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

Unveiling the eukaryotic N-glycoproteome

Dorota Zielinska

Aus
Leszno, Polen

2011

Erklärung

Diese Dissertation wurde im Sinne von § 13 Abs. 3 bzw. 4 der Promotionsordnung vom 29. Januar 1998 (in der Fassung der sechsten Änderungssatzung vom 16. August 2010) von Herrn Prof. Dr. Matthias Mann betreut.

Ehrenwörtliche Versicherung

Diese Dissertation wurde selbständig, ohne unerlaubte Hilfe erarbeitet.

München, am 10.06.2011

Dorota Zielinska
Digitally signed by Dorota Zielinska
DN: cn=Dorota Zielinska, o, ou, email=zielinsk@biochem.mpg.de, c=DK
Date: 2011.08.30 21:33:46 +02'00'
....

Dissertation eingereicht am 10.06.2011

1. Gutachter Prof. Dr. Matthias Mann

2. Gutachter Prof. Dr. Jacek R. Wiśniewski

Mündliche Prüfung am 27.07.2011

For my dear parents

Krystyna and Mieczyslaw Zielinscy

'As an experimentalist, you can go through life kicking over a lot of stones,

and, if you're lucky, you'll find something.'

John Fenn

Contents

SUMMARY 1
INTRODUCTION 5
  N-Glycosylation versus Other Post-translational Protein Modifications 5
  Biology of N-Glycosylation 6
  Shotgun Proteomics for Mapping of N-Glycosylation 7
    Sample Preparation 9
    Quantification 10
    Enrichment 11
    Fractionation 13
    Mass Spectrometry 13
    Peptide Detection 15
    MS Data Processing 16
    Bioinformatic Analysis 17
PUBLICATIONS 19
  Publication I: Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints 21
  Publication II: Evolution of N-glycosylation 35
  Publication III: Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry 43
  Publication IV: Proteomic and N-Glycoproteomic Profiling of Colorectal Cancer ... 59
  Publication V: Comparison of Ultrafiltration Units for Proteomic and N-Glycoproteomic Analysis by the FASP Method 67
CONCLUDING REMARKS AND PERSPECTIVES 73
REFERENCES 75
ACKNOWLEDGMENTS 81
APPENDIX: PhD publications not related to N-glycosylation 83
  Appendix I: Caenorhabditis elegans Has a Phosphoproteome Atypical for Metazoans That Is Enriched in Developmental and Sex Dermination Proteins 85
  Appendix II: Evolutionary Constraints of Phosphorylation in Eukaryotes, Prokaryotes and Mitochondria 99
CURRICULUM VITAE 113

SUMMARY

Among the approximately 200 known post-translational protein modifications, glycosylation is one of the most common ones in eukaryotes. Glycosylated proteins play a major role in cell-cell and receptor-ligand interactions, immune response, apoptosis, angiogenesis and pathogenesis of diseases (Varki, 1993; Woods et al., 1994). Despite great biological and clinical interest, our knowledge of *in vivo* N-glycosylation sites - a prerequisite for detailed functional understanding - is still very limited. Although in the past decade it has become possible to detect several thousands of phosphorylation sites and to illustrate the kinetics of phosphorylation in complex biological systems (Huttlin et al., 2010; Olsen et al., 2006; Wisniewski et al., 2010), large-scale N-glycoproteomics has so far remained unexplored. Due to the lack of large-scale data, our understanding of the extent and evolution of N-glycoproteomes is very limited.

The goal of our work was to develop a method of in-depth mapping of N-glycosylation with very high accuracy using recent advances in proteomics technology and to apply it to various biological questions. We developed a method based on the 'Filter Aided Sample Preparation' (FASP) technology (Wisniewski et al., 2009b) that allows highly efficient capture of glycopeptides, even from membrane proteins. We used an LTQ Orbitrap Velos, a very advanced mass spectrometer (Olsen et al., 2009), to acquire accurate mass spectrometric data at very high resolution. In our initial study we mapped 6,367 N-glycosylation sites in four mouse tissues and in mouse blood plasma. We found that sites almost always have the N-!P-[S|T]-!P motif (where !P is any amino acid except for proline) but sometimes also the N-X-C motif or a non-consensus sequence and that they are always oriented toward extracellular space or towards the lumen of ER, Golgi, lysosome or peroxisome (Zielinska et al., 2010).

Furthermore, we measured the N-glycoproteomes of an additional six organisms: *Schizosaccharomyces pombe, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster*, *Danio rerio* and *Arabidopsis thaliana* that representatively span more than a billion years in evolution. Combined with the mouse N-glycoproteome we identified 15,771 sites and found that all eukaryotic N-glycoproteomes have invariant characteristics including sequence recognition patterns, structural constraints and subcellular localization. We showed that a large percentage of the N-glycoproteome coevolved with the rise of extracellular processes that are specific within corresponding phylogenetic groups and that are essential for organismal development and body growth.

Finally, we applied the method to determine the differences in protein N-glycosylation between colorectal cancer and normal colonic mucosa using formalin-fixed paraffin-embedded human material. We identified 1,885 glycosylation sites, many on proteins of very low abundance. The majority of detected sites were on extracellular proteins with a predominant role in transport, response to stimulus and cell-cell communication. In addition to known colorectal cancer biomarkers including CEA and CD166, we identified novel candidate marker glycoproteins for colorectal cancer such as LRP-4 and CAT-1 in the set of regulated glycosylated proteins. We normalized the calculated ratios with the separately measured colon cancer proteome and we showed that as a general rule the glycoprotein expression, and not the glycosylation site occupancy, is altered in cancerous tissues. However, we also found evidence for differential regulation of N-glycosylation.

During my PhD study I additionally performed proteomic and phosphoproteomic analyses. The results of these unrelated to N-glycosylation side-projects are presented in appendix I-II.

I received the Young Scientist and Junior Scientist awards in recognition of my work. I believe these awards help to underline the success of my research and reflect the interest of the community in the field of N-glycosylation.

INTRODUCTION

N-Glycosylation versus Other Post-translational Protein Modifications

Site-specific covalent protein modifications play key roles in modulating protein activity and function in biological systems (Hann, 2006; Krishna and Wold, 1993). Over 200 kinds of protein post-translational modifications are known (Krishna and Wold, 1993). Phosphorylation and glycosylation are among the biologically most important types of protein modifications. Overall 30 % of proteins are phosphorylated and over 50% are estimated to be modified by covalent attachment of sugar molecules at some point in their life cycle (Van den Steen et al., 1998). Compared to the well-studied post-translational phosphorylation of proteins, little is known about glycosylation. While it was possible to map nearly 7,000 phoshorylation sites in one experiment in 2006 (Olsen et al., 2006), the largest reported N-glycoproteomic study had identified only 1,495 N-glycosylation sites – in the model organism *C. elegans* (Kaji et al., 2007). Other studies measured up to a few hundred N-glycosylation sites on cell surface proteins of the immune system (Wollscheid et al., 2009), of mouse C2C12 myoblasts (Gundry et al., 2009), in human blood plasma (Liu et al., 2005), in human serum (Bunkenborg et al., 2004), in human saliva (Ramachandran et al., 2006) and in rat liver (Lee et al., 2009). According to the most-comprehensive post-translational modification database PHOSIDA (Gnad et al., 2007) and the largest protein database Swiss-Prot (Wu et al., 2006) over 100,000 phosphorylation sites but less than 4,000 glycosylation sites have been experimentally observed in eukaryotes. This is likely to be a dramatic underestimate of the true extent of the N-glycoproteome. Indeed, in our study, by measuring the N-glycoproteomes of seven major model eukaryotic organisms and human colon tissue, we increase the number of known glycosylation sites by a factor of at least five compared to that known up to the present date.

Biology of N-Glycosylation

N-glycosylation plays a major role in cell-cell and receptor-ligand interactions, immune response, apoptosis, angiogenesis and pathogenesis (Varki, 1993; Woods et al., 1994). It is known to be involved in the proper folding of newly synthesised polypeptides, the protection of proteins from recognition by proteases and the modulation of interactions with other molecules. Aberrant glycosylation has been implicated in a number of diseases including neurodegenerative disorders and cancer (Liu et al., 2002).

Glycan molecules can be attached to proteins either N-linked or O-linked. The N-linked glycosylation is more common (Nalivaeva and Turner, 2001). N-linked saccharides are attached at the asparagine residue in the consensus sequence N-!P-[S|T] (where X represents any amino acid except for proline) via N-acetylglucosamine (GlcNAc). Additional sugar residues depend on the glycan type (Suzuki et al., 1995). The O-linked glycosylation occurs on serines and threonines. O-linked carbohydrates are typically attached to proteins via N- acetylgalactosamine (GalNAc). A variation of O-glycosylation is the O-GlcNAc modification that occurs on nucleocytoplasmatic proteins, and often maps to the same or adjacent sites as phosphorylation (Comer and Hart, 2000). My PhD study dealt exclusively with N-linked glycosylation.

The process of N-glycosylation starts in the endoplasmatic reticulum (ER) with the transfer of oligosaccharide precursor to the nascent polypeptide. This reaction is catalyzed by oligosacharyl transferase. After the protein is correctly folded, three glucose residues are removed and the glycoprotein is exported to the Golgi apparatus for further glycan processing. Eukaryotic N-linked glycans share a common core sugar sequence and are classified in three groups: high-mannose, complex and hybrid. The pentasaccharide core-structure of N-linked glycans is composed of two N-acetylglucosamines and three mannose residues. The core-structure can be modified - in vertebrates and invertebrates by adding α1,6 fucose to the asparagine-linked N-acetylglucosamine or in invertebrates and plants α1,3 fucose. Additionally, xylose may be transferred to the ß-linked mannose residue of the plant N-glycoproteins (Varki et al., 2008).

Up to now, a bacterial N-glycosylation system that is similar to the eukaryotic system was exclusively found in *C. jejuni* (Wacker et al., 2002). It is the only bacteria whose genome encodes a homologous protein to the eukaryotic oligosaccharyltransferase STT3. However, the core structure of *C. jejuni* N-glycans is known to be different from eukaryotes (Liu et al., 2006; Scott et al., 2009; Young et al., 2002).

Shotgun Proteomics for Mapping of N-Glycosylation

Mass spectrometers are instruments used to determine the masses of charged particles derived from molecules. The key steps in mass spectrometry (MS) are production of gas-phase ions in the ion-source, ion separation according to mass-to-charge ratio (m/z) in mass analysers, measurement of ion quantity in detectors and finally processing of the ion signals into mass spectra.

Since the introduction of the first mass spectrometer in the beginning of the twentieth century, instrumentation has rapidly developed and mass spectrometry has become the most popular platform in proteomics. It has become an ideal tool not only for studying protein expression and interactions but also for identification of modification sites. Particularly, liquid chromatography coupled to high-resolution mass spectrometry (LC-MS) has emerged as the key technology for the in-depth large-scale analysis of post-translational modifications in general and N-glycosylated proteins in particular (Aebersold and Mann, 2003; Jensen, 2006; Medzihradszky, 2005; Witze et al., 2007). For mapping of modifications the initial precursor mass is determined and the most abundant peptides are selected for fragmentation. In such tandem mass spectrometry experiments (MS/MS) the structural features are obtained from the masses of the individual fragments. In principle, any PTM can be detected by the corresponding increases or decreases of peptide masses.

The typical workflow for shotgun proteomics applied to PTM analysis is shown in Figure 1.The key steps in a mass spectrometry based PTM study are sample preparation, specific enrichment, fractionation, mass spectrometric measurement, MS data processing and bioinformatic analysis. These steps are described in detail below.

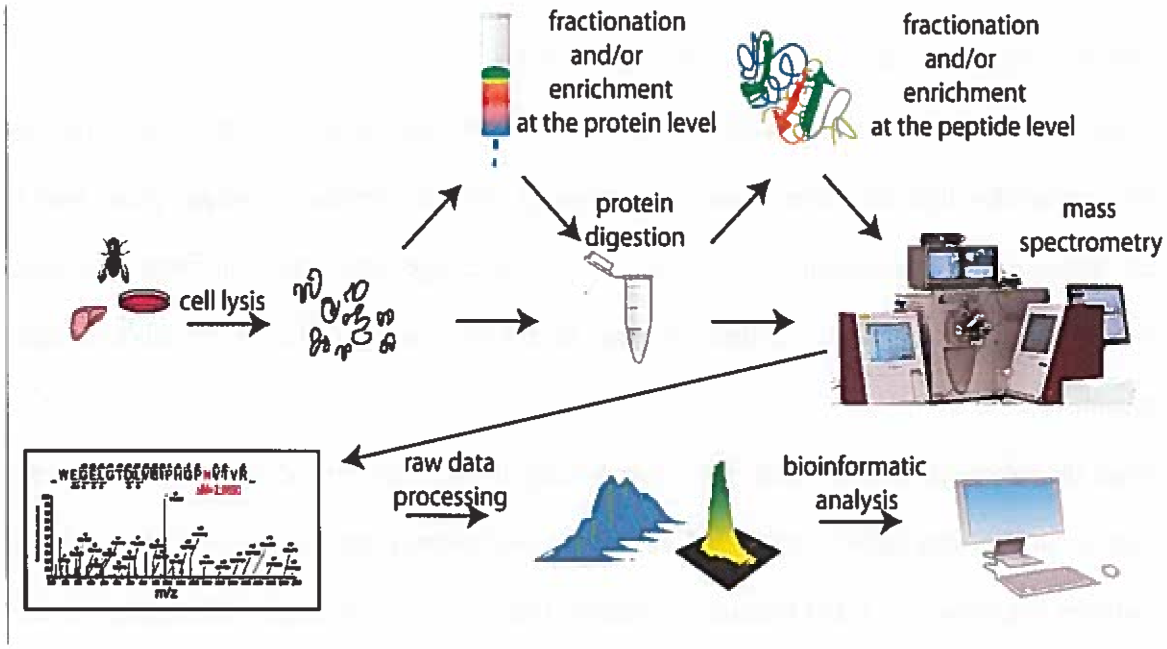

fractionation and/or enrichment at the protein level
fractionation and/or enrichment at the peptide level
protein digestion
mass spectrometry
cell lysis
raw data processing
bioinformatic analysis
m/z

Figure 1. General workflow of MS-based analysis of post-translational protein modifications
Cells, tissues or whole organisms are lysed. Extracted proteins can be supplied to proteolytic digestion either directly or after separation / enrichment at the protein level. Peptides can be further fractionated / enriched or directly subjected to mass spectrometry. The MS output is searched against protein sequence databases and acquired data is analyzed using bioinformatic tools.

Sample Preparation

Protein identification via MS can be carried out in the form of intact-protein analysis (top-down approach) or peptide analysis (bottom-up method) (Han et al., 2008). The top-down approach of analysing intact proteins is a challenging strategy when used for high-throughput studies because of the analyte complexity and limitation in high quality proteome fractionation (Siuti and Kelleher, 2007). Therefore, the bottom-up strategy is usually applied in modification-specific proteomic experiments, where proteins are digested into small peptides prior to MS analysis. Until recently two major strategies have been applied to generate peptides from proteins extracted from biological samples. The first strategy is based on protein solubilization with detergents, separation in the gel and 'in-gel' digestion (Shevchenko et al., 1996). This method is time-consuming and the peptides may have low recovery from the gel (Wisniewski et al., 2011). The advantages of this method are a low degree of contamination and an extreme robustness. The second strategy involves treatment of proteins with strong chaotropic reagents such as urea or thiourea, protein precipitation and 'in-solution' digestion (Washburn et al., 2001). This method is easier to carry out due to a higher degree of automation. However, not all proteins are solubilized and the solution may contain impurities, hindering digestion. The recently developed filter-aided sample preparation method (FASP) (Wisniewski et al., 2009b) combines the advantages of the above-mentioned techniques. It allows the extraction of even very poorly soluble proteins such as membrane proteins by using a very strong detergent sodium dodecyl sulphate, which is then exchanged by urea on a standard filtration unit. The approach eliminates impurities that can interfere with digestion and is very easy to handle, resulting in a pure peptide mixture in a very short period of time. This method is of particular interest in the study of N-glycosylated proteins that are in general localized on the hydrophobic plasma membranes.

Quantification

Mass spectrometry can be also applied to detect differences in PTM abundances. In general, most of the mass spectrometry based quantification methods employ stable isotope labeling that introduce an easily detectable mass tag (Pan and Aebersold, 2007). Stable isotopes can be incorporated by chemical reaction of isotope-coded affinity tags (ICAT) (Gygi et al., 1999), isobaric tags for relative and absolute quantification (iTRAQ) (Ross et al., 2004), enzymatic incorporation of ^{18}O into cleaved peptides (Reynolds et al., 2002) or metabolic labeling of whole cells during protein synthesis (Oda et al., 1999). Quantification using isotope labeling is based on the incorporation of a stable isotope signature in all proteins of one sample and using a different stable isotope signature in all proteins of another sample. The samples are then combined to serve as mutual references (Domon and Aebersold, 2006). The tandem mass tag method and the iTRAQ method are based on the incorporation of two labeled elements that have an overall constant mass. The relative intensity of the reporter group is measured (Thompson et al., 2003). For absolute quantification using internal standards a defined amount of labeled standard sample is added to the unlabeled sample allowing the absolute measurement of protein abundance by determining peak abundance ratios.

Stable isotope labeling with amino acids in cell culture (SILAC) (Ong et al., 2002) has become the method of choice for many quantification applications in proteomics. The method was initially developed for cell culture, and later extended to model organisms (Kruger et al., 2008). The need for complete labeling and the impossibility of labeling any organism with a diet containing isotopes prevented it from some applications – like studying diseases using human material, for example. These disadvantages can now be partially overcome by spike-in SILAC, where only one cell state (or experimental state of an animal) is subjected to SILAC and added to other states as reference (Geiger et al., 2010).

Whereas isotope-based methods are presumably the most robust and accurate, label-free quantification techniques are by far the simplest and most economical (Domon and Aebersold, 2006; Mueller et al., 2008). Recently emerging high-resolution mass spectrometers and sophisticated computer tools have enabled label-free quantification to be applied successfully to a number of proteomic studies including profiling of N-glycosite changes (Schiess et al., 2009).

Enrichment

Glycosylated proteins, like other post-translationally modified proteins, can often be present in very low levels of concentration. To detect such low abundance glycosylated proteins among the large excess of relatively high abundance proteins in complex mixtures, specific enrichment methods have to be applied (Corthals et al., 2000). Such enrichment can be achieved either at the intact protein level or at the peptide level (Jensen, 2004), most commonly based on lectin affinity (Bunkenborg et al., 2004) or other general physical and chemical properties of attached carbohydrates. Glycopeptides have higher masses and therefore can be easily separated by size exclusion chromatography (Alvarez-Manilla et al., 2006). Other approaches use the hydrophilic character of the glycans to interact with carbohydrate gel matrices such as cellulose or sepharose (Wada et al., 2004). Alternatively, hydrophilic interaction-liquid chromatography can be performed (HILIC) (Wuhrer et al., 2007), which may also be directly coupled to nano-electrospray mass spectrometry (Wuhrer et al., 2005). Other approaches are based on specific binding of sugar molecules to mannose-6-phosphate receptor, galectins or anti-carbohydrate monoclonal antibodies (Wuhrer et al., 2007). Recently many novel enrichment methods based on specific glycan interactions with different kinds of beads have been developed. Such methods include beads functionalized with di-boronic acid (Sparbier et al., 2005), hydrazide functionalized beads (Zhang et al., 2003) and graphitized carbon (Larsen et al., 2005). Despite the large number of available methods to enrich glycoproteins and glycopeptides, lectin affinity chromatography is most widely used. Using a combination of lectins that specifically recognize different classes of N-glycans, very high enrichment can be obtained (Kim et al., 2006). In this study we have developed a multilectin affinity chromatography based method, 'N-glyco-FASP' (Zielinska et al., 2010), that efficiently captures N-glycosylated peptides by simply adding a lectin mixture to the top of filtration units (Figure 2). This method is very easy to handle and does not require any affinity columns. It allows the in-depth detection of very low abundance glycoproteins and can be applied to glycoproteomic studies of very complex protein mixtures extracted from whole tissues or even organisms. It allows the identification of over 2,000 N-glycosylation sites in one LC-MS/MS run from a minute amount of protein starting material.

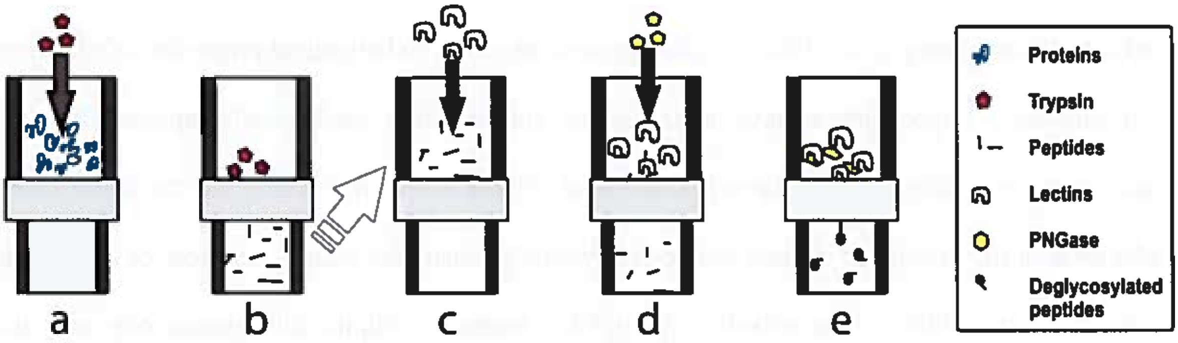


Figure 2. FASP-based N-linked glycopeptide capture method
A standard filtration device with 30k molecular cut-off is used as a proteomic reactor. Whole SDS-lysates of mouse tissues are processed according to the FASP protocol (a and b). To enrich for N-glycosylated peptides the digests are incubated with free lectins on the filter unit (c). Unbound peptides are removed by centrifugation. The bound glycopeptides are deglycosylated with PNGase (d), eluted from the filter (e), and identified using high accuracy LC-MS/MS.

Fractionation

The complexity of the protein mixture can be efficiently reduced by on-line or off-line approaches to fractionate the analyte. In the on-line approach, peptides are separated in an in-line set up within the instrument directly before injection into a mass spectrometer. In proteomics, reversed phase liquid chromatography is most commonly used, where peptides are separated on columns packed with C18 material. The off-line fractionation approach is uncoupled from mass spectrometry and can be performed at the protein and/or peptide level. Here, commonly used techniques include 1D- and 2D-gel electrophoresis. In particular two-dimensional electrophoresis, which separates proteins by isoelectric point (pI) and molecular weight, has been used in proteomics for many years (O'Farrell, 1975). This method provides good separation, but is not applicable to large-scale proteomic analysis. Peptide separation with immobilized pI strips is an alternative method. This approach is more reliable and results in an increased number of identifications (Horth et al., 2006; Hubner et al., 2008). Other approaches involve fractionating proteins by chromatofocusing (Chong et al., 2001), capillary isoelectric focusing (Zhou and Johnston, 2005), size exclusion chromatography (SEC) (Lu et al., 2009) or ion exchange (Sharma et al., 2007; Washburn et al., 2001; Wisniewski et al., 2009a).

Mass Spectrometry

All mass spectrometers consist of an ion source, at least one mass analyzer and a detector. Since only ions in the gas phase can be analyzed, proteins or peptides have to be evaporated and ionized prior to the mass spectrometric analysis. Ions are usually generated by electrospray ionization (ESI) (Fenn et al., 1989) or matrix-assisted laser desorption/ionization (MALDI) sources. The development of these soft ionization techniques allowing analysis of intact biomolecules, almost a century after the introduction of the first mass spectrometer, revolutionized the field of mass spectrometry and was honoured with the Nobel Prize in Chemistry in 2002. In this study we exclusively used electrospray ionization.

In LC-ESI-MS/MS, peptide mixture after chromatographic separation is directly sprayed into the mass spectrometer. Charged droplets are generated by an electric field and converted into intact ions in the gas phase. Their analysis in the mass spectrometer is divided into two parts, the survey scan (MS scan) allowing acquisition of peptide intensities, and MS/MS scans of fragment ions, performed after fragmentation of the highest abundance peaks from the survey scan. Information derived from both scans is required for the identification of peptides.

Many different types of mass spectrometers are commercially available. In shotgun proteomics the most popular is the combination of linear ion trap with high resolution mass analyzer (Domon and Aebersold, 2006). The ion cyclotron resonance instrument with a 7T magnet device was the first instrument that implemented this principle (Syka et al., 2004). Later, in 2005, the linear ion trap-orbitrap (LTQ Orbitrap) mass spectrometer with a powerful orbitrap analyzer that uses electrostatic field for trapping ions was commercially introduced (Hardman and Makarov, 2003; Hu et al., 2005; Makarov et al., 2006). Signal acquisition in the orbitrap and simultaneous isolation of the most intense ions, their fragmentation and measurement in the ion trap allow accurate mass measurements of proteins and high-confidence identification of post-translational modifications. The next generation of the linear ion trap-orbitrap instrument, LTQ Orbitrap Velos, with even greater sensitivity, dynamic range, mass accuracy, and sequencing speed when compared to the traditional orbitrap instrument, is particularly suitable for PTM analysis (Olsen et al., 2009). The LTQ Orbitrap Velos, unlike other mass spectrometers, is capable of measuring not only peptide masses but also fragment masses with mass accuracies in the part per million range while still retaining excellent sensitivity (Olsen et al., 2009).

In our work we performed all MS analyses using LTQ Orbitrap Velos mass spectrometers in high resolution precursor and high accuracy fragment mode ('high-high' strategy).

Peptide Detection

Post-translational modifications are very different regarding size and physiochemical properties, not only between the various types but also within the same group of modifications. For example, glycans are composed of different combinations of sugar molecules which form unique, branched structures. Furthermore, many glycosylation sites usually occur on a single protein and function together to regulate the structure and functions of proteins (Yang, 2005). The microheterogeneity and complex multi-site modification patterns make the identification of glycosylation sites and characterization of their structures very challenging. Therefore, for the determination of glycosylation sites – as opposed to the structure of the sugar – a universal deglycosylating enzyme is usually employed. Endoglycosidase (PNGase) that cleaves between innermost GlcNAc and asparagines is most common. It leads to the deamidation of the asparagine residue to aspartic acid and a mass increase of 0.9848 Da that can be detected by tandem mass spectrometry (MS/MS) as a mass shift of the precursor peptide and of its fragments. If deglycosylation is performed in ^{18}O-water, the resulting mass shift of 2.9890 Da can easily be distinguished from spontaneous deamidation, adding confidence to the site assignment (Kuster and Mann, 1999). Alternatively, deglycosylation can be performed by endo-N-acetylglucosamidase leaving a GlcNAc tag (+ 185 Da) attached to the peptide. Released glycans are usually characterized separately. Glycopeptides without prior deglycosylation can be also analyzed via MS/MS or MS^3 experiments; these approaches, however, still require further development in order to be applicable to complex protein mixtures.

MS Data Processing

A key step in mass spectrometry-based proteomics is the identification of peptides in sequence databases by their fragmentation spectra. Many approaches and algorithms for peptide and protein identification by searching a sequence database using MS data have been described in the literature (Sadygov et al., 2004). Although reported methods differ in their implementation, the general concept is similar. The experimental data are compared with peptide and peptide fragment mass values calculated on the basis of cleavage rules applied to the protein sequences in the specified database. To assign measured spectra to a peptide sequence, the probability-based search engine Mascot (Perkins et al., 1999) is a well established tool. It is primarily optimized for the identification of sequence peptides based on the presence of calculated fragment ions in the tandem spectra. Identified peptide sequences are assigned to protein entries afterwards. Modification sites are localized using an algorithm that calculates the PTM scores for all possible PTM combinations within a given modified peptide sequence by successively placing the modification on each candidate residue (Olsen et al., 2006). Recently, a novel peptide identification algorithm Andromeda was introduced (Cox et al., 2011). It accurately handles many modifications of the same peptide and also features a second peptide identification algorithm for co-fragmented peptide precursors.

To deduce the exact localization of PTM events within a given peptide along with the corresponding probabilities from the given spectrum, the algorithm is usually embedded into special software developed for MS data analysis (Beausoleil et al., 2006). In our laboratory, we use the MaxQuant software (Cox and Mann, 2008) publicly available at http://maxquant.org/.

Bioinformatic Analysis

To understand biological systems that build up complex networks, experimental approaches have to be complemented with bioinformatic analysis. Many bioinformatics tools employed for proteomic analysis have already been developed in the genomic or transcriptomic era. Most common among them are tools for data visualization and clustering. The hierarchical clustering and k-means algorithm, in particular, have found numerous applications in modern proteomics due to their clear and intuitive algorithmic assumptions (Kumar and Mann, 2009; Meunier et al., 2007). However, the unique nature of post-translationally modified proteins stimulated development of more sophisticated methods. The standard bioinformatic analyses used in proteomics were comprehensively reviewed by Kumar and Mann (Kumar and Mann, 2009). The high complexity of protein data with increased dimensionality caused by numerous modifications requires application of statistically robust machine learning approaches (Hastie, 2001). For data mining in large scale modification sets the most common techniques are gene ontology analysis (Barrell et al., 2009), KEGG pathway analysis (Kanehisa and Goto, 2000), motif extraction (Schwartz and Gygi, 2005) and secondary structure analysis (Wagner et al., 2005). The purpose of gene ontology and KEGG analysis is to find molecular functions, biological processes, cellular components and pathways that are over- or under-represented in the set of modified proteins. Motif extraction tools are widely used in modification-specific proteomics to identify post-translational modification sequence motifs from the vastly increasing volume of mass spectrometric input data. To derive structural constraints of modified sites from primary amino acid sequence, linear regression models and neural networks have proven to be very accurate. These types of analysis are easily accomplished using standard software tools, such as Cytoscape (Shannon et al., 2003) and its plugin BINGO (Maere et al., 2005), DAVID (Dennis et al., 2003), Motif-X (Schwartz and Gygi, 2005) or SABLE (Wagner et al., 2005). More sophisticated analysis can be performed using open source analysis programs. In life sciences the statistical environment R (http://cran.at.r-project.org/) together with associated BioConductor (Gentleman et al., 2004) are most commonly used. The R framework, traditionally used for micro-array data analysis is becoming more and more popular in proteomics, mainly due to its intuitive environment, statistical capability and excellent graphics (Kumar and Mann, 2009).

PUBLICATIONS

Publication I

Zielinska DF, Gnad F, Wiśniewski JR, Mann M (2010), *Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints*, Cell , May 28; 141(5) 897-907.

Publication II

Zielinska DF, Gnad F, Schropp K, Wiśniewski JR, Mann M, *Evolution of N-glycosylation* (submitted)

Publication III

Ostasiewicz P, **Zielinska DF**, Mann M, Wiśniewski JR (2010), *Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry*, J Proteome Res, Jul 2; 9(7) 3688-700.

Publication IV

Zielinska DF, Ostasiewicz P, Gnad F, Duś K, Mann M, Wiśniewski JR, *Proteomic and N-Glycoproteomic Profiling of Colorectal Cancer* (in preparation)

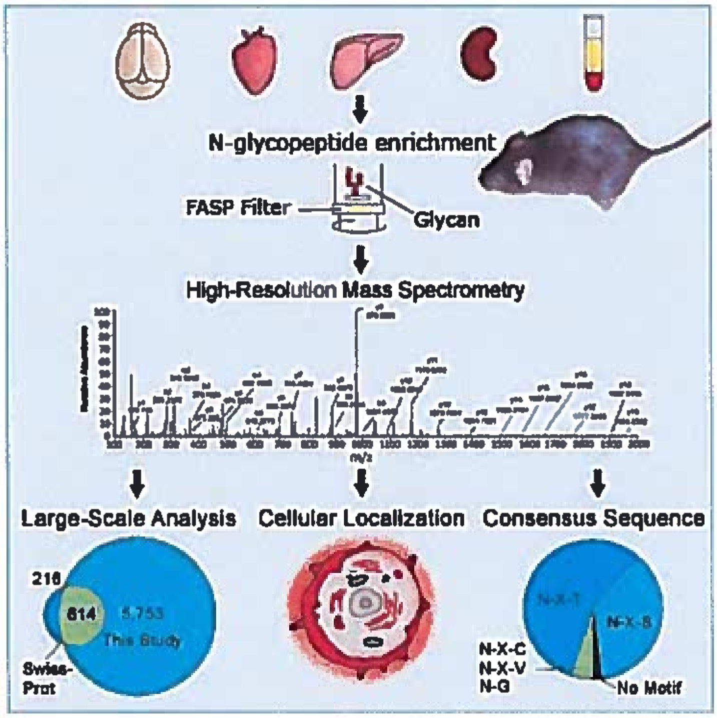
Publication V

Wiśniewski JR, **Zielinska DF**, Mann M, *Comparison of Ultrafiltration Units for Proteomic and N-Glycoproteomic Analysis by the FASP Method*, Anal Biochem, Mar 15; 410(2) 307-9

Publication I

Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints

Zielinska DF, Gnad F, Wiśniewski JR, Mann M



Cell , 2010

In this study we present the newly developed 'N-Glyco-FASP' method to precisely map N-glycosylation sites from complex protein mixtures. We applied the method to study N-glycosylation in five mouse tissues and mouse blood plasma. Our large-scale analysis resulted in the high confidence identification of 6,367 N-glycosylated sites. Bioinformatics revealed many interesting characteristics including constraints on sequence, structure and cellular localizations.

Cell

Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints

Dorota F. Zielinska,[1,3] Florian Gnad,[1,2,3] Jacek R. Wiśniewski,[1,*] and Matthias Mann[1,*]
[1]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, Martinsried D-82152, Germany
[2]Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA
[3]These authors contributed equally to this work
*Correspondence: jwisniew@biochem.mpg.de (J.R.W.), mmann@biochem.mpg.de (M.M.)

DOI 10.1016/j.cell.2010.04.012

SUMMARY

N-linked glycosylation is a biologically important protein modification, but only a small fraction of modification sites have been mapped. We developed a "filter aided sample preparation" (FASP)-based method in which glycopeptides are enriched by binding to lectins on the top of a filter and mapped 6367 N-glycosylation sites on 2352 proteins in four mouse tissues and blood plasma using high-accuracy mass spectrometry. We found 74% of known mouse N-glycosites and discovered an additional 5753 sites on a diverse range of proteins. Sites almost always have the N-!P-[S|T]-!P (where !P is not proline) and rarely the N-X-C motif or nonconsensus sequences. Combining the FASP approach with analysis of subcellular glycosite localization reveals that the sites always orient toward the extracellular space or toward the lumen of ER, Golgi, lysosome, or peroxisome. The N-glycoproteome contains a plethora of modification sites on factors important in development, organ-specific functions, and disease.

INTRODUCTION

N-glycosylation is one of the most prominent posttranslational protein modifications and plays a major role in the assembly of complex multicellular organs and organisms (Varki et al., 2009). This modification is involved in many cellular functions including cell-cell and receptor-ligand interactions, immune response, apoptosis, and pathogenesis of many diseases (Varki et al., 2009; Woods et al., 1994). N-glycosyltransferases are predominantly located in the lumen of the ER and Golgi apparatus and attach this modification cotranslationally in a complex series of processing steps to a subset of the sites with the consensus sequence N-!P-[S|T] (where !P signifies any amino acid except proline). This motif has been extended to N-!P-[S|T]-!P in *C. elegans* (Kaji et al., 2007). It is also possible that there are consensus motifs different from the canonical one.

Because of the topological location of the transferases, the modification is thought to be localized on secreted molecules, the extracellular part of plasma membrane proteins, and the lumenal part of proteins in compartments of subcellular organelles such as the endoplasmatic reticulum and the Golgi apparatus, endosomes, and lysosomes. A number of authors have raised the possibility that N-linked glycosylation may also be present in mitochondria (Chandra et al., 1998; Kung et al., 2009), in the nucleus (Reeves et al., 1981), and in the cytoplasm (Pedemonte et al., 1990). However, these studies do not map residue-specific N-linked glycosylation sites.

Despite great biological and clinical interest, our knowledge of in vivo N-glycosylation sites—a prerequisite for detailed functional understanding—is still very limited. Liquid chromatography coupled to high-resolution mass spectrometry (LC-MS) has emerged as the key technology for large-scale analysis of posttranslational modifications in general and N-glycosylated proteins in particular (Aebersold and Mann, 2003; Jensen, 2006; Medzihradszky, 2005; Witze et al., 2007). The large complexity of attached sugar molecules (North et al., 2010) and the low expression levels of many N-glycoproteins make the characterization of complete N-glycosylation structures very challenging. To detect low abundant N-glycosylated proteins or peptides in complex mixtures among the large excess of their nonglycosylated counterparts, specific enrichment methods have to be applied, most commonly based on lectin affinity (Bunkenborg et al., 2004) or chemical linkage of the sugar moiety to surfaces (Zhang et al., 2003). For determination of glycosylation sites—as opposed to the structure of the sugar—a universal deglycosylating enzyme (i.e., PNGase F) is used. This leads to deamidation of the asparagine residue to aspartic acid and a mass increase of 0.9848 Da of the modification site, which can be detected by tandem mass spectrometry (MS/MS) as a mass shift of the precursor peptide and of its fragments. If deglycosylation is performed in ^{18}O-water, the mass shift is 2.9890 Da, adding confidence to the site assignment (Kuster and Mann, 1999).

Among large-scale N-glycoproteomic studies, the largest reported 1495 N-glycosylation sites from *C. elegans* (Kaji et al., 2007). Others measured up to a few hundred N-glycosylation sites on cell surface proteins of the immune system (Wollscheid et al., 2009), of mouse C2C12 myoblasts (Gundry et al., 2009), in human blood plasma (Liu et al., 2005), in human serum (Bunkenborg et al., 2004), in human saliva (Ramachandran et al., 2006), and in rat liver (Lee et al., 2009).

The data of large-scale proteomics studies and some directed studies are combined in the Swiss-Prot database (Wu et al., 2006), which results in 830 mouse and 1998 human N-glycosylation sites. This is likely a drastic underestimate of the true extent of the mammalian N-glycoproteome. Notably, even though the Swiss-Prot database presents the most comprehensive resource of annotated N-glycosylation sites, it is not complete because of the difficulty in retrieving single sites from numerous literature studies.

Given its biomedical importance, we set out to map this modification in-depth and at very high accuracy using recent advances in proteomics technology. We have developed an N-glycopeptide enrichment method based on "filter aided sample preparation" (FASP) (Wisniewski et al., 2009b), which allows highly efficient capture of glycopeptides even from membrane proteins. We employ the ability of the recently introduced LTQ-Orbitrap Velos instrument to measure peptide fragments, and not only peptide precursor masses, with low ppm mass accuracy and at high sensitivity (Olsen et al., 2009). Our analysis of four different mouse tissues and blood plasma achieves very high confidence and covers a substantial part of the mouse N-glycoproteome—allowing in-depth characterization of this protein modification.

RESULTS

Development of a FASP-Based N-linked Glycopeptide Capture Method (N-Glyco-FASP)

Most N-linked glycosylations occur on membrane proteins, which have traditionally been difficult to analyze by proteomic methods. We have recently shown that the FASP method is especially well suited to analyze this class of proteins because it achieves complete protein solubilization in SDS while still allowing gel-free analysis (Wisniewski et al., 2009b). We reasoned that FASP could be combined with peptide affinity capture simply by adding the affinity reagent—in this case lectin—to the top of the filter after on-filter protein digestion. Glycosylated peptides are bound by lectin and thereby retained whereas nonglycosylated peptides can be washed through the filter. Next, glycopeptides are efficiently deglycosylated by PNGase F and released peptides are eluted, resulting in a peptide population of high purity (Figure 1A). We used two different endoproteinases, trypsin and Glu-C, to increase the number and localization confidence of glycosylation sites. In our experiments with the "N-glyco-FASP" method, sample amounts were typically 200 μg of total protein in 40 μl, but this can be scaled up or down as desired.

To capture all three classes of N-glycosylated peptides, multi-lectin enrichment can be employed (Yang and Hancock, 2004). In N-glyco-FASP, lectins do not need to be coupled to a solid support because they are retained by the filter, and therefore any lectin or mixture of lectins can be employed. We selected concanavalin A (ConA), which binds to mannose, wheat germ agglutinin (WGA), which binds to sialic acid, as well as N-acetylglucosamine and agglutinin RCA_{120}, which captures galactose modified at the 3-O position (e.g., with sialic acid or another galactose) as well as terminal galactose. Enrichment with this mixture of lectins was as efficient as separate experiments based on enrichment with all single lectins (Figure 1B). Overall, 63% of all N-glycosylation sites identified in a given tissue could be detected in a single LC-MS/MS experiment by multi-lectin enrichment. In comparison, 69% of a given tissue N-glycoproteome was covered when combining three LC-MS/MS experiments based on single lectin enrichment. WGA proved to have the highest binding efficiency among the applied lectins. The proportion of glycosylated peptides—measured as deamidated peptides after PNGase F digestion—to all identified peptides in single run analysis was 46%. In our experiments, this is comparable to or higher than the enrichment of phosphorylated peptides (Macek et al., 2009) and substantially higher than the enrichment of lysine acetylated peptides (Choudhary et al., 2009). Without lectin enrichment, glycopeptides were 0.5% of total peptides, indicating an enrichment factor of about 100-fold (Figure 1C). We also interrogated our datasets for other modifications but did not find large numbers of such peptides.

Precision Mapping of N-Glycosylation Sites

To identify deglycosylated peptides we used on-line liquid chromatography electrospray mass spectrometry (LC-MS/MS) on the recently introduced linear ion trap orbitrap instrument (LTQ-Orbitrap Velos). The LTQ-Orbitrap is capable of fragmenting peptides by "higher-energy dissociation" (HCD), in which the fragment mass spectrum is analyzed in the high-resolution part of the instrument without loss of low-mass ions (Olsen et al., 2007). The superior sequencing capabilities of HCD compared to ion trap fragmentation (CID) normally come at the cost of reduced sensitivity. However, the Velos instrument features 20-fold improved HCD performance (Olsen et al., 2009). We therefore tested if we could measure the N-glycosylation sites with HCD without loss of sensitivity. Comparison of orbitrap HCD and ion trap CID showed that HCD identified approximately the same number of glycosylated peptides and that it did not discriminate against low-abundance peptides (Figure S1 available online). We therefore performed all subsequent MS analyses in high-resolution precursor and high-accuracy fragment mode ("high-high" strategy).

We applied N-glyco-FASP combined with high-high MS measurement to four mouse organs (brain, liver, kidney, and heart) and blood plasma, which we group together with the other tissues for simplicity. Each tissue was independently prepared in triplicates and measured 11 times by single LC-MS/MS runs with 4 hr gradients after deglycosylation in ^{18}O-water by both single and multi-lectin enrichment. Furthermore we measured N-glycosylation sites in four subcellular fractions of liver cells. Together, 59 LC-MS/MS runs were acquired (Table S1A). Additionally, we performed 64 experiments without ^{18}O-water (Table S1B). Analysis of the data was performed with the MaxQuant software (Cox and Mann, 2008) specifying a false discovery rate of 1% at the peptide and site level. Average absolute mass deviation was 0.57 ppm for identified peptides and 3 ppm for all fragment ions contributing to peptide identification (Figure 2A). The median Mascot identification score for

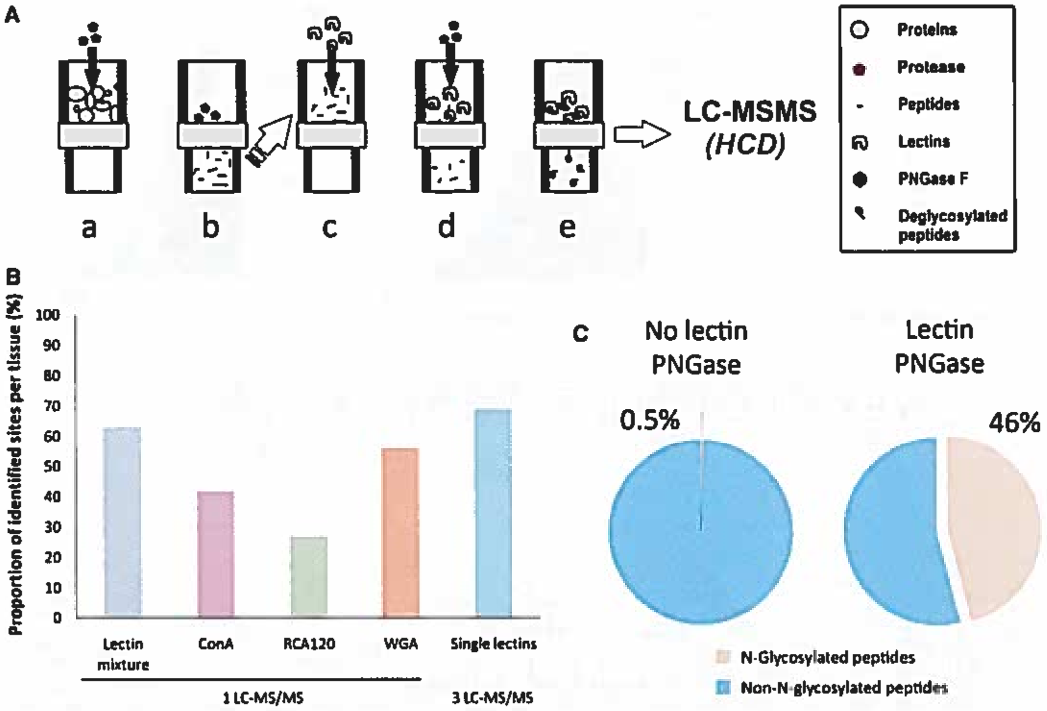


Figure 1. Sample Preparation and Enrichment of N-Glycosylated Peptides: N-Glyco-FASP

(A) A standard filtration device with 30k molecular cut-off is used as a proteomic reactor. Whole SDS-lysates of mouse tissues are processed according to the FASP protocol (a and b). To enrich for N-glycosylated peptides the digests are incubated with free lectins on the filter unit (c). Unbound peptides are removed by centrifugation. The bound glycopeptides are deglycosylated with PNGase F (d), eluted from the filter (e), and identified using high-accuracy LC-MS/MS. Peptide ions are fragmented via higher-energy dissociation (HCD).

(B) The proportion of N-glycosylated sites identified in a single LC-MS/MS experiment was lower in the case of single-lectin enrichment (ConA, RCA120, WGA) compared to multi-lectin enrichment. Using a mixture of lectins in one LC-MS/MS run was nearly as efficient as combining single lectin enrichment-based experiments in three LC-MS/MS runs.

(C) The proportion of detected N-glycosylated to unmodified peptides in an experiment without enrichment was 0.5%. With N-glycosylation enrichment the proportion increased by a factor of about 100 (to a median of 46%).

See also Table S1 and Table S2.

glycopeptides was 65 (Perkins et al., 1999) and the median posttranslational modification (PTM) score was 112 (Olsen et al., 2006) (Figures 2B and 2C). Because neither Mascot nor PTM scores incorporate the high fragment mass accuracy into database identification scores, confidence of glycopeptides identification is much higher than even those scores indicate. The average localization probability of all identified sites was 94.92%. This indicates that both peptide identification and localization of the modification in the peptide sequence with single amino acid resolution were unambiguous. Out of all identified N-glycosylated sites, we derived those that had a minimum residue localization probability of 95% and that were identified in two or more experiments with a minimum of one confirmation in the ^{18}O-water experiment to build a top confidence (class I) set (Table S2A). N-glycosites that do not satisfy these extremely strict criteria but show a residue localization probability higher than 90% are also high-confidence sites (class II, Table S2B), but they were not included in the global analyses described in this paper. Average localization probability of class I and II sites was 99.97%.

We performed PNGase digestion both with and without ^{18}O-water (see Extended Experimental Procedures). As mentioned above, ^{18}O-water deglycosylation events lead to a mass increment of 2.9890 Da that can be readily distinguished from spontaneous deamidation (mass increment of 0.9858 Da). Therefore the experimental confirmation with ^{18}O-water was considered as criteria for the definition of our top confidence set as described above. An example spectrum for N-glycopeptide identification is shown in Figure 2D.

