peptides (see below).

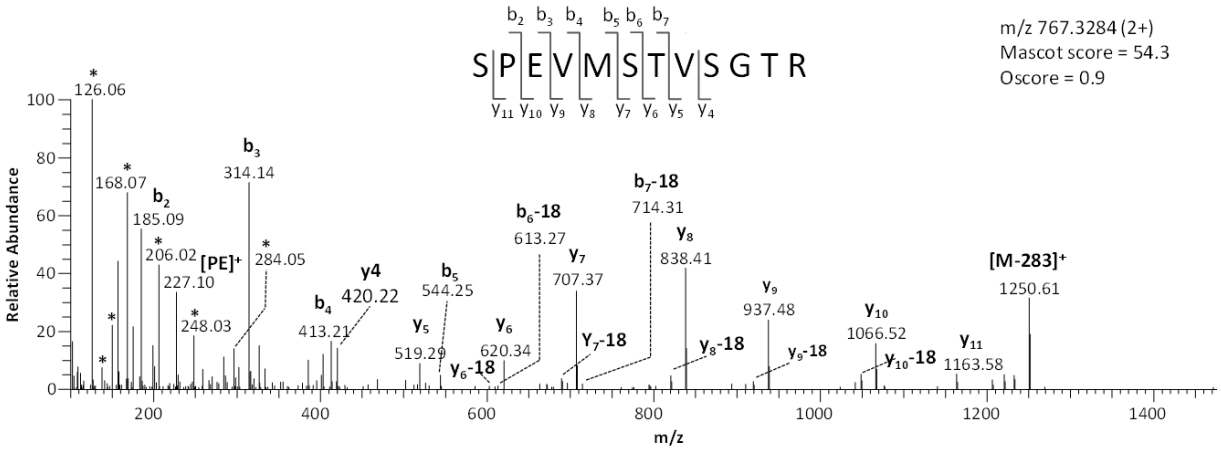

Figure 1 | HCD spectrum of an O-GlcNAc-phosphate peptide corresponding to the sequence SPEVMSTVSGTR of the protein 2 Shank2

Diagnostic fragment ions (*) and a low Oscore indicate that this peptide is O-GlcNAc-phosphate modified. The exact site of modification can however not be deduced with certainty.

Phosphorylation of O-GlcNAc likely occurs at position 6

The low m/z region of HCD spectra contains diagnostic reporter ions common and specific to GlcNAc and GlcNAc-phosphate modified peptides (Table 1). The high mass accuracy of HCD spectra readily allows the identification of the very mass deficient reporter ions containing the phosphate moiety. These ions are strong indicators for O-GlcNAc-phosphate peptides (Table 1, Figure 2A) and the most abundant ion is typically the phosphoHexNAc oxonium ion at m/z 284.0530.

Theoretically, the phosphate moiety may be attached to the 3, 4 or 6 hydroxy group of the sugar. Several lines of evidence however suggest that the phosphate group is attached to position 6. Based on ion count statistics for 220 experimental GlcNAc-phosphate spectra, the major fragmentation pathways (Figure 2) could be dissected. The typical fragmentation routes of a GlcNAc oxonium ion initially proceeds by the thermodynamically driven elimination of two water molecules to form an aromatic ring followed by the loss of a ketene (originating from the acetyl group) and a formaldehyde molecule (originating from the hydroxymethyl group) [15]. For the phosphoGlcNAc oxonium ion, a very similar fragmentation pattern can be formulated. First, two water molecules involving the hydroxyls at positions 3 and 4 are eliminated to form the aromatic species (m/z 248). From here, the fragmentation pathway branches. The m/z 248 species can lose HPO_3 from position 6 to form an aromatic oxonium ion (m/z 168). This ion can further eliminate a ketene or formaldehyde molecule, resulting in abundant signals at m/z 138 and 126, respectively. In an alternative pathway, the m/z 248 ion retains the phosphate moiety and gives rise to a fragment at m/z 206 which subsequently loses the HPO_3 group to generate the 126 ion. In contrast, a presumed GlcNAc-3-phosphate or GlcNAc-4-phosphate would be expected to fragment differently. To form an aromatic system, such molecules would have to lose the phosphate group along with water. Elimination of HPO_3 or H_3PO_4 would thus result in signals at m/z 204 and 186 respectively. However, these ions are barely observed. Further evidence for the proposed molecular structure of the modification comes from the analysis of a synthetic GlcNAc-6-phosphate standard. The fragmentation pathways deduced from the experimental phosphoGlcNAc peptide spectra (Figure 2A, upper spectrum) are exactly mirrored in HCD spectra of the GlcNAc-6-phosphate standard (Figure 2A, lower spectrum). It,

therefore, appears very likely that the phosphorylation is localized to the 6-position of the GlcNAc moiety.

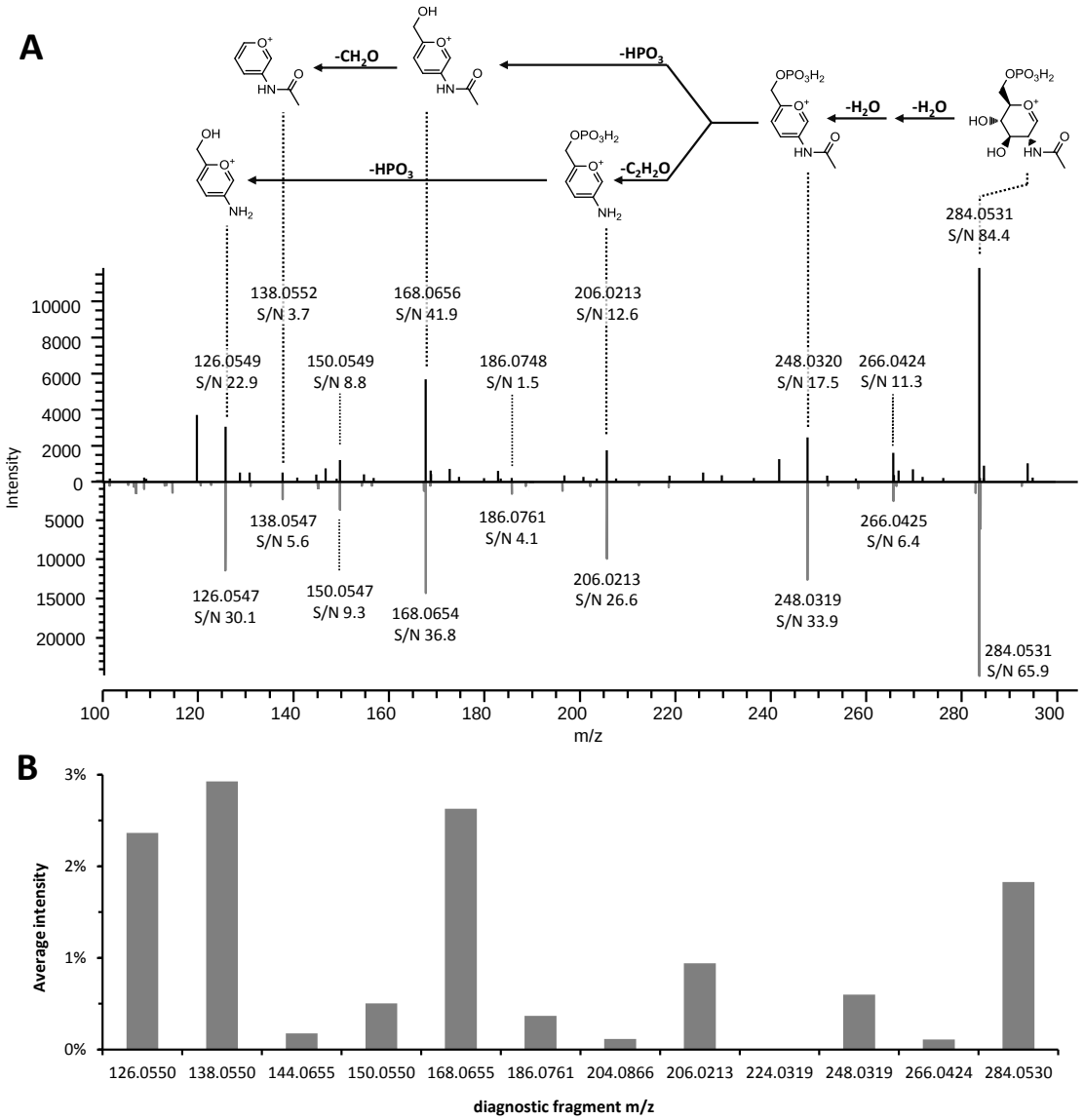

Figure 2 | Major fragmentation routes of GlcNAc-phosphate
A Upper spectrum: Low m/z region of a GlcNAc-phosphate peptide with the proposed structures and fragmentation of diagnostic ions. Lower spectrum: HCD spectrum of a synthetic GlcNAc-6-phosphate standard. The two fragmentation spectra are virtually identical supporting the assignment of GlcNAc-6-phosphate. **B** Average ion counts for the most abundant diagnostic fragment ions from 220 experimental GlcNAc-phosphate spectra. The intensity distribution of the fragment ions is consistent with the two proposed fragmentation pathways of GlcNAc-6-phosphate and intensity distribution of the synthetic standard.

The occurrence of O-GlcNAc-6-phosphate is closely linked to that of O-GlcNAc

Overall, 23 O-GlcNAc-6-phosphate peptides from 11 proteins along with 34 O-GlcNAc peptides from 25 proteins were identified (Table 2). For six of the O-GlcNAc peptides, the modified residues could be deduced from the HCD spectra. Unfortunately, this was only possible for a single O-GlcNAc-6-phosphate peptide, a shortcoming that will have to be addressed by ETD measurements in the future. Still, the identified O-GlcNAc-6-phosphate peptides include Thr-310 of clathrin coat assembly protein AP180 which is the only protein and only site reported thus far [16].

Table 2 | O-GlcNAc-6-phosphate modified proteins identified in this study

Protein (UniProt ID)	GlcNAc-6-phosphate Spectra	GlcNAc-6-phosphate Peptides	O-GlcNAc Spectra	O-GlcNAc Peptides	Other evidence	O-GlcNAc-6-phosphate site (range)
Piccolo (Q9QYX7)	11	6	1	1	dbOGAP	T2352-T2368; S2851-T2860; S2930-S2939; S2953-T2964; S3018-S3035; T3873-T3875
Bassoon (O88737)	9	6	14	5	dbOGAP	S1369-S1378; T1384-T1399; S1649-S1655; S2493; T2905-S2930; T2940-T2945
Serine/threonine-protein kinase WNK2 (Q3UH66)	6	2			dbOGAP*	T1514-T1540; T1514-S1553
Catenin δ-2 (O35927)	3	2	12	3	dbOGAP	S307-S324; T329-T359
Clathrin coat assembly protein AP180 (Q61548)	7	1			Yao *et al.* [26]*	S305-T317; T310 reported previously [16]
TOM1-like protein 2 (Q5SRX1)	4	1				T187-S218
Myocardin-related transcription factor B (P59759)	3	1	3	1	PhosphositePlus	S207-S225
Plakophilin-4 (Q68FH0)	2	1				S1086-S1099
CaMKIV (P08414)	1	1			dbOGAP*	T5-S33
Shank2 (Q80Z38)	1	1			dbOGAP	S1286-T1296
Unc-51-like kinase 2 (Q9QY01)	1	1	1	1		T727-T741
ZNF-462 [#] (Q96JM2)	1	1				S292-S309

* reported in human cells, [#] identified in phosphoproteome of hES and iPS cells [18]

Interestingly, the occurrence of O-GlcNAc-6-phosphate appears to be closely related to that of O-GlcNAc. Five out of the eleven O-GlcNAc-6-phosphate proteins identified here were also identified to contain O-GlcNAc-modified peptides. Another four O-GlcNAc-6-phosphate proteins harbour reported O-GlcNAcylation sites in mouse or human. In particular, the proteins Basson and Piccolo both of which are known to be highly O-GlcNAc-modified [21] were identified with six unique O-GlcNAc-6-phosphate peptides each. Similar observations can be made at the peptide level. Four peptides were identified in their respective O-GlcNAc-6-phosphate and O-GlcNAc forms, and six additional O-GlcNAc-6-phosphate peptides overlap with reported O-GlcNAc sites. Notably, O-GlcNAc as well as O-GlcNAc-6-phosphate modified peptides contain, on average, 6.5 serine or threonine residues, which is considerably higher than the 3.3 for phosphopeptides and the 1.5 for unmodified peptides [6], suggesting that O-GlcNAc-6-phosphate also preferentially occurs in protein regions of low compositional complexity. Taken together, almost 50% of all O-GlcNAc-6-phosphate peptides can also be found as O-GlcNAc-modified peptides, indicating that the biosynthetic route to GlcNAc-6-phosphate may proceed in two distinct steps. Presumably, the O-GlcNAc transferase (OGT) first

attaches the O-GlcNAc moiety to a target residue followed by a second step, in which a yet unknown kinase (possibly GlcNAc kinase, NAGK) phosphorylates the O-GlcNAc moiety at position 6. Given that GlcNAc-6-phosphate is also a precursor for UDP-GlcNAc [27], it cannot be precluded that OGT may accept a putative UDP-GlcNAc-6-phosphate as a substrate and thus directly transfers GlcNAc-6-phosphate to a protein target.

The occurrence of O-GlcNAc-6-phosphate does not only exhibit strong links to O-GlcNAc but also to phosphorylation. All O-GlcNAc-6-modified proteins identified here are known phosphoproteins and were indeed found to be phosphorylated in the present study. In fact, seven O-GlcNAc-6-phosphate-modified peptides were found, which are also modified with one or two phosphates on the same peptide and further two peptides, which overlap with phosphosites identified in the complete phosphoproteome data set.

O-GlcNAc-6-phosphate modified proteins in mouse brain

The proteins identified in this study to be O-GlcNAc-6-phosphate modified are of quite high abundance in mouse brain and may thus only represent the 'tip of the iceberg' of all proteins that may carry the modification. Interestingly though, many of the identified proteins serve multiple important functions in neurons, ranging from regulation of sodium/potassium-coupled chloride cotransporters (the serine threonine kinase WNK2) to the regulation of neurotransmitter release and retrieval (Bassoon, Piccolo, AP180, TOM1-like protein 2) as well as postsynaptic structural organization (the SH3 domain containing scaffold protein Shank2). Maybe the most striking finding is that CaMKIV can be O-GlcNAc-6-phosphate modified. CaMKIV belongs to the Ca^{2+}/calmodulin kinase signal cascade, which, in the nervous system, exerts key functions in signal transduction, gene transcription, synaptic plasticity, and behavior [28].

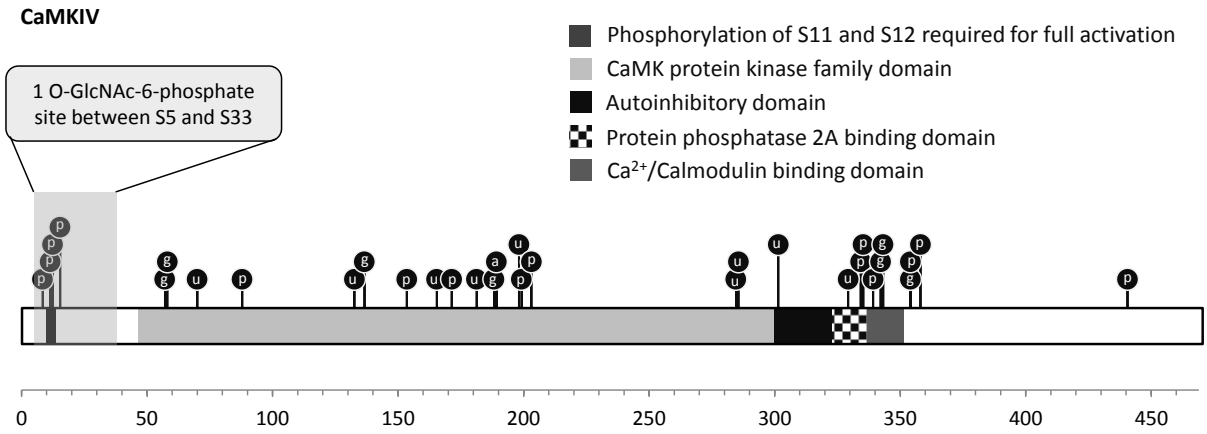

Figure 3 | Schematic representation of the post-translational modifications and protein domains of CaMKIV
The different posttranslational modifications are denoted as follows: a: acetylation; g: O-GlcNAc; p: phosphorylation; u: ubiquitination. O-GlcNAc-6-phosphate is located at a serine or threonine residue between Ser-5 and Ser-33 where also several phosphorylation sites are localized.

Located in the nucleus, CaMKIV directly and indirectly regulates the activity of several important transcription factors, including CREB. In addition, CaMKIV is highly O-GlcNAcylated, and its activity towards CREB can be reciprocally modulated by phosphorylation and O-GlcNAcylation of adjacent sites in the active site [29]. Further evidence suggests that CaMKIV specifically phosphorylates and activates OGT upon depolarization of neuronal cells, suggesting that OGT is a downstream target of

CaMKIV and activates the transcription factor AP-1 (also an O-GlcNAc protein [30, 31]). As depicted in Figure 3, the O-GlcNAc-6-phosphate modified residue is located between Thr-5 and Ser-33 at the amino-terminus of CaMKIV and the identified peptide overlaps with two serine residues (Ser-11 and Ser-12 in murine CamKIV) which are autophosphorylated and required for full activation of CaMKIV. CaMKIV regulation is complex and involves Ca^{2+}/calmodulin binding, phosphorylation in its activation loop as well as autophosphorylation [32, 33]. The modification with O-GlcNAc-6-phosphate in this region may therefore represent a novel, potentially functional feature in the regulation of CaMKIV. Clearly, future work needs to address if or to which extent the modification is functionally relevant for any of the identified proteins in general or during CaMK signalling in particular.

ZNF-462 is the first human O-GlcNAc-6-phosphate modified protein

Mouse brain is well known to contain relatively high O-GlcNAc levels, and given the apparent relationship between O-GlcNAc-6-phosphate and O-GlcNAc, the discovery of O-GlcNAc-6-phosphate proteins from this biological source may not be too surprising. In order to investigate if O-GlcNAc-6-phosphate is also found on human proteins, a number of published data sets obtained from different cancer and stem cell lines representing more than 36,000 phosphopeptides [18, 34, 35] were re-analyzed. From the hES and iPS cell line data [18], numerous O-GlcNAc proteins were identified [6] and O-GlcNAc-6-phosphate was identified on Zinc finger protein 462 (ZNF-462). Unfortunately, no clear function has been assigned to this protein yet (other than DNA binding and a putative involvement in transcriptional regulation), hence, precludes any speculation about any functional significance of its modification by O-GlcNAc-6-phosphate. However, the first identification of an O-GlcNAc-6-phosphate protein from human clearly indicates that the modification not only exists in rodents (and brain for that matter), but possibly represents a general novel post-translation protein modification in mammalian cells with potential functional significance.

Conclusion

The Oscore-based re-assessment of high resolution tandem mass spectra from published phospho-proteomic studies enabled the identification of 12 O-GlcNAc-6-phosphate modified proteins, including the first human O-GlcNAc-6-phosphate modified protein. This shows that O-GlcNAc-6-phosphate is not a singular protein modification [16] and that it is sufficiently stable and abundant to be detected in the presence of tens of thousands of phosphopeptides. Thus it can be expected that mining phosphoproteomic data will substantially increase the number of proteins that can be modified in this way. Still, more efficient biochemical enrichment tools as well as MS techniques such as ETD that preserves the modification will likely be required for the proteome-wide investigation of O-GlcNAc-6-phosphate in the future. In addition to merely enumerating modified peptides, the identification of the corresponding O-GlcNAc kinase(s) as well as potentially involved phosphatases will clearly be important steps towards a basic understanding of this novel post-translational modification.

Acknowledgments

The author is indebted to the originators of the mass spectrometry data used in this study for making their data available to the community.

