Supporting Information

NAPT, an unbiased approach for sequential analysis of protein N- and C-terminome

Zhongjie Wang,†[a] Wenjuan Yuan,†[a] Lei Zhang[a], Ying Zhang,[ab] Haojie Lu*[ab]

[a]Institutes of Biomedical Sciences and Shanghai Cancer Centre, Fudan University, Shanghai, 200032, P. R. China

[b]Department of Chemistry and Key Laboratory of Glycoconjugates Research Ministry of Public Health, Fudan University, Shanghai, 200433, P. R. China

†These authors contributed equally to this work.

*Corresponding author, E-mail: luhaojie@fudan.edu.cn

Contents

Materials and Chemicals

Guanidine hydrochloride (GndHCl), 1 M triethylammonium bicarbonate buffer (TEAB), dithiothreitol (DTT), iodoacetamide (IAA), d6-acetic anhydride, ammonium bicarbonate (ABC), trifluoroacetic acid (TFA), acetonitrile (ACN, HPLC grade), 2-morpholinoethanesulfonic acid monohydrate (MES), ethanolamine hydrochloride, N-(3-dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride (EDC), N-hydroxysuccinimide (NHS), and formic acid were bought from Sigma-Aldrich (St. Louis, MO). Halt™ protease inhibitor cocktail (100×) and 50% hydroxylamine (NH_2OH) were purchased from Thermo Fisher Scientific (Waltham, MA). Activated LysargiNase and LysN were obtained from Beijing Shengxia Proteins Scientific Ltd. (Beijing, China). Sodium hydroxide (NaOH), potassium phosphate monobasic (KH_2PO_4), potassium chloride (KCl), phosphoric acid (H_3PO_4), and di-potassium hydrogen phosphate (K_2HPO_4) were ordered from Sinopharm Chemical Reagent Co., Ltd. (China). SepPak C18 cartridges and Oasis MCX cartridges were provided by Waters Corp. (Milford, MA). Distilled water was purified with a Milli-Q system (Waters Corp., Milford, MA).

NAPT sample preparation protocol

Protein extraction

(1) Lyse the harvested Hela cells in 6 M GndHCl, 50 mM TEAB (pH 8.0) and 1% v/v protease inhibitor cocktail via sonication at 0 °C. After centrifugation at 12 000 g, 4 °C for 20 min, collect the supernatant, and measure the concentration of proteins by Bradford assay.

(2) Add 10 mM DTT to the protein lysates and incubate the solution at 37 °C for 1 h. Then add 30 mM IAA and incubate the solution in the dark at 25 °C for 40 min. Extinguish the excess IAA with an additional 10 mM DTT.

Protein d3-acetylation (optional)

(3) Add 4 μL d6-acetic anhydride to 400 μg protein lysate and vortex the solution. Adjust the solution to pH 8 with an adequate volume of 5 M NaOH. Carry out the reaction at room temperature for 1 h.

(4) Repeat (3) twice to ensure the completeness of derivatization.

(5) Terminate the reaction by adding 40 μL of 1 M ABC buffer and incubating the solution at room temperature for 30 min.

Protein amidation

(6) Transfer the solution to a 10-kDa-MWCO filter (Millipore). Centrifuge the filter at 12 000 g, 4 °C and washed it three times with carboxyl protection buffer (4 M GndHCl, 200 mM MES, and 2 M ethanolamine hydrochloride).

(7) Add 100 mM EDC and 50 mM NHS (both dissolved in the protection buffer) to the filter and carry out the reaction at room temperature for 2 h.

(8) Centrifuge the filter at 12 000 g, 4 °C to cast away the reaction buffer.

(9) Repeat (7) to achieve more completeness of derivatization.

Protein digestion

(10) Centrifuge the filter at 12 000g, 4 °C and wash it at least five times with 50 mM TEAB.

(11) Add 400 μL 50 mM TEAB and activated LysargiNase at an enzyme/substrate ratio of 1:20 into the filter. Carry out the digestion process at 37 °C overnight.

Note: The activated LysargiNase is a ~29 kDa segment containing Arg^{61}-Ala^{322} of the full-length LysargiNase. Thus, the preincubation step with Ca^{2+} in other LysargiNase-related protocols can be omitted. According to the Product Information kindly provided by the manufacturer, activated LysargiNase is vulnerable to 1 mM DTT, 0.1 mM GndHCl, 0.8 M urea, or 5% ACN (v/v). An enzyme/substrate ratio of 1:20 in 20~50 mM HEPES or Tris-HCl buffer, pH 7.5 at 37 °C, is recommended. Here, we used 50 mM TEAB, pH 8.0, as the digestion buffer to eliminate unwanted EDC-tyrosine adducts.

(12) Centrifuge the filter at 12 000g for 20 min and collect the filtrate. Then wash the filter by adding 400 μL of 50 mM TEAB and centrifuge it at 12 000g for 20 min. Collect the filtrate again and combine the obtained filtrates together.

(13) (If protein d3-acetylation was not conducted), add LysN at an enzyme/substrate ration of 1:50 and carry out the digestion process at 37 °C for another 12 h to reduce the missed cleavage rate of lysine residues.

Afterward treatments

(14) Concentrate the obtained peptide solution to approximately 400 μL through a vacuum drier. Then add 400 μL 1 M TEAB buffer and incubate the solution at 37 °C for 12 h to reverse the unwanted EDC-tyrosine adducts.

