Discovering the N-terminal Methylome by Repurposing of Proteomic Datasets

*Panyue Chen[1], Tiago Jose Paschoal Sobreira[2], Mark C. Hall[3,4] and Tony R. Hazbun[1,4]**

1. Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN 47907
2. Bindley Bioscience Center, Purdue University, West Lafayette, IN 47907
3. Department of Biochemistry, Purdue University, West Lafayette, IN 47907
4. Purdue Center for Cancer Research, Purdue University, West Lafayette, IN 47907

Corresponding Author:

*Tony R. Hazbun

Address: Purdue University, HANS 235, 201 South State St, West Lafayette, IN 47907

Tel: 765-496-8228

Email: thazbun@purdue.edu

Abstract

Protein α-N-methylation is an underexplored post-translational modification involving the covalent addition of methyl groups to the free α-amino group at protein N-termini. To systematically explore the extent of α-N-terminal methylation in yeast and humans, we reanalyzed publicly accessible proteomic datasets to identify N-terminal peptides contributing to the α-N-terminal methylome. This repurposing approach found evidence of α-N-methylation of established and novel protein substrates with canonical N-terminal motifs of established α-N-terminal methyltransferases, NTMT1/2 for humans, and Tae1 for yeast. NTMT1/2 has been implicated in cancer and aging processes. Moreover, α-N-methylation of non-canonical sequences was surprisingly prevalent, suggesting unappreciated and cryptic methylation events. Analysis of the amino acid frequencies of α-N-methylated peptides revealed a $[S]_1$-$[S/A/Q]_2$ pattern in yeast and $[A/N/G]_1$-$[A/S/V]_2$-$[A/G]_3$ in humans, which differs from the canonical motif. We delineated the distribution of the two types of prevalent N-terminal modifications, acetylation and methylation, on amino acids at the 1st position. We tested three potentially methylated proteins and confirmed the α-N-terminal methylation of Hsp31 by additional proteomic analysis and immunoblotting. The other two proteins, Vma1 and Ssa3, were found to be predominantly acetylated indicating proteomic searching for α-N-terminal methylation require careful consideration of mass spectra. This study demonstrates the feasibility of reprocessing proteomic data for global α-N-terminal methylome investigations.

The raw MS data that supports the findings of this study was deposited with PRIDE identifier: PXD022833.

Keywords: α-N-methylation, proteomics, repurposing, post-translational modification

Introduction

Protein N-terminal methylation is a novel post-translational modification (PTM) that was first reported decades ago in prokaryotes, whereas the eukaryotic methyltransferases were identified more recently. This PTM involves adding up to three methyl groups on the free α-amino group of the protein N-terminus. The fully methylated state is trimethylation except for proline dimethylation and can result in a pH insensitive positive charge on the protein N-termini.[1] The Tae1/NTMT1/NTMT2 α-N-terminal methyltransferases are highly conserved between yeast and humans, and the enzymes recognize substrates with similar N-terminal sequence motifs, X_1-P_2-[K/R]$_3$ (X position could be A, S, G or P), hereinafter referred to as the canonical N-terminal motif.[2,3] In the motif, the initiating methionine (iMet) is commonly removed during protein maturation, hence leaving the α-amino group on the X amino acid exposed which is subsequently targeted for methylation.

Although numerous putative substrates contain the canonical motif, most proteins containing the canonical motif have not been validated as methylated in yeast and humans. The major N-terminal methyltransferase in yeast that recognizes the X_1-P_2-[K/R]$_3$ motif, Tae1 (translational associated element 1), was previously shown to modify two ribosomal subunits, Rpl25a/b and Rps12ab, and one proteasome subunit

N-terminal methyltransferase	Substrate	N-terminal sequence	Reference
yTae1	Rpl12ab	PPKF	Webb, KJ. (2010)
	Rps25a/b	PPKQ	Webb, KJ. (2010)
	Rpt1 *	PPKE	Kimura, Y. (2013)
yNnt1/yEfm7	Tef1(eEF1A)	GKEK	Hamey, JJ. (2016)
hNTMT1/2	Rb	PPKT	Tooley, CS. (2010)
	CENP-A	GPKR	Bailey, AO. (2013) Sathyan, KM. (2017)
	CENP-B	GPKR	Dai, X. (2013)
	RCC1	SPKR	Tooley, CS. (2010)
	DDB2	APKK	Cai, Q. (2014)
	PARP3	APKP	Dai, X. (2015)
	SETα	APKR	Tooley, CS. (2010)
	KLHL31	APKK	Tooley, CS. (2010)
	MYL2	APKK	Tooley, CS. (2010)
	MYL3	APKK	Tooley, CS. (2010)
	MYL9	SSKR	Neivtt, C. (2018)
	Rpl23A	APKA	Tooley, CS. (2010)
	Rpl12	PPKF	Webb, KJ. (2010)
	Rps25	PPKD	Webb, KJ. (2010)
	OLA1	PPKK	Jia, K. (2019)
	MRG15	APKQ	Bade, D. (2021)
hMETTL13	eEF1α	GKEK	Jakobsson, ME. (2018)

Table 1. N-terminal methyltransferase and verified α-N-terminal methylated substrates in yeast and humans. * indicates the corresponding methyltransferase is not confirmed experimentally.

Rpt1.[4,5] These three substrates are involved in protein synthesis and degradation, but the modification's exact role is unclear. The prevailing view is that N-terminal methylation affects the assembly of the polysome and thus contributes to protein synthesis efficiency and fidelity.[6] Notably, a total of 45 proteins in the yeast proteome contain the canonical motif, and Tae1 is implicated in multiple biological pathways (Table S1). In humans, there are two primary enzymes responsible for N-terminal methylation, NTMT1 and NTMT2, that target a broad range of substrates associated with diverse biological pathways (Table 1) and have been implicated in cancer and aging.[4,5,7–17] N-terminal methylation has been implicated in regulating protein-protein interactions and protein-DNA interactions.[2] Trimethylation on CENP-A is crucial for constitutive centromere complex formation by recruiting CENP-T and CENP-I and contributes to cell cycle progress and cell survival.[9,10] Loss of the trimethylation on CENP-B prevents binding to its centromeric DNA motif.[11] Other research also suggests that N-terminal methylation might regulate protein-DNA interactions by creating a positive charge on the substrate protein, thus granting a strong ion-ion interaction with the negatively charged DNA backbone.[18] In addition, N-terminal methylation may cooperate with N-terminal acetylation to regulate protein localization and differential interactions because N-terminal methylation increases MYL9 binding to DNA in the nucleus while N-terminal acetylation contributes to its interaction with cytoskeletal proteins.[14] The other enzyme with known α-N-terminal methylation activity besides Tae1 in yeast is Nnt1/Efm7, which is also capable of lysine methylation. Nnt1 appears only to target a single substrate, the translational factor Tef1/eEF1A, and recognizes the Tef1 N-terminal G_1-K_2-E_3-K_4 sequence.[7] In humans, METTL13 methylates eEF1α and is the functional homolog to Nnt1 despite the lack of sequence similarity between the enzymes (Table 1).[16,19] To our knowledge, N-terminal methylation of proteins without the X_1-P_2-[K/R]$_3$ motif or G_1-K_2-E_3-K_4 sequence have not been previously reported in eukaryotes and the extended α-N-terminal methylation of non-canonical sequences is unclear.

Several methods have been used to confirm α-N-terminal methylation, including immunodetection and tandem mass spectrometry. Although antibodies have been developed for specific N-terminally methylated proteins, they have limited utility and are not suited for proteomic enrichment. Mass spectrometry is widely used for protein identification, detecting PTMs at a single protein level or proteomic level and studying protein N-terminomics. Numerous strategies have been developed for specific enrichment or quantitation of PTMs on protein N-termini, such as α-N-terminal acetylation. These techniques include quantitative isotopic labeling methods such as TAILS and SILAC,[20] and negative enrichment methods such as COFRADIC[21] and ChaFRADIC[22] analysis. Recently, there has been an effort to encourage archiving of proteomic datasets on various proteomic consortium databases with public accessibility, such as PRoteomics IDEntification Database (PRIDE)[23], ProteomeXchange[24] and iProx[25]. Proteomic datasets deposited in these databases were generated for various purposes, however, they contain information about N-terminal methylation that was not searched or exploited. Thus, by crafting the search parameters for N-terminal methylation in sequence-based searching engines, reanalysis of datasets or repurposing would be an unbiased approach to investigating the N-terminal methylome landscape. Besides, many platforms and tools are established to facilitate datasets reanalysis such as the trans proteomics pipeline/PTMProphet[26,27] and SearchGUI/PeptideShaker.[28]

This report demonstrates the utility of searching proteomic datasets generated by several techniques for methylated protein N-terminal peptides. We verified the presence of known α-N-methylated substrates and detected a collection of potentially α-N-terminal methylated proteins. We observed a close association between α-N-terminal methylation and the amino acid type at the first position of the protein sequence by analyzing the sequence pattern. By investigating the distribution of α-N-terminal methylation and α-N-terminal acetylation on various amino acids at the first position, we were able to characterize the sequence patterns of modifications across the proteome and between the two spices. Surprisingly, the majority of N-terminal peptides identified by MS search did not have the canonical motif. We endeavored to validate three of the potential protein hits with purified overexpressed protein from yeast. We found that detection of methylation from this dataset repurposing approach can be elusive because experimental analysis of two non-canonical candidate hits, Ssa3 and Vma1, were revealed to be fully acetylated rather than the methylation initially predicted. However, we confirmed that the canonical motif-containing protein, Hsp31, is methylated. This study reveals intriguing global patterns of α-

N-terminal methylation and serves as a foundation for further proteomic investigation of the N-terminal methylome.