Abbreviations

HCD	higher collision energy dissociation
HexNAc	N-acetylgalactosamine, N-acetylglucosamine
LC-MS/MS	Liquid chromatography – tandem mass spectrometry
MS	mass spectrometry
O-GlcNAc	O-linked β-N-acetylglucosamine
OGT	O-GlcNAc transferase
PhosphoGlcNAc	phosphorylated O-GlcNAc
PSM	peptide-spectrum-match

References

1. Hart, G. W., Housley, M. P., and Slawson, C. (2007) Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. Nature 446, 1017-1022.
2. Hart, G. W., Slawson, C., Ramirez-Correa, G., and Lagerlof, O. (2011) Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. Annu Rev Biochem 80, 825-858.
3. Hu, P., Shimoji, S., and Hart, G. W. (2010) Site-specific interplay between O-GlcNAcylation and phosphorylation in cellular regulation. FEBS Lett 584, 2526-2538.
4. Hanover, J. A. (2010) Epigenetics gets sweeter: O-GlcNAc joins the "histone code". Chem Biol 17, 1272-1274.
5. Yang, W. H., Kim, J. E., Nam, H. W., Ju, J. W., Kim, H. S., Kim, Y. S., and Cho, J. W. (2006) Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. Nat Cell Biol 8, 1074-1083.
6. Hahne, H., Gholami, A. M., and Kuster, B. (2012) Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. Mol Cell Proteomics [epub ahead of print 2012/06/05].
7. Trinidad, J. C., Barkan, D. T., Gulledge, B. F., Thalhammer, A., Sali, A., Schoepfer, R., and Burlingame, A. L. (2012) Global identification and characterization of both O-GlcNAcylation and phosphorylation at the murine synapse. Mol Cell Proteomics [epub ahead of print 2012/05/31].
8. Chalkley, R. J., and Burlingame, A. L. (2001) Identification of GlcNAcylation sites of peptides and alpha-crystallin using Q-TOF mass spectrometry. J Am Soc Mass Spectrom 12, 1106-1113.
9. Haynes, P. A., and Aebersold, R. (2000) Simultaneous detection and identification of O-GlcNAc-modified glycoproteins using liquid chromatography-tandem mass spectrometry. Anal Chem 72, 5402-5410.
10. Chalkley, R. J., and Burlingame, A. L. (2003) Identification of novel sites of O-N-acetylglucosamine modification of serum response factor using quadrupole time-of-flight mass spectrometry. Mol Cell Proteomics 2, 182-190.
11. Vosseller, K., Trinidad, J. C., Chalkley, R. J., Specht, C. G., Thalhammer, A., Lynn, A. J., Snedecor, J. O., Guan, S., Medzihradszky, K. F., Maltby, D. A., Schoepfer, R., and Burlingame, A. L. (2006) O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. Mol Cell Proteomics 5, 923-934.
12. Ozohanics, O., Krenyacz, J., Ludanyi, K., Pollreisz, F., Vekey, K., and Drahos, L. (2008) GlycoMiner: a new software tool to elucidate glycopeptide composition. Rapid Commun Mass Spectrom 22, 3245-3254.
13. Pompach, P., Chandler, K. B., Lan, R., Edwards, N., and Goldman, R. (2012) Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. J Proteome Res 11, 1728-1740.
14. Zhao, P., Viner, R., Teo, C. F., Boons, G. J., Horn, D., and Wells, L. (2011) Combining high-energy C-trap dissociation and electron transfer dissociation for protein O-GlcNAc modification site assignment. J Proteome Res 10, 4088-4104.
15. Hahne, H., and Kuster, B. (2011) A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. J Am Soc Mass Spectrom 22, 931-942.
16. Graham, M. E., Thaysen-Andersen, M., Bache, N., Craft, G. E., Larsen, M. R., Packer, N. H., and Robinson, P. J. (2011) A novel post-translational modification in nerve terminals: O-linked N-acetylglucosamine phosphorylation. J Proteome Res 10, 2725-2733.
17. Jedrychowski, M. P., Huttlin, E. L., Haas, W., Sowa, M. E., Rad, R., and Gygi, S. P. (2011) Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. Mol Cell Proteomics 10, M111 009910.
18. Phanstiel, D. H., Brumbaugh, J., Wenger, C. D., Tian, S., Probasco, M. D., Bailey, D. J., Swaney, D. L., Tervo, M. A., Bolin, J. M., Ruotti, V., Stewart, R., Thomson, J. A., and Coon, J. J. (2011)

Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat Methods 8, 821-827.
19. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24, 1285-1292.
20. Wang, Z., Udeshi, N. D., Slawson, C., Compton, P. D., Sakabe, K., Cheung, W. D., Shabanowitz, J., Hunt, D. F., and Hart, G. W. (2010) Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. Sci Signal 3, ra2.
21. Chalkley, R. J., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. Proc Natl Acad Sci USA 106, 8894-8899.
22. Myers, S. A., Panning, B., and Burlingame, A. L. (2011) Polycomb repressive complex 2 is necessary for the normal site-specific O-GlcNAc distribution in mouse embryonic stem cells. Proc Natl Acad Sci USA 108, 9490-9495.
23. Wang, J., Torii, M., Liu, H., Hart, G. W., and Hu, Z. Z. (2011) dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. BMC Bioinformatics 12, 91.
24. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 40, D261-270.
25. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics 8, 2759-2769.
26. Yao, P. J., and Coleman, P. D. (1998) Reduced O-glycosylated clathrin assembly protein AP180: implication for synaptic vesicle recycling dysfunction in Alzheimer's disease. Neurosci Lett 252, 33-36.
27. Vocadlo, D. J., Hang, H. C., Kim, E. J., Hanover, J. A., and Bertozzi, C. R. (2003) A chemical approach for identifying O-GlcNAc-modified proteins in cells. Proc Natl Acad Sci USA 100, 9116-9121.
28. Wayman, G. A., Lee, Y. S., Tokumitsu, H., Silva, A. J., and Soderling, T. R. (2008) Calmodulin-kinases: modulators of neuronal development and plasticity. Neuron 59, 914-931.
29. Dias, W. B., Cheung, W. D., Wang, Z., and Hart, G. W. (2009) Regulation of calcium/calmodulin-dependent kinase IV by O-GlcNAc modification. J Biol Chem 284, 21327-21337.
30. Tai, H. C., Khidekel, N., Ficarro, S. B., Peters, E. C., and Hsieh-Wilson, L. C. (2004) Parallel identification of O-GlcNAc-modified proteins from cell lysates. J Am Chem Soc 126, 10500-10501.
31. Song, M., Kim, H. S., Park, J. M., Kim, S. H., Kim, I. H., Ryu, S. H., and Suh, P. G. (2008) o-GlcNAc transferase is activated by CaMKIV-dependent phosphorylation under potassium chloride-induced depolarization in NG-108-15 cells. Cell Signal 20, 94-104.
32. Anderson, K. A., Means, R. L., Huang, Q. H., Kemp, B. E., Goldstein, E. G., Selbert, M. A., Edelman, A. M., Fremeau, R. T., and Means, A. R. (1998) Components of a calmodulin-dependent protein kinase cascade. Molecular cloning, functional characterization and cellular localization of Ca2+/calmodulin-dependent protein kinase kinase beta. J Biol Chem 273, 31880-31889.
33. Soderling, T. R. (1999) The Ca-calmodulin-dependent protein kinase cascade. Trends Biochem Sci 24, 232-236.
34. Hennrich, M. L., Groenewold, V., Kops, G. J., Heck, A. J., and Mohammed, S. (2011) Improving depth in phosphoproteomics by using a strong cation exchange-weak anion exchange-reversed phase multidimensional separation approach. Anal Chem 83, 7137-7143.

35. Hennrich, M. L., van den Toorn, H. W., Groenewold, V., Heck, A. J., and Mohammed, S. (2012) Ultra acidic strong cation exchange enabling the efficient enrichment of basic phosphopeptides. Anal Chem 84, 1804-1808.

Chapter 5

Proteome wide purification and identification of O-GlcNAc modified proteins using Click chemistry and mass spectrometry

Summary

The posttranslational modification of proteins with N-acetylglucosamine (O-GlcNAc) is involved in the regulation of a wide variety of cellular processes and associated with a number of chronic diseases. Despite its emerging biological significance, the systematic identification of O-GlcNAc proteins is still challenging. The present study introduces a novel O-GlcNAc protein enrichment procedure, which exploits metabolic labelling of cells by azide modified GlcNAc and copper mediated Click chemistry for purification of modified proteins on an alkyne-resin. On-resin proteolysis using trypsin followed by LC-MS/MS afforded the identification of >1500 O-GlcNAc proteins from a single cell line. Subsequent elution of covalently resin bound O-GlcNAc peptides using selective β-elimination enabled the identification of 125 O-GlcNAc modification sites on 80 proteins. To demonstrate the practical utility of the developed approach, the global effects of the O-GlcNAcase inhibitor GlcNAcstatin G was studied on the level of O-GlcNAc modification of cellular proteins. About 200 proteins including several key players involved in the hexosamine signalling pathway showed significantly increased O-GlcNAcylation levels in response to the drug which further strengthens the link of O-GlcNAc protein modification to cellular nutrient sensing and response.

Introduction

O-linked N-acetylglucosamine (O-GlcNAc) is an emerging dynamic posttranslational modification (PTM) of serine and threonine residues of proteins and was first discovered in 1984 by Torres and Hart [1]. Since then, O-GlcNAc has been found on a wide range of nuclear and cytoplasmic proteins involved in almost all cellular processes including signalling, cell cycle regulation, transcription and translation regulation, protein trafficking and protein quality control, as well as stress and survival [2-4]. The evolutionary conserved enzyme O-GlcNAc transferase (OGT) catalyzes the attachment of GlcNAc from uridine diphosphate N-acetylglucosamine (UDP-GlcNAc) to specific protein residues. This reaction can be reversed by the nuclear and cytoplasmic O-GlcNAcase (OGA) which is a bifunctional protein with an N-terminal glycosidase domain, a C-terminal histone acetyltransferase domain [5] and an interjacent caspase-3 cleavage site [6]. The O-GlcNAcase activity of OGA can be inhibited using small molecule inhibitors resulting in a dramatic increase in overall cellular O-GlcNAc levels [7]. OGA inhibitors have been extensively used to study the role of O-GlcNAc in nutrient-sensing and diabetes, protection from cellular stress or cellular signalling and the interplay of O-GlcNAcylation and phosphorylation [8].

Despite the emerging relevance of O-GlcNAc in a multitude of cellular processes, the systematic discovery of O-GlcNAc proteins is still challenging. Like for many other PTMs, liquid chromatography tandem mass spectrometry (LC-MS/MS) is the method of choice and abundant O-GlcNAc proteins can sometimes be identified directly from full proteome digests [9]. However, this task is more commonly achieved by adding a selective enrichment step. A number of different conceptual approaches have been developed for this purpose both at the level of modified peptides and proteins. For proteome wide applications, the most promising technologies so far are based on lectin affinity chromatography using wheat germ agglutinin [10, 11] and a chemoenzymatic approach that tags endogenous O-GlcNAc moieties with azide-labeled galactose and allows Click chemistry-based enrichment of the tagged proteins using a photocleavable biotin probe [12, 13]. The conceptual advantage of the above approaches are that they leave the modification on the peptide and thus, in principle, allow not only the identification of O-GlcNAc peptides and proteins but also the direct determination of the site of modification within the peptide or protein. Direct site determination is, however often complicated by the fact that the modification is labile during the standard mass spectrometric readout of *collision induced dissociation* (CID) and therefore the site information is frequently lost. This shortcoming can, in principle, be overcome by the use of *electron capture dissociation* (ECD) or *electron transfer dissociation* (ETD) mass spectrometry [10, 14, 15] but these techniques also have shortcomings, notably a rather poor overall sensitivity. As a result, alternative strategies that resort to semi-direct or even indirect measures of modification identification and site localisation have been developed.

For example several groups have employed metabolic labeling of O-GlcNAc proteins by azide or alkyne-tagged N-acetylglucosamine [16] (GlcNAz and GlcNAlk, respectively) and subsequent coupled the modified proteins to an affinity probe via copper-catalyzed azide/alkyne Click chemistry (CuAAC) or Staudinger ligation. The affinity enriched O-GlcNAc proteins can then be identified by mass spectrometry [17-20]. However these approaches did not enable the direct identification of a single O-GlcNAc site, hence, rendering the information regarding the O-GlcNAc modification rather indirect. As an alternative β-elimination of O-GlcNAc moieties followed by Michael addition (BEMAD)

has been employed for the enrichment and site identification of O-GlcNAc proteins [10, 21, 22]. In the BEMAD approach, O-GlcNAc moieties are eliminated under strong alkaline conditions resulting in an α,β-unsaturated carbonyl group (a so-called Michael system), which can subsequently be modified using a strong nucleophile. The addition of a stable nucleophile tags the former O-GlcNAc site which can be then recognized in the MS experiment. The BEMAD approach has been used frequently and has enabled the identification and quantification of numerous rodent brain proteins along with their sites [10, 21, 22]. A clear downside of the BEMAD approach is that phosphorylated and, to a lesser extent, unmodified serine, threonine and cysteine residues are also susceptible to β-elimination under certain experimental conditions [23, 24], necessitating additional means to control false-positive O-GlcNAc site assignments.

The present study utilizes elements of the above biochemical methods (notably metabolic GlcNAz labeling, Click chemistry, on resin proteolysis, selective β-elimination) and combine them in a novel way that when complemented with rigorous label-free quantification allowed the identification of >1,500 high confidence O-GlcNAc modified proteins from a single cell line along with >100 modification sites. Furthermore, the practical utility of the developed approach was demonstrated by studying the effect of the OGA inhibitor GlcNAcstatin G on the O-GlcNAc proteome which led to the identification of several key signaling proteins.

Experimental procedures

Peptide synthesis and assessment of β-elimination/Michael addition conditions

O-GlcNAc- and phosphopeptides for the systematic assessment of β-elimination/Michael addition conditions were synthesized in our laboratory using standard solid phase peptide synthesis [15, 25]. Beta-elimination reactions were performed on dried peptides using 1% triethylamine and 0.1% NaOH in 20% ethanol at different temperatures and for various amounts of time [21]. In addition, β-elimination was performed using the GlycoProfile β-elimination kit (Sigma-Aldrich, Taufkirchen Germany) according to the manufacturer's instructions. Michael addition was performed using β-mercaptoethanol, dithiothreitol or 1-propanethiol at different reagent concentrations (Table 1). The β-elimination/Michael addition reaction was quenched with 1% trifluoroacetic acid (TFA). Peptides were dried *in vacuo*, desalted using C_{18} StageTips [26], and reconstituted in 20 µl 0.1% formic acid (FA) prior to LC-MS/MS analysis.

Cell culture, metabolic labeling and inhibitor treatment

HEK293 cells were cultured in Dulbecco's modified Eagle's medium (DMEM; PAA, Pasching, Austria) containing 1.0 g/L glucose supplemented with 10% (v/v) fetal bovine serum (FBS; PAA, Pasching, Austria) at 37 °C with humidified air and 5% CO_2. For metabolic labeling, HEK293 cells were treated with 200 µM tetraacetylated GlcNAz (Ac_4GlcNAz; Life Technologies, Eugene, OR) for 18 hours. In case of GlcNAcstatin G-treated cells, HEK293 cells were metabolically labelled and treated with 20 µM GlcNAcstatin G for two hours before cell lysis [27].

Click chemistry-based enrichment of O-GlcNAc proteins

Cell lysis and Click chemistry-based enrichment were performed using the Click-iT protein enrichment kit according to the manufacturer's instructions with two relevant modifications. An ultracentrifugation step was incorporated to clear the lysate from fine insoluble matter and an

additional wash step with the strong copper chelator diethylene triamine pentaacetic acid (DPTA) after CuAAC. In order to control the selectivity of the O-GlcNAc protein purification procedure, the procedure was performed in parallel with GlcNAz-labelled and unlabelled HEK293 cells using the same amount of protein starting material for the CuAAC enrichment. Briefly, following cell lysis by sonification in a urea buffer (8 M urea, 200 mM TrisHCl pH8, 4% CHAPS, 1 M NaCl), cell debris were pelleted (10,000 x g, 15 min, 4 °C) and the samples were subjected to ultracentrifugation (145,000 x g, 60 min, 4 °C). Three mg of total protein in 800 µl lysis buffer was alkylated with 10 mM iodo acetamide (IAM, 60 min, room temperature) and used for the subsequent CuAAC reaction according to the manufacturer's instructions using 200 µl of the alkyne resin slurry. After overnight Click reaction, the resin was washed 3x with 1.5 ml 10 mM DPTA before proteins were reduced (10 mM dithiothreitol, 30 min, 55 °C) and alkylated (50 mM IAM, 60 min, room temperature). Following extensive washing with 5x 2 ml SDS wash buffer (100 mM TrisHCl pH 8, 1% SDS, 250 mM NaCl, 5 mM EDTA), 5x 2 ml urea buffer (8 M urea, 100 mM TrisHCl pH 8) and 5x 2 ml 20% ACN, the resin-bound proteins were digested in 50 mM TrisHCl (pH 7.6) in two steps, first for two hours and then overnight, using each time 0.5 µg trypsin. After on-resin digestion, the supernatant was aspirated and desalted using C_{18} StageTips before LC-MS/MS analysis.

On-resin dephosphorylation and elution of O-GlcNAc peptides by β-elimination

Following on-resin digestion, the resin was washed twice with 1.8 ml ddH_2O and once with 1.8 ml dephosphorylation buffer (50 mM TrisHCl pH 7.6, 100 mM NaCl, 1 mM DTT, 10 mM $MgCl_2$, 1 mM $MnCl_2$). Dephosphorylation was performed at 37 °C for 6 hours in 400 µl using 800 U λ phosphatase and 20 U calf intestine phosphatase (New England Biolabs, Frankfurt a. M., Germany). Following dephosphorylation, the resin was washed twice with 1.8 ml ddH_2O and the slurry volume was adjusted to 300 µl with ddH_2O before β-elimination using the GlycoProfile β-elimination kit. The β-elimination reaction was incubated on an end-over-end shaker with extensive mixing at 4 °C and quenched after 24 hours with 1% TFA. β-eliminated peptides were desalted and concentrated with C_{18} StageTips before LC-MS/MS analysis.

Liquid chromatography and mass spectrometry

Mass spectrometry was performed on an LTQ Orbitrap XL ETD or an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Germany) connected to a nanoLC Ultra 1D+ liquid chromatography system (Eksigent, CA) using an in-house packed precolumn (20 mm x 75 µm ReproSil-Pur C18, Dr. Maisch, Germany) and analytical column (400 mm x 50 µm ReproSil-Pur C18, Dr. Maisch, Germany). The mass spectrometer was equipped with a nano-electrospray ion source (Proxeon Biosystems, DK) and the electrospray voltage was applied *via* a liquid junction. The mass spectrometer was operated in data-dependent mode and all measurements were performed in positive ion mode. Intact peptide mass spectra were acquired at a resolution of 60,000 (at m/z 400) and an automatic gain control (AGC) target value of 10^6, followed by fragmentation of the most intense ions by collision-induced dissociation (CID; LTQ Orbitrap XL ETD) or higher energy-collision induced dissociation (HCD; LTQ Orbitrap Velos). Full scans were acquired in profile mode, whereas all tandem mass spectra were acquired in centroid mode. CID was performed for up to 15 MS/MS (2 h gradient) or 8 MS/MS (4 h gradient) per full scan with 35% normalized collision energy (NCE), and an AGC target value of 5000. HCD was performed for up to 10 MS/MS per full scan with 40% NCE, and an AGC target value of 35,000. Singly charged ions and ions without assigned charge state were excluded from fragmentation, and fragmented precursor ions were dynamically excluded (2h

gradient: 10 sec; 4h gradient: 30 sec). Internal calibration was performed using the poly-siloxane ion signal at m/z 445.1200 present in ambient laboratory air.

Protein identification and quantification

Protein identification and intensity based label free quantification from on-resin digestion experiments was achieved with Mascot version 2.3.02 (Matrix Science, UK) in combination with Progenesis LC-MS 4.0 (Nonlinear Dynamics, UK). LC-MS/MS experiments were manually pre-aligned and then submitted to the automated alignment routine in Progenesis. Features with two isotopes or less were discarded and tandem MS spectra were required to have a minimal ion count of 100. Further processing steps included de-isotoping and deconvolution of tandem MS spectra. Resulting peaklists were searched against the UniProtKB complete human proteome set (download date 26.10.2010, 110,550 sequences) combined with sequences of common contaminants using Mascot. The target-decoy option of Mascot was enabled and search parameters included a precursor tolerance of 10 ppm and a fragment tolerance of 0.5 Da for CID spectra and 0.02 Da for HCD spectra. Enzyme specificity was set to trypsin, and up to two missed cleavage sites were allowed. The Mascot ^{13}C option, which accounts for the mis-assignment of the monoisotopic precursor peak, was set to 1. The following variable modifications were considered: oxidation of Met, carbamidomethylation of Cys and HexNAcylation. The database search results were processed using the built-in Mascot Percolator option [28, 29] and filtered at a score of 13, which corresponds to the posterior error probability of 0.05. This results in a peptide FDR of 0.94% and 1.06% for the 'global O-GlcNAc profiling' data set and the 'OGA inhibition' data set, respectively. The Mascot results were imported into Progenesis LC-MS for protein grouping and quantification.

Protein identification from β-elimination experiments was performed with Mascot using the Mascot Distiller version 2.3 (Matrix Science, UK) for data processing. Search parameters were set as detailed above except that dehydration of Ser and Thr was used as variable modification instead of HexNAcylation. Mascot search results were processed using the Mascot Percolator stand-alone software [28, 29] and imported into Scaffold version 3.5.1 (Proteome Software, OR). Mascot Percolator results were filtered at a score of 13 and protein identifications based on only one β-eliminated peptide were manually reviewed. Ascore-based localization probabilities [30] for β-eliminated peptides were calculated with Scaffold PTM 2.0.0 (Proteome Software, Portland, OR).