The In Vivo Mouse N-Glycoproteome

Our large-scale analysis resulted in the high-confidence identification of 6367 N-glycosylated sites with a localization probability higher than 90% on 2352 proteins (class I and II; Table S2; all sites are available in the PHOSIDA database along with their corresponding spectra; Gnad et al., 2007). This dataset covers 74% of the 830 experimentally derived mouse N-glycosites recorded in Swiss-Prot (Version 57.12). Overall, 5753 sites of our set were not previously recorded as experimentally identified. Swiss-Prot also contains a large number of potential N-glycosites

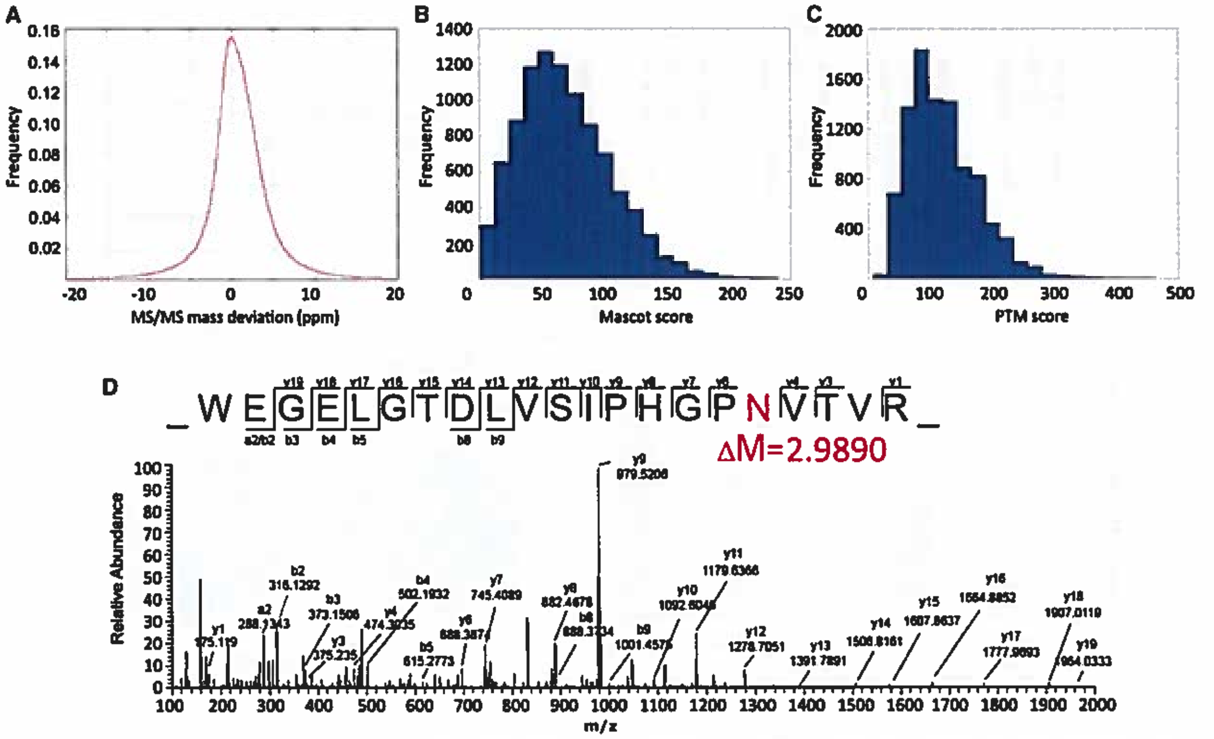


Figure 2. Identification of the Glycosylation Sites by High-Accuracy Tandem Mass Spectrometry using HCD Fragmentation

Distribution of MS/MS mass deviations (A), Mascot scores (B), and PTM scores (C) of the sequenced deglycosylated peptides.

(D) Representative MS/MS spectrum of the peptide WEGELGTDLVSIPHGPNVTVR of β-secretase 1, identified in ^{18}O-water. The mass increment of 2.9890 Da is indicated.

See also Figure S1.

(11,846), partly by similarity to the 1998 known human sites but mainly derived from bioinformatic prediction. Of these sites, we cover 33%, which is excellent agreement given the fact that algorithms are generally adjusted toward overprediction.

Several lines of experimental evidence suggest that our dataset achieves very deep coverage of the mouse N-glycoproteome: Technical and biological repeats of N-glyco-FASP enriched tissue samples resulted in largely overlapping N-glycosites (on average 80% between any two single runs) and added only a small number of additional sites (Figure 3A). Additional fractionation either on the protein level by size-exclusion chromatography (Zielinska et al., 2009) or on the peptide level by anion exchange chromatography (Wisniewski et al., 2009a) resulted in only a few percent of additional glycosites. In contrast, performing these additional fractionation steps for the tissue proteome more than doubled the number of detected proteins compared to a single run (Figure 3B). Both observations are consistent with a glycoproteome that is thoroughly sampled by our analysis.

Sequence Recognition Motifs, Structure Preference, and Occupancy

The canonical N-linked glycosylation motif is N-!P-[S|T]. We reasoned that our high precision and large-scale dataset might provide a good basis to test the generality of this motif and to discover further consensus sequences. We compared the position-specific amino acid frequencies of the surrounding sequences (six amino acids to both termini) of glycosylated asparagines that have serine or threonine on the second position to the C terminus with those of all asparagines that occur in the mouse proteome. We found that proline is drastically underrepresented not only in the first position relative to the modification site (0% compared to 6.16% expected) but also in the third position (0.54% compared to 5.16% expected; p = 0; Table S3A). Notably, cysteine was highly enriched in the surrounding sequences of N-glycosites that match with the canonical N-!P-[S|T] motif. Threonine occurs more frequently (1.4-fold) than serine at the second position—the reverse proportion as in the surrounding sequences of nonglycosylated asparagines that match with the consensus motif.

Next we asked if there were any motifs different from the canonical one. We applied the de novo method Motif-X (Schwartz and Gygi, 2005) to the surrounding sequences of all top confidence N-glycosylated asparagines (Extended Experimental Procedures). This resulted in the identification of three further significantly overrepresented consensus sequences (Figure 4A). Of 5052 sites, 177 did not match the N-!P-[S|T] motif. These sites turned out to be heavily enriched

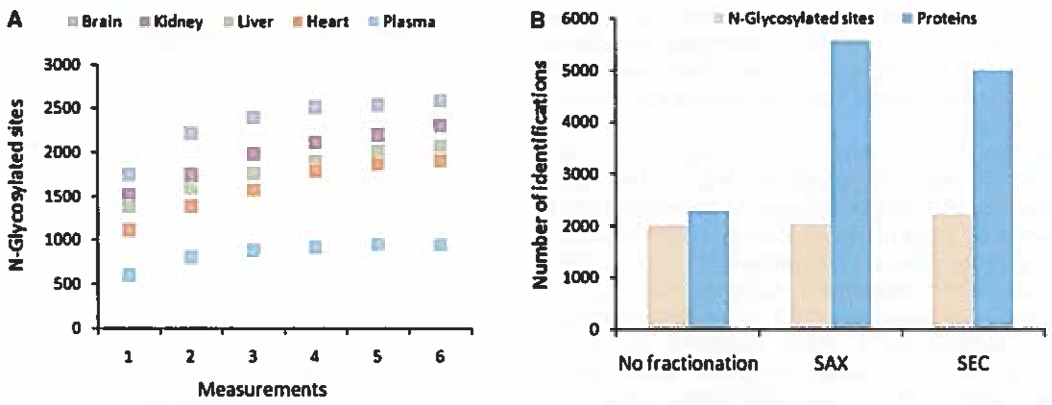


Figure 3. Depth of the Detected N-Glycoproteome

(A) Repeated measurements of each tissue yielded a minimal increase of identified N-glycosylation sites.

(B) Fractionation methods such as size-exclusion chromatography (SEC) and anion-exchange chromatography (SAX) did not result in a greatly increased number of identified N-glycosites, whereas the number of detected proteins in a proteome measurement increased more than 2-fold.

for cysteine or valine in place of S/T. In addition, glycine was enriched on the first position toward the C terminus. This result is interesting, as no other motifs except for N-!P-[S|T] and anecdotal evidence for the N-X-C motif are known (Zajonc et al., 2008). Overall, 112 N-glycosylated sites do not match with any of the consensus sequences. To further verify the existence of N-glycosylation on sites that do not match with the known motifs, we performed western blotting on Apolipoprotein A1 (ApoA1) and Apolipoprotein E (ApoE) (Extended Experimental Procedures). In our large-scale dataset we found both

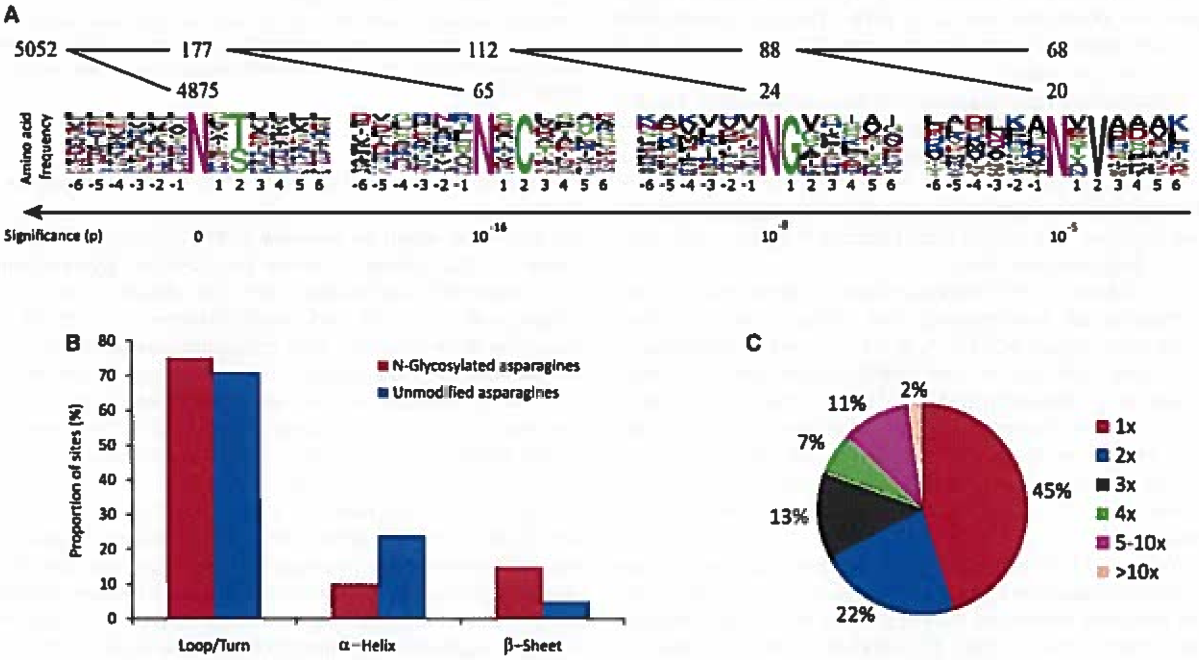


Figure 4. Sequence Recognition Motifs, Structure Preference, and Multiple Glycosylation

(A) N-glycosylation consensus sequence as derived using MotifX. WebLogo (Schneider and Stephens, 1990) was used to create relative frequency plots. The most significant sequence motif is the canonical one, with serine and threonine on position 2. In following iterative steps the consensus sequences N-X-C, N-G, and N-X-V were statistically identified.

(B) Proportion of N-glycosylated and non-N-glycosylated asparagines localized in loops, α helices, and β sheets.

(C) Distribution of singly and multiply glycosylated proteins.

See also Figure S2 and Table S3.

proteins to be N-glycosylated (ApoA1 on position 66, ApoE on position 130), even though they do not contain any asparagines that match with N-!P-[S|T] or N-X-C. In western blots, both proteins shifted their apparent molecular weights after PNGase treatment (Figure S2).

We predicted the secondary structure localizations and solvent accessibilities of N-glycosylated and non-N-glycosylated asparagines of N-linked mouse glycoproteins with SABLE 2.0 (Wagner et al., 2005) (Extended Experimental Procedures). As expected, N-glycosites are enriched on the protein surface. Like their unmodified counterparts N-glycosylated sites are mainly located in loop structures (71% versus 75%). Strikingly, we found evidence for a highly significant enrichment ($p < 10^{-10}$) in β sheets in comparison to non-N-glycosylated asparagines. Overall, 15% of glycosylated asparagines are predicted to be localized in β sheet structures in comparison to 5% of nonglycosylated asparagines (Figure 4B).

The percentage of proteins modified at a particular site ("site occupancy" or "stoichiometry") is often very low for reversible PTMs. N-glycosylation is thought to be stable and frequently serves structural roles, which would favor high site occupancy. Studies in the literature frequently report high N-glycosylation site occupancy. For example, transferrin and α_1-antitrypsin showed 98%–100% occupancy in human serum (Hulsmeier et al., 2007). However, experimental evidence for incomplete occupancies has also been reported. For example, in mouse brain the major prion protein has two different N-glycosylation sites with incomplete occupancy (66% of proteins glycosylated on both sites, 14% on one site, and 20% nonglycosylated) (Bradford et al., 2009).

If glycosylation mainly occurred in high stoichiometry, then the nonmodified counterparts of the brain N-glycopeptides should rarely be detected. In a separately measured brain proteome, we found that 98.6% of 2714 N-linked glycosylated peptides were not present in their unmodified form among 33,198 identified peptides. The lack of such peptides suggests high occupancy of glycosylation sites.

Of the total of 1938 N-glycosylated proteins from the top confidence set, approximately half carried a single N-linked sugar chain (Figure 4C). For 22% we detected two N-glycosylation sites, 13% carried three identified sites, and the average degree of glycosylation was 2.6. Notably, there was a group of 247 proteins that contained 5 or more N-glycosylation sites and 44 with at least 10. Low-density lipoprotein receptor-related protein I and II were the most heavily glycosylated proteins with 41 and 32 asparagine-linked glycosylation sites, respectively.

We applied cluster analysis to test whether N-glycosites occur in specific sequence segments of the proteins or whether they are randomly distributed on the primary sequence (Extended Experimental Procedures). Bootstrapping-based comparison of primary sequence distances between N-glycosylated and non-N-glycosylated asparagines of multiply glycosylated proteins did not reveal significant differences (Table S3B). Instead, the sites in some proteins such as sortilin (858 residues; N160, N272, N404, N682) are evenly distributed over the sequence, whereas they tightly cluster in other proteins such as Latrophilin-1 (1466 residues; N526, N635, N736, N795, N800).

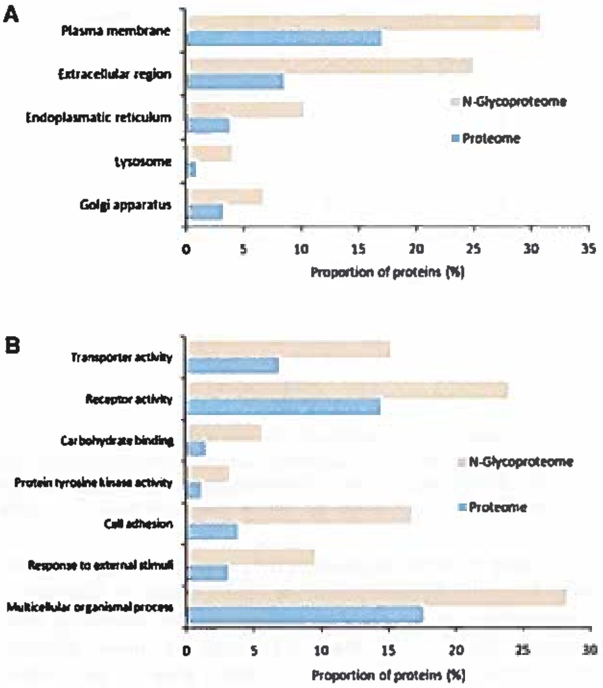


Figure 5. Gene Ontology Analysis
Cellular components (A) and molecular functions and biological processes (B) that are significantly overrepresented in the N-glycoproteome compared to the entire mouse proteome, according to Gene Ontology analysis. See also Table S4 and Table S5.

Cellular and Functional Classification of N-Glycosylated Proteins

We wished to obtain an overview of the subcellular compartments and the cellular functions that N-linked glycoproteins are preferentially associated with. We applied Cytoscape (Shannon et al., 2003) and BinGO (Maere et al., 2005) to determine Gene Ontology (GO) categories overrepresented in the glycoproteome compared to the entire mouse proteome.

A total of 31% of the N-linked glycoproteome was in the "plasma membrane" GO category and 25% in "extracellular region" (Table S4A). Taking into account nonexclusive localization in GO, 52% of the N-glycoproteome is located at the outside or beyond the plasma membrane (832 of 1594 N-glycoproteins with GO annotation). Furthermore, the ER, the Golgi apparatus, and the lysosome are overrepresented (Figure 5A). We also found N-glycosylated proteins associated with the peroxisome, endosome, and vacuole. Together, these intralumenal locations of cellular organelles accounted for 20% of the N-glycoproteome with GO annotation. Intriguingly, a number of proteins were assigned to compartments that are not topologically connected with the lumen of the ER or Golgi. However, in almost all cases, these annotations were nonexclusive or they were different in other databases such as Ensembl.

To directly address the long-standing question of N-glycosylation in unexpected cell compartments, we applied subcellular fractionation to mouse liver using sucrose gradient separation (Extended Experimental Procedures). Consistent with the above results, we did not find any N-glycosites in proteins that were annotated on the basis of experimental evidence to be localized exclusively in the nucleus, in the mitochondria, or in the cytosol.

Within the set of plasma membrane proteins, 99% of 912 N-glycosylated sites were located in the extracellular region based on predicted Swiss-Prot topology assignments. Our dataset contains five examples of N-glycosylation, annotated to occur in cytoplasmic domains (Table S4B). However, each of these topology assignments was based on bioinformatic prediction rather than experimental evidence. Given the uncertainties of such predictions and the fact that only 5 out of 912 sites were mapped to an unexpected topological location, we conclude that our dataset contains no statistically significant evidence for N-glycosylation on the cytoplasmic face of proteins.

Many functions that are known to be characteristic for N-glycoproteins were enriched in our set, including transporter activity, receptor activity, and carbohydrate binding (Figure 5B, Table S4C). Cell adhesion, response to external stimuli, and multicellular organismal processes were the major overrepresented biological processes (Figure 5B, Table S4D). Most of the functional categories appear to be secondary to the location of the proteins at the membrane. For example, "transmembrane protein tyrosine kinase activity" is significantly overrepresented ($p < 10^{-22}$).

To test the robustness of our bioinformatic enrichment analysis, we repeated it by matching only the brain N-glycoproteome against the brain proteome derived as part of this study (5880 proteins). We obtained very similar results, indicating that the results are not tissue specific nor affected by using the total proteome instead of tissue proteome as background (Table S4E).

KEGG pathway enrichment analysis (Kanehisa and Goto, 2000) using DAVID (Dennis et al., 2003) led to similar results and additionally revealed that our N-glycoproteome is enriched for proteins that are involved in neurodegenerative diseases such as Alzheimer's and prion disease (Table S5).

Tissue Distribution of the N-Glycoproteome

The function and extent of N-linked glycosylation are likely to be different between the tissues. To investigate this issue, we separately analyzed and overlapped the N-glycosites according to the tissue in which they had been identified.

The blood plasma N-glycoproteome comprised 1119 sites. According to GO annotation, they overwhelmingly mapped to "extracellular space" with only a few sites from lumenal organellar localizations. As blood is present in all tissues—even after perfusion of mice, as done here—sites that were identified in both blood plasma and another tissue cannot be unambiguously assigned to one of the tissues (marked in orange in Figure 6).

The highest number of N-glycosylation sites (3162) was observed in brain. Of these, 1140 were not identified in any other tissue and this group includes many brain-specific proteins. Heart had the lowest number of identified sites (2213 total and 93 exclusive). All tissues had a large proportion of N-glycosylation sites that were found in at least one other tissue (purple in Figure 6).

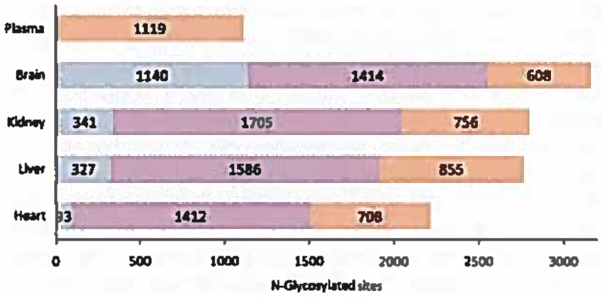


Figure 6. N-Glycoproteomes of Different Mouse Tissues
Number of identified N-glycosylated sites in blood plasma, brain, kidney, liver, and heart. Sites detected in blood plasma are in orange. Gray: sites only detected in one organ; purple: sites detected in at least two organs. See also Figure S3 and Table S6.

Accurate Relative Quantification of the In Vivo N-Glycoproteome

To demonstrate the capability of our approach to quantify N-glycoproteome changes under different conditions in vivo, we applied stable isotope labeling of amino acids in cell culture (SILAC) (Ong et al., 2002). Using brains from non-SILAC and SILAC mice (Kruger et al., 2008) we compared the N-glycosylation site pattern in old versus young mice (Extended Experimental Procedures). We performed two independent experiments, using different proteolytic enzymes, and each experiment was repeated after swapping the SILAC labels between old and young mice. We quantified 763 N-glycosylation sites using LysC and 1118 N-glycosylation sites using trypsin. In both sets we found the same N-glycosylation sites to be under- or overrepresented in forward and reverse experiments (Figure S3A, Table S6A). To determine if the apparent glycosylation changes were instead due to changes in protein abundance, we also quantified the brain proteins between the mice. Indeed, we found that the detected regulations occur not on the site but at the protein level (Table S6B and Table S6C). Although these experiments do not exclude subtle changes between the N-glycoproteome as a function of age, they suggest that there are no drastic changes. They also demonstrate that the N-glyco-FASP method is fully compatible with accurate SILAC-based quantification.

Evolution of N-Glycosylation

To derive orthologous proteins in 36 eukaryotic species, we assigned N-glycosylated proteins to their corresponding gene entries in Ensembl (Flicek et al., 2008; Gnad et al., 2009) and retrieved evolutionary annotation data from the Ensembl Compara Database (Vilella et al., 2009). We defined only "one-to-one" orthologs as interspecies homologs in our analysis. This strict definition excludes between-species paralogs and "one-to-many" or "many-to-many" orthologs in the homology set. Mouse N-glycoproteins have more orthologs in vertebrates compared to all mouse proteins (Figure S3B). However, in contrast to acetylated and phosphorylated proteins (Choudhary et al., 2009; Gnad et al., 2007), mouse N-glycosylated proteins have less orthologs in invertebrates and yeast. For example, 8% of mouse N-glycoproteins have orthologs in fly in comparison to 14% non-N-glycosylated mouse proteins. This finding also held true for each tissue-specific N-glycoproteome and may reflect the specific functions of many N-glycosylated proteins in multicellular organisms. Consistent with the drop in conservation to lower organisms, only 32 of the 829 N-glycosylated proteins identified in a *C. elegans* N-glycoproteome (Kaji et al., 2007) had one-to-one orthologs in our mouse N-glycosets. Finally, as expected, N-glycoproteins were not significantly conserved to prokaryotes compared to the entire proteome, as assessed by BLAST-based conservation analysis (Figure S3C).

DISCUSSION

A Proteomic Approach for In-Depth Mapping of the In Vivo N-Glycoproteome

We have developed and applied a strategy based on FASP, multi-lectin affinity, and high-accuracy mass spectrometric characterization. The N-glyco-FASP method allowed the determination of over 6000 N-glycosylation sites from five mouse tissues. In single LC-MS/MS runs it is possible to map over 2000 sites from 200 μg starting material, which opens up interesting areas of application. Enrichment efficiency is reflected by the detection of more than 1000 N-glycosites on almost 500 proteins in blood plasma—thus N-glyco-FASP may be an efficient method for plasma proteome characterization in a biomarker context (Zhang et al., 2005). Importantly, the difficult class of membrane proteins, on which most N-glycosylation occurs, was excellently covered in our dataset due to the use of SDS solubilization in the FASP protocol. Further highlighting the deep coverage of the brain N-glycoproteome, we found that 58% of 1296 identified N-glycoproteins were not contained in our brain proteome consisting of 5880 proteins. Although demonstrated for multi-lectin affinity here, the method is equally applicable to the enrichment of other modified peptides. For example, phosphotyrosine peptides could be enriched with a mixture of anti-phosphotyrosine antibodies, which would not have to be coupled to any support.

The data in our study were acquired with the LTQ-Orbitrap Velos mass spectrometer with which peptide ions can be efficiently and sensitively fragmented via HCD (Olsen et al., 2009). Mass accuracy for fragments was in the low ppm range, about a factor 100 better than what we achieved in recent large-scale studies of the phosphoproteome and the acetylome (Choudhary et al., 2009; Olsen et al., 2006). Thus a "high high" strategy with HCD is a powerful technology for in-depth posttranslational modification identification, as it allows very high-confidence identification of individual peptides and ensures precise localization of modifications within the peptide sequence.

N-Glycoproteome Characterization

Our dataset covers 74% of all known mouse N-glycosylation sites and increases the number of all experimentally identified N-glycosites by a factor of about seven compared to the Swiss-Prot reference database. Interestingly, the number of identified N-glycosylation sites does not increase significantly with repeated measurements and additional fractionation, suggesting that our dataset covers a substantial part of the mouse N-glycoproteome. This is in contrast to phosphoproteome analysis, in which repeated runs and fractionation add a large percentage of additional sites. Detection of 2352 N-linked glycoproteins demonstrates that more than 10% of the mouse proteome is N-glycosylated. The fact that glycosylated peptides are rarely found in their unmodified form points to high site occupancy (stoichiometry). In contrast, phosphorylation and acetylation are generally attached to the proteins in a substoichiometric manner, and nonmodified counterparts can be found for a large proportion of the phosphopeptides in proteomics studies (Olsen et al., 2010).

Biology and Evolution of the N-Glycoproteome

We find N-glycosites to be predominantly located in loops and turns on the protein surface, similarly to what we previously found for phosphorylation or acetylation sites (Choudhary et al., 2009; Gnad et al., 2007). However, unlike either acetylation sites or phosphorylation sites, N-glycosylation is also preferentially located in β sheets. This points to a more stable and rigid binding of cotranslationally attached sugar molecules in contrast to reversible phosphorylation for which sites have to be accessible posttranslationally to kinases and phosphatases.

Besides such structural constraints, the vast majority (96.5%) of N-glycosylated asparagines match the stringent glycosylation consensus sequence N-!P-[S|T]. Sequence motif analysis reveals additional constraints on the known sequence recognition pattern. We find that proline is also underrepresented on the third position relative to the asparagines. This has already been shown in *C. elegans* (Kaji et al., 2007) but not in vertebrates. Threonine occurs more often than serine on the second position, in contrast to nonglycosylated sites that match the motif by chance and to phosphorylation sites, which occur much more often on serine than threonine. Furthermore, there are scattered reports of N-X-C motif on individual proteins. Our results show that this motif is widespread but that it occurs with a low frequency of about 1.3%. Furthermore, 2.2% of the N-glycosites did not match with either of the two motifs. We found these sites to be enriched for either valine on the second position or glycine on the first position relative to the N-glycosylated asparagine. The N-G and N-X-V sequence overrepresentation is statistically significant (Figure 4), and it will be interesting to investigate if it represents further minor motifs for the N-glycosylation machinery. Additionally, western blotting validated the occurrence of N-glycosylation on sites that do not match with the known motifs (Figure S2). Thus our study provides in silico as well as experimental evidence for N-glycosylation on consensus sequences different from N-!P-[S|T] and N-X-C.

Evolutionary analysis reveals that N-glycosylated mouse proteins are more highly conserved throughout vertebrates but not invertebrates compared to non-N-glycosylated mouse proteins. This finding underlines the essential role of N-glycosylation on proteins that evolved with the rise of vertebrates, in concordance with the role of N-glycosylation in complex multicellular organisms. In particular, the adaptive immune system evolved in vertebrates and is thought to be intimately connected with recognition glycostructures by host and pathogen, providing a further reason why N-glycoproteins are more highly conserved than their nonglycosylated counterparts.

Cellular Organization of N-Glycosylation

We almost exclusively found N-glycosylation to occur on secreted molecules, on the extracellular face of membrane proteins, and on the lumenal side of ER, Golgi apparatus, or lysosomes. This is in agreement with the topology of glycosyltransferases that attach the sugar chains and usually have their active sites within the lumens of the ER and Golgi.

There have been suggestions regarding the presence of N-glycosylated proteins in the nuclear, cytoplasmatic, and mitochondrial compartments for many years (Pedemonte et al., 1990; Reeves et al., 1981; Chandra et al., 1998; Kung et al., 2009). Several experimental and conceptual attempts have been undertaken to explain and prove this hypothesis. Nonconventional soluble glycosyltransferases may exist in the cytoplasm or nucleus and directly modify the proteins in these compartment, or soluble N-glycosylated proteins may be flipped across membranes or originate from secretory pathways (Varki et al., 2009). Experimental evidence for such mechanisms was mostly based on protein binding to lectins or radioactive labeling but did not include mapping of the sugar attachment sites. For example, sodium/potassium-transporting ATPase subunit alpha from dog kidney was reported to contain N-glycan in its cytoplasmatic domain but without defining the sites (Pedemonte et al., 1990). Here, we detected a few N-glycosylation sites on parts of the proteins sequence that were annotated as cytoplasmic domains of plasma membrane proteins; however the topology of these proteins is based on prediction methods. Therefore this observation does not supply experimental evidence for the presence of N-glycosylation in the cytoplasm. Instead, despite in-depth mapping of these cellular compartments, we did not detect any proteins that were annotated to occur exclusively in the nucleus, in the mitochondria, or in the cytosol. In summary, our data point to a universal requirement of N-glycosylation sites to be in topological continuity with the lumen of the ER and Golgi.

Tissue-Specific and Disease-Related N-linked Glycoproteins

Many novel glycosylation sites have been detected for tissue-specific proteins, such as neurotransmitter receptors and contactins for brain or cubilin and megalin for kidney. For example, we found 30 kidney-specific sites on megalin (low-density lipoprotein receptor-related protein 2, 520 kDa), out of which only one was reported in a recent N-glycoproteomic study (Wollscheid et al., 2009).

Besides tissue-specific N-glycosylation sites, we also detected a number of N-glycoproteins that are associated with various diseases. For example, we found many N-glycosites on proteins that are involved in Alzheimer's disease (AD), the most common type of dementia (Price et al., 1998). Because N-glycosylation is involved in many processes impaired in AD, such as protein folding, protein anchoring to cell membranes, and protein delivery to organelles, it is possible that N-glycosylation is directly involved in cause or progress of AD (Selkoe, 2004; Suzuki et al., 2006). Surprisingly, using both high-accuracy mass spectrometry and western blotting, we found N-glycosylation on Apolipoprotein E, an important AD protein (Kim et al., 2009), which does not contain any asparagines that match the known motifs N-!P-[S|T| and N-X-C.

Summary and Outlook

Here we provide a practical and highly efficient method for mapping the attachment sites of N-glycostructures. Modification sites occur on many proteins of pivotal importance in development, multicellular communication, and many other basic biological functions. These high-confidence N-glycosylation sites can now be used by the community for detailed functional studies. We have also shown that N-glyco-FASP is fully compatible with quantitative proteomics methods such as SILAC, which could be used to study the many diseases that directly or indirectly involve N-glycosylation.

EXPERIMENTAL PROCEDURES

Tissue Preparation, Protein Extraction, and Digestion

Tissues used in our study were dissected from inbred mouse strain C57BL/6, males, aged 3 months and proceeded according to the FASP protocol (Wisniewski et al., 2009b). The resulting peptides were eluted in lectin binding buffer (1 mM $CaCl_2$, 1 mM $MnCl_2$, 0.5 M NaCl in 20 mM TrisHCl, pH 7.3). For more details, see the Extended Experimental Procedures.

Lectin Enrichment and Deglycosylation

Approximately 100 µg of digested peptides were mixed with lectin solution containing either a combination of ConA, WGA, and RCA_{120} (Sigma) or just one lectin at a time, resulting in mixtures of peptides and lectins with a mass proportion of 1:2. The mixtures were transferred to new YM-30 filter units (Microcon, Millipore). After 1 hr incubation at room temperature the unbound peptides were eluted by centrifugation. The captured peptides were washed followed by deglycosylation with PNGase F (Roche). After incubation for 3 hr at 37°C deglycosylated peptides were eluted. A more detailed protocol is available in the Extended Experimental Procedures section.

Mass Spectrometric Analysis

The eluted peptides were purified on StageTips (Rappsilber et al., 2003) and analyzed on the LTQ-Orbitrap Velos mass spectrometer; (Thermo Fisher Scientific, Germany) coupled to HPLC via a nanoelectrospray ion source. Full MS scans were acquired with resolution of 30,000 at *m/z* 400 in the orbitrap analyzer. The ten most intense ions were fragmented by higher-energy collisional dissociation (HCD) in the octopole collision cell. The HCD fragment ion spectra were acquired in the orbitrap analyzer with resolution of 7500 at m/z 400. The MS data were analyzed using MaxQuant (Cox and Mann, 2008). Proteins were identified by searching MS and MS/MS data using the MASCOT search engine (Matrix Science, UK) (Perkins et al., 1999). See the Extended Experimental Procedures for further information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, three figures, and six tables and can be found with this article online at doi:10.1016/j.cell.2010.04.012.

ACKNOWLEDGMENTS

We thank Alexandra Zougman, Nagarjuna Nagaraj, Korbinian Mayr, Mara Monetti, Ina Rohwedder, Markus Moser, and Pawel Ostasiewicz for help and fruitful discussions. This work was supported by PROSPECT, a 7th framework program of the European Union (grant agreement HEALTH-F4-2008-201648/PROSPECTS).

Received: October 27, 2009
Revised: February 8, 2010
Accepted: April 7, 2010
Published: May 27, 2010

REFERENCES

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

Bradford, B.M., Tuzi, N.L., Feltri, M.L., McCorquodale, C., Cancellotti, E., and Manson, J.C. (2009). Dramatic reduction of PrP C level and glycosylation in peripheral nerves following PrP knock-out from Schwann cells does not prevent transmissible spongiform encephalopathy neuroinvasion. J. Neurosci. *29*, 15445–15454.

Bunkenborg, J., Pilch, B.J., Podtelejnikov, A.V., and Wisniewski, J.R. (2004). Screening for N-glycosylated proteins by liquid chromatography mass spectrometry. Proteomics *4*, 454–465.

Chandra, N.C., Spiro, M.J., and Spiro, R.G. (1998). Identification of a glycoprotein from rat liver mitochondrial inner membrane and demonstration of its origin in the endoplasmic reticulum. J. Biol. Chem. *273*, 19715–19721.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science *325*, 834–840.

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.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, 3.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. (2008). Ensembl 2008. Nucleic Acids Res. *36*, D707–D714.

Gnad, F., Oroshi, M., Birney, E., and Mann, M. (2009). MAPU 2.0: high-accuracy proteomes mapped to genomes. Nucleic Acids Res. *37*, D902–D906.

Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol. *8*, R250.

Gundry, R.L., Raginski, K., Tarasova, Y., Tchernyshyov, I., Bausch-Fluck, D., Elliott, S.T., Boheler, K.R., Van Eyk, J.E., and Wollscheid, B. (2009). The mouse C2C12 myoblast cell surface N-linked glycoproteome: identification, glycosite occupancy, and membrane orientation. Mol. Cell. Proteomics *8*, 2555–2569.

Hulsmeier, A.J., Paesold-Burda, P., and Hennet, T. (2007). N-glycosylation site occupancy in serum glycoproteins using multiple reaction monitoring liquid chromatography-mass spectrometry. Mol. Cell. Proteomics *6*, 2132–2138.

Jensen, O.N. (2006). Interpreting the protein language using proteomics. Nat. Rev. Mol. Cell Biol. *7*, 391–403.

Kaji, H., Kamiie, J., Kawakami, H., Kido, K., Yamauchi, Y., Shinkawa, T., Taoka, M., Takahashi, N., and Isobe, T. (2007). Proteomics reveals N-linked glycoprotein diversity in Caenorhabditis elegans and suggests an atypical translocation mechanism for integral membrane proteins. Mol. Cell. Proteomics *6*, 2100–2109.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kim, J., Basak, J.M., and Holtzman, D.M. (2009). The role of apolipoprotein E in Alzheimer's disease. Neuron *63*, 287–303.

Kruger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C.A., Forner, F., Schmidt, S., Zanivan, S., Fassler, R., and Mann, M. (2008). SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. Cell *134*, 353–364.

Kung, L.A., Tao, S.C., Qian, J., Smith, M.G., Snyder, M., and Zhu, H. (2009). Global analysis of the glycoproteome in Saccharomyces cerevisiae reveals new roles for protein glycosylation in eukaryotes. Mol. Syst. Biol. *5*, 308.

Kuster, B., and Mann, M. (1999). 18O-labeling of N-glycosylation sites to improve the identification of gel-separated glycoproteins using peptide mass mapping and database searching. Anal. Chem. *71*, 1431–1440.

Lee, A., Kolarich, D., Haynes, P.A., Jensen, P.H., Baker, M.S., and Packer, N.H. (2009). Rat liver membrane glycoproteome: enrichment by phase partitioning and glycoprotein capture. J. Proteome Res. *8*, 770–781.

Liu, T., Qian, W.J., Gritsenko, M.A., Camp, D.G., 2nd, Monroe, M.E., Moore, R.J., and Smith, R.D. (2005). Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. J. Proteome Res. *4*, 2070–2080.

Macek, B., Mann, M., and Olsen, J.V. (2009). Global and site-specific quantitative phosphoproteomics: principles and applications. Annu. Rev. Pharmacol. Toxicol. *49*, 199–221.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448–3449.

Medzihradszky, K.F. (2005). Characterization of protein N-glycosylation. Methods Enzymol. *405*, 116–138.

North, S.J., Huang, H.H., Sundaram, S., Jang-Lee, J., Etienne, A.T., Trollope, A., Chalabi, S., Dell, A., Stanley, P., and Haslam, S.M. (2010). Glycomics profiling of Chinese hamster ovary cell glycosylation mutants reveals N-glycans of a novel size and complexity. J. Biol. Chem. *285*, 5759–5775.

Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell *127*, 635–648.

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.

Olsen, J.V., Schwartz, J.C., Griep-Raming, J., Nielsen, M.L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., et al. (2009). A dual pressure linear ion trap - Orbitrap instrument with very high sequencing speed. Mol. Cell. Proteomics *8*, 2759–2769.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., et al. (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci. Signal. *3*, ra3.

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.

Pedemonte, C.H., Sachs, G., and Kaplan, J.H. (1990). An intrinsic membrane glycoprotein with cytosolically oriented n-linked sugars. Proc. Natl. Acad. Sci. USA *87*, 9789–9793.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551–3567.

Price, D.L., Tanzi, R.E., Borchelt, D.R., and Sisodia, S.S. (1998). Alzheimer's disease: genetic studies and transgenic models. Annu. Rev. Genet. *32*, 461–493.

Ramachandran, P., Boontheung, P., Xie, Y., Sondej, M., Wong, D.T., and Loo, J.A. (2006). Identification of N-linked glycoproteins in human saliva by glycoprotein capture and mass spectrometry. J. Proteome Res. *5*, 1493–1503.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. *75*, 663–670.

Reeves, R., Chang, D., and Chung, S.C. (1981). Carbohydrate modifications of the high mobility group proteins. Proc. Natl. Acad. Sci. USA *78*, 6704–6708.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. *23*, 1391–1398.

Selkoe, D.J. (2004). Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases. Nat. Cell Biol. *6*, 1054–1061.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Suzuki, T., Araki, Y., Yamamoto, T., and Nakaya, T. (2006). Trafficking of Alzheimer's disease-related membrane proteins and its participation in disease pathogenesis. J. Biochem. *139*, 949–955.

Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W., and Etzler, M.E. (2009). Essentials of Glycobiology (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Wagner, M., Adamczak, R., Porollo, A., and Meller, J. (2005). Linear regression models for solvent accessibility prediction in proteins. J. Comput. Biol. *12*, 355–369.

Wisniewski, J.R., Zougman, A., and Mann, M. (2009a). Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J. Proteome Res. *8*, 5674–5678.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009b). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

Witze, E.S., Old, W.M., Resing, K.A., and Ahn, N.G. (2007). Mapping protein post-translational modifications with mass spectrometry. Nat. Methods *4*, 798–806.

Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., Aebersold, R., and Watts, J.D. (2009). Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. Nat. Biotechnol. *27*, 378–386.

Woods, R.J., Edge, C.J., and Dwek, R.A. (1994). Protein surface oligosaccharides and protein function. Nat. Struct. Biol. *1*, 499–501.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., et al. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. *34*, D187–D191.

Yang, Z., and Hancock, W.S. (2004). Approach to the comprehensive analysis of glycoproteins isolated from human serum using a multi-lectin affinity column. J. Chromatogr. A *1053*, 79–88.

Zajonc, D.M., Striegl, H., Dascher, C.C., and Wilson, I.A. (2008). The crystal structure of avian CD1 reveals a smaller, more primordial antigen-binding pocket compared to mammalian CD1. Proc. Natl. Acad. Sci. USA *105*, 17925–17930.

Zhang, H., Li, X.J., Martin, D.B., and Aebersold, R. (2003). Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nat. Biotechnol. *21*, 660–666.

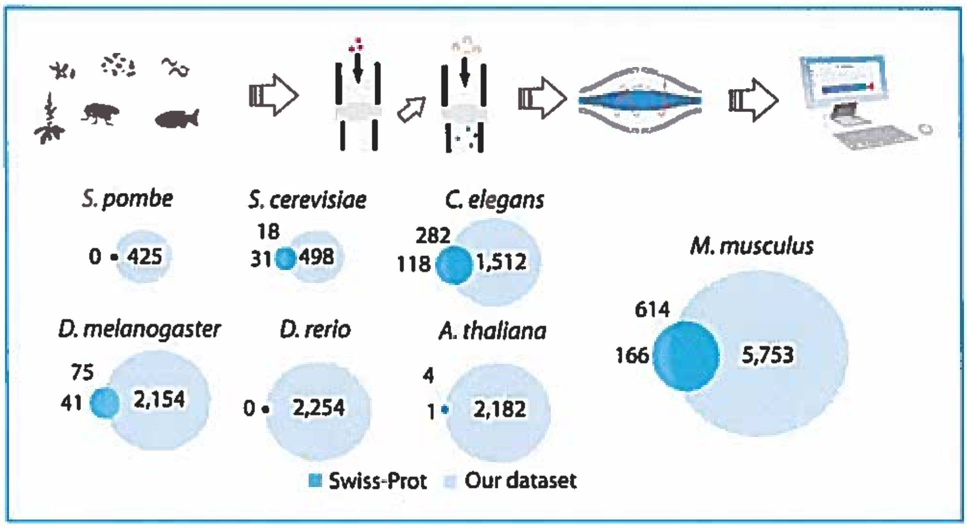
Zhang, H., Yi, E.C., Li, X.J., Mallick, P., Kelly-Spratt, K.S., Masselon, C.D., Camp, D.G., 2nd, Smith, R.D., Kemp, C.J., and Aebersold, R. (2005). High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. Mol. Cell. Proteomics *4*, 144–155.

Zielinska, D.F., Gnad, F., Jedrusik-Bode, M., Wisniewski, J.R., and Mann, M. (2009). Caenorhabditis elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. J. Proteome Res. *8*, 4039–4049.

Publication II

Evolution of N-glycosylation

Zielinska DF, Gnad F, Schropp K, Wiśniewski JR, Mann M



(Submitted)

Here we used the previously developed technology to answer the question: how did the N-glycoproteomes evolve? We detected nearly 10,000 N-glycosylation sites from six evolutionary distant model species that span more than a billion years in evolution. Surprisingly, we found that N-glycoproteomes, unlike phosphoproteomes or acetylomes, are more conserved within the same phylogenetic group compared to the corresponding proteomes, but less in distant species.

Evolution of the N-glycoproteome

Dorota F. Zielinska*, Florian Gnad*, Katharina Schropp, Jacek R. Wiśniewski & Matthias Mann****

Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany
* These authors contributed equally to this study
** Correspondence: jwisniew@biochem.mpg.de; mmann@biochem.mpg.de

N-linked glycosylation is an important posttranslational modification involved in diverse biological processes and it plays a major role in the development of complex multicellular organisms[1]. The core mechanism of N-glycosylation is highly conserved in eukaryotic species[2,3] but almost nothing is known about the evolution of N-glycoproteomes. Mass spectrometric methods now allow in-depth and highly accurate mapping of the sites of N-glycosylation[4-8]. Here, we measure N-glycoproteomes of the major model organisms *Arabidopsis thaliana, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster* and *Danio rerio* providing a high quality resource to the respective communities. Together with the recently determined in-depth mouse N-glycoproteome[8] these data sets representatively span the eukaryotic domain of life. The number of detected N-glycosylation sites varied between 425 in fission yeast, 516 in budding yeast, 1,794 in worm, 2,186 in plant, 2,229 in fly and 2,254 in zebrafish. We find that all eukaryotic N-glycoproteomes have invariant characteristics including sequence recognition patterns, structural constraints and subcellular localization. However, a surprisingly large percentage of the N-glycoproteome evolved after the phylogenetic divergences between plants, fungi, nematodes, insects and vertebrates. Many N-glycosylated proteins coevolved with the rise of extracellular processes that are specific within corresponding phylogenetic groups and essential for organismal development, body growth and organ formation.

To rapidly characterize the sites of N-glycan attachment in multiple species, we applied the recently described N-glyco-FASP method that combines highly efficient 'on-filter' protein digestion[9], multiple-lectin enrichment and PNGase deglycosylation[8]. We chose to investigate the established model organisms *Arabidopsis thaliana, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster* and *Danio rerio* (Fig. 1a). Whole organisms were homogenized in strongly denaturating buffer and processed on filter, which enable the removal of detergent, digestion enzyme, lectins and PNGase. Lectin mixture containing Con A, WGA and RCA_{120} were added to the top of the filters to capture the three N-glycan classes (high-mannose, complex and hybrid). Peptides from non-glycosylated proteins as well as non-glycopeptides from glycoproteins were efficiently eluted resulting in better than 100-fold enrichment. Deglycosylation was performed in $H_2{}^{18}O$, followed by high resolution, high mass accuracy measurement of the resulting peptide precursors and MS/MS fragments on a linear ion trap Orbitrap mass spectrometer[10], ensuring extremely low false positive rates for peptide identification and site determination[8]. A slightly modified protocol was applied to plant material.

To maximize coverage of the N-glycoproteome we used two different enzymes – trypsin and GluC – for protein digestion. This yielded a 25% to 50% increase in the number of identified sites compared to tryptic digestion alone. For each enzyme the measurement of multiple replicates only added few novel sites, suggesting that essentially all detectable N-glycopeptides were indeed identified. Furthermore, site identification was highly reproducible between runs showing that the applied method was robust and applicable to entire, homogenized organisms. Cumulatively and including the previously measured mouse N-glycoproteome[11] , we determined 15,771 sites across the seven model organisms at high confidence (99% certainty of identification and 95% certainty of single amino acid localization).

In fission and budding yeast, 425 and 516 high-confident sites were mapped, while in multicellular eukaryotes the number ranged from 1,794 in worm to 2,186 in plant, 2,229 in fly and 2,254 in zebrafish. These 9,404 distinct sites occurred on 4,900 proteins (Supplementary Table 1,2). To date little was known about the N-glycoproteomes of plant, yeast, fly and zebrafish and essentially all the sites identified here are novel. *C. elegans* is the only model organism for which a large-scale glycoproteome has previously been reported[12]. Overall, according to the Swiss-Prot database[13] 96% of the 9,404 sites identified in the six species are novel and we extend the number of known sites in these species more than fifteen-fold (Fig. 1b).

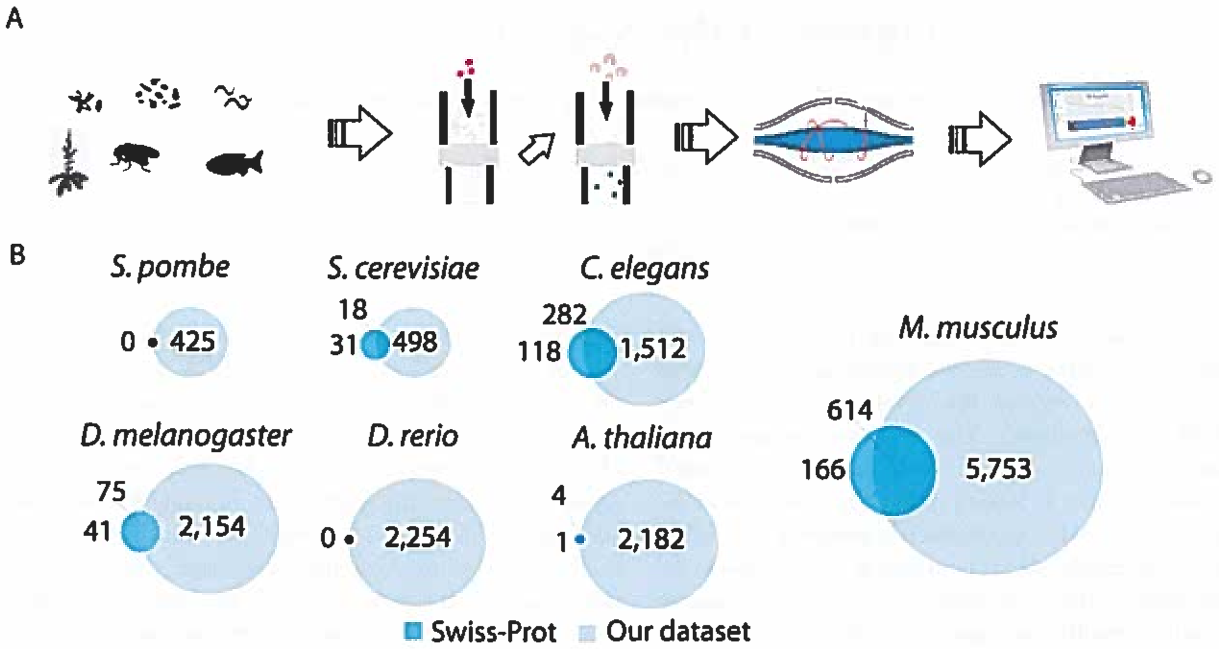


Figure 1: Workflow and identified N-glycoproteomes. a, Proteins were extracted from *A. thaliana, S. cervisiae, S. pombe, C. elegans, D. melanogaster and D.rerio* and processed using the N-glyco-FASP method. Peptides were measured on high resolution mass spectrometer with high mass accuracy for precursor and fragment ions. **b,** Overlaps between N-glycosylation sites annotated in the Swiss-Prot database and detected in this study.