Experimental Procedures.

Script for Mascot automatic searching

A Perl script was created to automate searching and avoid overloading of the server. The Perl script allows downloading of raw files from the ftp server such as PRIDE or iProx by wget and subsequently convert them into .mgf files by msconverter tool, Proteowizard. MINE files with parameters described below were generated for corresponding datasets and used for Mascot searching by the command "nph-mascot.exe". Mascot searching results were exported into .csv file by the command "export_dal_2.pl". Finally, the methylated peptides were parsed and retrieved into .csv files. The script will be made available upon request.

Yeast culturing and protein overexpression

Candidate proteins were purified using the movable open reading frame (MORF) collection from Dharmacon, where yeast genes were Gateway cloned into pBG1805 in frame with a C-terminal triple affinity tag, 6xHis-HA-3C protease site-Protein A. Protein expression is regulated by the *GAL* promotor.[29] Yeast storage, culture and protein purification protocol was slightly modified from the Dharmacon technical manual.[29] Yeast strains were streaked and stored on a synthetic complete agar plate with uracil dropout (SC-URA). On day 1, a single colony was inoculated into 2 mL SC-URA plus 2% glucose liquid media for overnight culture. On day 2, the culture was diluted into 20 mL SC-URA plus 2% raffinose at OD_{600}=0.1 and incubated overnight. The culture was diluted into 800 mL of SC-URA plus 2% raffinose at day 3. When OD_{600} reached 0.8-1.2, 400 mL of 3X YP (3% yeast extract, 6% peptone and 6% galactose) media was added to induce the protein expression for 6 h. The yeast pellet was harvested by centrifuging at 4200 rpm for 10 min at 4 °C. The pellet was washed with 10 mL sterilized deionized water once and stored at -80 °C.

Protein purification

Frozen yeast pellets were thawed from -80 °C and resuspended in CE buffer (50 mM pH 7.5 Tris-Cl, 1 mM EDTA, 4 mM $MgCl_2$, 5 mM DTT, 10% glycerol, 0.75 M NaCl). PMSF (Promega) was added to 1 mM. The cell suspension was distributed into microcentrifuge tubes with an equal volume of 0.5 mm glass beads (Biospec). Yeast cells were lysed by bead beater (Biospec) for 7 cycles of 25 s interspersed with cooling on ice for 30 s. The lysate was centrifuged at 13000 rpm for 15 min. Clarified lysate supernatant was diluted into IPP0 buffer (10 mM pH 8 Tris-Cl, 0.1% NP40) and incubated with triple-washed IgG Sepharose 6 Fast Flow bead (Cytiva) for 2 h. IgG bead was separated, washed twice with IPP150 buffer (10 mM pH 8 Tris-Cl, 0.1% NP40, 150 mM NaCl) and subsequently wash twice with 3C cleavage buffer (10 mM pH8 Tris, 150 mM NaCl, 0.5 mM EDTA, 1 mM DTT, and 0.1% NP40). Bead was resuspended in 3C cleavage buffer and protein of interest was cleaved from bead using 3C protease (Acro Biosystems) with overnight digestion. The 3C protease was removed by glutathione agarose (Scientific). The eluent was passed through a 0.22 μM Spin-X centrifuge tube filter (Corning). Protein concentration was determined by BCA assay (Thermo Scientific).

In solution digestion for Mass Spectrometry

The sample preparation protocol was adjusted from the literature.[30] Protein purified and eluted from protein A agarose bead was used for sample preparation. Cold acetone (-20 °C) was added into the protein solution and stored at -20 °C overnight. Precipitated protein was collected by centrifugation at 15,000 x g for 10 min at 4 °C. The protein pellet was washed once with cold acetone and resuspended in 10 μl of 10 mM dithiothreitol (DTT, Sigma) in 25 mM ammonium bicarbonate and incubate at 37 °C for 1 h. 10 μl of alkylation reagent mixture (97.5% acetonitrile, 0.5% triethylphosphine and 2% iodoethanol) was added to each sample and incubated at 37 °C for 1 h. The sample was dried in a vacuum centrifuge and digested using a barocycler (PBI). Hsp31, Ssa3 and Vma1 was digested by AspN (Sigma), GluC (Promega) and Trypsin (Sigma-Aldrich), respectively. Digested protein was cleaned with a UltraMicroSpin C18 column (The Nest Group) and dried into a pellet. The peptide pellet was reconstituted in 97% deionized water, 3% acetonitrile and 0.1% formic acid (v/v). An aliquot of 2-3 μl was run into the Q Exactive HF hybrid Quadrupole Orbitrap instrument (Thermo Scientific).[30] The α-N-acetylation on Ssa3 was also confirmed by Orbitrap Fusion Lumos Tribrid mass spectrometer (ThermoFisher). Detailed protocols are included in Supporting data.[31–34]

N-terminal methylation searching parameters in Mascot Daemon

Datasets were searched by Mascot Daemon with customized parameters to search for N-terminal methylated peptides in four different datasets as outlined in Table 2 and Table 3. For the ChaFRADIC dataset, iProx SILAC labeling dataset and MISL dataset, a manual search with Mascot Daemon (2.5.1 or 2.7.0.1) was conducted and the N-terminal peptide hit lists were matched with the results from a Perl-driven automatic searching process. A fourth dataset which we termed the Chemical Labeling dataset was searched using the automated script to identify the methylated N-terminal peptides. The exact search parameters used for each dataset is outlined in Table 2 including the variable and fixed modifications, tolerance level, peptide charge state and missed cleavages. Decoy databases were searched in all four cases. Only matched N-terminal peptides with scores greater than 20 were used for hits selection and subsequent bioinformatics analysis. The false discovery rate (FDR) of each Mascot search for the ChaFRADIC dataset, iProx dataset and MISL dataset were manually checked and collected into an excel file (Supporting information Table S2).

Bioinformatic analysis of methylated peptides and proteins

Sequence conservation was graphically visualized using two tools, WebLogo[35] and iceLogo[36]. Four types of sequence logo were generated for each dataset or dataset combinations with the redundant peptide list including iMet, redundant peptide list omitting iMet, nonredundant peptide list including iMet or nonredundant peptide list without iMet. The logo range was limited from the first amino acid position to the third of identified peptides. To generate the redundant list of peptide sequences, methylated peptide sequences were directly extracted from Mascot search export .csv file for each dataset and trimmed to the first three N-terminal amino acids. Duplicated peptide sequences were removed to generate non-redundant lists. For logo analysis on iMet cleaved proteins, peptides containing iMet were excluded from each peptide list. In the WebLogo application, the small sample correction option was utilized to generate the frequency and content logo plots. In the iceLogo web application, the percentage scoring system was used to generate both logos and heat maps (p-value 0.05). The proteome background reference was utilized based on Swiss-Prot database composition of yeast or human proteins. The α-N-terminal methylation sequence patterns of yeast and humans were extracted from heat maps of yeast and humans combined peptide lists (nonredundant and iMet omitted).

Venn diagrams were generated using an online tool, Venny 2.1[37], to analyze the reproducibility across datasets of each species. A list of non-redundant gene names of corresponding protein hits was extracted from the Mascot search export .csv file for each dataset. Three gene name lists of yeast datasets, or two gene name lists of human datasets were used to generate the corresponding Venn diagram of yeast and humans.

Western blot analysis for α-N-terminal methylation on Hsp31

The pBG1805 plasmid with Hsp31 ORF from Dharmacon was transformed into BY4741 WT yeast, *tae1Δ* BY4741 and *tae1-/-*BY4743 yeast. Protein expression and purification was conducted in a similar manner as described in the previous method section. Purified Hsp31 protein was analyzed via SDS-PAGE and western blot. Primary antibody capable of detecting mono-/di-N-terminal methylation (anti-N-me2Ser 2 antibody, a gift from Dr. Schaner-Tooley) was used at 1:1000 dilution.[38] Chemiluminescent anti-rabbit (1:10,000) was used as secondary anybody. The western blot image was developed with ECL reagents and followed by X-ray film exposure.

Results

Discovery of N-terminal methylation of proteins with canonical motif and non-canonical motifs from four proteomic datasets

The following properties are important when considering repurposing a dataset to discover α-N-methylation: N-terminal peptides enrichment methods applied, chemical or isotopic labeling methods employed, type of higher resolution instrumentation, and appropriate protease. Enrichment for protein N-terminal peptides increases the representation of α-N-methylation in the sample. Though employing positive enrichment is hindered by the lack of appropriate N-terminal methylation antibodies, many negative enrichment processes have been developed that deplete internal peptides and have successfully been applied to investigate α-N-acetylation. Hence, we focused on MS datasets using negative selection methods that eliminate free α-N-terminal peptides and internal peptides and consequently increase the α-N-methylome representation. Isobaric modifications, such as α-N-terminal trimethylation (42.047 Da) and α-N-terminal acetylation (42.011 Da), require extra efforts to differentiate with fidelity. MS datasets generated by high-resolution instruments, isotopic labeling of either methyl group or acetyl group, and efficient fragmentation methods are thus favored. Trypsin is the most used protease in proteomics studies but is not optimal for identifying substrates containing the X_1-P_2-$[K/R]_3$ canonical sequences. It cleaves at the carboxyl side of $[K/R]_3$ and yields N-terminal peptides consisting of three amino acids that are poorly detectable.[39] Hence, MS datasets generated by alternative proteases are more optimal for detecting N-terminal peptides.

