research articles Journal of proteome research

Targeted Analysis of Protein Termini

Wilma Dormeyer,† Shabaz Mohammed,† Bas van Breukelen, Jeroen Krijgsveld, and Albert J. R. Heck*

Department of Biomolecular Mass Spectrometry, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Sorbonnelaan 16, 3584 CA Utrecht, the Netherlands

Received June 15, 2007

We describe a targeted analysis of protein isoforms by selective enrichment and identification of in vivo acetylated protein N-termini and protein C-termini. Our method allows the characterization of these protein termini regardless of their annotation in protein databases and requires no chemical derivatization. Using an iterative database search strategy that takes account of the enrichment protocol, 263 IPI annotated and 87 unpredicted acetylated N-termini were identified in the crude membrane fraction of human embryonic carcinoma cells. The N-acetylated peptides conform to the reported criteria for in vivo modification. In addition, 168 IPI annotated and 193 unpredicted C-termini were identified. Additionally, and for the first time, we also report on in vivo N-terminal propionylation. The significant number of unknown protein N- and C-termini suggests a high degree of novel transcription independent of annotated gene boundaries and/or specific protein processing. Biological relevance of several of these unpredicted protein termini could be curated from the literature, adding further weight to the argument to go beyond routine database search strategies. Our method will improve the correct annotation of genes and proteins in databases.

Keywords: N-Terminal acetylation • N-terminal propionylation • gene annotation • database search strategies • bioinformatics

Introduction

For several reasons, the human proteome is much more diverse than the genome. For instance, proteins with altered termini are created by alternative initiation of transcription within genes or by transcription independent of annotated gene boundaries.[1–3] Recent analyses of the human genome showed that alternative initiation of transcription takes place in an unexpectedly large number of genes.[4] Cell-specific mRNA splicing[5–7] and enzymatic protein processing[8–10] can further alter the termini of proteins. For the definition of the human proteome, it is therefore essential to identify and annotate the termini of proteins.

Here we focus on the targeted analysis of alternatively terminated protein isoforms, that is, proteins that are coded by the same gene and have identical sequences but shortened N- or C-termini. In such an analysis, it must be considered that 80% of cellular proteins are estimated to be N-acetylated in vivo.[11,12] Several "terminal" proteomics approaches have been described.[8,13–17] Many of these approaches are labor-intensive since they require extensive chemical modification and/or repeated separation of the sample. Gevaert et al. used combined fractional diagonal chromatography (COFRADIC) in which free amino groups are chemically acetylated on the protein level, and N-termini of tryptic cleavage products are blocked with trinitrobenzene (TNBS) on the peptide level. N-Acetylated peptides could then be separated from TNBS-blocked peptides by RP-HPLC.[15] McDonald et al. chose a similar procedure but blocked the N-termini of the cleavage products by biotinylation. N-Acetylated peptides were subsequently recovered by subtractive binding of biotinylated peptides to streptavidin.[18] Both approaches allow the identification of in vivo N-terminal acetylated proteins by omitting the acetylation step or the discrimination between naturally and chemically α-acetylated N-terminal peptides if stable isotopic labeled reagents were used for the acetylation of free amino groups.

An alternative strategy for the enrichment of N-acetylated and C-terminal peptides is to take advantage of their reduced basicity in comparison to non-N-acetylated and tryptic peptides. Gorman et al. have shown that strong cation exchange (SCX) fractionation can prefractionate blocked N-termini and unmodified C-termini in tryptic digests of small quantities of viral proteins.[19] The same principle has been demonstrated to be highly efficient for phosphopeptide enrichment. In such phospho-proteomic studies, modified N-terminal and unmodified C-terminal peptides coeluted with phosphopeptides but were not analyzed in detail.[20–22] Here, we fully exploited SCX fractionation for the selective enrichment and identification of in vivo acetylated protein N-termini and protein C-termini in the crude membrane fraction of human embryonic carcinoma (NT2/D1) cells.

One general drawback for the determination of alternatively terminated protein isoforms by proteomic approaches is the use of routine database strategies, that is, the computational comparison of the peptide (tryptic fragment) masses that were

* Corresponding author. E-mail: a.j.r.heck@uu.nl.
† These authors contributed equally to this work.

10.1021/pr070375k CCC: $37.00
Published on Web 10/10/2007

detected by the mass spectrometer against the peptide (tryptic fragment) masses derived from a database of known proteins.[23] Since unpredicted protein termini resulting from alternative transcription, mRNA splicing, or protein processing often remain unknown or are incompletely annotated in protein databases, the standard computational comparison will cause unpredicted terminal peptides to be overlooked.[24,25] Here, we overcome this drawback for protein isoforms with shortened and in vivo acetylated N-termini and for shortened unmodified C-termini. We have used SCX fractionation for the enrichment of terminal peptides, followed by an iterative search against specifically designed databases. This novel database search strategy proved of great value for the simultaneous identification of annotated and unpredicted in vivo acetylated protein N-termini and unmodified protein C-termini.

Methods

Preparation of the Membrane-Enriched Fraction of Nt2/d1 Cells. Five $\times$ 10^6 human teratocarcinoma Nt2/d1 cells were washed twice with PBS and harvested by scraping. Harvested cells were pelleted at 1100*g* for 5 min at room temperature, resuspended in lysis buffer (50 mM Tris pH 7.8, 250 mM sucrose, 2 mM EDTA, protease inhibitor cocktail (Roche-Diagnostics, Netherlands) and incubated on ice for 10 min. Cells were lysed by 30 passes through a 30-½ gauge needle at 4 °C. Cell debris, unbroken nuclei, ER membranes, and mitochondrial membranes were partially removed by centrifugation at 1000*g* for 10 min at 4 °C. The postnuclear supernatant was layered onto a 60% sucrose cushion and and centrifuged (Optima, Beckman-Coulter, Netherlands) at 160 000*g* in a MLS50 rotor for 1 h at 4 °C. The membrane fraction on top of the sucrose cushion was collected, diluted 1:2 with cold 50 mM Tris pH 7.8, and pelleted (Optima, Beckman-Coulter, Netherlands) at 100 000*g* in a TLA55 rotor for 1 h at 4 °C. The supernatant was discarded, and the membrane pellet was washed with 100 mM Na_2CO_3 pH 11.5 for 1 h at 4 °C, rinsed twice with cold H_2O, and pelleted at 20 000*g* for 30 min at 4 °C.

Proteolytic Cleavage. The membrane-enriched pellet was incubated for 10 min at 95 °C in denaturation solution (20 mM NH_4HCO_3 pH 8.0, 0.2% SDS, 135 mM βME). After the sample cooled down, 1 U PNGase F (Sigma-Aldrich, Netherlands) per million cells was added, and deglycosylation was performed for 2 h at 37 °C. The deglycosylated protein pellet was washed with ice-cold H_2O, pelleted at 20 000*g* for 30 min at 4 °C, and dissolved in 8 M urea in 50 mM NH_4HCO_3 pH 8.0. Reduction and subsequent alkylation were performed with 45 mM DTT for 30 min at 56 °C and 100 mM iodoacetamide for 30 min at room temperature in the dark, respectively. A total of 0.5 μg of endoproteinase Lys-C (Roche-Diagnostics, Netherlands) was added, and digestion was performed overnight at 37 °C. After dilution to 2 M urea with 50 mM NH_4HCO_3 0.5 μg of trypsin (Roche-Diagnostics, Netherlands) was added, and digestion was performed for 8 h. The remaining membranes were sedimented at 20 000*g* for 30 min at 4 °C, and the supernatant was stored at −80 °C. The pellet was redissolved in 80% acetonitrile and redigested with 0.5 μg of trypsin overnight at 37 °C. The second supernatant was pooled with the first and desalted and concentrated using Aqua-C18 material (Phenomenex, Torrance, CA) packed into a ZipTip. The eluate was dried in a vacuum centrifuge and reconstituted in 20% acetonitrile, 0.05% formic acid.

SCX Fractionation. SCX fractionation was performed on a system consisting of two Zorbax BioSCX-Series II columns (i.d., 0.8 mm; l, 50 mm; particle size, 3.5 μm), a Famos autosampler (LC packings, Amsterdam, The Netherlands), a Shimadzu LC-9A binary pump, and a SPD-6A UV-detector (Shimadzu, Tokyo, Japan). In the first 10 min after injection, unbound material was washed from the column with 100% solvent A (0.05% formic acid in 8/2 (v/v) water/acetonitrile, pH 3.0). The subsequent linear gradient increased with 1.3%/min solvent B (500 mM NaCl in 0.05% formic acid in 25% acetonitrile, pH 3.0) with a flow rate of 50 μL/min. Fractions of 50 μL volume were manually collected, dried in a vacuum centrifuge, and reconstituted in 0.1% acetic acid.