Data analysis

Biochemical O-GlcNAc protein enrichment factors were determined based on label-free protein quantification comparing GlcNAz-labeled samples and samples without metabolic labelling. Missing values were replaced with arbitrarily chosen small intensity values (0.008) to avoid zero and infinite ratios. Furthermore, proteins representing biochemical noise (i. e. very low intensity in all samples of an experiment) were discarded, when they did not show a minimum intensity in at least one sample. In case of the global O-GlcNAc proteome profiling of HEK293 cells, the minimum intensity was set to 10^5. For the O-GlcNAc proteome profiling in response to OGA inhibition, the minimum intensity was set to 10^4. The differences are due to the different mass spectrometers used.

A positive predictive value (PPV) of representing an O-GlcNAc protein was calculated for each quantified protein based on the bimodal $\log_2$ distribution of biochemical enrichment factors. Briefly, the $\log_2$ distribution of biochemical enrichment factors can be approximated with two Gaussian distributions, one for background proteins and one for specifically enriched O-GlcNAc proteins. The

resulting sum of Gaussian distributions (a 'Gaussian mixture model') was fitted to the observed distribution of $\log_2$ enrichment factors using the method of least squares. A PPV was then calculated based on the resulting modelled Gaussian distributions of background proteins and O-GlcNAc proteins. Similarly, a PPV for the GlcNAcstatin G-responsive proteins was calculated based on the distribution of $\log_2$ ratios (+/- GlcNAcstatin G) with a mixture model approximating non-responsive proteins with a Gaussian distribution and GlcNAcstatin G-responsive proteins with a Gamma distribution.

To assess the novelty in the data, a comprehensive list of reported O-GlcNAc proteins was compiled from a number of sources [11, 20, 31-37]. To reduce redundancy in terms of identifiers and orthologs, the proteins were processed as well as compared to O-GlcNAc proteins identified using IPA (Ingenuity Systems, www.ingenuity.com). Biological functions and pathways enriched in the generated O-GlcNAc data were assessed using IPA. A right-tailed Fisher's exact test was used to calculate a *p*-value determining the probability that each biological function assigned to that data set is due to chance alone. The *p*-value was corrected for multiple hypothesis testing using the method of Benjamini-Hochberg.

Results and discussion

Experimental strategy for global O-GlcNAc proteome-profiling in HEK293 cells

To enable the efficient enrichment and identification of O-GlcNAc proteins along with their sites of modification, a straightforward experimental strategy was developed based on azide/alkyne Click chemistry using a commercially available alkyne resin (Figure 1) [38]. O-GlcNAc proteins are metabolically labeled with GlcNAz, covalently conjugated to an alkyne agarose via CuAAC and are then purified from the vast background of unmodified proteins. To minimize the copper-mediated protein background, which is frequently observed during CuAAC [39], it turned out that the washing step with a strong copper chelator is absolutely required. The resin bound O-GlcNAc proteins are subsequently digested with trypsin, thereby allowing for MS-based identification of those parts of O-GlcNAc proteins, which are not covalently bound to the resin. To minimize the risk of false-positive O-GlcNAc site assignments during β-elimination, the remaining resin bound O-GlcNAc peptides are extensively dephosphorylated before they are released by β-elimination, which enables the concomitant tagging and MS-based identification of the former O-GlcNAc sites. The selectivity of this procedure is mainly conferred by the metabolic labeling of O-GlcNAc proteins and the bioorthogonality of the Click chemistry reaction.

To evaluate the merits of the strategy, initially the global O-GlcNAc proteome of HEK293 cells was profiled. To assess the selectivity of the enrichment procedure, the experiment was performed in parallel with GlcNAz-labeled and unlabeled HEK293 cells. Label-free intensity-based quantification was used to obtain a biochemical enrichment factor for every protein identified in the on-resin digest. The summed intensities of proteins identified from the GlcNAz-labeled sample was $1.9 \cdot 10^{10}$ which is >60-fold higher than the summed intensity of the negative control (Figure 2A) and the median protein enrichment factor across all proteins was 260. Together, these figures clearly show that O-GlcNAc modified proteins can be efficiently enriched from metabolically GlcNAz-labelled samples.

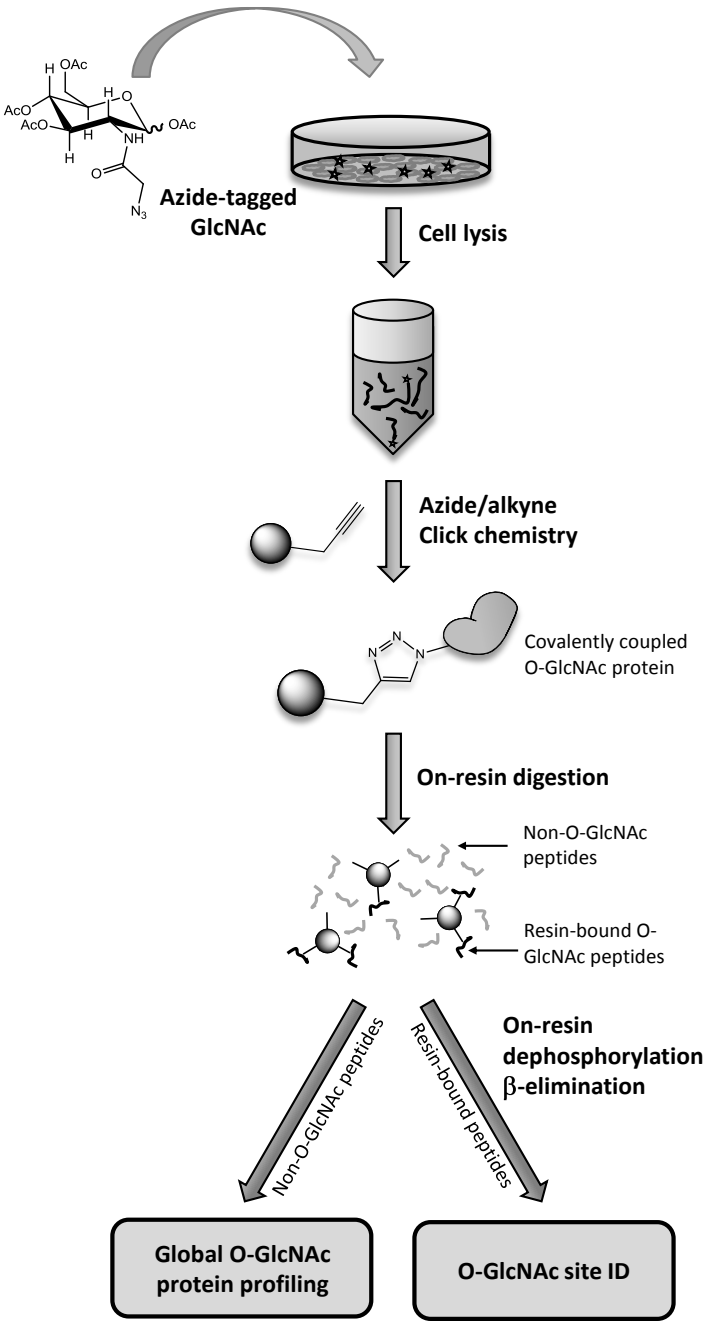

Figure 1 | Experimental strategy for the Click chemistry-based enrichment and identification of O-GlcNAc modified proteins.
Briefly, O-GlcNAc proteins are metabolically labeled with GlcNAz, covalently conjugated to an alkyne agarose, and subsequently enriched from the vast background of unmodified proteins. For the global O-GlcNAc protein profiling, the proteins are proteolytically digested on the resin and the released peptides are analyzed by LC-MS/MS. To obtain O-GlcNAc protein site information, the resin-bound peptides are subjected to on-resin dephosphorylation and β-elimination before LC-MS/MS analysis.

Interestingly, the biochemical enrichment factors follow a bimodal distribution on a logarithmic scale (Figure 2B). The distribution centered around zero likely represents the background proteome unselectively bound by the alkyne resin. This background binding might be in part copper-mediated, but proteins may also unspecifically bind as a result of non-covalent interactions with the alkyne moieties on the beads or the agarose backbone of the beads themselves. In contrast, the specifically

enriched O-GlcNAc proteome is represented by the strongly right-shifted distribution of biochemical enrichment factors. To describe the selectivity of the enrichment more quantitatively, the observed bimodal distribution was approximated by a Gaussian mixture model which enabled the calculation of a positive predictive value (PPV) for O-GlcNAc proteins. For instance, a PPV of 0.99 indicates that 99% of all proteins with 82-fold biochemical enrichment originate from the distribution of O-GlcNAc proteins and only 1% from the distribution of background proteins. Following this rationale, 1,536 O-GlcNAc proteins with a PPV > 0.99 (1747 proteins at PPV >0.95) were identified, representing the largest collection of O-GlcNAc proteins identified in a single experiment so far.

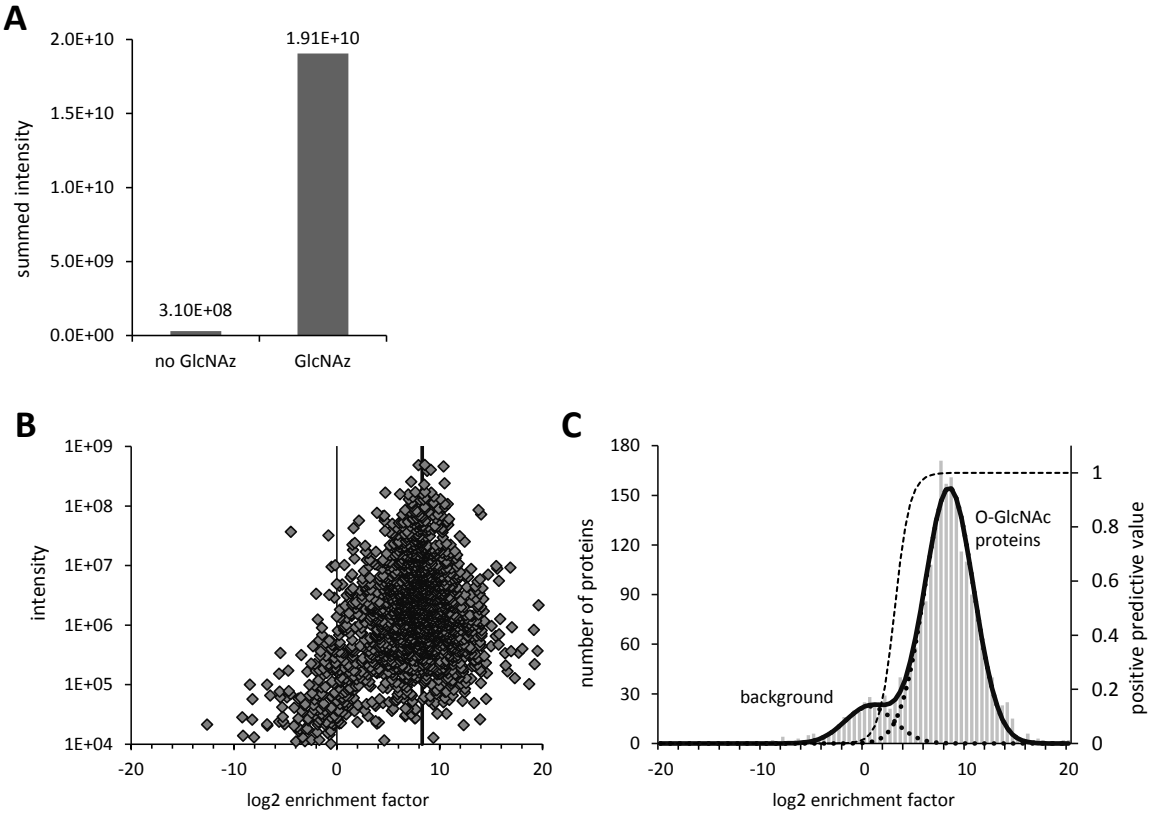

Figure 2 | Global identification of O-GlcNAc proteins from HEK293 cells
A Summed intensities of enriched proteins with and without metabolic labeling of cells by GlcNAz. **B** Scatterplot of intensity and $\log_2$ biochemical enrichment of identified O-GlcNAc proteins. **C** The bimodal distribution of biochemical enrichment factors allows the calculation of positive predictive values for CuAAC-enriched O-GlcNAc proteins.

A comparison to 1,269 O-GlcNAc proteins reported in recent studies from mouse and human tissue and cell lines [11, 20, 31-37] revealed that 74 of the 100 most abundant O-GlcNAc proteins in the data set (or their mouse orthologs) have been previously identified again showing that the biochemical approach works. The total overlap of the data with the cited studies is 27% (338 reported O-GlcNAc proteins), which appears reasonable given that the data stems from a single human cell line whereas the reported 1,269 O-GlcNAc proteins were identified from different mammalian species, tissues and cell types thus representing a mixture of cell type specific O-GlcNAcylation profiles.

The results of the GO term analysis for subcellular localization of the identified proteins is consistent with the common notion that O-GlcNAc is primarily a nuclear and cytoplasmic PTM. Indeed, 39% of

all identified proteins are supposedly nuclear and a further 41% cytoplasmic (Figure 3A). A broad range of biological functions (Figure 3B) is associated with the identified O-GlcNAc proteins. Notably, the data comprises a wealth of biologically very interesting proteins that are often not identified in global proteome profiling studies which adds further evidence that the described enrichment protocol is efficient. Examples for such proteins include the transcription factors p53, SP1, FoxO3, CREB, STAT1 and NFκB, proteins involved in epigenetic regulation such as HCF-1, Sirtuin-1, NCOA-1, -2 and -3, HDAC1, HDAC2, MLL and proteins required for microRNA maturation (e. g. Dicer and TARPB2). Other modified proteins are involved in ubiquitination, RAN-mediated nuclear transport, aminoacyl-tRNA synthesis and several signalling pathways. Taken together, the identified O-GlcNAc proteins reflect the large variety of regulatory functions O-GlcNAc may exert in the cell [4, 40-42] and provides a method by which these roles may be further studied in the future.

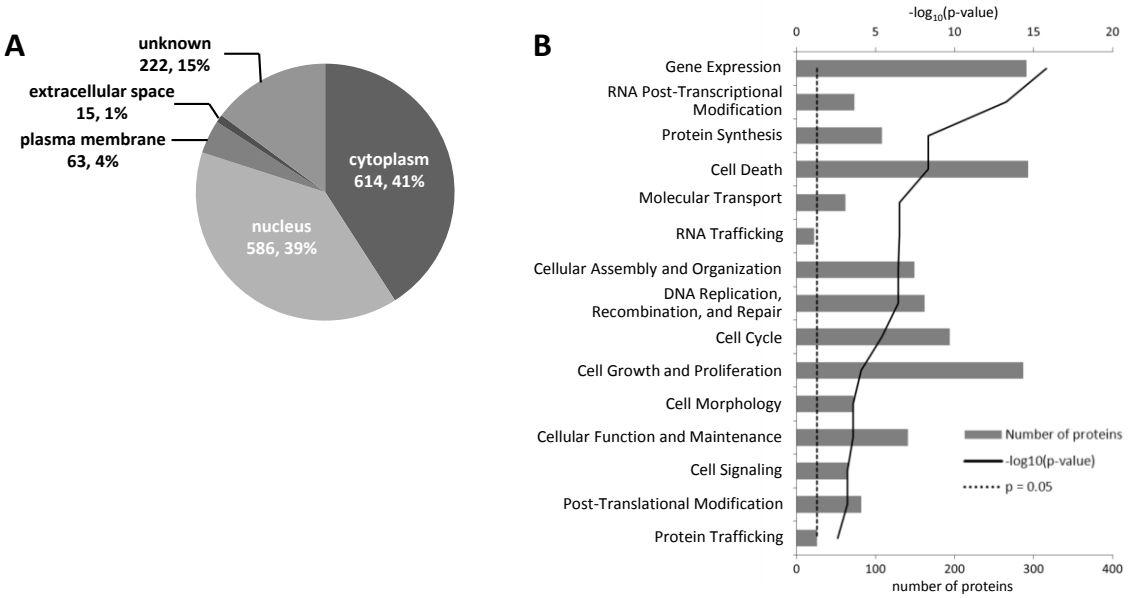

Figure 3 | Subcellular localization and functional categories of identified O-GlcNAc proteins
A Presumed subcellular localization and **B** significantly enriched cellular functions as revealed by Ingenuity Pathways Analysis ($p < 0.05$).

Identification of O-GlcNAc sites

While the previous section provides good evidence for the selectivity of the developed O-GlcNAc protein enrichment method, this section addresses the identification of the corresponding modification sites based on selective β-elimination. Chemical β-elimination is a rather unselective procedure because it does not discriminate well between O-GlcNAc and phosphate moieties on Ser and Thr residues. This is further complicated by the frequent co-occurrence of both modifications and the approximately ten-fold higher abundance of phosphorylation over O-GlcNAc-ylation [11, 34]. To discriminate between phosphorylation and O-GlcNAcylation, an on-resin dephosphorylation step was incorporated between the on-resin proteolytic digest and the on-resin β-elimination. By way of the Click reaction, all peptides bound to the alkyne resin (ideally) should be O-GlcNAc modified but could also be additionally phosphorylated. The latter case could lead to misinterpretation of the data obtained after β-elimination while removing the phosphate groups first largely eliminates this issue. To obtain an efficient as well as selective β-elimination procedure, a broad range of β-elimination and Michael addition conditions were screened with respect to

reaction time, temperature and Michael addition donors (Table 1) using a synthetic reference peptide library comprising 72 O-GlcNAc peptides [15] and 48 phosphopeptides [25]. From this work, it could be concluded that the commercial GlycoProfile β-elimination kit represents a sound compromise between efficiency and selectivity as it enables efficient O-GlcNAc β-elimination while maintaining a low number of false-positive identifications resulting from β-elimination of phospho-sites (Table 1). Importantly, BEMAD conditions often led to massive degradation of the protein backbone (Figure 4) whereas this was not the case for the β-elimination kit. From this set of experiments, there was no advantage in adding a nucleophile after β-elimination as this step is not required for the detection of the modified amino acid by mass spectrometry (which can be done by detecting the de-hydro form of the amino acid) and, instead only increases the chemical complexity of the sample (e. g. as a result of incomplete or side reactions).

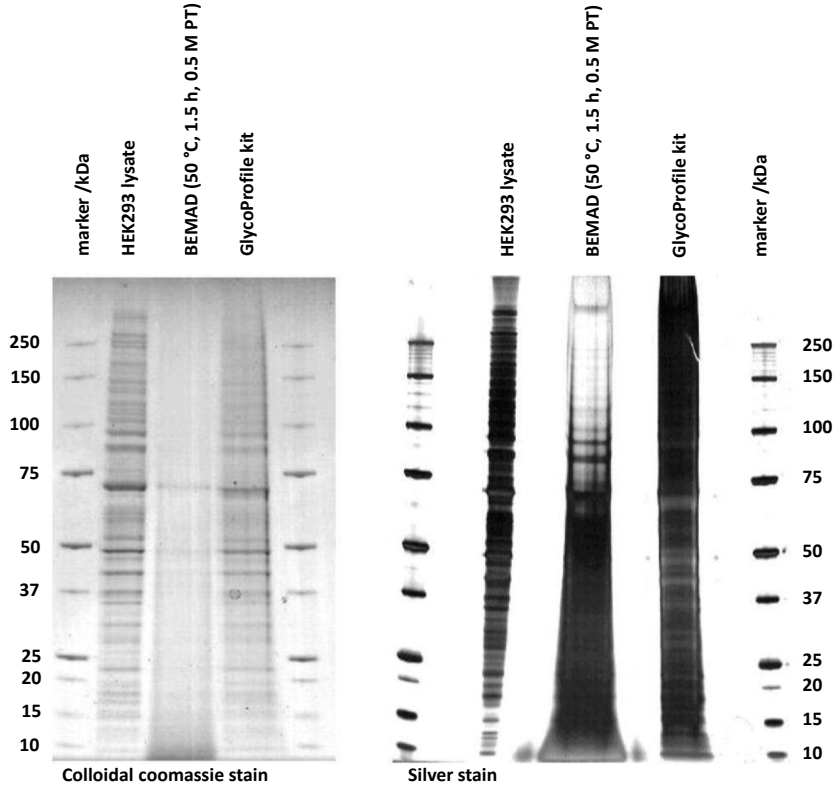

Figure 4 | BEMAD compromises protein integrity
This comparison of the BEMAD approach and the GlycoProfile kit using an HEK293 cell lysate reveals that BEMAD leads to significant protein degradation, while proteins remain intact during the GlycoProfile kit β-elimination procedure.