	ChaFRADIC dataset (PXD000292)	iProx SILAC labeling (IPX0001550000)	MISL dataset (PXD000606)
Species (Repurposed in this paper)	*S. cerevisiae (*WT and *icp55Δ)*	*S. cerevisiae* (*met6Δ*) and Hela	S-adenosylmethionine (SAM) auxotroph *YDR502C S. cerevisiae (sam1Δ, sam2Δ)*
Database	Uniprot_S288C	Uniprot_S288C; Uniprot_human	Uniprot_S288C
Variable modification	Trimethyl (protein N-term) Dimethyl (protein-N-terminal proline) Dimethyl (K) Dimethyl: 2H(4) (K)	Methyl: 2H(3)13C(1)(protein N-term) Dimethyl: 2H(6)13C(2) (protein N-term) Trimethyl: 2H(9)13C(3) (protein N-term)	For light SAM labeled MS files: Acetyl (protein N-term) Methyl (protein N-term) Dimethyl (protein N-term) Trimethyl (protein N-term) For heavy SAM labeled MS files: Acetyl (Protein N-term) Methyl: 2H(3) (protein N-term) Dimethyl: 2H(6) (protein N-term) Trimethyl: 2H(9) (protein N-term)
Fixed modification	Carbamidomethyl (C)	Carbamidomethyl (C) Label: 13C(1)2H(3)(M)	Carbamidomethyl (C)
MS tolerance	10 ppm	10 ppm	50 ppm
MS/MS tolerance	0.02 Da	0.5 Da	0.8 Da
Enzyme	semi ArgC	trypsin/P	trypsin/P
Instrument	Q-Exactive mass spectrometer (Thermo Scientific)	Orbitrap Fusion mass spectrometer (Thermo Scientific)	LTQ-Orbitrap XL mass spectrometer (Thermo)
Peptide charge state	Charge +1 or Charge +2 or Charge +3 or Charge +4 or Charge +5 or Based on which raw file is searched	+2, +3 and +4	+2, +3 and +4
# missed cleavage	2	2	2

Table 2. Searching parameters for ChaFRADIC, iProx and MISL datasets. The parameters used for repurposing each dataset is same as used in the original study except that the variable modifications are crafted for α-N-methylation.

Based on the above criteria, we focused on three types of datasets generated by the following methods: (1) Negative selective N-terminal enrichment method based on charge shift, such as ChaFRADIC method (e.g., ChaFRADIC dataset repurposed in this study); (2) Negative selection N-terminal enrichment method that depletes internal peptides, such as COFRADIC and chemical labeling methods (e.g., Chemical labeling dataset in this work); (3) Quantitative isotopic labeling methods such as ^{13}C labeling (e.g., iProx dataset and MISL datasets in this work). Based on the above factors, we selected four datasets to demonstrate the prevalence of α-N-terminal methylation in yeast and humans. We performed manual searching for three experimental datasets in Mascot Daemon (herein referred to as the ChaFRADIC dataset[22], iProx dataset[40] and MISL dataset[41]). The parameters used for repurposing each dataset were identical to the original study except that variable modification was crafted individually to detect α-N-methylation (Table 2). A more complex Chemical Labeling dataset[42] was searched using a Perl script to automate the searching process with the Mascot server (Table 3). All the datasets in ChaFRADIC had FDRs in the 2-8% range under specified significance level. Most Mascot searches of the iProx dataset had an FDR of 1-4% and most MISL searches had an FDR in the range of 4-9%. The frequency of α-N-terminal methylated peptides identified from decoy databases was monitored for ChaFradic, iProx and MISL datasets and was summarized in Table S2.

	Group 1	Group 2	Group 3	Group 4
Species	HEK293T	HEK293T	HEK293T	HEK293T
Variable modification	1. Methyl (protein N-term) 2. Dimethylation (protein N-terminal) 3. Trimethyl (protein N-term) 4. D6-acetylation: 2H(3) (protein N-terminal)		1. Methyl (protein N-term) 2. Dimethylation (protein N-terminal) 3. Trimethyl (protein N-term) 4. Propionlyation (protein N-terminal)	
Fixed modification	1. Carbamidomethyl (C) 2. D6-acetylation: 2H(3) (K)		1. Carbamidomethyl (C) 2. Propionlyation (K)	
MS tolerance	15 ppm			
MS/MS tolerance	0.5 Da			
Enzyme	Trypsin/P	GluC	Trypsin/P	GluC
Instrument	Q-Exactive mass spectrometer (Thermo Scientific)			
Peptide charge state	+1, +2, +3			
# missed cleavage	1			

Table 3. Parameter settings for Chemical Labeling dataset (PXD0055831). Four subsets are generated with different combination of enzyme and blocking reagent. Each of them is searched with same parameter settings as in original study but the variable modification is crafted for α-N-methylation.

To verify the site of methylation on protein N-termini, the MS/MS spectra of the ChaFRADIC, iProx and MISL datasets were manually examined, and the spectra of yeast hits are available in Figure S1. The number and identity of b/y ions, the maximum number of consecutive b/y ions and peak intensity were checked and recorded in Supporting information (Table S3). All the MS2 spectra have at least two b/y ion matches, although some spectra have weak peak signals. The MS2 spectra for the Chemical Labeling set were not inspected, because it was not feasible to check all the spectra associated with this dataset manually. For most proteins identified, the MS2 spectra support the site of modification on protein N-termini. We refer to the identified proteins as hits in this study due to the limited number of peptides and relatively weak signal of ion fragmentation. Many software have functions available to measure PTM localization probability when more than one modification sites might be present in one peptide, such as Mascot site analysis and MaxQuant PTM scoring[43], and they all utilize MS/MS data. However, it is not possible to calculate the probability of methylation on protein N-termini without introducing other potential modification sites. Considering that lysine and arginine are the two sites most preferred to be methylated, we included mono-/di-/tri-methylation on lysine and mono-/di-methylation on arginine into the Mascot searches and monitored the site analysis probability score

Protein name	Modified peptide Counts in Mascot (pep. score>20)	N-terminal peptide sequence detected
Ssa3	13	S$_{\text{me3}}$RAVGIDLGTTYS S$_{\text{me3}}$RAVGIDLGTTY
Por1	1	S$_{\text{me3}}$PPVYSDISR
Gpm1	1	P$_{\text{me2}}$K$_{\text{me3}}$LVLVR
Rpl15a	1	G$_{\text{me3}}$AYKYLEELQR
Rpl16b	1	S$_{\text{me3}}$QPVVVIDAK$_{\text{me2}}$DHLLGR
Hom2	2	A$_{\text{me3}}$GK$_{\text{me2?}}$K$_{\text{me2?}}$IAGVLGATGSVGQR
Rpl12a*	5	P$_{\text{me2}}$PK$_{\text{me3}}$FDPNEVKYLYLR
Rpl24a	2	**M**$_{\text{me3}}$KVEIDSFSGAK$_{\text{me2}}$IYPGR
Rpl24b	1	**M**$_{\text{me3}}$KVEVDSFSGAK$_{\text{me2}}$IYPGR
Rpl26a	6	A$_{\text{me3}}$KQSLDVSSDRR
Rpl27a	8	A$_{\text{me3}}$K$_{\text{me?}}$FLK$_{\text{me?}}$AGK$_{\text{me?}}$VAVVVR
Tef1	1	G$_{\text{me3}}$K$_{\text{me2}}$EK$_{\text{me2}}$SHINVVVIGHVDSGKSTTTGHLIY

Table 4. Methylated hits from the ChaFRADIC dataset. 12 proteins are tri-methylated while none has the canonical motif recognized by Tae1. The localization of modifications are determined by Mascot site analysis and MS2 spectra. *Rpl12a peptide is detected to have 5 methyl groups which is consistent with reported proline$_1$ dimethylation and lysine$_3$ trimethylation. Ambiguous localization of lysine methylation is labeled with "?".

of Mascot. We were able to recapitulate part of our initial hits with very high probability of protein N-terminal localization (Table S4). These site localization analysis results are consistent with the manual interpretation based on MS/MS spectra and confirm the effectiveness of our manual examination on MS2 spectra and b/y ion series.

In the ChaFRADIC dataset, N-termini of yeast proteins are negatively enriched from *S. cerevisiae* spheroplasts based on charge reduction and two consecutive SCX separations. Briefly, free α- and ε-amino groups on lysine side chains or protein N-termini in the samples were blocked with a dimethyl group by formaldehyde and cyanoborohydride. N-terminal proline is the only exception and can only be mono-methylated during the blocking process.[44] Trypsin was used for sample digestion with ArgC specificity in this case because trypsin recognition and subsequent cleavage are blocked by dimethylated lysine. Digested peptides were fractionated based on their charge state by the 1st SCX (Strong Cation Exchange) separation. Internal peptides in each fraction with free α-amino group released by trypsin digestion were deuteron-acetylated and had their positive charge reduced by one. The change in the charge states of internal peptides lead to a peak shift in the 2nd SCX separation and would be depleted. Only the fractions remaining unchanged in the two SCX separations were used for further MS/MS identification by Q-Exactive mass spectrometry, which contains increased representation of protein N-termini, either protected by extraneous dimethylation or innate modifications.