(15) (If protein d3-acetylation was conducted), add 10 μL 50% NH_2OH to reverse the unwanted partial d3-acetylation on Ser/Thr/Tyr residues. Carry out the reaction at room temperature for 30 min.

(16) Desalt the peptide samples with SepPak C18 cartridges and lyophilize them in a vacuum drier.

Peptide fractionation through SCX-SPE column.

(17) Oasis MCX Cartridges were used for peptide fractionation. Specifically, 900 μL acidic buffer 7 was added to wash the SCX-SPE. Then it was equilibrated by adding 2 × 900 μL acidic buffer 1. Lyophilized sample peptides were dissolved in 900 μL acidic buffer 1 and loaded onto SCX-SPE. After sample loading, another 600 μL acidic buffer 1 was added. These two fractions were combined and marked as “fraction A1”. Then 2 × 750 μL acidic buffer 2~4 and 2 × 750 μL basic buffer 1~8 was added in sequence and the resulting fractions were marked as “fraction A2~A4” and “fraction B1~B8”,

respectively. The detailed compositions of all buffers mentioned above were listed in Table S1. Afterwards all twelve fractions were lyophilized in a vacuum drier, redissolved in 900 μL 0.1% TFA (v/v) and desalted with SepPak C18 cartridges. The peptide concentration of each fraction was measured by BCA assay. A total of 1 μg peptides from each fraction were preserved for LC-MS/MS analysis.

Data acquisition and processing

LC-MS/MS analysis

LC-MS/MS analysis of fraction samples were performed on a nano-HPLC chromatography system connected to a hybrid trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer (TIMS-TOF Pro, Bruker Daltonics) via a CaptiveSpray nano-electrospray ion source. A total of 200 ng peptides dissolved in solvent A (0.1% formic acid) was loaded onto the analytical column (75 μm i.d. × 25 cm) and separated with a 60 min gradient (2−22% solvent B (ACN with 0.1% formic acid) for 45 min, 22−37% B for 5 min, 37−80% B for 5 min, and then 80% B for 5 min). The flow rate was maintained at 300 nL/min. For MS analysis, the accumulation and ramp time were set as 100 ms each. Survey full-scan MS spectra (m/z 100–1700) were obtained in positive electrospray mode. The ion mobility was scanned from 0.7 to 1.3 Vs/cm2. The overall acquisition cycle of 1.16s comprised one full TIMS-MS scan and 10 parallel accumulation-serial frag-mentation (PASEF) MS/MS scans. During PASEF MSMS scanning, the collision energy was ramped linearly as a function of the mobility from 59 eV at 1/K0 = 1.6 Vs/cm2 to 20 eV at 1/K0 = 0.6 Vs/cm2.

Data analysis

Raw MS data files were searched against the Swiss-Prot database (downloaded on October 25, 2021, containing 20 386 protein sequence entries) by using PEAKS Online Xpro Software (v1.4) for peptide and protein identifications. Only peptides with length between 6 to 45 amino acid residues were taken into consideration. Mass tolerances were set as 15 ppm for parent ions and 0.05 Da for fragments. LysargiNase and ArgN were selected as the protease for non-d3-acetylated and d3-acetylated samples, respectively, allowing three missed cleavages. The digestion mode was set as semi-specific. Carbamidomethylation on cysteine (+57.021464 Da) and d3-acetylation on lysine (+45.029395 Da for d3-acetylated samples only) were selected as fixed modifications. Oxidation on methionine (+15.994915 Da), acetylation on protein N-term (+42.010565 Da), formylation on protein N-term (+27.994915 Da), d3-acetylation on protein N-term (+45.029395 Da for d3-acetylated samples only), amidation by

ethanolamine (ETA) on aspartate, glutamate, and peptide C-term (+43.042199 Da), and EDC-tyrosine adduct (+155.142248) were selected as variable modifications, allowing a maximum of six variable modifications per peptide. False discovery rate (FDR) at PSM and protein level was controlled below 1%.

Data Availability.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD030791. Reviewers may get access through https://www.iprox.cn/page/PSV023.html;?url=1643447260287dOGC. Reviewer's password: EcsY.

Optimization of NAPT workflow

EDC-tyrosine adducts and missed cleavage sites remained two unignorable factors twisting the behavior of internal peptides by bringing them extra positive charges, elongating their retention time on SCX SPE cartridges, and consequently mixing them up with C-terminal peptides (Fig. S2a). Initially discovered by Li et al.,[1] EDC-tyrosine adducts could effectively be erased through incubation with TEAB buffer. At first, we selected 50 mM Tris-HCl, pH 7.5 as the digestion buffer. As shown in Fig. S2b, while EDC-tyrosine adducts occurred on 49.8% of the tyrosine-containing peptides, the proportion was diminished to 20.6% if the peptides were incubated with 500 mM TEAB for 12 h after digestion. Further shifting the digestion buffer from 50 mM Tris-HCl to 50 mM TEAB for another 12-hour incubation, we found that only 5.5% of the tyrosine-containing peptides were modified, which was in accordance with Li et al.'s report.