NanoLC-MS/MS. The SCX fractions were subsequently analyzed by nanoLC-LTQ-Orbitrap-MS (Thermo, San Jose, CA). An Agilent 1100 series LC system was equipped with a 20 mm Aqua C18 (Phenomenex, Torrance, CA) trapping column (packed in-house, i.d., 50 μm; resin, 5 μm) and a 254 mm ReproSil-Pur C18-AQ (Dr. Maisch GmbH, Ammerbuch, Germany) analytical column (packed in-house, i.d., 50 μm; resin, 3 μm). Trapping was performed at 5 μL/min for 10 min, and elution was achieved with a gradient of 0–45% B in 45 min, 45–100% B in 1 min, 100% B for 4 min The flow rate was passively split from 0.4 mL/min to 100 nL/min. Nanospray was achieved using a distally coated fused silica emitter (New Objective, Cambridge, MA) (o.d., 360 μm; i.d., 20 μm, tip i.d. 10 μm) biased to 1.8 kV. The mass spectrometer was operated in the data dependent mode to automatically switch between MS and MS/MS. Survey full scan MS spectra were acquired from m/z 350 to m/z 1500 in the FT-Orbitrap with a resolution of $R = 60\,000$ at m/z 400 after accumulation to a target value of 500 000 in the linear ion trap with lock-mass. The two most intense ions were fragmented in the linear ion trap using collisionally induced dissociation at a target value of 10 000.

Protein Identification. Spectra were processed with Bioworks 3.2 (Thermo, Bremen, Germany), and the subsequent data analysis was carried out using the Mascot (version 2.1.0) software platform (Matrix Science, London, UK).

For the initial screen, the IPI human database was searched with trypsin allowing two missed cleavages, carbamidomethyl (C) as fixed modification and oxidation (M) and N-acetylation (protein N terminus) as variable modifications. The peptide tolerance was set to 5 ppm with 1^+, 2^+ and 3^+ peptide charges and the MS/MS tolerance to 0.9 Da. The significance thresholds for the identifications was set to $p < 0.05$. For the comprehensive analysis using the semi-tryptic search strategy, the IPI human database was searched with semi-trypsin allowing no missed cleavages, carbamidomethyl (C) as fixed modification and oxidation (M), N-acetylation (protein N terminus), N-propionylation (protein N terminus) as variable modifications. The peptide tolerance was set to 2.5 ppm with 1^+, 2^+, and 3^+ peptide charges and the MS/MS tolerance to 0.9 Da. The stringent tolerance of 2.5 ppm excludes the possible interference of ^{13}C peaks of peptides that potentially interfere with N-acetylated or propionylated peptides, such as carbamylated or carbamidomethylated peptides that have a mass-shift of 43 and 57 Da, respectively. Moreover, the precursor MS spectra would clearly reveal such possible interferences, which in the presented data were not observed. The significance thresholds for the identifications was set to $p < 0.01$. For the comprehensive analysis using the iterative search strategy, the four section databases and the augmented IPI human database that were created as described below in "iterative search strategy" were

searched with trypsin allowing no missed cleavages, carbamidomethyl (C) as fixed modification and oxidation (M), N-acetylation (protein N terminus), N-propionylation (protein N terminus) as variable modifications. In the case of the C-terminal database, the variable modification N-acetylation (protein N-Terminus) was omitted. The peptide tolerance was set to 2.5 ppm with 1^+, 2^+, and 3^+ peptide charges and the MS/MS tolerance to 0.9 Da. The significance thresholds for the identifications were set to $p < 0.01$.

False-positive rates were determined by searching databases concatenated of the original target database and the reversed target database. For the semi-tryptic search, the concatenated database was composed of the IPI human and the reversed IPI human databases. For the iterative search, the concatenated database was composed of the augmented IPI human and the reversed augmented IPI human databases.

Iterative Search Strategy. In the first step the data are searched against a databases consisting solely of unpredicted terminal peptides derived from the IPI human database, thus using a smaller search space than the semi-tryptic strategy. Identified peptides are then added to the original IPI human database to create an augmented database. In the second step, the data are searched against the augmented database to confirm the identity of the identified terminal peptides using strict criteria. Details are provided in the text.

All the resulting MS/MS spectra and information on protein and peptide scores in the various searches mentioned in this study are available at https://bioinformatics.chem.uu.nl/supplementary/dormeyer_JPR.

Results

Selective Enrichment of Protein Termini. Proteins from a membrane-enriched fraction of human embryonic carcinoma NT2/D1 cells were digested using the endoproteinases Lys-C and trypsin, and the resulting peptides were separated by SCX fractionation. An initial liquid chromatography tandem mass spectrometry (LC-MS/MS) screen of all collected SCX fractions revealed that the majority of N-acetylated and C-terminal peptides eluted after 6–7 min from the SCX column (Figure 1a), while the later fractions contained the bulk of the tryptic digest. Solely in the screen, 116 N-acetylated and 26 C-terminal peptides were identified in these two fractions by Mascot database searches of the IPI human database. The terminal peptides represented 92% of the total number (154) of unique peptides identified in these two fractions. There is a second "elution" of N-acetylated peptides during the SCX gradient (figure 1a), which represent those N-terminal acetylated peptides that contain a single mis-cleavage. Unfortunately, this second distribution of acetylated peptides has the same net charge as regular tryptic peptides but represents a minute fraction of the identified unique peptides in the bulk of the digest (Supporting Information Table 1).

To increase the number of identifications of in vivo acetylated protein N-termini and unmodified protein C-termini, we performed a more comprehensive second analysis of the SCX fractions 6 and 7 using more material and longer LC gradients (fraction 6* and 7*). Although such an analysis did lead to an improvement of identifications, more than 3-fold more identifications, we noted that the majority of the acquired tandem mass spectra remained unidentified. First, a semi-tryptic search strategy of the IPI human database was applied using a peptide tolerance of 2.5 ppm and a significance threshold of $p < 0.01$. Unfortunately, N-acetylation must be chosen as a variable peptide modification and thus applies not only to the real N-terminal peptides but also needlessly to the inner and C-terminal peptides. In the semi-tryptic search, the Mascot score corresponding to a p value of 0.01 was 40 identifying 301 N-acetylated proteins. Manual validation of continuous *b*- and *y*-ions series and rejection of low-quality spectra reduced this number to only 222. The false-positive rate of the semi-tryptic search was assessed by performing a search in a concatenated target and reversed IPI human database [26–28] and amounted to 3.8% for the N-acetylated peptides (Table 1 and Supporting Information Table 2).

Design of an Iterative Database Search Strategy. To identify more unpredicted protein isoforms with higher confidence, we set up a two-step iterative database search strategy as an alternative for semi-tryptic searches (Figure 2). The main premise of this strategy is to create a semi-tryptic database containing components we know to be present in the mass spectrometric data and eliminate peptides that could not possibly be present such as N-acetylated internal and C-terminal peptides.

Each protein in the IPI human database was in silico digested using the rules for trypsin digestion. The resulting tryptic fragments were copied several times, with each time one residue from the N-terminus removed resulting in a peptide ladder down to a minimum length of six residues, the minimum length that will be observable by the mass spectrometer. To remove redundancy and distinguish the ladder peptides from the original IPI human entries, all original protein N-termini were removed. The newly generated FASTA database contained all possible shortened N-termini of all proteins in the IPI human database that could be detected by the mass spectrometer. The size of this database was approximately 6 GB compared to the initial human IPI database of 40 MB. To reduce size, false-positive rates and increase search speeds the database was split into three sections. Each peptide was given an identifier based on tryptic fragment and number of amino acids removed from the original peptide (IPIidentifier_tf_#tryptic-fragment_#removed residues_name of section database). The first section database contained all possible N-termini for the first 150 amino residues of the proteins. The second database contained the N-termini for amino acids at postions 150–300, and the third database contained the remaining set of termini from residue 300 to the end. An additional database was created in a similar fashion for potential new C-termini, but this was restricted to the final 150 amino acids of each protein. The four section databases were used for the identification of unpredicted N-acetylated and C-terminal peptides from the LC-MS/MS analyses as described above in "protein identification". The 239 N-acetylated and 227 C-terminal peptides identified in the search against the new databases were added to the human IPI database to create an augmented database. This augmented IPI human database was used for a second search using the above mentioned stringent parameters, and the results were evaluated manually.