The ChaFRADIC method selectively enriches modified N-terminal peptides and excludes the internal peptides. However, the blocking method obscures the origin of monomethylation and dimethylation (except dimethylated proline) on protein α-N-amino groups; hence we focused on identifying native trimethylation at protein α-N-termini and dimethylation on protein N-terminal proline. We identified a total of 42 N-terminal methylated peptides corresponding to 12 yeast proteins from 5 SCX fraction MS files, including Ssa3, Por1, Gpm1, Rpl15a, Rpl16b, Hom2, Rpl12a, Rpl24a, Rpl24b, Rpl26a, Rpl27a and Tef1. It should be noted that Ssa3 and Tef1 don't have a C-terminal tryptic end but were included as possible hits. The frequency of identified peptides carrying a non-tryptic end in the complete dataset is ~16%-20%, possibly due to non-specific digestion, solution degradation or impurity of the endonuclease.[45] All of them were putatively identified as protein N-terminal trimethylated, except that Gpm1 and Rpl12a are dimethylated on P_1. The α-N-trimethylation for all proteins was on the first amino acid after iMet is excised, except for Rpl24a/b (Figure S1). The site localization of each hit was confirmed by MS2 spectra and the site analysis score of Mascot (Table S4). Two proteins, Tef1 and Rpl12ab, have previously been reported to be α-N-methylated. Tef1 is the yeast translation elongation factor 1a (eEF1A) known to be α-N-trimethylated by the dual lysine and protein α-N-terminal methyltransferase, Nnt1 (Table 4 and Table S5).[46] Rpl12ab is encoded by two separate genes, *RPL12A* and *RPL12B*, resulting in identical polypeptide sequences. Five methyl groups were assigned to the first three amino acid positions of Rpl12ab in our Mascot searching, which is equivalent to the fully methylation state on both protein N-terminal and K_3 side chain. This observation agrees with the previous report on Rpl12ab.[4,47] To our knowledge, this is the first report of N-terminal trimethylation for the other five hits and surprisingly, they do not contain the canonical motif X_1-P_2-$[K/R]_3$ motif recognized by the predominant α-N-terminal methyltransferases in yeast, Tae1.

The iProx isotopic labeling dataset was designed initially for identifying lysine methylation using $^{13}CD_3$-methionine in *E. coli*, yeast and Hela cell line.[40] We examined $^{13}CD_3$ labeled N-terminal methylation in the datasets for both eukaryotes (yeast and Hela cell line), which differentiate the α-N-acetylation (42.011 Da, UNIMOD) from α-N-terminal trimethylation (54.113 Da, UNIMOD). This iProx dataset and the MISL dataset described below are the two proteomic datasets in our report that used heavy labeled methyl donors. Here, we searched for all heavy α-N-terminal methylation species, isotopic labeled mono-/di-/tri-methylation, whereas only α-N-monomethylation was searched in the original publication.[40] We discovered 11 unique N-terminal methylated peptides (iMet cleaved) from nine unique protein methylation events in the HeLa dataset including seven α-N-monomethylation, three α-N-dimethylation and one α-N-trimethylation. Only the ATG3 (Q_1-N_2-V_3) protein was found to have two α-N-methylation species, monomethylation and dimethylation (Table S5). Identification of multiple methylation states increases the confidence that these proteins are biologically methylated. For the yeast dataset, 11 unique α-N-methylated peptides (iMet cleaved) were detected from two unique proteins including three α-N-monomethylated peptides and eight α-N-dimethylated peptides. One protein, the gene product of *ARB1* (P_1-P_2-V_3) had evidence of multiple N-terminal peptide methylation states (Table S5). Arb1 was the only protein previously reported to be α-N-monomethylated in the associated publication from this dataset and the detection of methylation of Bdh2 (R_1-A_2-L_3) is novel.[40] Notably, none of the peptides contained the canonical recognition motif. Arb1 has an N-terminal sequence of P_1-P_2-V_3, which has similarity to the P_1-P_2-K_3 canonical motif but is missing the crucial lysine residue at the third position.

The MISL dataset was generated using a strategy called Methylation by Isotope Labeled SAM (MISL). Briefly, *S. cerevisiae* (*sam1Δ* and *sam2Δ* BY4741 background) SAM deficient cells were metabolically labeled with either heavy CD_3-SAM or light CH_3-SAM, and subsequent lysates were mixed at 1:1 ratio and digested with trypsin before MS/MS analysis. The original study only searched for methylation on amino acid side chains and did not investigate α-N-terminal methylation. This data was repurposed by searching for light α-N-methylation species and heavy α-N-methylation species with the same search parameters as in the original study. We focused on heavy labeled methylation because it easily distinguishes trimethylation from acetylation and demonstrates *in vivo* methylation. 49 unique heavy N-terminal methylated peptides (iMet cleaved) were found for 48 unique proteins, including 23 α-N-monomethylated peptides, 13 α-N-dimethylated peptides and 13 α-N-trimethylated peptides. Several proteins were found as multiple methylation species or in multiple datasets. Ahp1, with the S_1-D_2-L_3 amino acid sequence, was found to be α-N-monomethylated and α-N-trimethylated with heavy label and α-N-trimethylated with the light label. Rpn13 (S_1-M_2-S_3) was α-N-trimethylated with heavy label and α-N-dimethylated with the light label. Asc1 (A_1-S_2-N_3) was found to be α-N-trimethylation in both light and heavy label. Only one protein, Rps25a/b (P_1-P_2-K_3) contained the canonical motif and was shown to be heavy α-N-monomethylated. Interestingly, four more proteins containing the canonical motif were shown to be heavy α-N-methylated with low peptide scores, including Rpl12ab, Rpt1, Ola1 and Hsp31. Rpl12ab and Rpt1 were previously reported to be α-N-dimethylated.[4,5] The use of trypsin in this study likely leads to suboptimal detection of canonical motif-containing peptides as discussed earlier.

Another large proteomic data set was investigated that was generated using a negative N-terminal enrichment method. Four different individual sample preparations depending on the combination of endoprotease and chemical blocking reagent. Briefly, the free α- and ε-amino groups in HEK293T cell lysate were blocked (by propionylation or D3-acetylation), including protein α-amino group and amino groups on lysine side chains. The proteomic sample was then subjected to enzymatic digestion by GluC or Trypsin. Internal peptides with the free amino group were removed by NHS activated agarose and thus protected protein N-termini were negatively selected. Results of this search are listed in the Table S5. The majority of the hits identified did not contain the canonical motif while five unique protein hits contained the canonical X-P-K motif (Table S6). Two of these hits have previously been demonstrated to be methylated, SET (A_1-P_2-K_3) was reported to be α-N-terminally trimethylated and RPL23A (A_1-P_2-K_3) has previously been demonstrated to be α-N-terminally dimethylated and trimethylated.[8] In addition, four other protein hits containing the canonical motif were identified that had corresponding N-terminal methylated peptides with scores near the cutoff.

Overall, we screened for potential N-terminal methylation events by searching and parsing three yeast datasets and two human datasets generated by various techniques. We identified three yeast proteins and five human proteins containing known α-N-terminal recognition motifs (X_1-P_2-$[K/R]_3$ or G_1-K_2-E_3-K_4), in which all the yeast proteins and two human proteins were verified in earlier studies. This result indicates the efficacy of the repurposing method in identifying potential α-N-methylation hits with canonical motifs and yields a large set of proteins without the canonical motif. These results suggest a greater prevalence of α-N-terminal methylation in the yeast and human proteomes than previously appreciated. However, careful verification is needed to confirm potential hits due to possible false localization of a methyl group on lysine/arginine side chains as opposed to the N-terminus or because of ambiguity between isobaric modifications.

Sequence analysis of methylated peptides shows semi-specific conservation at the first position of the α-N-terminal methylome

The identification of numerous peptides in multiple datasets allowed the investigation of global conserved pattern of α-N-terminal methylation events. We applied both WebLogo[35] and iceLogo[48] to analyze the sequence conserveness of α-N-methylated proteins in yeast and humans (Figure 1, Figure 2 and Figure S2).

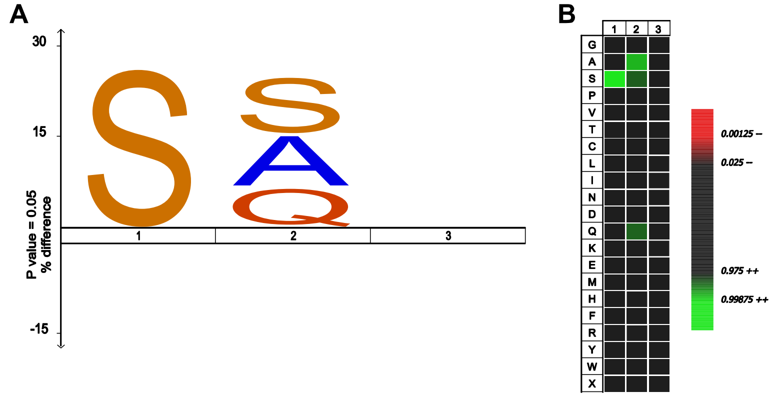


Figure 1. N-terminal amino acid conservation for methylated peptides in the yeast datasets.
A. IceLogo were generated by iceLogo web application. Over-presented amino acids are show above center line and under-represented amino acids are below the center line. A S_1-$[S/A/Q]_2$ pattern is clearly enriched in yeast dataset.
B. Heatmap demonstrates over-represented amino acids and under-represented amino acids colored green and red, respectively.

To avoid over-interpretation from the sequence analysis and the uncertainty introduced by isobaric modifications (especially the prevalent α-N-terminal acetylation), unlabeled trimethylated protein hits are excluded from the logo analysis. For yeast datasets, we only analyzed the heavy labeled hits from iProx yeast subset and MISL datasets to determine sequence pattern in yeast. Hits from ChaFRADIC are excluded from logo analysis. We used the heavy labeled methylation hits from the iProx Hela subset and unlabeled monomethylated or dimethylated hits from the Chemical Labeling datasets for the human proteome logo analysis (Table S7). Our subsequent investigations focused on the list of nonredundant peptides with iMet removed and other logo visualizations can be found in supporting information (Figure S2).