As for the events of missed cleavage, we then fully investigated a peptide list generated from LysargiNase digestion (n=45 630). It was found that among 6754 missed cleavage sites corresponding to 8426 peptides (18.5% for missed cleavage rates), 5621 were accounted for lysine residues, which was in agreement with the previous report (Fig. S2c, Fig. S3).[2] To address this problem, we introduced LysN digestion rightly after LysargiNase digestion. Consequently, among 51 131 identified peptides, 5076 missed cleavage sites corresponding to 5952 peptides (11.6% for missed cleavage rates) were obtained, with 3852 accounted for lysine residues. It should also be noticed that the procedure of LysN digestion allowed further 12-hour incubation of peptides with TEAB buffer, reducing the ratio of EDC-tyrosine adducts to 2.5% (Figure S2b).

A crucial parameter for NAPT workflow was the time point shifting pH when N-terminal peptides were eluted while internal and C-terminal peptides were still bound on the SPE cartridges. As already mentioned in the manuscript, internal peptides and C-terminal peptides produced by LysargiNase generally possess two positive charges at pH 2.7, and N-terminal peptides ordinarily bear two less positive charges, leaving one-positive-charge tolerance for various accidents. Therefore, we mainly focused on proportion of peptides with charge states at +1 or lower and ratio of N-terminal peptides

from fractions eluted by acidic buffer 1~7 (referred as fraction A1~A7), which contained 0~500 mM KCl, respectively (Table S1). The charge state of a peptide at pH 2.7 was determined through the following equation:

$$z_{2.7} = (1 - n_{formyl} - n_{acetyl}) + n_{Lys} + n_{Arg} + n_{His} + n_{EDC-Tyr} \text{ (Equation S1)}$$

where "$z_{2.7}$" represented "charge state at pH 2.7", "1" implied the positive charge on a peptide's N-terminus, "n" was for "the number of", "formyl" meant formylation on protein N-term, which was the second commonly occurred protein N-terminal post translational modification (PTM)[3, 4] neutralizing positive charges of amine groups like acetylation. All other related chemical groups or amino acid residues were expressed by their abbreviations.

As shown in Fig. S4, direct analysis of an unfractionated Hela digest showed that only 5.4% of the identified peptides possessed charge states at +1 or lower and 1.3% were annotated to N-terminal peptides. As for fractionated samples, with the increase of elution strength, proportion of peptides with charge states at +1 or lower, as well as ratio of N-terminal peptides, decreased drastically. We noticed that fraction A4 (eluted by acidic buffer 4 containing 30 mM KCl) was the last fraction enriching both peptides with charge states at +1 or lower (22.4%) and N-terminal peptides (2.5%). Therefore, we determined to shift pH from 2.7 to 8.5 rightly after fraction A4 was eluted.

Once the pH was changed, charge states of internal peptides were reduced while those of modified C-terminal peptides remained the same. Hence, peptides with charge states at +2 or higher at pH 8.5 were monitored instead. The charge state of a peptide at pH 8.5 was determined through the following equation:

$$z_{8.5} = (1 - n_{formyl} - n_{acetyl}) + n_{Lys} + n_{Arg} + n_{EDC-Tyr} - (n_{Asp} + n_{Glu} + 1 - n_{ETA})$$

(Equation S2)

where "$z_{8.5}$" represented "charge state under pH 8.5", the former "1" implied the positive charge on a peptide's N-terminus, and the latter "1" was for the negative charge on a peptide's C-terminus. All related chemical groups or amino acid residues were expressed by their abbreviations.

Determination of N-terminal acetylation levels

In order to further evaluate to which extent were these stabilizing N-terminal residues acetylated, a semi-quantitative assessment based on peptide-spectrum matches (PSMs) was established as follows: 1) If all the identified N-terminal peptides corresponding to a certain N-terminus were acetylated, it was considered 100% acetylated, and vice versa. 2) If a certain N-terminus was observed both in its acetylated and unacetylated forms, its acetylation level was determined through the equation below:

$$\alpha_{acetyl} = \frac{n_{acetylated\ PSMs}}{n_{all\ PSMs}} \times 100\% \qquad \text{(Equation S3)}$$

where α_{acetyl} was for the acetylation level of a certain protein N-terminus, $n_{acetylated\ PSMs}$ represented number of PSMs corresponding to its acetylated forms, and $n_{all\ PSMs}$ meant number of PSMs corresponding to the protein N-terminus.

Fig. S7 exhibited the acetylation level of all identified N-termini initiated with stabilizing N-terminal residues or sequences except for MC- and MW- (n<10). The acetylation level of certain N-terminal residue/sequence was calculated by averaging acetylation levels of all identified protein termini initiated with it, as listed in Table S4. Overall, based on the occurrence of each N-terminal residue/sequence, the average acetylation level of the human proteome was estimated to be 65%.