Identification of Protein Termini Regardless of Their Annotation in Protein Databases. The iterative search of the fractions 6* and 7* using a peptide tolerance of 2.5 ppm and a significance threshold of $p < 0.01$ resulted in the identification of 263 IPI annotated and 87 unpredicted N-acetylated and 168 IPI annotated and 193 unpredicted C-terminal peptides (Figure 1b). Note, a number of identifications made in the first search against the modified semi-tryptic database were lost. The Mascot score corresponding to a p value of 0.01 was 25, and

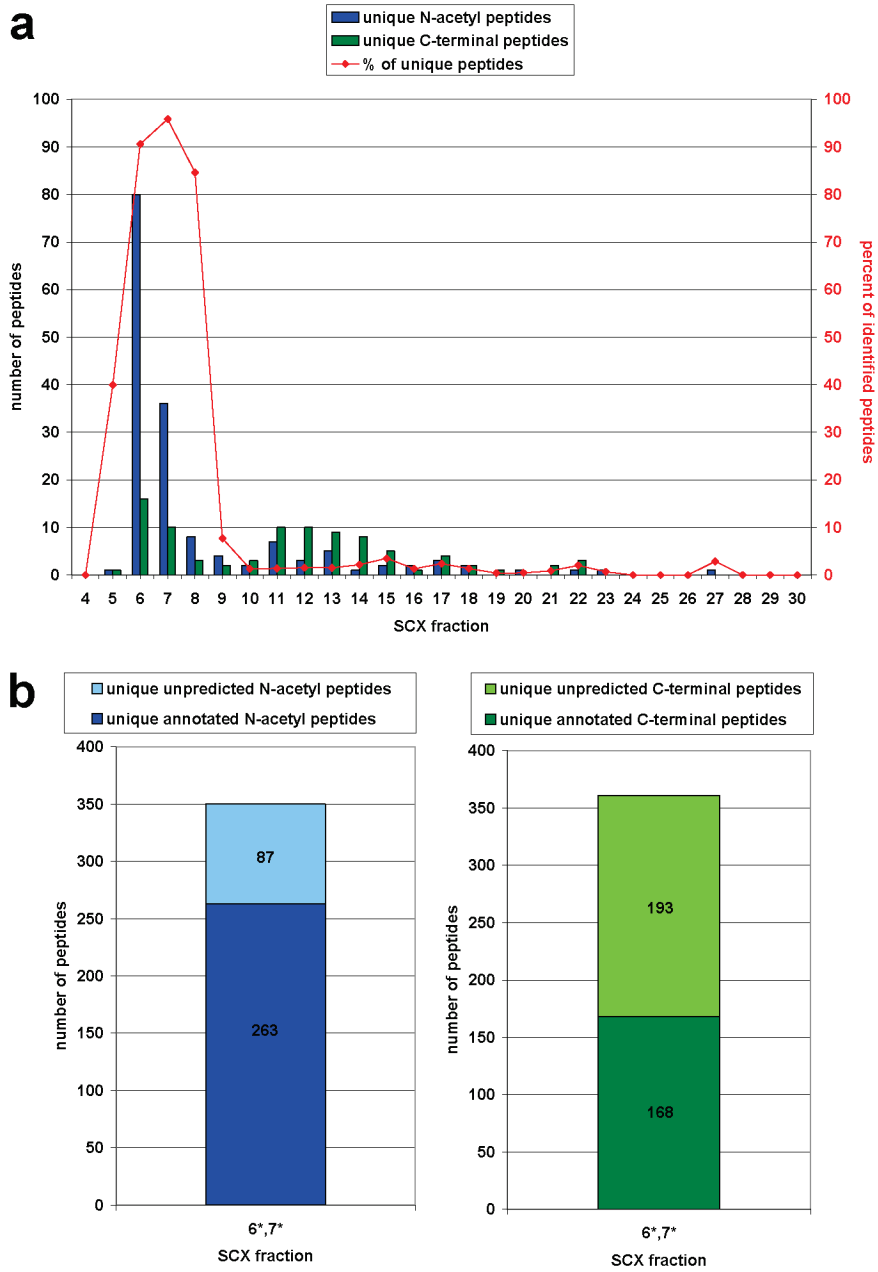

Figure 1. Targeted analysis of in vivo acetylated N-termini and unmodified C-termini. (a) Membrane proteins of human embryonic carcinoma NT2/D1 cells were digested using the endoproteinases Lys-C and trypsin. The proteolytic peptides were fractionated by SCX at low pH, analyzed by LC-MS/MS, and identified by Mascot searches of the IPI human database. The bulk of N-acetylated and C-terminal peptides eluted after 6–7 min (= fractions 6 and 7) from the SCX column. Eighty and 36 predicted N-acetylated (blue) and 16 and 10 predicted C-terminal (green) peptides were identified, respectively (left *y*-axis). Together, they represented 91 and 96% (red, right *y*-axis) of the total number of unique peptides in each fraction. Eighteen predicted N-acetylated and 42 predicted C-terminal peptides eluted not until 11–15 min from the SCX column. These peptides carried more than one positive charge due to tryptic miscleavages or His residues. Their appearance in the later fractions underscores that the selective recovery of N-acetylated and C-terminal peptides by SCX fractionation depends on the integrity of the tryptic digest. (b) For a comprehensive analysis of IPI annotated and unpredicted N-acetylated and C-terminal peptides 5 times more material of the SCX fractions 6 and 7 was reanalyzed by LC-MS/MS using a longer RP-HPLC gradient (120 min) and identified by Mascot searches of the augmented IPI human database (Figure 2). In total, 350 N-acetylated (blue) and 361 C-terminal (green) peptides were identified representing 88% of the total number (804) of identified unique peptides in these fractions (6*,7*).

peptides with a score below 30 were evaluated manually. Identifications with significant *b*- and *y*-ion series were accepted, while low-quality spectra were rejected. The exact numbers of peptides identified in the subsequent steps of the iterative search are given in Table 1. Performing a search in a database concatenated of the augmented and the reversed augmented IPI human database, we found no N-acetylated reversed peptide sequence (an apparent false discovery rate of 0%) (Table 1). In comparison, 10 N-acetylated reversed peptide sequences were found in the semi-tryptic search of the concatenated target and reversed IPI human database (3.8% false discovery rate). Hence, our iterative database search

Table 1. Summary of the Number of Identifications and the (Mascot Defined) Confidence Levels for the Semitryptic and the Iterative Database Search Strategies[a]

	semi-tryptic search in the human IPI database		
details in Supporting Information Table 2A	all peptides	N-acetyl peptides	C-terminal peptides
total number of unique peptides	887	222 (301 before evaluation)	93
Mascot score according to $p < 0.01$	>40		
	iterative search in the augmented human IPI database		
details in Supporting Information Table 2B	all peptides	N-acetyl peptides	C-terminal peptides
total number of unique peptides	1084	350 (372 before evaluation)	361
number of unique annotated peptides	804	263	168
number of unique unpredicted peptides	280	87	193
Mascot score according to $p < 0.01$	>25		
	semi-tryptic search in a database concatenated of the human IPI and the reversed human IPI database		
details in Supporting Information Table 2C	all peptides	N-acetyl peptides	C-terminal peptides
total number of identified unique peptides	762	263	81
false-identified reversed sequences	38	10	0
false-positive rate (%)	4.99	3.80	0.00
Mascot score according to $p < 0.01$	>43		
	iterative search in a database concatenated of augmented human IPI and reversed augmented human IPI database		
details in Supporting Information Table 2D	all peptides	N-acetyl peptides	C-terminal peptides
total number of unique peptides	729	324	351
false-identified reversed sequences	19	0	1
false-positive rate (%)	2.61	0.00	0.28
Mascot score according to $p < 0.01$	>28		

[a] Given are the numbers of peptide identifications from the semi-tryptic search in the human IPI database and the iterative search in the augmented human IPI database. Given for the determination of the false-positive rates are the numbers of peptide identifications from the semi-tryptic search of the database concatenated of the IPI human and the reversed IPI human database and the iterative search of the database concatenated of the augmented IPI human and the reversed augmented IPI human. The full underlying Mascot search results are listed in Supporting Information Table 2, sheets (A), (B), (C), and (D). Please note that the number of N-acetylated peptides identified by the iterative search drops from 350 in the augmented IPI human to 324 in the database concatenated from the augmented IPI human and the reversed augmented IPI human. This was due to the increased size of the database and the rise of the score according to $p < 0.01$ from 25 to 28. In the latter search, the missed peptides were identified with scores below 28.

strategy performs better with regard to the number and the confidence of the peptide identifications. The unexpected high number of identified unpredicted terminal peptides substantiates the problem of incomplete annotation of protein termini in databases and the need for search strategy that takes into account truncated protein N termini.

Identified Unpredicted N-Acetylated Peptides. To test the reliance of the identifications in the augmented database, we compared the sequences of the identified IPI annotated and the unpredicted N-acetylated peptides. The full sequences of all identified N-acetylated peptides are given in Supporting Information Table 3A,B as well as part of a comprehensive database that includes spectra which is downloadable (see Methods). In the eukaryotic cell, Met cleavage by methionine aminopeptidases (MAPs) and/or N-terminal acetylation by N-terminal acetyltransferases (NATs) occur cotranslationally when 20–50 amino acid residues extrude from the ribosome.[29,30] Only general sequence requirements for N-acetylation by three of the five known eukaryotic NAT enzymes have been suggested[31,32] with most notably Met, Ala, Ser being abundant as the N-terminal residue and Asp and Glu being abundant as penultimate residues. We determined the occurrence of each amino acid at the N-terminus and the three penultimate residues in the 350 identified N-acetylated peptides. The sequences of the 263 IPI annotated N-acetylated peptides strongly confirmed the suggested criteria for N-acetylation by NATs (Figure 3a). Met, Ala, Ser and to a lesser extent Thr, Cys, Val, Gly showed an abundant occurrence as N-acetylated residues. Asp and Glu were frequent residues at the penultimate position. For the residues in positions 3 and 4, no contributive influence on the N-acetylation could be deduced. We next evaluated the sequences of the 23 unpredicted N-acetylated termini that originated from the first 40 residues of an annotated protein isoform because signal cleavage generally takes place within residue 15–40 of a protein. Interestingly, the sequences also matched the N-terminal acetylation criteria. Ala, Phe, Met were the most frequent N-acetylated residues, whereas Pro, Asp, Gln, Ala, Glu were prevalent in the penultimate position. The variability of the sequences was more prominent in the 64 unpredicted N-acetylated peptides that originated from cleavage between residue 41 to the C-terminus of an IPI annotated protein isoform (Supporting Information Figure 1). Since N-acetylation by NATs takes place at the ribosome while signal cleavage occurs in various cellular compartments, we hypothesize that the observed unpredicted N-acetylated termini are generated by two different biological processes: (1) peptides matching the NAT acetylation criteria result from cotranslational N-acetylation by NATs after alternative intronic transcription or mRNA splicing, and (2) termini not matching the NAT acetylation criteria result from post-translational processing. Evidence for post-translational N-acetylation exists, but the nature of the responsible enzymes is largely unknown.[31,33,34] Interestingly, 7 of the 23 unpredicted N-acetylated termini that originated from the first 40 residues of an IPI annotated protein isoform and matched the NAT acetylation criteria carry a Met at the position preceding the unpredicted N-acetyl terminus in the IPI annotated sequence (Supporting Information Table 3B, column G), indicating post-