Table 1 | Beta-elimination and Michael addition of standard O-GlcNAc and phospho-peptides

				Spectral Counts[#]				Unique peptides[#]			
				O-GlcNAc		phospho		O-GlcNAc		phospho	
	Temperature and time	beta-elimination	Michael addition	BE	MA	BE	MA	BE	MA	BE	MA
1st round	25 °C, 24 h	BE	no Michael addition	9	-	4	-	9	-	4	-
			10mM BME	11	18	1	5	10	15	1	2
			10mM DTT	10	18	1	1	10	12	1	1
	50 °C, 1.5 h		no Michael addition	20	-	0	-	13	-	0	-
			10mM BME	15	17	3	5	11	12	3	5
			10mM DTT	22	15	5	3	15	12	3	3
	50 °C, 4 h		no Michael addition	37	-	6	-	22	-	4	-
			10mM BME	29	12	0	1	15	12	0	5
			10mM DTT	19	28	3	7	11	16	3	4
2nd round	50 °C, 1.5 h	BE	0.5M BME	18	102	6	5	16	29	5	3
			0.5M PT	41	139	1	23	24	40	1	4
			1.0M PT	161	265	3	21	35	49	3	4
3rd round	4 °C, 18 h	BE	0.5M PT	0	73	3	3	0	23	3	3
		GlycoProfile kit	no Michael addition	141	-	3	-	39	-	2	-
			0.5M PT	20	132	3	5	14	32	2	4

Mascot ion score > 20; BE: beta-elimination mix; BME: beta-mercaptoethanol; DTT: dithiothreitol; MA: Michael addition; PT: 1-propanethiol

Following these results, HEK293 cells were analyzed and, overall, 585 O-GlcNAc spectra representing 125 O-GlcNAc sites on 80 proteins were identified (triplicate analysis on an LTQ Orbitrap XL and a single analysis on an LTQ Orbitrap Velos platform). For 84 of the O-GlcNAc sites, the site could be determined with a localization probability of better than 90%, while for 41 sites the former O-GlcNAc sites could not be mapped with certainty. More than 70% of the identified O-GlcNAc proteins and 31% of the identified sites have been reported previously, which is well within expectation and underpins the validity of the employed procedure.

Even though the methods reported in this study led to the so far largest number of O-GlcNAc sites identified from a single cell line without artificially elevated O-GlcNAc levels, there still is a striking discrepancy between the number of identified O-GlcNAc proteins and the corresponding modification sites. While the on-resin digest revealed 1536 high confidence O-GlcNAc proteins (PPV >0.99), site evidence was obtained only for 80 proteins. These 80 proteins are, by and large, among the most abundant O-GlcNAc proteins (Figure 5) indicating that the identification of CuAAC-captured O-GlcNAc proteins can be readily accomplished while the identification of their O-GlcNAc sites is rather challenging.

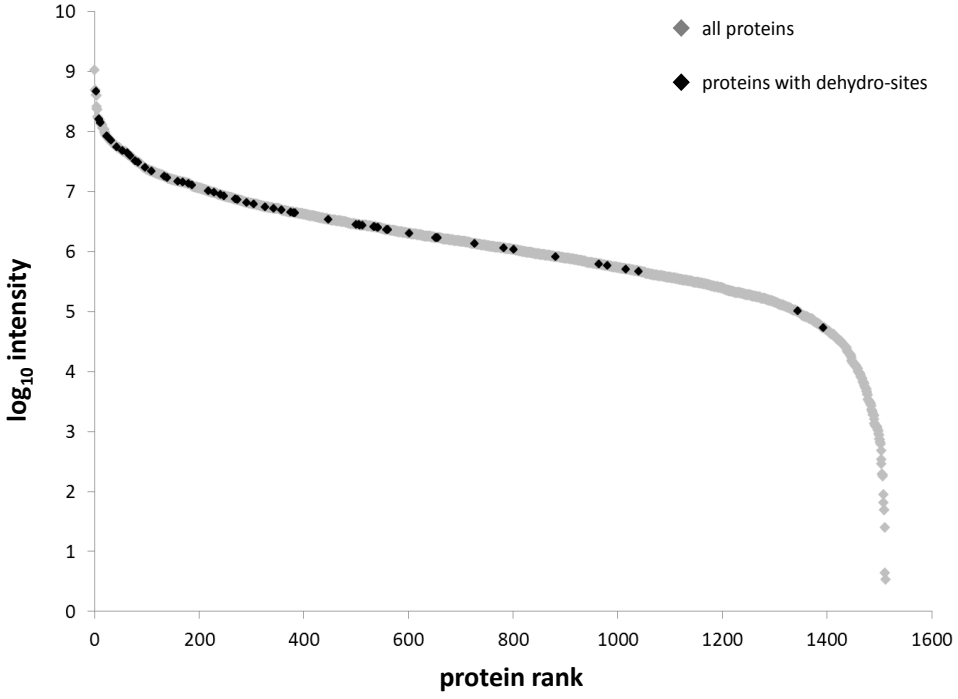

Figure 5 | Intensity distribution of proteins identified in the on-resin digest (gray) and additionally identified with dehydro-peptides in the β-elimination fraction (black)
The distribution underscores that the dehydro-peptides were, in most cases, identified from highly abundant O-GlcNAc proteins.

Several general and experiment-specific reasons were identified which may explain the observed bias. Clearly, the identification of resin-bound O-GlcNAc proteins is strongly facilitated by the large number and diversity of tryptic peptides generated from intact O-GlcNAc proteins. While any of the peptides support an identification, the modification site assignment clearly requires the detection of the specific modified peptide. In complex samples, the chance of this happening is likely less than 10-20%. In addition, azide-tagged O-GlcNAc proteins may be conjugated to the alkyne resin via any one of their O-GlcNAc sites but very unlikely by all sites. While this would lead to a stoichiometry of

at least one for the captured protein, the stoichiometry of a particular captured O-GlcNAc peptide will likely be fairly (or even very) low, rendering the identification of O-GlcNAc sites difficult. Furthermore, O-GlcNAc sites have been found to occur predominantly in regions with low compositional complexity [11, 34], suggesting that O-GlcNAc sites may often not be amenable for common proteomic workflows using trypsin. And, last but not least, the employed β-elimination procedure has been optimized for O-GlcNAc peptides in solution. However, the on-resin reaction may not proceed with the same efficiency owing to kinetic or spatial imperfections. Still, compared to previous approaches utilizing some form of β-elimination/Michael addition for the identification of O-GlcNAc proteins and sites [10, 21, 22], the approach represents a considerable improvement. Because the reagents are commercially available, the methods should also be easily adaptable by other laboratories.

Global identification of GlcNAcstatin G-responsive proteins

To demonstrate the practical utility of the developed O-GlcNAc protein enrichment procedure for the global analysis of O-GlcNAc proteins, the effect of OGA inhibition on a proteome-wide scale using GlcNAcstatin G [27] was studied. GlcNAcstatin G is a potent OGA inhibitor which exhibits selectivity over related lysosomal hexosaminidases and has been shown to induce hyper-O-GlcNAcylation of cellular proteins in the nanomolar range but the spectrum of proteins actually responding to this treatment has not yet been systematically determined. To do so, HEK293 cells were labeled with GlcNAz, treated with GlcNAcstatin G or vehicle control (DMSO) for two hours and O-GlcNAc proteins were enriched and quantified as above. The selectivity of the enrichment was controlled using unlabelled HEK293 cells as negative control. In this experiment, analysed on a LTQ Orbitrap XL, 1,692 proteins, of which 1,039 can be considered high confidence O-GlcNAc proteins (PPV > 0.99) were identified. Note that in this experiment, the PPV for O-GlcNAc modified proteins was calculated based on the OGA inhibitor treated sample. Otherwise O-GlcNAc proteins with an initially low O-GlcNAc stoichiometry (and, hence, low biochemical enrichment factor), but significantly increased O-GlcNAcylation upon treatment would not be taken into account, thus leaving out a very relevant group of GlcNAcstatin G-responsive proteins. Following a two hour drug treatment, a significant number of proteins exhibit clearly increased O-GlcNAc levels (Figure 6A). To obtain a reliable list of affected proteins, an additional PPV for GlcNAcstatin G-responsive proteins was calculated using a mixture model assuming a Gaussian distribution around zero for unaffected proteins and a Gamma distribution for GlcNAcstatin G-responsive proteins. This mixture model can be rationalized as follows: The intensity of unaffected proteins reflects solely biological and technical variation that can be approximated by a Gaussian distribution. In contrast, the intensity of GlcNAcstatin G-responsive proteins also increases to a varying extent upon OGA inhibition. To account for the inherent skewness of such a distribution, the distribution of affected proteins was modeled using a right-tailed Gamma distribution. Overall, this model provides a good fit for the observed right-tailed distribution of fold changes on a logarithmic scale (Figure 6B). Following this rationale, 189 O-GlcNAC proteins were affected by the drug treatment (PPV >0.9). The fact that less than 20% of all high confidence O-GlcNAc proteins were GlcNAcstatin G-responsive, suggests that most of these proteins are already modified to a high stoichiometry. This is consistent with recent findings that O-GlcNAc sites of the most abundant O-GlcNAc proteins exhibit an relative site occupancy of around 90% [9] and also with *in vitro* experiments showing that some OGT substrates can be constitutively modified [43]. An important technical aspect of this experiment is that the

identification of 189 GlcNAcstatin G-responsive proteins further underscores the validity of developed O-GlcNAc enrichment approach.

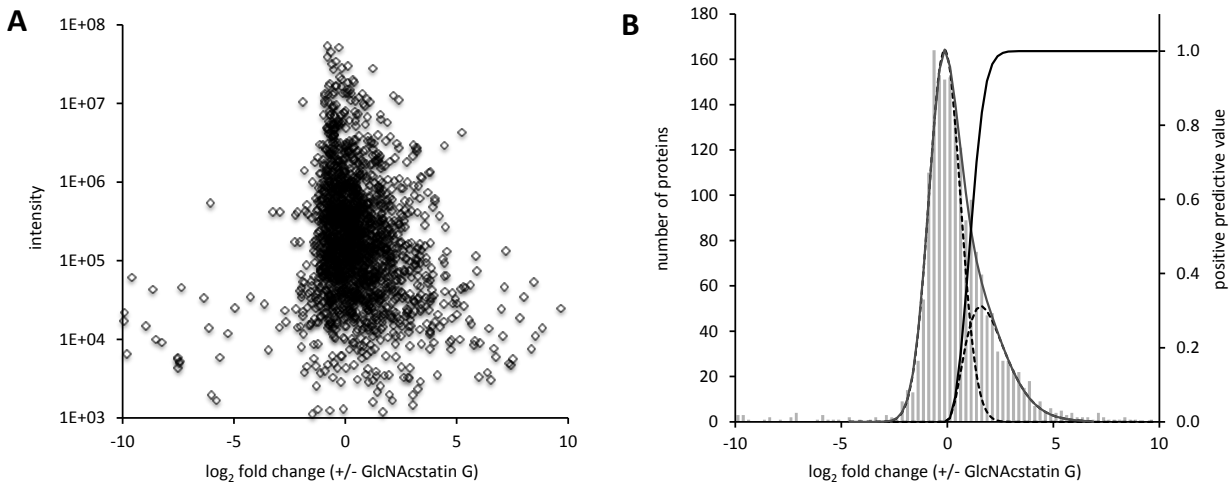

Figure 6 | Global profiling of GlcNAcstatin G-responsive proteins
A Scatterplot of intensity and $\log_2$ fold change of identified O-GlcNAc proteins. **B** The distribution of $\log_2$ ratios (+/- GlcNAcstatin G) of identified O-GlcNAc proteins was used to calculate positive predictive values for GlcNAcstatin G-responsive proteins.

GlcNAcstatin G-responsive pathways and selected proteins

Cellular pathways significantly overrepresented among the GlcNAcstatin G-responsive proteins are depicted in Figure 7A. Interestingly, there are several key players involved in nutrient responsive O-GlcNAc cycling, some of which have not been reported as O-GlcNAc modified before. Several metabolic enzymes, most of which catalyse committed steps of anabolic pathways requiring glutamine as amino donor, were identified to be OGA inhibitor-responsive. Notably, glutamine-fructose-6-phosphate transaminase (GFPT) was identified, which controls the flux of glucose into the HBP and, eventually, the intracellular level of UDP-GlcNAc. AMP-activated protein kinase (AMPK) also showed increased O-GlcNAc levels following inhibitor treatment. AMPK is the central hub in energy responsive AMPK signalling and down-regulates multiple anabolic pathways upon energy deprivation [44, 45]. This is consistent with previous results that an increased flux through the HBP by overexpression of GFPT results in increased AMPK phosphorylation and O-GlcNAcylation leading to the activation of energy replenishing pathways [46]. Another drug responsive protein is the deacetylase Sirtuine-1. Interestingly, it has been hypothesized that O-GlcNAcylation and sirtuin-dependent deacetylation may exert opposing functions in situations of nutrient excess or starvation [41]. A striking finding is that almost all members of the RAN-dependent nuclear transport system, which mediates the transport of proteins, tRNAs and ribosomal subunits across the nuclear membrane are modified by O-GlcNAc and several of the key proteins show considerable increase in O-GlcNAc levels upon OGA inhibition (Figure 7B). O-GlcNAc residues on nuclear transport factors have been associated with important recognition events in nuclear transport [47]. It is, therefore, tempting to speculate that RAN signalling may represent an additional module of the nutrient responsive O-GlcNAc cycling system. Clearly, further work is required to identify the site(s) of modification on these proteins as well as and their potential functional significance.

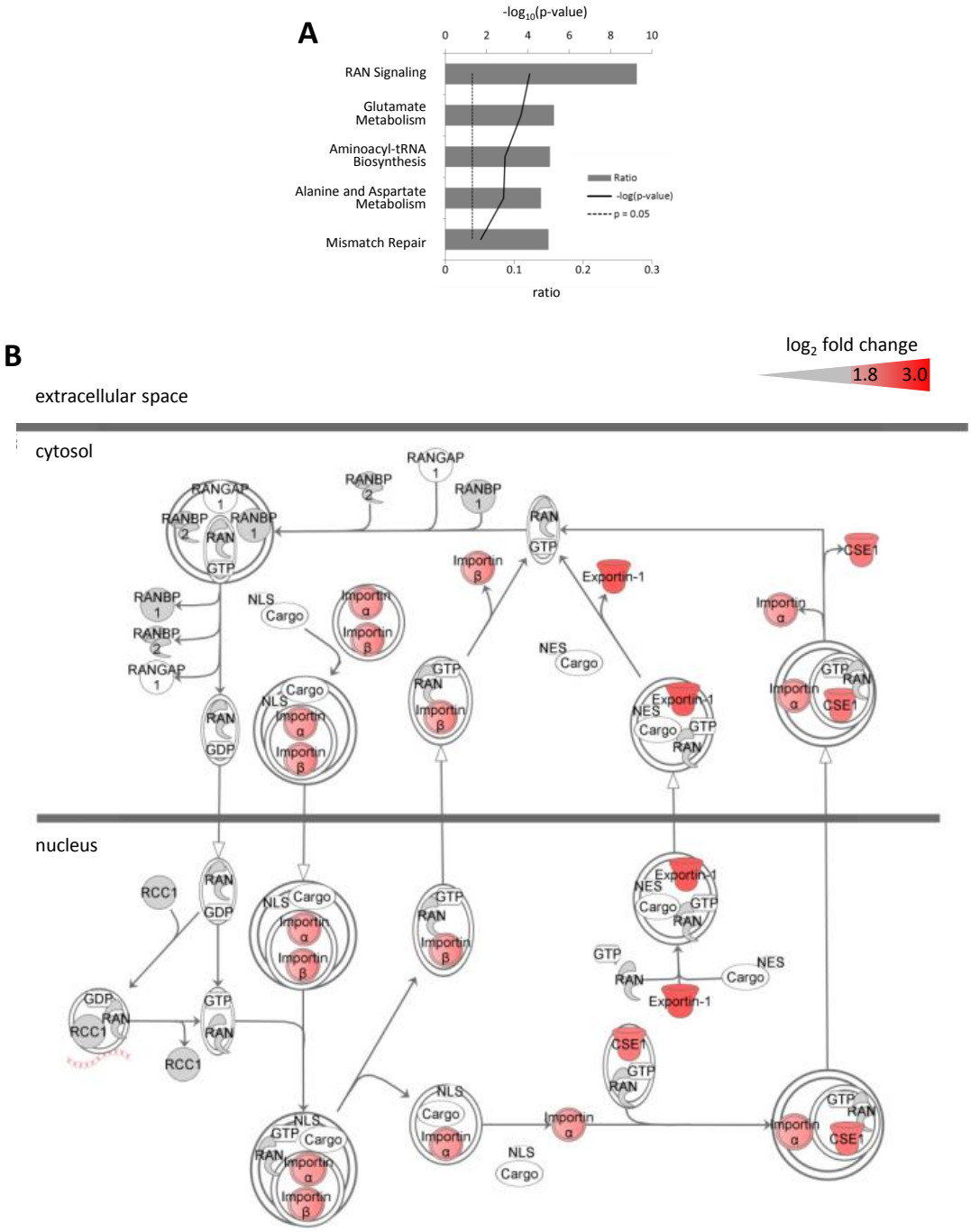

Figure 7 | Cellular pathways responding to OGA inhibition by GlcNAcstatin G
A Significantly enriched biochemical and signalling pathways as revealed by Ingenuity Pathways analysis ($p<0.05$). **B** RAN nuclear transport pathway. Almost all proteins of this pathway were found to be O-GlcNAc modified (grey) and several members of this cellular machinery show increased levels of O-GlcNAc upon OGA inhibition (red).

Conclusion

This study introduces a novel and valid method for the enrichment of O-GlcNAc modified proteins based on metabolic labelling, Click chemistry and quantitative mass spectrometry that represents a significant improvement over and a useful complement to existing methods. The approach routinely enables the identification of >1,000 modified proteins which opens up many lines of investigation into the cellular role of this emerging post translational protein modification. Because the reagents are all commercially available, the approach should be readily adoptable by other laboratories and may even be combined with existing O-GlcNAc enrichment approaches such as chemoenzymatic O-GlcNAc tagging. Indeed, the reach of the method may extend even further to any type of azide-tagged protein or modification thereof.

Acknowledgments

The author is indebted to Nadine Sobotzki who performed the initial development of this approach during her MSc thesis and to Dominic Helm who was working on alternative Click-based enrichment approaches during his MSc thesis. The author also thanks Tamara Nyberg and Brian Agnew from Life Technologies (Eugene, Oregon, USA) for fruitful discussions and providing reagents and alkyne resin, and Daan van Aalten (Division of Molecular Microbiology, College of Life Sciences, University of Dundee, Dundee, Scotland) for GlcNAcstatin G.

Abbreviations

BEMAD	β-elimination followed by Michael addition
CID	collision induced dissociation
CuAAC	copper-catalyzed azide/alkyne Click chemistry
ECD	electron capture dissociation
ETD	electron transfer dissociation
HBP	hexosamine biosynthetic pathway
MS	mass spectrometry
LC-MS/MS	liquid chromatography coupled to tandem mass spectrometry
O-GlcNAc	O-linked β- N-acetylglucosamine
PTM	posttranslational modification
UDP-GlcNAc	uridine diphosphate N-acetylglucosamine

References

1. Torres, C. R., and Hart, G. W. (1984) Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. J Biol Chem 259, 3308-3317.
2. Love, D. C., and Hanover, J. A. (2005) The hexosamine signaling pathway: deciphering the "O-GlcNAc code". Sci STKE 2005, re13.
3. Hart, G. W., Housley, M. P., and Slawson, C. (2007) Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. Nature 446, 1017-1022.
4. Hart, G. W., Slawson, C., Ramirez-Correa, G., and Lagerlof, O. (2011) Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. Annu Rev Biochem 80, 825-858.
5. Toleman, C., Paterson, A. J., Whisenhunt, T. R., and Kudlow, J. E. (2004) Characterization of the histone acetyltransferase (HAT) domain of a bifunctional protein with activable O-GlcNAcase and HAT activities. J Biol Chem 279, 53665-53673.
6. Wells, L., Gao, Y., Mahoney, J. A., Vosseller, K., Chen, C., Rosen, A., and Hart, G. W. (2002) Dynamic O-glycosylation of nuclear and cytosolic proteins: further characterization of the nucleocytoplasmic beta-N-acetylglucosaminidase, O-GlcNAcase. J Biol Chem 277, 1755-1761.
7. Haltiwanger, R. S., Grove, K., and Philipsberg, G. A. (1998) Modulation of O-linked N-acetylglucosamine levels on nuclear and cytoplasmic proteins in vivo using the peptide O-GlcNAc-beta-N-acetylglucosaminidase inhibitor O-(2-acetamido-2-deoxy-D-glucopyranosylidene)amino-N-phenylcarbamate. J Biol Chem 273, 3611-3617.
8. Macauley, M. S., and Vocadlo, D. J. (2010) Increasing O-GlcNAc levels: An overview of small-molecule inhibitors of O-GlcNAcase. Biochim Biophys Acta 1800, 107-121.
9. Hahne, H., Gholami, A. M., and Kuster, B. (2012) Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. Mol Cell Proteomics [epub ahead of print 2012/06/05].
10. Vosseller, K., Trinidad, J. C., Chalkley, R. J., Specht, C. G., Thalhammer, A., Lynn, A. J., Snedecor, J. O., Guan, S., Medzihradszky, K. F., Maltby, D. A., Schoepfer, R., and Burlingame, A. L. (2006) O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. Mol Cell Proteomics 5, 923-934.
11. Trinidad, J. C., Barkan, D. T., Gulledge, B. F., Thalhammer, A., Sali, A., Schoepfer, R., and Burlingame, A. L. (2012) Global identification and characterization of both O-GlcNAcylation and phosphorylation at the murine synapse. Mol Cell Proteomics [epub ahead of print 2012/05/31].
12. Wang, Z., Udeshi, N. D., O'Malley, M., Shabanowitz, J., Hunt, D. F., and Hart, G. W. (2010) Enrichment and site mapping of O-linked N-acetylglucosamine by a combination of chemical/enzymatic tagging, photochemical cleavage, and electron transfer dissociation mass spectrometry. Mol Cell Proteomics 9, 153-160.
13. Alfaro, J. F., Gong, C. X., Monroe, M. E., Aldrich, J. T., Clauss, T. R., Purvine, S. O., Wang, Z., Camp, D. G., 2nd, Shabanowitz, J., Stanley, P., Hart, G. W., Hunt, D. F., Yang, F., and Smith, R. D. (2012) Tandem mass spectrometry identifies many mouse brain O-GlcNAcylated proteins including EGF domain-specific O-GlcNAc transferase targets. Proc Natl Acad Sci U S A 109, 7280-7285.
14. Mirgorodskaya, E., Roepstorff, P., and Zubarev, R. A. (1999) Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. Anal Chem 71, 4431-4436.
15. Hahne, H., and Kuster, B. (2011) A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. J Am Soc Mass Spectrom 22, 931-942.
16. Vocadlo, D. J., Hang, H. C., Kim, E. J., Hanover, J. A., and Bertozzi, C. R. (2003) A chemical approach for identifying O-GlcNAc-modified proteins in cells. Proc Natl Acad Sci USA 100, 9116-9121.