Supplementary Figures

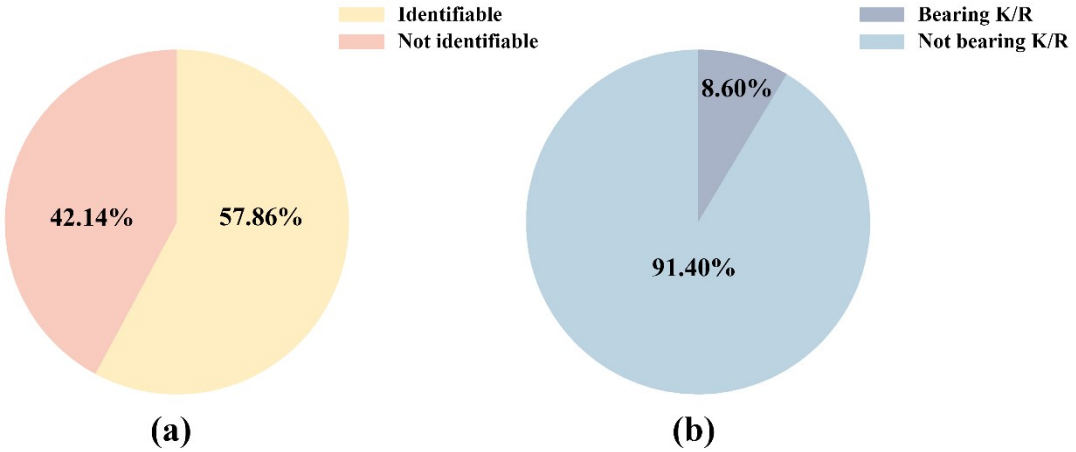


Figure S1. (a) In silico analysis of the "Identifiable" and "Not identifiable" protein N termini in the human proteome digested with LysargiNase. Peptides with 6-45 amino acid residues were considered identifiable by MS. Swiss-Prot database (Homo sapiens, 20 386 entries) was used. (b) Proportion of "peptides containing lysine or arginine residues", and "peptides not containing lysine and arginine residues" to identifiable N-termini generated in (a).

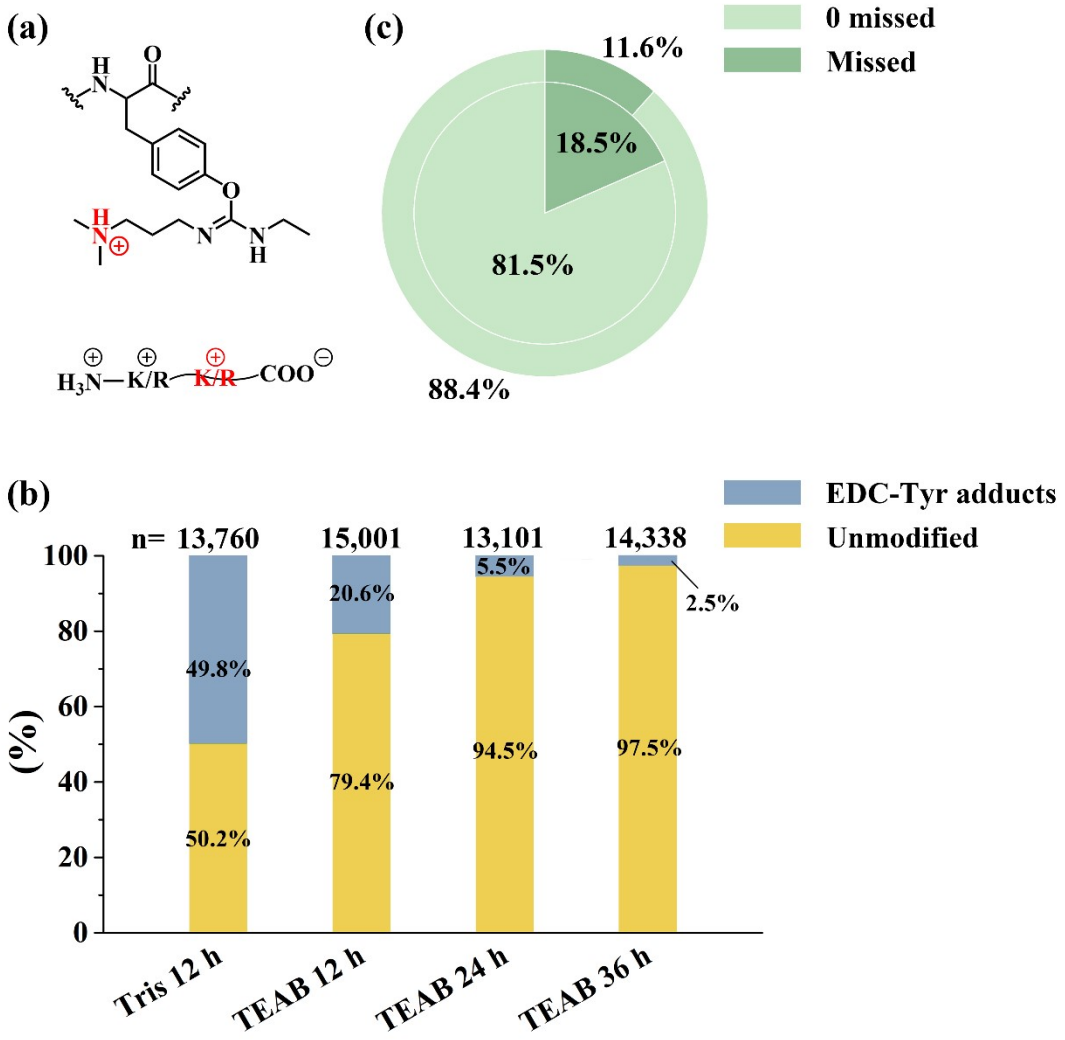


Figure S2. (a) Structure of EDC-tyrosine adducts (upper panel) and peptides with missed cleavage sites generated by LysargiNase digestion (lower panel). Substructures which brought extra positive charges were labeled red. (b) Ratio of peptides with EDC-tyrosine adducts to tyrosine-containing peptides under different incubating conditions. (c) Missed cleavage rates of peptides produced by LysargiNase digestion alone (inner circle) or followed by LysN digestion (outer circle).