N-glycosylation might also occur in prokaryotes[14] and we therefore used our methodology to investigate this possibility. Up to now, a bacterial N-glycosylation system that is similar to the eukaryotic one was exclusively found in *C. jejuni*[15]. It is the only bacteria whose genome encodes a homologous protein to the eukaryotic oligosacharyltransferase STT3. This gene was reported to be acquired by lateral gene transfer from Achaea or eukaryotes[15]. However, the core structure of *C. jejuni* N-glycans is different from eukaryotes, implying their resistance to PNGase digestion[16-18]. Consistently, in the bacterial model systems *E. coli* and *B. subtilis* we found no firm evidence for N-glycosylation sites – at least those recognized by the lectins employed here and liberated by PNGase F.

In mice, N-glycosylation adheres to very rigid topological and sequence constraints[8]. In all model systems, N-glycosylated proteins were annotated to be located in the cell outside or in the expected intracellular compartments such as ER and Golgi (Supplementary Fig. 1a, Supplementary Table 3). The extracellular region was particularly over-represented according to Gene Ontology (GO) analysis[19]. Based on experimentally validated annotations we did not find evidence for the occurrence of N-glycosylation in non-expected topological locations, such as mitochondria, cytosol and nucleus. Specifically, we did not find mitochondrial N-glycoproteins as had been observed in a recent study based on protein microarrays[20].

In the investigated species 97 to 99% of the identified N-glycosites match the known canonical motif N-!P-[S|T], where !P represents any amino acid except proline. Threonine occurs more frequently than serine at the second position (Supplementary Fig. 1b) and proline is not only underrepresented at the first position, but also at the third position (Supplementary Table 4). No additional constraints around the canonical motif could be derived by statistical tests or machine learning approaches (Supplementary Information). Only a minute proportion of sites match with other motifs including the previously reported N-G and N-X-C motifs[8] (Supplementary Fig. 2). Approximately half of the identified glycoproteins carry a single sugar chain, about 20% of the proteins have two identified N-glycosylation sites and 25% three or more sites (Supplementary Fig. 1c). N-glycosylation sites were enriched in β-sheets and depleted in α-helices in all organisms (Supplementary Fig. 1d). Thus our analysis of model species separated by more than 1,000 million years revealed that rigid topological and sequence constraints on N-glycosylation are universal in eukaryotes. This finding is in concordance with the fact that the underlying core 'N-glycosylation machinery' consists of relatively few proteins, which are highly conserved in all eukaryotic species, whereas the different eukaryotic glycosyltransferase families are more complex and species specific[1].

In the extensively studied *S. cerevisiae* and *S. pombe* model organisms, in total 49 N-glycosylation sites were known before. We detected approximately 500 distinct sites pro analyzed species, demonstrating that these unicellular organisms extensively employ N-glycosylation. In addition to the classical cellular localizations, a large proportion of sites occur on fungi specific proteins incorporated into the yeast cell wall. Consequently many yeast N-glycoproteins are involved in the organization and biogenesis of the cell wall such as the mannoprotein PST1 in *S. cerevisiae* or alpha-1,3-glucan synthase AGS1 in *S. pombe*. Despite its unicellularity, yeast encodes secreted proteins some of which we found to be N-glycosylated, including the canonical example *S. cerevisiae* invertase 2.

In accordance with the pivotal role of N-glycosylation in extracellular processes such as cell-cell communication, organ development and body growth, the number of detected sites was substantially higher in all multicellular organisms compared to yeast (Supplementary Table 5). For example in all studied animals, we detected an average of 45 N-glycosylation sites on laminins and 30 on integrins. We found species specific glycosylated subunits of these proteins, such as the insect specific protein integrin alpha-PS2, which functions in sensory perception and in morphogenesis[21]. In addition to response to environmental conditions N-glycosylated proteins of multicellular animals play a fundamental role in development. In fly, smooth and timp are involved in the development of the wings[22]. We mapped sites on the N-terminal extracellular domain of smooth and the NTR domain of the secreted protein timp, which is involved in axon guidance. Zebrafish is an important vertebrate model of development and disease but no N-glycosylation sites have been mapped experimentally so far. Among the 2,254 high confidence sites reported here, many map to developmentally important proteins, such as three sites in the vertebrate specific neural cadherin (Chd2)[23]. Deficiency in delta-sarcoglycan is linked to cardiac dysfunction[24]. Standard bioinformatics methods did not predict this integral membrane protein to be glycosylated and also incorrectly predicted its membrane topology (Supplementary Fig. 3), illustrating the usefulness of our experimental data. Amyloid beta protein is an intensively studied Alzheimer's disease protein[25] whose N-glycosylation status was not known in the zebrafish model before this study.

In green plant, despite its biotechnological and industrial potential, only few N-glycosylation sites are known. In contrast to fungi and animals, green plants contain chloroplasts. Initially it was thought that proteins can only enter the chloroplast in an organelle-specific way[26,27]. Recently, it was shown that chloroplast-located proteins can take an alternative route through the secretory pathway and become N-glycosylated before entering the chloroplast[28]. Our dataset with several examples of N-glycosylated chloroplast-located proteins, including three identified sites on alpha-carbonic anhydrase that was reported to be exclusively located in the chloroplast stroma[28], proves this model.

As already mentioned for yeast, the different N-glycoproteomes contain a large number of proteins specific to the phylogenetic class of the model organism. This was unexpected given the common core N-glycosylation machinery and raised the question of the extent of conservation of the N-glycoproteome. Glycosyltransferase responsible for the modification of sugar chains and consequently the glycan families co-developed with the evolution of multicellularity[1], but lack of experimental data has so far prevented any investigation of the evolution of N-glycosylated substrates.

Global bioinformatic analysis clearly showed that the identified N-glycoproteomes are higher conserved than the non-N-glycosylated proteomes within their corresponding phylogenetic group, but less in other phyla (Fig. 2, Supplementary Fig. 4, Supplementary Table 6). Relative to their proteomes, the *S. pombe* and *S. cerevisiae* N-glycoproteomes are highly conserved in other fungi, whereas only a low proportion of yeast N-glycoproteins has orthologs in multicellular species. Likewise the *C. elegans* N-glycoproteome is higher conserved in nematodes compared to the whole proteome, but shows relatively low conservation in other eukaryotic species. The *A.thaliana*, *D. melanogaster*, zebrafish and mouse data proved that the same rule held true for the N-glycoproteomes of plants, insects and of vertebrates ($p < 10^{-100}$ for all species).

To check if the observed phenomenon is secondary to the extracellular localization of glycoproteins, we compared the conservation degrees of extracellular N-glycosylated proteins with extracellular non-N-glycosylated proteins that are known to be poorly conserved. We found that N-glycosylated and non-N-glycosylated proteins are similarly week

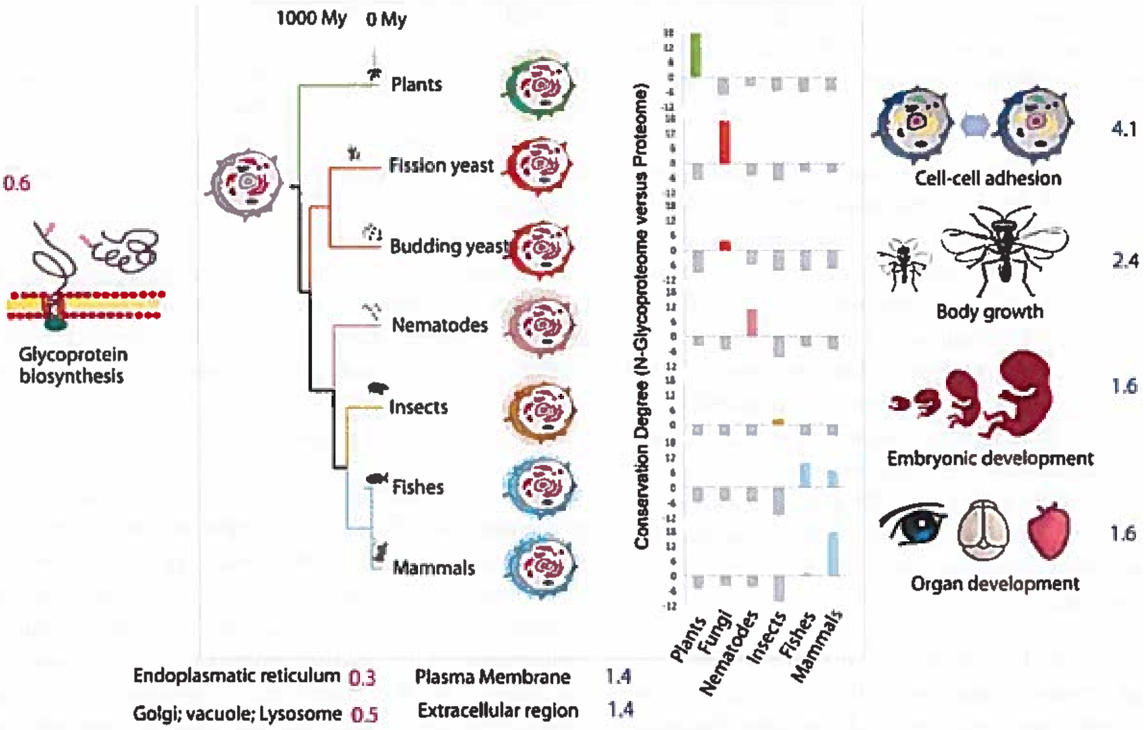


Figure 2: Evolution of N-glycoproteomes. For each analyzed organism the proportion of N-glycoproteins with orthologs in other species was calculated. In addition the proportion of the whole proteome with orthologs was derived. For each phylogenetic group (plants, fungi, nematodes, insects, fishes, and mammals) the average of these proportions was calculated and compared between the N-glycoproteome and the whole proteome. Positive conservation degree values signify that the N-glycoproteome has more orthologs compared to the whole proteome. The *A. thaliana* N-glycoproteome is significantly conserved within plants (green), the *S. cerevisiae* and *S. pombe* N-glycoproteomes in fungi (orange), the *C. elegans* N-glycoproteome in nematodes (pink), and the *D. melanogaster* N-glycoproteome in insects (yellow). The zebra fish and the mouse N-glycoproteomes have significantly more orthologs than the corresponding proteomes in vertebrates (blue). Negative N-glycoproteome conservation degree values indicate a lower proportion of N-glycoproteins with orthologs and are represented in gray. These observations show that a large proportion of N-glycoproteins evolved after the respective phylogenetic divergence and are therefore specific to the corresponding phyla. Phylogenetic group specific N-glycoproteins are overrepresented in the plasma membrane and in the extracellular region compared to non-specific N-glycoproteins that have at least one ortholog in another group. These proteins predominantly play a role in cell-cell adhesion, body growth, embryonic development and organ development. Non-specific proteins are overrepresented in the endoplasmatic reticulum, in golgi, vacuole and lysosomes and are mainly involved in glycoprotein biosynthesis. The median enrichment ratio of group specific to non specific N-glycoproteins is given in blue for overrepresentations and in red for underrepresentations.

conserved across phyla, however extracellular N-glycosylated proteins are significantly higher conserved within the same phylogenetic group. This suggests that the majority of glycoproteins developed after the species divergence and remained conserved maintaining its pivotal role in the phyla.

The high proportion of phylogenetically specific N-glycoproteins in each species (Supplementary Fig. 5) suggests the co-evolution of the N-glycoproteome with the adaption of complex eukaryotic species to their present-day ecological niches. The functional roles of phylogenetically specific N-glycoproteins are heavily overrepresented in cell-cell communication processes that are essential for the development and growth of organisms and accordingly their cellular localizations are generally on the outer region of the cell ($p < 10^{-6}$; Fig. 2). The evolutionarily conserved N-glycoproteins are instead overrepresented in the intracellular compartments involved in the attachment and processing of glycans ($p < 10^{-4}$; Fig. 2).

In contrast to N-glycosylated proteins, the vast majority of phosphorylated proteins is located in the cytoplasm and is often involved in intracellular signal transduction. Phosphoproteins already occurred in the earliest life forms, whereas their regulation via phosphorylation evolved much later[29]. A similar pattern holds for lysine acetylated proteins[30]. Therefore, the evolution of the N-glycoproteome with its high interspecies diversity presents an outlier among post-translational modifications. Our analysis suggests that the mechanism responsible for this outlier status is that the core N-glycosylation machinery decorates proteins with N-glycosylation in strict concordance with the sequence motifs and topological locations of the substrates. Novel N-glycosylated proteins are easily evolved by providing them with sequence motifs that direct them along the ER/Golgi route and that ensure N-glycosylation by the core machinery. As discussed above, common cellular functions tend to take place in intracellular compartments and cell type specific functions in extracellular compartments, which therefore provide a rational for the preferential non-conservation of the N-glycoproteome.

In summary, N-glycosylation occurs on many proteins of pivotal importance in development, multicellular communication and many other basic biological functions. Besides shedding light on the evolution of the N-glycoproteome, our results should be useful for biotechnological applications, for example those that involve the production of human glycoproteins in insect cells, and in general for the study of human disease proteins in model organisms. The modification sites determined here have been uploaded to the Posttranslational Modification Database PHOSIDA[31] and can now be used by the community for detailed functional biological as well as bioinformatic studies.

METHODS SUMMARY

A. thaliana, S. pombe, S.cerevisiae, C.elegans, D.melagonaster and *D.rerio* were lysed and processed according to the N-glyco-FASP protocol[8]. For details see Supplementary Information. Briefly, lysis buffer containing 0.1 M DTT, 4% SDS in 0.1 M Tris/HCl pH 7.6 was added to the frozen organism samples, followed by homogenization, sonication and subsequent boiling. Clarified supernatants containing approximately 400 µg protein were transferred to 30 kDa filtration units (Millipore). Filters were rinsed three times with 8 M urea followed by 20 min incubation with iodoacetamide. Subsequently, three additional washes with 8 M urea and three washes with 40 mM ammonium bicarbonate were performed. Finally, proteins were digested using 4 µg trypsin or 20 µg GluC at 37°C or 25°C. After 12 h incubation, peptides were eluted in lectin binding buffer (1 mM CaCl2, 1 mM MnCl2, 0.5 M NaCl in 20 mM TrisHCl, pH 7.3) and transferred to a new 30 kDa filtration unit. Lectin mixture containing Con A, WGA and RCA_{120} in a mass proportion to peptides of 2:1 was added to the top of the filters. Not bound peptides were eluted. Captured peptides were washed with the binding buffer prior to PNGase F addition. Deglycosylation was performed in $H_2{}^{18}O$ at 37°C. Deglycosylated peptides were measured on a linear ion trap Orbitrap mass spectrometer (LTQ-Orbitrap Velos)[10]. MS data were analyzed using MaxQuant[32]. Slightly modified protocol was applied to plant material. General characteristics of the N-glycoproteomes were determined on the basis of common bioinformatic tools as described in Supplementary Information. Phylogenetic relationships were derived from the Ensembl Compara database[33]. In the case of *S. pombe* and *A. thaliana* homologs were derived using BLAST[34]. Global sequence alignments between orthologous proteins were generated using the Needle software[35].

References

1 Varki, A. *et al. Essentials of Glycobiology*. (New York: Cold Spring Harbor Laboratory Press, 2008).

2 Silberstein, S. & Gilmore, R. Biochemistry, molecular biology, and genetics of the oligosaccharyltransferase. *FASEB J* **10**, 849-858 (1996).

3 Knauer, R. & Lehle, L. The oligosaccharyltransferase complex from yeast. *Biochim Biophys Acta* **1426**, 259-273 (1999).

4 Kaji, H. *et al.* Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol* **21**, 667-672 (2003).

5 Bunkenborg, J., Pilch, B. J., Podtelejnikov, A. V. & Wisniewski, J. R. Screening for N-glycosylated proteins by liquid chromatography mass spectrometry. *Proteomics* **4**, 454-465 (2004).

6 Liu, T. *et al.* Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *J Proteome Res* **4**, 2070-2080 (2005).

7 Wollscheid, B. *et al.* Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol* **27**, 378-386 (2009).

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

9 Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359-362 (2009).

10 Olsen, J. V. *et al.* A dual pressure linear ion trap - Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics* (2009).

11 Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).

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

13 Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **34**, D187-191 (2006).

14 Upreti, R. K., Kumar, M. & Shankar, V. Bacterial glycoproteins: functions, biosynthesis and applications. *Proteomics* **3**, 363-379 (2003).

15 Wacker, M. *et al.* N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. *Science* **298**, 1790-1793 (2002).

16 Scott, N. E. *et al.* Mass spectrometric characterization of the surface-associated 42 kDa lipoprotein JlpA as a glycosylated antigen in strains of Campylobacter jejuni. *J Proteome Res* **8**, 4654-4664 (2009).

17 Liu, X. *et al.* Mass spectrometry-based glycomics strategy for exploring N-linked glycosylation in eukaryotes and bacteria. *Anal Chem* **78**, 6081-6087 (2006).

18 Young, N. M. *et al.* Structure of the N-linked glycan present on multiple glycoproteins in the Gram-negative bacterium, Campylobacter jejuni. *J Biol Chem* **277** (2002).

19 Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-3449 (2005).

20 Kung, L. A. *et al.* Global analysis of the glycoproteome in Saccharomyces cerevisiae reveals new roles for protein glycosylation in eukaryotes. *Mol Syst Biol* **5**, 308 (2009).

21 Bloor, J. W. & Brown, N. H. Genetic analysis of the Drosophila alphaPS2 integrin subunit reveals discrete adhesive, morphogenetic and sarcomeric functions. *Genetics* **148**, 1127-1142 (1998).

22 Tweedie, S. *et al.* FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37**, D555-559 (2009).

23 Rieger, S., Senghaas, N., Walch, A. & Koster, R. W. Cadherin-2 controls directional chain migration of cerebellar granule neurons. *PLoS Biol* **7**, e1000240 (2009).

24 Moreira, E. S. *et al.* A first missense mutation in the delta sarcoglycan gene associated with a severe phenotype and frequency of limb-girdle muscular dystrophy type 2F (LGMD2F) in Brazilian sarcoglycanopathies. *J Med Genet* **35**, 951-953 (1998).

25 Malenka, R. C. & Malinow, R. Alzheimer's disease: Recollection of lost memories. *Nature* **469**, 44-45 (2011).

26 Soll, J. Protein import into chloroplasts. *Curr Opin Plant Biol* **5**, 529-535 (2002).

27 Jarvis, P. & Soll, J. Toc, Tic, and chloroplast protein import. *Biochim Biophys Acta* **1541**, 64-79 (2001).

28 Villarejo, A. *et al.* Evidence for a protein transported through the secretory pathway en route to the higher plant chloroplast. *Nat Cell Biol* **7**, 1224-1231 (2005).

29 Gnad, F. *et al.* Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. *Mol Cell Proteomics* **9**, 2642-2653.

30 Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**, 834-840 (2009).

31 Gnad, F. *et al.* PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* **8**, R250 (2007).

32 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).

33 Flicek, P. *et al.* Ensembl's 10th year. *Nucleic Acids Res* **38**, D557-562.

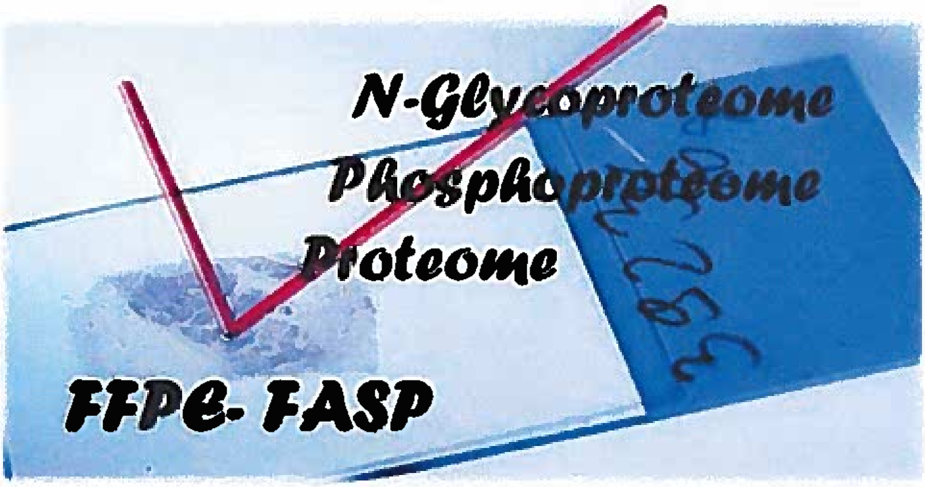
34 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

35 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).

Publication III

Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry

Ostasiewicz P, Zielinska DF, Mann M, Wiśniewski JR



J Proteome Res, 2010

This publication presents the newly developed 'FFPE-FASP' method to quantitatively analyse proteomes, phosphoproteomes and glycoproteomes from formalin fixed and paraffin embedded material.

research articles
Journal of proteome research

Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry

Paweł Ostasiewicz,[†,‡] Dorota F. Zielinska,[†] Matthias Mann,[*,†] and Jacek R. Wiśniewski[*,†]

Department for Proteomics and Signal Transduction at Max-Planck Institute for Biochemistry, 82152 Martinsried, Germany, and Department of Pathology at Wroclaw Medical University, 50368 Wroclaw, Poland

Received March 15, 2010

Tissue samples in biobanks are typically formalin-fixed and paraffin-embedded (FFPE), in which form they are preserved for decades. It has only recently been shown that proteins in FFPE tissues can be identified by mass spectrometry-based proteomics but analysis of post-translational modifications is thought to be difficult or impossible. The filter aided sample preparation (FASP) method can analyze proteomic samples solubilized in high concentrations of SDS and we use this feature to develop a simple protocol for FFPE analysis. Combination with simple pipet-tip based peptide fractionation identified about 5000 mouse liver proteins in 24 h measurement time—the same as in fresh tissue. Results from the FFPE-FASP procedure do not indicate any discernible changes due to storage time, hematoxylin staining or laser capture microdissection. We compared fresh against FFPE tissue using the SILAC mouse and found no significant qualitative or quantitative differences between these samples either at the protein or the peptide level. Application of our FFPE-FASP protocol to phosphorylation and N-glycosylation pinpointed nearly 5000 phosphosites and 1500 N-glycosylation sites. Analysis of FFPE tissue of the SILAC mouse revealed that these post-translational modifications were quantitatively preserved. Thus, FFPE biobank material can be analyzed by quantitative proteomics at the level of proteins and post-translational modifications.

Keywords: FFPE • FASP • post-translational modification • phosphorylation • N-glycosylation • Orbitrap Velos • high resolution mass spectrometry

Introduction

Formalin-fixation and paraffin-embedding (FFPE) is a routine procedure for the preservation of human specimens for histopathological analyses. Since FFPE samples can be stored over long periods, many pathology laboratories possess vast banks of human samples with associated patient, treatment, and outcome information. These could in principle be used for identification of disease biomarkers and drug targets. Currently, such investigations rely on immunohistochemistry (IHC) with paraffin compatible antibodies, but this method is laborious, is not truly quantitative,[1] and is only applied to already known or suspected markers.

It would be attractive to complement IHC with the powerful methods of high resolution, mass spectrometry-based proteomics.[2−4] However, proteins in FFPE are cross-linked and have therefore long been thought to be unsuitable to MS-based proteomics. Indeed, the identification of a few hundred or thousands of proteins in FFPE has only been reported during the past few years.[5−11]

Several of these studies compared protein identifications from fresh or frozen tissue to FFPE tissue and generally agreed that similar numbers could be identified in both cases. However, there was less agreement on the efficiency of protein retrieval from FFPE compared to fresh tissue, discrimination against certain amino acids—especially lysines—or the effect of the fixation duration. Almost all previous studies on FFPE tissue have been performed by low-resolution mass spectrometry and without quantification by stable isotope labeling, which may explain some of these disagreements. Taken together with the fact that reported numbers of identifications are lower than those that can be obtained with fresh tissues with optimized sample preparation and high resolution mass spectrometry,[12,13] it is still not clear if proteome analysis from FFPE tissue and fresh tissue are qualitatively and quantitatively equivalent.

Furthermore, the fundamental question whether it is possible to analyze the phosphoproteome after FFPE has remained unanswered. In particular, it is not known if phosphorylation sites are preserved in FFPE samples and, if so, whether they can be measured by modern mass spectrometric methods. This is a very important issue because large-scale analysis of PTMs such as phosphorylation is one of the most important and

---

[*] To whom correspondence should be addressed. Matthias Mann and Jacek R. Wiśniewski, Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18 D-82152 Martinsried, Germany. E-mail: mmann@biochem.mpg.de or jwisniew@biochem.mpg.de. Fax: +49 89 8578 2219.

[†] Max-Planck Institute for Biochemistry.

[‡] Wroclaw Medical University.

10.1021/pr100234w

© 2010 American Chemical Society

Published on Web 05/18/2010

active areas of proteomics. It is conceivable that the quantitation of the phosphoproteome may yield as much diagnostic information for cancer and other diseases as quantitation of protein expression levels.

It has been known since the early '90s that application of heat during sample preparation (termed AR method for "antigen retrieval") is advantageous in unmasking epitopes for IHC, presumably by reversing cross-links.[14] Furthermore, sodium dodecyl sulfate (SDS) based extraction has generally been found to be advantageous for FFPE,[6] but this detergent was not compatible with downstream MS analysis. We have recently described the Filter Aided Sample Preparation Method (FASP), which combines advantages of gel-based and gel-free proteomic workflows and utilizes high concentration of SDS while still producing pure peptides for MS analysis.[13] We reasoned that this method might be particularly well suited to the analysis of FFPE and developed a simple protocol that combines boiling in SDS with a FASP-based proteomic workflow.

The recently described SILAC-mouse[15] provides a fully labeled control tissue proteome to quantitatively compare results from fresh and FFPE tissues. We show that the FFPE-FASP approach allows unbiased and quantitative investigation of fixed tissue. Importantly, it is also suitable for comparative mapping of thousands of phosphorylation sites. Our results demonstrate that the phosphoproteome is preserved in FFPE samples and that it is analyzable by high resolution MS-based proteomics.

Experimental Procedures

SILAC Mouse. For SILAC experiments labeled mice (C57BL/6) were used.[15] Briefly, mice were raised on a protein free diet for several generations, which, in addition to the other amino acids, contained "heavy" L-$^{13}C_6$-lysine (Silantes GmbH, Munich, Germany). We chose lysine because it is an essential amino acid and is not converted to other amino acids. SILAC-mice are morphologically and behaviorally normal.[15] In the experiments reported here we used liver tissue from a third generation (F3) SILAC-mouse, ensuring complete labeling.

Tissue Dissection and Fixation. Liver samples were obtained from 5-months old male mice (C57BL/6), euthanized by chloroform. From each mouse half of the liver was immediately immersed in 10 mL of 10% buffered formalin solution and fixed overnight (unless indicated otherwise) whereas the second half was subjected to immediate protein extraction or protein extraction after freezing in liquid nitrogen (see below). To embed a fixed sample in paraffin, samples were equilibrated in turn in a 50% (v/v) solution of ethanol in 10% formalin (1 h), then 70% ethanol (30 min), 95% ethanol (2 × 40 min), absolute ethanol (2 × 40 min), xylene (2 × 40 min), and paraffin (3 × 30 min).

Microdissection. For microdissection, FFPE-samples were sliced in a microtome (5 μm sections), mounted on standard histological slides (Menzel Glaser, Braunschweig, Germany), deparaffinized with xylene, and hydrated via an ethanol/water series. Sections were stained with Mayer's hematoxylin for 20 s and again dehydrated. Tissue (40 mm^2) was dissected either with a needle or with Laser Pressure Catapulting (LPC) using a PALM Instrument (Zeiss, Göttingen, Germany). Liver areas subjected to laser microdissection were marked using a 10× objective and dissected in Auto-LPC mode. Dissected tissue was collected in Adhesive Caps (Zeiss, Göttingen, Germany) and then suspended in the lysis buffer.

Protein Extraction from Fresh Tissue. Liver tissue was briefly washed in PBS solution and homogenized in 10 mL lysis buffer (0.1 M Tris-HCl, pH 8.0, 0.1 M DTT using a blender (T 10 basic Ultra, IKA, Staufen, Germany). The homogenate was sonicated using Branson Sonifier 250 (Heinemann, Schwäbisch Gmünd, Germany) (output control 5; duty cycle 20%) on ice for 3 min. SDS was then added to the suspension to a final concentration of 4% and the mixture was incubated for 3 min at 95 °C. The crude extract was then clarified by centrifugation at 16 000× *g* at 18 °C for 10 min.

Protein Extraction from FFPE. Paraffin was removed from tissue slices by successive incubations in xylene (2×) and absolute ethanol (2×) and then the samples were rehydrated in series of 97, 80, 70, and 50% ethanol solution in water. Following vacuum-drying, the tissue was homogenized in 0.1 M Tris-HCl, pH 8.0, 0.1 M DTT in a tissue to buffer ratio of 1:20 using the blender and sonifier as described above. After addition of SDS to a final concentration of 4% the homogenate was incubated for 60 min at 99 °C in a heating block with agitation (600 rpm). The crude extract was then clarified by centrifugation at 16 000× *g* at 18 °C for 10 min. A detailed protocol for performing FFPE-lysis is described in Supplementary Data (FFPE-FASP-Protocol).

Protein Digestion by FASP. Detergent was depleted from the lysates and the proteins were digested with trypsin (qualitative experiments) or LysC (SILAC experiments) using the FASP protocol.[13] Briefly, to YM-30 filter units (Millipore, Carrigtwohill, Ireland) containing protein concentrates, 200 μL of 8 M urea in 0.1 M Tris/HCl, pH 8.5 (UA), was added and the samples were centrifuged at 14 000× *g* at 20 °C for 15 min. This step was repeated once. Then 50 μL of 0.05 M iodoacetamide in 8 M urea was added to the filters and the samples were incubated in darkness for 20 min. Filters were washed twice with 100 μL of 8 M UA followed by two washes with 100 μL of 40 mM NH_4HCO_3. Finally, trypsin (Promega, Madison, WI) or LysC (Wako, Richmond, VA) were added in 40 μL of 40 mM NH_4HCO_3 to each filter. The protein to enzyme ratio was 100:1 for trypsin and 50:1 for LysC. The samples were incubated overnight at 37 °C and peptides were collected by centrifugation.

Total Protein Determination. Protein content was determined using Cary Eclipse Fluorescence Spectrometer (Varian, Palo Alto, CA) as described previously.[16] Briefly, 1−2 μL of sample or tryptophan standard was mixed with 2 mL of 8 M urea in 10 mM Tris-HCl pH 8.0. Fluorescence was measured at 295 nm for excitation and 350 nm for emission. The slits were set to 10 nm.

Peptide Fractionation. Peptides were fractionated according to a StageTip[17]-based fractionation protocol (SAX).[12] Briefly, peptides were loaded into tip-columns made by stacking six layers of a 3 M Empore Anion Exchange disk (1214−5012 Varian, Palo Alto, USA) into a 200 μL micropipet tip. For column equilibration and elution of fractions, we used Britton & Robinson universal buffer (BRUB) composed of 20 mM acetic acid, 20 mM phosphoric acid, and 20 mM boric acid titrated with NaOH to the desired pH. Peptides were loaded at pH 11 and fractions were subsequently eluted with buffer solutions of pH 8, 6, 5, 4, and 3. Phosphopeptides were separated using a four-step elution with buffer of pH 6, 4.5, 3.5, and 3.5 containing 0.4 M NaCl.

Phosphopeptide Enrichment. Peptide solutions were acidified with CF_3COOH and phosphopeptides were enriched on TiO_2 beads[18,19] according to a previously described protocol.[20] Twenty-five milligrams of 10 μm titansphere TiO_2 (GL Sciences, Tokyo, Japan) were suspended in 50 μL of 3% (m/v) 2,5-dihydroxy benzoic acid (DHB), 80% (v/v) CH_3CN, 0.1% (v/v) CF_3COOH and were diluted with 200 μL of water before use. Ten microliters of the TiO_2 slurry were added to the peptide solution and incubated under continuous agitation for 20 min. The titanium beads were then sedimented by centrifugation at 5000× *g* for 1 min and the supernatants were collected and mixed with another portion of the beads and incubated as above. The pellets were resuspended in 150 μL of 30% (v/v) CH_3CN containing 3% (v/v) CF_3COOH and were transferred to a 200 μL pipet tip containing one layer of Empore-C8 filter (Millipore, Carrigtwohill, Ireland). The beads were washed 3× with 30% (v/v) CH_3CN, 3% (v/v) CF_3COOH (v/v) solution and 3× with 80% (v/v) CH_3CN, 0.3% (v/v) CF_3COOH solution. Finally, peptides were eluted from the beads with 100 μL of 40% (v/v) CH_3CN and 15% (m/v) NH_4OH and concentrated in a speed-vac to 3 μL. Phosphopeptides were diluted with 20 mM BRUB pH 6 and fractionated according to the SAX protocol described above.

N-Glycopeptide Enrichment. N-glycopeptides were enriched according to the N-Glyco-FASP procedure.[21] Briefly, 100 μg peptides were incubated with a mixture of ConA, WGA and RCA120 (Sigma, St. Louis, MO) at a peptide to lectin ratio 2:1. The mixtures were transferred to new YM-30 filter units and after 1 h incubation at 25 °C, the samples were centrifuged at 14 000× *g* at 18 °C for 10 min and the elutates were discarded. The filters were rinsed with 200 μL binding buffer and then with 50 μL 40 mM NH_4HCO_3 in $H_2{}^{18}O$ water. Peptides were deglycosylated with 2 units of PNGase F (Roche, Basel, Switzerland) in 40 mM NH_4HCO_3 in $H_2{}^{18}O$ at 37 °C for 3 h, which marks former sites of N-glycosylation by a 2.9890 Da mass difference.[22] The deglycosylated peptides were eluted by centrifugation.

LC–MS/MS Analysis. LC was performed on a Proxeon Easy-nLC System (Proxeon Biosystems, Odense, Denmark). Peptides were separated on a 15 cm fused silica emitter (Proxeon Biosystems, Odense, Denmark) packed in-house with the reverse phase material ReproSil-Pur C_{18}-AQ, 3 μm resin (Dr. Maisch, Ammerbuch-Entringen, Germany) with a 230 min gradient from 2 to 80% of 80% (v/v) CH_3CN, 0.5% (v/v) acetic acid.

For investigating the effects of storage and processing on FFPE samples, peptide mixtures were analyzed on an LTQ-Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany). The recently introduced LTQ-Orbitrap Velos mass spectrometer[23] was used for proteome and PTM analysis. The lock-mass at *m*/*z* 445.120025 provided internal mass calibration.[24] In the LTQ-Orbitrap XL full mass range scans at a resolution of 60 000 were acquired while up to 10 MS/MS spectra were acquired at low resolution in the linear ion trap ("top10" method). On the LTQ-Orbitrap Velos, MS scans were combined either with tandem mass spectrometry by CID in the linear ion trap or in a "high–high" strategy with Higher Energy collisional Dissociation (HCD)[25] with high accuracy analysis of the fragment ions in the orbitrap analyzer. For HCD analysis, transients of 100 ms were acquired, corresponding to a resolution of 7500. Because HCD does not operate in parallel with the acquisition of the MS spectrum, transients were limited to 0.5 s for the MS spectra in the Velos instrument (30 000 resolution). Because of the high sequencing speed of the Velos instrument, we employed a "top20" method for CID (target value for MS/MS 5000) and a "top10" method for HCD (target values for MS/MS 50 000).

Data Analysis. The MS data were analyzed using the software environment MaxQuant,[26,27] version 1.0.12.36. Proteins were identified by searching MS and MS/MS data of peptides against a decoy version of the International Protein Index (IPI) database for Mouse (v. 3.46). Carbamidomethylation of cysteines was set as fixed modification and oxidation of methionines, acetylation of the protein N-terminus and phosphorylation of serines, threonines, and tyrosines were variable modifications. For non-SILAC experiments trypsin was used and accordingly trypsin allowing for cleavage N-terminal to proline was chosen as enzyme specificity. For SILAC experiments only lysine was labeled, LysC was used for protein digestion and LysC was chosen as enzyme specificity. A maximum of two missed cleavages were allowed and only fully tryptic or LysC peptides, respectively, were considered. The minimum peptide length was specified to be 6 amino acids. The initial maximal mass tolerance in MS mode was set to 7 ppm, whereas MS/MS tolerance was set to 0.5 Da for CID data and 0.02 Da for HCD data. The maximum peptide and site false discovery rates were specified as 0.01. The PTM score was used for assignment of the phosphorylation sites as described.[19] The PTM localization score reflects the normalized probability that the phosphorylation is indeed localized at the specified amino acid position. It is calculated for all combinatorial phosphorylation site possibilities from the overlap of assigned *b* and *y* ions and observed matches. The algorithm is integrated into MaxQuant. The "class I phosphorylation sites" discussed in the text are defined by a localization probability of at least 0.75. To adjust for unequal peptide loading, the phosphopeptide and glycopeptide SILAC histograms were centered at zero on the $\log_2$ scale.

Statistical Analysis. We used Student test for testing differences between the means of two populations. For small sample sizes, two tailed hypothesis test at the 0.05 level of significance was applied.[28] Significance of outliers from the quantitative distributions was corrected for multiple hypothesis testing by the Benjamini-Hochberg method.

Results

FASP-based FFPE Sample Analysis Protocol. In the development of the method for filter aided sample preparation for formalin-fixed paraffin embedded tissues (FFPE-FASP), we started from the success of the "Antigen Retrieval" (AR) method in immunohistochemistry, whose key point is heating FFPE material in water or buffer solution. Furthermore, sodium dodecyl sulfate (SDS) is a powerful and near universal solubilization agent in proteomics and we therefore sought to include it in the protocol. Previous work recognized the desirability of employing SDS,[29] but due to its incompatibility with subsequent MS analysis, researchers have explicitly striven to develop protocols that substituted other denaturants and solubization agents (see for example ref 7). In FASP, however, use of high concentrations of SDS is the defining feature as the method is based on the observation that SDS can be quantitatively exchanged by urea on the matrix of a filter unit while still leaving proteins accessible to a proteolytic enzyme.[13] We therefore combined a procedure of boiling FFPE sample in SDS buffer with subsequent loading onto 30k filter units and digestion by trypsin or LysC. Note that these filters still retain very small proteins because these are completely unfolded under the denaturing conditions used.[13] Figure 1 shows an overview of the FFPE-FASP procedure, including downstream proteomic analysis.

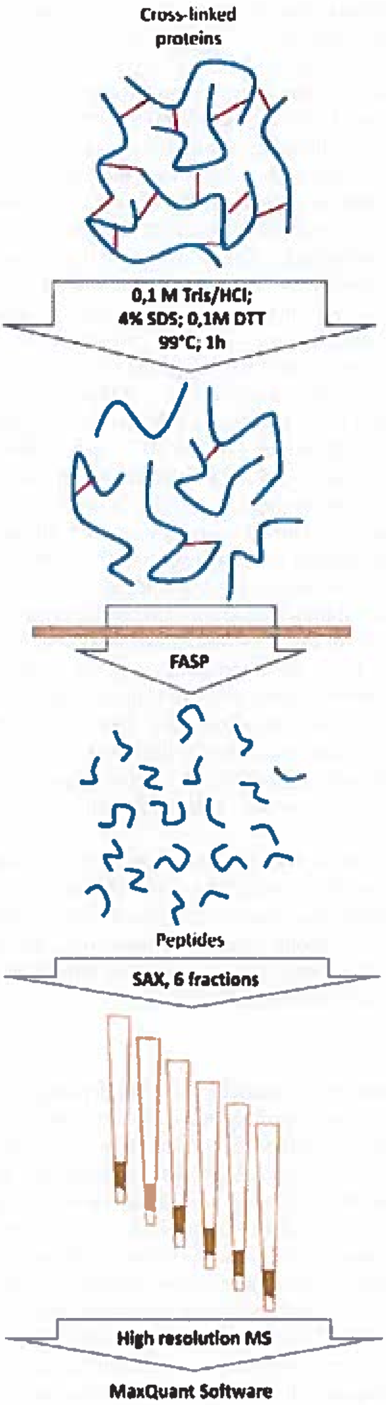


Figure 1. Schematic overview of the FFPE-FASP method. Formalin-fixed and paraffin embedded samples are deparaffinized and heated in SDS-containing lysis buffer. The resulting protein mixture is applied to the top of a spin filter where SDS is completely exchanged with urea, followed by on-filter proteolytic digestion (FASP). Peptides pass through the filter into another pipet containing strong anion exchange filter material (SAX). Six fractions are eluted by pH steps and analyzed by high resolution LC−MS/MS on a linear ion trap orbitrap instrument.

FASP allows straightforward determination of the yield of the method by measuring UV absorbance of total eluted peptide. We used this feature to optimize lysis and digestion conditions. Best yields were achieved when the deparaffinized tissues was homogenized in a SDS and DTT-containing buffer and then lysed at 99 °C for 1 h. To compare the effectiveness of protein extraction from fresh and fixed samples, we divided fresh liver into eight parts and weighed them. Four were subjected to immediate extraction whereas the other four were fixed and embedded in paraffin before extraction. From 1 mg of fresh tissue we obtained 140 ± 18 μg total protein and from the same amount of FFEP tissue we extracted 155 ± 16 μg (Figure 2A). Thus protein extraction was very high in both cases (given an estimated protein content of around 10% by weight in tissues) and not statistically different from each other.

We next determined the digestion efficiency of FFPE tissue boiled in SDS and processed by FASP. We found that 60% of protein amount was converted to pure peptides ready for MS analysis. This value was even somewhat higher than the corresponding value for fresh tissue (Figure 2B); however, we assume this is due to experimental variability. We visualized the proteome extracted from FFPE and fresh liver by Coomassie staining on a 1D gel (Figure 2C). The protein gels appear similar except for a somewhat increased staining in the high molecular weight region of the FFPE lane, which we attribute to proteins that remain cross-linked to another protein at least at one position. Note that this appears to affect a small percentage of the proteome and that a small number of cross-links would not necessarily influence overall protein extraction efficiency or digestion yield. A step-by-step protocol for FFPE-FASP is provided as Supporting Information.

Qualitative Comparison of the FFPE and Fresh Liver Proteome. To compare the numbers of identifiable proteins, we first performed a pipet-based strong anion exchange fractionation step (FASP-SAX)[12] on the peptides retrieved by FASP. Six fractions were measured by LC−MS/MS on the LTQ-Orbitrap Velos with 230 min gradients and the results were analyzed in the MaxQuant environment,[26] specifying a 1% False Discovery Rate (FDR). For each of the two samples, we performed a technical replicate. Together this resulted in the identification of 5203 proteins in the FFPE sample and 5426 proteins in the fresh sample. Proteins were identified with seven nonredundant peptides on average. The average absolute mass deviation between calculated and measured peptide masses was below 1 ppm for these and the experiments described below (Supplemental Table 1, Supporting Information).

Between fixed and fresh tissue 91% of the identifications were identical (Figure 2D). We suspected that this level of overlap in identified proteins was as high as it would be between identical samples and that it only reflected stochastic aspects of which peptides are chosen for MS/MS in different LC runs. Indeed, when we compared the proteins identified in the technical replicates of the fresh sample, the overlap was 85%. (The slightly lower overlap than that of fresh and fixed tissue is presumably due to the smaller depth of coverage of the proteome achieved in two single experiments.) Furthermore, when we additionally matched sequenced peptides in one condition to unsequenced peptides in the other condition ("match between runs" feature in MaxQuant), we obtained a protein identification overlap between fresh and fixed tissue samples of 96%. These results are compatible with an assumption that FFPE tissue can be analyzed as readily as fresh tissue by MS-based proteomics.

The chemistry of formalin fixation is not fully understood but it is believed that the primary amines of lysine side chains

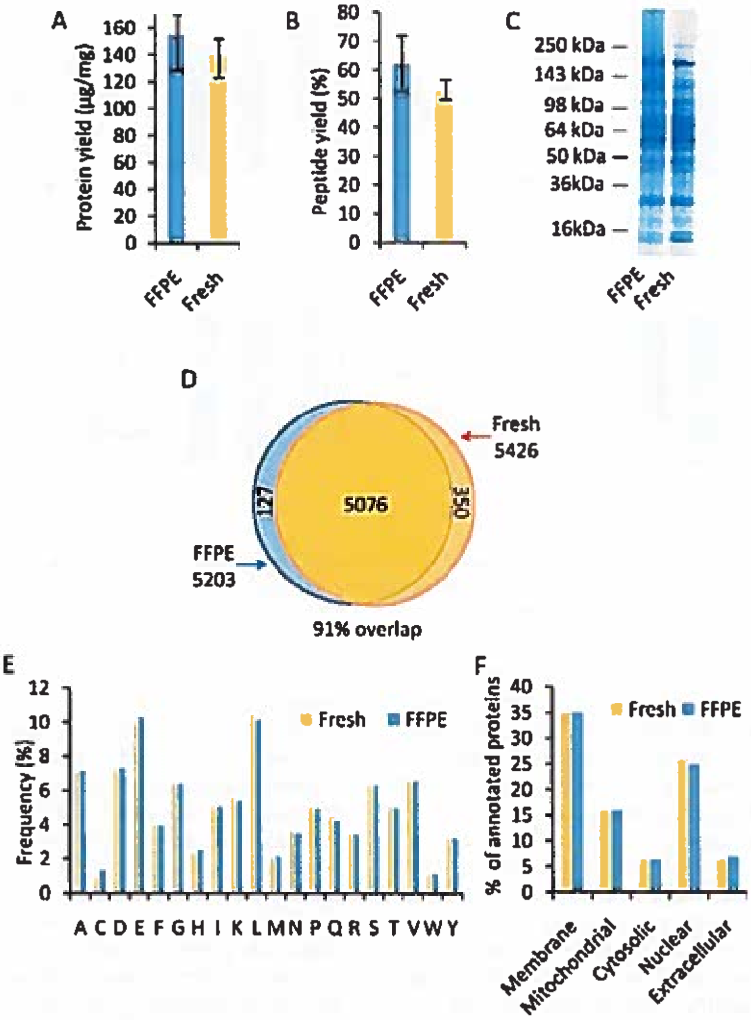


Figure 2. Comparison of FFPE-FASP and FASP protocols using FFPE and fresh tissue samples. Yields of (A) proteins and (B) peptides obtained from FFPE and fresh and samples. Error bars represent standard deviation ($n = 4$). (C) Protein extracts from 100 µg wet tissue that was either FFPE treated or fresh were separated by SDS-PAGE and Coomassie stained. (D) Overlap of proteins identified from FFPE and fresh samples. (E) Frequencies of amino acid residues in identified peptides. (F) Subcellular distribution of identified proteins using GeneOntology annotations.

are mainly involved in protein−protein cross-links and that several residues like arginine, cysteine, serine and threonine may also contribute.[30] Since the reversal of these fixation-induced cross-links may only be partial and some residues may remain modified, peptides containing them could be underrepresented in the proteomic data sets. Similar to previous studies, we therefore analyzed the frequency distribution of all residues in the presence and absence of formalin fixation, but in our data we could not detect any such preferences or biases (Figure 2E). Furthermore, because trypsin should not cleave at modified lysine residues, the number of tryptic peptides with a C-terminal lysine might be a sensitive indicator of extensive lysine modification.[5] However, we found that 59% of the identified peptides had a C-terminal lysine in both the fresh and the FFPE tissue, compatible with the absence of large-scale lysine modification (Supplemental Table 1, Supporting Information).

Next, we compared the subcellular distribution of proteins identified in both types of samples and again did not observe significant differences (Figure 2F). Notably, the percentage of membrane proteins is high and is close to the value predicted from genetic data[5] demonstrating that FFPE-FASP does not discriminate against this class of proteins.

So far, our qualitative analysis had indicated no differences between the analysis of fresh or fixed material by FFPE-FASP. In particular, large scale proteomic analysis was equally possible from fresh material and material from paraffin blocks.

Effect of Storage and Processing FFPE Samples. Since the major advantage of using FFPE material is the possibility to analyze and compare archival samples, possible effects of storage time and sample processing need to be evaluated for each new sample preparation procedure. We studied four key aspects of FFPE tissue analysis: the storage time, the length of fixation and a possible effect on sample harvesting of the hematoxylin staining procedure or of laser microdissection. For each of these we compared the results of the analysis of two mouse liver samples, prepared identically except for the parameter to be evaluated. After processing by the FFPE-FASP procedure, each liver sample was measured in duplicate by single 230 min LC−MS/MS runs on an LTQ-Orbitrap XL. In

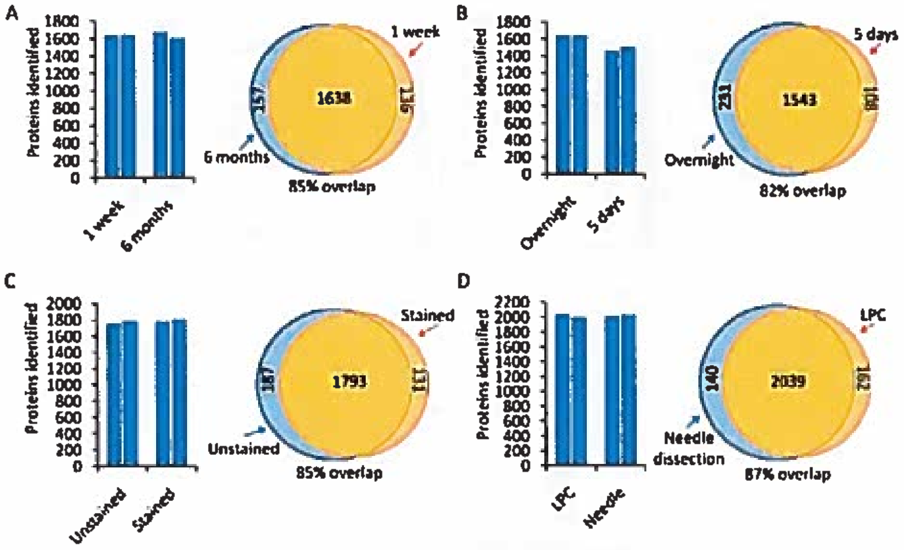


Figure 3. Effects of key processing steps in FFPE analysis. Liver samples were analyzed in duplicate by single LC–MS/MS runs and each run was analyzed separately (bars) or together (Venn diagrams). (A) Effects of storage period, (B) fixation time, (C) Hematoxilin staining, and (D) laser vs needle scraping were analyzed.