17. Sprung, R., Nandi, A., Chen, Y., Kim, S. C., Barma, D., Falck, J. R., and Zhao, Y. (2005) Tagging-via-substrate strategy for probing O-GlcNAc modified proteins. J Proteome Res 4, 950-957.
18. Nandi, A., Sprung, R., Barma, D. K., Zhao, Y., Kim, S. C., and Falck, J. R. (2006) Global identification of O-GlcNAc-modified proteins. Anal Chem 78, 452-458.
19. Gurcel, C., Vercoutter-Edouart, A. S., Fonbonne, C., Mortuaire, M., Salvador, A., Michalski, J. C., and Lemoine, J. (2008) Identification of new O-GlcNAc modified proteins using a click-chemistry-based tagging. Anal Bioanal Chem 390, 2089-2097.
20. Zaro, B. W., Yang, Y. Y., Hang, H. C., and Pratt, M. R. (2011) Chemical reporters for fluorescent detection and identification of O-GlcNAc-modified proteins reveal glycosylation of the ubiquitin ligase NEDD4-1. Proc Natl Acad Sci U S A 108, 8146-8151.
21. Wells, L., Vosseller, K., Cole, R. N., Cronshaw, J. M., Matunis, M. J., and Hart, G. W. (2002) Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. Mol Cell Proteomics 1, 791-804.
22. Vosseller, K., Hansen, K. C., Chalkley, R. J., Trinidad, J. C., Wells, L., Hart, G. W., and Burlingame, A. L. (2005) Quantitative analysis of both protein expression and serine / threonine post-translational modifications through stable isotope labeling with dithiothreitol. Proteomics 5, 388-398.
23. Li, W., Backlund, P. S., Boykins, R. A., Wang, G., and Chen, H. C. (2003) Susceptibility of the hydroxyl groups in serine and threonine to beta-elimination/Michael addition under commonly used moderately high-temperature conditions. Anal Biochem 323, 94-102.
24. Herbert, B., Hopwood, F., Oxley, D., McCarthy, J., Laver, M., Grinyer, J., Goodall, A., Williams, K., Castagna, A., and Righetti, P. G. (2003) Beta-elimination: an unexpected artefact in proteome analysis. Proteomics 3, 826-831.
25. Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., and Kuster, B. (2011) Confident phosphorylation site localization using the Mascot Delta Score. Mol Cell Proteomics 10, M110 003830.
26. Rappsilber, J., Mann, M., and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2, 1896-1906.
27. Dorfmueller, H. C., Borodkin, V. S., Schimpl, M., Zheng, X., Kime, R., Read, K. D., and van Aalten, D. M. (2010) Cell-penetrant, nanomolar O-GlcNAcase inhibitors selective against lysosomal hexosaminidases. Chem Biol 17, 1250-1255.
28. Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4, 923-925.
29. Brosch, M., Yu, L., Hubbard, T., and Choudhary, J. (2009) Accurate and sensitive peptide identification with Mascot Percolator. J Proteome Res 8, 3176-3181.
30. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24, 1285-1292.
31. Wang, Z., Udeshi, N. D., Slawson, C., Compton, P. D., Sakabe, K., Cheung, W. D., Shabanowitz, J., Hunt, D. F., and Hart, G. W. (2010) Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. Sci Signal 3, ra2.
32. Wang, J., Torii, M., Liu, H., Hart, G. W., and Hu, Z. Z. (2011) dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. BMC Bioinformatics 12, 91.
33. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 40, D261-270.
34. Hahne, H., Gholami, A. M., and Kuster, B. (2012) Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. Mol Cell Proteomics [epub ahead of print 2012/06/05].

35. Myers, S. A., Panning, B., and Burlingame, A. L. (2011) Polycomb repressive complex 2 is necessary for the normal site-specific O-GlcNAc distribution in mouse embryonic stem cells. Proc Natl Acad Sci USA 108, 9490-9495.
36. Chalkley, R. J., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. Proc Natl Acad Sci USA 106, 8894-8899.
37. Zhao, P., Viner, R., Teo, C. F., Boons, G. J., Horn, D., and Wells, L. (2011) Combining high-energy C-trap dissociation and electron transfer dissociation for protein O-GlcNAc modification site assignment. J Proteome Res 10, 4088-4104.
38. Hart, C., Chase, L. G., Hajivandi, M., and Agnew, B. (2011) Metabolic labeling and click chemistry detection of glycoprotein markers of mesenchymal stem cell differentiation. Methods Mol Biol 698, 459-484.
39. Speers, A. E., and Cravatt, B. F. (2004) Profiling enzyme activities in vivo using click chemistry methods. Chem Biol 11, 535-546.
40. Hu, P., Shimoji, S., and Hart, G. W. (2010) Site-specific interplay between O-GlcNAcylation and phosphorylation in cellular regulation. FEBS Lett 584, 2526-2538.
41. Hanover, J. A., Krause, M. W., and Love, D. C. (2010) The hexosamine signaling pathway: O-GlcNAc cycling in feast or famine. Biochim Biophys Acta 1800, 80-95.
42. Butkinaree, C., Park, K., and Hart, G. W. (2009) O-linked beta-N-acetylglucosamine (O-GlcNAc): Extensive crosstalk with phosphorylation to regulate signaling and transcription in response to nutrients and stress. Biochim Biophys Acta 1800, 96-106.
43. Shen, D. L., Gloster, T. M., Yuzwa, S. A., and Vocadlo, D. J. (2012) Insights into O-GlcNAc processing and dynamics through kinetic analysis of O-GlcNAc transferase and O-GlcNAcase activity on protein substrates. J Biol Chem 287, 15395-15408.
44. Carling, D., Mayer, F. V., Sanders, M. J., and Gamblin, S. J. (2011) AMP-activated protein kinase: nature's energy sensor. Nat Chem Biol 7, 512-518.
45. Hardie, D. G., Ross, F. A., and Hawley, S. A. (2012) AMPK: a nutrient and energy sensor that maintains energy homeostasis. Nat Rev Mol Cell Biol 13, 251-262.
46. Luo, B., Parker, G. J., Cooksey, R. C., Soesanto, Y., Evans, M., Jones, D., and McClain, D. A. (2007) Chronic hexosamine flux stimulates fatty acid oxidation by activating AMP-activated protein kinase in adipocytes. J Biol Chem 282, 7172-7180.
47. Yu, S. H., Boyce, M., Wands, A. M., Bond, M. R., Bertozzi, C. R., and Kohler, J. J. (2012) Metabolic labeling enables selective photocrosslinking of O-GlcNAc-modified proteins to their binding partners. Proc Natl Acad Sci U S A 109, 4834-4839.

Chapter 6

Carbonyl-reactive tandem mass tags for the proteome-wide quantification of N-linked glycans

Summary

N-linked protein glycosylation is one of the most prevalent post-translational modifications and is involved in essential cellular functions such as cell-cell interactions and cellular recognition as well as in chronic diseases. This study explored stable isotope labelled carbonyl-reactive tandem mass tags (glyco-TMTs) as a novel approach for the quantification of N-linked glycans. Glyco-TMTs bearing hydrazide- and aminooxy-functionalized groups were compared for glycan reducing end derivatization efficiency and quantification merits. Aminooxy TMTs outperform the hydrazide reagents in terms of labelling efficiency (>95% vs. 65% at 0.1 μM) and mass spectrometry based quantification using heavy/light-TMT labelled glycans enabled accurate quantification in MS1 spectra (CV < 15%) over a broad dynamic range (up to 1:40). In contrast, isobaric TMT labelling with quantification of reporter ions in tandem mass spectra suffered from severe ratio compression already at low sample ratios. To demonstrate the practical utility of the developed approach, the global N-linked glycosylation profiles of the isogenic human colon carcinoma cell lines SW480 (primary tumor) and SW620 (metastatic tumor) were quantitatively characterized. The data revealed significant down-regulation of high-mannose glycans in the metastatic cell line.

Introduction

Glycosylation is one of the most abundant post-translational modifications found on more than half of all secreted and cellular proteins [1]. The modification is involved in many cellular processes including cell-cell interactions and cellular recognition, proliferation and development, immune response, and protein stability [2]. It is therefore not surprising that aberrant protein glycosylation has also been reported in the context of numerous inherited and acquired diseases and that altered protein glycosylation may represent potential biomarkers for diagnosis or prognosis [3].

Mass spectrometry (MS) currently is the method of choice for the compositional and structural analysis of protein glycosylation [4, 5] and quantitative mass spectrometry has, therefore, developed into an indispensable tool for the investigation of the role of glycosylation in physiological and pathological processes [6, 7]. Traditionally, glycan quantification was performed using light absorbing derivatization reagents or the mass spectrometric signal intensity of native and derivatized N-linked glycans. More recently, MS-based quantification has become increasingly popular. It has been found that N-linked glycans exhibit similar signal responses, which enables the estimation of the relative proportions of each glycan in a mixture [7, 8]. However, for the quantification of the same glycan species between two or more different biological states, stable isotope labelling akin to common current proteomic methods has become widely used and typically follows one of two general ideas: Either, glycan labelling is performed by permethylation [9-11] or at the reducing end, e. g. using reductive amination [12-15], hydrazone [16] or oxime [17] formation. As interesting alternative, a metabolic labelling approach has also been devised [18], which enables stable isotope labelling of glycans *via* the biosynthetic incorporation of ^{15}N from glutamine into aminosugars.

During permethylation, stable isotopes can be incorporated at any hydroxyl, amine, and carboxyl moiety of glycan using light ($^{12}CH_3$) and heavy-labelled methyl iodide ($^{13}CH_3$, $^{12}CDH_2$, $^{12}CD_2H$ and/or $^{12}CD_3$) resulting in a mass shift of 1 to 3 Da per methylation site [9, 10]. A nominally isobaric labelling can be achieved when using a near isobaric pair of methyl iodide ($^{13}CH_3$ and $^{12}CH_2D$) [11]. The small mass increment of 0.002922 Da per methyl group adds up during permethylation. However, a resolving power of 50,000 (full width half maximum, FWHM) is required for base line separation of the two molecular species, limiting the approach to very high resolution mass spectrometers such as Orbitraps and FT-ICR instruments. Permethylation has long been used because of a number of inherent advantages such as the stabilization of acidic glycans during positive ion mode matrix-assisted laser desorption/ionization (MALDI) MS [19], an increased overall MS signal strength and the ability to separate the glycans by reversed phase chromatography. However, there are also distinct disadvantages. In order to achieve accurate quantification, the permethylation efficiency has to be precisely the same for the samples that are to be compared and the reaction has to be exhaustive. For instance, permethylation of a small glycan with 30 methylation sites would require a reaction efficiency of 99.9% at each site to achieve an overall yield of 97%. Any decrease in reaction efficiency (say 0.2%) would result in a significant (here 6%) decrease in the yield of the permethylated glycan. Obviously, the larger the glycan the more severe this issue becomes.

As stated earlier, the alternative to permethylation is to derivatize glycans at the reducing end. During reductive amination, the free aldehyde at the reducing end can be converted into an imine (or Schiff base) using deuterated or ^{13}C-isotope coded aromatic amines [13-15]. The imine is then reduced using sodium borohydride or similar reagents to stabilize the compound. This strategy is

straightforward (labelling one of one functional group) and, unlike for permethylation, results in a fixed mass increment for any glycan. In combination with differentially labeled aniline (D_0, D_4, D_8, D_{12}) this approach allows multiplexed quantification of up to four different samples in the same spectrum [14], but the 4 Da mass increment afforded by the label often necessitates deconvolution of overlapping isotope patterns to achieve accurate quantification.

Alternatively, stable isotopes can be introduced into glycans *via* hydrazone formation using isotope-coded hydrazides [16] or *via* oxime formation using aminooxy reagents [17]. In contrast to reductive amination, hydrazone and oxime formation do not require further sample clean-up prior to MS analysis. The simultaneous quantification of reducing end labelled neutral and acidic glycans requires prior stabilization of acidic glycans because the latter are prone to spontaneous gas phase decomposition involving the loss of sialic acid or CO_2 [20]. This stabilization can be elegantly achieved e. g. *via* selective methyl esterification [21].

Tandem mass tags (TMTs) have been originally developed for the isobaric labeling and quantification of peptides in proteomics experiments [22, 23]. TMTs have a modular structure consisting of a mass reporter group, a mass normalizer group and a reactive functional group (Figure 1A). In the standard configuration, differential TMT labelling results in isobaric molecules for which quantification is achieved *via* up to six different reporter ion signals in the low mass region of tandem mass spectra. TMT labelling reagents can also be configured to introduce a mass difference of the intact analyte molecule, which allows for quantification in ordinary (intact) mass spectra. Amine-reactive TMTs belong to the standard repertoire of quantitative proteomics methods [24], and, more recently, thiol-reactive TMTs have also become available for this purpose [25, 26].

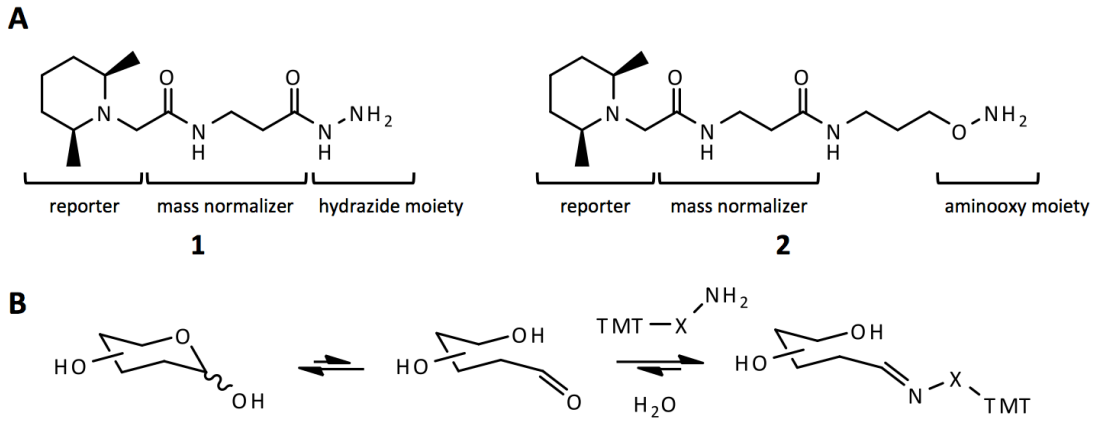

Figure 1 | Glyo-TMT reagents for reducing end-labelling of sugars
A Carbonyl-reactive TMT compounds used in this study. **B** Reaction scheme for hydrazone and oxime formation using carbonyl-reactive TMTs. X denotes nitrogen for the hydrazide or oxygen for the aminooxy reagent.

The current study is the first demonstration of carbonyl-reactive tandem mass tags (glyco-TMTs) as a novel approach for the quantification of N-linked glycans. Glyco-TMTs were investigated as hydrazide- and aminooxy-functionalized reagents (Figure 1A) and an extensive comparison revealed that aminooxy TMTs are superior to their hydrazide counterparts. In addition, the direct comparison of isobaric quantification vs. quantification using heavy/light TMT-labelled glycans and MALDI time-of-flight mass spectrometry disclosed advantages and shortcomings of either approach. To demonstrate the practical utility of the developed approach, the N-linked glycosylation profiles of

two isogenic human colon carcinoma cell lines were characterized in a quantitative fashion. The data revealed significant differences in the quantities of high abundant high-mannose glycans in the metastatic and non-metastatic lines. Collectively, the data suggest that glyco-TMT labelling is a promising new approach for the quantitative N-linked glycan profiling with significant potential in basic and biomedical research.

Experimental procedures

Preparation of glycans from model glycoproteins and cell lines

Chicken egg ovalbumin, bovine α_1-acid glycoprotein and disialylacto-N-tetraose (DSLNT) were from Sigma-Aldrich (Taufkirchen, Germany), and maltooctaose (MO) from Carbosynth Limited, UK, Human cancer cell lines SW480 and SW620 were cultured and lyzed following standard procedures. 500 µg protein from whole cell lysates or 100 µg of standard glycoproteins were reduced with dithiothreitol, alkylated with iodoacetamide, and digested with trypsin (Sigma-Aldrich). (Glyco-)peptides were purified *via* Sep-Pak C_{18} solid-phase extraction cartridges (Waters, Eschborn, Germany) and dried *in vacuo*. N-linked glycans were released from glycopeptides using peptide N-glycosylase F (PNGase F, Roche Diagnostics, Mannheim, Germany). Released N-linked glycans were separated from the peptide fraction using Sep-Pak C_{18} solid-phase extraction cartridges (Waters, Eschborn, Germany) and the glycan containing flow-through was dried *in vacuo*.

TMT-labelling and methyl esterification of standard glycans

Glycans from standard glycoproteins as well as standard oligosacharides were prepared at a concentration of 5 µg/µl glycoprotein or 100 µM oligosaccharide, respectively. TMT-labelling was performed in four different buffers containing 80% (v/v) methanol and (1) 50 mM TEAB, (2) only 80% (v/v) methanol, (3) 5% (v/v) acetic acid, and (4) 20% (v/v) acetic acid. Hydrazide or aminooxy-TMTs were added from a 100 mM TMT stock in dimethyl sulfoxide (DMSO) to the desired concentration (0.1 to 10 mM). The the labelling reaction was incubated at 75 °C for 4 hours under constant shaking. Reduction and HILIC purification of hydrazide-TMT labelled glycans was performed using sodium borohydride and ZIC-HILIC columns in StageTip format. Prior to MALDI TOF/TOF analyses, TMT-labelled glycans were dried down and reconstituted in ddH_2O.Methyl esterification of sialic acids using methyl iodide essentially followed the procedure described by Powell and Harvey [21]. Briefly, standard glycans from α_1-acid glycoprotein or DSLNT were converted into their sodium salts, dried *in vacuo* and dissolved in dry DMSO and methyl iodide (1:1 v/v). The reaction was incubated at 4 °C for 30 minutes, and glycans were directly analyzed from the reaction mixture by MALDI-TOF MS. The labelling efficiency of both, TMT-labelling and methylation was determined based on MALDI MS peak areas of sodium and potassium adducts of converted and unreacted standard glycans.

TMT-labelling and methyl esterification of whole cell glycans

SW480 and SW620 cells were harvested at three different passage numbers to obtain biological triplicates and subsequently subjected to cell lysis, protein digestion and glycan purification. Unlabelled, dry glycans were spiked with 5 pmol DSLNT and converted into their sodium form before the TMT-labelling was accomplished with aminooxy-TMT^0 or TMT^6. Heavy and light TMT-labelled glycans from the two cell lines were mixed in equal quantities (based on equal protein amounts) and dried *in vacuo*. Selective methylation of sialic acid-containing glycans was done as described above.