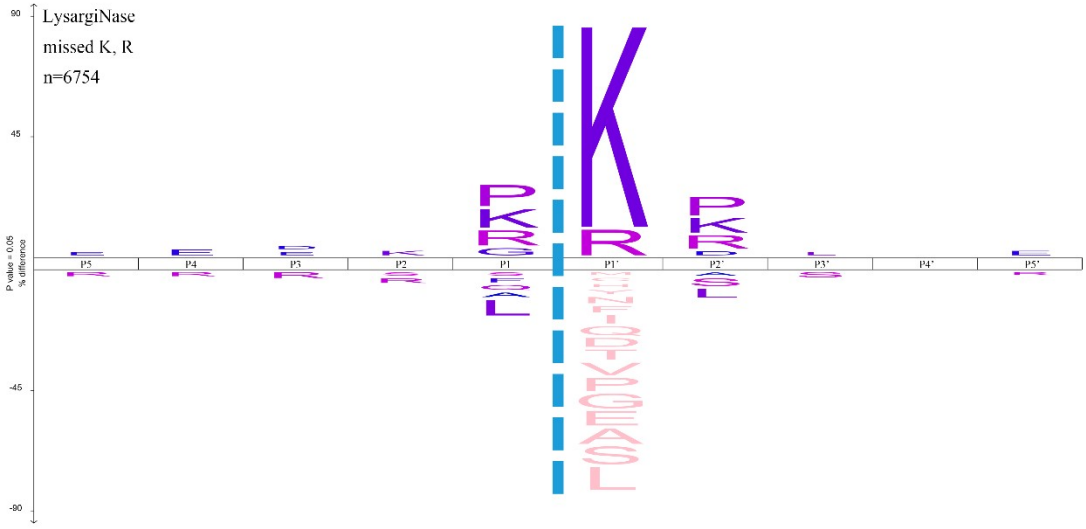


Figure S3. IceLogo[5] showing missed cleavage sites generated by LysargiNase.

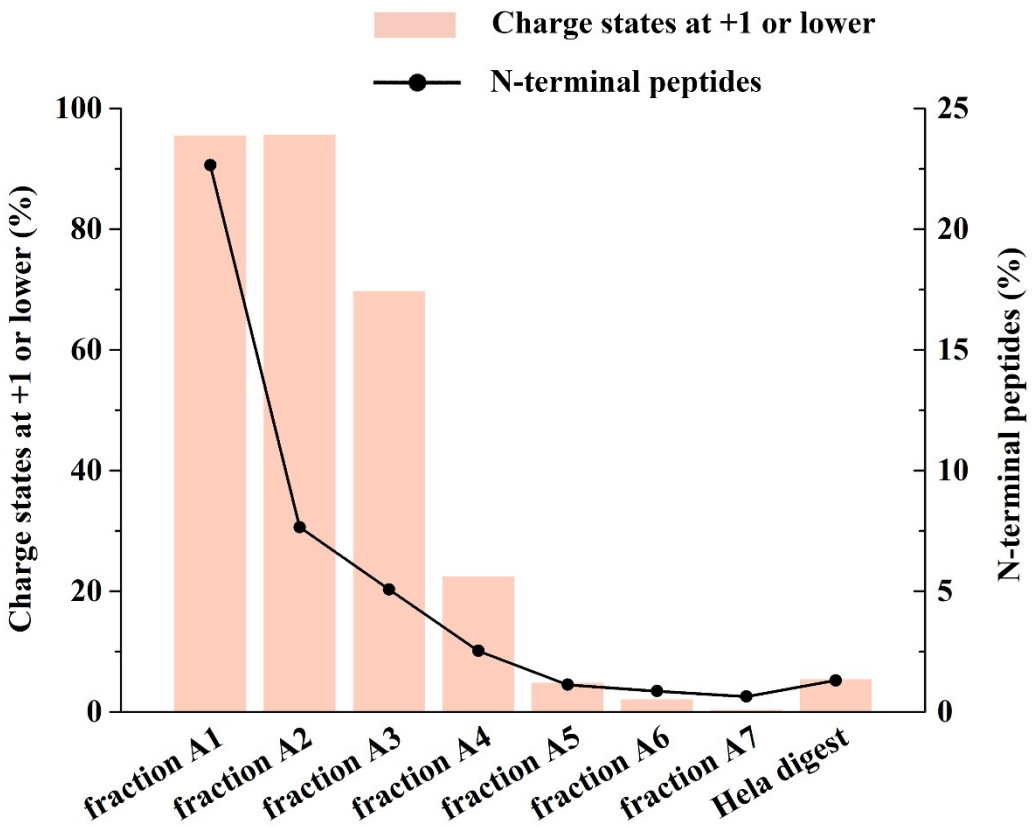


Figure S4. Proportion of peptides with charge states at +1 or lower and ratio of N-terminal peptides identified from each fraction and an unfractionated Hela digest.