MaxQuant, data files were analyzed separately to determine the number of proteins identified in each of the single runs as well as together to determine the overlap of the data sets. This enabled separate assessment of technical reproducibility and the influence of the experimental variability under study (Supplemental Table 2, Supporting Information).

To address the effects of storage time, we stored a fixed and paraffin embedded mouse liver for six months before analysis whereas another liver was analyzed after one week after embedding. In duplicate analysis, we indentified 1774 proteins in the sample stored for one week and 1795 in the sample stored for six months (Figure 3A). Of these proteins, 1638 (85%) were identified in both samples, which is good agreement at this limited depth of analysis.

Since fixation time is not standardized in the preparation of FFPE tissue archives but usually takes between 24 and 48 h, we checked if prolonged fixation would affect proteomics results. For this purpose we compared two fragments of the same liver. The first sample was fixed overnight and the second one for five days. In duplicate analysis we identified 1774 proteins in the sample that was fixed overnight and 1651 in the sample fixed for five days (Figure 3B). Eighty-two percent of the proteins were identified in both samples. These results showed that proteomic analysis was still possible after prolonged fixation time. The difference in the numbers of proteins identified in both samples is relatively low (7%), but since it is reproducible in the two duplicate analyses, we cannot exclude that it is significant.

Staining of FFPE tissue is necessary to visualize cellular and tissue morphology. The hematoxylin or hematoxylin and eosin (H&E) stain is the standard staining method used routinely in pathology laboratories. To test whether hematoxylin staining has any effect on the number of identified proteins in mouse tissue we analyzed adjacent liver sections that were stained or unstained. In duplicate analysis of the unstained sections we identified 1981 proteins and in duplicate analysis of the stained sections we identified 1924. In addition to this very similar total number of identifications, 85% of the identifications were identical, demonstrating that the staining procedure has little if any influence on the depth of proteome analysis in our procedure (Figure 3C).

Lastly, it was possible that laser cutting of FFPE tissue, as employed in laser microdissection, could negatively affect proteome analysis. We therefore analyzed adjacent tissue sections that were either excised by Laser Capture Microdissection (LCM) or by needle scraping. In duplicate analysis 2178 proteins were identified in the scraped samples and 2201 in the samples obtained with LPC, with 87% overlap between identifications (Figure 3D). This indicates that LCM does not affect bulk proteomic analysis.

Together, the analyses in this section show that key storage and sample processing parameters have little or no influence on the number and type of proteins identified by FFPE-FASP, with the possible exception of prolonging fixation times up to several days.

Accurate Quantitative Proteomics of FFPE Samples. The experiments described above were qualitative rather than directly quantifying fresh and fixed tissue against each other. Accurate quantitative analysis is challenging because of the limited amounts that are usually obtained from tissue sections. As we had already excluded large changes between the fresh and FFPE tissues, we therefore turned to the SILAC method, which is very accurate and sensitive.[31,32] SILAC is normally thought to be restricted to cell culture but here we made use of the recently developed SILAC-mouse, in which an entire animal is labeled by a heavy amino acid.[15] This makes peptides from any mouse tissue of the heavy label mouse distinguishable from peptides from the same tissue of a control, "light" labeled mouse. When combining tissue from heavy ($^{13}C_6$-Lys) labeled mice and light ($^{12}C_6$-Lys) control mice, and digesting with LysC, every peptide will occur as a pair (except the C-terminal peptide of the protein), allowing accurate quantification of the parent protein.

Figure 4 shows the scheme for quantifying potential differences between fresh and FFPE tissues. To eliminate experimental variability between individual mice, we obtained a liver from a control mouse and from a SILAC-mouse and cut them in half. In each case, we fixed and paraffin embedded one-

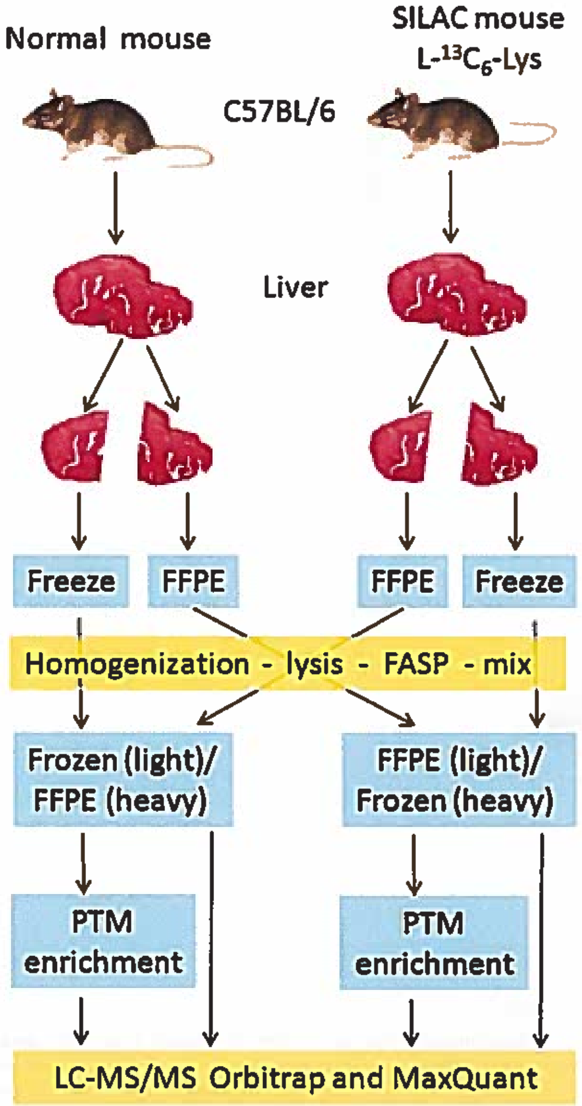

Normal mouse
SILAC mouse
L-$^{13}C_6$-Lys
C57BL/6
Liver
Freeze
FFPE
FFPE
Freeze
Homogenization - lysis - FASP - mix
Frozen (light)/ FFPE (heavy)
FFPE (light)/ Frozen (heavy)
PTM enrichment
PTM enrichment
LC-MS/MS Orbitrap and MaxQuant

Figure 4. Quantitative analysis of normal tissue samples vs FFPE samples in the SILAC mouse. Mouse livers were taken from normal and heavy SILAC-labeled mice, halved and each half either analyzed fresh (after freezing) or after FFPE. SILAC comparisons were done between fresh and FFPE livers. Peptides are either quantified directly after FFPE-FASP or after a PTM enrichment step. Note that the crossover experiment eliminates individual differences between mice and exposes differences in sample preparation.

half of the liver and processed the other half after freezing. We then mixed the FFPE heavy liver sample with the fresh liver sample of the control mouse. In a reverse labeling control experiment, we mixed the light FFPE liver sample with the heavy labeled fresh liver sample. In this way, genuine quantitative changes introduced by the FFPE-FASP protocol would need to be reversed in the control experiment, independent of any biological differences between the livers of the two mice.

For the forward and for the reverse experiment, we prepared fresh and FFPE liver by the FFPE-FASP protocol followed by six fraction pipet-based SAX.[12] Each of the resulting 12 fractions was measured by LC−MS/MS with 230 min gradients on the LTQ-Orbitrap Velos, resulting in the quantification of 19 145 identified SILAC-peptide pairs and 3740 proteins in MaxQuant (Supplemental Table 3, Supporting Information). Quantification results for proteins are depicted in Figure 5. Plotting fold-change vs protein abundance (total peptide signal per protein) revealed that the vast majority of proteins are located at the 1:1 axis (zero on the $\log_2$ scale) in both the forward and the reverse experiment, regardless of abundance of the protein. A 2-fold change in protein amount is chosen as cutoff for biological significance in many proteomic experiments. A full 96% of the proteins were within this interval (−1 to 1 interval on the $\log_2$ scale). This already demonstrated the lack of widespread quantitative changes to the proteome due to FFPE when using the FFPE-FASP protocol.

There are some statistically significant outliers in the plot, colored red in Figure 5A and B. These could in principle be due to the fixation and paraffin embedding or due to biological differences in the samples. To distinguish these two possibilities, we determined if any of the proteins changed in a manner consistent with an effect of FFPE. We plotted fold changes of heavy SILAC labeled fresh tissue to light (normal) labeled FFPE tissue (H/L fold change) on the x-axis and light labeled fresh tissue to heavy labeled SILAC FFPE tissue (L/H fold change) on the y-axis (Figure 5C). If there were any proteins that are reduced due to FFPE they must be outliers in the positive fold change direction on both axes. However, the upper right quadrant of Figure 5C is empty, indicating that there are no such proteins. Likewise there are no proteins whose quantity is increased due to FFPE (empty lower left quadrant).

The outliers in the upper left quadrant are proteins that are increased in the control liver sample compared to the SILAC liver sample. When preparing liver samples, it is possible to include small amounts of adjacent tissue in some cases, and we suspected that this was the cause of these outliers. Indeed, when inspecting the identities of the outlier proteins, we found that some of them were myosin isoforms or creatine kinase (belonging to the most abundant proteins in muscle tissue), suggesting that some adjacent muscle tissue was included with the control liver but not the SILAC liver. Additionally, some of the outliers were mouse keratins. These are readily explained by extensive sequence identity to human protein contaminants. (These human contaminant proteins had been removed before analysis.) Our results therefore do not reveal any quantitative changes between fresh and FFPE tissue when analyzed by high resolution mass spectrometry after FFPE-FASP).

In addition to ruling out general or specific quantitative changes due to FFPE in our protocol, these results imply that accurate quantification is possible from FFPE tissues. The ratio distributions are quite similar to other large-scale data sets that we have obtained on fresh SILAC material analyzed at equal depths of proteome coverage.[33,34]

Identification of Phosphorylation Sites. Analysis of changes in posttranslational modifications can be crucial for understanding pathological processes. Having established the feasibility of analyzing proteome changes both qualitatively and quantitatively, we therefore wished to investigate if the phosphoproteome was preserved in FFPE and whether it was analyzable by MS-based proteomics.

We prepared peptides by FFPE-FASP from 2.5 mg of fixed mouse liver protein. The resulting peptides were enriched for phosphopeptides on TiO_2 beads in a pipet and were separated into 4 fractions using the SAX approach (Experimental Procedures). They were analyzed in 230 min gradients on the LTQ-Orbitrap Velos using HCD fragmentation.[25,35] In duplicate

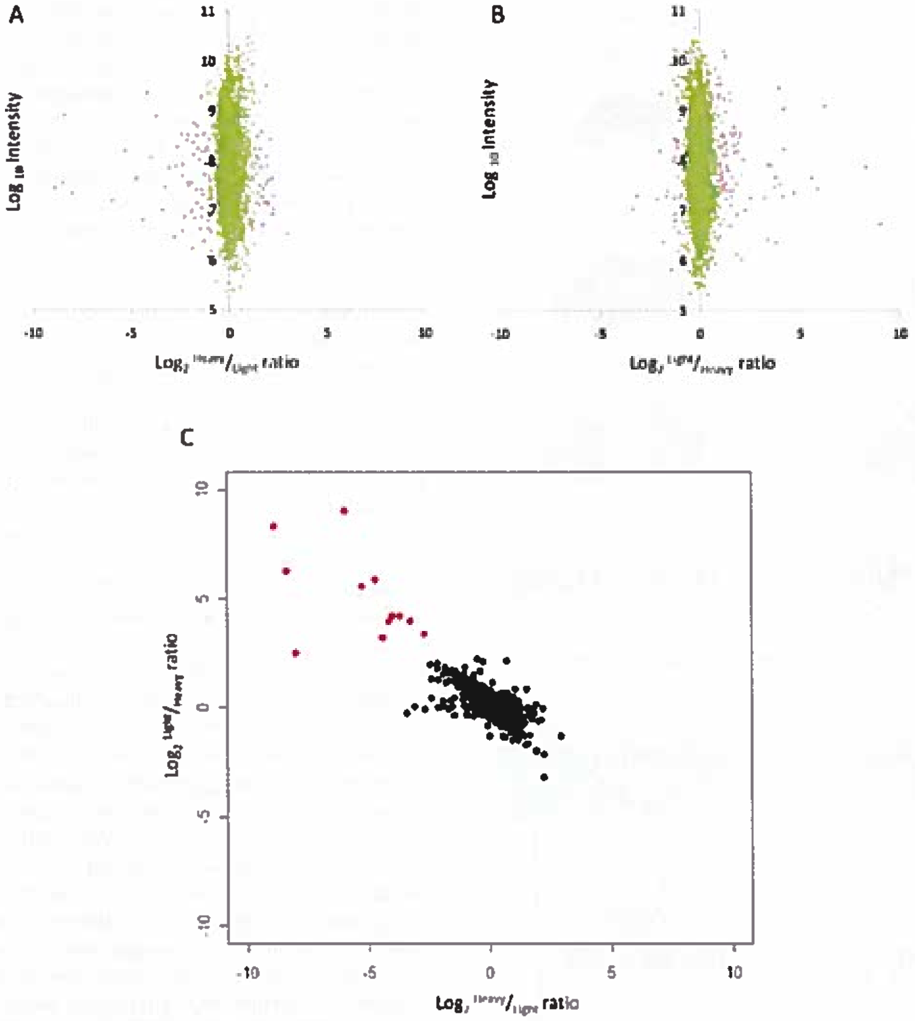


Figure 5. SILAC quantification of fresh against FFPE tissue samples. Quantification was done as outlined in Figure 4. In each panel plotted ratios are fresh divided by FFPE protein signal. (A) Fresh, heavy labeled tissue sample against fixed, normal labeled tissue sample. SILAC-protein ratios are plotted as a function of added peptide intensity of each protein. Statistically significant outliers are colored in red ($p < 0.01$). (B) Crossover labeling experiment. Fresh, normal labeled tissue sample against fixed, heavy labeled tissue sample. (C) Ratios of the upper two panels plotted against each other. None of the statistically significant outliers are in the upper right-hand quadrant or the lower left-hand quadrant, which would have indicated selective influence of FFPE treatment. The outliers in the upper left-hand corner are due to differences in sample handling (red dots).

experiments with this minimal preparation protocol we identified 7718 different phosphopeptides in the fresh and 6870 different phosphopeptides in the FFPE tissue material (Supplemental Table 4, Supporting Information). These reduced to 5799 class I phospho-sites in fresh liver tissue and 4991 class I phospho-sites in FFPE (Figure 6A). Class I sites have a localization probability of at least 75% by definition, but median and mean localization probability was 0.99 and 0.87, respectively.

We found that 4121 (62%) sites were identified in both experiments (Figure 6A). This limited overlap probably is not due to differences in the phosphoproteome of fresh and PPFE tissue because a separate experiment indicated a very similar overlap of measured phosphosites between two fresh samples (59%; Figure 6B). Rather, the explanation for this relatively low degree of overlap is likely to be the large size of the phosphoproteome, of which we are only sampling a fraction. These experiments demonstrated that large numbers of phosphorylation sites are preserved in FFPE tissues and that they can be analyzed by LC−MS/MS to the same degree as fresh samples.

Next we investigated quantification of the phosphoproteome in FFPE against fresh mouse liver tissue. We again made use of the SILAC mouse and the experimental scheme indicated in Figure 4, with the difference that we added a phosphopeptide enrichment step. In the MaxQuant analysis of the combined forward and reverse experiments we quantified 2417 different phosphopeptides. As shown in the histograms in Figure 6C and D, the quantitative distributions are roughly symmetrical. Although slight bias toward light peptides can be observed it could be explained rather by biological differences between two samples (livers) than by differences caused by fixation, since light liver was either fixed or unfixed for forward and reverse experiment respectively. This indicates that FFPE does not impede retrieval of phosphorylated peptides by FFPE-FASP. The distribution of fold changes was somewhat broader than it was at the protein level (Figure 6C, D). This is because typically several peptides contribute to protein quantification whereas usually only one phosphopeptide contributes to the

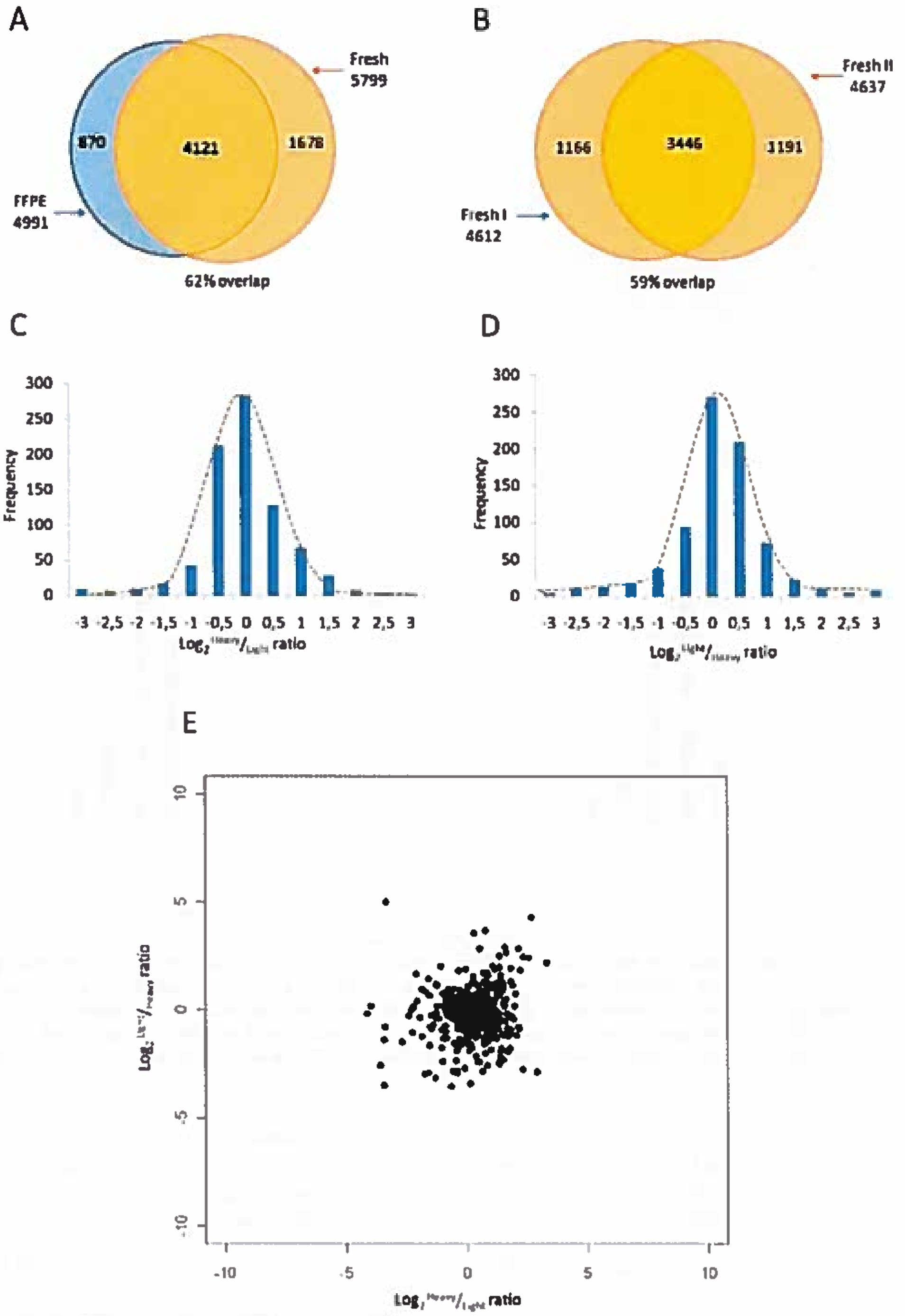


Figure 6. Phosphoproteomic analysis of FFPE tissue material. One and a half milligrams of peptides obtained from FFPE and fresh liver tissue samples were subjected to phosphopeptide enrichment, separated into four fractions by SAX and analyzed by LC–MS/MS using HCD on an LTQ-Orbitrap Velos. (A) Overlap in unique phosphorylation sites after duplicate analysis of FFPE and fresh material. (B) Overlap in unique phosphorylation sites after duplicate analysis of one fresh sample compared to another fresh sample. (C) Histogram of the fold changes observed between phosphopeptides extracted from fresh, heavy labeled and FFPE, normal labeled samples. (D) Histogram from the crossover experiment with ratios for FFPE, heavy labeled and fresh, normal labeled material. (E) Ratios of panel (C) and panel (D) plotted against each other.

quantification of a single phosphosite. Therefore, to decrease the probability of random error in quantitation of single peptide analysis of phosphosites requires more repetitions than in the case of protein quantitation.

To determine any potential changes in particular phosphopeptides due to FFPE, we plotted phosphopeptide ratios in the same way as we had done for the protein ratios before. As can be seen in Figure 6E, almost all phosphorylation sites are close

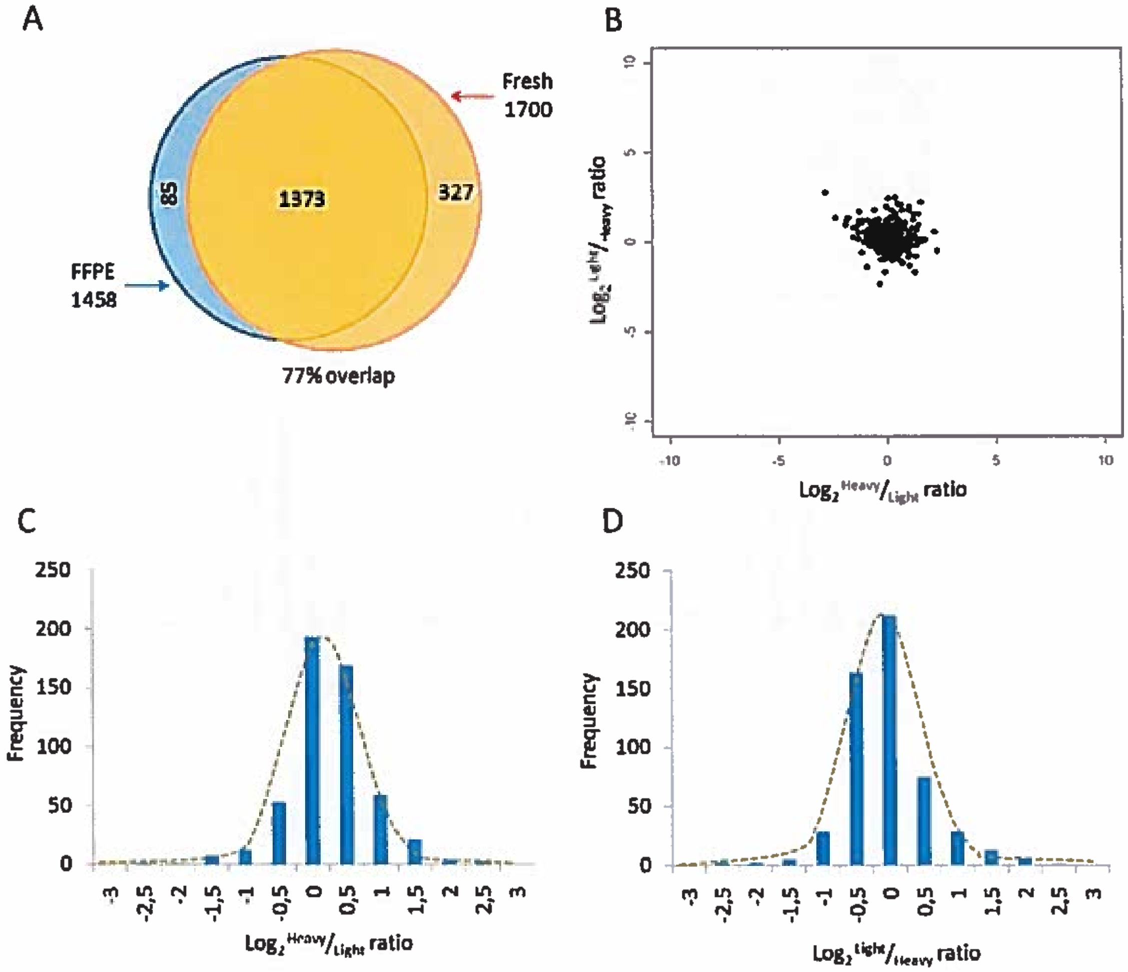


Figure 7. Analysis of N-glycosylation sites. One-hundred micrograms of peptides obtained from FFPE and fresh liver tissue samples were subjected to glycopeptide enrichment by N-Glyco-FASP and analyzed by LC−MS/MS on the LTQ-Orbitrap Velos using HCD. Cumulative results of four or three LC−MS/MS runs for the qualitative and SILAC analysis, respectively, are shown. (A) Overlap of unique glycosylation sites between FFPE and fresh tissue. (B) Ratios of the forward and crossover SILAC experiments show no significant outliers. (C) Histogram of the fold changes observed between glycopeptides extracted from fresh, heavy labeled and FFPE, normal labeled samples. (D) Histogram from the crossover experiment with ratios for FFPE, heavy labeled and fresh, normal labeled material.

to the origin. Few if any sites displayed an apparent change in phosphorylation due to FFPE of the tissue. Thus, our results demonstrate that qualitative and quantitative analysis of phosphorylation in FFPE material is feasible.

Mapping and Quantification of N-Glycosylation Sites. Glycosylation is one of the most important protein modifications[36] and changes in glycosylation patterns seem to be extremely frequent in neoplastic tissue.[37] A previous study has already reported identification of 168 sites in FFPE lung sample.[38]

Here, we applied the recently developed N-Glyco-FASP protocol[21] to the analysis of FFPE tissues. Peptides were processed as in the proteome analysis of FFPE samples described above except that peptides were allowed to interact with unbound lectins on top of the filters used in FASP. Peptides that were not captured by the lectins were washed through the filter. Deglycosylation in ^{18}O-water ($H_2{}^{18}O$) by PNGase F led to a mass shift of 2.9890 Da on the former N-glycosylated asparagines, making them easily identifiable to high resolution MS and HCD MS/MS on the LTQ-Orbitrap Velos. In four single runs, this approach allowed us to identify 1458 N-glycosylated sites in FFPE tissue sample and 1700 in fresh samples with an overlap 77% at the site level (Figure 7A, Supplemental Table 5, Supporting Information).

The analysis of 1:1 mixtures of nonlabeled and heavy SILAC labeled mouse liver tissue again showed a nearly symmetrical distribution in both the forward and the reverse experiment (Supplemental Table 5, Supporting Information, Figure 7C and D). As in the case of the phosphoproteome, a bias toward heavy peptides is likely due to biological differences between samples. Likewise, there are no specific outliers that are due to FFPE treatment of the sample (Figure 7B). Thus, glycosylation, in addition to phosphorylation, can be quantitatively analyzed in FFPE tissues.

In both the phosphorylation and glycosylation experiments, slightly more peptides were identified from fresh tissues. Even though this difference is small, we cannot exclude that it may in fact be significant.

Discussion

For more than a hundred years, biopsies have been preserved by fixing them in formalin and embedding them in paraffin blocks. In this form they are quite inert and easily stored for years or decades. The flip side of this preservation and stability is the inaccessibility of FFPE samples to analytical procedures. In a pioneering study a few years ago, Hood et al, first showed that FFPE material is in principle accessible to proteomic analysis.[5] Since then several parameters have been evaluated, mainly with a view to comparing analysis of FFPE and fresh tissue material.[6−8,10,11,29,39] These studies found that best results were achieved when the samples were lysed at high temperature and in the presence of a strong detergent, such as SDS.[29] Unfortunately, SDS even in small concentrations impairs enzymatic digestion of proteins and is not compatible with mass spectrometric analysis. Since effective removal of SDS from minute amounts of biological samples using standard biochemical methods is not easy, in the previous analyses of FFPE material strong detergents were avoided.[7,11] In contrast, in this work we used high concentrations of SDS for tissue lysis and processed the samples using the recently developed FASP protocol (Figure 1). This FFPE-FASP protocol allowed ready analysis of fixed and paraffin embedded sample.

Comparing our results with previously published data,[5,6,8,9,11] shows that we obtained the largest collection of proteins identified from single FFPE samples so far. The depth of analysis reached in this study is due to the method introduced here as well as to state of the art MS analysis equipment and the computational pipeline employed. Unlike most previous studies, we employed high resolution, high mass accuracy proteomics, resulting in subppm average absolute mass deviations. Proteins in our study were typically identified with many peptides (seven on average for the qualitative comparison of fresh and FFPE tissue). In many experiments, we employed high mass accuracy measurements not only in the MS but also in the MS/MS dimension. This demonstrates that very high data quality can be obtained from proteomes extracted from FFPE tissue material.

Accurate quantitation of thousands of proteins is still a challenge for proteomics, especially for samples that do not derive from cell lines and which therefore are less amenable to metabolic labeling.[40] Here, we used the completely labeled SILAC-mouse as a tissue source for quantitative comparison of fresh and fixed tissue. We found that quantitative accuracy in these experiments was equivalent to SILAC experiments comparing freshly prepared cell lines and conclude that quantitative accuracy is not compromised when analyzing FFPE material. While the SILAC-mouse is not a possible solution for the analysis of human tissue, we believe that the recently described super-SILAC strategy[41] would allow accurate quantitation of human FFPE tissues. In super-SILAC, a superset of labeled cell lines provides an accurate internal standard for a human tissue. For FFPE analysis, the super-SILAC mix would be added to the SDS-boiled protein extract during the FASP procedure.

For possible future use of the protocol in biomarker discovery, it is encouraging that our procedure- already at its current stage of development - allowed identification of more than 5000 proteins form a single sample with about one day of measurement time employing a relatively simple and fast pipet-tip SAX fractionation method.[12] In contrast, previously published protocols for in-depth proteome analysis of FFPE samples required less straightforward sample fractionation systems such as isoelectric focusing[11] or isotachophoresis.[9] Another practically important aspect of our method is its extremely low reagent cost, which is in marked contrast to commercial tissue preparation kits (which we did not discuss here because their composition is not known to the public).

Using the above technologies, we investigated whether and how the tissue proteome might be affected by the FFPE procedure. Both qualitatively and quantitatively we were not able to detect differences in protein levels between fresh or fixed and paraffin embedded proteomes. This indicates that the protein modifications induced by FFPE are not permanent (at least at a bulk level that can be probed by proteomics experiments) and that FFPE-FASP is efficient at reversing those modifications before protein analysis. Our findings contrast with several previous reports using different protocols and LC−MS/MS analyses schemes. For example, the ratio of tryptic peptides with lysine vs arginine at their C-terminus had been reported to be affected by FFPE,[5,11] whereas we found no such effect. Moreover, we found no evidence for reduced protein extraction or digestion efficiency with FFPE-FASP compared to fresh tissue, in contrast to other reports.[11] These differences may be due to the nearly complete solubilization efficiency obtained by SDS, which is also reflected in the high proportion of membrane proteins (30%) that we observed in fixed and paraffin embedded mouse liver.

In agreement with other studies,[9−11] we found that it is possible to analyze FFPE material after prolonged storage. Although we did not extend our observation period for more than half a year, we believe this result applies to longer times, too, because the process of formalin fixation is thought to be completed in 1 month.[30] In clinical settings fixation time is not standardized and often exceeds 24 h. We tested the ability of our protocol to retrieve proteins from samples fixed for five days and found that prolonged fixation (5 days) somewhat decreased the number of identifications. However, even after fixation for such a long time, we still obtained a high overlap in protein identifications with fresh tissue.

Microdissection is now a method of choice for obtaining homogeneous population of cells from clinical samples. For precise dissection of a region of interest it has to first be visualized by staining. Since hematoxylin, a standard dye in histopathology, can bind to DNA and thereby interfere with PCR,[42] it was important to check whether hematoxylin can impair mass spectrometric analysis, too. As the number of identifications was not reduced by hematoxylin staining, we conclude this standard stain can be used in conjunction with FFPE-FASP and proteomic analysis.

Although the impact of a focused laser beam on tissue should be restricted to the cutting line,[43] some authors prefer not to use lasers in sample preparation since possible side-effects of the laser on proteomics analysis are difficult to rule out.[6] We compared samples dissected either with a laser or scraped-out with a needle and found no significant difference in the number of identifications.

Using the FFPE-FASP protocol we demonstrated for the first time that the phosphoproteome is preserved in fixed and paraffin embedded tissue. Using straightforward pipet-based fractionation and TiO_2 enrichment, we identified 4991 phosphorylation sites with high localization accuracy. This number is similar to the number of identifications from fresh tissue and those that we have recently reported from cultured liver cells[44] or from cancer tissues in mouse models.[45]

Our glycoproteomic analysis of FFPE sample demonstrates that other PTMs than phosphorylation can also be efficiently analyzed by the FFPE-FASP protocol. A previous study aimed at identification of N-glycosylation in FFPE material reported 168 sites in FFPE lung sample.[38] With our protocol we were able to identify about 10 times mores N-glycosylation sites in a simple experiment without any additional fractionation of proteins or peptides before LC−MS/MS analysis. Our data set contains more than 1400 high-confidence N-glycosylation sites from FFPE liver tissue, measured with high accuracy at the MS/MS level by HCD.

A goal of the analysis of FFPE samples is to discover new biomarkers useful for a more personalized therapy, for better survival prediction or for an earlier diagnosis. Our work provides a straightforward technique for quantitative analysis of FFPE material. The FFPE-FASP procedure enables quantitative mapping of phosphorylation and N-glycosylation sites to a depth that has previously been achieved using fresh biological material. The FFPE-FASP method described here has the potential for opening new perspectives in the proteomic exploration of archival clinical samples at the proteome and the PTM level.

Acknowledgment. We thank Dr. Sara Zanivan for providing the SILAC-labeled mouse liver and Nagarjuna Nagaraj, Dr. Alexandre Zougman, and Korbinian Mayr for help and fruitful discussions. This work was supported by the Max-Planck Society for the Advancement of Science, from the European Commission's seventh Framework Programme (grant agreement HEALTH-F4-2008-201648/PROSPECTS) and the Munich Center for Integrated Protein Science (CIPSM).

Supporting Information Available: FFPE-FASP-Protocol.

Supplemental Table 1 FFPE and fresh proteome identification data.

Supplemental Table 2 Influence of key FFPE processing parameters.

Supplemental Table 3 SILAC proteome fresh against FFPE tissue.

Supplemental Table 4 Phosphopeptide data.

Supplemental Table 5 Glycopeptide data. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Idikio, H. A. Immunohistochemistry in diagnostic surgical pathology: contributions of protein life-cycle, use of evidence-based methods and data normalization on interpretation of immunohistochemical stains. *Int. J. Clin. Exp. Pathol.* 2009, *3* (2), 169–76.
(2) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* 2003, *422* (6928), 198–207.
(3) Cravatt, B. F.; Simon, G. M.; Yates, J. R., 3rd. The biological impact of mass-spectrometry-based proteomics. *Nature* 2007, *450* (7172), 991–1000.
(4) Mann, M.; Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U.S.A.* 2008, *105* (47), 18132–8.
(5) Hood, B. L.; Darfler, M. M.; Guiel, T. G.; Furusato, B.; Lucas, D. A.; Ringeisen, B. R.; Sesterhenn, I. A.; Conrads, T. P.; Veenstra, T. D.; Krizman, D. B. Proteomic analysis of formalin-fixed prostate cancer tissue. *Mol. Cell. Proteomics* 2005, *4* (11), 1741–53.
(6) Shi, S. R.; Liu, C.; Balgley, B. M.; Lee, C.; Taylor, C. R. Protein extraction from formalin-fixed, paraffin-embedded tissue sections: quality evaluation by mass spectrometry. *J. Histochem. Cytochem.* 2006, *54* (6), 739–43.
(7) Jiang, X.; Jiang, X.; Feng, S.; Tian, R.; Ye, M.; Zou, H. Development of efficient protein extraction methods for shotgun proteome analysis of formalin-fixed tissues. *J. Proteome Res.* 2007, *6* (3), 1038–47.
(8) Guo, T.; Wang, W.; Rudnick, P. A.; Song, T.; Li, J.; Zhuang, Z.; Weil, R. J.; DeVoe, D. L.; Lee, C. S.; Balgley, B. M. Proteome analysis of microdissected formalin-fixed and paraffin-embedded tissue specimens. *J. Histochem. Cytochem.* 2007, *55* (7), 763–72.
(9) Xu, H.; Yang, L.; Wang, W.; Shi, S. R.; Liu, C.; Liu, Y.; Fang, X.; Taylor, C. R.; Lee, C. S.; Balgley, B. M. Antigen retrieval for proteomic characterization of formalin-fixed and paraffin-embedded tissues. *J. Proteome Res.* 2008, *7* (3), 1098–108.
(10) Balgley, B. M.; Guo, T.; Zhao, K.; Fang, X.; Tavassoli, F. A.; Lee, C. S. Evaluation of archival time on shotgun proteomics of formalin-fixed and paraffin-embedded tissues. *J. Proteome Res.* 2009, *8* (2), 917–25.
(11) Sprung, R. W., Jr.; Brock, J. W.; Tanksley, J. P.; Li, M.; Washington, M. K.; Slebos, R. J.; Liebler, D. C. Equivalence of protein inventories obtained from formalin-fixed paraffin-embedded and frozen tissue in multidimensional liquid chromatography-tandem mass spectrometry shotgun proteomic analysis. *Mol. Cell. Proteomics* 2009, *8* (8), 1988–98.
(12) Wisniewski, J. R.; Zougman, A.; Mann, M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.* 2009, *8*, 5674–8.
(13) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* 2009, *6* (5), 359–62.
(14) Shi, S. R.; Key, M. E.; Kalra, K. L. Antigen retrieval in formalin-fixed, paraffin-embedded tissues: an enhancement method for immunohistochemical staining based on microwave oven heating of tissue sections. *J. Histochem. Cytochem.* 1991, *39* (6), 741–8.
(15) Kruger, M.; Moser, M.; Ussar, S.; Thievessen, I.; Luber, C. A.; Forner, F.; Schmidt, S.; Zanivan, S.; Fassler, R.; Mann, M. SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* 2008, *134* (2), 353–64.
(16) Nielsen, P. A.; Olsen, J. V.; Podtelejnikov, A. V.; Andersen, J. R.; Mann, M.; Wisniewski, J. R. Proteomic mapping of brain plasma membrane proteins. *Mol. Cell. Proteomics* 2005, *4* (4), 402–8.
(17) Rappsilber, J.; Ishihama, Y.; Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* 2003, *75* (3), 663–70.
(18) Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell. Proteomics* 2005, *4* (7), 873–86.
(19) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 2006, *127* (3), 635–48.
(20) Zielinska, D. F.; Gnad, F.; Jedrusik-Bode, M.; Wisniewski, J. R.; Mann, M. Caenorhabditis elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J. Proteome Res.* 2009, *8* (8), 4039–49.
(21) Zielinska, D. F.; Gnad, F.; Jedrusik-Bode, M.; Wisniewski, J. R.; Mann, M. Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints. *Cell* 2010, in press (doi: 10.1016/j.cell.2010.04.012.
(22) Kuster, B.; Mann, M. 18O-labeling of N-glycosylation sites to improve the identification of gel-separated glycoproteins using peptide mass mapping and database searching. *Anal. Chem.* 1999, *71* (7), 1431–40.
(23) Olsen, J. V.; Schwartz, J. C.; Griep-Raming, J.; Nielsen, M. L.; Damoc, E.; Denisov, E.; Lange, O.; Remes, P.; Taylor, D.; Splendore, M.; Wouters, E. R.; Senko, M.; Makarov, A.; Mann, M.; Horning, S. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* 2009, *8* (12), 2759–69.
(24) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* 2005, *4* (12), 2010–21.
(25) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* 2007, *4* (9), 709–12.
(26) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, *26* (12), 1367–72.

(27) Cox, J.; Matic, I.; Hilger, M.; Nagaraj, N.; Selbach, M.; Olsen, J. V.; Mann, M. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat. Protoc.* 2009, *4* (5), 698–705.

(28) Zar J. *Biostatistical Analysis*; Prentice Hall: Upper Saddle River, NJ, 1999.

(29) Fowler, C. B.; Cunningham, R. E.; O'Leary, T. J.; Mason, J. T. 'Tissue surrogates' as a model for archival formalin-fixed paraffin-embedded tissues. *Lab. Invest.* 2007, *87* (8), 836–46.

(30) Kiernan, J. *Histological and Histochemical Methods: Theory and Practice*; 4th ed.; Cold Spring Harbor Laboratory Press: Long Island, NY, 2008.

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

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

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

(34) Graumann, J.; Hubner, N. C.; Kim, J. B.; Ko, K.; Moser, M.; Kumar, C.; Cox, J.; Scholer, H.; Mann, M. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol. Cell. Proteomics* 2008, *7* (4), 672–83.

(35) Olsen, J. V.; Schwartz, J. C.; Griep-Raming, J.; Nielsen, M. L.; Damoc, E.; Denisov, E.; Lange, O.; Remes, P.; Taylor, D.; Splendore, M.; Wouters, E. R.; Senko, M.; Makarov, A.; Mann, M.; Horning, S. A dual pressure linear ion trap - Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* 2009, *8*, 2759–69.

(36) Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E. *Essentials of Glycobiology* 2008.

(37) Peracaula, R.; Barrabes, S.; Sarrats, A.; Rudd, P. M.; de Llorens, R. Altered glycosylation in tumours focused to cancer diagnosis. *Dis. Markers* 2008, *25* (4−5), 207–18.

(38) Tian, Y.; Gurley, K.; Meany, D. L.; Kemp, C. J.; Zhang, H. N-linked glycoproteomic analysis of formalin-fixed and paraffin-embedded tissues. *J. Proteome Res.* 2009, *8* (4), 1657–62.

(39) Crockett, D. K.; Lin, Z.; Vaughn, C. P.; Lim, M. S.; Elenitoba-Johnson, K. S. Identification of proteins from formalin-fixed paraffin-embedded cells by LC-MS/MS. *Lab. Invest.* 2005, *85* (11), 1405–15.

(40) Ong, S. E.; Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* 2005, *1* (5), 252–62.

(41) Geiger, T.; Cox, J.; Ostasiewicz, P.; Wisniewski, J.; Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* 2010, *7* (5), 383–5.

(42) Gjerdrum, L. M.; Hamilton-Dutoit, S. Laser-assisted microdissection of membrane-mounted tissue sections. *Methods Mol. Biol.* 2005, *293*, 127–38.

(43) Micke, P.; Ostman, A.; Lundeberg, J.; Ponten, F. Laser-assisted cell microdissection using the PALM system. *Methods Mol. Biol.* 2005, *293*, 151–66.

(44) Pan, C.; Kumar, C.; Bohl, S.; Klingmueller, U.; Mann, M. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type specific functions. *Mol. Cell. Proteomics* 2009, *8*, 443–50.

(45) Zanivan, S.; Gnad, F.; Wickstrom, S. A.; Geiger, T.; Macek, B.; Cox, J.; Fassler, R.; Mann, M. Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. *J. Proteome Res.* 2008, *7*, 5314–26.

(46) Burlone, M. E.; Budkowska, A. Hepatitis C virus cell entry: role of lipoproteins and cellular receptors. *J. Gen. Virol.* 2009, *90* (Pt 5), 1055–70.

(47) Nguyen, D. H.; Ludgate, L.; Hu, J. Hepatitis B virus-cell interactions and pathogenesis. *J. Cell Physiol.* 2008, *216* (2), 289–94.

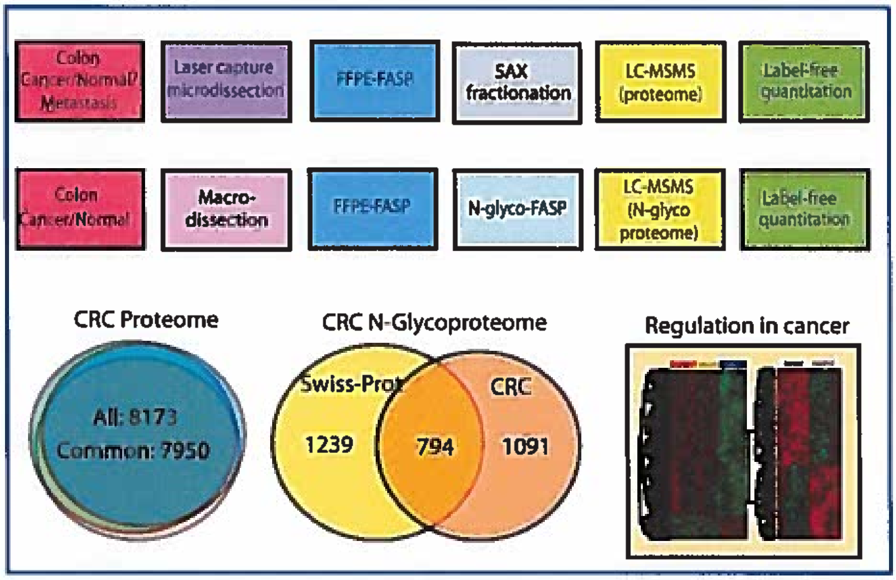
(48) Schiess, R.; Wollscheid, B.; Aebersold, R. Targeted proteomic strategy for clinical biomarker discovery. *Mol. Oncol.* 2009, *3* (1), 33–44.

PR100234W

Publication IV

Proteomic and N-Glycoproteomic Profiling of Colorectal Cancer

Zielinska DF, Ostasiewicz P, Gnad F, Duś K, Mann M, Wiśniewski JR



(In preparation)

This publication describes proteomic and N-glycoproteomic quantitative analyses of colorectal cancer and normal colonic mucosa using formalin-fixed paraffin-embedded human material. We identified 8,173 proteins and 1,885 N-glycosylation sites and compared their abundances. We show that as a general rule the glycoprotein expression, and not the glycosylation site occupancy, is altered in cancer. However, in several proteins we also find evidence for differentially regulated site occupancy.

Proteomic and N-Glycoproteomic Profiling of Colorectal Cancer

Dorota F. Zielinska[1], Pawel Ostasiewicz[1,2], Florian Gnad[3], Kamila Duś[1,2], Matthias Mann[1]* and Jacek R. Wiśniewski[1]*

[1]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany
[2]Department of Pathology, Wroclaw Medical University, PL-50368 Wroclaw, Poland
[3]Department of Bioinformatics, Genentech, Inc., CA 94080 South San Francisco, USA
Correspondence: jwisniew@biochem.mpg.de or mmann@biochem.mpg.de

SUMMARY

Colorectal cancer (CRC) is one of the most common cancers and each year more than half a million people die of this disease. CRC arise from the colonic mucosa and spreads throughout the body forming metastases. Although these processes have been extensively studied with biochemical and genetic methods only little progress has been done in characterization of the changes occurring at the proteome level. Here we report results obtained in proteomic analyses of laser capture microdissected formalin fixed tissues and N-glycoproteomic analysis of macrodissected formalin fixed tissues. Proteomic analysis of patient-matched sets of colonic mucosa, the primary cancer and its nodal metastasis allowed identification of more than 8,000 proteins and their quantitative comparisons between the three stages. We found 1,800 proteins which are significantly up or down regulated in cancer ($p < 0.05$), but did not observe any significant differences between cancer and metastasis. The N-glycoproteomic analysis of patient-unmatched samples of colonic mucosa and primary cancer enabled detection of nearly 2,000 N-glycosylation sites on 1,000 proteins. We observed that 400 glycopeptides were changed in cancer and we showed that as a general rule the glycoprotein expression, and not the glycosylation site occupancy, is altered in cancerous tissues. Furthermore, we found evidence for different N-glycosylation patterns between cancer and normal colonic tissues, which could be of potential clinical interest. Our datasets contains all known CRC protein biomarkers and provide a resource of differentially expressed proteins.

RESULTS

Identification of 8,000 Proteins from Laser Capture Microdissected Colonic Mucosa, Primary Cancer and its Nodal Metastasis

To survey and compare the proteomes of colonic mucosa (N), primary cancer (C) and its nodal metastasis (M) archival formalin fixed and paraffin embedded clinical samples originating from eight patients were analyzed. Laser capture microdissection was used to obtain enriched populations of enterocytes, primary cancer, and metastasizing cells. From each sample 175 nl of cells were collected and processed using recently developed FFPE-FASP procedure (Ostasiewicz et al., 2010). It yielded 3.1±0.9 µg total peptide per sample. The peptides were fractionated by anion exchange chromatography into 6 fractions and analyzed by LC-MS/MS on Orbitrap-Velos instrument as described previously (Wisniewski et al., 2011a) (Figure 1A). The obtained spectral data were analyzed using the *Andromeda* searching engine (Cox et al., 2011) and the *MaxQuant* software (Cox and Mann, 2008) using a label free quantitation option (Cox et al., submitted) allowing quantitative comparison of the individual samples. Our analysis allowed identification of 72,000 unique peptides (Supplementary Table 1) corresponding to 8,173 proteins (Supplementary table 2). Overall, 841 of the latter were reported as the components of human plasma (Schenk et al., 2008) or erythrocytes (Pasini et al., 2006) and hence may partially originate from blood (Figure 1C). Comparison of the N, C, and M proteomes revealed that 99% of the identified proteins were common in all the three samples (Figure 1D).