MALDI sample preparation and analysis

The sample solution (0.5 µl) was mixed with 2 µl of matrix (20mg/ml 2,5-dihydroxybenzoic acid [Bruker Daltonik, Bremen, Germany] in 30% [v/v] acetonitrile and 0.1% [v/v] trifluoroacetic acid and 7 mM triethylammonium citrate) on a stainless steel target (Bruker Daltonik, Bremen, Germany). MALDI spots from methylated glycans were allowed to crystallize overnight. Mass spectra were acquired in positive ion reflectron mode on an ultrafleXtreme MALDI-TOF/TOF mass spectrometer equipped with a 1 kHz Smartbeam-II laser (Bruker Daltonik, Bremen, Germany). Each spectrum was externally calibrated using the Peptide Calibration Standard II (Bruker Daltonik, Bremen, Germany) and the "cubic enhanced" calibration function. In addition, glycan spectra from cell lines were internally re-calibrated based on the signals from high-mannose glycans. Tandem mass spectra were acquired in positive ion reflectron mode using the LIFT technique [27] at an increased laser intensity (+40%) accumulating 1000 laser shots per parent ion spectrum and 10,000 shots per fragment spectrum. The isolation window for precursor ions was set to 0.5% of the precursor m/z value. Mass spectra were analyzed using the flexAnalysis software (version 3.3) (Bruker Daltonik, Bremen, Germany). Peak picking was performed manually and glycan quantification used the integrated peak areas. Glycan compositions were assigned based on their measured masses using the theoretical glycan library constructed by Kronewitter *et al.* [28].

Results and discussion

Novel carbonyl-reactive tandem mass tags for glycan quantification

Tandem mass tags have been exploited for the isobaric relative and absolute quantification of peptides in proteomic studies [24]. This study presents the assessment of two novel TMT compounds designed for stable isotope-based quantification in comparative glycoprotein analysis.

Table 1 | Isotope codes of TMT compounds

	Reagent composition	Reagent mass /Da	Induced mass shift /Da	Reporter ion composition	Reporter ion m/z
Hydrazide compounds					
TMT^0-126	$C_{12}H_{24}N_4O_2$	256.189926	238.179361	$[C_8H_{16}N]^+$	126.1277259
TMT^2-126	$^{13}C_1C_{11}H_{24}N_4O_2$	257.193281	239.182716	$[C_8H_{16}N]^+$	126.1277259
TMT^2-127	$^{13}C_1C_{11}H_{24}N_4O_2$	257.193281	239.182716	$[^{13}C_1C_7H_{16}N]^+$	127.1310808
TMT^6-127	$^{13}C_4C_8H_{24}{}^{15}N_1N_3O_2$	261.200380	243.189816	$[^{13}C_1C_7H_{16}N]^+$	127.1310808
Aminooxy compounds					
TMT^0-126	$C_{15}H_{30}N_4O_3$	314.231791	296.221226	$[C_8H_{16}N]^+$	126.1277259
TMT^6-128	$^{13}C_4C_{11}H_{30}{}^{15}N_1N_3O_3$	319.242245	301.231680	$[^{13}C_2C_6H_{16}N]^+$	128.1344356
TMT^6-130	$^{13}C_4C_{11}H_{30}{}^{15}N_1N_3O_3$	319.242245	301.231680	$[^{13}C_4C_4H_{16}N]^+$	130.1411453
TMT^6-131	$^{13}C_4C_{11}H_{30}{}^{15}N_1N_3O_3$	319.242245	301.231680	$[^{13}C_4C_4H_{16}{}^{15}N]^+$	131.1381802

Essentially, glyco-TMT reagents are derivatives of the original TMT compounds but are functionalized with carbonyl-reactive groups involving either hydrazide chemistry forming hydrazones or aminooxy chemistry leading to oxime conjugates (Figure 1). In addition, the glyco-

TMT compounds are coded with stable isotopes (Table 1) to enable i) isobaric quantification in tandem mass spectra and ii) quantification in MS1 spectra using heavy/light pairs.

The mass difference of 5.0105 Da between the light TMT^0 and the heavy TMT^6 reagents is sufficient to separate the isotopic patterns of practically all commonly existing N- glycans. When calculating all possible mass differences of a glycan library composed of 331 glycan compositions from the human N-linked serum glycome [28], only 14 out of 54,615 possible mass differences (0.026%) fall within a window of 4.5 and 5.5 Da. The possible compositional assignment ambiguity is therefore negligible.

TMT labelling efficiency and stabilization of acidic glycans

First labelling conditions for both, hydrazide and aminooxy TMTs as well as the stability of the labelled glycans under neutral and acidic conditions were assessed. High labelling efficiency could be obtained for both reagents (Figure 2A). In particular, quantitative aminooxy TMT labelling was achieved at sub-milimolar TMT concentration. To avoid acidic hydrolysis of TMT-sugar conjugates and additional sample preparation steps associated with some of the reaction conditions (Figure 2B), all subsequent labelling reactions were performed in 80% (v/v) methanol and spotted labelled samples onto MALDI plates immediately following an experiment. To achieve simultaneous detection of neutral and acidic glycans by MALDI MS, sialic acids were selectively esterified using methyl iodide as described by Powell and Harvey [21]. Notably, when combined with aminooxy TMT labelling for quantification, it is important to perform the TMT labelling before methylation. Otherwise, the procedure is inefficient and results in numerous by-products. A further important experimental detail, particularly for small quantities of (acidic) glycans, is to convert them into their sodiated form prior to TMT labelling and methylation (see also below).

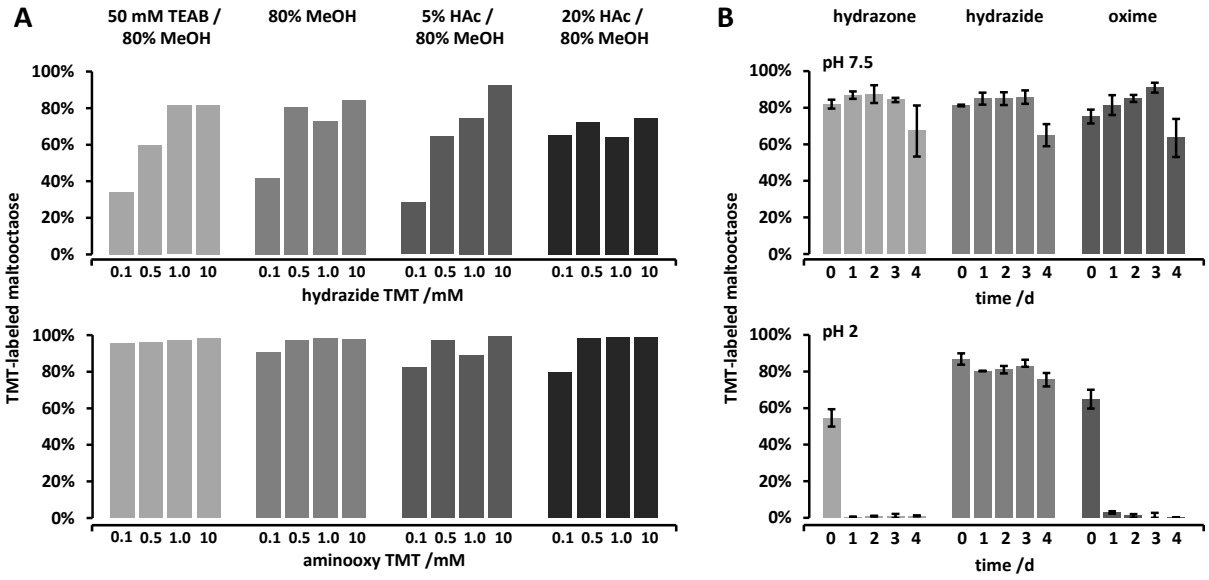

Figure 2 | Reaction efficiencies and stability of hydrazide and aminooxy TMT under different experimental conditions

A Reaction efficiency of hydrazide and aminooxy TMT in different buffer systems. **B** Stability of hydrazone, hydrazide and oxime TMT-maltooctaose conjugates. Labelling efficiency and stability was estimated using the MALDI MS peak areas of labelled and unlabelled maltooctaose.

Quantification of heavy/light-labelled glycans

In order to explore the quantification accuracy and dynamic range of glyco-TMT quantification, N-linked glycans from ovalbumin were labeled with light (TMT^0) and heavy (TMT^6) forms of both hydrazide and aminooxy TMT. Figure 3A shows a MALDI MS spectrum displaying the typical pattern of high mannose and hybrid glycans of this protein and its characterized impurities [29]. Notably, although the TMT label contains a basic tertiary amine and should thus be readily protonated, the TMT-conjugated glycans are almost exclusively observed as sodiated species.

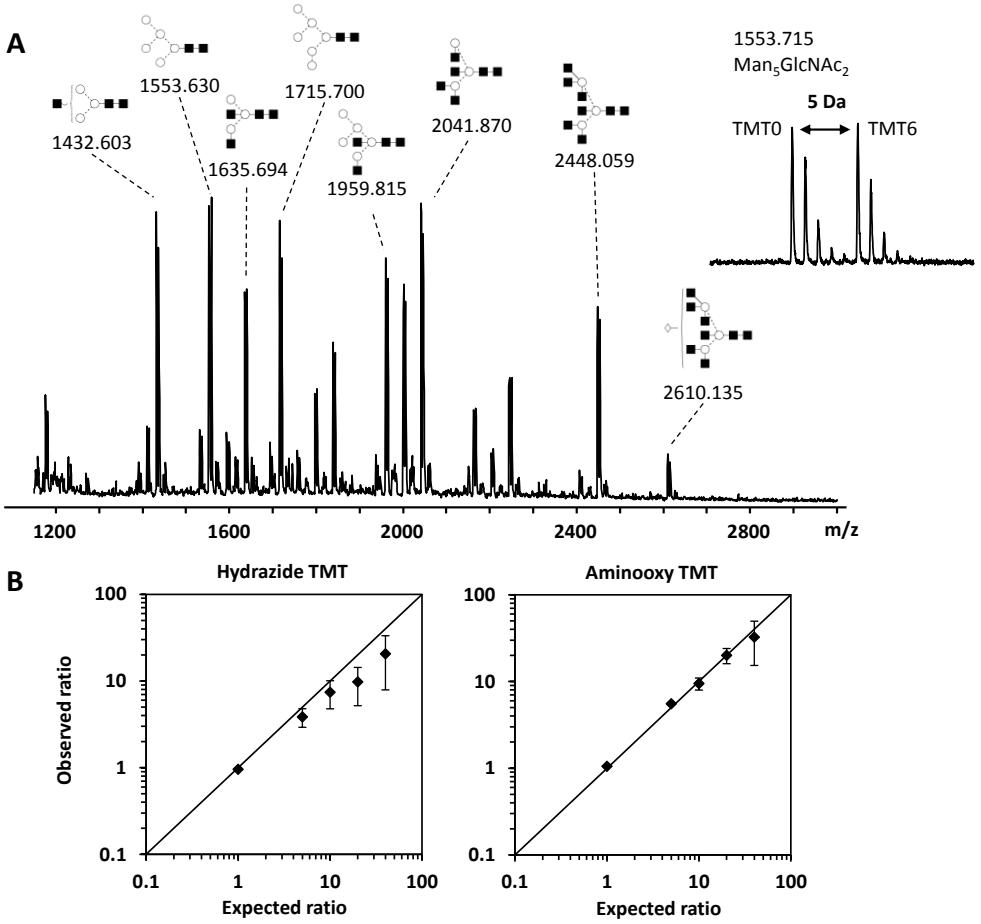

Figure 3 | Quantification of heavy and light labelled N-glycans from ovalbumin
A MALDI TOF spectrum of heavy and light labelled glycans (TMT^0/TMT^6 1:1). Structural assignments of glycans were taken from Harvey *et al.* [29]. The inset shows labelled $Man_5GlcNAc_2$. **B** Dynamic range and accuracy of hydrazide and aminooxy TMT quantification. N-linked glycans were mixed in known ratios (1:1, 1:5, 1:10, 1:20, and 1:40) and the experiment was performed in triplicates.

Heavy/light labelling with glyco-TMT enables accurate quantification in MALDI MS spectra. Coefficients of variations (CVs) for 1:1 ovalbumin mixtures were as low as 3% and 1% for hydrazide and aminooxy TMT, respectively. As one would expect, quantification accuracy decreases with increasing mixing ratio, and CVs of 36% (hydrazide) and 15% (aminooxy) were observed in the 1:10 samples. The accessible dynamic range of quantification with heavy/light glyco-TMT was determined to be 1:10 for hydrazide TMT and 1:20 for aminooxy TMT (Figure 3B). Beyond these values, ratio compression is increasingly observed. The quantification accuracy is primarily limited by two factors. First, the signal-to-noise ratio (S/N) of the MALDI TOF spectra and, second, the labelling efficiency. For the 1:20 (hydrazide TMT) and 1:40 (aminooxy TMT) samples, the intensities of the TMT-glycan

signals were below a S/N of 2 at which quantification is no longer reliably possible. Beyond this fundamental point, the differences in quantification accuracy and dynamic range between hydrazide and aminooxy TMT can (at least partially) be attributed to significant differences in labelling efficiency. The lower labelling efficiency observed for hydrazide TMT increases the spectrum complexity, which in turn may results in diminished quantification accuracy and ratio compression. Still, the dynamic range observed in MALDI TOF spectra of glyco-TMT labelled glycans is comparable to what is commonly observed in peptide quantification experiments. Even for shotgun proteomics studies using stable isotope labelling by amino acids in cell culture (SILAC) and LC-MS/MS analysis [30], ratios beyond 1:20 are rarely observed.

Quantification of isobarically labelled glycans

The accuracy and dynamic range of reporter ion-based quantification in tandem mass spectra using isobaric glyco-TMTs were assessed essentially as described for the heavy/light labelling but using the aminooxy TMT^6-128 and TMT^6-131 as well as the hydrazide TMT^2-126 and TMT^2-127 reagents. A conceptual advantage of isobaric tagging over heavy/light labelling is that the former does not lead to increased spectrum complexity because the precursor masses of the differentially labelled samples are exactly the same. Figure 4A shows a LIFT spectrum of an aminooxy TMT-labelled high-mannose glycan exhibiting the typical fragmentation pattern of a reducing end derivatized N-linked glycan [20]. A similar fragmentation pattern was observed for the hydrazide label (not shown). The depicted spectrum reveals a significant loss of the label during fragmentation by cleavage of the oxime bond. Such a loss is not unusual, and has, for instance, been reported by Lattova *et al.* [31]. It does, however, negatively impact on reporter ion quantification because it reduces the yield of these ions. Still, the quantification accuracy of the 1:1 samples is quite good for aminooxy TMT (CV < 10%), but rather poorer for hydrazide TMT (CV > 20%) (Figure 4B). In addition, the dynamic range of quantification is limited for both labels and substantial ratio compression can already be observed at mixing ratios as low as 1:5. Two reasons for the observed shortcomings can be identified. Owing to the high lability of O-glycosidic bonds in the gas phase, extensive fragmentation along the glycosidic bonds occurs and the recovery of TMT reporter ions is low (Figure 4A).

The low TMT ion intensity along with a low S/N inevitably compromises quantification accuracy and dynamic range. A particular shortcoming of the isobaric hydrazide TMT label is that the TMT^2-126 reporter ion interferes with a HexNAc fragment of the formula $[C_6H_8O_2N+H]^+$ [32] (mass difference of 0.0727 Da) which cannot be resolved by the medium resolution technique used here. Co-detection of both ions obviously leads to a distortion of the quantification data for this TMT channel. A further limitation is that the number of glycan precursors selected for tandem MS from a MALDI spectrum is generally limited to about 10 resulting in numerous detectable glycans that cannot be quantified. At this stage, it can be concluded that glycan quantification using heavy and light glyco-TMTs is superior to reporter ion quantification in tandem MS spectra.

A systematic investigation of other fragmentation techniques using MALDI ionization or extending it further to ESI-MS/MS was beyond the scope of this initial study. However, the above shortcomings may be (partially) overcome by a number of measures such as alternative fragmentation techniques (e. g. CID on qTOF instruments or HCD on Orbitrap platforms), that provide more efficient recovery of low mass reporter ions. High resolution MS/MS data will also be of benefit because it enables the resolution of interfering fragment ions. It may further be argued that a LC-MS/MS workflow could lead to improvements because it would enable the fragmentation of many more precursors and

avoid issues of signal suppression by the presence of too many analytes. And, finally, higher order MSn approaches akin to recently developed isobaric quantification methods in proteomics [33], may further improve quantitative accuracy. Clearly, future studies are required to fully explore the significant potential of isobaric tags for the quantification of glycans.

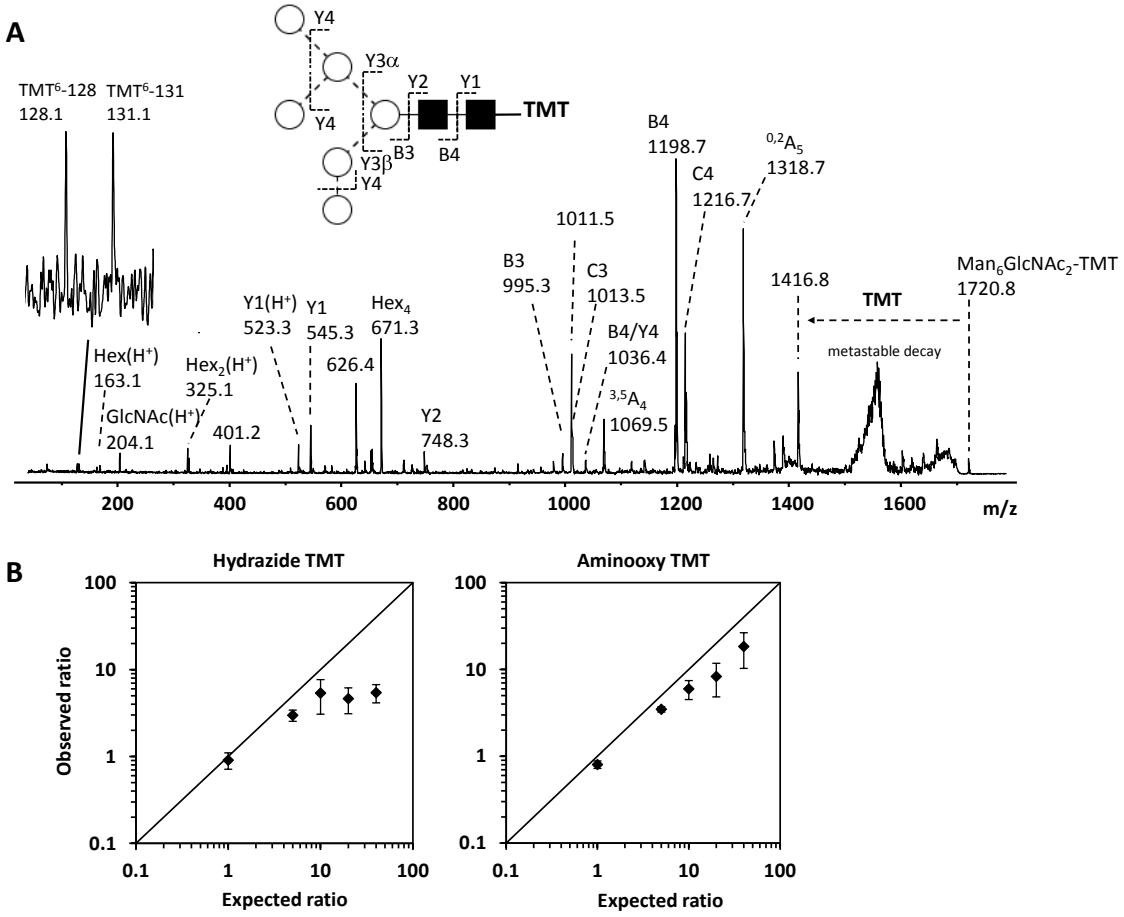

Figure 4 | Reporter ion-based quantification of isobarically labelled N-linked ovalbumin glycans
A MALDI TOF/TOF spectrum of labelled $Man_6GlcNAc_2$ (m/z 1720.708, aminooxy TMT^6-128/131, 1:1). Fragment ions are labelled following common nomenclature [34]. Ions are sodiated unless otherwise denoted. **B** Dynamic range and accuracy of hydrazide and aminooxy TMT quantification.

Quantitative glycan profiling of isogenic colorectal cancer cell lines

To demonstrate the practical utility of the developed glyco-TMT quantification approach, the N-linked glycan complement of two isogenic human colon carcinoma cell lines was profiled. The primary tumor cell line SW480 and its lymph node metastatic variant SW620 are a commonly used model system for studying metastasis and, given the potential involvement of cell surface glycosylation in cell migration and invasion of cancer cells [35], represents an interesting use case for the developed methodology. Overall, 41 heavy/light glycan pairs were detected that could be assigned to different glycan compositions based on their accurate masses. Qualitatively, the global glycan profiles of the two cancer cell lines are very similar (Figure 5A). However, while most glycans showed minor or no differences between both cell lines (Figure 5B), six glycan compositions exhibited statistically significant changes in abundance (t-test p-value < 0.05; Figure 5C and Table 2). Among these were three high-mannose glycans and two doubly fucosylated glycan compositions, which decreased significantly in the metastatic cell line, and one rather low abundant fucosylated

glycan exhibiting significantly increased levels. Both cell lines display a rather unexpectedly high proportion of high-mannose type glycans (60% and 50% of the total MS signal of SW480 and SW620 cells, respectively). In addition, a small but noticeable quantity of $Hex_{10}GlcNAc_2$ was observed, which corresponds to $Man_{7\text{-}9}GlcNAc_2$ with one to three terminal glucoses and which probably originates from the endoplasmatic reticulum (ER). Interestingly, high mannose sugars have also recently been reported as major glycan constituents of the cell surface of several leukemia cells [36], human embryonic stem cell lines [37] and significantly elevated levels of these glycans have been found on invasive and non-invasive breast cancer cell lines [38] as well as in human serum during breast cancer progression [39].