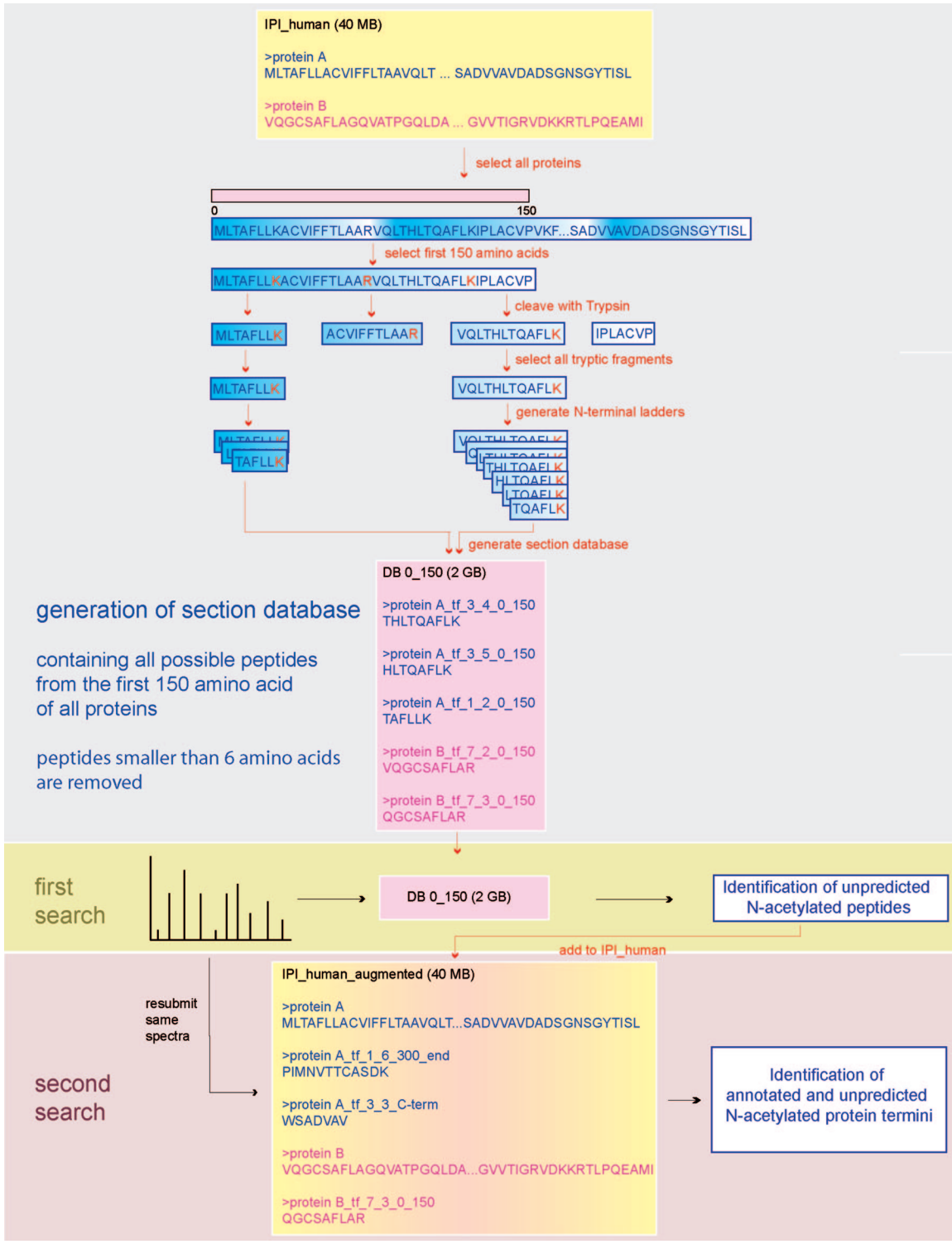

Figure 2. Iterative database search strategy for the targeted analysis of acetylated N-termini and unmodified C-termini of unpredicted protein isoforms. (Upper panel) Initially, a new database was created containing all possible N-termini from the first 150 amino acids of all proteins in the human IPI database. Each protein in the human IPI database was cleaved in silico using trypsin. From the tryptic fragments, single residues were cleaved off in a stepwise manner resulting in N-terminal ladders of peptides with a minumum length of six residues. Each peptide was given an identifier based on the IPI entry, the tryptic fragment, the number of amino acids removed from the original peptide and the split section or terminus. (Middle panel) The new database was used in the primary searches for the identification of N-acetylated and C-terminal peptides from the more comprehensive LC-MS/MS analyses. To ascertain that these unpredicted termini did not match an annotated peptide in the original human IPI database, the identified sequences were added to the human IPI database. (Bottom panel) The augmented database was used in the final search of the spectra from the comprehensive LC-MS/MS analysis for the conclusive identification of acetylated N-termini and unmodified C-termini of annotated and unpredicted protein isoforms (bottom panel). The process was repeated for identifying potential N-termini originating from protein residues 151–300 and 301 to the C-terminus (not shown). Additionally, a similar search was performed for unpredicted C-termini using the final 150 amino acid residues of each IPI protein (not shown).