Mapping of 1,885 N-Glycosylation Sites in the Colonic Mucosa and the Primary Cancer

To identify N-glycosylation sites on proteins occurring in N and C we measured the N-glycoproteomes of FFPE samples using the N-glyco-FASP method (Zielinska et al., 2010) (Figure 1B). The suitability of this methods for analysis of FFPE tissues was recently demonstrated (Ostasiewicz et al., 2010). Since identification of posttranslational modifications requires affinity enrichment and hence larger starting amounts of total protein, the profiling of N-glycoproteomes was performed using macro-dissected material. The tissues were processed according to the FFPE-FASP protocol (Ostasiewicz et al., 2010) using 30 kDa filtration units (Wisniewski et al., 2011b). N-glyco peptides were enriched by adding a lectin mixture to the top of the filters. After deglycosylation with PNGase F the peptides were identified by LC-MS/MS on Orbitrap-Velos instrument.

We analyzed eight samples per N and C in duplicates originating from unmatched patients (in total 16 measurements per state). On average 1,000 highly confident sites were identified per single LC-MS/MS run. In total we identified 1,885 N-glycosylation sites with a minimum localization probability of 95% and at least three identifications per N or C state. These sites were distributed on 962 proteins. It is the largest N-glycoproteomic dataset for human tissue published to date. According to Swiss-Prot (Wu et al., 2006) we cover 40% of all known human N-glycosylation sites and increase the number of experimentally validated sites from 2,000 to 3,000 (Figure 1E). Identified sites are listed in Supplementary Table 3 and are available in the post-translational modification database PHOSIDA (Gnad et al., 2007) (http://www.phosida.com).

The comparison of the proteomic and N-glycoproteomic datasets revealed that 38% of the N-glycoproteins were not detected in the proteome measurement (Figure 1F). We found that glycoproteins that were absent in the proteomic dataset were significantly enriched for the low abundance proteins ($p < 10^{-16}$) (Figure 1G).

Global Characterization of the Proteomes of the Colonic Mucosa, Primary Cancer and Metastasis

Using Gene Ontology annotations we calculated the percentage of identified proteins per 'cellular compartment' and 'molecular function'. The C, M, and N proteomes contain high percentage of proteins annotated as 'integral to membrane' and 'plasma membrane', which emphasizes the utility of the FFPE-FASP approach for studying tissues (Fig 2A). Based on the abundance of serum albumin, we estimated that 4% of detected proteins may originate from body fluids. Comparing the N-glycoproteomic and proteomic datasets, we estimated that 8% of proteins identified in measurement of microdissected normal mucosa are N-glycosylated (Fig 2C).

To describe the protein content constituting defined cell compartments or molecular functions and to provide insight into differences between states we calculated the total protein amount using normalized spectral intensities provided by the MaxQuant software. Figure 2 shows representative examples of the protein content in selected organelles and functions. These results demonstrate clear alterations between N and C, and no obvious differences between C and M. Whereas the N proteome contains 20-30% more extracellular and integral to plasma membrane proteins and about 30% less nuclear proteins then the C proteome (Figure 2A), there are no clear differences between N and C in the content of proteins belonging to cytoplasm and its major components such as the Golgi body, mitochondrion and endoplasmic reticulum. The differences in subcellular composition are reflected by the functional protein content (Figure 2B). Partial loss of integral membrane proteins in cancer is accompanied with a 50 and 30% drop in the content of channel and transporter proteins in N when comparing to C or M. Further, cancer cells have a two fold increase in the content of proteins with transcription factor activity, which is in concordance with elevated expression of nuclear proteins in cancer. The increased transcriptional activity is in connection with decreased compactness of chromatin, and therefore core histones are more abundant in cancer whereas the levels of linker histone remain unchanged between the normal colon and cancer. Additionally, elevated expression of proteins involved in cell cycle regulation in C reflects higher cell division rates of cancer cells in comparison to normal cells. The elevated expression of ribosomal proteins in C may reflect a portion of ribosomes occurring in the nuclear envelope.

In all three states we found similar amounts of N-glycoproteins that could partially or solely reflect blood contamination in the microdissected material (app. 4% of total protein) (Figure 3). Correct values are difficult to calculate because many cellular proteins can be found in plasma and are common to blood and solid tissue cells. The content of N-glycosylated proteins in colonic mucosa (at least 8 % of total cell protein) appears to be two times higher than in cancer and metastatic cells (~ 4 %) (Figure 2C), which is secondary to decreased glycosylation of extracellular proteins in cancer (Figure 2D). In C, M and N samples no obvious differences were detected for N-glycoprotein content in Golgi body and endoplasmic reticulum. Interestingly, whereas a two-fold drop in the content of N-glycoproteins with transporter activity was observed for cancer cells, a tenfold decrease was found for glycosylated proteins with channel activity (Figure 2E).

Proteome and N-glycoproteome changes in colon cancer

To identify potential driver proteins of cancer, we first compared the intensities of detected proteins between colonic mucosa, primary cancer and its nodal metastasis. Hierarchical clustering of the proteomic data reveals that the intensities of a substantial proportion of measured proteins are different between tumors and healthy samples, and that there are no differences between cancer and metastasis (Figure 3A). Student's t test was used to determine the significance of protein differences. In total 1,800 proteins were significantly down- or upregulated in tumor tissues ($p < 0.05$) (Figure 3B). Hierarchical clustering was used to visualize the differential expression of regulated proteins (Fig 3C). Next, we compared the intensities of detected glycopeptides between colonic mucosa and primary cancer. In total 398 N-glycopeptides on 288 proteins were significantly down- or upregulated in tumor tissues ($p < 0.01$) (Figure 3D, E).

We compared the calculated glycopeptide ratios with the ratios obtained from proteomic measurement and we showed that as a general rule the glycoprotein expression, and not the glycosylation pattern, is altered in cancerous tissues (Figure 3F). For example, we identified MUC2 to be downregulated in proteome measurement and we detected its 15 glycosylation sites in N-glycoproteome measurement that are consistently down-regulated. Similarly, DPEP1 and all mapped sites on DPEP1 are up-regulated, which is in agreement with its known protein overexpression in cancer (Toiyama et al., 2011).

Overall 75% of the 419 multiply glycosylated proteins identified in this study contained all sites strictly either down- or up-regulated (Figure S1A). In total 5% of identified proteins had differentially regulated site occupancy between the cases (Figure S1B). For example, we identified both highly up- and down-regulated sites on macroglobulin alpha 2, which was reported to have a decreased expression level in malignant breast cancer cells (Liang et al., 2006). Similarly, CFR1, a marker protein for the detection of precancerous and cancerous lesions (Brandlein et al., 2003), had two down- and one upregulated sites. Clusterin, known for differential expression of multiple isoforms in cancer (Rodriguez-Pineiro et al., 2006), also contains glycopeptides that are differentially regulated between the cancer and normal state.

REFERENCES

Brandlein, S., Beyer, I., Eck, M., Bernhardt, W., Hensel, F., Muller-Hermelink, H.K., and Vollmers, H.P. (2003). Cysteine-rich fibroblast growth factor receptor 1, a new marker for precancerous epithelial lesions defined by the human monoclonal antibody PAM-1. Cancer Res *63*, 2052-2061.

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.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res *10*, 1794-1805.

Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol *8*, R250.

Liang, X., Zhao, J., Hajivandi, M., Wu, R., Tao, J., Amshey, J.W., and Pope, R.M. (2006). Quantification of membrane and membrane-bound proteins in normal and malignant breast cancer cells isolated from the same patient with primary breast carcinoma. J Proteome Res *5*, 2632-2641.

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

Pasini, E.M., Kirkegaard, M., Mortensen, P., Lutz, H.U., Thomas, A.W., and Mann, M. (2006). In-depth analysis of the membrane and cytosolic proteome of red blood cells. Blood *108*, 791-801.

Rodriguez-Pineiro, A.M., de la Cadena, M.P., Lopez-Saco, A., and Rodriguez-Berrocal, F.J. (2006). Differential expression of serum clusterin isoforms in colorectal cancer. Mol Cell Proteomics *5*, 1647-1657.

Schenk, S., Schoenhals, G.J., de Souza, G., and Mann, M. (2008). A high confidence, manually validated human blood plasma protein reference set. BMC Med Genomics *1*, 41.

Toiyama, Y., Inoue, Y., Yasuda, H., Saigusa, S., Yokoe, T., Okugawa, Y., Tanaka, K., Miki, C., and Kusunoki, M. (2011). DPEP1, expressed in the early stages of colon carcinogenesis, affects cancer cell invasiveness. J Gastroenterol *46*, 153-163.

Wisniewski, J.R., Ostasiewicz, P., and Mann, M. (2011a). High Recovery FASP Applied to the Proteomic Analysis of Microdissected Formalin Fixed Paraffin Embedded Cancer Tissues Retrieves Known Colon Cancer Markers. J Proteome Res.

Wisniewski, J.R., Zielinska, D.F., and Mann, M. (2011b). Comparison of ultrafiltration units for proteomic and N-glycoproteomic analysis by the filter-aided sample preparation method. Analytical biochemistry *410*, 307-309.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res *34*, D187-191.

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

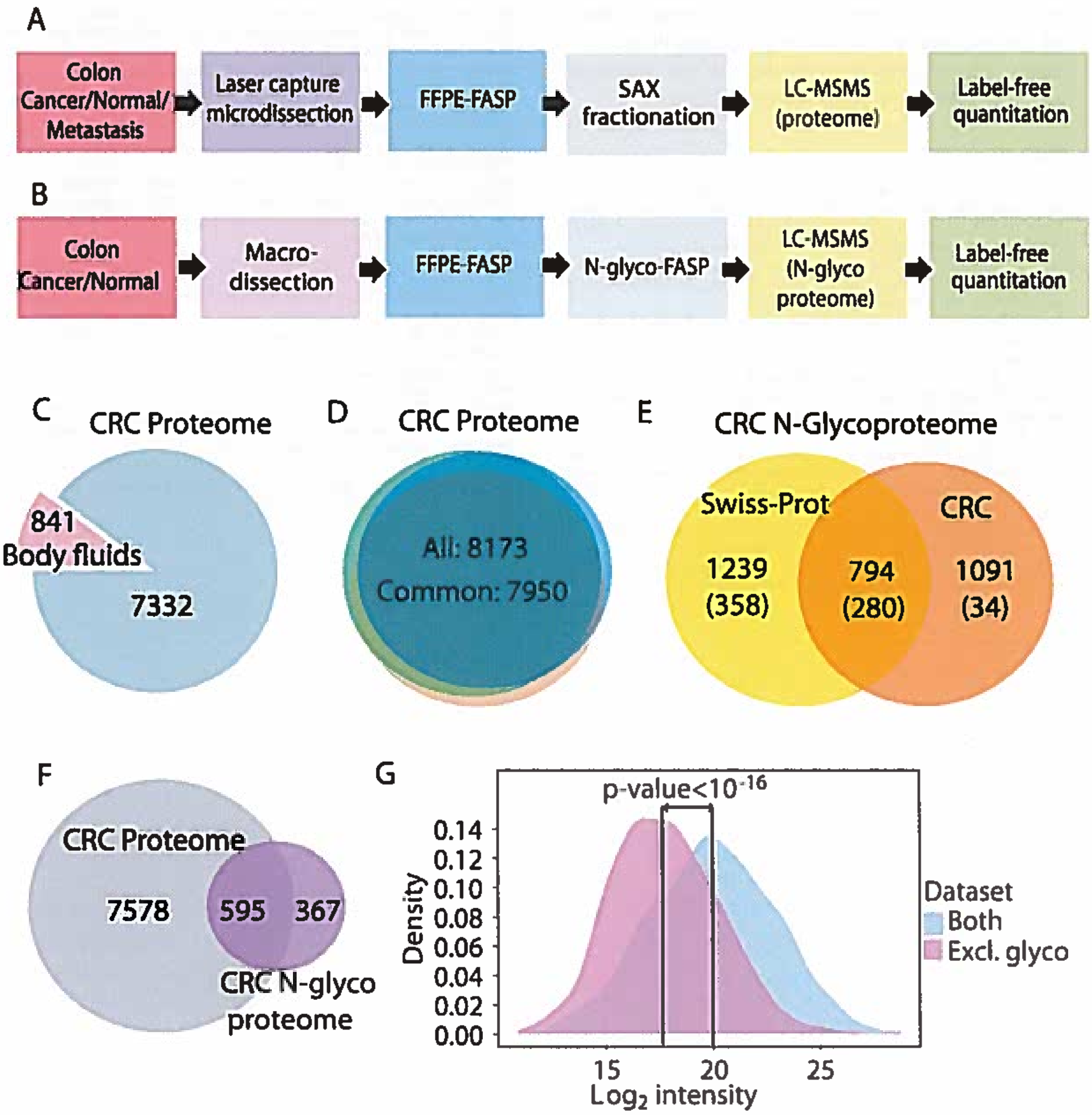


Figure 1. Proteomic and N-glycoproteomic profiling of colorectal cancer

(A) Proteomic workflow applied to microdissected samples of colonic mucosa, cancer, and metastasis (CMN).
(B) Workflow for mapping of N-glycosylation sites in normal colon and cancer.
(C) Identified proteins in proteome measurement. Proteins that may originate from blood are marked in pink.
(D) Overlap of the proteins identified from colonic mucosa, cancer, and metastasis.
(E) Overlap between here identified N-glycosylation sites and all experimentally derived sites in human according to Swiss-Prot. The glycosylated sites that may originate from blood are given in brackets.
(F) Overlap between detected colon cancer proteome and N-glycoproteome.
(G) Distributions of $\log_2$ MS intensities for glycopeptides found on proteins detected in both datasets (blue) and exclusively in N-glycoproteomic measurement (pink). The enrichment of low abundant glycosylated proteins in N-glycoproetomic dataset is represented by the significant difference in the means between the two distributions ($p < 10^{-16}$).

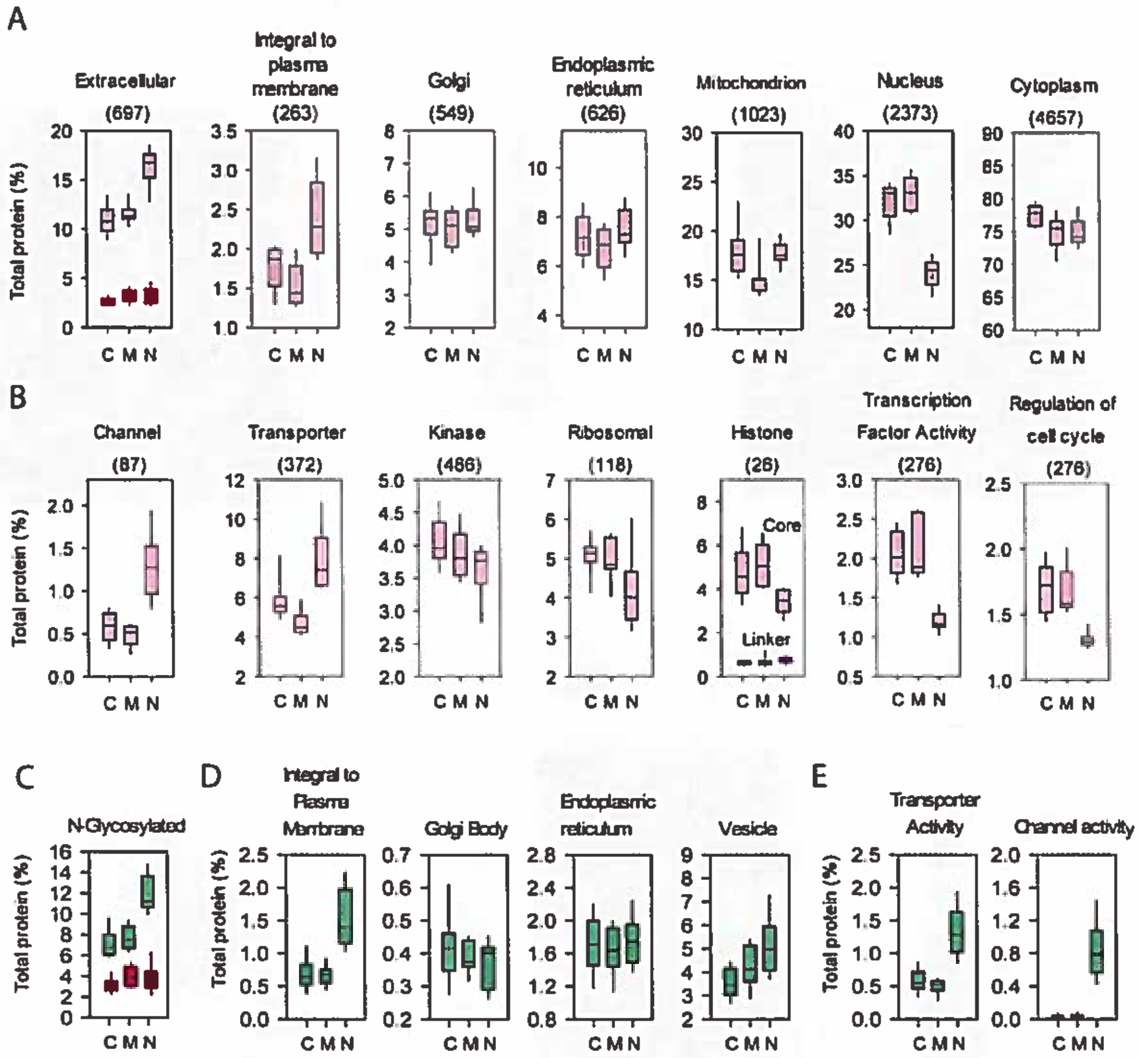


Figure 2. Comparison of protein and N-glycoprotein contents between cancer, metastasis and normal colon

(A) The underrepresentation of extracellular proteins and integral proteins to plasma membrane, unchanged expression pattern in Golgi, ER and cytoplasm, and overrepresentation of nuclear proteins in cancer compared to normal colon is pictured. In all states the amount of proteins that could potentially originate from blood (in red) is constant.

(B) Channels and transporters are underrepresented in cancer, kinases and mitochondrial proteins unchanged, whereas ribosomal proteins, core histones and transcriptions factors overrepresented.

(C) In normal colonic tissue, after subtraction of blood proteins (red), glycoproteins amounts approximately 8 % of total proteins which is nearly twice as much as in cancer and metastatic tissues (app. 4%).

(D) The amount of total glycoprotein integral to plasma membrane is twice as low in cancer as in normal colon and the amount of glycoprotein in ER and Golgi apparatus is unchanged.

(E) The expression of glycosylated transporters and channels is two-fold and ten-fold decreased in cancer when compared to colonic muscosa.

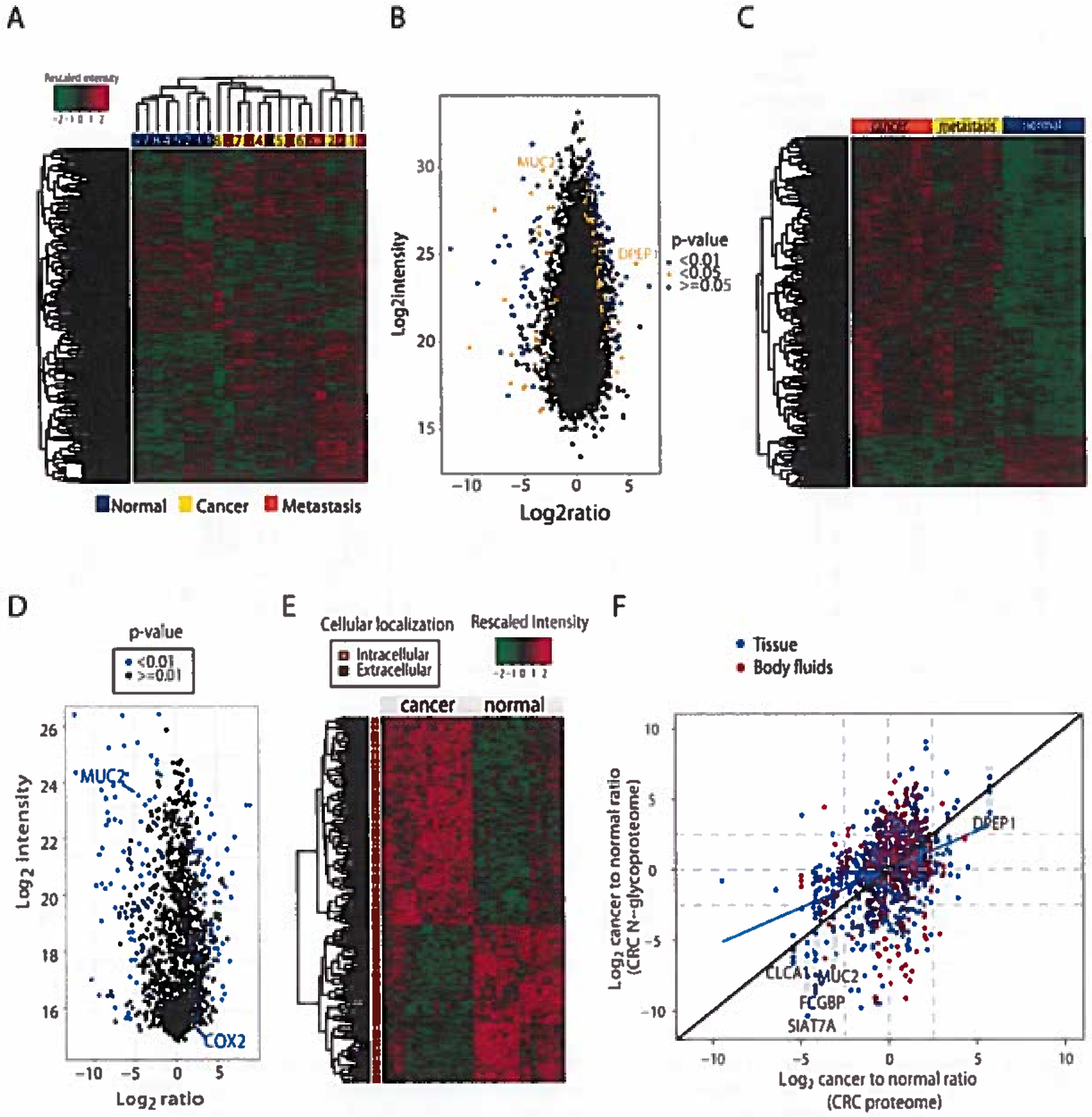


Figure 3. Proteome and N-glycoproteome changes in colon cancer

(A) Hierarchical clustering of the identified proteins in CRC proteome measurement. The samples are presented along the x-axis and the identified proteins along the y-axis. The detected intensities are rescaled. Unsupervised hierarchical clustering of the samples reveals that the protein intensity profiles differ sufficiently to clearly distinguish the normal state from cancer/metastasis. For cancer and metastasis, the replicates from individual patients are correctly clustered together.

(B) The intensity plot of identified proteins in CRC proteome measurement. Median of $\log_2$ ratios of cancer to normal intensities is presented along the x-axis and the median of $\log_2$ intensities along the y-axis.

(C) Hierarchical clustering of the regulated proteins ($p < 0.05$) in CRC proteome.

(D) The intensity plot of identified glycopeptides in CRC N-glycoproteome measurement. Log_2 ratio of the median intensity in cancer vs normal state is presented along the x-axis and the median of $\log_2$ intensities along the y-axis. Among the regulated proteins we find a number of proteins known to be involved in colorectal cancer like for example CEA, COX-2, MUC-2 and CaCC-1.

(E) Clustering of regulated glycopeptides intensities. Columns represent samples and rows differentially expressed glycopeptides. The color code on the right reveals that the majority of the regulated proteins is annotated to the extracellular region (brown) and the minority belongs to intracellular proteins (orange).

(F) Ratio vs ratio plot. Log_2 ratio of the median intensity in cancer vs normal state is presented along the x-axis and the median of $\log_2$ ratios of cancer to normal intensities along the y-axis. Proteins that may originate from blood are shown in red.

Publication V

Comparison of Ultrafiltration Units for Proteomic and N-Glycoproteomic Analysis by the FASP Method

Wiśniewski JR, Zielinska DF, Mann M

Anal Biochem, 2011

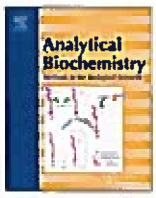
This publication describes the performance of various commercially available filtration units that are used as proteomic reactors in FASP and N-glyco-FASP protocols.

Analytical Biochemistry 410 (2011) 307–309

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio



ELSEVIER

Notes & Tips

Comparison of ultrafiltration units for proteomic and N-glycoproteomic analysis by the filter-aided sample preparation method

Jacek R. Wiśniewski *, Dorota F. Zielinska, Matthias Mann

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

ARTICLE INFO

Article history:
Received 1 October 2010
Received in revised form 3 December 2010
Accepted 3 December 2010
Available online 6 December 2010

ABSTRACT

The filter-aided sample preparation (FASP) method allows gel-free processing of biological samples solubilized with detergents for proteomic analysis by mass spectrometry. In FASP detergents are removed by ultrafiltration, and after protein digestion peptides are separated from undigested material. Here we compare the effectiveness of different filtration devices for analysis of proteomes and glycoproteomes. We show that Microcon and Vivacon filtration units with nominal molecular weight cutoffs of 30,000 and 50,000 (30 and 50 k, respectively) are equally suitable for FASP, whereas Microcon 30 k units are most appropriate for mapping of N-glycosylation sites. The use of filters with these relatively large cutoffs facilitates depletion of detergents.

© 2010 Elsevier Inc. All rights reserved.

Filter-aided sample preparation (FASP)[1] [1] is a technique allowing processing of lysates containing large amounts of sodium dodecyl sulfate (SDS) or other detergents for mass spectrometry-based proteome analysis. In FASP ultrafiltration units serve as "reactors" for chemical modification and digestion of proteins. This concept has also been used in several proteomic approaches [2–5] that were, however, incompatible with strong detergents for protein solubilization or did not allow complete detergent removal. In contrast, in FASP proteins are quantitatively depleted from detergents in the presence of concentrated urea [6], enabling analysis of both membrane and soluble proteins in the same way. FASP has already been successfully applied to mammalian [7–11], invertebrate [12], and yeast [13] samples for protein identification and mapping of phosphorylation [13–15] and N-glycosylation [14,16] sites.

FASP operates in a proteomic reactor modus where the ultrafiltration membrane allows separation of proteins from detergents, salts, and small molecular weight reagents before enzymatic digestion. After the digestion, FASP enables isolation of pure peptides that are free of undigested components such as nucleic acids and large polypeptides without specific digestion sites. Recently, the filtration principle of FASP has also been applied for lectin affinity enrichment of N-glycosylated peptides (N-glyco FASP) [16] and isolation of phosphotyrosine-containing peptides using uncoupled monoclonal anti-phosphotyrosine antibodies [15]. In both procedures, properties of the ultrafiltration membranes are essential for successful separation of individual components, leading to isolation of pure peptide mixtures that can be injected into a mass spectrometer.

FASP was originally developed using filters with small pores with nominal molecular weight cutoffs of 3000 and 10,000, but further developments showed that filters with larger pores significantly facilitate the sample preparation process [9]. Here we describe the performances of various commercially available filtration units applied to the FASP and N-glyco FASP protocols. We compared the performances of Microcon (Millipore) and Vivacon (Sartorius Stedim Biotech) with nominal cutoffs of 10,000, 30,000, and 50,000 (10, 30, and 50 k, respectively) using equal amounts of total cell lysate containing 45 μg of total protein. In all experiments, we used the "FASP II" protocol (described in Ref. [1]) that allows protein digestion in the absence of urea. Briefly, the SDS-solubilized samples were mixed with UA solution (0.2 ml of 8 M urea in 0.1 M Tris–HCl, pH 8.5), loaded into the filtration devices, and centrifuged at 14,000*g* for 15 min (or 45 min for 10 k devices). The concentrates were diluted in the devices with 0.2 ml of UA solution and centrifuged again. After centrifugation, the concentrates were mixed with 0.1 ml of 50 mM iodoacetamide in UA solution and incubated in darkness at room temperature for 30 min. Following centrifugation, the concentrate was diluted with 0.2 ml of UA solution and concentrated again. This step was repeated twice. Next, the concentrate was diluted with 0.1 ml of 40 mM $NaHCO_3$ and concentrated again. This step was repeated once. Subsequently, 0.45 μg of trypsin in 30 μl of 40 mM $NaHCO_3$ was added to the filter, and the samples were incubated at 37 °C overnight. Peptides were collected by centrifugation of the filter units, followed by two additional 30-μl washes with 40 mM $NaHCO_3$.

All tested ultrafiltration units led to isolation of similar amounts of tryptic peptides with yields ranging between 47% and 50% of the

---

* Corresponding author.
E-mail address: jwisniew@biochem.mpg.de (J.R. Wiśniewski).

[1] *Abbreviations used:* FASP, filter-aided sample preparation; SDS, sodium dodecyl sulfate; 10, 30, and 50 k, molecular weight cutoffs of 10,000, 30,000, and 50,000, respectively; UV, ultraviolet; MS, mass spectrometry; LC–MS/MS, liquid chromatography–tandem mass spectrometry; FDR, false discovery rate; ConA, concanavalin A; WGA, wheat germ agglutinin; CMC, critical micelle concentration; 100 k, 100,000.

0003-2697/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2010.12.004

loaded protein as measured by ultraviolet (UV) absorbance (Fig. 1A). The UV spectra also showed that overnight digestion in the Microcon filters results in contamination detectable in UV spectra ("empty Microcon" line in Fig. 1A). For the calculation of the protein concentration in the Microcon eluates, the background absorption originating from the "empty" filter units was subtracted. This contamination did not affect the mass spectrometry (MS) analysis of the samples, likely because it was removed by desalting of the samples on StageTips [17] before the analysis. To quantify low amounts of peptides, we measure the peptide concentration by fluorescence of tryptophan [15,18].

Each sample digest was analyzed in duplicate. Aliquots of 5-µg peptides were desalted on StageTips and analyzed by liquid chromatography–tandem mass spectrometry (LC–MS/MS) using 4-h gradients on a linear ion trap Orbitrap hybrid mass spectrometer (LTQ Orbitrap XL, Thermo Fisher) as described previously [1]. Single runs allowed identification of 6568–12,631 peptides representing 2507–2868 proteins at a false discovery rate (FDR) of 0.01 using MaxQuant software [19] (see Supplementary Table 1 in supplementary material). Between the individual filters, we did not observe any obvious bias against a specific group of proteins. Mass distributions of the peptides and proteins identified in individual experiments are shown in Fig. 1B and C, respectively. Even filters with large molecular weight cutoffs (50 k) retained small proteins; this can be explained by the "bulky" nature of unfolded proteins that prevents them from passing the filter membrane. Presumably for similar reasons, the distribution of the isolated peptides showed that the 30 and 50 k units passed more peptides with molecular weights greater than 1500 Da than did the 10 k filters. Thus, the 30 and 50 k ultrafiltration units are better suited for FASP. An additional practical advantage of these higher molecular weight cutoff filters is that the centrifugation time needed to concentrate samples is three to four times shorter than that with the 10 k units.

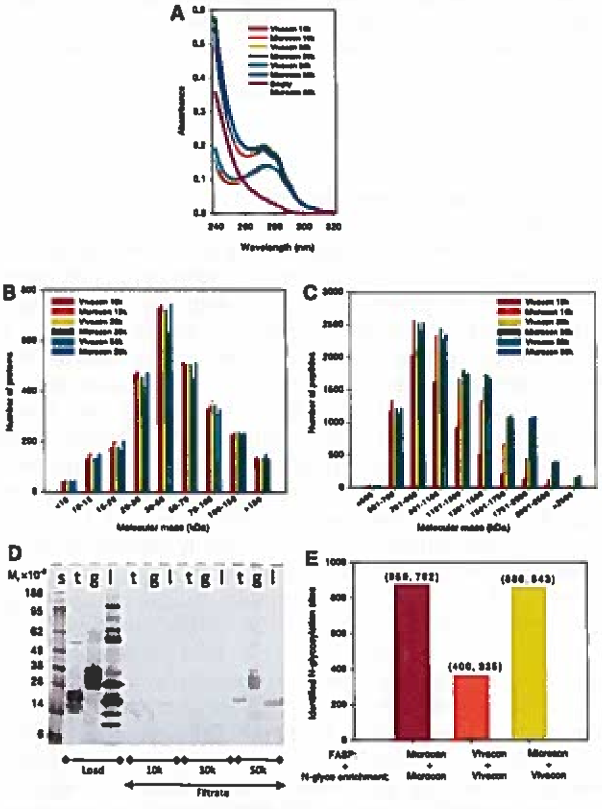


Fig. 1. Performances of Microcon and Vivacon protein concentration units in FASP and N-glyco FASP. Total lysates of MCF7 cells were processed according to the FASP procedure using different filtration units. The final eluates containing tryptic peptides were analyzed by LC–MS/MS. (A) UV spectra of the eluates. (B and C) Distribution of the molecular masses of identified proteins (B) and peptides (C). (D) Retention of trypsin (t), GluC (g), and lectin mixture (l) containing ConA, WGA, and RC-20 on Microcon filter with nominal molecular weight cutoffs of 10, 30, and 50 k. (E) Total lysates of mouse liver were processed according to the N-glyco FASP procedure using Microcon 30 k and Vivacon 30 k filters, and deglycosylated peptides were analyzed by LC–MS/MS. Each analysis was performed in duplicate. The bars represent the mean values, and individual results are shown in parentheses.

The N-glyco FASP protocol uses the ultrafiltration device twice: once for the generation of peptide mixtures (FASP) and once for affinity enrichment and deglycosylation of the glycan-containing peptides. The goal of the first step is to obtain the maximum possible amount of glycan-containing peptides. Because the peptides are incubated with lectins in the second step, the FASP eluate must be free from the proteinase used for digestion (e.g., trypsin, GluC). In the second step, lectins and the deglycosylation enzyme (PNGase F) must be fully retained on the filters because any leakage of them would affect the downstream MS analysis. To select ultrafiltration devices with pore sizes fulfilling these requirements, we tested the Microcon and Vivacon devices for their permeability to the digestion enzymes and the lectins (under the native conditions used in the protocol). We found that among the tested units, only the Microcon 50 k filters are not suitable for the N-glyco FASP method because the membrane retains neither proteinases nor the lectins (Fig. 1D).

To compare the performances of the Microcon and Vivacon filters in the N-glyco FASP protocol, we analyzed whole mouse liver lysates using 30 k filters. Briefly, the samples containing 0.4 mg of total protein were processed according to the FASP II protocol. Here 0.1 mg of the eluted peptides was transferred to new filtration units and incubated with 150 µg of each concanavalin A (ConA) and wheat germ agglutinin (WGA) in the buffer containing 1 mM $CaCl_2$, 1 mM $MnCl_2$, and 0.5 M NaCl in 20 mM Tris–HCl (pH 7.3) for 1 h. The unbound peptides were eluted by centrifugation. The bound peptides were washed four times with 200 µl of binding buffer and twice with $H_2{}^{18}O$. The captured fractions were deglycosylated by PNGase F in $H_2{}^{18}O$ as described previously [16]. With the Microcon filter, we identified 959 and 792 glycosylation sites in two independent experiments (see Supplementary Table 2 in supplementary material). The analyses with Vivacon filters identified only 400 and 335 sites in these experiments (Fig. 1E). Combining both filters by using Microcon for protein digestion and Vivacon for the N-glyco enrichment, we identified 886 and 843 sites in these two experiments. This result suggests that the Vivacon 30 k filter has tighter pores and prevents elution of some of the glycosylated peptides.

During one period of the development and application of the N-glyco FASP, we observed that the filtration properties of the Microcon 30 k ultrafiltration units were variable between batches. Apparently, some batches had larger than usual pores that led to the elution of trypsin and GluC in the peptide mixture and impaired the N-glyco FASP analysis. As a result, we tested each batch of the Microcon 30 k units for their performance using a standard mouse liver lysate before using them in experiments.

The strength of the FASP procedure is its ability to remove SDS from protein lysates in a straightforward way. To demonstrate this, we loaded 100 µl of the 2% SDS solution into ultrafiltration devices and concentrated the solution to less than 5 µl. The retentate was diluted with 100 µl of buffer without SDS, and the detergent concentration was measured according to Rusconi and coworkers [20]. Because micelles of SDS are relatively small (~18,000–40,000 Da) [21] and the critical micelle concentration (CMC) of this detergent is relatively high (1–10 mM) [22], SDS easily passes the 30 k ultrafiltration membrane and two washes with 8 M urea were sufficient for quantitative depletion of the detergent (Fig. 2A). Washing of the filters with buffer solution without urea, as done

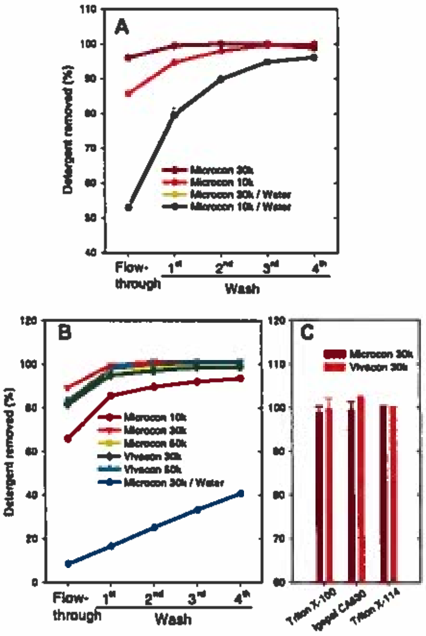


Fig.2. Detergent depletion. (A) Removal of 0.2% SDS using Microcon 10 and 30 k concentrators by repeated elution with 8 M urea or water. SDS was determined according to Ref. [20]. (B) Removal of 0.2% (w/v) solutions of Triton X-100 (Fluka, product No. 93427), Triton X-114 (Sigma–Aldrich, product No. X114), and Igepal CA630 (Nonidet P-40) (Sigma–Aldrich, product No. I3021) by repeated centrifugation-mediated elutions and sample dilutions. The concentration of the detergents was determined by absorbance measurement at 275 nm. (C) Extent of detergent removal after three washing steps.

by Manza and coworkers [4], is less effective and does not enable complete depletion of SDS (Fig. 2A). Therefore, that method is not useful for analysis of samples containing detergents. In contrast to SDS, detergents with large micelles and low CMCs, such as Triton X-100, Triton X-114, and Igepal CA630 (Nonidet P-40), could remain in the filter during FASP, resulting in inhibition of the digesting enzymes and/or generation of detergent-contaminated peptides. The Tritons and Igepal are frequently used for partial solubilization of cells and membrane fractionation at concentrations of 1 to 20 mM. Already at submillimolar concentrations, these detergents aggregate into micelles of 100,000 (100 k) molecular weight [23]. Therefore, we tested the retention of this group of detergents in the FASP procedure as we did for SDS. We loaded 100 μl of the 0.2% detergent solution (3.2 mM for Triton X-100) into ultrafiltration devices and measured their concentrations after the initial centrifugation and the washes with the buffer. Fig. 2B shows that Triton X-100 can be quantitatively removed from the 30 and 50 k filters by three or four elution steps, whereas the 10 k filters required additional elution steps for complete depletion of the detergent. Triton X-114 and Igepal CA630 behaved in the same way as Triton X-100. Three washing steps allowed complete removal of all these detergents (Fig. 2C). In the absence of urea, removal of Triton X-100 was very slow and in practice it cannot be removed quantitatively.

In conclusion, filters with larger nominal molecular weight cutoffs of 30 and 50 k, as opposed to the originally described 10 k ones, have advantages in peptide yield and sample preparation time. In addition, depletion of detergents with low CMCs is more efficient on filtration devices with larger cutoffs. We have shown that filters of both tested brands can be used in the FASP protocol with similar results. However, for N-glycoproteomic studies, the 30 k filters from Millipore are more suited than the Vivacon ones. We hope that this note will prove to be useful for the many researchers currently adopting the FASP protocol.

Acknowledgments

The work was supported by the Max Planck Society for the Advancement of Science, Munich Center for Integrated Protein Science (CIPSM), and PROSPECT, a Seventh Framework Program of the European Union (HEALTH-F4-2008-201648/PROSPECTS)

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2010.12.004.

References

[1] J.R. Wisniewski, A. Zougman, N. Nagaraj, M. Mann, Universal sample preparation method for proteome analysis, Nat. Methods 6 (2009) 359–362.
[2] M. Ethier, W. Hou, H.S. Duewel, D. Figeys, The proteomic reactor: a microfluidic device for processing minute amounts of protein prior to mass spectrometry analysis, J. Proteome Res. 5 (2006) 2754–2759.
[3] J. Vasilescu, D.R. Zweitzig, N.J. Denis, J.C. Smith, M. Ethier, D.S. Haines, D. Figeys, The proteomic reactor facilitates the analysis of affinity-purified proteins by mass spectrometry: application for identifying ubiquitinated proteins in human cells, J. Proteome Res. 6 (2007) 298–305.
[4] L.L. Manza, S.L. Stamer, A.J. Ham, S.G. Codreanu, D.C. Liebler, Sample preparation and digestion for proteomic analyses using spin filters, Proteomics 5 (2005) 1742–1745.
[5] R. Tian, S. Wang, F. Elisma, L. Li, H. Zhou, L. Wang, D. Figeys, Rare cell proteomic reactor applied to SILAC based quantitative proteomic study of human embryonic stem cell differentiation, Mol. Cell. Proteomics, in press.
[6] N. Nagaraj, A. Lu, M. Mann, J.R. Wisniewski, Detergent-based but gel-free method allows identification of several hundred membrane proteins in single LC–MS runs, J. Proteome Res. 7 (2008) 5028–5032.
[7] C.G. Spruijt, S.J. Bartels, A.B. Brinkman, J.V. Tjeertes, I. Poser, H.G. Stunnenberg, M. Vermeulen, CDK2AP1/DOC-1 is a bona fide subunit of the Mi-2/NuRD complex, Mol. Biosyst. 6 (2010) 1700–1706.
[8] D.E. Gordon, L.M. Bond, D.A. Sahlender, A.A. Peden, A targeted siRNA screen to identify SNAREs required for constitutive secretion in mammalian cells, Traffic 11 (2010) 1191–1204.
[9] J.R. Wisniewski, A. Zougman, M. Mann, Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome, J. Proteome Res. 8 (2009) 5674–5678.
[10] T. Geiger, J. Cox, P. Ostasiewicz, J.R. Wisniewski, M. Mann, Super-SILAC mix for quantitative proteomics of human tumor tissue, Nat. Methods 7 (2010) 383–385.
[11] M.P. Weekes, R. Antrobus, J.R. Lill, L.M. Duncan, S. Hor, P.J. Lehner, Comparative analysis of techniques to purify plasma membrane proteins, J. Biomol. Tech. 21 (2010) 108–115.
[12] D.F. Zielinska, F. Gnad, M. Jedrusik-Bode, J.R. Wisniewski, M. Mann, *Caenorhabditis elegans* has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins, J. Proteome Res. 8 (2009) 4039–4049.
[13] B. Soufi, C.D. Kelstrup, G. Stoehr, F. Frohlich, T.C. Walther, J.V. Olsen, Global analysis of the yeast osmotic stress response by quantitative proteomics, Mol. Biosyst. 5 (2009) 1337–1346.
[14] P. Ostasiewicz, D.F. Zielinska, M. Mann, J.R. Wisniewski, Proteome, phosphoproteome and *N*-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry, J. Proteome Res. 9 (2010) 3688–3700.
[15] J.R. Wisniewski, N. Nagaraj, A. Zougman, F. Gnad, M. Mann, Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology, J. Proteome Res. 9 (2010) 3280–3289.
[16] D.F. Zielinska, F. Gnad, J.R. Wisniewski, M. Mann, Precision mapping of an in vivo *N*-glycoproteome reveals rigid topological and sequence constraints, Cell 141 (2010) 897–907.
[17] J. Rappsilber, Y. Ishihama, M. Mann, Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics, Anal. Chem. 75 (2003) 663–670.
[18] P.A. Nielsen, J.V. Olsen, A.V. Podtelejnikov, J.R. Andersen, M. Mann, J.R. Wisniewski, Proteomic mapping of brain plasma membrane proteins, Mol. Cell. Proteomics 4 (2005) 402–408.
[19] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies, and proteome-wide protein quantification, Nat. Biotechnol. 26 (2008) 1367–1372.
[20] F. Rusconi, E. Valton, R. Nguyen, E. Dufourc, Quantification of sodium dodecyl sulfate in microliter-volume biochemical samples by visible light spectroscopy, Anal. Biochem. 295 (2001) 31–37.
[21] K.J. Mysels, L.H. Princen, Light scattering by some lauryl sulfate solution, J. Phys. Chem. 83 (1959) 1699–1700.
[22] A. Chattopadhyay, E. London, Fluorimetric determination of critical micelle concentration avoiding interference from detergent charge, Anal. Biochem. 139 (1984) 408–412.
[23] R.J. Robson, E.A. Dennis, The size, shape, and hydration of nonionic surfactant micelles: Triton X-100, J. Phys. Chem. 81 (1977) 1075–1078.

CONCLUDING REMARKS AND PERSPECTIVES

The aim of this PhD study was to develop a simple and straightforward yet powerful N-glycopeptide capture method and to apply it to biological and biomedical questions that involve N-glycosylation. When taking into account relatively modest success but great focus over the past years of several research groups including those exclusively dedicated to N-glycosylation, this was a very challenging PhD project. N-glycosylation has been extensively studied on a case by case basis but, mainly due to a lack of suitable methods, it has until now remained largely unexplored at a global scale. The N-glyco-FASP method developed here allows efficient capture of glycopeptides even from membrane proteins. Since it enriches glycopeptides by simply adding a lectin mixture to the top of a filtration unit, no homogenous phase affinity columns are required. This technique can easily be applied to cell culture, tissues or even whole organisms and allows identification of over 2,000 N-glycosylation sites from a minute amount of protein starting material within one day. This is nearly as many as all the glycosylation sites in eukaryotes experimentally derived over the last decades. Our method is compatible with formalin-fixed paraffin-embedded material, which enables N-glycoproteomic investigation of the vast banks of clinical samples available in many pathology laboratories.

We have successfully applied N-glyco-FASP to precisely map the N-glycoproteomes of seven eukaryotic model organisms and to study N-glycosylation changes in colorectal cancer. Altogether we identified nearly 18,000 different N-glycosylation sites. This large volume of input data enabled detailed biological, evolutionary and biomedical characterization.

In summary, we provide a method that can now be applied to elucidate glycoproteins even from very low abundance clinical material and thereby has a potential to be used in biomarker discovery. Our dataset allows general insights into N-glycosylation in evolutionary and cancer contexts and contains a catalogue of novel glycosylation sites on many biologically important and disease-related proteins that can now be used by the respective communities for further functional and clinical studies. Since it is well known that most of the clinical biomarkers are glycoproteins, I hope that our method will become a standard technique used for biomarker discovery and that our work will contribute to a better understanding of the role of N-glycosylation in diseases and our lives in general.

REFERENCES

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198-207.

Alvarez-Manilla, G., Atwood, J., 3rd, Guo, Y., Warren, N.L., Orlando, R., and Pierce, M. (2006). Tools for glycoproteomic analysis: size exclusion chromatography facilitates identification of tryptic glycopeptides with N-linked glycosylation sites. J Proteome Res *5*, 701-708.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res *37*, D396-403.

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.

Bunkenborg, J., Pilch, B.J., Podtelejnikov, A.V., and Wisniewski, J.R. (2004). Screening for N-glycosylated proteins by liquid chromatography mass spectrometry. Proteomics *4*, 454-465.

Chong, B.E., Yan, F., Lubman, D.M., and Miller, F.R. (2001). Chromatofocusing nonporous reversed-phase high-performance liquid chromatography/electrospray ionization time-of-flight mass spectrometry of proteins from human breast cancer whole cell lysates: a novel two-dimensional liquid chromatography/mass spectrometry method. Rapid Commun Mass Spectrom *15*, 291-296.

Comer, F.I., and Hart, G.W. (2000). O-Glycosylation of nuclear and cytosolic proteins. Dynamic interplay between O-GlcNAc and O-phosphate. J Biol Chem *275*, 29179-29182.

Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., and Sanchez, J.C. (2000). The dynamic range of protein expression: a challenge for proteomic research. Electrophoresis *21*, 1104-1115.

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.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda - a peptide search engine integrated into the MaxQuant environment. J Proteome Res.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4*, P3.

Domon, B., and Aebersold, R. (2006). Mass spectrometry and protein analysis. Science *312*, 212-217.

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.

Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R., and Mann, M. (2010). Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat Methods *7*, 383-385.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol *8*, R250.

Gundry, R.L., Raginski, K., Tarasova, Y., Tchernyshyov, I., Bausch-Fluck, D., Elliott, S.T., Boheler, K.R., Van Eyk, J.E., and Wollscheid, B. (2009). The mouse C2C12 myoblast cell surface N-linked glycoproteome: Identification, glycosite occupancy, and membrane orientation. Mol Cell Proteomics.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol *17*, 994-999.

Han, X., Aslanian, A., and Yates, J.R., 3rd (2008). Mass spectrometry for proteomics. Curr Opin Chem Biol *12*, 483-490.

Hann, S.R. (2006). Role of post-translational modifications in regulating c-Myc proteolysis, transcriptional activity and biological function. Seminars in cancer biology *16*, 288-302.

Hardman, M., and Makarov, A.A. (2003). Interfacing the orbitrap mass analyzer to an electrospray ion source. Anal Chem *75*, 1699-1705.

Hastie, T., Tibshirani, R. Friedman, J. (2001). The Elements of Statistical Learning
Data Mining, Inference, and Prediction, Second Edition.