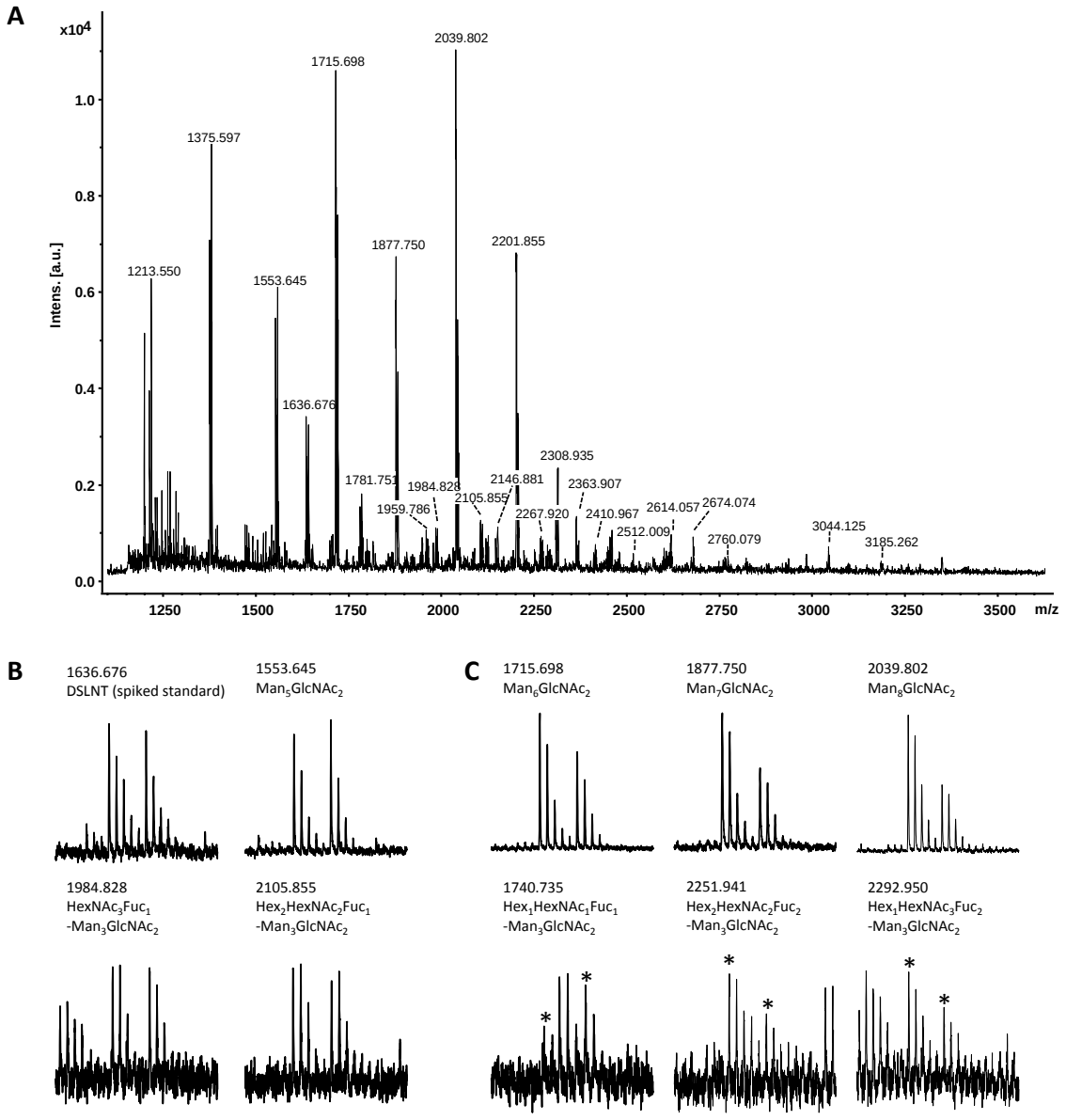

Figure 5 | Aminooxy TMT$^{0/6}$ quantification of neutral and acidic glycans from the colorectal primary tumor cancer cell line SW480 and its isogenic metastatic variant SW620

A MALDI MS spectrum of the mixed N-glycan pools. Based on signal intensity, the detected glycan compositions comprised 56% high mannose-glycans, 15% complex, 4% hybrid glycans and 25% compositions which can be assigned to both hybrid and complex glycans. **B** Expanded MALDI spectra for the spiked standard DSLNT and selected unregulated or **C** significantly regulated N-glycans. Light and heavy monoisotopic glycan signals are indicated (*) where required.

Table 2 | Significantly differential glycan compositions from SW480 and SW620 cells

Composition	Type	Normalized intensity (std. dev.) SW480	SW620	log2 ratio SW480/SW620	*p* value
$Man_6GlcNAc_2$	High Mannose	590 (17)	326 (74)	-0.9	0.032
$Man_7GlcNAc_2$	High Mannose	439 (3)	239 (29)	-0.9	0.010
$Man_8GlcNAc_2$	High Mannose	653 (102)	326 (75)	-1.0	0.026
$Hex_1HexNAc_1Fuc_1$-$Man_3GlcNAc_2$	Complex/Hybrid	13 (3)	34 (3)	1.3	0.003
$Hex_2HexNAc_2Fuc_2$-$Man_3GlcNAc_2$	Complex/Hybrid	48 (3)	37 (1)	-0.4	0.023
$Hex_1HexNAc_3Fuc_2$-$Man_3GlcNAc_2$	Complex/Hybrid	52 (8)	26 (7)	-1.0	0.025

There is currently no clear explanation for the observed differences in the N-linked glycan profiles of SW480 and SW620 cells but a number of discussion threads can be picked up from the literature. Multiple enzymes are responsible for the sequential trimming of N-linked Glc_3-capped high mannose glycans in the ER and Golgi [40]. Interestingly, mRNA expression data of two studies [41, 42] deposited in the Gene Expression Atlas [43] indicate a general down-regulation of key enzymes (e. g., α-glucosidase, ER α-mannosidase, Golgi α-mannosidase II) involved in high mannose trimming in the primary tumor cell line SW480 compared to SW620 is consistent with our observation of over-representation of high-mannose glycans in SW480 cells. Although the lines are isogenic, they display considerable differences in their migration and invasion behaviour. While the cell line derived from the primary tumor (SW480) shows the greater migration potential [44], the metastatic cell line is more invasive [45]. Several studies (e. g., [46, 47]) using inhibitors against high-mannose glycan trimming glucosidases and mannosidases revealed that treated cancer cells typically exhibit reduced adhesion to extracellular matrix components, which is a prerequisite for migratory behaviour. Apparently, artificially increased levels of high-mannose glycans also often alleviate invasiveness of cancer cells [48, 49]. Clearly, more work is required to understand how changes in gene and glycoprotein expression influences the behaviour of cancer cells, but it is tempting to speculate that quantitative N-glycan profiling may be able to serve as a global molecular readout for some of these cellular phenotypes.

Future directions

This work describes the first isobaric labelling reagents for glycans and demonstrates that the technology has merit. Future work will have to include systematic investigations of ionization and fragmentation techniques as well as the compatibility of TMT-labelled glycans with chromatographic and electrophoretic separation techniques. Although we were able to show that it is possible to profile glycans from cell lines, important future work should extend this to tissues including considerations about sample quantity and biological matrix. A particular technical area that needs improving is the development of appropriate software. In order to achieve accurate quantification, the data analysis including peak picking, glycan composition assignments and matching of heavy/light pairs was mostly performed manually in this work using the vendor's software and a standard spreadsheet application. This approach will likely not be scalable with increasing complexity of the data (e. g. as obtained from LC-MS experiments or with considerably higher numbers of samples and/or replicates) and, hence, software solutions will be required. Software for glycan analysis lags behind tools available for proteomics and there are a number of specific requirements that should be taken into account when analysing TMT-labelled sugars. Peak picking

should, in addition to the average elemental composition of glycans, also consider the elemental composition of the tag and the introduced mass shift to facilitate the matching of heavy/light pairs. The assignment of glycan compositions based on publically available databases such as GlycoMod [50] or theoretical glycan libraries has to consider that isobaric structural and compositional glycan isomers cannot be resolved, when only intact glycan spectra are available. Moreover, depending on mass accuracy and resolution, also different glycans with distinct exact masses may not be resolvable. Here, internal re-calibration of the m/z-axis based on unambiguously assigned glycan compositions might be helpful.

Conclusion

This study introduced carbonyl-reactive TMTs are novel tools for the quantitative analysis of protein glycosylation from single proteins through to complex biological samples such as cancer cell lines. The data show that this approach does not only enable a accurate quantification of N-glycans across a broad dynamic range, it also allows the quantification of neutral and sialylated glycans alongside without the need for permethylation. In conjunction with MALDI MS analysis, the approach also has the potential for substantial sample throughput, which would enable a multitude of investigations including but not limited to glyco-typing of model cell lines, primary (human) tissue and body fluids. Future work will address the potential of this approach for such applications and an extension to N-glycan profiling by LC-MS which likely offers a more in-depth coverage of the many compositional glycan isomers that exist in nature. As a result of this study, Thermo Fisher Scientific is planning to commercialize the reagents so that the technology becomes more broadly accessible.

Acknowledgments

The author is indebted to Patrick Neubert, who performed the initial characterization of glycoTMTs during his MSc thesis. The author further thanks Karsten Kuhn (Proteome Science, Frankfurt, Germany) and Chris Etienne and Ryan Boomgarden (Thermo Fisher Scientific Pierce Protein Research, Thermo Fisher Scientific, Rockford, IL, USA) for synthesis of glycoTMTs reagents and John C. Rogers (also Thermo) for insightful discussions.

Abbreviations

DSLNT	disialylacto-*N*-tetraose
ER	Endoplasmatic reticulum
FWHM	full width half maximum
glyco-TMT	carbonyl-reactive tandem mass tag
MALDI	matrix-assisted laser desorption/ionization
MS	mass spectrometry
S/N	signal-to-noise ratio
SCX	strong cation exchange
TEAB	triethylammonium bicarbonate
TOF	time-of-flight
TMT	tandem mass tag

References

1. Apweiler, R., Hermjakob, H., and Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1473, 4-8.
2. Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E., eds. (2009) *Essentials of Glycobiology*, Second Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
3. Packer, N. H., von der Lieth, C. W., Aoki-Kinoshita, K. F., Lebrilla, C. B., Paulson, J. C., Raman, R., Rudd, P., Sasisekharan, R., Taniguchi, N., and York, W. S. (2008) Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics* 8, 8-20.
4. Marino, K., Bones, J., Kattla, J. J., and Rudd, P. M. (2010) A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol* 6, 713-723.
5. Pan, S., Chen, R., Aebersold, R., and Brentnall, T. A. (2011) Mass spectrometry based glycoproteomics--from a proteomics perspective. *Mol Cell Proteomics* 10, R110 003251.
6. Wada, Y., Azadi, P., Costello, C. E., Dell, A., Dwek, R. A., Geyer, H., Geyer, R., Kakehi, K., Karlsson, N. G., Kato, K., Kawasaki, N., Khoo, K. H., Kim, S., Kondo, A., Lattova, E., Mechref, Y., Miyoshi, E., Nakamura, K., Narimatsu, H., Novotny, M. V., Packer, N. H., Perreault, H., Peter-Katalinic, J., Pohlentz, G., Reinhold, V. N., Rudd, P. M., Suzuki, A., and Taniguchi, N. (2007) Comparison of the methods for profiling glycoprotein glycans--HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study. *Glycobiology* 17, 411-422.
7. Zaia, J. (2004) Mass spectrometry of oligosaccharides. *Mass Spectrom Rev* 23, 161-227.
8. Naven, T. J., and Harvey, D. J. (1996) Effect of structure on the signal strength of oligosaccharides in matrix-assisted laser desorption/ionization mass spectrometry on time-of-flight and magnetic sector instruments. *Rapid Commun Mass Spectrom* 10, 1361-1366.
9. Alvarez-Manilla, G., Warren, N. L., Abney, T., Atwood, J., 3rd, Azadi, P., York, W. S., Pierce, M., and Orlando, R. (2007) Tools for glycomics: relative quantitation of glycans by isotopic permethylation using 13CH3I. *Glycobiology* 17, 677-687.
10. Kang, P., Mechref, Y., Kyselova, Z., Goetz, J. A., and Novotny, M. V. (2007) Comparative glycomic mapping through quantitative permethylation and stable-isotope labeling. *Anal Chem* 79, 6064-6073.
11. Atwood, J. A., 3rd, Cheng, L., Alvarez-Manilla, G., Warren, N. L., York, W. S., and Orlando, R. (2008) Quantitation by isobaric labeling: applications to glycomics. *J Proteome Res* 7, 367-374.
12. Yuan, J., Hashii, N., Kawasaki, N., Itoh, S., Kawanishi, T., and Hayakawa, T. (2005) Isotope tag method for quantitative analysis of carbohydrates by liquid chromatography-mass spectrometry. *J Chromatogr A* 1067, 145-152.
13. Bowman, M. J., and Zaia, J. (2007) Tags for the stable isotopic labeling of carbohydrates and quantitative analysis by mass spectrometry. *Anal Chem* 79, 5777-5784.
14. Bowman, M. J., and Zaia, J. (2010) Comparative glycomics using a tetraplex stable-isotope coded tag. *Anal Chem* 82, 3023-3031.
15. Xia, B., Feasley, C. L., Sachdev, G. P., Smith, D. F., and Cummings, R. D. (2009) Glycan reductive isotope labeling for quantitative glycomics. *Anal Biochem* 387, 162-170.
16. Walker, S. H., Budhathoki-Uprety, J., Novak, B. M., and Muddiman, D. C. (2011) Stable-isotope labeled hydrophobic hydrazide reagents for the relative quantification of N-linked glycans by electrospray ionization mass spectrometry. *Anal Chem* 83, 6738-6745.
17. Uematsu, R., Furukawa, J., Nakagawa, H., Shinohara, Y., Deguchi, K., Monde, K., and Nishimura, S. (2005) High throughput quantitative glycomics and glycoform-focused proteomics of murine dermis and epidermis. *Mol Cell Proteomics* 4, 1977-1989.

18. Orlando, R., Lim, J. M., Atwood, J. A., 3rd, Angel, P. M., Fang, M., Aoki, K., Alvarez-Manilla, G., Moremen, K. W., York, W. S., Tiemeyer, M., Pierce, M., Dalton, S., and Wells, L. (2009) IDAWG: Metabolic incorporation of stable isotope labels for quantitative glycomics of cultured cells. *J Proteome Res* 8, 3816-3823.
19. Morelle, W., and Michalski, J. C. (2007) Analysis of protein glycosylation by mass spectrometry. *Nat Protoc* 2, 1585-1602.
20. Harvey, D. J. (1999) Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrom Rev* 18, 349-450.
21. Powell, A. K., and Harvey, D. J. (1996) Stabilization of sialic acids in N-linked oligosaccharides and gangliosides for analysis by positive ion matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* 10, 1027-1032.
22. 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.
23. Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F., Burkhard, P. R., and Sanchez, J. C. (2008) Relative Quantification of Proteins in Human Cerebrospinal Fluids by MS/MS Using 6-Plex Isobaric Tags. *Anal Chem* 80, 2921-2931.
24. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 389, 1017-1031.
25. Giron, P., Dayon, L., Turck, N., Hoogland, C., and Sanchez, J. C. (2011) Quantitative analysis of human cerebrospinal fluid proteins using a combination of cysteine tagging and amine-reactive isobaric labeling. *J Proteome Res* 10, 249-258.
26. Murray, C. I., Uhrigshardt, H., O'Meally, R. N., Cole, R. N., and Van Eyk, J. E. (2011) Identification and quantification of S-nitrosylation by cysteine reactive tandem mass Tag switch assay. *Mol Cell Proteomics* 2012, 2.
27. Suckau, D., Resemann, A., Schuerenberg, M., Hufnagel, P., Franzen, J., and Holle, A. (2003) A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Anal Bioanal Chem* 376, 952-965.
28. Kronewitter, S. R., An, H. J., de Leoz, M. L., Lebrilla, C. B., Miyamoto, S., and Leiserowitz, G. S. (2009) The development of retrosynthetic glycan libraries to profile and classify the human serum N-linked glycome. *Proteomics* 9, 2986-2994.
29. Harvey, D. J., Wing, D. R., Kuster, B., and Wilson, I. B. (2000) Composition of N-linked carbohydrates from ovalbumin and co-purified glycoproteins. *J Am Soc Mass Spectrom* 11, 564-571.
30. 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.
31. Lattova, E., Snovida, S., Perreault, H., and Krokhin, O. (2005) Influence of the labeling group on ionization and fragmentation of carbohydrates in mass spectrometry. *J Am Soc Mass Spectrom* 16, 683-696.
32. Hahne, H., and Kuster, B. (2011) A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. *J Am Soc Mass Spectrom* 22, 931-942.
33. Wenger, C. D., Lee, M. V., Hebert, A. S., McAlister, G. C., Phanstiel, D. H., Westphall, M. S., and Coon, J. J. (2011) Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat Methods* 8, 933-935.
34. Domon, B., and Costello, C. E. (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal* 5, 397-409.
35. Zhao, Y. Y., Takahashi, M., Gu, J. G., Miyoshi, E., Matsumoto, A., Kitazume, S., and Taniguchi, N. (2008) Functional roles of N-glycans in cell signaling and cell adhesion in cancer. *Cancer Sci* 99, 1304-1310.

36. Nakano, M., Saldanha, R., Gobel, A., Kavallaris, M., and Packer, N. H. (2011) Identification of glycan structure alterations on cell membrane proteins in desoxyepothilone B resistant leukemia cells. *Mol Cell Proteomics* 10, M111 009001.
37. An, H. J., Gip, P., Kim, J., Wu, S., Park, K. W., McVaugh, C. T., Schaffer, D. V., Bertozzi, C. R., and Lebrilla, C. B. (2011) Extensive determination of glycan heterogeneity reveals an unusual abundance of high-mannose glycans in enriched plasma membranes of human embryonic stem cells. *Mol Cell Proteomics* [epub ahead of print 2011/12/08].
38. Goetz, J. A., Mechref, Y., Kang, P., Jeng, M. H., and Novotny, M. V. (2009) Glycomic profiling of invasive and non-invasive breast cancer cells. *Glycoconj J* 26, 117-131.
39. de Leoz, M. L., Young, L. J., An, H. J., Kronewitter, S. R., Kim, J., Miyamoto, S., Borowsky, A. D., Chew, H. K., and Lebrilla, C. B. (2011) High-mannose glycans are elevated during breast cancer progression. *Mol Cell Proteomics* 10, M110 002717.
40. Stanley, P., Schachter, H., and Taniguchi, N. (2009) N-Glycans. In: Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E., eds. *Essentials of Glycobiology*, Second Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
41. Provenzani, A., Fronza, R., Loreni, F., Pascale, A., Amadio, M., and Quattrone, A. (2006) Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. *Carcinogenesis* 27, 1323-1333.
42. Jakobsen, C. H., Storvold, G. L., Bremseth, H., Follestad, T., Sand, K., Mack, M., Olsen, K. S., Lundemo, A. G., Iversen, J. G., Krokan, H. E., and Schonberg, S. A. (2008) DHA induces ER stress and growth arrest in human colon cancer cells: associations with cholesterol and calcium homeostasis. *J Lipid Res* 49, 2089-2100.
43. Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N., Kurnosov, P., Malone, J., Melnichuk, O., Petryszak, R., Pultsin, N., Rustici, G., Tikhonov, A., Travillian, R. S., Williams, E., Zorin, A., Parkinson, H., and Brazma, A. (2012) Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res* 40, D1077-1081.
44. Kubens, B. S., and Zanker, K. S. (1998) Differences in the migration capacity of primary human colon carcinoma cells (SW480) and their lymph node metastatic derivatives (SW620). *Cancer Lett* 131, 55-64.
45. Kim, H. R., Wheeler, M. A., Wilson, C. M., Iida, J., Eng, D., Simpson, M. A., McCarthy, J. B., and Bullard, K. M. (2004) Hyaluronan facilitates invasion of colon carcinoma cells in vitro via interaction with CD44. *Cancer Res* 64, 4569-4576.
46. von Lampe, B., Stallmach, A., and Riecken, E. O. (1993) Altered glycosylation of integrin adhesion molecules in colorectal cancer cells and decreased adhesion to the extracellular matrix. *Gut* 34, 829-836.
47. Atsumi, S., Nosaka, C., Ochi, Y., Iinuma, H., and Umezawa, K. (1993) Inhibition of experimental metastasis by an alpha-glucosidase inhibitor, 1,6-epi-cyclophellitol. *Cancer Res* 53, 4896-4899.
48. Reddy, B. V., and Kalraiya, R. D. (2006) Sialilated beta1,6 branched N-oligosaccharides modulate adhesion, chemotaxis and motility of melanoma cells: Effect on invasion and spontaneous metastasis properties. *Biochim Biophys Acta* 1760, 1393-1402.
49. Seftor, R. E., Seftor, E. A., Grimes, W. J., Liotta, L. A., Stetler-Stevenson, W. G., Welch, D. R., and Hendrix, M. J. (1991) Human melanoma cell invasion is inhibited in vitro by swainsonine and deoxymannojirimycin with a concomitant decrease in collagenase IV expression. *Melanoma Res* 1, 43-54.
50. Cooper, C. A., Gasteiger, E., and Packer, N. H. (2001) GlycoMod--a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* 1, 340-349.