We combined a total of 47 nonredundant peptides (iMet cleaved) identified from the heavy labeled MISL yeast dataset and iProx yeast datasets for yeast logo analyses (Table S7). IceLogo visualization suggests that the first two amino acid positions might be more determinant in α-N-terminal methylation events. Serine is significantly enriched at the 1st position while the second position appears to have a more diverse preference for serine, alanine and glutamine (Figure 1A and 1B). The 3rd, 4th, and 5th positions are not conserved in the nonredundant and likely do not factor in influencing non-canonical N-terminal methylation (data not shown after the 3rd position). No amino acid is under-represented. This result reveals a pattern of $[S]_1$-$[S/A/Q]_2$ for N-terminal methylation events in the yeast proteome.

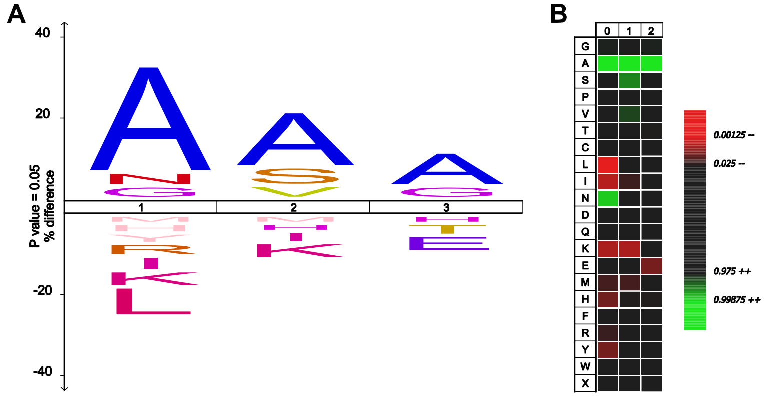


Figure 2. N-terminal amino acid conservation for methylated peptides in the human datasets.
A. IceLogo generated by iceLogo web application as in Figure 1. Over-represented amino acids are show above center line and under-represented amino acids are below center line. A $[A/N/G]_1$-$[A/S/V]_2$-$[A/G]_3$ pattern is clearly enriched in the human dataset.
B. Heatmap demonstrates over-represented amino acids and under-represented amino acids colored green and red, respectively.

In the human proteome methylation data, we analyzed a list of 222 nonredundant peptides (iMet cleaved) from the heavy labeled iProx human dataset and from the Chemical Labeling dataset. Similar to the yeast IceLogo, the first three amino acids are more relevant to the α-N-terminal methylation events in humans. IceLogo analysis demonstrates that alanine is significantly enriched at the 1st position along with glycine and asparagine. In addition, several amino acids are significantly under-represented at the first three positions, which indicates that they are less favored for the α-N-terminal methylation. A pattern of $[A/N/G]_1$-$[A/S/V]_2$-$[A/G]_3$ was indicated for α-N-terminal methylation events in the human proteome (Figure 2A and 2B).

Evidence for the NTMT1/2 canonical sequence which prefers proline at the second position and lysine at the third position in human global N-terminal methylome was not detected. Eight protein hits containing the canonical motif were identified with peptide scores greater than 20, and an additional eight proteins with canonical motifs were identified with less confidence (Supporting information Table S6). However, these were not sufficient to influence the overall methylation pattern uncovered in our approach. Methylated Rpl12ab peptides were detected in ChaFRADIC and MISL dataset with high score and low score, respectively. Thus, in total there are 15 nonredundant protein hits containing canonical motifs from all four datasets. Nine out of 15 were identified in the HEK cell line, and 6 were from the yeast datasets. 6 out of the total 15 protein hits have been verified in earlier literature, including Rpl12ab, Rps25a/b, eEF1 and Rpt1 in yeast and Rpl23a, SET in humans[4]. Overall, our motif analysis suggests the presence of an additional methylation process that is not restricted to the X_1-P_2-[K/R]$_3$ motif and is mostly dependent on the first amino acid for semi-specificity. The NTMT1/2 enzymes are likely not responsible for this non-canonical methylation, because extensive peptide specificity studies show strong specificity for the X_1-P_2-[K/R]$_3$ motif or related motif.[4,8]

Comparison of N-terminal methylation with N-terminal acetylated proteins

Two prevalent modifications on the protein α-N-termini are α-N-terminal acetylation and α-N-terminal methylation. There is evidence that the competition between these two modifications on the same protein and regulate its dual localization, such as MYL9.[14] Here we investigate the overlap of α-N-terminal methylation events and α-N-terminal acetylation events. By integrating our α-N-terminal methylation repurposing results with α-N-terminal acetylation reported from the original papers, we classified the identified protein hits into four categories based on their N-termini modification states: only α-N-terminal acetylated, only α-N-terminal methylated, dual modification or not modified by either of the two. Each category was counted, and the frequency of each category was plotted in a pie chart (Figure 3A). The proportion for each class indicates the distribution of modifications for the proteome. To further explore the extent of both modifications in the global proteome, we plotted the frequency of each category over the 1st amino acid type of the protein hits. The distribution plots reveal that subsets of proteins with some types of amino acids might be under dual control or prefer either of the two modifications. (Figure 3B and 3C).

We utilized the nonredundant proteins from our pattern analysis for analyzing the distribution of modifications. Analysis that included protein hits modified on iMet (which consist of more than 40% of total identified protein hits) showed that methionine did not change the overall distribution of α-N-terminal modifications in the proteome, but iMet is predominantly modified by both methylation and acetylation (Figure S3). Both yeast and humans datasets had 60%-70% of protein hits that were α-N-terminal acetylated, consistent with other estimates of N-terminal acetylated proteins.[49,50] The α-N-terminal methylated protein hits consist of 7-8% of the total proteome in both yeast and humans. Only 2% out of the 7-8% methylated protein hits could be under dual control of α-N-terminal methylation and α-N-terminal acetylation. In addition, around 21% and 35% proteome for humans and yeast, respectively, did not have evidence of either modification. The proportion of α-N-terminal acetylation is approximately 10 times as much as α-N-terminal methylation (Figure 3A). The high prevalence of α-N-terminal acetylation could be related to the close association of acetyltransferases with ribosomes.[49] There is no current evidence that α-N-terminal methyltransferases are in close spatial proximity with ribosomes.

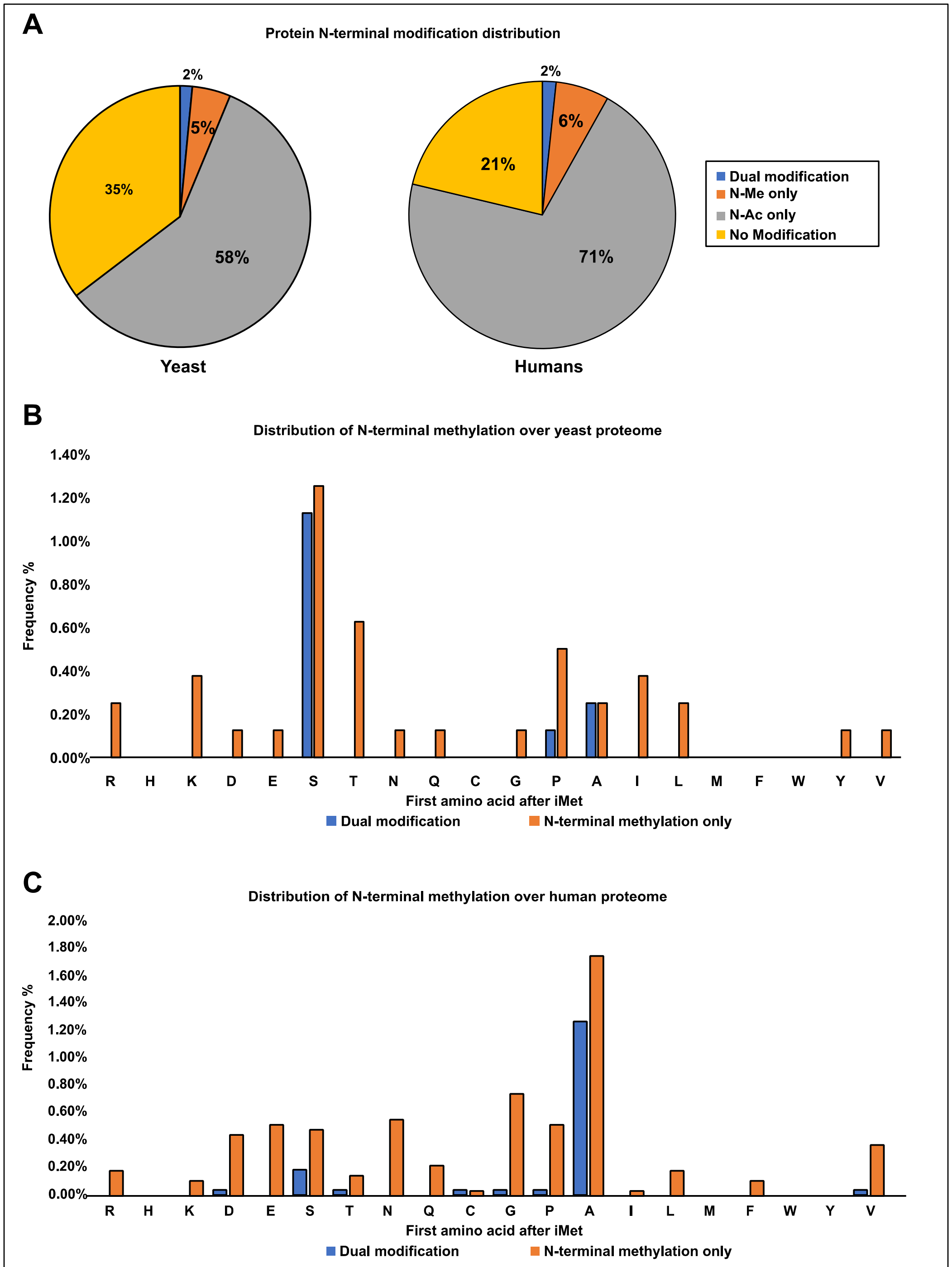


Figure 3. The frequency of proteins found to have only α-N-terminal methylation or both modifications are plotted according to the type of amino acid. Only iMet cleaved proteins are used for this graph. (A) Pie charts shows the proportion of yeast or human proteome undergo α-N-terminal methylation or α-N-terminal acetylation. (B) The frequency plot for yeast datasets, combining hits from MISL and iProx datasets. (C) The frequency plot for human datasets, combining hits from Chemical Labeling and iProx datasets. The frequency of N-terminal acetylation only was not depicted in the frequency plots because high levels of acetylation for several amino acids results in a scale that obscures the visualization of lower frequency methylation levels. The frequency plots including iMet retained proteins are in Figure S3.