Horth, P., Miller, C.A., Preckel, T., and Wenz, C. (2006). Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. Mol Cell Proteomics *5*, 1968-1974.

Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005). The Orbitrap: a new mass spectrometer. J Mass Spectrom *40*, 430-443.

Hubner, N.C., Ren, S., and Mann, M. (2008). Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. Proteomics *8*, 4862-4872.

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.

Jensen, O.N. (2004). Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Biol *8*, 33-41.

Jensen, O.N. (2006). Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol *7*, 391-403.

Kaji, H., Kamiie, J., Kawakami, H., Kido, K., Yamauchi, Y., Shinkawa, T., Taoka, M., Takahashi, N., and Isobe, T. (2007). Proteomics reveals N-linked glycoprotein diversity in Caenorhabditis elegans and suggests an atypical translocation mechanism for integral membrane proteins. Mol Cell Proteomics *6*, 2100-2109.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.

Kim, S.C., Sprung, R., Chen, Y., Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L., *et al.* (2006). Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. Mol Cell *23*, 607-618.

Krishna, R.G., and Wold, F. (1993). Post-translational modification of proteins. Adv Enzymol Relat Areas Mol Biol *67*, 265-298.

Kruger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C.A., Forner, F., Schmidt, S., Zanivan, S., Fassler, R., and Mann, M. (2008). SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. Cell *134*, 353-364.

Kumar, C., and Mann, M. (2009). Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Lett *583*, 1703-1712.

Kuster, B., and Mann, M. (1999). 18O-labeling of N-glycosylation sites to improve the identification of gel-separated glycoproteins using peptide mass mapping and database searching. Anal Chem *71*, 1431-1440.

Larsen, M.R., Hojrup, P., and Roepstorff, P. (2005). Characterization of gel-separated glycoproteins using two-step proteolytic digestion combined with sequential microcolumns and mass spectrometry. Mol Cell Proteomics *4*, 107-119.

Lee, A., Kolarich, D., Haynes, P.A., Jensen, P.H., Baker, M.S., and Packer, N.H. (2009). Rat liver membrane glycoproteome: enrichment by phase partitioning and glycoprotein capture. Journal of proteome research *8*, 770-781.

Liu, F., Zaidi, T., Iqbal, K., Grundke-Iqbal, I., Merkle, R.K., and Gong, C.X. (2002). Role of glycosylation in hyperphosphorylation of tau in Alzheimer's disease. FEBS letters *512*, 101-106.

Liu, T., Qian, W.J., Gritsenko, M.A., Camp, D.G., 2nd, Monroe, M.E., Moore, R.J., and Smith, R.D. (2005). Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. J Proteome Res *4*, 2070-2080.

Liu, X., McNally, D.J., Nothaft, H., Szymanski, C.M., Brisson, J.R., and Li, J. (2006). Mass spectrometry-based glycomics strategy for exploring N-linked glycosylation in eukaryotes and bacteria. Anal Chem *78*, 6081-6087.

Lu, A., Wisniewski, J.R., and Mann, M. (2009). Comparative proteomic profiling of membrane proteins in rat cerebellum, spinal cord, and sciatic nerve. Journal of proteome research *8*, 2418-2425.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448-3449.

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.

Medzihradszky, K.F. (2005). Characterization of protein N-glycosylation. Methods Enzymol *405*, 116-138.

Meunier, B., Dumas, E., Piec, I., Bechet, D., Hebraud, M., and Hocquette, J.F. (2007). Assessment of hierarchical clustering methodologies for proteomic data mining. J Proteome Res *6*, 358-366.

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

Nalivaeva, N.N., and Turner, A.J. (2001). Post-translational modifications of proteins: acetylcholinesterase as a model system. Proteomics *1*, 735-747.

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. The Journal of biological chemistry *250*, 4007-4021.

Oda, Y., Huang, K., Cross, F.R., Cowburn, D., and Chait, B.T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. Proc Natl Acad Sci U S A *96*, 6591-6596.

Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell *127*, 635-648.

Olsen, J.V., Schwartz, J.C., Griep-Raming, J., Nielsen, M.L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., *et al.* (2009). A dual pressure linear ion trap - Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics.

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.

Pan, S., and Aebersold, R. (2007). Quantitative proteomics by stable isotope labeling and mass spectrometry. Methods Mol Biol *367*, 209-218.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551-3567.

Ramachandran, P., Boontheung, P., Xie, Y., Sondej, M., Wong, D.T., and Loo, J.A. (2006). Identification of N-linked glycoproteins in human saliva by glycoprotein capture and mass spectrometry. Journal of proteome research *5*, 1493-1503.

Reynolds, K.J., Yao, X., and Fenselau, C. (2002). Proteolytic 18O labeling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. Journal of proteome research *1*, 27-33.

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., *et al.* (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics *3*, 1154-1169.

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

Schiess, R., Mueller, L.N., Schmidt, A., Mueller, M., Wollscheid, B., and Aebersold, R. (2009). Analysis of cell surface proteome changes via label-free, quantitative mass spectrometry. Mol Cell Proteomics *8*, 624-638.

Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat Biotechnol *23*, 1391-1398.

Scott, N.E., Bogema, D.R., Connolly, A.M., Falconer, L., Djordjevic, S.P., and Cordwell, S.J. (2009). Mass spectrometric characterization of the surface-associated 42 kDa lipoprotein JlpA as a glycosylated antigen in strains of Campylobacter jejuni. J Proteome Res *8*, 4654-4664.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Sharma, S., Simpson, D.C., Tolic, N., Jaitly, N., Mayampurath, A.M., Smith, R.D., and Pasa-Tolic, L. (2007). Proteomic profiling of intact proteins using WAX-RPLC 2-D separations and FTICR mass spectrometry. Journal of proteome research *6*, 602-610.

Shevchenko, A., Wilm, M., Vorm, O., Jensen, O.N., Podtelejnikov, A.V., Neubauer, G., Shevchenko, A., Mortensen, P., and Mann, M. (1996). A strategy for identifying gel-separated proteins in sequence databases by MS alone. Biochem Soc Trans *24*, 893-896.

Siuti, N., and Kelleher, N.L. (2007). Decoding protein modifications using top-down mass spectrometry. Nat Methods *4*, 817-821.

Sparbier, K., Koch, S., Kessler, I., Wenzel, T., and Kostrzewa, M. (2005). Selective isolation of glycoproteins and glycopeptides for MALDI-TOF MS detection supported by magnetic particles. J Biomol Tech *16*, 407-413.

Suzuki, T., Kitajima, K., Inoue, S., and Inoue, Y. (1995). N-glycosylation/deglycosylation as a mechanism for the post-translational modification/remodification of proteins. Glycoconj J *12*, 183-193.

Syka, J.E., Marto, J.A., Bai, D.L., Horning, S., Senko, M.W., Schwartz, J.C., Ueberheide, B., Garcia, B., Busby, S., Muratore, T., *et al.* (2004). Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. J Proteome Res *3*, 621-626.

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.

Van den Steen, P., Rudd, P.M., Dwek, R.A., and Opdenakker, G. (1998). Concepts and principles of O-linked glycosylation. Critical reviews in biochemistry and molecular biology *33*, 151-208.

Varki, A. (1993). Biological roles of oligosaccharides: all of the theories are correct. Glycobiology *3*, 97-130.

Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W., and Etzler, M.E. (2008). Essentials of Glycobiology (New York: Cold Spring Harbor Laboratory Press).

Wacker, M., Linton, D., Hitchen, P.G., Nita-Lazar, M., Haslam, S.M., North, S.J., Panico, M., Morris, H.R., Dell, A., Wren, B.W., *et al.* (2002). N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. Science *298*, 1790-1793.

Wada, H., Sekine, Y., Yoshida, S., Yasufuku, K., Iyoda, A., Iizasa, T., Saitoh, Y., Fujisawa, T., Shibuya, K., Hiroshima, K., *et al.* (2004). Dramatic improvement of respiratory condition after lobectomy for localized bullous emphysema. Ann Thorac Cardiovasc Surg *10*, 293-296.

Wagner, M., Adamczak, R., Porollo, A., and Meller, J. (2005). Linear regression models for solvent accessibility prediction in proteins. J Comput Biol *12*, 355-369.

Washburn, M.P., Wolters, D., and Yates, J.R., 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol *19*, 242-247.

Wisniewski, J.R., Nagaraj, N., Zougman, A., Gnad, F., and Mann, M. (2010). Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. J Proteome Res *9*, 3280-3289.

Wisniewski, J.R., Ostasiewicz, P., and Mann, M. (2011). High Recovery FASP Applied to the Proteomic Analysis of Microdissected Formalin Fixed Paraffin Embedded Cancer Tissues Retrieves Known Colon Cancer Markers. J Proteome Res.

Wisniewski, J.R., Zougman, A., and Mann, M. (2009a). Combination of FASP and StageTip-based Fractionation Allows In-Depth Analysis of the Hippocampal Membrane Proteome. J Proteome Res.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009b). Universal sample preparation method for proteome analysis. Nat Methods *6*, 359-362.

Witze, E.S., Old, W.M., Resing, K.A., and Ahn, N.G. (2007). Mapping protein post-translational modifications with mass spectrometry. Nat Methods *4*, 798-806.

Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., Aebersold, R., and Watts, J.D. (2009). Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. Nature biotechnology *27*, 378-386.

Woods, R.J., Edge, C.J., and Dwek, R.A. (1994). Protein surface oligosaccharides and protein function. Nat Struct Biol *1*, 499-501.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res *34*, D187-191.

Wuhrer, M., Catalina, M.I., Deelder, A.M., and Hokke, C.H. (2007). Glycoproteomics based on tandem mass spectrometry of glycopeptides. Journal of chromatography *849*, 115-128.

Wuhrer, M., Deelder, A.M., and Hokke, C.H. (2005). Protein glycosylation analysis by liquid chromatography-mass spectrometry. Journal of chromatography *825*, 124-133.

Yang, X.J. (2005). Multisite protein modification and intramolecular signaling. Oncogene *24*, 1653-1662.

Young, N.M., Brisson, J.R., Kelly, J., Watson, D.C., Tessier, L., Lanthier, P.H., Jarrell, H.C., Cadotte, N., St Michael, F., Aberg, E., *et al.* (2002). Structure of the N-linked glycan present on multiple glycoproteins in the Gram-negative bacterium, Campylobacter jejuni. J Biol Chem *277*, 42530-42539.

Zhang, H., Li, X.J., Martin, D.B., and Aebersold, R. (2003). Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nature biotechnology *21*, 660-666.

Zhou, F., and Johnston, M.V. (2005). Protein profiling by capillary isoelectric focusing, reversed-phase liquid chromatography, and mass spectrometry. Electrophoresis *26*, 1383-1388.

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

ACKNOWLEDGMENTS

Matthias Mann, my doctoral advisor. Thank you for accepting me as a PhD student in the Proteomics and Signal Transduction Group, for the challenging projects and support. I really appreciate everything you have done for me and I am really happy that I had the opportunity to be a member of your group and learn from you. I cannot imagine a better place to do a PhD.

Jacek Wisniewski, my supervisor. Thank you for all your good ideas and help throughout my PhD time. It was great to have you as a supervisor. I very much enjoyed working with you. I had the best projects under the best supervision in the whole world ☺

Florian Gnad, Danke für die tolle Zusammenarbeit und die tolle Zeit, die wir zusammen erlebt haben. Ich habe viel von dir gelernt und ich werde dich nie vergessen.

Other people from the lab, for help and nice atmosphere. Especially, Korbi, Naga and Sasha for help with mass spectrometry. A big thanks to Alison and Theresa for taking care of all organisational matters. Richard, for introducing me into R. And finally thanks to Pawel, Kamila and Mario for a really nice time at work. I really enjoyed spending time with you.

Friedrich Altmann and Kalle Hult, My Bachelor and Master thesis supervisors. Thank you for introducing me in to protein research and developing my interest in the field of protein science.

Kasia, Kuba and Markus, my best, best, best friends. I cannot imagine my life without you! I don't know who I would be now, if I wouldn't have met you. You are part of my life and I hope you all know how much I love you!

My whole family, Dziekuje babciu za milosc i opieke i dobre jedzenie, ktore sprawialo ze odrazu nabieralam koloru na twarzy☺. Ania, za wszystkie chwile spedzone razem, ktore do dzis milo wspominam. Wiedz, ze zawsze mozesz na mnie liczyc, i ze kocham cie z calego serca!!! Teresa and Chris for all the nice presents, holidays and help with English. I am very grateful for everything you have done for me.

Mamo i tato, Dziekuje za wasze wsparcie i milosc przez cale moje zycie. Nic nie jest w stanie wyrazic mojej wdziecznosci i radosci ze mam tak wspanialych rodzicow. Wam zawdzieczam ta prace, gdyz to wy mnie wspieraliscie od mlodosci w rozwiazywaniu zadan logicznych, wzbudziliscie we mnie ambicje, ukierunkowaliscie w wyborze studiow.... Bez was nie bylabym teraz nawet w jednej setnej ta osoba ktora jestem teraz. Kocham was z calego serca!

APPENDIX: PhD publications not related to N-glycosylation

Appendix I

Zielinska DF, Gnad F, Jedrusik-Bode M, Wiśniewski JR, Mann M (2009), *Caenorhabditis elegans Has a Phosphoproteome Atypical for Metazoans That Is Enriched in Developmental and Sex Dermination Proteins*, **J Proteome Res**, Aug 7; 8(8) 4039-49.

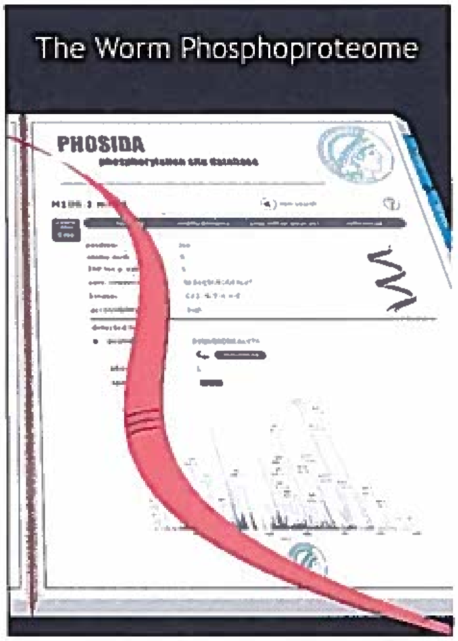
Appendix II

Gnad F, Forner F, **Zielinska DF**, Birney E, Gunawardena J, Mann M (2010), *Evolutionary Constraints of Phosphorylation in Eukaryotes, Prokaryotes and Mitochondria*, **Mol Cell Proteomics**, Dec 9; 9(12) 2642-53.

Appendix I

Caenorhabditis elegans Has a Phosphoproteome Atypical for Metazoans That Is Enriched in Developmental and Sex Dermination Proteins

Zielinska DF, Gnad F, Jedrusik-Bode M, Wiśniewski JR, Mann M



J Proteome Res, 2009

Here we describe a high accuracy *C. elegans* phosphoproteome of almost 7,000 sites, 99% of which are novel. We find distinct characteristics when comparing to phosphoproteomes of other multicellular eukaryotes.

research articles Journal of proteome research

Caenorhabditis elegans Has a Phosphoproteome Atypical for Metazoans That Is Enriched in Developmental and Sex Determination Proteins

Dorota F. Zielinska,[†,#] Florian Gnad,[‡,#] Monika Jedrusik-Bode,[§] Jacek R. Wiśniewski,[*,†] and Matthias Mann[*,†]

Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany, Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, and Laboratory of Chromatin Biochemistry, Max-Planck-Institute for Biophysical Chemistry, Am Faßberg 11, D-37077 Göttingen, Germany

Received May 1, 2009

In eukaryotic species, signal transduction is often mediated by posttranslational modifications that can serve as regulatory switches. Although nematodes have usually been studied by genetic rather than biochemical methods, PTMs such as phosphorylation are thought to control all aspects of biological functions including sex determination and development. Here, we apply high accuracy mass spectrometry and comprehensive bioinformatic analysis to determine and characterize the *in vivo Caenorhabditis elegans* phosphoproteome for the first time. We detect nearly 7000 phosphorylation sites on 2400 proteins, which are disproportionately involved in development and sex determination. Interestingly, the worm phosphoproteome turns out to be very distinct from phosphoproteomes of other multicellular eukaryotes as judged by its phylogenetic conservation, kinase substrate motifs and site analysis by a support vector machine. This result agrees with the large proportion of worm specific kinases previously discovered by genome sequencing. Furthermore, our data show that the *C. elegans* specific dosage complex can be phosphorylated on most subunits, suggesting its regulation by kinases. Availability of the *C. elegans* phosphoproteome should add a novel dimension to functional data obtained by genetic screens in this organism.

Keywords: *C. elegans* • phosphorylation • mass spectrometry • bioinformatics

Introduction

Protein phosphorylation plays a crucial role in the regulation of cellular signal transduction processes and is the most widely studied type of post-translational modification in eukaryotes.[1−4] Recent advances in mass spectrometry allow the high accuracy identification of phosphoproteomes of various eukaryotic and prokaryotic organisms in a large-scale manner.[5,6] Phosphoproteomes of yeast, fly, mouse, and human have already been reported and analyzed.[7−10] *Caenorhabditis elegans* is the only major model organism whose phosphoproteome has not been characterized yet. *C. elegans* is a small, approximately 1 mm long, free-living, round worm which has proven an ideal system to study gene regulation and function due to its short life cycle, well-defined developmental pattern of its 959 somatic cells and relatively small genome (3×10^9 bp), which was the first metazoan genome to be fully sequenced.[11] Worms employ many different reproductive strategies, often related to their parasitic life histories, and can be either gonochoristic (male/female) or hermaphroditic (self-fertilizing female). In the laboratory, *C. elegans* is usually grown as self-fertilizing hermaphrodite. The essential signal for sex determination in *C. elegans* is the ratio of X chromosomes to autosomes (X/A): hermaphrodites have two X chromosomes, whereas males only have one. Thus, the control of expression of X chromosome linked genes is crucial for sex determination. The dosage compensation complex, which equalizes the gene expression levels between sexes by lowering the transcription levels of genes on the X chromosome, is therefore a key system in *C. elegans.*

In favorable conditions, nematode eggs go through four larval stages in three days, and then they live for another 2−3 weeks.[12] In unfavorable conditions, worms adopt an alternative life form, termed the dauer larva stage. This stage is specialized in structure and behavior for long-term survival and dispersal.[13] Apart from its interest to the developmental community, *C. elegans* is the organism in which the seminal discoveries of specific longevity mutants have been made and it continues to be a prominent model for aging.[14] Recently, the worm proteome has been measured using a shotgun proteomics approach.[15,16]

Here, we analyze the phosphoproteome of *C. elegans* using high accuracy MS in combination with chromatographic

---

[*] To whom correspondence should be addressed: Matthias Mann (mmann@biochem.mpg.de) or Jacek R. Wiśniewski (jwisniew@biochem.mpg.de).

[†] Max-Planck-Institute for Biochemistry.

[#] These authors contributed equally to this work.

[‡] Harvard Medical School.

[§] Max-Planck-Institute for Biophysical Chemistry.

10.1021/pr900384k CCC: $40.75 © 2009 American Chemical Society

Published on Web 06/16/2009

fractionation, Filter Aided Sample Preparation (FASP) based protein digestion,[17] and phosphopeptide enrichment. We found close to 7000 phosphorylation sites on 2400 proteins and investigated the identified worm phosphoproteome using comprehensive bioinformatic data analysis. We explored main features of the biological role of phosphorylation in the worm through gene ontology annotations. Furthermore, we elucidated structural constraints and evolutionary conservation of phosphorylation sites.

A recent evolutionary study on the phosphoproteomes of yeast, fly, mouse, and human has shown that the phosphosites of single cell eukaryotes are not significantly conserved in higher eukaryotes.[18] In contrast, phosphosites of human, mouse, and also fly are significantly more conserved than their nonphosphorylated counterparts throughout higher eukaryotes. This suggests that many phosphorylation patterns evolved after the speciation event that separated yeast from higher eukaryotes. These observations are in concordance with the existence of large kinase families that are specific to either single cellular or multicellular species.[19]

The collection of worm kinases, the worm kinome, differs from the ones of other multicellular eukaryotes to a high degree. There is evidence for a dramatic expansion of worm specific kinases that resulted in a kinome tree that is nearly twice as large as the one of fly.[19] Therefore, there may be nematode specific phosphorylation events and we were interested to confirm this hypothesis on the basis of a comprehensive evolutionary study of our worm phosphoproteome along with consensus sequence motif analysis. Indeed, bioinformatic analysis of our *C. elegans* phosphoproteome provides direct evidence that the phosphoproteome is very different from that of other metazoan model organisms.

Materials and Methods

Culturing of *C. elegans*. The samples of synchronized adult wild-type worms were prepared by growing the worms in liquid culture at 20 °C using concentrated *Escherichia coli* OP50 as the food source.[20] The worm culture was monitored, and additional *E. coli* OP50 was added when needed. The adult worms were washed in M9 buffer [22 mM KH_2PO_4, 42 mM Na_2HPO_4, 0.1 M NaCl, 1 mM $MgSO_4$ in H_2O] and frozen in liquid nitrogen.

Protein Extraction. A total 240 mg of frozen worms (estimated protein amount 6 mg) was lysed in 1 mL of lysis buffer (0.1 M Tris-HCl, pH 7.6, 0.125 M DTT containing "Complete, EDTA free" Protease Inhibitor Cocktail Tablet (Roche) and "PhosSTOP" Phosphatase Inhibitor cocktail Tablet (Roche)) using Branson SONIFIER 250 (G-HEINEMANN Ultraschall- and Labortechnik, Germany) for 3 min on ice (output control 5; duty cycle 20%). A total of 250 μL of 20% (w/v) SDS in 0.1 M Tris-HCl, pH 7.6, was added to the suspension and the mixture was incubated for 3 min at 95 °C. The crude extract was then clarified by centrifugation at 16 000*g* at 18 °C for 10 min. Subsequently, 36 mg of iodoacetamide was added to the supernatant, and after 30 min incubation, the sample was centrifuged at 16 000*g* at 18 °C for 10 min. The protein content was determined using Cary Eclipse Fluorescence Spectrometer (Varian) as described previously.[21] Briefly, 1–2 μL of sample or tryptophan standard was mixed with 2 mL of 8 M urea in 10 mM Tris-HCl pH 7.6. Fluorescence was measured at 295 nm for excitation and 350 nm for emission. The slits were set to 10 nm.

Protein Fractionation. The protein extract was concentrated on Microcon filters YM-30 (Millipore) to 200 μL. A total of 100 μL, containing approximately 2 mg of protein, was loaded onto a Superdex 200 10/300GL column (GE Bioscience). The proteins were eluted using 25 mM Tris-HCl, pH 8.0, 0.1 M NaCl, and 0.2% (w/v) SDS, and 400 μL fractions were collected. Fractions containing proteins were pooled together into 5 fractions of similar protein content and concentrated on Microcon filters YM-30 to 30 μL.

Protein Digestion. With the use of the FASP protocol,[17] the proteins were subjected to trypsin digestion. Briefly, to YM-30 filter units containing protein concentrates, 200 μL of 8 M urea in 0.1 M Tris/HCl, pH 8.5, was added and the samples were centrifuged at 14 000*g* at 18 °C for 15 min. This step was repeated three times. Then, 100 μL of 40 mM NH_4HCO_3 was added to the filters and the samples were centrifuged for 10 min under the same conditions as before. This step was repeated once. Finally, 7 μg of trypsin (Promega) in 40 μL of 40 mM NH_4HCO_3 was added to each filter (resulting in approximately 1:100 (w/w) ratio of enzyme to protein). The samples were incubated overnight at 37 °C. Peptides were collected by centrifugation. The filters were rinsed with 50 μL of 40 mM NH_4HCO_3. Subsequently, the peptide solutions were acidified with 5 μL of 10% (v/v) CF_3COOH.

Phosphopeptide Enrichment. Acidified protein digest was enriched for phosphopeptides using a TiO_2 enrichment protocol[9,22] with slight modifications. Twenty-five milligrams of titansphere TiO_2 10 μm (GL Sciences, Inc., Japan) was suspended in 50 μL of 3% (m/v) DHB, 80% (v/v) CH_3CN, 0.1% CF_3COOH and diluted 1:4 with Milli-Q water before use. Ten microliters of this TiO_2 slurry was added to each sample and incubated under continuous agitation for 20 min. Then, the titanium beads were sedimented by centrifugation at 5000*g* for 1 min and the supernatants were collected and mixed with another portion of the beads and incubated as above. The pellets were resuspended in 150 μL of 30% (v/v) CH_3CN containing 3% (v/v) CF_3COOH and were transferred to a 200 μL pipet tip plugged with one layer of Empore-C_8 filter. The beads were washed 3 times with 30% (v/v) CH_3CN, 3% CF_3COOH (v/v) solution and 3 times with 80% CH_3CN (v/v), 0.3% CF_3COOH (v/v) solution. Finally, the peptides were eluted from the beads with 100 μL of 40% CH_3CN (v/v) and 15% NH4OH (m/v) and vacuum-concentrated to 3 μL.

Mass Spectrometric Analysis. Peptide mixtures were analyzed on the LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Germany) coupled with HPLC via a nanoelectrospray ion source. Sample solutions were applied to a 15 cm fused silica emitter (Proxeon Biosystems, Denmark) packed in-house with the reverse phase ReproSil-Pur C18 -AQ, 3 μm resin (Dr. Maisch, GmbH). A 230 min gradient from 2% to 80% of 80% (v/v) CH_3CN, 0.5% (v/v) CH_3COOH was used to separate peptides. Lock-masses were used to improve accuracy[23] and MS scans were acquired with resolution of 60 000. Each full scan performed in Orbitrap was followed by MS fragmentation events in the LTQ for the eight most intense ions as described.[24] Multistage activation was enabled upon detection of a neutral loss of phosphoric acid (97.97, 48.88, or 32.66 Th) for further ion fragmentation.[25] Annotated spectra of all detected phosphopeptides are available via the Phosphorylation Site Database (www.phosida.com).

Data Analysis. The MS data were analyzed using MaxQuant, version 1.0.12.36. Proteins were identified by searching MS and MS/MS data of peptides against Wormbase 200 (containing

23 973 entries) combined with 175 common contaminants and concatenated with reversed versions of all sequences using the MASCOT search engine (Matrix Science, U.K.). Carbamidomethylation of cysteines was set as fixed modification and methionine oxidations, protein N-terminal acetylation, phosphorylation of serines, threonines, and tyrosines as variable modifications. Trypsin allowing for cleavage N-terminal to proline was chosen as enzyme specificity. A maximum of two miscleavages were allowed. The minimal peptide length was specified to be 6 amino acids. The initial maximal mass deviation of parental ions was set to 7 ppm, whereas for fragment ions it was set to 0.5 Da. The maximum peptide false discovery rate was set to 0.01.

The PTM score was used for assignment of the phosphorylation site(s) as described.[9] Briefly, the PTM localization score reflects the normalized probability that the phosphorylation is indeed localized at the specified amino acid position. It is calculated for all combinatorial phosphorylation site possibilities from the overlap of assigned b and y ions and observed matches. The algorithm is integrated into MaxQuant. Class I phosphorylation sites are defined by a localization probability of 0.75. Phosphorylation sites were made nonredundant with regards to their surrounding peptide sequence and all alternative proteins that match a particular phosphosite were reported as one group.

Gene Ontology Enrichment Analysis. We used Cytoscape[28] and BinGO[29] to derive biological functions, processes and cellular components that were significantly overrepresented in the phosphoset. Corresponding gene ontology annotations were derived from WormBase 200 (www.wormbase.org). The significance of overrepresented gene ontology annotations in the phosphoset compared to entire worm proteome was based on the hypergeometric model and the Benjamini Hochberg false discovery rate correction. A probability value of 0.0001 was considered significant.

Genome Annotation. Identified phosphorylated peptides were assigned to gene products annotated in the Ensembl database (http://www.ensembl.org) (Version 53.190).[30] Each peptide was assigned to a maximum of one gene product.[31] If a given phosphopeptide matched to more than one Ensembl database entry, it was assigned to the gene product that contained the highest number of matching phosphopeptides in total. With the Proserver Version 2.8 technology,[32] we established a PHOSIDA DAS source that enables Web users of the genome database Ensembl to retrieve genome or gene segments that encode phosphorylation sites in *C. elegans*.

Bootstrapping. We created bootstrap distributions from 10 000 sets of randomly selected *C. elegans* proteins annotated as "known" in Ensembl. Each random set contained as many proteins as the given phosphoset. We calculated the proportion of one-to-one orthologs in the chosen species and the histogram of these 10 000 proportions provided the bootstrap distribution.[18]

Next, we created a bootstrap distribution for a given species from 10 000 sets of serines, threonines, and tyrosines that were randomly selected from phosphorylated proteins that had one-to-one orthologs according to the Ensembl Compara annotation. Only those residues were included in the analysis that showed the same predicted structural constraints as phosphorylated residues (high accessibility and localization in loops) using SABLE 2.0.[33] For each set, we calculated the proportion of conserved residues in the chosen species and the histogram of these 10 000 proportions provided the bootstrap distribution. Resulting histograms were illustrated via MatLab (The MathWorks).

Phosphorylation Site Prediction. The SVM was trained on the primary sequence comprising the site and its 12 surrounding residues as described before.[34] Identified worm phosphorylation sites (5427 pS, 1230 pT, and 123 pY) made up the positive training sets. We randomly took sites from worm proteins that were annotated in WormBase Web site, http://www.wormbase.org, release WS200, and not reported to be phosphorylated, to create negative sets of the same size. The positive and negative sets were iteratively split into a training set (90%) and a test set (10%) to estimate the accuracy of prediction (5-fold cross-validation). We optimized the parameters and chose the Kernel function as reported before.[34]

Results

A *C. elegans* Phosphoproteome. Our *C. elegans* phosphoproteome study combines chromatographic protein separation, FASP based protein digestion, TiO_2-resin based phosphopeptide enrichment, and high accuracy mass spectrometry. We directly extracted proteins from *C. elegans* using 4% SDS and fractioned them by gel filtration into five fractions, which were concentrated and processed in 30k filtration units according to the FASP protocol.[17] The resulting peptides were enriched for phosphopeptides using TiO_2 affinity chromatography and were analyzed using high accuracy LTQ-Orbitrap mass spectrometry.

With this approach, we detected 6780 phosphorylation sites from 2373 proteins with an estimated false positive rate of less than 1%. We only considered class I phosphorylation sites (99% identification probability and localization probability to a single residue of at least 75%).[9] The average localization probability for class I phosphosites was 97%, which means that a majority of sites were detected with close to absolute certainty. Because of the very high and peptide dependent mass accuracy achieved by analysis in MaxQuant,[26] even peptides with relatively low PTM score reached 99% identification significance.[24] Nevertheless, overall, 86% of the corresponding phosphopeptides have Mascot identification scores higher than 14. All phosphorylation sites identified in our experiment are listed in Supplementary Table 1 and in Phosida. Around one-third of all identified peptides were phosphorylated. A majority of the identified phosphopeptides were singly phosphorylated, whereas only 20% of the detected peptides were multiply phosphorylated (Supplementary Figure 1). The proportion of phosphorylated serines, threonines, and tyrosines was 80%, 18.2%, and 1.8%, respectively (Supplementary Figure 2).

Protein kinases are one of the largest gene families and are estimated to phosphorylate at least one-third of the proteome. Even though around 80% of the human kinome have orthologs in *C. elegans*, around half of the worm kinome is nematode-specific.[19] The KinBase database (www.kinase.com/kinbase) contains kinases of several eukaryotic species including worm. Among the phosphorylated substrate proteins, we found 54 kinases (highlighted in Supplementary Table 1). One-third of the AGC kinases, including PKC and GRK, were phosphorylated, as were members of known phosphatase families such as IA2 and PP2A. As expected from the regulatory roles of the phosphoproteome, many phosphorylated worm proteins contain binding domains such as pleckstrin homology (PH), phosphotyrosine (PTB), and Src homology 2/3 (SH2/SH3) domains. For instance, the insulin receptor substrate (*ist-1*) homologue was phosphorylated on a serine at position 611

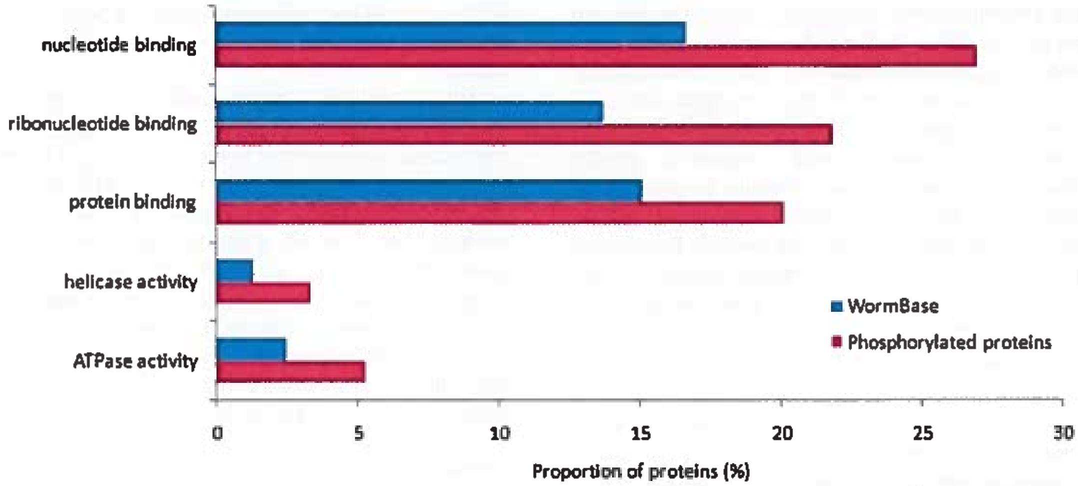


Figure 1. Biological functions that are significantly overrepresented in the phosphoset compared to the entire worm proteome according to gene ontology annotation. Cytoscape and BinGO were used for the enrichment analysis. Relative proportions of phosphoproteins with corresponding functions (according to gene ontology annotation) were compared with the proportions of all proteins that are annotated in the WormBase database.

(S611) that is located directly C-terminal to the pleckstrin homology domain. An example of a phosphorylated, uncharacterized, and nematode-specific protein is the one encoded by *C. elegans* gene Y37D8A.4 with phosphorylation on S102 directly N-terminal to the SH2 domain.

Using Cytoscape[28] and the BinGO plugin,[29] we derived protein functions and cellular component localizations that were overrepresented or underrepresented in the phosphoset according to GO (gene ontology) annotation (Supplementary Table 2, see also Materials and Methods). We found that many signaling related functions such as protein binding and ATPase activity were significantly enriched in the set of phosphoproteins compared to the entire Wormbase database (Figure 1). Additionally, transcriptional and translational regulator activities including helicase activity and DNA/RNA binding were overrepresented in the phosphoset.

Furthermore, identified phosphorylated proteins were significantly overrepresented in membranes, on the cytoskeleton, and in the nucleus (Supplementary Figure 3). In contrast, only a small subset of phosphoproteins was located in the extracellular region. This serves as a useful biological control because extracellular proteins are not usually controlled by phosphorylation and probably become substrates only during biogenesis. The enrichment of membrane proteins in our phosphoset was interesting because this is a compartment that is very much involved in cell signaling but which had been underrepresented in previous studies.[9]

To associate phenotypes from *C. elegans* genomic screens[35,36] with the identified phosphorylated proteins, we overlaid the phenotype analysis on our data using corresponding accession numbers from WormBase. We found that phosphorylation is present in all phenotype classes suggesting an important role of phosphorylation throughout all developmental stages of the worm (Supplementary Table 3).

Chromosome Localization of Genes That Encode Phosphoproteins. Genomic studies have found evidence for differences in gene type or function across the worm chromosome.[37−39] Chromosome arms excluding tips are more prone to recombination and mutation compared to chromosome centers. To compare the chromosome localization distribution of genes encoding phosphorylated proteins with the distribution of all worm genes, we assigned detected phosphopeptides to Ensembl entries as described in Gnad et al.[31] (see Materials and Methods). This gene assignment approach allowed the annotation of the worm genome in Ensembl via the Proserver technology.[32] Interestingly, the proportion of phosphoprotein encoding genes that were localized on the chromosome arms was somewhat higher than the one of all other genes in the case of four out of six chromosomes (II, IV, V, and X). Supplementary Figure 4 shows the difference between the distribution of genes encoding phosphoproteins and the one of all other genes localized on defined proportional ends of the sex chromosome. The investigated chromosome regions range from middle of the chromosome to 1% of the telomeres (Supplementary Table 4). Supplementary Figure 5 depicts the total number of all genes (left panel) and genes encoding phosphoproteins (right panel) on the six *C. elegans* chromosomes.

Distinct Characteristics of the Worm Phosphoproteome in Comparison to Higher Eukaryotes.

1. Derivation of Unknown Putative Kinase Motifs. We estimated the kinase correspondence of phosphorylated substrates by applying a sequence motif analysis as described in Zanivan et al.[40] We calculated the proportion of phosphorylated sites in worm that match with published motif sequences of human kinases[34] and compared it with the proportion of matching sites that one would expect by chance. We found that sequence motifs of kinase families which are highly conserved throughout eukaryotes matched significantly with sites that are phosphorylated in *C. elegans* (Supplementary Table 5). The PKA motif R-R/K-X-pS/T, for example, was enriched by a factor of 10 compared to the distribution that is expected by chance. Furthermore, sequence motifs of other conserved kinases such as AKT, CAMK, CDK, and ERK are also significantly enriched.

In addition, we applied a *de novo* method that identifies significantly overrepresented sequence motifs from large-scale data sets without any a priori knowledge.[41] The main purpose of this approach is to find potential motifs that correspond to worm specific kinases and therefore have not been reported yet. In addition to detected sequence motifs that matched with the ones of conserved eukaryotic kinases, we found consensus sequences (such as pS-X-X-X-X-P) that have not been reported to be kinase motifs and therefore present potential worm-

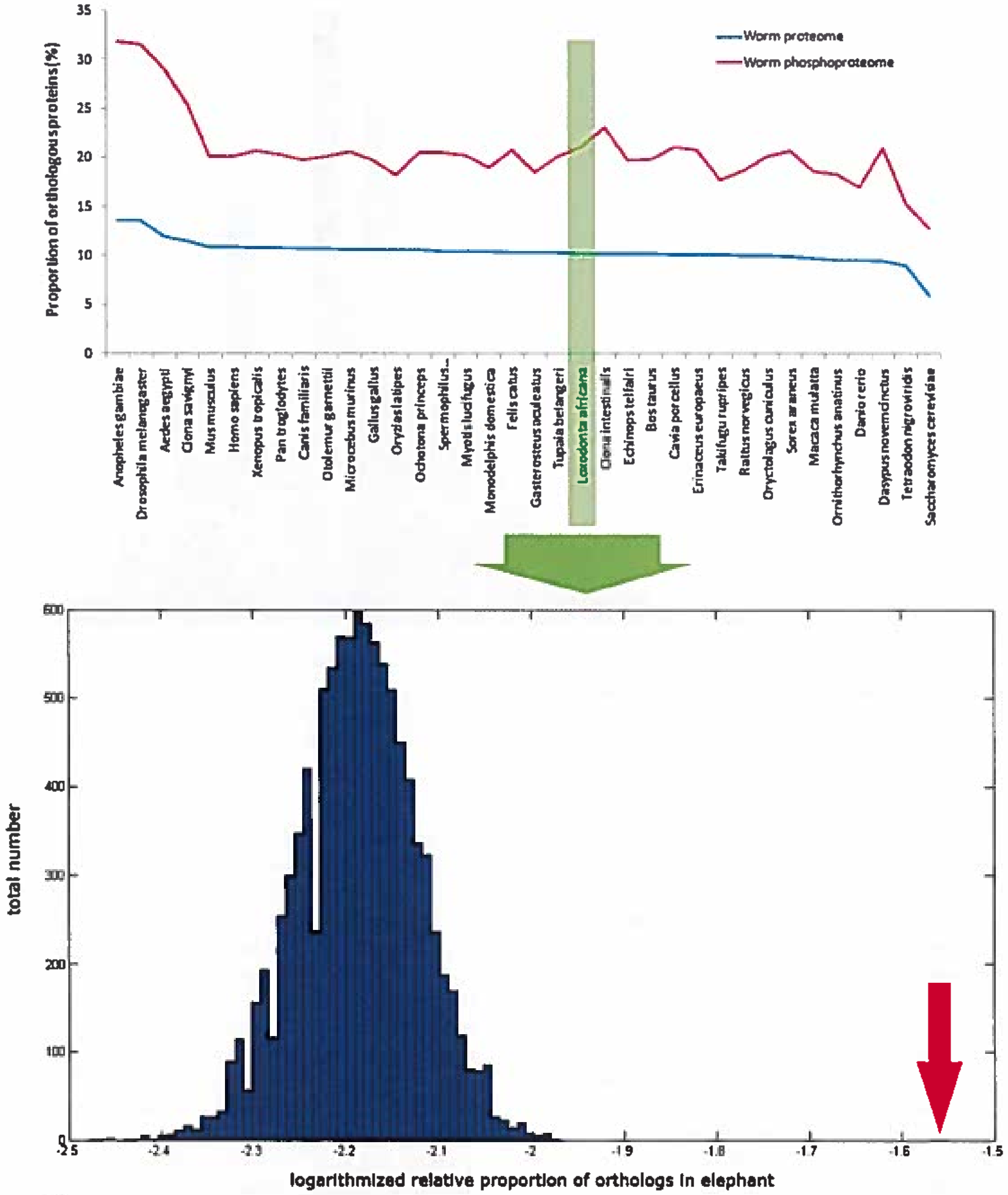


Figure 2. Proportion of worm genes that have orthologs in other eukaryotes. Genes that encode phosphorylated proteins have more orthologs than other worm genes (upper panel). The significance of the difference was confirmed on the basis of a bootstrapping approach (lower panel for elephant). The proportion of phosphogenes that have orthologs in elephant is marked in red.

specific kinase motifs (Supplementary Table 6). We applied the same approach to phosphorylation sites identified in fly using the same parameters. Strikingly, the number of significantly overrepresented unknown motifs in worm was 2× higher than that in fly.

2. Low Conservation of Worm Phosphosites in Higher Eukaryotes. With phosphorylated sites assigned to their gene products in hands, we next used the phylogenetic information provided by the comparative genome database Ensembl Compara to investigate the proportion of genes that have orthologs in other eukaryotic species. For each investigated species, worm genes that encode phosphoproteins had proportionally more orthologs than all other genes (Supplementary Table 7). Overall, 21% of phosphorylated worm proteins had orthologs in the elephant, for example, in comparison to 10% of all other proteins (Figure 2, upper panel). To confirm the statistical significance of this observation, we applied a bootstrapping approach as described before.[16] The resulting histograms reflecting the distribution of logarithmized proportions of orthologs illustrate the significant difference between the set of interest (phosphoset) and randomly selected sets (Figure 2, lower panel).

Next, we investigated the conservation of phosphorylated sites. We calculated the proportions of phosphorylated and nonphosphorylated serines, threonines, and tyrosines that are conserved in other eukaryotic species. We only considered sites of phosphorylated proteins that have orthologs. Furthermore, only sites located on nonregularly structured regions on the

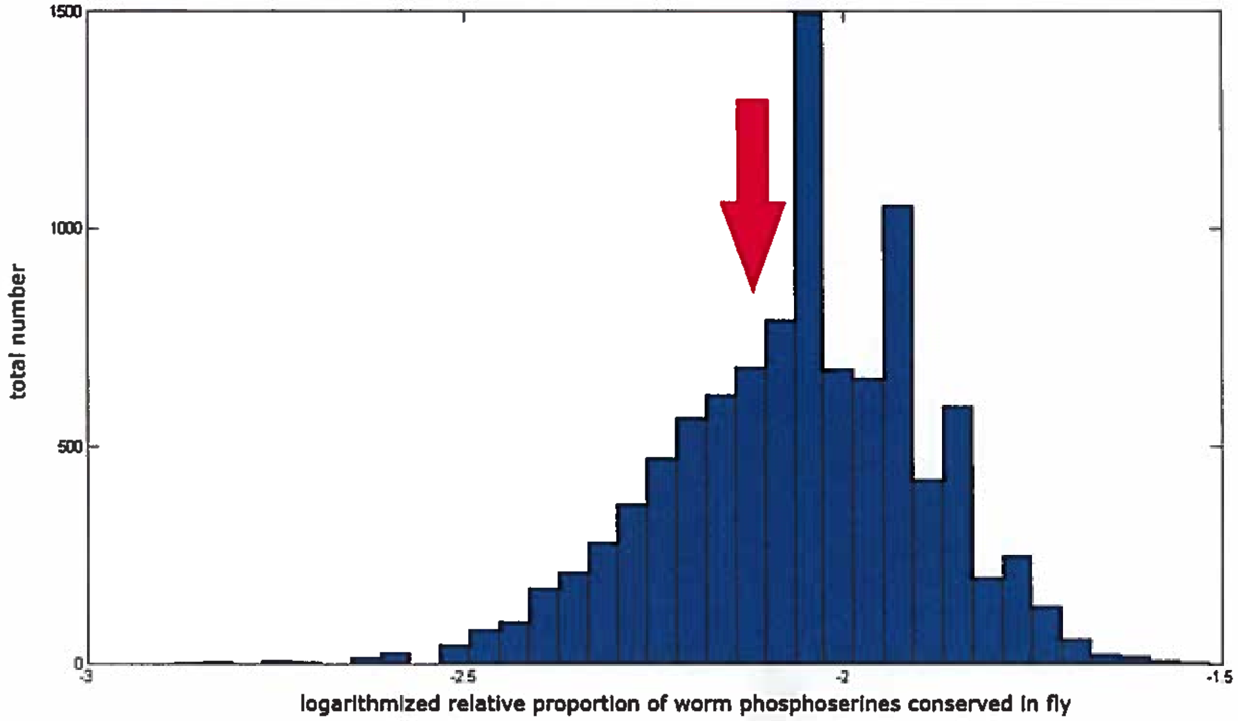


Figure 3. Bootstrapping distribution reflecting the conservation of worm serines in fly. The proportion of conserved phosphoserines (highlighted in red) is close to the average value of the bootstrapping distribution, which is created by iterative random selection of serine sets. This indicates that phosphorylated serines are not significantly more conserved than any other serines of the same worm protein.

protein surfaces according to SABLE 2.0 structure predictions were included in our analysis, as previous experience has shown that phosphorylation sites are predominantly located in those fast evolving regions.[34] Structure prediction on phosphorylated *C. elegans* proteins confirmed preferential phosphorylation events in loops also in worm (Supplementary Figure 6). To reduce the structure effects on the conservation analysis, we excluded sites that occur in the hydrophobic core of the proteins in our evolutionary study.

We found that phosphorylated residues are as conserved as their nonphosphorylated counterparts. To confirm the statistical significance of our observation, we applied a bootstrapping approach (Figure 3). Unexpectedly, phosphorylated residues were even less conserved than nonphosphorylated residues in some species, which can be explained by the structural localization of phosphorylated worm residues (loops with high flexibility).

3. The Requirement for Worm-Specific Phosphosite Prediction and Its Application To Establish Species-Specific Phosphorylation Patterns. We have previously established yeast, fly, mouse, and human specific phosphorylation site predictors[34] on the basis of support vector machines (SVMs), a machine learning technology.[42] Each species-specific prediction tool is freely available via the Phosphorylation Site Database (www.phosida.com). We compared the accuracies of predicting worm phosphosites and human phosphosites on the basis of a SVM that we had specifically trained on human sites.[9] As expected, the human site predictor determined human phosphosites much better than it did worm phosphosites (Supplementary Figure 7). This underlines the fact that the worm phosphoproteome is distinct as already concluded from the conservation analysis described above and supports the usefulness of species-specific prediction. Interestingly, the accuracy of predicting fly phosphosites with the human phosphosite predictor was better than the accuracy for predicting worm phosphosites.

Therefore, we created a worm-specific SVM-based phosphosite predictor using the large number of *in vivo* phosphorylation sites identified in *C. elegans*. The main concepts of the training, testing, and application of the predictor have been described before[34] (see also Materials and Methods). To our knowledge, the resulting online application provides the first *C. elegans* specific predictor that is capable of identifying worm phosphorylation sites *in silico* on the basis of the protein sequence.

Iteratively, we used 90% of the positive and negative sets for training and 10% for testing (5-fold cross-validation). Overall, 85% of the phosphoserines were predicted correctly in the test set (on a 1:1 phospho vs nonphospho data set). High accuracy was also observed for the prediction of phosphothreonines (80%). The resulting Precision-Recall curves for the worm-specific phosphoserine and phosphothreonine prediction are illustrated in Figure 4. Precision is the proportion of true positives out of all predicted positives, whereas recall describes the number of true positives to the sum of true positives and false negatives. Because of the low number of phosphotyrosines, it was not possible to create a worm-specific pY predictor. The prediction of phosphorylated serines and threonines on any input protein sequence was integrated into PHOSIDA.