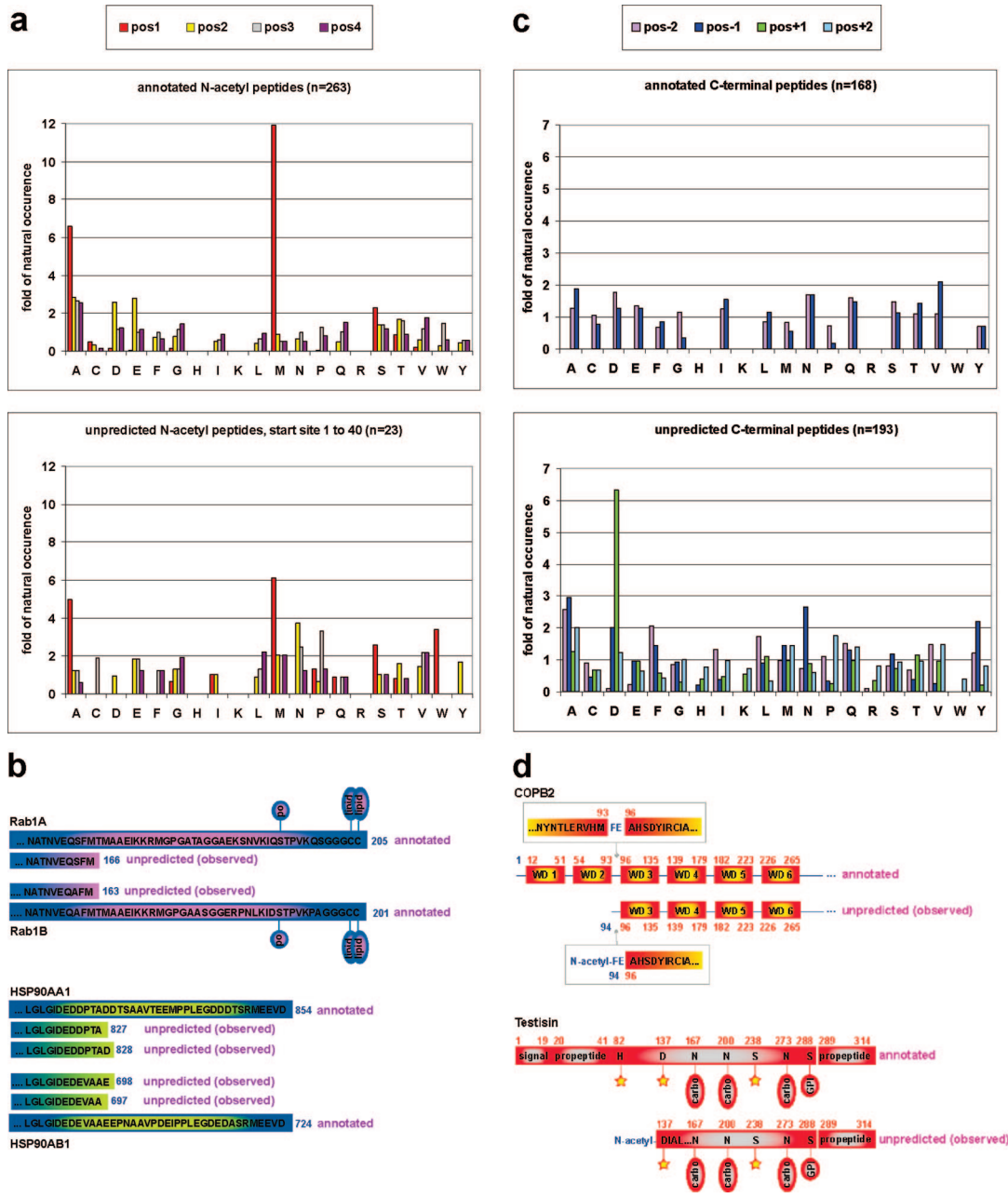

Figure 3. Sequence comparison of annotated and unpredicted N-acetylated and C-terminal protein isoforms. (a) The occurrence of each amino acid in the first four positions of the identified N-acetylated peptides was determined and normalized using the natural amino acid occurrence in all proteins in the human IPI database. (Upper diagram) The sequences of the identified annotated N-acetylated peptides confirm the criteria for in vivo N-acetylation with most notably Met, Ala, Ser being abundant and His being rare as N-acetylated residue and Asp and Glu being abundant as penultimate residues. (Bottom diagram) The sequences of the identified unpredicted N-acetylated peptides that originated from the first 40 residues of an annotated protein isoform also matched the N-terminal acetylation criteria. In particular, Met and Ala were observed frequently and His rarely as N-acetylated residue indicating in vivo cleavage of localization signals and subsequent N-acetylation of the identified unpredicted isoforms. (b) Two examples of unpredicted N-acetylated isoforms that might serve a biological role. (COPB2) The observed unpredicted N-acetylated isoform of the coatomer subunit beta 2 lacks the first two of the six WD repeat of the annotated isoform domains which might affect the ability of the protein to form a propeller structure. (Testisin) The observed N-acetylated isoform lacks the N-terminal signal and propeptide sequences and starts exactly at the second of the three residues that form the charge relay system which might suppress the Testisin protease activity. ☆, charge relay system; carbo, glycosylation site; GPI, glycosylphosphatidylinositol-anchor site. (c) The occurrence of each amino acid in the four positions surrounding the potential C-terminus of the identified C-terminal peptides was determined and normalized using the natural amino acid occurrence in all proteins in the human IPI database. The two final positions of the identified annotated (upper diagram) and unpredicted (bottom diagram) C-terminal peptides show high variability, whereas Asp is frequently observed in the first position of the missing C-terminal sequence. (d) Two examples of unpredicted C-terminal isoforms that indicate a role of C-terminal processing in vivo. (Rab1) The two observed unpredicted C-terminal isoforms result from truncation at a conserved site and lack the two lipid anchors that are required for the localization of the protein to certain cellular compartments. Po, phosphorylation site; lipid, geranylgeranylation site. (HSP90A) The 2x2 observed unpredicted C-terminal isoforms originate from a combination of truncation and/or processing at a conserved site of the protein family indicating specific in vivo truncation.

translational cleavage and N-acetylation by enzymes similar to the cotranslational MAPs and NATs.

The significance of in vivo protein N-acetylation is not fully understood. Only a subset of N-acetylated proteins requires acetylation for their individual function, localization, or stability, whereas others are acetylated only because their terminus happens to match the N-acetylation criteria.[29,31] No N-terminal deacetylation enzymes are known. To assess the potential biological significance of the 87 unpredicted N-acetylated isoforms, we examined the annotation of the proteins in databases and sorted them into five categories (Supporting Information Table 3B, columns H–L). For 10 unpredicted isoforms, the observed N-terminus was previously described but not correctly annotated in the human IPI database (category 1). Four of these 10 known isoforms were even reported to be N-acetylated. For example, a short isoform of RACK1, a protein that regulates major nervous system pathways involved in brain aging and various neurodegenerative diseases,[35] was described but not IPI annotated. The short RACK1 isoform is known to undergo Met cleavage and N-terminal acetylation at the peptide (M-)N-acetyl-TEQMTLR (Bienvenut et al., 2006, unpublished observations), which we indeed found as unpredicted N-acetylated terminus. These cases emphasize that our database search strategy leads to the identification of genuine in vivo isoforms. Our data might therefore be used to improve the human IPI database. For four unpredicted isoforms, it was suggested that the first amino acid in the IPI annotated sequence is not the N-terminal residue of the protein in vivo (category 2). For example, the N-terminus of the jerky homologue protein, a putative DNA-binding protein involved in the development of epilepsy, is unknown.[36] Our study revealed that the protein starts with N-acetyl-SYEVLK at residue 27 of the annotated sequence. For four of the observed unpredicted N-acetylated proteins isoforms were suggested to exist but not fully characterized (category 3). The suggested isoform for endothelin converting enzyme 1[37–39] differed only in three residues from the start site N-acetyl-QGLGLQR, which we have detected. Again, this might reveal wrong annotation of the protein termini in the IPI human database. Ten unpredicted isoforms represented hypothetical proteins, cDNA fragments, or chromosome open reading frames (category 4). Little is known about their appearance in vivo, and accordingly the annotation of their N-terminus in the IPI human database is prone to be incorrect. For example, the first 18 residues of the annotated sequence of the hypothetical protein LOC346689[40] might not be synthesized in vivo. We found the protein to start at residue 19 with the sequence N-acetyl-MEPVGSLVPTLEQPQV-PAK, which matches the NAT acetylation criteria. Our strategy may help increasing the confidence of such insecure annotations. Fifty-nine unpredicted N-acetylated isoforms were to our knowledge not reported previously (category 5). Seven of these 59 unknown isoforms start at vital regions of the annotated protein such as post-translational modification sites, SNP positions, or boundaries of functional domains. These cases might indicate that an isoform exists in vivo, which dispenses with the function of this vital region. For example, the coatomer subunit beta is essential for vesicular trafficking from the Golgi and contains six tryptophan-aspartic acid (WD) repeats that can form a circularized beta propeller structure.[41,42] The isoform we have observed starts with N-acetyl-FEAHSDYIR at residue 94, which is the exact end of the second WD domain (Figure 3b). The lack of two of the six WD repeats in the identified N-acetylated isoform will likely affect the ability of the protein to form the propeller. Another example is the serine protease testisin, a tumor suppressor of testicular cancer that is highly abundant in ovarian cancer.[43,44] The active site of testisin consists of a charge relay system composed of the residues His82, Asp137, and Ser238.[45] We identified an isoform starting with N-acetyl-DIALVK at residue 137, which is the second of these critical residues (Figure 3b). Truncation and N-acetylation at this essential residue are very likely to suppress the proteolytic function of testisin needed for malignant transformation. The identified truncated protein might therefore represent a nonmalignant isoform.

Identification of a Novel Post-Translational Modification: N-Propionylation. So far we have shown that N-acetylated peptides are prominent in early SCX fractions. However, the strategy is tailored for enrichment of any N-terminally blocked peptide as long as the net charge of the peptide remains +1. This is becoming a highly relevant feature as highlighted by the recent finding that histones may be in vivo modified by lysine propionylation and butyrylation, thought to be carried out by acetyltransferases.[46] Therefore, we applied our iterative strategy on the the search for N-terminal protein propionylation. The analyses lead us to confidently identify five N-propionylated peptides in our data set with a significance threshold of $p < 0.01$ and a peptide score above 35 where four were present at the IPI annotated N-terminus of a protein and one not at the expected N-terminus (Supporting Information Table 4, Figure 4). Their sequences match the NAT modification criteria, and three of the identified proteins are nucleic acid binding proteins. This result suggests that N-terminal propionylation may play a role in vivo and underscores the utility of our iterative search strategy for the analysis of protein termini

Identified Unpredicted C-Terminal Peptides. Similar to the unpredicted acetylated N-termini, the high number of identified unpredicted C-termini might arise from alternative transcription, splicing, or post-translational cleavage, but no cotranslational processing of protein C-termini is known.[47] In addition, unspecific cleavage by proteases during in vivo degradation might occur. Unspecific cleavage during in vitro sample handling was suppressed by the use of protease inhibitor cocktails but cannot be fully excluded. However, cleavage of C-terminal signals is biologically relevant as illustrated by peroxisomal proteins which carry their localization sequence at the C-terminus.[48] To assess the origin of the unpredicted protein C-termini, we determined the occurrence of each amino acid at the two final positions of the identified 361 identified protein C-termini. The full sequences of all identified C-terminal peptides are given in Supporting Information Table 3C,D. The variability in these positions was high for both the annotated and unpredicted C-terminal peptides, and no prevalence was observed (Figure 3c). Using the annotated protein sequences, we next assessed the first two residues in the sequence that was potentially cleaved off of the protein. Interestingly, Asp showed a high occurrence in the first position after the observed cleavage site indicating the presence of an enzyme with Asp-N activity. However, Asp-N endoproteinase has not been described in mammals. In human, proteases such as memaprin-2, cathepsin-L, procollagen-C peptidase, and various caspases cleave N-terminally of Asp but only in a specific set of proteins.[49] Detailed analyses are required to determine the origin and potential biological impact of the observed C-terminal isoforms. It can not be fully excluded that the identified unpredicted C-termini result from unspecific