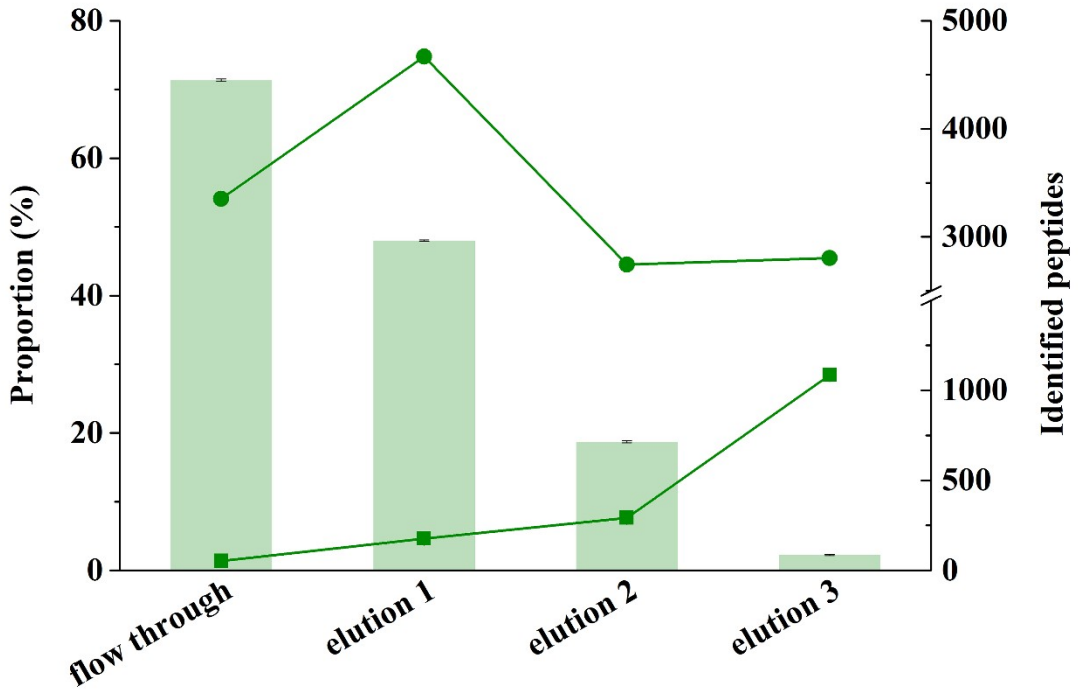


Figure S5. Proportion of peptides with charge states at +1 or lower (green bar), number of identified terminal peptides (green line plus circle) and histidine-containing terminal peptides (green line plus square) from four fractions in SAPT workflow.[4]

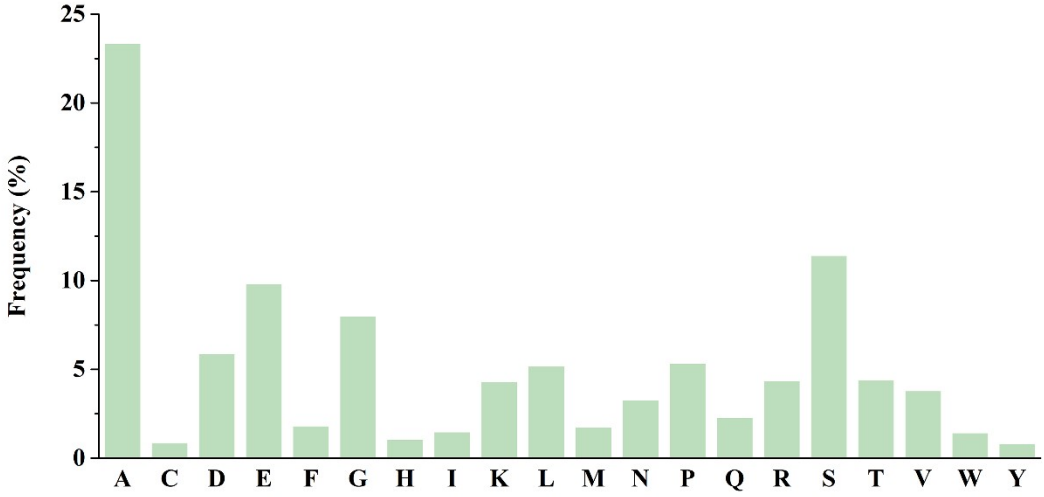
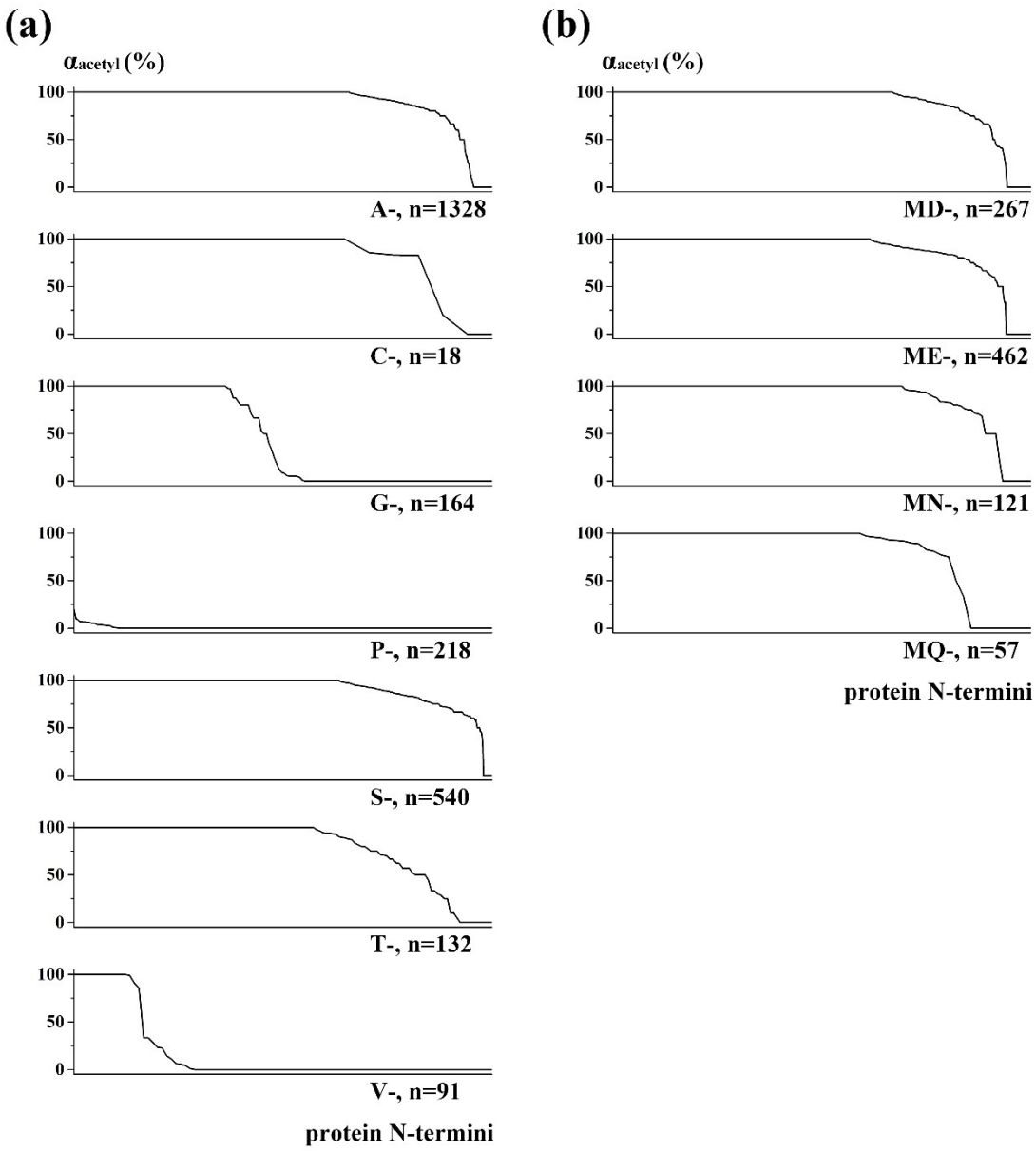