Phosphorylation in Worm Sex Determination, Sex Differentiation, and Development. The gene ontology enrichment analysis for biological processes described above revealed that a significantly higher proportion of proteins that play fundamental roles in the sex determination and development of *C. elegans* was found to be phosphorylated compared to the entire

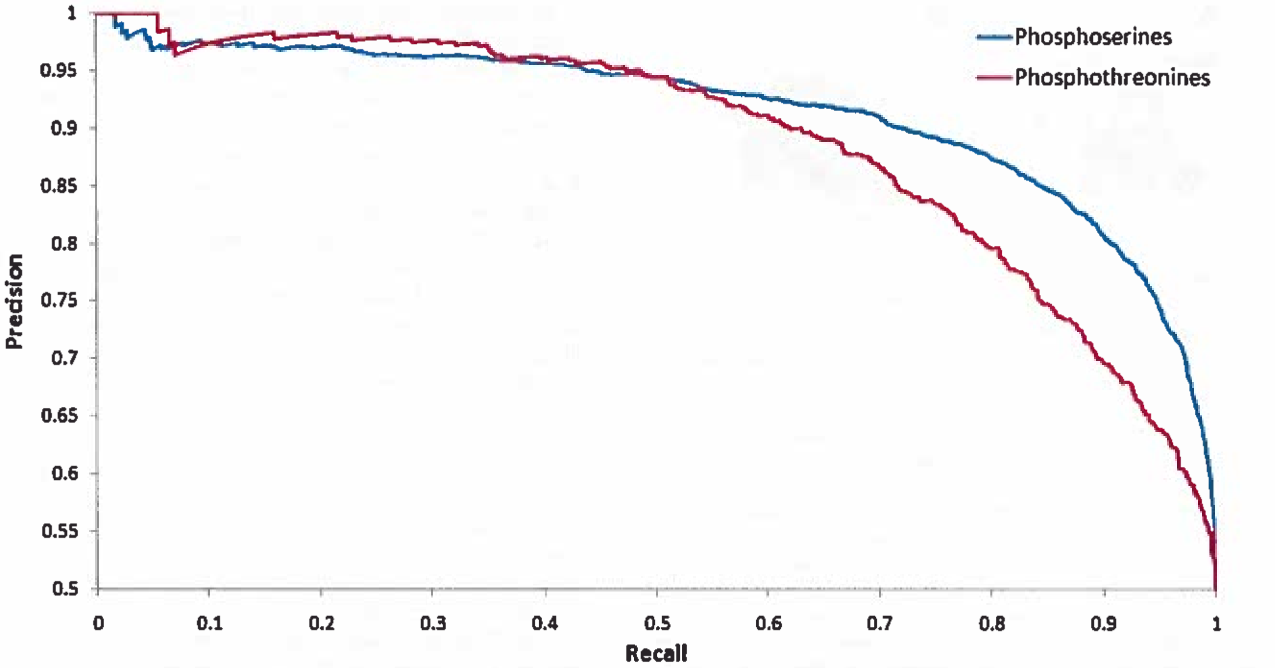


Figure 4. Precision Recall curves for phosphoserine and phosphothreonine prediction. The two lines present the accuracies for the prediction of phosphorylated serines (blue) and phosphorylated threonines (red) in worm.

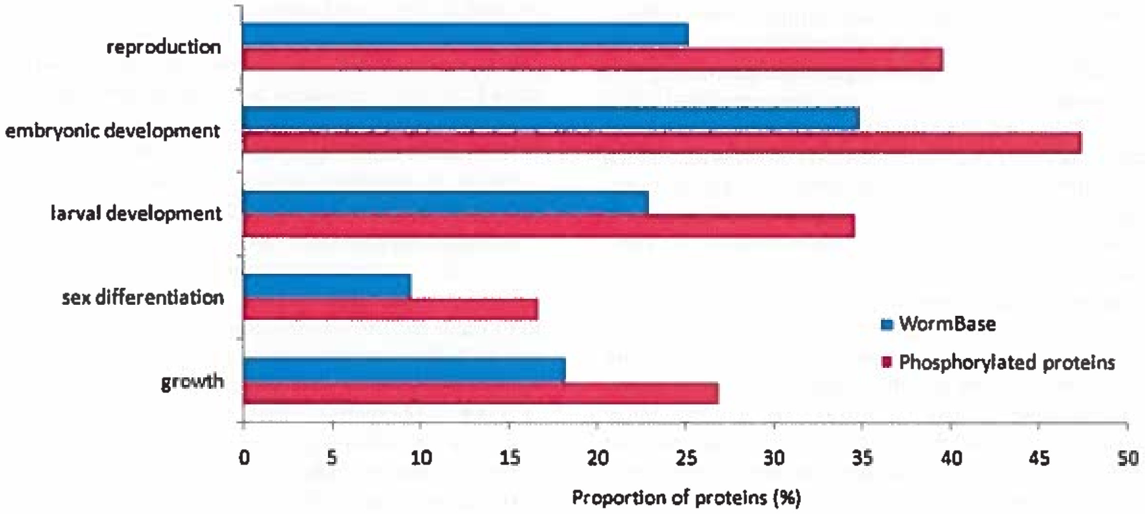


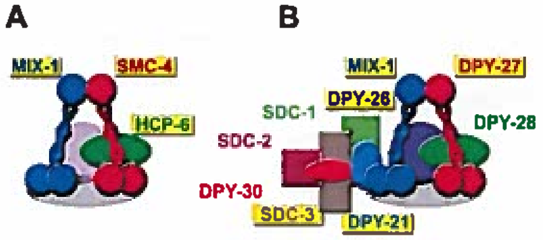
Figure 5. Biological processes that are significantly overrepresented in the phosphoset compared to the entire worm proteome according to gene ontology annotation. On the basis of gene ontology enrichment analysis, phosphoproteins were found to be involved in the development and sex differentiation to a significantly higher degree.

proteome (Figure 5). For example, 40% of phosphorylated proteins compared to 25% of all worm proteins are involved in reproduction. Furthermore, many genes essential for sex determination were found to be phosphorylated in our study (Table 1).[43,44]

The gene *xol-1* plays the key role in sex determination in worm, as it 'counts' the X chromosomes of a given individual and thereby controls both sex determination and dosage compensation.[45] Its repressor SEX-1 (signal element on X) was found to be phosphorylated on a serine at position 238 (S238), which matches with the CKI kinase motif. To our knowledge, this is the first evidence for phosphorylation of SEX-1. Interestingly, most of the proteins that are involved in the X chromosome dosage compensation process,[46] which plays a crucial role in the sex specific gene regulation of the worm, were

Table 1. Examples of Phosphorylated Sex Determination Proteins

gene	role in sex determination
gld-1	germline translational repressor of *tra-2*
mag-1	germline repressor of male promoting genes
mog-1	global repressor of *fem-3* translation
mog-5	global repressor of *fem-3* translation
nos-3	germline cofactor of *FBF-1/2*, repressor of *fem-3* translation
sex-1	X-dosage counting element
sdc-3	X-dosage compensation complex component, her-1 transcriptional repressor
tra-2	receptor of HER-1, repressor of fem gens
tra-3	receptor of HER-1, repressor of fem gens



Worm Mitotic/Meiotic Condensin Complex

Worm Dosage Compensation Complex

Figure 6. Two condensin complexes in *C. elegans*. (A) The condensin complex specialized for the resolution, compaction, and segregation of mitotic and meiotic chromosomes is conserved in all eukaryotes ranging from yeast to human. All known subunits of the complex are detected to be phosphorylated in our set (the phosphorylated proteins are marked yellow). (B) Nematode specific X chromosome dosage compensation complex that binds to both X chromosomes of hermaphrodites and equalizes the gene expression between males and females. Accordingly, most of the involved proteins were found to be phosphorylated in our study. The color code represents homologous proteins in these two complexes. Proteins in light gray and light magenta are assumed to be present in the complexes (original source: http://www.wormbook.org; Reprinted with permission from ref 48. Copyright 2005 Barbara J. Meyer.)

detected to be phosphorylated in our study (Figure 6B). MIX-1 (mitosis and X) is an essential dosage compensation component[47] and was identified to be phosphorylated on a CK2 motif matching S366. The chromosome segregation protein DPY-27 was phosphorylated on S1419 (matching kinase motifs of GSK3, CAMK2, PLK, PLK1) and S1423 (matching kinase motifs of CK1). Additional important dosage compensation proteins that match with kinase motifs of diverse kinase families and were phosphorylated in our study are DPY-26 (S28, S612, T617, S985, T1060, T1062), SDC-3 (T205, T1581, S1647) and DPY-21 (T565, S776, T1329).

Furthermore, as highlighted in Figure 6A, all known components of the condensin complex (MIX-1, SMX-4 and HCP-6), which is specialized for the resolution, compaction, and segregation of mitotic and meiotic chromosomes, were phosphorylated. All proteins involved in the condensing complex are conserved throughout the entire eukaryotic domain and homologous to members of the dosage compensation complex, which is nematode-specific.

Notably, the phosphorylation sites of both worm complexes are not conserved throughout the eukaryotic domain, raising the possibility that regulation by phosphorylation of this complex is nematode-specific.

Discussion

The main goal of this work was to provide a phosphoproteome of the model organism *C. elegans* to the community as a resource and to obtain functional insights from it. Our study identified 6780 phosphosites from 2373 proteins within a single experiment using 6 mg of *C. elegans* protein starting material. Overall, 99% of the sites were novel as judged by their Swiss-Prot database entries. Gene ontology analysis of phosphorylated proteins revealed enrichment in membrane proteins, in contrast to our previous phosphoproteomic studies.[9] This is presumably due to the FASP method, which allows the solubilization of proteins in strong detergents such as SDS and therefore does not penalize membrane proteins in proteomic analysis.

As expected, some basic features of the worm phosphoproteome are shared with the phosphoproteomes of other eukaryotes. For example, the distribution of phosphorylation events on serine, threonine, and tyrosine (80%, 18.2%, and 1.8%, respectively) in worm is similar to the distribution in other eukaryotes.[9] Moreover, phosphatases and kinases were identified among the phosphorylated proteins. For instance, the mitogen activated protein kinase was found to be phosphorylated on T191 and Y193. Both of these phosphorylation events have been suggested by sequence similarity analysis according to Uniprot, but lacked experimental evidence in *C. elegans* so far. The putative cell division cycle protein kinase 7 was also determined to be phosphorylated at S4 and S17. These phosphosites have not been reported or predicted so far. The protein kinase C-like kinase was phosphorylated on six residues that have not been reported yet. Three phosphorylation sites (S533, T534, S538) are located in the protein kinase domain and may therefore be involved in the regulation of this kinase. We also found evidence for enriched biological functions that are characteristic for signaling proteins. Binding functions and transcription/translation regulative functions including DNA/RNA binding were significantly overrepresented in our phosphoset. This has already been reported for other eukaryotic species and underlines the universal regulatory impact of phosphorylation throughout all domains of life. Structure analysis showed that worm phosphosites are predominantly located in loops and turns on the protein surface, again in concordance with previous experience of other eukaryotic phosphoproteomes. Furthermore, phosphorylated worm proteins have more orthologs than nonphosphorylated proteins. This observation is consistent with evolutionary analyses of other eukaryotic phosphoproteomes[18] suggesting evolutionary pressure to maintain phosphorylated proteins with crucial regulatory functions in signaling during evolution.

Even though general features of the worm phosphoproteome are similar to other eukaryotes and many phosphorylated substrates were already present in common ancestors, we found that the site specific phosphorylation patterns in worm are quite distinct from other eukaryotic phosphoproteomes. For example, phosphorylated amino acids were as conserved as nonphosphorylated amino acids of the same protein. The relatively low conservation of worm phosphosites compared to other metazoans indicates different phosphorylation patterns between nematodes and higher eukaryotes. Previous work had already shown that the phosphoproteomes of yeast, a single cellular species, and higher eukaryotes are very distinct.[49] While this is perhaps not surprising, our study demonstrates that signaling functions via post-translational phosphorylation can also differ drastically within multicellular model eukaryotes, as worm phosphosites are not significantly conserved compared to nonphosphorylated residues in other higher eukaryotes such as human. We had earlier seen evidence for this lack of conservation in another large-scale evolutionary study of eukaryotic phosphorylation sites which showed that phosphorylation sites identified in human cells were significantly conserved in other higher eukaryotes, but not in worm.[18] Furthermore, the relatively low conservation of worm phosphosites is in agreement with the fact that around half of the worm kinome is nematode-specific. The *C. elegans* kinome expanded dramatically from a relatively small number of kinase classes. It is nearly 2× larger than the fly kinome, for example. Even though our motif analysis showed that a large proportion of worm phosphosites likely are substrates of kinases that are conserved in eukaryotes, the many novel sequence motifs that were identified via *de novo* motif analysis point to phosphorylation by *C. elegans* specific kinases.

Species-specific phosphophosite predictors constructed and trained by machine learning provide further evidence of the 'outlier status' of the *C. elegans* phosphoproteome. The human phosphosite predictor is more accurate on human sequences and even on those in fly, which has roughly the same evolutionary distance from man as the nematode. Together, our data and analyses provide strong evidence of nematode-specific phosphorylation, which cannot be captured by the human phosphosite predictor. This finding motivated us to use the large number of *in vivo* phosphorylation sites to establish a *C. elegans* specific phosphorylation site predictor on the basis of a support vector machine. Overall, in 1:1 nonphospho and phosphosets, its accuracy is 85% for predicting phosphorylated serines and 80% for predicting phosphorylated threonines. Since this predictor is relatively accurate and its application is purely *in silico* without any experimental effort, it may be of use to the *C. elegans* community, for example, to generate putative target sites for functional experiments.

Biological process analysis suggested a highly significant overrepresentation of phosphoproteins involved in nematode development and sex determination. For example, the phosphorylated protein, W04D2.6, is required for normal growth, embryonic viability and oogenesis.[50] Repressors such as *tra-2* and *mog-5* are essential regulators for sex determination in nematodes and we found them to be phosphorylated in our set.[51−53] The X-chromosome counting factor *sex-1* and the pre-mRNA splicing factors *mag-1* and *mog-1* are further essential sex determination regulators[54−56] and we found them to be phosphorylated in worm, too.

In nematodes as well as in other multicellular eukaryotes, sex is determined by the ratio of X chromosomes to autosomes (X/A).[43] Female worms have two X chromosomes, whereas male worms have only one X chromosome. The interpretation of the X/A is very sensitive in *C. elegans* as evident from triploid and tetraploid worms: it has been shown that worms with two chromosomes and three sets of autosomes ($X/A = 0.67$) develop as males. In contrast, tetraploid worms with consequently higher X/A ratios ($X/A = 0.75$) develop as females.[57,58] There are diverse strategies for dosage compensation, a process that yields equalized expression levels of genes located on the X chromosome between males and females.[46] In male flies, for example, the expression level of genes on the X chromosome is boosted to compensate the lack of another X chromosome. In contrast, in worm, the dosage compensation complex binds to both chromosomes of hermaphrodites (females) and reduces the expression level on both chromosomes by a factor of 2. Interestingly, we found the majority of the subunits of the DCC complex to be phosphorylated suggesting a putative control of the mechanism by phosphorylation. The *C. elegans* DCC complex is similar to the well-conserved 13 S condensing complex that resolves and condenses mitotic and meiotic chromosomes. All subunits of this complex were detected to be phosphorylated, suggesting a regulatory role of phosphorylation.

The DCC complex and the 13 S complex share the phosphorylated protein MIX-1. Mutation in MIX-1 was reported to cause not only disruption of dosage compensation, but also incorrect chromosome resolution and condensation resulting in death of both female and male nematodes.[47,59,60] MIX-1 binds to hermaphrodite X-chromosomes when it associates with the dosage compensation protein DPY-27 or to the mitotic female/male chromosomes after association with mitotic/meiotic-specific protein SMC-4. Both DPY-27 and SMC-4 were identified to be phosphorylated in our study. Another essential phosphorylated dosage compensation protein that is recruited to X by binding to other proteins is DPY-21.[61] Mutation studies showed that amino acid changes in these proteins lead to the disruption of dosage compensation and therefore to reduced viability of hermaphrodite, but not of male nematodes.[46] All these observations suggest putative regulatory roles of phosphorylation in sex determination and development of worm and provide a starting point for functional studies at a site specific level.

Conclusion and Outlook

Our study demonstrates the capability of rapidly developing mass spectrometry based proteomics to measure an *in vivo* phosphoproteome of a complex sample such as whole worms with a starting material of less than 6 mg of protein within less than 1 week. Recently developed methods including FASP and TiO_2 chromatography even allow the in-depth detection of low-abundance or insoluble phosphoproteins such as membrane proteins.

Bioinformatic analyses uncovered the role of the *C. elegans* phosphorylated proteins in sex differentiation and development and proved our hypothesis that the worm phosphoproteome is very distinct from phosphoproteomes of other eukaryotes with respect to its role in biological processes, kinase specificity, and evolutionary conservation.

We hope that this work will stimulate research on some of the identified phosphorylation sites to determine their biological functions in *C. elegans*. One of the most interesting results of our study is the high degree of phosphorylation in the chromosome dosage complex. It would be interesting to investigate identified phosphorylation sites of this fundamental complex or to compare phosphorylation site patterns in males versus females or in different developmental and phenotype stages.

The *C. elegans* phosphoproteome along with the prediction application was integrated into PHOSIDA (www.phosida.com). The retrieval of phosphorylated peptides and sites along with cross-links to biological annotations including evolution and structure in the phosphorylation site database is explained in detail in Gnad et al.[34] Alternatively, the background sections of the database provide information about the application of PHOSIDA. In addition, our *C. elegans* phosphoproteome data set will be made available within WormBase.

Abbreviations: DCC, dosage compensation complex; DHB, dihydroxybenzoic acid; FASP, filter aided sample preparation; HPLC, high performance liquid chromatography; LTQ, linear trap quadrupole; MS, mass spectrometry; SDS, sodium dodecyl sulfate; SEC, size exclusion chromatography; SVM, support vector machine.

Acknowledgment. Nadin Neuhauser helped in the analysis of large-scale data sets and Shubin Ren contributed to the construction of the *C. elegans* phosphorylation site predictor. We thank Jeremy Gunawardena and Gerard Manning for fruitful discussions and Alexandre Zougman for help with mass spectrometry. This work was partially supported by CIPS, the Center for Integrated Protein Science Munich.

Supporting Information Available: Supplementary Table 1, phosphorylation sites identified in *C. elegans*. Phosphorylation sites of kinsases are marked in blue. Phosphosites of phosphatases are presented in green. Supplementary Table 2, enriched functions, processes and cellular component localizations of phosphoproteins. Supplementary Table 3, phenotype analysis of phosphoproteins. Supplementary Table 4, differences between the distribution of genes encoding phosphoproteins and the distribution of all other genes of the worm chromosomes from the ending 1% to 50% of the chromosomes on both sides (relative to the entire chromosome lengths). Supplementary Table 5, proportion of worm phosphosites that match with annotated eukaryotic kinase motifs. Proportions of worm phosphosites that match with annotated eukaryotic kinase motifs are compared with the proportions of worm phosphosites that would match with kinase motifs by chance. The resulting χ^2 values estimate the significant difference between those two proportions. Supplementary Table 6, identified significantly overrepresented sequence motifs. Supplementary Table 7, proportion of worm genes that have orthologs in other eukaryotic species. Supplementary Figure 1, distribution of identified singly and multiply phosphorylated peptides. Supplementary Figure 2, distribution of Class I phosphorylation sites by amino acids. Supplementary Figure 3, enriched cellular component localizations of phosphoproteins. Supplementary Figure 4, difference between the distribution of genes encoding phosphoproteins and the distribution of all other genes of the X chromosome from the ending 1% to 50% of the chromosome on both sides (relative to the entire chromosome length). Genes encoding phosphoproteins are located on the chromosome ends to a higher degree than all other genes. Supplementary Figure 5, distribution of all worm genes (left) and worm genes that encode phosphoproteins (right) on each chromosome. Supplementary Figure 6, average accessibilities (a) and proportional loop localizations (b) of phosphorylated and nonphosphorylated serines/threonines/tyrosines in *C. elegans* (according to structure prediction using SABLE 2.0) Supplementary Figure 7, Precision Recall curves for worm (red), fly (green) and human (blue) phosphoserine prediction based on a human phosphosite predictor. We applied the human phosphosite predictor separately on phosphorylated worm serines (and their nonphosphorylated counterparts), on phosphorylated fly serines (and their nonphosphorylated counterparts) and on phosphorylated human serines (and their nonphosphorylated counterparts). On the basis of cross-validation, the performance for human phosphosite prediction was better than the one for worm phosphosite prediction. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hunter, T. Signaling—2000 and beyond. *Cell* 2000, *100* (1), 113-27.

(2) Cohen, P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur. J. Biochem.* 2001, *268* (19), 5001–10.

(3) Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* 2000, *103* (2), 211–25.

(4) Schreiber, T. B.; Mausbacher, N.; Breitkopf, S. B.; Grundner-Culemann, K.; Daub, H. Quantitative phosphoproteomics—an emerging key technology in signal-transduction research. *Proteomics* 2008, *8* (21), 4416–32.

(5) Macek, B.; Mijakovic, I.; Olsen, J. V.; Gnad, F.; Kumar, C.; Jensen, P. R.; Mann, M. The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. *Mol. Cell. Proteomics* 2007, *6* (4), 697–707.

(6) Witze, E. S.; Old, W. M.; Resing, K. A.; Ahn, N. G. Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* 2007, *4* (10), 798–806.

(7) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol.* 2002, *20* (3), 301–5.

(8) Pan, C.; Gnad, F.; Olsen, J. V.; Mann, M. Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics* 2008, *8* (21), 4534–46.

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

(10) Bodenmiller, B.; Malmstrom, J.; Gerrits, B.; Campbell, D.; Lam, H.; Schmidt, A.; Rinner, O.; Mueller, L. N.; Shannon, P. T.; Pedrioli, P. G.; Panse, C.; Lee, H. K.; Schlapbach, R.; Aebersold, R. PhosphoPep—a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.* 2007, *3*, 139.

(11) C. elegans Sequence Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 1998, *282* (5396), 2012–8.

(12) Madi, A.; Mikkat, S.; Koy, C.; Ringel, B.; Thiesen, H. J.; Glocker, M. O. Mass spectrometric proteome analysis suggests anaerobic shift in metabolism of Dauer larvae of Caenorhabditis elegans. *Biochim. Biophys. Acta* 2008, *1784* (11), 1763–70.

(13) Cassada, R. C.; Russell, R. L. The dauerlarva, a post-embryonic developmental variant of the nematode Caenorhabditis elegans. *Dev. Biol.* 1975, *46* (2), 326–42.

(14) Kenyon, C.; Chang, J.; Gensch, E.; Rudner, A.; Tabtiang, R. A C. elegans mutant that lives twice as long as wild type. *Nature* 1993, *366* (6454), 461–4.

(15) Merrihew, G. E.; Davis, C.; Ewing, B.; Williams, G.; Kall, L.; Frewen, B. E.; Noble, W. S.; Green, P.; Thomas, J. H.; MacCoss, M. J. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res.* 2008, *18* (10), 1660–9.

(16) Schrimpf, S. P.; Weiss, M.; Reiter, L.; Ahrens, C. H.; Jovanovic, M.; Malmstrom, J.; Brunner, E.; Mohanty, S.; Lercher, M. J.; Hunziker, P. E.; Aebersold, R.; von Mering, C.; Hengartner, M. O. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol.* 2009, *7* (3), e48.

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

(18) Gnad, F.; Gunawardena, J.; Forner, F.; Birney, E.; Mann, M., Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes and mitochondria. *Mol. Syst. Biol.*, submitted for publication, 2009.

(19) Manning, G.; Plowman, G. D.; Hunter, T.; Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* 2002, *27* (10), 514–20.

(20) Lewis, J. A.; Fleming, J. T. Basic culture methods. *Methods Cell Biol.* 1995, *48*, 3–29.

(21) Nielsen, P. A.; Olsen, J. V.; Podtelejnikov, A. V.; Andersen, J. R.; Mann, M.; Wisniewski, J. R. Proteomic mapping of brain plasma membrane proteins. *Mol. Cell. Proteomics* 2005, *4* (4), 402–8.

(22) Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell. Proteomics* 2005, *4* (7), 873–86.

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

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

(25) Schroeder, M. J.; Shabanowitz, J.; Schwartz, J. C.; Hunt, D. F.; Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal. Chem.* 2004, *76* (13), 3590–8.

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

(27) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20* (18), 3551–67.

(28) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a

software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003, *13* (11), 2498–504.

(29) Maere, S.; Heymans, K.; Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005, *21* (16), 3448–9.

(30) Flicek, P.; Aken, B. L.; Beal, K.; Ballester, B.; Caccamo, M.; Chen, Y.; Clarke, L.; Coates, G.; Cunningham, F.; Cutts, T.; Down, T.; Dyer, S. C.; Eyre, T.; Fitzgerald, S.; Fernandez-Banet, J.; Graf, S.; Haider, S.; Hammond, M.; Holland, R.; Howe, K. L.; Howe, K.; Johnson, N.; Jenkinson, A.; Kahari, A.; Keefe, D.; Kokocinski, F.; Kulesha, E.; Lawson, D.; Longden, I.; Megy, K.; Meidl, P.; Overduin, B.; Parker, A.; Pritchard, B.; Prlic, A.; Rice, S.; Rios, D.; Schuster, M.; Sealy, I.; Slater, G.; Smedley, D.; Spudich, G.; Trevanion, S.; Vilella, A. J.; Vogel, J.; White, S.; Wood, M.; Birney, E.; Cox, T.; Curwen, V.; Durbin, R.; Fernandez-Suarez, X. M.; Herrero, J.; Hubbard, T. J.; Kasprzyk, A.; Proctor, G.; Smith, J.; Ureta-Vidal, A.; Searle, S. Ensembl 2008. *Nucleic Acids Res.* 2008, *36* (Database issue), D707–14.

(31) Gnad, F.; Oroshi, M.; Birney, E.; Mann, M. MAPU 2.0: high-accuracy proteomes mapped to genomes. *Nucleic Acids Res.* 2009, *37* (Database issue), D902−6.

(32) Jenkinson, A. M.; Albrecht, M.; Birney, E.; Blankenburg, H.; Down, T.; Finn, R. D.; Hermjakob, H.; Hubbard, T. J.; Jimenez, R. C.; Jones, P.; Kahari, A.; Kulesha, E.; Macias, J. R.; Reeves, G. A.; Prlic, A. Integrating biological data—the Distributed Annotation System. *BMC Bioinf.* 2008, *9* (8), S3.

(33) Wagner, M.; Adamczak, R.; Porollo, A.; Meller, J. Linear regression models for solvent accessibility prediction in proteins. *J. Comput. Biol.* 2005, *12* (3), 355–69.

(34) Gnad, F.; Ren, S.; Cox, J.; Olsen, J. V.; Macek, B.; Oroshi, M.; Mann, M. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* 2007, *8* (11), R250.

(35) Fraser, A. G.; Kamath, R. S.; Zipperlen, P.; Martinez-Campos, M.; Sohrmann, M.; Ahringer, J.; Functional genomic analysis of, C. elegans chromosome I by systematic RNA interference. *Nature* 2000, *408* (6810), 325–30.

(36) Simmer, F.; Moorman, C.; van der Linden, A. M.; Kuijk, E.; van den Berghe, P. V.; Kamath, R. S.; Fraser, A. G.; Ahringer, J.; Plasterk, R. H. Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions. *PLoS Biol.* 2003, *1* (1), E12.

(37) Hillier, L. W.; Miller, R. D.; Baird, S. E.; Chinwalla, A.; Fulton, L. A.; Koboldt, D. C.; Waterston, R. H. Comparison of C. elegans and C. briggsae genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biol.* 2007, *5* (7), e167.

(38) Barnes, T. M.; Kohara, Y.; Coulson, A.; Hekimi, S. Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. *Genetics* 1995, *141* (1), 159–79.

(39) Rockman, M. V.; Kruglyak, L. Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet.* 2009, *5* (3), e1000419.

(40) Zanivan, S.; Gnad, F.; Wickstrom, S. A.; Geiger, T.; Macek, B.; Cox, J.; Fassler, R.; Mann, M. Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. *J Proteome Res* 2009, *7* (12), 5314–26.

(41) Gnad, F.; Gunawardena, J. A bootstrapping based frequent itemset mining approach to the identification of protein phosphorylation motifs. *Bioinformatics*, submitted for publication, 2009.

(42) Noble, W. S. What is a support vector machine. *Nat. Biotechnol.* 2006, *24* (12), 1565–7.

(43) Hansen, D.; Pilgrim, D. Sex and the single worm: sex determination in the nematode C. elegans. *Mech. Dev.* 1999, *83* (1−2), 3–15.

(44) Haag, E. S.; Kimble, J. Regulatory elements required for development of caenorhabditis elegans hermaphrodites are conserved in the tra-2 homologue of C. remanei, a male/female sister species. *Genetics* 2000, *155* (1), 105–16.

(45) Meyer, B. J. Sex in the wormcounting and compensating X-chromosome dose. *Trends Genet.* 2000, *16* (6), 247–53.

(46) Meyer, B. J.; Casson, L. P. Caenorhabditis elegans compensates for the difference in X chromosome dosage between the sexes by regulating transcript levels. *Cell* 1986, *47* (6), 871–81.

(47) Lieb, J. D.; Albrecht, M. R.; Chuang, P. T.; Meyer, B. J. MIX-1: an essential component of the C. elegans mitotic machinery executes X chromosome dosage compensation. *Cell* 1998, *92* (2), 265–77.

(48) Meyer, B. J. X-Chromosome dosage compensation (June 25, 2005), *WormBook*, The C. elegans Research Community, WormBook, ed., doi/10.1895/wormbook.1.8.1, http://www.wormbook.org.

(49) Gnad, F.; Godoy de, L.M. F.; Cox, J.; Ren, S.; Olsen, J. V.; Mann, M., High accuracy identification and bioinformatic analysis of in-vivo protein phosphorylation sites in yeast. *Proteomics*, submitted for publication, 2009.

(50) Parry, D. H.; Xu, J.; Ruvkun, G. A whole-genome RNAi Screen for C. elegans miRNA pathway genes. *Curr. Biol.* 2007, *17* (23), 2013–22.

(51) Puoti, A.; Kimble, J. The hermaphrodite sperm/oocyte switch requires the Caenorhabditis elegans homologs of PRP2 and PRP22. *Proc. Natl. Acad. Sci. U.S.A.* 2000, *97* (7), 3276–81.

(52) Wilkins, A. S. Moving up the hierarchy: a hypothesis on the evolution of a genetic sex determination pathway. *BioEssays* 1995, *17* (1), 71–7.

(53) Kuwabara, P. E.; Kimble, J. A predicted membrane protein, TRA-2A, directs hermaphrodite development in Caenorhabditis elegans. *Development* 1995, *121* (9), 2995–3004.

(54) Li, W.; Boswell, R.; Wood, W. B. mag-1, a homolog of Drosophila mago nashi, regulates hermaphrodite germ-line sex determination in Caenorhabditis elegans. *Dev. Biol.* 2000, *218* (2), 172–82.

(55) Puoti, A.; Kimble, J. The Caenorhabditis elegans sex determination gene mog-1 encodes a member of the DEAH-Box protein family. *Mol. Cell. Biol.* 1999, *19* (3), 2189–97.

(56) Carmi, I.; Kopczynski, J. B.; Meyer, B. J. The nuclear hormone receptor SEX-1 is an X-chromosome signal that determines nematode sex. *Nature* 1998, *396* (6707), 168–73.

(57) Hodgkin, J.; Horvitz, H. R.; Brenner, S. Nondisjunction mutants of the nematode Caenorhabditis elegans. *Genetics* 1979, *91* (1), 67–94.

(58) Madl, J. E.; Herman, R. K. Polyploids and sex determination in Caenorhabditis elegans. *Genetics* 1979, *93* (2), 393–402.

(59) Chu, D. S.; Dawes, H. E.; Lieb, J. D.; Chan, R. C.; Kuo, A. F.; Meyer, B. J. A molecular link between gene-specific and chromosome-wide transcriptional repression. *Genes Dev.* 2002, *16* (7), 796–805.

(60) Hagstrom, K. A.; Holmes, V. F.; Cozzarelli, N. R.; Meyer, B. J., C. elegans condensin promotes mitotic chromosome architecture, centromere organization, and sister chromatid segregation during mitosis and meiosis. *Genes Dev.* 2002, *16* (6), 729–42.

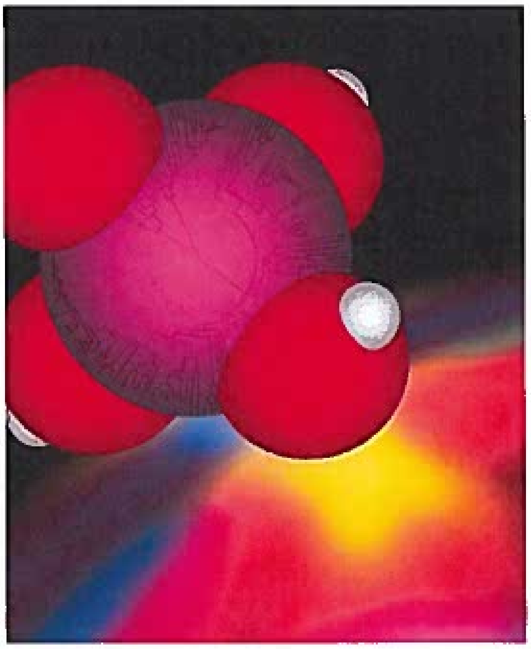
(61) Yonker, S. A.; Meyer, B. J.; Recruitment of, C. elegans dosage compensation proteins for gene-specific versus chromosome-wide repression. *Development* 2003, *130* (26), 6519–32.

PR900384K

Appendix II

Evolutionary Constraints of Phosphorylation in Eukaryotes, Prokaryotes and Mitochondria

Gnad F, Forner F, Zielinska DF, Birney E, Gunawardena J, Mann M



Mol Cell Proteomics, 2010

This paper is a combined proteomics and bioinformatic study of the mitochondrial mouse phosphoproteome and addresses, for the first time, the evolution of phosphorylation from prokaryotes to mammals. We identified 174 phosphorylation sites in mitochondria and showed that mitochondrial phosphorylation sites are not conserved throughout prokaryotes, consistent with the notion that serine/threonine phosphorylation in prokaryotes occurred relatively recently in evolution.

Evolutionary Constraints of Phosphorylation in Eukaryotes, Prokaryotes, and Mitochondria*

Florian Gnad‡, Francesca Forner§, Dorota F. Zielinska§, Ewan Birney¶, Jeremy Gunawardena‡, and Matthias Mann§||

High accuracy mass spectrometry has proven to be a powerful technology for the large scale identification of serine/threonine/tyrosine phosphorylation in the living cell. However, despite many described phosphoproteomes, there has been no comparative study of the extent of phosphorylation and its evolutionary conservation in all domains of life. Here we analyze the results of phosphoproteomics studies performed with the same technology in a diverse set of organisms. For the most ancient organisms, the prokaryotes, only a few hundred proteins have been found to be phosphorylated. Applying the same technology to eukaryotic species resulted in the detection of thousands of phosphorylation events. Evolutionary analysis shows that prokaryotic phosphoproteins are preferentially conserved in all living organisms, whereas site specific phosphorylation is not. Eukaryotic phosphosites are generally more conserved than their non-phosphorylated counterparts (with similar structural constraints) throughout the eukaryotic domain. Yeast and *Caenorhabditis elegans* are two exceptions, indicating that the majority of phosphorylation events evolved after the divergence of higher eukaryotes from yeast and reflecting the unusually large number of nematode-specific kinases. Mitochondria present an interesting intermediate link between the prokaryotic and eukaryotic domains. Applying the same technology to this organelle yielded 174 phosphorylation sites mapped to 74 proteins. Thus, the mitochondrial phosphoproteome is similarly sparse as the prokaryotic phosphoproteomes. As expected from the endosymbiotic theory, phosphorylated as well as non-phosphorylated mitochondrial proteins are significantly conserved in prokaryotes. However, mitochondrial phosphorylation sites are not conserved throughout prokaryotes, consistent with the notion that serine/threonine phosphorylation in prokaryotes occurred relatively recently in evolution. Thus, the phosphoproteome reflects major events in the evolution of life. ***Molecular & Cellular Proteomics 9:2642–2653, 2010.***

Reversible protein phosphorylation on serines, threonines, and tyrosines plays a crucial role in regulating processes in all living organisms ranging from prokaryotes to eukaryotes (1). Traditionally, phosphorylation has been detected in single, purified proteins using *in vitro* assays. Recent advances in mass spectrometry (MS)-based proteomics now allow the identification of *in vivo* phosphorylation sites with high accuracy (2–7). On-line databases such as PhosphoSite (8), Phospho.ELM (9), and PHOSIDA[1] (10) have collected and organized thousands of identified phosphosites. These databases as well as dedicated analysis environments such as NetworKIN (11, 12) offer and use contextual information including structural features, potential kinases, and conservation. They constitute resources that should allow the derivation of general patterns for phosphorylation events. Specifically, the recent availability of data for archaeal, prokaryotic, and diverse eukaryotic phosphoproteomes in these databases should enable investigation of the evolutionary history of this post-translational modification.

Prokaryotes have two separate classes of phosphorylation events. Apart from the canonical histidine/aspartate phosphorylation, which has been studied for decades, serine/threonine/tyrosine phosphorylation is also present and has recently become amenable to analysis by MS (13). Bacterial phosphoproteins are involved in protein synthesis, carbohydrate metabolism, and the phosphoenolpyruvate-dependent phosphotransferase system. Recent phosphoproteomics studies of *Bacillus subtilis*, *Escherichia coli*, and *Lactococcus lactis* described around 100 phosphorylation sites on serine, threonine, and tyrosine in each of these species (13–15). Bacterial phosphorylation sites can change in response to environmental conditions (16).

Interestingly, even archaea have serine/threonine and tyrosine phosphorylation. A recent study of *Halobacterium salinarum* described 75 serine/threonine/tyrosine phosphorylation sites on 62 proteins involved in a wide range of cellular processes including a variety of metabolic pathways (17).

Although only a few hundred phosphorylation events have been found in prokaryotic species, similar experimental conditions and effort have yielded the detection of thousands of phosphorylation events in eukaryotes ranging from yeast to

From the ‡Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, §Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany, and ¶European Bioinformatics Institute (European Molecular Biology Laboratory-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Received, July 4, 2010, and in revised form, August 4, 2010

Published, MCP Papers in Press, August 5, 2010, DOI 10.1074/mcp.M110.001594

---

[1] The abbreviations used are: PHOSIDA, posttranslational modification database; IPI, International Protein Index; FDR, false discovery rate; PTM, post-translational modification; LTQ, linear trap quadrupole.

© 2010 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

human (7, 18–21). Before the advent of large scale phosphoproteomics, serine/threonine/tyrosine phosphorylation has been estimated to affect one-third of all proteins (22). Recent large scale phosphoproteomics studies now suggest that more than half of all eukaryotic proteins are phosphorylated (23).

A key event in evolution was the endosymbiosis of prokaryotes that enabled the development of a much more complex type of life, the eukaryotic cell. Analyses of mitochondrial genes suggest that the α-proteobacterium *Rickettsia prowazekii* is the endosymbiotic precursor leading to modern mitochondria (24). Almost all of the mitochondrial genes have migrated to the nuclear genome during subsequent evolution, and it is predicted that 10–15% of eukaryotic nuclear genes of organisms encode mitochondrial proteins (25).

Thus, mitochondria with their unique evolutionary position between prokaryotes and eukaryotes form an interesting link for the evolutionary analysis of phosphorylation. Several studies investigated the mitochondrial phosphoproteome in different organisms using gel electrophoresis or specific enrichment methods coupled with mass spectrometry (26–28). Those studies established potential mitochondrial phosphoproteins. Three large scale studies based on affinity enrichment of phosphopeptides and mass spectrometry obtained direct experimental evidence of phosphorylation sites in mitochondria. Lee *et al.* (29) used a combination of different peptide enrichment strategies and found 80 phosphorylation sites of 48 different proteins from mouse liver. Very recently, a study by Deng *et al.* (30) characterized the murine cardiac mitochondrial mouse phosphoproteome, covering 236 phosphosites on 181 proteins. Investigating yeast, Reinders *et al.* (31) assigned 84 phosphorylation sites in 62 proteins.

To enable comparative analysis of phosphoproteomes between all domains of life and mitochondria, here we experimentally determined a high accuracy mitochondrial mouse phosphoproteome based on technology conditions similar to those applied to the identification of prokaryotic and eukaryotic phosphoproteomes. We then performed a detailed evolutionary study of the conservation of the identified phosphoproteins and phosphorylation sites in prokaryotes and in eukaryotes. This allowed an initial comparison of the phosphoproteomes of prokaryotes, mitochondria, and eukaryotes.

EXPERIMENTAL PROCEDURES

Cell Culture and Primary Cell Isolation—3T3-L1, brown preadipocytes, C2C12, and Hepa 1-6 cell lines were subcultured and differentiated in DMEM supplemented with 10% fetal bovine serum (Invitrogen) and antibiotics in 5% CO_2 at 37 °C. Stable isotope labeling by amino acids in cell culture was performed as described with L-[$^{13}C_6$,$^{15}N_2$]lysine and L-[$^{13}C_6$,$^{15}N_4$]arginine. 3T3-L1 preadipocytes were grown and differentiated as described previously (32). Brown preadipocytes were differentiated as described (33). Confluent C2C12 were differentiated into myotubes by reducing the percentage of serum to 2%. For the isolation of primary brown adipocytes, interscapular brown adipose tissue from 20 C57BL/6 newborn mice (1–2 days) was excised, immersed in Hank's buffered salt solution, cleaned free of connective tissue under a binocular microscope, minced, and digested with 1 mg/ml collagenase A (Roche Applied Science) at 37 °C for 30 min. After digestion, the slurry was passed through 250-μm-mesh opening fiber material (Sefar) and centrifuged at 500 × *g* for 1 min. The floating adipocytes and the supernatant were discarded. The stromal vascular fraction was resuspended in DMEM supplemented with 10% fetal bovine serum (Invitrogen) and antibiotics and transferred into 7-cm-diameter Petri dishes. Preadipocytes were grown to confluence and differentiated as described (33).

Isolation of Mitochondria—Cells were harvested with trypsin (Invitrogen) and diluted with DMEM supplemented with protease inhibitors (Roche Applied Science) and a 5 mM concentration of each of the following phosphatase inhibitors: sodium fluoride, 2-glycerol phosphate, sodium vanadate, and sodium pyrophosphate (Sigma). Cells were centrifuged at 1000 × *g* for 10 min. Cells were then resuspended with SEH buffer (250 mM sucrose, 10 mM Hepes, pH 7.4, 0.1 mM EGTA) supplemented with protease and phosphatase inhibitors and washed twice. The cell suspensions were homogenized in a 7-ml Dounce homogenizer. Membrane disruption was checked with trypan blue staining. The tissue homogenate was centrifuged twice at 800 × *g* for 10 min at 4 °C. The supernatant was centrifuged at 10,000 × *g* for 10 min at 4 °C. The crude mitochondrial pellet was resuspended in SEH buffer supplemented with protease inhibitors (Roche Applied Science). The suspension was further centrifuged at 7000 × *g* for 10 min at 4 °C and purified with Percoll gradients as described (34). 3T3-L1 cell mitochondria were purified by means of protease treatment of crude mitochondrial fractions with trypsin as described previously (35).

Phosphopeptide Enrichment—Mitochondrion-enriched fractions were resuspended in SEH buffer supplemented with inhibitors. 500 μg of mitochondrial proteins were dissolved in 6 M urea, 2 M thiourea, 20 mM Hepes, pH 7.4; reduced with dithiothreitol; alkylated with iodoacetamide; digested for 3 h with endoproteinase Lys-C (1:50, w/w) (Waco); diluted 4 times with 10 mM ammonium hydrogen carbonate; and further digested overnight with trypsin (1:50, w/w) (Promega). Trypsin cleaves peptide chains at the carboxyl side lysine or arginine except when either is followed by proline. Lys-C cleaves at lysine residues. The resulting peptide mixture was captured on TiO_2 beads (GL Sciences). Briefly, TiO_2 beads were preincubated with 2,5-dihydroxybenzoic acid (final concentration, 5 g/liter). About 10 mg of TiO_2 beads were added to each sample and incubated for 60 min at room temperature. After washing twice with 80% acetonitrile, 0.2% trifluoroacetic acid, peptides were eluted from the beads with 0.5% ammonium hydroxide solution in 40% acetonitrile (pH 10.5), almost completely dried in a vacuum centrifuge, and resuspended in 10 μl of 1% trifluoroacetic acid, 2% acetonitrile in water for LC-MS analysis.

Mass Spectrometry—Liquid chromatography was performed on a 1100 nano-HPLC system (Agilent) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher). Peptides (5 μl) were eluted over 140-min water/acetonitrile gradients. The LTQ-Orbitrap was operated in the positive ion mode with the following acquisition parameters. A full scan recorded in the Orbitrap analyzer (resolution, 60,000) was followed by MS/MS of the five most intense peptide ions in the LTQ analyzer. Multistage activation was enabled in all MS/MS events to improve fragmentation spectra of phosphopeptides. Raw MS spectra were processed in Quant.exe, the first module of our in-house built software MaxQuant (Version 1.0.13.13) (36). The derived peak list was searched with the MASCOT search engine (Matrix Science, London, UK) (Version 2.1.0). The following search criteria were used: tryptic specificity was required; carbamidomethylation was set as a fixed modification; oxidation (Met), *N*-acetylation (protein), and phosphorylation (Ser/Thr/Tyr) were set as variable modifications. We performed the MASCOT search against the International

Protein Index (IPI) mouse database Version 3.37 containing 51,292 proteins to which 175 commonly observed contaminants and all the reversed sequences had been added (37). Maximally two missed cleavages were allowed. Initial mass tolerance for precursor and fragments ions was 7 ppm and 0.5 Thompson, respectively. All spectra and all sequence assignments made by MASCOT were imported into MaxQuant.

The derived peptides and their assigned proteins were further processed in Identify.exe, the second module of MaxQuant. The posterior error probability and false discovery rate (FDR) were used for statistical evaluation. The FDR is derived from the number of identifications from reversed protein sequences. All phosphopeptide identifications suggested by MASCOT were filtered in MaxQuant by applying thresholds on MASCOT score, peptide length, and mass error. We accepted peptides based on the criteria that the number of forward hits in the database was at least 100-fold higher than the number of reversed database hits (incorrect peptide sequences), which gives an estimated FDR of less than 1%. To achieve highly reliable identifications, the following criteria were used: maximal peptide posterior error probability of 0.1, maximal peptide FDR of 0.01, and minimal peptide length of 6.

In case the identified peptides of one protein included all peptides of another protein, these proteins (*e.g.* homologs and isoforms) were combined by MaxQuant and reported as one protein group. Phosphorylation sites were made non-redundant with regard to their surrounding sequence. The PTM score was used for assignment of the phosphorylation sites as described Ref. 18. Phosphosites fulfilling the following two criteria are defined as unambiguously identified sites: 1) their localization probability for the assignment is at least 0.75, and 2) the PTM score difference from the second possible localization assignment is 5 or higher. These cutoffs proved to yield highly accurate results in previous studies (18). The localization probability of such class I sites is almost always higher than 0.98. Phosphorylated peptides were uploaded to PHOSIDA (10). Finally, to further reduce the probability of including non-mitochondrial proteins in the data set, we retained only those proteins that were also defined as mitochondrial in the MitoCarta repository or in the gene ontology annotation.

Conservation Analysis—To derive homologous proteins between mouse and prokaryotic species, we used BLASTP (38) (supplemental Fig. 1). To check the phosphorylation site conservation, we generated global protein alignments between homologous protein pairs via Needle (39). To derive orthologs between eukaryotic species, we checked the phylogenetic relationships and conservation of triplets encoding phosphosites in comparison with triplets that encode non-phosphorylated serines or threonines localized on exposed loop structures of phosphorylated proteins throughout 37 eukaryotes on the basis of cDNA alignments as provided by Ensembl.

Our previous large scale phosphoproteomics study demonstrated that phosphorylation sites are predominantly located in non-regularly structured regions on the protein surface (10). Therefore, the surrounding sequence regions may diverge to such an extent that the structural effect (fast sequence evolution in loop regions) effectively competes with the constraining pressure of function (slow sequence evolution of kinase substrate motifs). To correctly assess the degree of conservation of phosphosites, it is therefore important to take the structural effect into account. We did this by choosing only sites located in loop regions according to SABLE (40) predictions for the comparison set, which should isolate the functional, evolutionary constraints on the phosphosite itself.

Bootstrapping—We created a bootstrap distribution for a given species from 10,000 sets of randomly selected human proteins annotated as "known" in Ensembl. Each random set contained as many proteins as the given phosphoset. For each set, we calculated the proportion of orthologs in the chosen species, and the histogram of these 10,000 proportions provided the bootstrap distribution. The resulting histograms reflecting the distribution of logarithmized proportions of orthologs illustrate the significant difference between the set of interest (phosphoset) and randomly selected sets.

Next, we created a bootstrap distribution for a given species from 10,000 sets of serines, threonines, and tyrosines that were randomly selected from phosphorylated proteins that had an ortholog. Only those residues were included in the analysis that showed the same predicted structural constraints as phosphorylated residues (high accessibility and localization in loops). For each set, we calculated the proportion of conserved residues in the chosen species, and the histogram of these 10,000 proportions provided the bootstrap distribution. Resulting histograms were illustrated using MatLab (MathWorks).