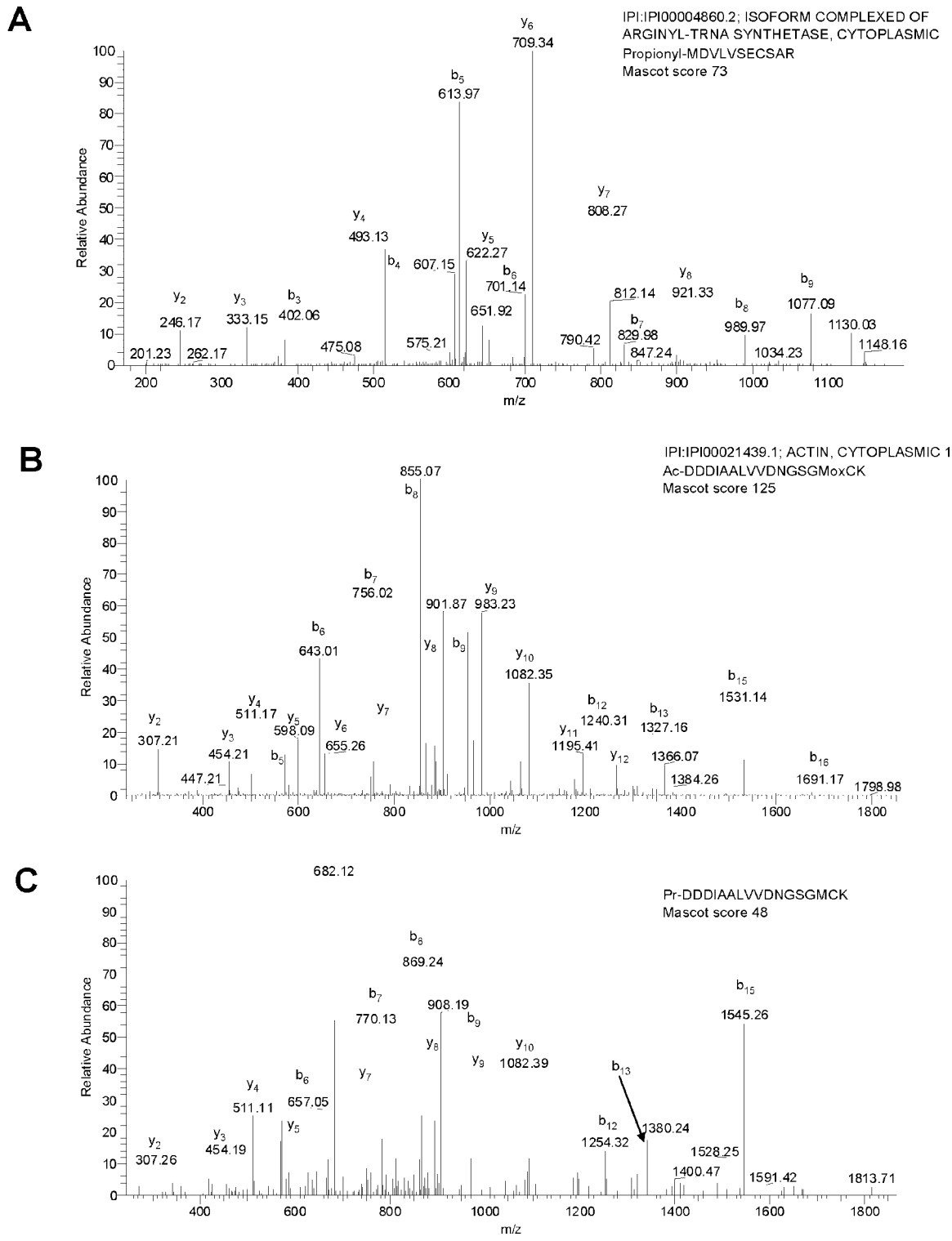

Figure 4. (A) Annotated spectrum of the identified propionylated N-terminus of arginyl-tRNA synthetase that was identified with a peptide score of 73 by our iterative database search strategy. Annotated spectra of the identified (B) acetylated and (C) propionylated N-terminus of actin. The similarity of the spectra underscores that the identification of the N-terminus and its two modifications by our strategy is valid.

proteolytic cleavage during sample handling, chymotryptic side-activity of trypsin during the digest, or collision-induced dissociation fragmentation in the source of the mass spectrometer. Still, the following examples of unpredicted C-termini that lie in vital regions of proteins or are conserved among members of protein families provide evidence that C-terminal processing might serve a role in vivo.

The B-cell receptor-associated protein 31 which potentially regulates intracellular trafficking in neutrophils[50] was observed to end with DGPM(-KKEE) at position 240. The neighboring signal KKEE in positions 242–245 was reported to prevent the secretion of the protein from the ER.[51,52] Therefore, the detected isoform might represent the secreted protein after KKEE signal cleavage. Remarkably, it was reported previously

that another C-terminal cleavage fragment of this protein can direct pro-apoptotic signals between the ER and mitochondria.[53] The selenoprotein K was observed in our analysis to end with MAGG(-^{seleno}CGR) at position 91. The missing C-terminus is special since it carries the rare amino acid selenocysteine at the beginning. The role of selenocysteine in this protein remains unclear, and no selenocysteine-specific proteases are known. Strikingly, 3′ untranslated regions of selenoprotein genes carry a selenocysteine insertion sequence that is necessary for the recognition of the codon UGA as a selenocysteine codon rather than as a stop signal.[54] The C-terminally truncated isoform we detected could therefore be coded by a gene that lacks the selenocysteine insertion sequence so that the UGA codon was read as a stop signal. The two members Rab1A and Rab1B of the RAS oncogene family[55] were detected with the same unpredicted C-terminus (Figure 3d). Their sequences QAFM(-TMAA) and QSFM(-TMAA) at the observed truncation sites are highly conserved but differ at position −3. Interestingly, the truncation leads to the loss of the two C-terminal geranylgeranylated Cys residues that are essential but not sufficient for the regulated recruitment of Rab1 proteins to certain membranes.[56] The detected truncated variants also lack the conserved GTPase region and might therefore represent unfunctional isoforms or degradation species. Our strategy could therefore help to decode cellular localization signals or conserved degradation pathways. A double-truncation site was observed for the molecular chaperones and potential anticancer targets HSP90A A1 and B1.[57] For both protein, we detected two conserved isoforms that differed only by one acidic residue (Figure 3d). The isoforms might therefore originate from a combination of specific in vivo truncation and Asp cleavage of HSP90A family members at the conserved site.

In the early SCX fractions, each protein isoform should be represented by both terminal peptides. Indeed, for some proteins we found an annotated and an unpredicted acetylated N- or C-terminus, respectively (Supporting Information Table 5). For the voltage-dependent anion channel 1 (VDAC1), the primary transporter of nucleotides, ions, and metabolites across the outer mitochondrial membrane,[58] we even observed an annotated (N-acetyl-^{1}AVPP) and an unpredicted (N-acetyl-^{53}VTGS) acetylated N-terminus and a annotated (EFQA^{283}) and an unpredicted (DACF^{233}-(SAKV)) C-terminus. From our experiment, there is no information about which N-terminus belongs to which C-terminus. In recent VDAC1 topology models,[59,60] the unpredicted acetylated N-terminus is located in the mitochondrial intermembrane space at the turn between the second and the third membrane-spanning domain, whereas the unpredicted C-terminus lies in the middle of the tenth membrane-spanning strand. No such VDAC1 isoforms were reported previously, but remarkably both truncations remove one potential ATP binding site from the protein while leaving the vital binding sites for kinases and the adenine nucleotide translocase complex intact.

Discussion and Conclusions

Shotgun proteomics provides the most comprehensive identification of proteins from cellular lysates but generally fails to characterize protein isoforms. This is most often due to the scarce detection of terminal peptides during the mass spectrometric analysis of complex peptide mixtures and the incomplete annotation of protein termini in protein databases. Typically, researchers try to overcome the difficulties with incomplete databases by using non- or semi-tryptic search parameters. However, such strategies produce a reduced level of identification with poorer confidence in particular when N-acetylation is considered. In a direct comparison to our iterative two-step search method, semi-tryptic search conditions indeed revealed to be far inferior with regard to the number of identifications and the false-positive rate (Supporting Information Table 2). Our targeted analysis of SCX-fractionated terminal peptides in the proteolytic digest of a crude membrane fraction revealed the existence of a wide variety of annotated and unpredicted in vivo acetylated N-termini and unmodified C-termini. While the identified unpredicted C-terminal peptides could in theory result from unspecific protein degradation during sample handling, it is unlikely that the unpredicted N-acetylated and N-propionylated peptides are artifacts. It is against the odds that unspecific N-terminal degradation followed by enzymatic transfer of an acetyl group from acetyl-coenzyme A to the artifactual N-terminus by a specific NAT enzyme takes place in the sample or the source of the mass spectrometer.