Figure S6. Frequency of each amino acid residue at position 2 across the human proteome. Swiss-Prot database (Homo sapiens, 20 386 entries) was used.


(a)
αacetyl (%)
100
50
0
A-, n=1328
C-, n=18
G-, n=164
P-, n=218
S-, n=540
T-, n=132
V-, n=91
protein N-termini
(b)
αacetyl (%)
MD-, n=267
ME-, n=462
MN-, n=121
MQ-, n=57
protein N-termini

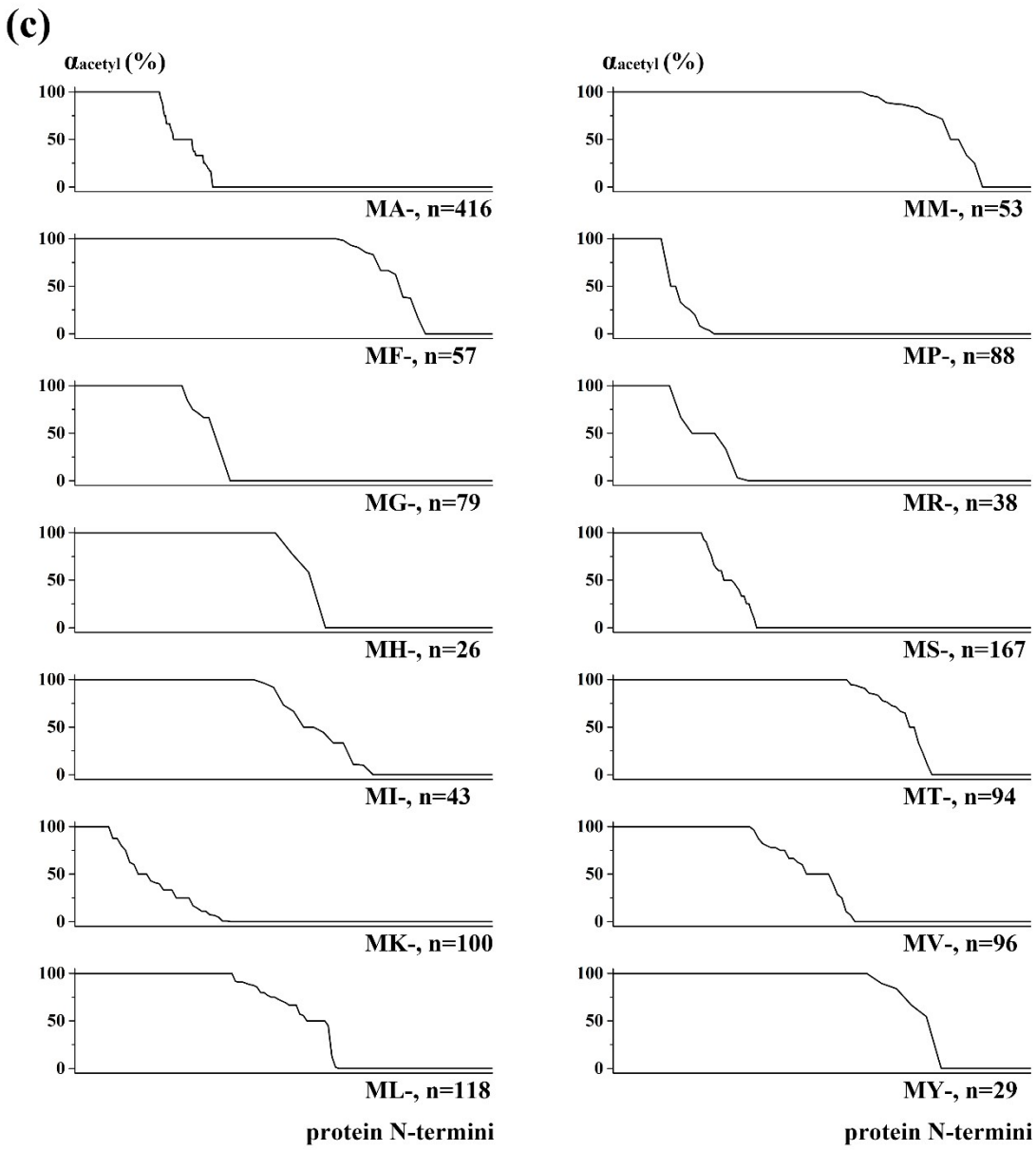


Figure S7. Acetylation level of identified protein N-termini initiated with different species. (a) Substrates of N-terminal acetyltransferase A (Nat A). (b) Substrates of Nat B. (c) Substrates of Nat C/E/F. Nt-species corresponding to which no less than 10 identified protein N-termini were selected.

Supplementary Tables

Table S1. Composition of buffer used during the SCX fractionation

Buffer	Composition
Acid 1	5 mM KH_2PO_4, pH 2.7, 30% ACN
Acid 2	5 mM KH_2PO_4, 10 mM KCl, pH 2.7, 30% ACN
Acid 3	5 mM KH_2PO_4, 20 mM KCl, pH 2.7, 30% ACN
Acid 4	5 mM KH_2PO_4, 30 mM KCl, pH 2.7, 30% ACN
Acid 5	5 mM KH_2PO_4, 40 mM KCl, pH 2.7, 30% ACN
Acid 6	5 mM KH_2PO_4, 50 mM KCl, pH 2.7, 30% ACN
Acid 7	5 mM KH_2PO_4, 500 mM KCl, pH 2.7, 30% ACN
Basic 1	5 mM K_2HPO_4, pH 8.5, 30% ACN
Basic 2	5 mM K_2HPO_4, 10 mM KCl, pH 8.5, 30% ACN
Basic 3	5 mM K_2HPO_4, 20 mM KCl, pH 8.5, 30% ACN
Basic 4	5 mM K_2HPO_4, 30 mM KCl, pH 8.5, 30% ACN
Basic 5	5 mM K_2HPO_4, 40 mM KCl, pH 8.5, 30% ACN
Basic 6	5 mM K_2HPO_4, 50 mM KCl, pH 8.5, 30% ACN