We examined each modification's distribution and preference on the 1st amino acid for both yeast and humans. The identified proteins with various α-N-terminal modification states were clustered by the 1st amino acid, and the fraction of each α-N-terminal modification states (only α-N-terminal methylated, only α-N-terminal acetylated, dual modification or not modification) was determined by dividing the modified population of each cluster against the total identified protein population.

Yeast α-N-methylation was detected on all amino acids with less frequency for H, C, M, F and W. As expected, methylation was detected on the amino acids A, S and P as in the Tae1/NTMT canonical motif and amino acids with methylatable side chains such as R and K. Unexpectedly, methylation was also detected on D, E, N, and Q which are rarely reported to be N-methylated. The dual modification was observed predominantly on A and S (Figure 3B). In the human proteome, all amino acids are methylated except H, M, W and Y. The A and G constitute the majority of the N-terminal methylation detected. The dual-modification types were observed predominantly on A and S. Interestingly, although N-terminal proline is thought to block acetylation,[51] we note that proline acetylation was identified in the original paper and an additional recent report has demonstrated proline acetylation (Figure 3C).[52]

To summarize, in both human and yeast datasets, around 7-8% of proteins are subject to α-N-methylation, which is one-tenth of the proportion of N-terminal acetylation. N-methylation is found to be more prevalent in both species and not strictly conserved on the initial amino acid type. Both charged amino acids (D, N, Q, E, K and R) or hydrophobic amino acids (L, V, A and G) could be α-N-methylated. A and S at protein N-termini are the most frequent for α-N-methylation, while the combination of α-N-methylation and α-N-acetylation is also prevalent. Yeast and human show similar pattern in the distribution of α-N-methylation and α-N-acetylation.

Ssa3 and Vma1 are predominantly α-N-terminal acetylated instead of methylated

PTM events identified from a proteomic dataset maybe not accurate and difficult to locate because of peptide misassignment and limited fragmentation detection. Thus, verification is necessary for ensuring α-N-terminal methylation is correctly identified from the repurposed datasets because of the closely matched mass of trimethylation and acetylation modifications. We demonstrate that initial identification of trimethylation for some peptides could in fact be acetylation but also validated a novel methylation of a protein that was not previously reported. We attempted to verify three hits detected using our repurposed datasets; two yeast protein hits without the canonical motif (Ssa3 and Vma1) and one protein with a canonical motif were purified (Hsp31). To verify the results from proteomic dataset searching and semi-quantify the ratio between different N-terminal PTMs, we purified the proteins using the Dharmacon yeast ORF collection. The C-terminally tagged proteins of interest were overexpressed using the *GAL* promoter. Subsequently, the overexpressed protein was affinity-purified with Protein A agarose beads. All proteins were purified to greater than 95% purity as assessed by SDS-PAGE (Figure 4A) and examined by western blot (Figure S4). Ssa3 is a member of stress-inducible

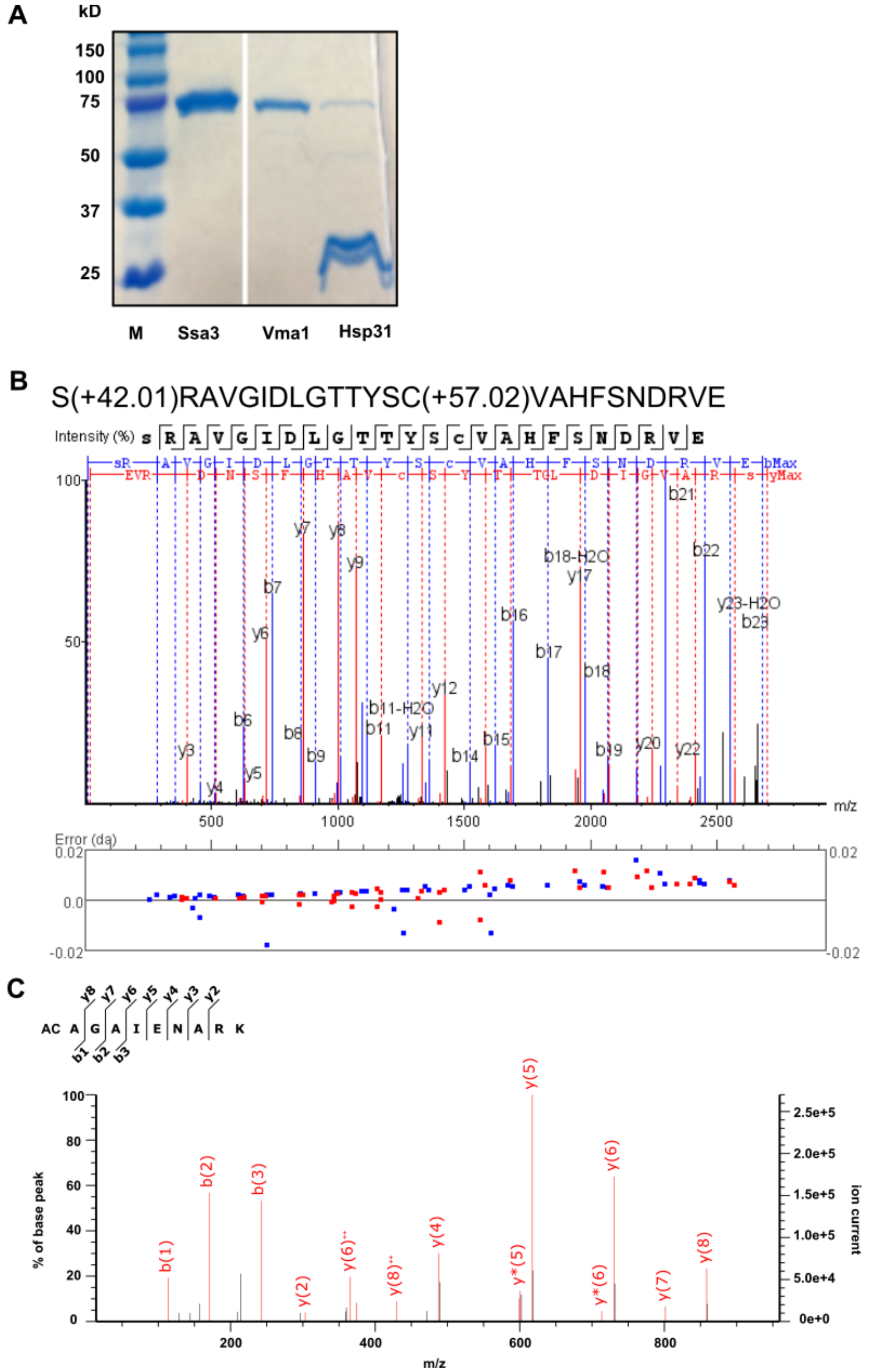


Figure 4. Verification of N-terminal modifications potentially methylated substrates. (A) Vma1-MORF, Ssa3-MORF and Hsp31-MORF were purified by protein A affinity purification and released from the beads using protease 3C. Purified proteins were analyzed by Coomassie blue staining. (B) Fragmentation of Ssa3 confirmed N-terminal acetylation on Ssa3 Serine_1. Error is within 0.02 Da range. (C) MS2 spectrum indicates that Vma1 is acetylated on Alanine_1.

member of the heat shock protein 70 family. The Ssa3 band matches the predicted molecular weight and the Vma1 band matches the molecular weight predicted after intein splicing (Figure 4; Figure S5 and S6).[53] The Vma1 intein splicing was verified based on the protein coverage map (Figure S7). Ssa3 was digested by GluC, while Vma1 was digested by trypsin in solution followed by intact protein analysis and tandem MS.