The identified unpredicted terminal peptides, that is, protein termini not annotated in the human IPI database, are missed by routine strategies. Although approaches for the comprehensive characterization of terminal peptides from cellular lysates have been described before, our method is unique in that it leads to the simultaneous identification of annotated and unpredicted N-acetylated and C-terminal peptides without laborious chemical pretreatment or separation of the sample. We aimed at the identification of protein termini of membrane proteins of teratocarcinoma cells using the human IPI database, but the method is generally applicable. SCX-enrichment can be used for complex peptide samples from any source, and our iterative search strategy can be easily adapted and automated for any protein database.

N-Terminal protein acetylation and its biological impact are the subject of an increasing number of studies, while protein C-termini have received relatively little attention so far. However, the determination of the C-terminus of a protein is important not only for the identification of protein isoforms but also in the research of enzymatic protein processing. A growing number of proteases are known, but their biological role in vivo is largely unknown.[49] The identification of their substrates and the characterization of their cleavage criteria are an ongoing task in the new area of degradomics.[61–63] Tailored methods for elucidating C-termini are necessary to completely unravel the “degradome”, and our method can provide a useful base for the analysis of protein processing.

The characterization of protein termini is not only essential for the correct annotation of proteins in protein databases but is also of great value for the determination of protein-coding gene boundaries and alternative transcription sites in genome and transcriptome sequencing projects. Conclusively, our targeted analysis of alternatively terminated protein isoforms can support the correct annotation of genes and proteins in databases and thereby help to further define the human genome and proteome.

Acknowledgment. We thank Dennis van Hoof and Christine Mummery for NT2/D1 cells and Sylvia Neumann and Joost Holthuis for advice concerning membrane preparations. We also thank Martina O’Flaherty and Martijn Pinkse for technical support with LTQ-Orbitrap and SCX and Joost Holthuis and Peter van der Sluijs for critically reading the manuscript. This work

was supported by The Netherlands Proteomics Centre (www.netherlandsproteomicscentre.nl). The authors declare that they have no competing financial interests.

Supporting Information Available: Table summarizing the number of peptides identified in the initial LC-MS/MS screen of all collected SCX fractions (Table 1A); table summarizing the number of peptides that were identified in the individual steps of the iterative database search during the comprehensive analysis of SCX fraction 6* and 7* (Table 1B); Tables representing the full underlying Mascot search results (Table 2); tables of the identified (A) IPI annotated and (B) unpredicted N-acetylated peptides and the (C) IPI annotated and (D) unpredicted C-terminal peptides (Table 3); table summarizing the identified N-propionylated peptides (Table 4); table categorizing the proteins for which an annotated and/or an unpredicted acetylated N-terminus and/or C-terminus was found (Table 5); graph showing the occurrence of each amino acid at the N-terminus and the three penultimate residues in the 64 unpredicted N-acetylated peptides that originated from cleavage between residue 41 to the C-terminus of an IPI annotated protein isoform. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bertone, P.; Stolc, V.; Royce, T. E.; Rozowsky, J. S.; Urban, A. E.; Zhu, X.; Rinn, J. L.; Tongprasit, W.; Samanta, M.; Weissman, S.; Gerstein, M.; Snyder, M. Global identification of human transcribed sequences with genome tiling arrays. *Science* **2004**, *306* (5705), 2242–2246.

(2) Johnson, J. M.; Edwards, S.; Shoemaker, D.; Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **2005**, *21* (2), 93–102.

(3) Emanuelsson, O.; Nagalakshmi, U.; Zheng, D.; Rozowsky, J. S.; Urban, A. E.; Du, J.; Lian, Z.; Stolc, V.; Weissman, S.; Snyder, M.; Gerstein, M., Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Res.* **2006**.

(4) Xing, Y.; Resch, A.; Lee, C. The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.* **2004**, *14* (3), 426–441.

(5) Xing, Y.; Xu, Q.; Lee, C. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett.* **2003**, *555* (3), 572–578.

(6) Godovac-Zimmermann, J.; Kleiner, O.; Brown, L. R.; Drukier, A. K. Perspectives in spicing up proteomics with splicing. *Proteomics* **2005**, *5* (3), 699–709.

(7) Kuroyanagi, H.; Kobayashi, T.; Mitani, S.; Hagiwara, M. Transgenic alternative-splicing reporters reveal tissue-specific expression profiles and regulation mechanisms in vivo. *Nat. Methods* **2006**, *3* (11), 909–915.

(8) Gevaert, K.; Van Damme, P.; Ghesquiere, B.; Vandekerckhove, J. Protein processing and other modifications analyzed by diagonal peptide chromatography. *Biochim. Biophys. Acta* **2006**, *1764* (12), 1801–1810.

(9) Kleyman, T. R.; Myerburg, M. M.; Hughey, R. P. Regulation of ENaCs by proteases: An increasingly complex story. *Kidney Int.* **2006**, *70* (8), 1391–1392.

(10) Nesvizhskii, A. I.; Roos, F. F.; Grossmann, J.; Vogelzang, M.; Eddes, J. S.; Gruissem, W.; Baginsky, S.; Aebersold, R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* **2006**, *5* (4), 652–670.

(11) Brown, J. L.; Roberts, W. K. Evidence that approximately eighty percent of the soluble proteins from Ehrlich ascites cells are Nalpha-acetylated. *J. Biol. Chem.* **1976**, *251* (4), 1009–1014.

(12) Polevoda, B.; Sherman, F. The diversity of acetylated proteins. *Genome Biol.* **2002**, *3* (5), reviews0006.

(13) Sechi, S.; Chait, B. T. A method to define the carboxyl terminal of proteins. *Anal. Chem.* **2000**, *72* (14), 3374–3378.

(14) Kosaka, T.; Takazawa, T.; Nakamura, T. Identification and C-terminal characterization of proteins from two-dimensional polyacrylamide gels by a combination of isotopic labeling and nano-electrospray Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **2000**, *72* (6), 1179–1185.

(15) Gevaert, K.; Goethals, M.; Martens, L.; Van Damme, J.; Staes, A.; Thomas, G. R.; Vandekerckhove, J. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **2003**, *21* (5), 566–569.

(16) Martens, L.; Van Damme, P.; Van Damme, J.; Staes, A.; Timmerman, E.; Ghesquiere, B.; Thomas, G. R.; Vandekerckhove, J.; Gevaert, K. The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics* **2005**, *5* (12), 3193–3204.

(17) Samyn, B.; Sergeant, K.; Castanheira, P.; Faro, C.; Van Beeumen, J. A new method for C-terminal sequence analysis in the proteomic era. *Nat. Methods* **2005**, *2* (3), 193–200.

(18) McDonald, L.; Robertson, D. H.; Hurst, J. L.; Beynon, R. J. Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat. Methods* **2005**, *2* (12), 955–957.

(19) Gorman, J. J.; Shiell, B. J. Isolation of carboxyl-termini and blocked amino-termini of viral proteins by high-performance cation-exchange chromatography. *J. Chromatogr.* **1993**, *646* (1), 193–205.

(20) Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (33), 12130–12135.

(21) Gruhler, A.; Schulze, W. X.; Matthiesen, R.; Mann, M.; Jensen, O. N. Stable isotope labeling of Arabidopsis thaliana cells and quantitative proteomics by mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (11), 1697–1709.

(22) Trinidad, J. C.; Specht, C. G.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L. Comprehensive identification of phosphorylation sites in postsynaptic density preparations. *Mol. Cell. Proteomics* **2006**, *5* (5), 914–922.

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

(24) Kersey, P.; Hermjakob, H.; Apweiler, R. VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics* **2000**, *16* (11), 1048–1049.

(25) Hoehenwarter, W.; Ackermann, R.; Zimny-Arndt, U.; Kumar, N. M.; Jungblut, P. R. The necessity of functional proteomics: protein species and molecular function elucidation exemplified by in vivo alpha A crystallin N-terminal truncation. *Amino Acids* **2006**, *31* (3), 317–323.

(26) Moore, R. E.; Young, M. K.; Lee, T. D. Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (4), 378–386.

(27) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2003**, *2* (1), 43–50.

(28) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214.

(29) Polevoda, B.; Sherman, F. Nalpha -terminal acetylation of eukaryotic proteins. *J. Biol. Chem.* **2000**, *275* (47), 36479–36482.

(30) Frottin, F.; Martinez, A.; Peynot, P.; Mitra, S.; Holz, R. C.; Giglione, C.; Meinnel, T. The proteomics of N-terminal methionine cleavage. *Mol. Cell. Proteomics* **2006**, *5* (12), 2336–2349.

(31) Polevoda, B.; Sherman, F. N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. *J. Mol. Biol.* **2003**, *325* (4), 595–622.

(32) Kiemer, L.; Bendtsen, J. D.; Blom, N. NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics* **2005**, *21* (7), 1269–1270.

(33) Michel, H.; Hunt, D. F.; Shabanowitz, J.; Bennett, J. Tandem mass spectrometry reveals that three photosystem II proteins of spinach chloroplasts contain N-acetyl-O-phosphothreonine at their NH_2 termini. *J. Biol. Chem.* **1988**, *263* (3), 1123–1130.