RESULTS

The basis of our evolutionary studies is high accuracy phosphoproteomes that have been published during the last 3 years and that are deposited in PHOSIDA (10). The data were acquired using the technology described in detail in Olsen *et al.* (18). Briefly, cellular proteomes are digested to peptides, and phosphopeptides are enriched using a TiO_2 metal affinity matrix in the presence of 2,5-dihydrobenzoic acid (41) and measured by tandem mass spectrometry on a linear ion trap Orbitrap mass spectrometer using multistage activation (42). Phosphopeptide assignment including localization of the phosphogroups to particular amino acids is performed in MaxQuant (36). Mass accuracies are in the ppb range, and the false discovery rate of each data set is lower than 1%. Note that the high accuracy of detection of phosphorylation sites refers to the minimization of false positives. The percentage of false negatives, on the other hand, is not known in phosphoproteomics data sets. Site-specific localization assignments of the phosphogroup peptides from MS/MS fragmentation spectra were determined in MaxQuant, which contains an algorithm to predict phosphorylation site localization (18, 43). We used a threshold probability of 75%; however, median localization scores on phosphopeptides were typically about 98% (44). Eleven high quality phosphorylation sets from nine species that are available in PHOSIDA were used in our evolutionary analysis. Together they comprise 39,574 phosphorylation sites from *E. coli* (14), *B. subtilis* (13), *L. lactis* (15), *H. salinarum* (17), *Saccharomyces cerevisiae* (46), *Caenorhabditis elegans* (44), *Drosophila melanogaster* (47), *Mus musculus* (48, 49) and *Homo sapiens* (18, 50).

The generation of each phosphoproteome with the same generic work flow ensured that analysis on the phosphoproteome was performed to a comparable depth for each data set. Because no mitochondrial phosphorylation data set was available that was acquired in the same way, we performed a high accuracy large scale mass spectrometry study of the mitochondrial phosphoproteome as described below. This extends the compendium of evolutionary data for these phos-

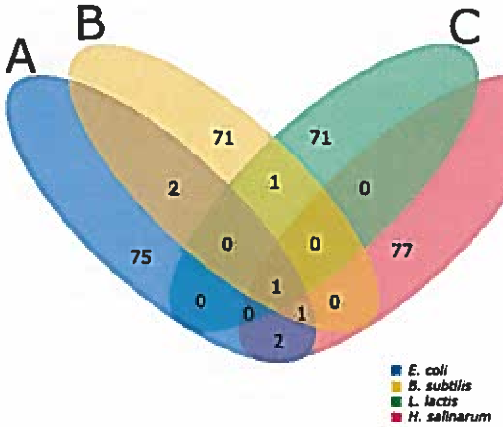


Fig. 1. **Overlap of prokaryotic phosphorylation sites.** The overlap between phosphorylation sites of *E. coli* (*A*), *B. subtilis* (*B*), *L. lactis* (*C*), and *H. salinarum* (*D*) is very low.

phoproteomes and in particular allowed us to analyze the mitochondrial phosphoproteome, the link between the domains of life, including its conservation.

Conservation of Prokaryote and Eukaryote Phosphoproteomes—The size of the phosphoproteomes of the four prokaryotes analyzed here are remarkably similar. They range from 73 sites for *L. lactis* to 81 sites for *E. coli* (www.phosida.com). Note that these numbers do not reflect the true size of the prokaryote phosphoproteome but rather the extent that can be readily probed with current technology. Nevertheless, the fact that the same technology results in about the same size phosphoproteome clearly indicates that the phosphoproteome is similarly sparse across different prokaryotic species.

Furthermore, the prokaryotic phosphoproteomes hardly overlap with each other. We found that a considerable proportion of proteins including elongation factors and glycolytic enzymes are phosphorylated in prokaryotes and highly conserved throughout the domain. However, these proteins, although conserved in prokaryotes, are not phosphorylated in all prokaryotic species, and if a given phosphoprotein occurs in more than one species, the phosphorylation sites do not overlap (Fig. 1).

In eukaryotes, phosphorylation regulates many key processes including cell growth, proliferation, differentiation, and immune response (51, 52), many of which do not occur in prokaryotes. Reflecting the fundamentally different biological roles of phosphorylation in eukaryotes and prokaryotes, the phosphoproteomes obtained with current technology are vastly larger. Using the technology described here they range from more than 3500 sites for yeast (46) to more than 20,000 sites for mammalian species such as human (23).

In previous bioinformatics analyses of the eukaryotic phosphoproteomes, we determined that phosphorylation sites, especially those on serine and threonine, tend to be located in fast evolving loop and hinge regions of proteins. We considered the PHOSIDA serine/threonine/tyrosine phosphoproteomes of mouse and other eukaryotes including human, fly, worm, and yeast, representing the most comprehensive evolutionary study on eukaryotic phosphoproteomes so far. The results on the protein level showed commonality between all eukaryotes. Within the eukaryotic domain, identified phosphoproteins have more orthologs than non-phosphorylated proteins (supplemental Table 1). This held true in a large variety of experimental systems throughout the eukaryotic domain. For each analyzed species, based on a bootstrap approach, the proportion of orthologous phosphoproteins to all phosphoproteins was clearly outside the distribution, and the resulting χ^2 values indicated extremely high statistical significance (see "Experimental Procedures").

Analysis of phosphoproteome conservation at the level of individual sites is potentially much more informative than analysis of entire substrate proteins. This is because secondary effects such as protein abundance cannot skew comparison between sites on orthologous proteins. More importantly, individual phosphorylation sites are specific substrates of one or more kinases and phosphatases, and they mediate the functionality of this post-translational modification.

To study the conservation of phosphorylation that occurs throughout the entire mouse cell and other eukaryotic cells at the site level, we created bootstrap distributions by repeated random selections of non-phosphorylated serines, threonines, and tyrosines from proteins of each species from the Ensembl database. For these thousands of random sets, the proportion of conserved residues in the given species was derived, and the histogram of the proportions of conserved sites then provided the bootstrap distribution. The χ^2 test assigns a p value to the difference between the proportion of conserved and non-conserved phosphosites and the proportion of conserved and non-conserved counterparts (non-phosphorylated Ser/Thr/Tyr). In some cases, the conservation of phosphothreonines differs from the conservation of phosphoserines, probably because of the roughly 10-fold lower number of phosphothreonines in the data sets. For this reason, only the analysis on serine residues is discussed in more detail. For none of the investigated organisms was it possible to find a significant pattern regarding the conservation of highly accessible phosphotyrosines in alignments of orthologous proteins because of the low number of these residues present in the data sets.

We found that the proportion of conserved phosphorylation sites was clearly outside the bootstrap distribution, and the calculated χ^2 values were higher than 6, which corresponds to p values lower than 0.01, for all analyzed species except for yeast and worm (supplemental Table 2). The yeast conservation study showed that ~13% of phosphorylated serines were conserved in human orthologs, the same number as for non-phosphorylated serines. The same trend was observed in worm. In these two species, the p values was not statistically significant, which indicates that phosphorylated and non-phosphorylated residues have the same degree of conservation.

Interestingly, in fly, which has roughly the same phylogenetic distance to mammals as worm, the phosphorylated serines were more conserved than non-phosphorylated serines throughout higher eukaryotes (*e.g.* 20% of fly phosphoserines and 16% of non-phosphorylated fly serines were conserved in human) (supplemental Fig. 2). The same pattern was observed for all mammalian phosphoproteomes. Note that conservation at the residue level can be very high between mammalian species in general. For example, 92% of human serines are conserved in proteins that are orthologous in chimpanzee and humans. However, even in this case, our study of human phosphosites (18) found that the conservation of phosphorylated residues is even higher; for example, 97% of human phosphoserines are conserved in chimpanzee alignments (supplemental Fig. 3). The different mouse and other mammalian phosphodata sets resulted in similar numbers.

Together, phosphorylation sites of fly, mouse, and human are more conserved than non-phosphosites of the same proteins with the same structural features throughout higher eukaryotes. In contrast, yeast and worm phosphorylation sites are not highly conserved with respect to higher eukaryotes. The bootstrapping approach and the χ^2 test verified the statistical significance of these observations. They agree with recent findings that many of the known human kinases evolved after the divergence between yeast and higher eukaryotes (54). Therefore, there are a considerable number of yeast-specific kinases that are not present in higher eukaryotes and vice versa. The worm kinome also differs significantly from kinomes of other eukaryotes. It is 2 times larger than the fly kinome, and half of the worm kinases are nematode-specific. This shows that the majority of phosphorylation events have evolved after the divergence of higher eukaryotes from yeast, which is in concordance with the unusually large number of nematode-specific kinases.

Characterization of Mitochondrial Phosphoproteome—To obtain a map of mitochondrial phosphorylated proteins across different cell types in basal conditions, we isolated and purified mitochondria from different mouse cell lines (C2C12, Hepa 1-6, 3T3-L1, and brown adipose tissue) and from mouse primary brown adipocytes. From these mitochondrial fractions, we enriched phosphorylated peptides using TiO_2 chromatography following the work flow of Olsen *et al.* (18). Peptides were analyzed via nano-LC-MS/MS on the high mass accuracy LTQ-Orbitrap. Raw data were processed with in-house developed MaxQuant software (36).

To further distinguish between mitochondrial proteins and the contaminants resulting from the biochemical preparation, mitochondrial localization of these proteins was manually verified against the MitoCarta database (55) and gene ontology localization description. In total, we identified and confidently assigned 174 phosphorylation sites on 74 mitochondrial proteins with an estimated false positive rate of less than 1% for phosphopeptide identification and median phosphorylation localization probability greater than 99.9% (Table I and supplemental Table 3). The median MASCOT score and the median PTM score were 47.36 and 147.94, respectively. The median absolute mass deviation was 0.29 ppm. Further parameters for quality assurance in MaxQuant are described by Cox and Mann (36). The MitoCarta database comprises around 1100 mitochondrial proteins. Thus, the proportion of phosphorylated to all mitochondrial proteins is around 7%, which is similar to that of bacterial phosphoproteomes. In *E. coli*, for example, around 100 of 4000 (3%) proteins are known to be phosphorylated. In comparison, more than half of all proteins in eukaryotic cells are estimated to be phosphorylated.

Around 40% of the identified mitochondrial phosphosites were identified in at least two cell types (supplemental Fig. 4); this is a relatively high value given the comparatively low number of phosphorylation sites. The phosphorylation distribution on serines, threonines, and tyrosines was 89.7, 9.2, and 1.1%, respectively (Fig. 2*A*). This distribution is similar to the distributions we previously measured for eukaryotic and prokaryotic cells (18, 20). Ptpra (receptor-type tyrosine-protein phosphatase α precursor) and Pgrmc1 (receptor-type tyrosine-protein phosphatase α) were found to be phosphorylated on tyrosine residues. Overall, 80 of 174 phosphorylation sites are novel according to the Swiss-Prot database. For example, elongation factor Ts (Ser-269), pyruvate carboxylase (Ser-22, Ser-143, and Ser-1033), and long chain-specific acyl-CoA dehydrogenase (Ser-55) are not annotated to be phosphorylated on these sites in the Swiss-Prot database. Overall, 11 of 84 mitochondrial phosphorylation sites identified by Lee *et al.* (29) overlap with our set.

Gene ontology analysis shows that the mitochondrial phosphoproteome is enriched for proteins that are involved in protein binding, transporter activity, ATP binding, nucleotide binding, and ribonucleotide binding (Fig. 2*B*) compared with the entire mouse proteome. Supplemental Table 4 lists all overrepresented gene ontology categories along with the corresponding mitochondrial phosphoproteins based on the DAVID analysis tool (56). Furthermore, we checked the surrounding sequences of the identified mitochondrial phosphosites to derive the putative corresponding kinases. The phosphorylation sites match significantly with motifs of various kinases including casein kinase II, calcium/calmodulin-dependent protein kinase type II, and protein kinase D (supplemental Table 5). The wide spectrum of corresponding kinases is also reflected in the overall position-specific amino acid frequency as illustrated in Fig. 2C (57).

Evolutionary Conservation of Mitochondrial Phosphoproteome in Prokaryotes—Mitochondria and bacteria probably share an α-proteobacterial ancestor and are therefore evolutionarily related. Phylogenetic analyses suggest that the closest relative of the mitochondrial ancestor is *Rickettsia*, an

Table I
Mitochondrial mouse phosphorylation sites

IPI accession no.	UniProt accession no.	Gene symbol	Phosphosites
IPI00108147	P70335	*Rock1*	Ser-1102, Ser-1105
IPI00108685	P18052	*Ptpra*	Tyr-825
IPI00109033	Q3UJK3	*Fam54b*	Ser-100, Ser-235
IPI00111770	Q06185	*Atp5i*	Ser-68
IPI00112339	Q9ERG0	*Lima1*	Ser-367, Ser-372, Ser-374, Ser-488, Ser-580, Ser-581, Ser-615, Ser-617, Ser-620
IPI00113052	Q9CZR8	*Tsfm*	Ser-269
IPI00114209	P26443	*Glud1*	Ser-128, Thr-410
IPI00114710	Q05920	*Pc*	Ser-22, Ser-143, Ser-1033
IPI00116138	Q99KK9	*Hars2*	Ser-66
IPI00116414	Q99L88	*Sntb1*	Ser-86, Ser-218
IPI00117689	O54724	*Ptrf*	Ser-21, Ser-42, Ser-169, Ser-368, Ser-389, Ser-391
IPI00119114	P51174	*Acadl*	Ser-55
IPI00119320	Q91W06	*Tcirg*	Ser-693
IPI00120715	Q9DCC8	*Tomm20*	Ser-135, Ser-138
IPI00121190	Q01279	*Egfr*	Thr-695
IPI00122547	Q60932	*Vdac2*	Ser-116
IPI00122549	Q60932	*Vdac1*	Ser-117, Thr-120
IPI00123183	Q02013	*Aqp1*	Ser-262
IPI00123410	A2APH4	*Usp24*	Ser-2044
IPI00123709	Q9WTQ5	*Akap12*	Ser-631
IPI00123891	P97315	*Csrp1*	Ser-192
IPI00126253	Q924Z4	*Lass2*	Ser-341, Ser-346, Ser-348, Ser-349
IPI00126634	Q91VA6	*Poldip2*	Thr-292
IPI00126939	Q8C064	*Prkcdbp*	Ser-165, Ser-166, Ser-188, Ser-194
IPI00128522	P14602	*Hspb1*	Ser-86, Ser-87
IPI00130280	Q03265	*Atp5a1*	Ser-53, Ser-65, Ser-76, Ser-451
IPI00130444	Q9Z1T1	*Ap3b1*	Ser-276
IPI00132076	O88587	*Comt*	Ser-260, Ser-261
IPI00132478	Q9CQF4		Ser-106, Ser-116
IPI00133440	P67778	*Phb*	Ser-151
IPI00133608	Q9CRD0	*Ociad*	Ser-108, Ser-147
IPI00135660	Q63918	*Sdpr*	Ser-203, Ser-204, Ser-218, Ser-359, Ser-363
IPI00137194	P53986	*Slc16a1*	Ser-210, Ser-213, Ser-461, Ser-462, Ser-477, Ser-482, Ser-491
IPI00138190	P55288	*Cdh11*	Ser-714, Ser-788
IPI00139795	P99027	*Rplp2*	Ser-79, Ser-86, Ser-102, Ser-105
IPI00153842	Q9JLR9	*Higd1a*	Ser-8
IPI00169862	Q8K1Z0	*Coq9*	Thr-78, Ser-81
IPI00222496	Q922R8	*Pdia6*	Ser-433
IPI00228826	P54310	*Lipe*	Ser-77, Ser-600, Ser-602, Ser-694
IPI00229548	P51912	*Slc1a5*	Ser-509, Ser-521
IPI00230283	Q8K0D5	*Gfm1*	Ser-92
IPI00272690	P09470	*Ace*	Ser-1305
IPI00308161	Q8BGV8	*Smcr7l*	Ser-55, Thr-58, Ser-59
IPI00308885	P63038	*Hspd1*	Ser-70
IPI00315135	Q9CPQ3	*Tomm22*	Ser-15, Ser-45
IPI00317684	Q8CCJ3-3	*Kiaa0776*	Ser-458
IPI00318935	Q9D8T7	*Slirp*	Thr-104, Ser-105
IPI00319973	O55022	*Pgrmc1*	Tyr-180, Ser-181
IPI00321499	P59017	*Bcl2l13*	Ser-387, Thr-389
IPI00331555	P50136	*Bckdha*	Ser-338, Thr-339, Ser-348
IPI00337893	P35486	*Pdha1*	Thr-231, Ser-232, Ser-293, Ser-295, Ser-300
IPI00344090	Q5ND52	*Rnmtl1*	Ser-42
IPI00352475	Q3UHX2	*Pdap1*	Ser-60, Ser-63
IPI00356904	Q3UY58	*Farp1*	Thr-24, Ser-373, Ser-892
IPI00378438	Q3U0Q9	*Tns1*	Ser-790, Ser-792, Thr-1343, Ser-1346, Ser-1468, Ser-1470, Ser-1535

TABLE I—*continued*

IPI accession no.	UniProt accession no.	Gene symbol	Phosphosites
IPI00379441	XP_485703	*EG433968*	Thr-123
IPI00387392	Q9CQS4	*Slc25a46*	Thr-45
IPI00387480	Q8CGN5	*Plin*	Ser-130, Ser-174, Ser-384, Ser-410
IPI00408119	P27546	*Map4*	Ser-475, Ser-506, Ser-517, Ser-901, Ser-1046
IPI00453589	O70404	*Vamp8*	Ser-5
IPI00453688	Q8C0T5-1	*Sipa1l1*	Ser-1528
IPI00459443	P58871	*Tnks1bp1*	Thr-533, Ser-807, Ser-974, Ser-1370, Ser-1375, Ser-1611, Ser-1612, Ser-1657
IPI00480432	Q6ZQ33	*Rab11fip5*	Ser-307
IPI00551411	Q8BX70	*Vps13c*	Ser-871, Ser-873
IPI00556700	Q2YDW1	*Eif3j*	Ser-14, Ser-16
IPI00620800	Q3UM62	*Fam38a*	Thr-632, Ser-634
IPI00623114	Q6ZQ33	*Fat1*	Ser-4483
IPI00649184	Q9WTI7	*Myo1c*	Ser-408
IPI00649326	Q9JMH9	*Myo18a*	Ser-1950, Ser-1978, Ser-1982, Ser-2049, Ser-2051
IPI00678532	Q3UJU9	*Fam82c*	Ser-44, Ser-46
IPI00751137	Q3TSX8	*Tomm70a*	Ser-94, Ser-99
IPI00754876	Q3UFQ5	*Ttc7b*	Ser-732, Ser-733
IPI00757909	Q9D0L7	*Armc10*	Ser-41, Ser-43
IPI00848443	Q5SWU9	*Acaca*	Ser-63, Ser-647

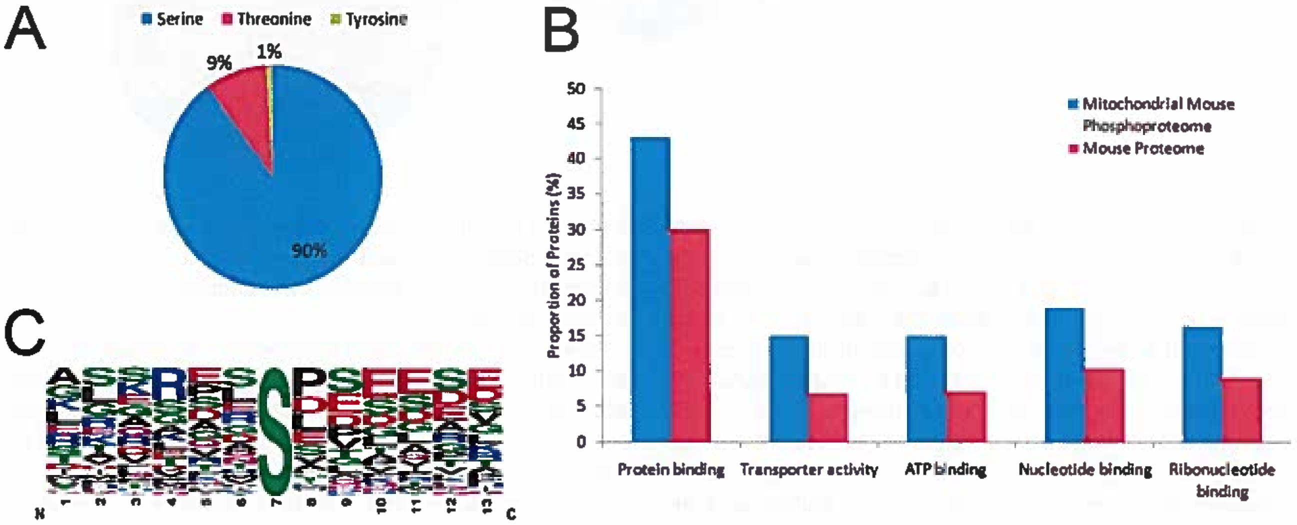


FIG. 2. **Characterization of mitochondrial mouse phosphoproteome.** *A*, serine/threonine/tyrosine distribution. *B*, molecular functions that are enriched in the mitochondrial phosphoset compared with the entire mouse proteome. *C*, relative position-specific amino acid frequency around phosphorylated sites.

obligate intracellular parasitic α-proteobacterium that could have initiated the endosymbiotic event (24). Our study provides independent evidence for the endosymbiotic hypothesis through the comparison of the conservation of mitochondrial *versus* non-mitochondrial proteins in prokaryotes (supplemental Fig. 5 and supplemental Table 6) (see "Experimental Procedures"). Mitochondrial mouse proteins are more highly conserved in prokaryotes than mouse proteins that are located in other organelles (Fig. 3). On average, around 35% of the mitochondrial proteins (MitoCarta data set) are conserved in any prokaryote in comparison with 13% of the non-mitochondrial proteins. The same pattern was observed for phosphorylated proteins. Overall, 28% of the phosphorylated mitochondrial proteins (from our set as well as from the Lee *et al.* (29) data set) are conserved in a prokaryotic species compared with 20% of the phosphorylated non-mitochondrial proteins. These observations are in concordance with the prokaryotic character of mitochondria expected from the endosymbiotic hypothesis.

However, the mitochondrial and the prokaryotic phosphoproteomes of *E. coli* (14), *B. subtilis* (13), and *L. lactis* (15) hardly overlap. Elongation factor G and elongation factor Ts were found to be phosphorylated in mitochondria as well as in all investigated bacteria. However, phosphorylation events on

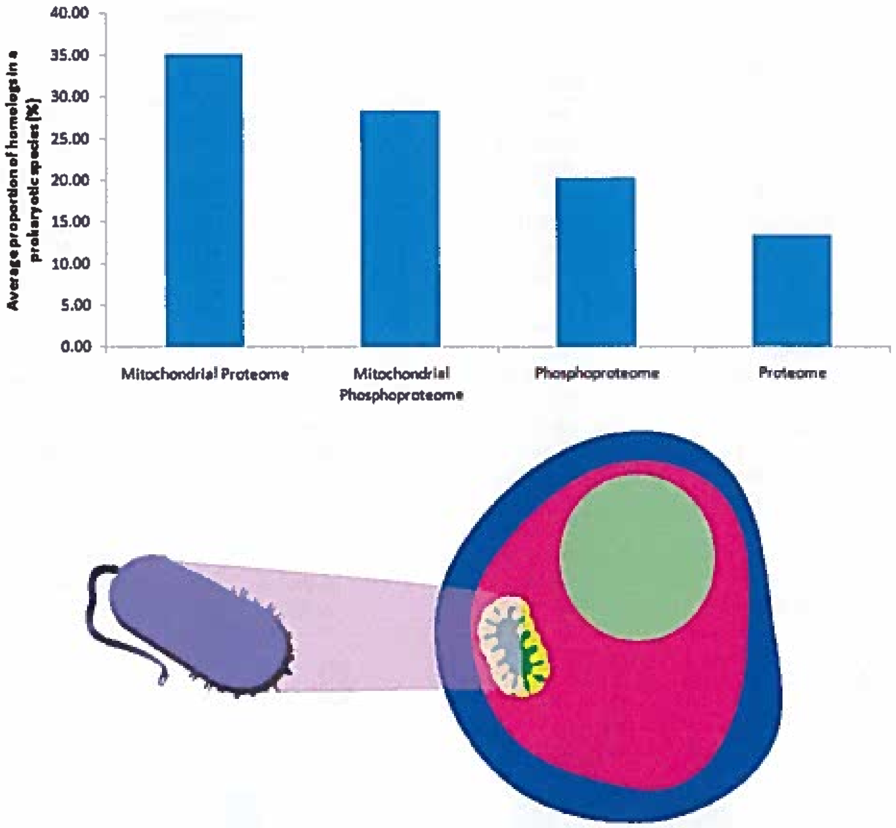


FIG. 3. **Proportion of mouse proteins that are orthologous to prokaryotic proteins.** Phosphorylated (our data set) and non-phosphorylated (MitoCarta data set) mitochondrial mouse proteins are more highly conserved in prokaryotes than phosphorylated (PHOSIDA data set) and non-phosphorylated (Swiss-Prot Database) proteins, which are located in other compartments of the cell. These observations are in concordance with the endosymbiotic scenario as illustrated in the *lower panel*. As a measure for conservation, we used the average proportion of orthologs in 62 prokaryotes.

these elongation factors were identified on different residues of the proteins. Other homologous proteins that were phosphorylated in mitochondria and in at least one of the selected bacteria were histidyl-tRNA synthetase, 60-kDa heat shock protein, and a "protein similar to nucleoside-diphosphate kinase." This protein is the only one with a phosphorylation site (threonine 123) that was found to be phosphorylated in mitochondria as well as in bacteria (serine 123 in *B. subtilis*). These findings are in agreement with a model that regulation of mitochondrial proteins via phosphorylation evolved after the endosymbiotic event.

Evolutionary Conservation of Mitochondrial Phosphoproteome in Eukaryotic Domain—In the eukaryotic evolutionary study, we studied the conservation of the mitochondrial phosphoproteome in 36 eukaryotic species that are contained in the Ensembl database and that span the eukaryotic domain. Based on χ^2 statistics, mitochondrial mouse phosphoproteins derived from our study and the one by Lee *et al.* (29) have significantly more homologs in the eukaryotic domain than other mouse proteins (supplemental Table 7). Even for yeast, the species most evolutionarily distant from mouse, this was the case.

Previous studies have already shown that non-mitochondrial phosphoproteins of other eukaryotes are also significantly conserved in a defined set of a few species (10). To check in more detail whether the high proportion of orthologs is a common feature of different phosphoproteomes in more detail, we studied the conservation of non-mitochondrial proteins of different phosphoproteomes in 36 eukaryotic species.

To study the conservation of mitochondrial phosphorylation at the site level, we derived the conservation of phosphorylated serines/threonines and non-phosphorylated serines/threonines of the same mitochondrial phosphoproteins in orthologous proteins of different eukaryotic species that are annotated in the Ensembl database. Because the identification of eukaryotic phosphoproteomes is far from being complete, we restricted our analysis to amino acid conservation, which means that in most of the cases there is no experimental evidence for phosphorylation of sites that are phosphorylated in mouse and conserved in another species. We found that mitochondrial phosphosites tend to be more conserved than non-phosphorylated serines/threonines in other eukaryotes (supplemental Table 8). Although the high conservation of mitochondrial phosphosites is evident in all eukaryotic species, the significance decreases for lower eukaryotes, probably because of the low number of homologous proteins in species such as *Ciona intestinalis*.

DISCUSSION

Despite a large number of phosphoproteome studies of diverse organisms, our knowledge of each of them is not comprehensive. The huge and continuing increase of measured mammalian phosphorylation sites makes it clear that the identification of the whole phosphoproteome is far from being complete. In fact, although previously one-third of all mammalian proteins were estimated to be phosphorylated, recent developments even suggest that the majority of all proteins are phosphorylated at least under some circumstances. As a practical matter, all evolutionary studies on phosphorylation are limited to the currently identified phosphoproteins. However, the number of known phosphosites turns out to very similar for different prokaryotic species and for mitochondria, whereas phosphoproteomes for eukaryotes range from over 3000 to more than 20,000 sites. This argues that large scale comparisons of phosphoproteomes can usefully be undertaken at this time. To this end, we took advantage of the recent availability of high resolution phosphorylation data and the PHOSIDA environment, which integrates biological context to quantify constraints of phosphorylation on a proteome-wide scale. We focused on phosphosets from PHOSIDA because it contains phosphoproteomes of a wide range of organisms all obtained with high accuracy and analyzed with the same stringent criteria. Other phosphorylation databases such as PhosphoSite and Phospho.ELM focus on mammalian phosphoproteomes only. Although they are more comprehensive than PHOSIDA, they also contain low resolution data sets with difficult to control overall false positive rates. More importantly, those data sets were obtained with widely diverging experimental work flows precluding comparative analysis. The data sets derived from PHOSIDA, although analyzed in the same way, are derived from different investigations. To exclude that this could cause any bias, we compared changing and non-changing subsets of the phosphoproteome, which showed that the results from our evolutionary study are valid for the whole phosphoproteome. For example, human phosphorylation sites that are unaffected by the stimulation of EGF show nearly the same conservation patterns as all other human phosphosites. Furthermore, our phosphoproteomes almost always contain the "basally phosphorylated peptide" for each phosphopeptide that is found to change upon a stimulus. This latter observation also argues that we have already covered a sizable proportion of the stimulus-specific phosphoproteome. We conclude that the general trends regarding the conservation of phosphorylated residues should already be contained in our analysis.

For all analyzed prokaryotes, the phosphoproteome is very sparse. Prokaryotic phosphorylation sites are hardly conserved, and only one site is identical in all four investigated prokaryotic phosphoproteomes. This finding is compatible with a model in which site-specific phosphorylation coevolved with the adaption of the bacterial species to their present-day ecological niches (15). The presence of phosphorylated prokaryotic proteins might also partially result from gene transfers as many eukaryote-like kinases have been found in prokaryotes by metagenomics projects (45). This suggests that regulation via phosphorylation was a relatively recent development that occurred late in prokaryotic evolution. Future studies will likely lead to the identification of larger prokaryotic phosphorylation data sets, but it seems unlikely that these studies will result in numbers comparable with eukaryotic phosphoproteomes or to a large overlap between prokaryotic phosphoproteomes.

In contrast to the mitochondrial and prokaryotic phosphoproteomes, which comprise a few hundred phosphosites, the application of high accuracy mass spectrometry to eukaryotic cells yields the identification of thousands of phosphorylation sites. A combined analysis of the phosphoproteomes of human, mouse, and fly demonstrates that phosphorylated serines and threonines are more highly conserved than non-phosphorylated serines and threonines that are also localized in loops or turns on the protein surface. In contrast, yeast and worm phosphorylation sites are not highly conserved with respect to higher eukaryotes. These observations are in concordance with recent findings that many of the known human kinases evolved after the divergence between yeast and higher eukaryotes (54). Therefore, there are a considerable number of yeast-specific kinases that are not present in higher eukaryotes and vice versa. The worm kinome also differs significantly from kinomes of other eukaryotes. It is 2 times larger than the fly kinome, and half of the worm kinases are nematode-specific.

In this study, we were particularly interested in the evolution of phosphorylation in the mitochondrion, the organelle that presents the phylogenetic link between prokaryotes and eukaryotes according to the endosymbiotic theory. Phosphorylated peptides from mitochondria were enriched from different mouse cell types (myotubes, hepatocytes, and adipocytes). The experiments were performed in cell lines (3T3-L1, C2C12, Hepa 1-6, and brown preadipocytes) and tissue samples (brown adipose tissue). These cells play a pivotal role in essential body functions such as energy expenditure, management, and storage, which are associated with main mitochondrial processes (*e.g.* pyruvate/acetyl-CoA utilization, oxidative phosphorylation, and fatty acid oxidation). The phosphoproteome of mouse mitochondria represents a rich source of novel roles of protein phosphorylation in this organelle and forms the basis for comparison of mitochondria with prokaryotes and eukaryotes.

Mitochondria from the different cell types were enriched by biochemical approaches. Proteins identified as phosphorylated by MS were further filtered based on validated mitochondria repositories and annotations (MitoCarta and gene ontology). This was done because all biochemical methods have some limitations, making it impossible to purify any organellar fraction to complete homogeneity. Without this filtering, phosphorylation sites originating from relatively minor cytosolic contaminations of the mitochondrial fraction would otherwise dominate the phosphoset, especially when adding the phosphoproteome from several cell types.

The measured mitochondrial phosphoproteome consists of 174 phosphorylation sites on 74 mitochondrial proteins. Although mitochondria are part of eukaryotic cells, our analysis thus clearly shows that in terms of the size of their phosphoproteome they are similar to prokaryotes. The phosphoproteome is enriched for essential proteins that are involved in binding and transport. The comparatively sparse mitochondrial phosphoproteome underlines the difficulty to identify phosphorylation sites in this organelle. Thus, the overlaps between different identified mitochondrial phosphoproteomes are not expected to be high. In fact, only 13% of our phosphosites overlap with data from a previous mitochondrial study by Lee *et al.* (29). In comparison, the overlap between our data sets derived from different cell types is relatively high (around 40%). This motivated us to confirm our conclusions with other mitochondrial data sets. We found that mitochondrial proteins are indeed relatively more conserved in prokaryotes than the non-mitochondrial eukaryotic proteome (Fig. 3). This is consistent with the endosymbiotic theory. However, the overlap between mitochondrial and bacterial phosphoproteomes is very low, which is in agreement with a model in which regulation of mitochondrial proteins via phosphorylation may have evolved after the endosymbiotic event. This finding is also in concordance with the fact that prokaryotic phosphoproteomes themselves have a very low overlap and with the idea that regulation by phosphorylation is a relatively recent introduction during evolution.

Mitochondrial phosphoproteins also proved to have significantly more orthologs in the eukaryotic domain. This was expected because of their essential roles in the eukaryotic cell. This pattern has also been reported for other phosphoproteomes.

Within the eukaryotic domain, we found that mitochondrial as well as non-mitochondrial phosphoproteins have more orthologs than non-phosphorylated proteins. This held true in a large variety of experimental systems; for example, for the analysis of mammalian phosphorylation conservation, we used data obtained from measurements of two different human cell lines as well as a mouse cell line and a mouse tissue. Our analysis of the global alignments of orthologs in 37 eukaryotes shows that mitochondrial mouse phosphorylation sites are more conserved than non-phosphosites of the same proteins with the same structural features throughout higher eukaryotes. The significance of this observation decreases with more distantly related lower eukaryotes.

The overall picture that emerges from our evolutionary analysis of serine/threonine/tyrosine phosphorylation contains several major discontinuities. One is between bacteria and eukaryotes: bacteria have only a few percent of the phosphorylation events that eukaryotes have. In support of the endosymbiotic hypothesis of mitochondrial origin, we found this organelle to be sparsely phosphorylated, just like bacteria. The second discontinuity of the phosphoproteome is between yeast and other eukaryotes. Here our phosphoproteomics analysis parallels the notion that much of the signaling function of phosphorylation involves cell-cell communication. This is evident in the high conservation of phosphoproteins and phosphosites within metazoans but low conservation in a single cell organism. Furthermore, the worm phosphoproteome is poorly conserved throughout the eukaryotic domain, which is in concordance with its distinct kinome evolution and overrepresentation of phosphorylation events on substrates that are involved in the sex determination and development of *C. elegans* (44).

Taken together, our analysis establishes the phylogenetic study of the mitochondrial and non-mitochondrial phosphoproteome as a fruitful adjunct to the evolutionary study of the kinase-encoding genes. The recent increase in the number of studies on the evolution of phosphorylation and its regulation shows the current interest in this field. For example, a very recent comparison of phosphorylation patterns across yeast species showed that kinase-substrate interactions change 2 orders of magnitude more slowly than transcription factor-promoter interactions and that protein kinases are important for the phenotypic diversity (53).

In the future, the ever increasing quality and depth of proteomics studies will help to identify more complete phosphoproteomes and will also provide important quantitative information, for example on phosphorylation site stoichiometry. Combined with functional studies at the site level, which is still a major bottleneck, as well as more in-depth knowledge of kinase-substrate relationships, this will allow the derivation of a more detailed picture of the evolution of phosphorylation and its role in the cell. Nevertheless, the major features of the evolution of the kinome (54) and the phosphoproteome are already clear.

Acknowledgments—We thank Johannes Graumann, Juergen Cox, Albert Vilella Bertran, and Sudhakaran Prabhakaran for fruitful discussions.

* This work was supported, in whole or in part, by National Institutes of Health Grant R01 GM081578-02 on "Complex dynamics in multisite phosphorylation." This work was also supported by the Max Planck Society and the HepatoSys Network.

[S] This article contains supplemental Figs. 1–5 and Tables 1–8.

‖ To whom correspondence should be addressed. Tel.: 49-89-8578-2557; Fax: 49-89-8578-2219; E-mail: mmann@biochem.mpg.de.

REFERENCES

1. Hunter, T. (2000) Signaling—2000 and beyond. *Cell* 100, 113–127
2. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207
3. Salomon, A. R., Ficarro, S. B., Brill, L. M., Brinker, A., Phung, Q. T., Ericson, C., Sauer, K., Brock, A., Hom, D. M., Schultz, P. G., and Peters, E. C. (2003) Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* 100, 443–448
4. Bimey, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E.,

Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T., and Clamp, M. (2004) An overview of Ensembl. *Genome Res.* 14, 925–928

5. Nita-Lazar, A., Saito-Benz, H., and White, F. M. (2008) Quantitative phosphoproteomics by mass spectrometry: past, present, and future. *Proteomics* 8, 4433–4443
6. Mann, M., and Kelleher, N. L. (2008) Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18132–18138
7. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villén, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12130–12135
8. Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 4, 1551–1561
9. Diella, F., Cameron, S., Gem0nd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N., and Gibson, T. J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 5, 79
10. Gnad, F., Ren, S., Cox, J., Olsen, J. V., Macek, B., Oroshi, M., and Mann, M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* 8, R250
11. Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A., Jørgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., Yaffe, M. B., and Pawson, T. (2007) Systematic discovery of in vivo phosphorylation networks. *Cell* 129, 1415–1426
12. Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M. B., and Pawson, T. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* 36, D695–D699
13. Macek, B., Mijakovic, I., Olsen, J. V., Gnad, F., Kumar, C., Jensen, P. R., and Mann, M. (2007) The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. *Mol. Cell. Proteomics* 6, 697–707
14. Macek, B., Gnad, F., Soufi, B., Kumar, C., Olsen, J. V., Mijakovic, I., and Mann, M. (2008) Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol. Cell. Proteomics* 7, 299–307
15. Soufi, B., Gnad, F., Jensen, P. R., Petranovic, D., Mann, M., Mijakovic, I., and Macek, B. (2008) The Ser/Thr/Tyr phosphoproteome of Lactococcus lactis IL1403 reveals multiply phosphorylated proteins. *Proteomics* 8, 3486–3493
16. Eymann, C., Becher, D., Bernhardt, J., Gronau, K., Klutzny, A., and Hecker, M. (2007) Dynamics of protein phosphorylation on Ser/Thr/Tyr in Bacillus subtilis. *Proteomics* 7, 3509–3526
17. Aivaliotis, M., Macek, B., Gnad, F., Reichelt, P., Mann, M., and Oesterhelt, D. (2009) Ser/Thr/Tyr protein phosphorylation in the archaeon Halobacterium salinarum—a representative of the third domain of life. *PLoS One* 4, e4777
18. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127, 635–648
19. Villén, J., Beausoleil, S. A., Gerber, S. A., and Gygi, S. P. (2007) Large-scale phosphorylation analysis of mouse liver. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1488–1493
20. Macek, B., Mann, M., and Olsen, J. V. (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol. Toxicol.* 49, 199–221
21. Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D., Lam, H., Schmidt, A., Rinner, O., Mueller, L. N., Shannon, P. T., Pedrioli, P. G., Panse, C., Lee, H. K., Schlapbach, R., and Aebersold, R. (2007) PhosphoPep—a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.* 3, 139
22. Cohen, P. (2001) The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur. J. Biochem.* 268, 5001–5010
23. Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A., Brunak, S., and Mann, M. (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* 3, ra3
24. Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C., Podowski, R. M., Näslund, A. K., Eriksson, A. S., Winkler, H. H., and Kurland, C. G. (1998) The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* 396, 133–140
25. Neupert, W., and Herrmann, J. M. (2007) Translocation of proteins into mitochondria. *Annu. Rev. Biochem.* 76, 723–749
26. Thomson, M. (2002) Evidence of undiscovered cell regulatory mechanisms: phosphoproteins and protein kinases in mitochondria. *Cell. Mol. Life Sci.* 59, 213–219
27. Lewandrowski, U., Sickmann, A., Cesaro, L., Brunati, A. M., Toninello, A., and Salvi, M. (2008) Identification of new tyrosine phosphorylated proteins in rat brain mitochondria. *FEBS Lett.* 582, 1104–1110
28. Hopper, R. K., Carroll, S., Aponte, A. M., Johnson, D. T., French, S., Shen, R. F., Witzmann, F. A., Harris, R. A., and Balaban, R. S. (2006) Mitochondrial matrix phosphoproteome: effect of extra mitochondrial calcium. *Biochemistry* 45, 2524–2536
29. Lee, J., Xu, Y., Chen, Y., Sprung, R., Kim, S. C., Xie, S., and Zhao, Y. (2007) Mitochondrial phosphoproteome revealed by an improved IMAC method and MS/MS/MS. *Mol. Cell. Proteomics* 6, 669–676
30. Deng, N., Zhang, J., Zong, C., Wang, Y., Lu, H., Yang, P., Wang, W., Young, G. W., Wang, Y., Korge, P., Lotz, C., Doran, P., Liem, D. A., Apweiler, R., Weiss, J. N., Duan, H., and Ping, P. (September 7, 2010) Phosphoproteome analysis reveals regulatory sites in major pathways of cardiac mitochondria. *Mol. Cell. Proteomics* 10.1074/mcp.M110.000117
31. Reinders, J., Wagner, K., Zahedi, R. P., Stojanovski, D., Eyrich, B., van der Laan, M., Rehling, P., Sickmann, A., Pfanner, N., and Meisinger, C. (2007) Profiling phosphoproteins of yeast mitochondria reveals a role of phosphorylation in assembly of the ATP synthase. *Mol. Cell. Proteomics* 6, 1896–1906
32. Kratchmarova, I., Kalume, D. E., Blagoev, B., Scherer, P. E., Podtelejnikov, A. V., Molina, H., Bickel, P. E., Andersen, J. S., Fernandez, M. M., Bunkenborg, J., Roepstorff, P., Kristiansen, K., Lodish, H. F., Mann, M., and Pandey, A. (2002) A proteomic approach for identification of secreted proteins during the differentiation of 3T3-L1 preadipocytes to adipocytes. *Mol. Cell. Proteomics* 1, 213–222
33. Fasshauer, M., Klein, J., Kriauciunas, K. M., Ueki, K., Benito, M., and Kahn, C. R. (2001) Essential role of insulin receptor substrate 1 in differentiation of brown adipocytes. *Mol. Cell. Biol.* 21, 319–329
34. Forner, F., Foster, L. J., Campanaro, S., Valle, G., and Mann, M. (2006) Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol. Cell. Proteomics* 5, 608–619
35. Forner, F., Arriaga, E. A., and Mann, M. (2006) Mild protease treatment as a small-scale biochemical method for mitochondria purification and proteomic mapping of cytoplasm-exposed mitochondrial proteins. *J. Proteome Res.* 5, 3277–3287
36. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367–1372
37. Elias, J. E., Haas, W., Faherty, B. K., and Gygi, S. P. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* 2, 667–675
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
39. Needleman, S. B., and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453
40. Wagner, M., Adamczak, R., Porollo, A., and Meller, J. (2005) Linear regression models for solvent accessibility prediction in proteins. *J. Comput. Biol.* 12, 355–369
41. Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P., and Jørgensen, T. J. (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell. Proteomics* 4, 873–886
42. 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

43. Olsen, J. V., and Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 13417–13422
44. Zielinska, D. F., Gnad, F., Jedrusik-Bode, M., Wisniewski, J. R., and Mann, M. (2009) C. elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J. Proteome Res.* **8,** 4039–4049
45. Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S., Li, H., Mashiyama, S. T., Joachimiak, M. P., van Belle, C., Chandonia, J. M., Soergel, D. A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B. J., Bafna, V., Friedman, R., Brenner, S. E., Godzik, A., Eisenberg, D., Dixon, J. E., Taylor, S. S., Strausberg, R. L., Frazier, M., and Venter, J. C. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5,** e16
46. Gnad, F., de Godoy, L. M., Cox, J., Neuhauser, N., Ren, S., Olsen, J. V., and Mann, M. (2009) High accuracy identification and bioinformatic analysis of in-vivo protein phosphorylation sites in yeast. *Proteomics* **9,** 4642–4652
47. Hilger, M., Bonaldi, T., Gnad, F., and Mann, M. (2009) Systems-wide analysis of a phosphatase knock down by quantitative proteomics and phosphoproteomics. *Mol. Cell. Proteomics* **8,** 1908–1920
48. Pan, C., Gnad, F., Olsen, J. V., and Mann, M. (2008) Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics* **8,** 4534–4546
49. Zanivan, S., Gnad, F., Wickström, S. A., Geiger, T., Macek, B., Cox, J., Fässler, R., and Mann, M. (2008) Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. *J. Proteome Res.* **7,** 5314–5326
50. Daub, H., Olsen, J. V., Bairlein, M., Gnad, F., Oppermann, F. S., Körner, R., Greff, Z., Kéri, G., Stemmann, O., and Mann, M. (2008) Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Mol. Cell* **31,** 438–448
51. Schlessinger, J. (2000) Cell signaling by receptor tyrosine kinases. *Cell* **103,** 211–225
52. Pawson, T., and Scott, J. D. (2005) Protein phosphorylation in signaling—50 years and counting. *Trends Biochem. Sci.* **30,** 286–290
53. Beltrao, P., Trinidad, J. C., Fiedler, D., Roguev, A., Lim, W. A., Shokat, K. M., Burlingame, A. L., and Krogan, N. J. (2009) Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol.* **7,** e1000134
54. Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S. (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* **27,** 514–520
55. Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A., and Mootha, V. K. (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112–123
56. Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57
57. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190

CURRICULUM VITAE

DOROTA ZIELINSKA

27 July 1983
Polish

EDUCATION

1.11.2008 – 27.07.2011 *Munich, Germany*	Max Planck Institute of Biochemistry (PhD) *Unveiling the eukaryotic N-glycoproteome*
01.09.2007 – 30.06.2008 *Stockholm, Sweden*	Royal Institute of Technology (Master Thesis) *Design of a lipase with low affinity towards water as nucleophile*
01.10.2006 – 30.06.2008 *Vienna, Austria*	University of Natural Resources and Applied Life Sciences (Master's degree study in Biotechnology)
01.10.2003 – 30.06.2006 *Vienna, Austria*	University of Natural Resources and Applied Life Sciences (Bachelor's degree study in Food and Biotechnology)

INTERNSHIPS

01.03.2006 – 31.03.2006 *Vienna, Austria*	University of Natural Resources and Applied Life Sciences, Department of Biochemistry, *Glycomic and proteomic investigation of glycoproteins*
01.02.2006 – 28.02.2006 *Vienna, Austria*	University of Natural Resources and Applied Life Sciences, Department of Biochemistry *Bachelor's thesis: Analysis and structural determination of N-glycosidically linked oligosaccharides*
01.08.2005 – 30.09.2005 *Poznan, Poland*	Jutrzenka S.A., Department of Technology *Sensor assessment & Controlling*
01.06.2005 – 30.06.2005 *Vienna, Austria*	University of Veterinary Medicine, Department of Livestock and Genetics *Embryo manipulation and in-vitro fertilisation*
01.02.2005 – 28.02.2005 *Vienna, Austria*	University of Veterinary Medicine, Department of Medical Chemistry *Protein purification and determination*

AWARDS

Junior Scientist Award (2011)
Young Scientist Award (2010)
Master's degree passed with honours (2008)
Swedish Institute Scholarship (2007)
Finalist of the V National Polish Chemistry Olympics (1998)
First place in the provincial Chemistry competition in Wielkopolska (1998)

PUBLICATIONS

Zielinska DF, Gnad F, Wiśniewski JR, Mann M (2010), *Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints*, **Cell** , May 28; 141(5) 897-907.

Ostasiewicz P, **Zielinska DF**, Mann M, Wiśniewski JR (2010), *Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry,* **J Proteome Res**, Jul 2; 9(7) 3688-700.

Wiśniewski JR, **Zielinska DF**, Mann M, *Comparison of Ultrafiltration Units for Proteomic and N-Glycoproteomic Analysis by the FASP Method*, Analytical Biochemistry, **Anal Biochem,** Mar 15; 410(2) 307-9

Zielinska DF, Gnad F, Jedrusik-Bode M, Wiśniewski JR, Mann M (2009), *Caenorhabditis elegans Has a Phosphoproteome Atypical for Metazoans That Is Enriched in Developmental and Sex Dermination Proteins*, **J Proteome Res**, Aug 7; 8(8) 4039-49.

Gnad F, Forner F, **Zielinska DF**, Birney E, Gunawardena J, Mann M (2010), *Evolutionary Constraints of Phosphorylation in Eukaryotes, Prokaryotes and Mitochondria*, **Mol Cell Proteomics**, Dec 9; 9(12) 2642-53.

Larsen MW, **Zielinska DF**, Martinelle M, Hidalgo A, Jensen LJ, Bornscheuer UT, Hult K (2010), *Suppression of Water as a Nucleophile in Candida antarctica Lipase B Catalysis*, **Chembiochem**, Apr 2; 11(6) 796-801.

Zielinska DF, Gnad F, Schropp K, Wiśniewski JR, Mann M, *Evolution of N-glycosylation* (submitted)

Zielinska DF, Ostasiewicz P, Gnad F, Duś K, Mann M, Wiśniewski JR, *Proteomic and N-glycoproteomic profiling of colorectal cancer* (in preparation)