Analysis of the purified Ssa3 by intact mass spectrometry suggested a major mass shift (+40.75Da) corresponding to trimethylation (+42.047Da) or acetylation (+42.0311Da) (Figure S7). Furthermore, the tandem MS analysis using an Orbitrap Fusion Lumos Tribrid mass spectrometer (ThermoFisher) confirmed that the N-terminal peptides were fully acetylated instead of trimethylated as predicted with a protein sequence coverage of 99% (Figure 4B & Figure S7). Analysis of the b ions and y ions collected demonstrated the acetylation on the N-terminal serine. We performed both database searches and de novo searches with Mascot Daemon (MATRIX SCIENCE), PEAKS X plus (Bioinformatics Solutions Inc.) and BioPharma Finder software (ThermoFisher). In PEAKS search with 10ppm tolerance on the precursor and 0.02 Da fragment error, 309 peptides were matched to Ssa3 and cover 99% of the protein sequence. 95 peptide-spectrum matches (PSM) of Ssa3 were found to be exclusively acetylated at α-N-terminus, consistent with the major peak detected using intact mass spectrometry (Figure S7). All the PSM were +3 charge. The finding that Ssa3 is acetylated rather than trimethylated using Mascot is probably a result of utilizing the more sensitive Lumos instrument at higher resolution (60K) than that used in the ChaFRADIC study. Interestingly, Ssa1 was also one of the top hits indicating it is a copurifying protein, and the peptides of Ssa1 also were identified to be α-N-terminal acetylated. Ssa1 is predicted to be acetylated, and mutations preventing N-terminal acetylation appear to decrease the ability of binding to prions.[54] Although Ssa1 is assumed to be acetylated in the literature, the acetylation evidence is indirect, whereas our MS data is the first direct evidence confirming that heat shock protein 70s are N-terminal acetylated.[54,55] The purified Vma1 has 55% protein coverage and 21 N-terminal peptides detected. All N-terminal peptides were α-N-terminal acetylated and had PSMs with +2 charge. The NatA dependent acetylation of Vma1 is consistent with earlier reports and annotations in PTM databases.[56] Although the repurposed MS dataset suggested that Vma1 is α-N-monomethylated, our MS analysis of Vma1 did not detect α-N-monomethylation possibly due to the overexpression strategy employed to purify the protein or the low representation of Vma1 methylation under physiological conditions. Vma1 is predominantly acetylated, but methylation could occur at a low frequency because Vma1 monomethylation was detected in two separate repurposed datasets. These results demonstrate the importance of instrument sensitivity and approach in the verification of these modifications. Further comprehensive studies are needed to examine and verify methylation frequencies and levels but are beyond the scope of this study.

Identification of methylation of the canonical motif-containing protein, Hsp31

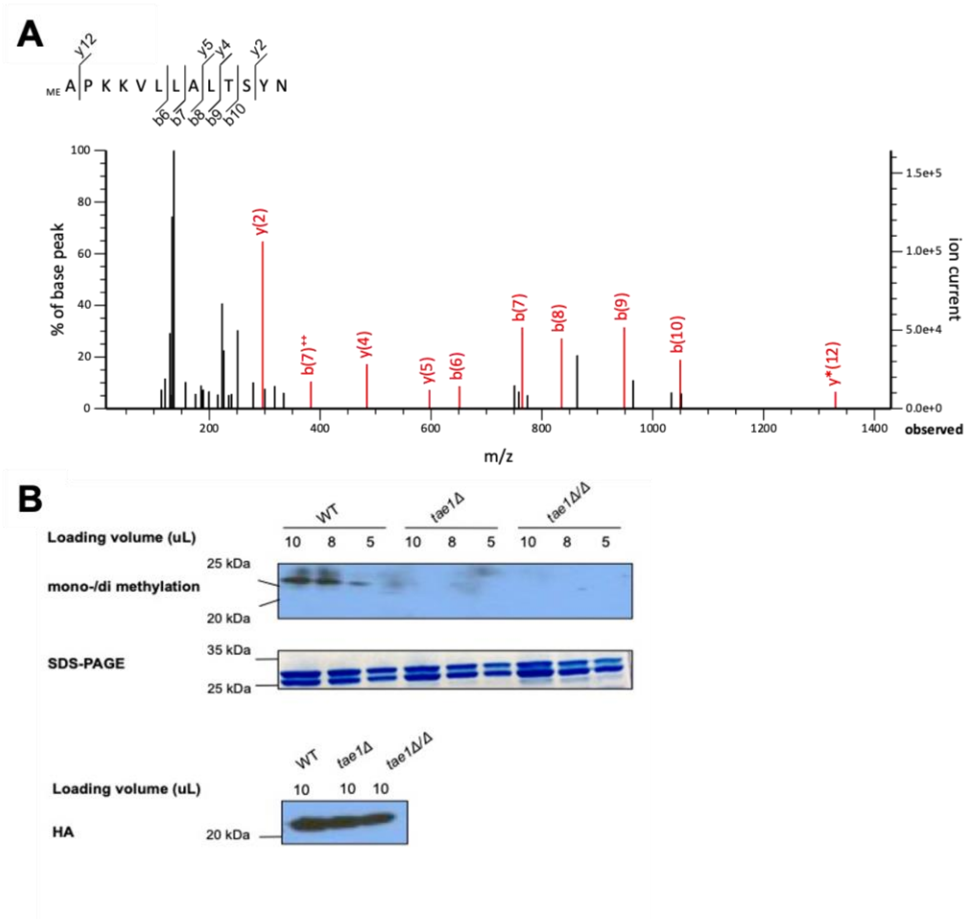


Figure 5. Hsp31 is N-terminally methylated. A) MS2 spectrum consistent with monomethylation. Mascot score for this PSM is 44. An additional monomethylation MS2 spectra is depicted in Figure S8. (B) A polyclonal antibody recognizing mono and dimethylated N-termini was used to probe Hsp31 purified from WT, haploid *tae1Δ* and diploid *tae1Δ*/ *tae1Δ* yeast strains. The same samples were stained with coomassie blue after separate SDS-PAGE to show purity – an extra band is evident due to proteolysis during purification. Samples were also probed with anti-HA antibody which is a tag retained on the C-terminus after purification.

We then investigated another potential substrate consistent with the canonical α-N-terminal motif. Hsp31 was suggested to be α-N-methylated in the MISL dataset, although the peptide score was below the score cutoff of 20. We predicted that Tae1 is likely responsible for methylation of Hsp31. Here, we determined the N-terminal modification on Hsp31 by tandem mass spectrometry and immunodetection. Hsp31 was purified from the yeast MORF collection as described earlier (Figure S4). For tandem mass spectrometry, purified Hsp31 was digested with AspN for MS/MS analysis. Protein coverage was 75%, and MS fragmentation was consistent with α-monomethylation (Figure 5A and Figure S8). 16 PSM corresponded to the Hsp31 N-terminus in total. Two PSM were consistent with monomethylated Hsp31 N-termini. The y12 ion in 1 PSM indicates the presence of monomethylation on Ala_1 site. Five PSM were consistent with either N-terminal dimethylation or formylation, but MS2 spectra matched formylation with lower ppm error (~3.6 ppm for formylation and ~21.6 for dimethylation. All the N-terminal PSM had +2 charge or +3 charge, which could be explained due to the presence of multiple lysines. Hsp31 purified from wild type BY4741 yeast was shown to be methylated by a polyclonal antibody designed to recognize dimethylated SPK N-terminal sequences (a gift from Dr. Schaner-Tooley)[38]. This antibody also has cross-reactivity for mono-methylated N-termini and similar sequences such as A_1-P_2-K_3 based on synthetic peptide dot blot assays (data not shown). The loss of methylation was observed in Hsp31 purified from both haploid *tae1Δ* and diploid *tae1Δ/ tae1Δ* yeast (Figure 5B). This data further supports monomethylation on the Hsp31 N-terminus, albeit at a low level under these culturing and expression conditions. The lack of signal in the *TAE1*-deficient yeast strains indicates that methylation is contributed mainly by Tae1. In conclusion, we were able to verify novel N-terminal methylation of a protein with a canonical motif demonstrating the repurposing approach's validity.

Discussion

Protein N-terminal methylation is a novel protein modification type that has garnered increasing attention recently. Two classes of α-N-methyltransferases found in yeast and humans recognize distinct protein N-terminal sequence patterns. Yeast Tae1 and the human orthologs, NTMT1/NTMT2, recognize a canonical X_1-P_2-[K/R]$_3$ motif on substrate proteins and contribute to most of the N-terminal methylation events reported to date. Another yeast N-terminal methyltransferase, Nnt1, is a more specific enzyme that appears only to have a sole target, Tef1.[46] Tef1 contains a G_1-K_2-E_3-K_4 N-terminal sequence and does not share similarity with the X_1-P_2-[K/R]$_3$ motif. The functional, but not the sequence ortholog, of Nnt1 in humans appears to be METTL13 which also recognizes a similar sequence and methylates the orthologous substrate, eEF1a.[16,19] Besides these two classes of N-terminal methyltransferases, the detection of N-terminal methylation events usually relies on assuming that the canonical motif sequence has the potential to be methylated, and proteins without the motif are unmethylated and hence excludes the potential of identifying non-canonical methylation events. An *in vitro* ^{3}H-methylation assay with 9-mer synthetic N-terminal peptides suggests the X_1-P_2-[K/R]$_3$ is the most favorable substrate for yeast and human N-terminal methyltransferases but also demonstrates varying amino acid preferences.[4] There is an urgent need to develop an unbiased method that can screen potential protein N-terminal methylation events regardless of the N-terminal sequence. Repurposing current datasets verify previously characterized protein modifications at a proteomic level and enables discovery of potential N-terminal methylation events in an unbiased way.