Basic 7	5 mM K_2HPO_4, 100 mM KCl, pH 8.5, 30% ACN
Basic 8	5 mM K_2HPO_4, 500 mM KCl, pH 8.5, 30% ACN

Table S2. List of identified N-terminal and C-terminal peptides from Hela cell using LysargiNase as protease.

Table S3. List of identified N-terminal and C-terminal peptides from Hela cell using LysargiNase as protease after protein d3-acetylation.

Table S4 Acetylation levels of different N-terminal species

Substrates of NatA (including P-)		Substrates of NatB	
Nt-residues	Acetyl(%)	Nt-sequences	Acetyl(%)
A-	89.5	MD-	88.4
C-	81.8	ME-	87.9
G-	45.0	MN-	87.3
P-	0.6	MQ-	79.7
S-	91.0		
T-	79.3		
V-	18.2		
Substrates of NatC/E/F			
Nt-residues	Acetyl(%)	Nt-sequences	Acetyl(%)
MA-	26.3	MM-	79.3
MF-	76.1	MP-	15.9
MG-	32.5	MR-	22.5
MH-	55.2	MS-	27.9
MI-	57.2	MT-	69.4
MK-	19.0	MV-	47.1
ML-	54.2	MY-	72.2

References

(1) Q. Li, Y. Zhang, Z. Wu, J. Huang, N. Yue, L. Huang and X. Zhang, *Anal. Chem.*, 2021, **93**, 697-703.

(2) P. F. Huesgen, P. F. Lange, L. D. Rogers, N. Solis, U. Eckhard, O. Kleifeld, T. Goulas, F. X. Gomis-Ruth and C. M. Overall, *Nat. Methods*, 2015, **12**, 55-58.

(3) S. Ju, Y. Kwon, J. Kim, D. Park, S. Lee, J. Lee, C. Hwang and C. Lee, *Anal. Chem.*, 2020, **92**, 6462-6469.

(4) Z. Wang, L. Zhang, W. Yuan, Y. Zhang and H. Lu, *Anal. Chem.*, 2021, **93**, 10553-10560.

(5) N. Colaert, K. Helsens, L. Martens, J. Vandekerckhove and K. Gevaert, *Nat. Methods*, 2009, **6**, 786-787.