General Conclusions

General Conclusions

Mass spectrometry has evolved as the key technology for the proteome-wide analysis of O-GlcNAc-modified proteins and N-linked glycosylation. The primary objective of this thesis was to develop novel mass spectrometry (MS)-based approaches for studying N-linked and O-GlcNAc protein glycosylation. This thesis addressed three different issues of current technologies, namely the detection and identification of O-GlcNAc peptides from tandem MS data, the enrichment of O-GlcNAc proteins from complex samples, and the MS-based quantification of N-linked glycans.

O-GlcNAc is a regulatory post-translational modification (PTM) with emerging roles in a variety of cellular processes and associated with acquired diseases such as cancer, diabetes type II and Alzheimer's. Moreover, O-GlcNAc exhibits an intimate interplay with protein phosphorylation and, in some cases, also with other important PTMs. However, despite its biological relevance, the analysis of O-GlcNAc-modified proteins remains highly challenging.

Undoubtedly, the detection of O-GlcNAc peptides by tandem MS is impaired by the high lability of the O-glycosidic bond in the gas phase. In particular, collision-induced dissociation (CID)-based fragmentation techniques suffer from a low yield of sequence-specific fragment ions, which are required for the identification and site localization of O-GlcNAc peptides. However, O-GlcNAc peptides also give rise to abundant O-GlcNAc-specific fragment ions, including the GlcNAc oxonium ions, numerous fragments thereof and 'charge loss' species. Together, these fragment ions constitute a highly specific signature, which can be utilized to detect tandem mass spectra of O-GlcNAc peptides. In order to obtain an objective measure for the presence or absence of the O-GlcNAc modification on the precursor ion, a scoring scheme was developed based on these features. This score, termed Oscore, enables the automated interrogation of tandem mass spectra and ranks spectra according to their probability of representing an O-GlcNAc spectrum. Although the Oscore readily outperforms alternative approaches to classify tandem mass spectra in O-GlcNAc and non-O-GlcNAc spectra, it is most powerful in combination with high resolution/high mass accuracy tandem mass spectra. This is mostly because the number of (intense) interfering peptide fragment ions is considerably lower compared to low resolution mass spectrometry techniques. Moreover, the Oscore can be readily extended to other labile modifications which give rise to intense reporter ions. In particular, N-linked or O-linked glycopeptide fragmentation results in abundant monosaccharide oxonium ions. And in combination with smart data-dependent acquisition schemes akin to commonly employed decision-tree approaches, the Oscore might significantly improve the identification of O-GlcNAc peptides from complex samples.

In combination with higher energy collision-induced dissociation (HCD), the Oscore-based re-analysis of publically available large-scale proteomic and phosphoproteomic data sets enabled the identification of hundreds of O-GlcNAc peptides and more than hundred. The results from this study suggest that O-GlcNAcylation is even more abundant than previously anticipated, but considerably less frequent than phosphorylation. At the same time, it could be observed that the occupancy of many O-GlcNAc sites is rather high, indicating that many of the observed (i. e., abundant) O-GlcNAc proteins are stably modified under physiological conditions. Somewhat in contrast to common notion, the discovery of numerous doubly modified peptides (i. e., peptides with one or multiple O-GlcNAc and phosphate moieties) suggests that O-GlcNAc and phosphorylation are not necessarily mutually exclusive, but can occur simultaneously at adjacent sites.

Another remarkable link between O-GlcNAc and phosphorylation has been discovered during the re-analysis of a publically available mouse brain phosphoproteomic data set. Besides numerous O-GlcNAc and phosphoproteins, 23 peptides corresponding to 11 proteins could be discovered, which were modified with a phosphorylated O-GlcNAc moiety, a modification so far only known for a single rat brain protein. The comparison to a GlcNAc-6-phosphate standard and the systematic dissection of major fragmentation routes of this modification strongly suggest that the phosphorylated O-GlcNAc moiety is actually O-GlcNAc-6-phosphate. By re-analyzing mass spectrometric data from human embryonic and induced pluripotent stem cells, our study also identified Zinc finger protein 462 as the first human O-GlcNAc-6-phosphate modified protein, suggesting that O-GlcNAc-6-phosphate is a general post-translation modification of mammalian proteins. O-GlcNAc-6-phosphate appears to be intimately associated with O-GlcNAc as almost all O-GlcNAc-6-phosphate sites are known O-GlcNAc sites. This immediately poses the question how biosynthesis of the O-GlcNAc-6-phosphate modification proceeds. Based on the given data, the most likely scenario is a two-step process involving OGT and one or multiple yet unknown kinases. A candidate kinase might be the metabolic N-acetylglucosamine kinase (NAGK), an enzyme of the hexosamine salvage pathway. A one-step attachment process appears unlikely in light of the fact there is currently no evidence of the existence of UDP-GlcNAc-6-phosphate in cells. Clearly, either mechanism indicates that there are yet undiscovered enzymes which catalyze the attachment reaction. The removal of O-GlcNAc-6-phosphate could be, in principle, catalyzed by OGA, which has a relatively large active site pocket that tolerates various GlcNAc-related substrates including azide-tagged O-GlcNAc [1]. But also here a two-stage process involving one or multiple yet unknown O-GlcNAc-6-phosphatases and OGA is possible. Clearly, future work does not only involve the identification of key enzymes for O-GlcNAc-6-phosphate attachment and removal, but also the development of efficient enrichment and detection approaches to study possible functional roles of this novel mammalian post-translational modification.

Even though some abundant O-GlcNAc proteins can be identified without prior biochemical enrichment, the proteome-wide analysis necessitates efficient enrichment approaches, which are not widely available yet. Metabolic labeling of O-GlcNAc proteins with azide-tagged GlcNAc followed by Cu(I)-catalyzed azide/alkyne Click chemistry (CuAAC) using alkyne-tagged fluorescent dyes is a commonly employed method to globally visualize O-GlcNAc proteins from cells in culture. In combination with affinity probes, this approach is also gaining popularity for the enrichment of O-GlcNAc proteins, but so far did not afford the identification of a single modified site. Metabolic labeling, CuAAC and a commercially alkyne-resin was employed in combination to purify O-GlcNAc proteins from cell lysates. On-resin proteolysis routinely enabled the identification of >1000 O-GlcNAc proteins including numerous important transcription factors and other low abundant proteins. Subsequent elution of covalently-resin bound O-GlcNAc peptides using selective β-elimination concomitantly tagged the former O-GlcNAc site and enabled the identification of >100 O-GlcNAc sites. Clearly, further work is required to map O-GlcNAc sites for most of the identified O-GlcNAc proteins. However, this approach has significant future potential for the study of the various emerging regulatory roles of O-GlcNAc and will be readily adaptable by other laboratories. Moreover, the method can be easily extended to purify chemoenzymatically tagged O-GlcNAc proteins as well as any other type of azide-tagged protein or modification thereof.

N-linked protein glycosylation is involved in a variety of fundamental biological processes. It is, therefore, not surprising that aberrant glycosylation of secreted and membrane proteins has been found in numerous inherited and acquired diseases and that altered glycosylation profiles may serve as disease marker or reveal potential drug targets. Chapter 6 delineates how carbonyl-reactive tandem mass tags (glycoTMTs) in conjunction with matrix-assisted laser desorption/ionization (MALDI) MS can be used to study and quantify N-linked protein glycosylation from single proteins to complex biological samples such as cancer cell lines. The developed approach does not only enable accurate quantification of N-linked glycans across a broad dynamic range, it also affords the concomitant quantification of neutral and acidic glycans without the need for permethylation. In addition, this study served as initial characterization of glycoTMTs for the isobaric quantification of glycans, which is bears significant potential. Clearly, glycoTMTs also have substantial potential for high throughput applications in conjunction with MALDI MS, for the in-depth quantification of glycan isoforms when coupled to graphitized carbon or hydrophilic interaction chromatography LC-MS as well as for the quantification of O-linked glycans when released by non-reducing β-elimination.

The study of protein glycosylation has seen considerable progress during the time I was working on my thesis. For example, 261 research articles on O-GlcNAc have been published between January 2009 and June 2012, representing about 46% of all O-GlcNAc articles indexed in PubMed. Novel technologies now enable the identification more than thousand O-GlcNAc sites [2, 3] or the quantification of *in vivo* O-GlcNAc levels and stoichiometry of proteins [4]. The structures of OGT and its complex with a peptide substrate have been reported [5] and will significantly facilitate the investigation of OGT's function or its potential as therapeutic target. The discovery of O-GlcNAc as part of the 'histone code' [6, 7], its emerging role in epigenetics [8], the discovery of the O-GlcNAc modification of cyclic AMP-response element binding protein (CREB) and its implications on gene expression and memory formation [9] as well as the discovery of O-GlcNAc-modified master regulators of pluripotency and stem cell renewal [10, 11] established novel possible roles of O-GlcNAc in essential cellular and physiological functions. These and other findings underscore the hypothesis that O-GlcNAc regulation represents a superimposed level of cellular signaling and modulates a variety of processes, presumably in response to nutrient availability and disease.

Abbreviations

CID	collision-induced dissociation
CuAAC	Cu(I)-catalyzed azide/alkyne Click chemistry
glycoTMT	carbonyl-reactive tandem mass tag
HBP	hexosamine biosynthetic pathway
HCD	higher energy-collision induced dissociation
MALDI	matrix-assisted laser desorption/ionization
MS	mass spectrometry
NAGK	N-acetylglucosamin kinase
OGA	nuclear/cytoplasmic O-GlcNAcase/acetyltransferase
O-GlcNAc	β-N-acetylglucosamine
OGT	O-GlcNAc transferase
PTM	post-translational modification

References

1. Macauley, M. S., Chan, J., Zandberg, W. F., He, Y., Whitworth, G. A., Stubbs, K. A., Yuzwa, S. A., Bennet, A. J., Varki, A., Davies, G. J., and Vocadlo, D. J. (2012) Metabolism of vertebrate amino sugars with N-glycolyl groups: intracellular O-GlcNGc, UDP-GlcNGc, and the biochemical and structural rationale for the substrate tolerance of O-GlcNAcase. *J Biol Chem* [epub ahead of print 2012/06/14].
2. Alfaro, J. F., Gong, C. X., Monroe, M. E., Aldrich, J. T., Clauss, T. R., Purvine, S. O., Wang, Z., Camp, D. G., 2nd, Shabanowitz, J., Stanley, P., Hart, G. W., Hunt, D. F., Yang, F., and Smith, R. D. (2012) Tandem mass spectrometry identifies many mouse brain O-GlcNAcylated proteins including EGF domain-specific O-GlcNAc transferase targets. *Proc Natl Acad Sci U S A* 109, 7280-7285.
3. Trinidad, J. C., Barkan, D. T., Gulledge, B. F., Thalhammer, A., Sali, A., Schoepfer, R., and Burlingame, A. L. (2012) Global identification and characterization of both O-GlcNAcylation and phosphorylation at the murine synapse. *Mol Cell Proteomics* [epub ahead of print 2012/05/31].
4. Rexach, J. E., Rogers, C. J., Yu, S. H., Tao, J., Sun, Y. E., and Hsieh-Wilson, L. C. (2010) Quantification of O-glycosylation stoichiometry and dynamics using resolvable mass tags. *Nat Chem Biol* 6, 645-651.
5. Lazarus, M. B., Nam, Y., Jiang, J., Sliz, P., and Walker, S. (2011) Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* 469, 564-567.
6. Sakabe, K., Wang, Z., and Hart, G. W. (2010) Beta-N-acetylglucosamine (O-GlcNAc) is part of the histone code. *Proc Natl Acad Sci U S A* 107, 19915-19920.
7. Fujiki, R., Hashiba, W., Sekine, H., Yokoyama, A., Chikanishi, T., Ito, S., Imai, Y., Kim, J., He, H. H., Igarashi, K., Kanno, J., Ohtake, F., Kitagawa, H., Roeder, R. G., Brown, M., and Kato, S. (2011) GlcNAcylation of histone H2B facilitates its monoubiquitination. *Nature* 480, 557-560.
8. Capotosti, F., Guernier, S., Lammers, F., Waridel, P., Cai, Y., Jin, J., Conaway, J. W., Conaway, R. C., and Herr, W. (2011) O-GlcNAc transferase catalyzes site-specific proteolysis of HCF-1. *Cell* 144, 376-388.
9. Rexach, J. E., Clark, P. M., Mason, D. E., Neve, R. L., Peters, E. C., and Hsieh-Wilson, L. C. (2012) Dynamic O-GlcNAc modification regulates CREB-mediated gene expression and memory formation. *Nat Chem Biol* 8, 253-261.
10. Myers, S. A., Panning, B., and Burlingame, A. L. (2011) Polycomb repressive complex 2 is necessary for the normal site-specific O-GlcNAc distribution in mouse embryonic stem cells. *Proc Natl Acad Sci USA* 108, 9490-9495.
11. Jang, H., Kim, T. W., Yoon, S., Choi, S. Y., Kang, T. W., Kim, S. Y., Kwon, Y. W., Cho, E. J., and Youn, H. D. (2012) O-GlcNAc Regulates Pluripotency and Reprogramming by Directly Acting on Core Components of the Pluripotency Network. *Cell Stem Cell* 11, 62-74.

List of publications

[1] **Hahne H**., Kuster B. (2011) A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. *J. Am. Soc. Mass Spectrom.* 22(5):931-42

[2] **Hahne H.**, Neubert P., Kuhn K., Etienne C., Bomgarden R., Rogers J. C., Kuster B. (2012) Carbonyl-reactive tandem mass tags for the proteome-wide quantification of N-linked glycans. *Anal. Chem*. 84(8):3716-24

[3] **Hahne H.**, Gholami A. M., Kuster B. (2012) Discovery of O-GlcNAc-modified proteins in published large-scale proteome data *Mol. Cell. Proteomics*, online published 1 June 2012

[4] **Hahne H.**, Kuster B. Discovery of O-GlcNAc-6-phosphate-modified proteins in large-scale phosphoproteomics data *Mol. Cell. Proteomics,* online published 23 July 2012

[5] **Hahne H.**, Sobotzki N., Nyberg T., Helm D., van Aalten D., Agnew B., Kuster B. Proteome wide purification and identification of O-GlcNAc modified proteins using Click chemistry and mass spectrometry [manuscript submitted]

Danksagung | Acknowlgedment

Keiner promoviert allein und diese Dissertation wäre ohne die vielfältigen Beträge von Kollegen, Freunden und Familie nicht möglich gewesen. An dieser Stelle möchte ich daher alldenjenigen meinen Dank erweisen, die in den letzten dreineinhalb Jahren zu dieser Dissertation einen grossen oder vielleicht auch nur einen kleinen Beitrag geleistet haben.

Zuvorderst gilt mein Dank natürlich Bernhard, der als Doktorvater nicht nur daran schuld ist, dass diese Dissertation über O-GlcNAc geschrieben wurde – eine posttranslationale Modifikation, von der ich bis dato nicht wusste, dass es sie gibt –, sondern der auch jederzeit hervorragende Ideen hatte und gute von schlechten zu unterscheiden wusste.

Natürlich gilt mein Dank auch Prof. Dr. Matthias Mann – dafür, dass er sich bereiterklärt hat, das Zweitgutachten zu übernehmen.

Desweiteren gilt mein Dank meinen ehemaligen Studenten, die als Forschungspraktikanten und während ihrer Masterarbeiten einen wichtigen Beitrag zu dieser Dissertation geleistet haben. Nadine wird hoffentlich bald für ihre hervorragende Arbeit, die die Grundlage für Kapitel 5 gelegt hat, mit einem JPR-Paper belohnt. Auch Patricks Experimente rund um Tandem Mass Tags zur Quantifizierung von Glykanen haben aus gutem Grund ihren Weg in diese Dissertation und eine gemeinsame Publikation gefunden. Dominics Ergebnisse sind zwar nicht in diese Arbeit eingeflossen, bilden stattdessen aber die Grundlage für einen hoffentlich erfolgreichen Drittmittelantrag.

Darüberhinaus gilt mein Dank meiner Vielzahl an Kooperationspartnern für das Zurverfügungstellen von wertvollen Reagentien und aufschlussreiche Diskussionen, auch wenn am Ende nicht alles davon den Weg in diese Dissertation geschafft hat: Tamara Nyberg und Brian Agnew für das Alkin-Harz (Life Technologies); Karsten Kuhn (Proteome Science), Chris Etienne, Ryan Boomgarden und John C. Rogers (Thermo Fisher Scientific) für die glycoTMT-Reagentien; Chin-Fen Teo, Margreet Wolfert und Geert-Jan Boons (Complex Carbohydrate Research Center, University of Georgia) für O-GlcNAc-Antikörper; Dirk Eick und Roland Schüller (Helmholtz Zentrum München) für RNA Polymerase II IPs; Juri Rappsilber und Adam Belsom (TU Berlin) für die BAC-Säule; und nicht zu letzt Steven Verhelst und Yinliang Yang (TUM) für ihre Click-Reagentien.

Mein Dank gilt natürlich auch der Studientstiftung des deutschen Volkes für ideelle und finanzielle Förderung im Rahmen meines Promotionsstipendiums und meinen Vertrauensdozenten Eckhard Wolf und Anna Friedl für ihre freundliche Unterstützung. Großzügig gewährte finanzielle Förderung von Tagungsreisen in In- und Ausland kam neben der Studienstiftung auch von der Gesellschaft für Biochemie und Molekularbiologie (GBM) und der TUM Graduate School/Graduiertenzentrum Weihenstephan.

Besonderer Dank gilt allen, die am Lehrstuhl Massenspektrometer oder Server oder andere Dinge am Laufen halten und immer für lehrreiche Diskussionen zu haben sind: Simone, Amin, Guillaume, Zhixiang, Fiona, Harald, Stefan, Ben und Dominic, ausserdem Andrea, Michaela und Andy für ihre Unterstützung in Sachen Zellkultur, Peptidsynthese und in-Gel Verdau.

Zu guter Letzt möchte ich meinen Großeltern, meiner Familie und Sandras Familie von Herzen für ihre beständige Unterstützung danken. Dabei gilt natürlich Sandra, Beeke und Merle mein ganz besonderer Dank für die vielen, vielen Wochenenden und Abende, die sie mit großer Langmut ohne mich verbracht haben, und für die vielen, vielen schönen Momente, die wir in den letzten dreieinhalb Jahren trotzdem zusammen hatten.

Curriculum vitae

Personal Information

E-mail	hannes.hahne@gmail.com
Phone	+49 8161 4924615 (home), +49 171 5072172 (mobile)
Address	Jagdstrasse 13, 85356 Freising, Germany
Date of birth	06.10.1981
Place of birth	Kassel

Professional Experience

since 01/2009	**Doctoral candidate** at the Chair of Proteomics and Bioanalytics, Technische Universität München Thesis in biochemistry under the supervision of Prof. Dr. Bernhard Küster *Studies towards the proteome-wide detection, identification and quantification of protein glycosylation*
01/2008 – 09/2008	**Research associate** at the Institute for Microbiology, University of Greifswald

Education

10/2006 – 10/2007	**Diploma thesis** at the Institute for Microbiology, University of Greifswald under the supervision of Prof. Michael Hecker
10/2002 – 10/2007	**Studies in biochemistry** at the University of Greifswald
01/2006 – 04/2006	**Research internship** at the Institute for Cell and Molecular Biosciences, Newcastle University, UK with Prof. Dr. Colin Harwood (Molecular Microbiology)
07/2005 – 09/2005	**Research internship** at the Friederich-Löffler-Institut/Federal Research Institute for Animal Health, Greifswald-Riems with Prof. Martin Groschup

Membership

since 2004	German Society for Biochemistry and Molecular Biology (GBM)

Awards

01/2009	Doctoral scholarship of the Studienstiftung des deutschen Volkes e. V.
09/2004	Awarded among the four best pre-diploma biochemistry students in Greifswald by the Gesellschaft Deutscher Chemiker (GDCh)