(34) Sheff, D. R.; Rubenstein, P. A. Isolation and characterization of the rat liver actin N-acetylaminopeptidase. *J. Biol. Chem.* **1992**, *267* (28), 20217–20224.

(35) Sklan, E. H.; Podoly, E.; Soreq, H. RACK1 has the nerve to act: structure meets function in the nervous system. *Prog. Neurobiol.* **2006**, *78* (2), 117–134.

(36) Totoki, Y.; Toyoda, A.; Takeda, T.; Sakaki, Y.; Tanaka, A.; Yokoyama, S.; Ohara, O.; Nagase, T.; Kikuno, R. F. Homo sapiens protein coding cDNA. Direct submission to Genbank. http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=62087584, 2005.

(37) Valdenaire, O.; Rohrbacher, E.; Langeveld, A.; Schweizer, A.; Meijers, C. Organization and chromosomal localization of the

human ECEL1 (XCE) gene encoding a zinc metallopeptidase involved in the nervous control of respiration. *Biochem. J.* **2000**, *346*, 611–616.
(38) Orzechowski, H. D.; Richter, C. M.; Funke-Kaiser, H.; Kroger, B.; Schmidt, M.; Menzel, S.; Bohnemeier, H.; Paul, M. Evidence of alternative promoters directing isoform-specific expression of human endothelin-converting enzyme-1 mRNA in cultured endothelial cells. *J. Mol. Med.* **1997**, *75* (7), 512–521.
(39) Schmidt, M.; Kroger, B.; Jacob, E.; Seulberger, H.; Subkowski, T.; Otter, R.; Meyer, T.; Schmalzing, G.; Hillen, H. Molecular characterization of human and bovine endothelin converting enzyme (ECE-1). *FEBS Lett.* **1994**, *356* (2–3), 238–243.
(40) Strausberg, R. L.; Feingold, E. A.; Grouse, L. H.; Derge, J. G.; Klausner, R. D.; Collins, F. S.; Wagner, L.; Shenmen, C. M.; Schuler, G. D.; Altschul, S. F.; Zeeberg, B.; Buetow, K. H.; Schaefer, C. F.; Bhat, N. K.; Hopkins, R. F.; Jordan, H.; Moore, T.; Max, S. I.; Wang, J.; Hsieh, F.; Diatchenko, L.; Marusina, K.; Farmer, A. A.; Rubin, G. M.; Hong, L.; Stapleton, M.; Soares, M. B.; Bonaldo, M. F.; Casavant, T. L.; Scheetz, T. E.; Brownstein, M. J.; Usdin, T. B.; Toshiyuki, S.; Carninci, P.; Prange, C.; Raha, S. S.; Loquellano, N. A.; Peters, G. J.; Abramson, R. D.; Mullahy, S. J.; Bosak, S. A.; McEwan, P. J.; McKernan, K. J.; Malek, J. A.; Gunaratne, P. H.; Richards, S.; Worley, K. C.; Hale, S.; Garcia, A. M.; Gay, L. J.; Hulyk, S. W.; Villalon, D. K.; Muzny, D. M.; Sodergren, E. J.; Lu, X.; Gibbs, R. A.; Fahey, J.; Helton, E.; Ketteman, M.; Madan, A.; Rodrigues, S.; Sanchez, A.; Whiting, M.; Madan, A.; Young, A. C.; Shevchenko, Y.; Bouffard, G. G.; Blakesley, R. W.; Touchman, J. W.; Green, E. D.; Dickson, M. C.; Rodriguez, A. C.; Grimwood, J.; Schmutz, J.; Myers, R. M.; Butterfield, Y. S.; Krzywinski, M. I.; Skalska, U.; Smailus, D. E.; Schnerch, A.; Schein, J. E.; Jones, S. J.; Marra, M. A. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (26), 16899–16903.
(41) Pelham, H. R.; Rothman, J. E. The debate about transport in the Golgi--two sides of the same coin. *Cell* **2000**, *102* (6), 713–719.
(42) Li, D.; Roberts, R. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell. Mol. Life Sci.* **2001**, *58* (14), 2085–2097.
(43) Tang, T.; Kmet, M.; Corral, L.; Vartanian, S.; Tobler, A.; Papkoff, J. Testisin, a glycosyl-phosphatidylinositol-linked serine protease, promotes malignant transformation in vitro and in vivo. *Cancer Res.* **2005**, *65* (3), 868–878.
(44) Mitsui, S.; Okui, A.; Kominami, K.; Konishi, E.; Uemura, H.; Yamaguchi, N. A novel serine protease highly expressed in the pancreas is expressed in various kinds of cancer cells. *FEBS J.* **2005**, *272* (19), 4911–4923.
(45) Ishida, T.; Kato, S. Role of Asp102 in the catalytic relay system of serine proteases: a theoretical study. *J. Am. Chem. Soc.* **2004**, *126* (22), 7111–7118.
(46) Chen, Y.; Sprung, R.; Tang, Y.; Ball, H.; Sangras, B.; Kim, S. C.; Falck, J. R.; Peng, J.; Gu, W.; Zhao, Y. Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol. Cell. Proteomics* **2007**, *6* (5), 812–819.
(47) Chung, J. J.; Yang, H.; Li, M. Genome-wide analyses of carboxyl-terminal sequences. *Mol. Cell. Proteomics* **2003**, *2* (3), 173–181.
(48) Baker, A.; Sparkes, I. A. Peroxisome protein import: some answers, more questions. *Curr Opin Plant Biol* **2005**, *8* (6), 640–647.
(49) Rawlings, N. D.; Morton, F. R.; Barrett, A. J. MEROPS: the peptidase database. *Nucleic Acids Res.* **2006**, *34*, D270–D272.
(50) Zen, K.; Utech, M.; Liu, Y.; Soto, I.; Nusrat, A.; Parkos, C. A. Association of BAP31 with CD11b/CD18. Potential role in intracellular trafficking of CD11b/CD18 in neutrophils. *J. Biol. Chem.* **2004**, *279* (43), 44924–44930.
(51) Townsley, F. M.; Pelham, H. R. The KKXX signal mediates retrieval of membrane proteins from the Golgi to the ER in yeast. *Eur. J. Cell Biol.* **1994**, *64* (1), 211–216.
(52) Nguyen, M.; Breckenridge, D. G.; Ducret, A.; Shore, G. C. Caspase-resistant BAP31 inhibits fas-mediated apoptotic membrane fragmentation and release of cytochrome c from mitochondria. *Mol. Cell. Biol.* **2000**, *20* (18), 6731–6740.
(53) Chandra, D.; Choy, G.; Deng, X.; Bhatia, B.; Daniel, P.; Tang, D. G. Association of active caspase 8 with the mitochondrial membrane during apoptosis: potential roles in cleaving BAP31 and caspase 3 and mediating mitochondrion-endoplasmic reticulum cross talk in etoposide-induced cell death. *Mol. Cell. Biol.* **2004**, *24* (15), 6592–6607.
(54) Copeland, P. R. Regulation of gene expression by stop codon recoding: selenocysteine. *Gene* **2003**, *312*, 17–25.
(55) Martinez, O.; Goud, B. Rab proteins. *Biochim. Biophys. Acta* **1998**, *1404* (1–2), 101–112.
(56) Leung, K. F.; Baron, R.; Seabra, M. C. Thematic review series: lipid posttranslational modifications. geranylgeranylation of Rab GTPases. *J. Lipid Res.* **2006**, *47* (3), 467–475.
(57) Cullinan, S. B.; Whitesell, L. Heat shock protein 90: a unique chemotherapeutic target. *Semin. Oncol.* **2006**, *33* (4), 457–465.
(58) Colombini, M. VDAC: the channel at the interface between mitochondria and the cytosol. *Mol. Cell. Biochem.* **2004**, *256–257* (1–2), 107–115.
(59) Casadio, R.; Jacoboni, I.; Messina, A.; De Pinto, V. A 3D model of the voltage-dependent anion channel (VDAC). *FEBS Lett.* **2002**, *520* (1–3), 1–7.
(60) Yehezkel, G.; Hadad, N.; Zaid, H.; Sivan, S.; Shoshan-Barmatz, V. Nucleotide-binding sites in the voltage-dependent anion channel: characterization and localization. *J. Biol. Chem.* **2006**, *281* (9), 5938–5946.
(61) Overall, C. M.; Dean, R. A. Degradomics: systems biology of the protease web. Pleiotropic roles of MMPs in cancer. *Cancer Metastasis Rev* **2006**, *25* (1), 69–75.
(62) Overall, C. M.; Tam, E. M.; Kappelhoff, R.; Connor, A.; Ewart, T.; Morrison, C. J.; Puente, X.; Lopez-Otin, C.; Seth, A. Protease degradomics: mass spectrometry discovery of protease substrates and the CLIP-CHIP, a dedicated DNA microarray of all human proteases and inhibitors. *Biol. Chem.* **2004**, *385* (6), 493–504.
(63) Lopez-Otin, C.; Overall, C. M. Protease degradomics: a new challenge for proteomics. *Nat. Rev. Mol. Cell Biol.* **2002**, *3* (7), 509–19.

PR070375K