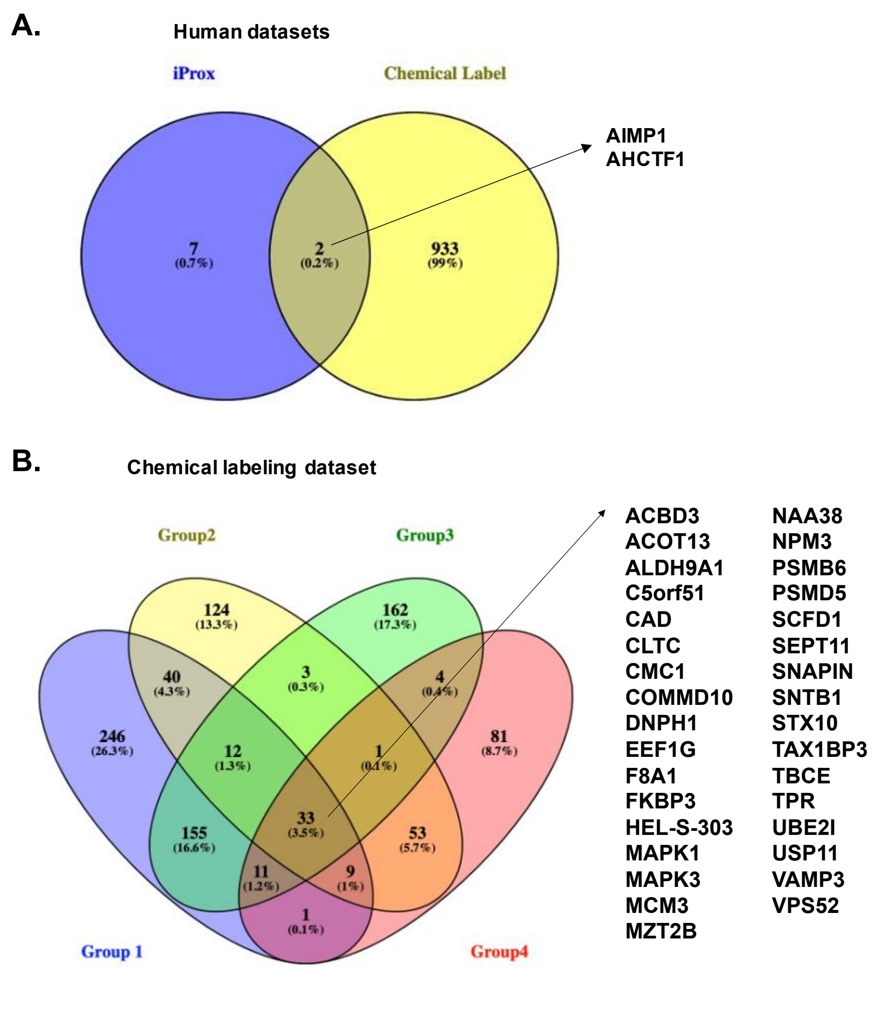


Figure 6. Venn diagram demonstrating the number of potential methylated proteins and the overlap from human datasets. (A) The human datasets yielded 2 overlapping proteins. (B) The Chemical Labeling subsets yielded 33 overlapping proteins present in all four subsets and 289 proteins that appeared in two or more subsets. Yeast datasets don’t have reproduced hit.

In this study, we expanded our knowledge of the extent of the N-terminal methylome in yeast cells and human cell lines. Only a small subset of the potential hits contains canonical methylation motif. In the ChaFRADIC dataset, we identified two known α-N-methylated substrates, Rpl12ab (P_1-P_2-K_3) and Tef1 (G_1-K_2-E_3-K_4). In the MISL dataset, we identified one known substrate Rps25ab (P_1-P_2-K_3) with peptide score >20 and two known substrates with lesser scores, Rpl12ab (P_1-P_2-K_3) and Rpt1 (P_1-P_2-K_3).[4,5] We also identified two low score hits with the canonical motif but have not been demonstrated to be methylated, which are Hsp31 (A_1-P_2-K_3) and Ola1 (P_1-P_2-K_3). We confirmed α-N-monomethylation and α-N-dimethylation of Hsp31 by mass spectrometry. From the Chemical Labeling dataset, we identified two known substrates, Rpl23a (A_1-P_2-K_3) and SET (A_1-P_2-K_3) with peptide score >20. Three more protein hits with the canonical motif were identified that had not been previously reported, which are CMTM2 (A_1-P_2-K_3), XPNPEP1 (P_1-P_2-K_3) and ZC3H15 (P_1-P_2-K_3). In addition, we found 4 other protein hits with lower score, RGS11 (P_1-P_2-R_3), PCSK6 (P_1-P_2-R_3), MORF4L1 (A_1-P_2-K_3) and TOMM34 (A_1-P_2-K_3) (Table S6). There are 45 yeast proteins with the canonical motif in the yeast proteome and many were not identified by our repurposing approach demonstrating that we are under-sampling the N-terminome (Table S1). The remaining hits identified by our study appeared to have moderate conservation in the first position but minimal conservation on adjacent amino acid positions. Only Arb1 contains a related sequence, P_1-P_2-V_3, while the majority contain unrelated sequences. This indicates N-terminal methylation might be more widely occurring and is not restricted to the X_1-P_2-$[K/R]_3$ motif-containing proteins. We found serine in yeast and alanine/serine/asparagine in humans are over-represented compared to proteome background at the first position, indicating that proteins with these amino acids are the most methylated in humans and yeast, respectively.[49] We also note that some amino acids are under-represented, such as Met/His/Trp/Arg/Ile/Lys/Leu at the first position after iMet. These N-terminal peptides might be less favored, or they are not optimal for MS detection because of peptide length or low representation.

It should be noted that protein hits from the repurposing result should be carefully investigated and further confirmed. Our attempt to validate two proteins with non-canonical sequences indicated that they were predominantly acetylated. In the case of Ssa3, the repurposing approach initially indicated trimethylation of the protein, but our experiment with a high-resolution Orbitrap Fusion Lumos Tribrid mass spectrometer demonstrates that the protein is α-N-acetylated. In addition, Vma1 was detected as monomethylated in two different isotopic labeling datasets, but our experiment indicates that it was predominantly acetylated. We posit that Vma1 may be monomethylated at a nominal level relative to acetylation, and hence it is difficult to identify these low frequency methylated peptides. The lack of prevalent α-N-terminal methylation in these two proteins with non-canonical motifs implies that α-N-terminal methylation is a relatively rare modification for these proteins. Due to possible ambiguity of identifying α-N-terminal methylation from α-N-terminal acetylation as shown in Ssa3, we excluded any nonlabelled trimethylation hits from our downstream bioinformatics analysis for the sake of accuracy. Future proteomic assessment of α-N-terminal methylation representation and function would be enhanced by combining N-terminal enrichment techniques with isotopic labeling methods or depleting the α-N-terminal acetylated proteins.

The source of the non-canonical methylation events is unclear. N-terminal methyltransferases that prefer the canonical motifs may be responsible for a rare level of background methylation of proteins with non-canonical motifs. Another possibility is that an additional methyltransferase enzyme is performing this activity. In prokaryotes, the function of N-terminal methylation is found to be relatively promiscuous and does not follow strong canonical motif recognition.[2] PrmA is the nonessential N-terminal methyltransferase in both *E.coli* and *T. thermophilus* and is reported to modify three amino acids on Rpl11 including N-terminal alanine, Lys3 and Lys39 while Rpl11.[57–59] Notably, there is minimal sequence similarity between PrmA and the eukaryotic N-terminal transferases, NTMT1/2 and Tae1.[4] Our studies raise the intriguing possibility that a functionally corresponding eukaryotic methyltransferase contributes to these general methylation events detected by our study. The source of these methylation events could be revealed by proteomic studies of genetically modified cell lines or yeast strains (such as methyltransferase gene deletions or functionally compromised strains). We expect non-canonical methylation is enzymatically driven because various isotopic labels resulted in detectable methylation, but another possibility is that nonenzymatic methylation of proteins can occur by reaction with SAM.[60] We also note that bioinformatic analysis of non-canonical motif methylated proteins, including GO term analysis did not show enrichment in any category (data not shown). An additional question is if these methylation events have direct biological roles or are possibly present at a very nominal and persistent level or are sporadic events. Although the source of non-canonical methylation is unclear, our studies have shown it to be persistent and widespread across the proteome, indicating that this type of methylation warrants further investigation.

The reproducibility of hits across different samples is low. No protein hits shared by the three yeast datasets and two proteins are reproduced in two human datasets (Figure 6). The low reproducibility in both human and yeast is likely because of the limited number of protein hits, exclusion of the majority of trimethylated hits from bioinformatics analysis and insufficiency in the detection of the α-N-terminome. Besides, up to 2-3 peptide occurrences were identified for the same PTM for most protein hits (Table S5). Also of note is that we observed different peptide cleavage forms that encompassed the same modification which increases the confidence in correct assignment. The low reproducibility across datasets demonstrates a need for the optimized design of experimental approaches and more comprehensive dataset searching strategies to improve α-methylation detection. Within the Chemical Labeling dataset, where four sample preparation methods (two proteases and two chemical labeling methods) were individually used to analyze the α-N-terminal methylome in the same sample, 33 proteins are found in all four subsets and 289 proteins were found in two or more subsets (Figure 6). Moving forward, we demonstrate the utility of building an automated searching platform to repurposing appropriate datasets and provide the path to generating a compendium of potentially α-N-terminal methylated proteins.

To avoid false positive hits and overinterpretation, extra attention should be paid to sample preparation techniques, resolution of mass spectrometry instruments, and the ambiguity caused by α-N-terminal acetylation. In some N-terminal enrichment techniques, dimethylation reagent is used to block the free α-N-terminus so that the free α-N-terminus of the internal peptides can be selectively removed. Only α-N-trimethylation would be meaningful in these datasets, while the α-N-monomethylation and α-N-dimethylation could be chemically formed during sample preparation. The sensitivity of an instrument is also another crucial factor. It is essential that high-resolution mass spectrometry instruments are used that are capable of differentiating α-N-trimethylation from acetylation which have a small mass difference, such as the Q-Exactive or higher resolution LTQ instruments. In addition, instruments may not be run at a high resolution for scanning speed reasons, and thus they cannot accurately differentiate acetylation from trimethylation. Another concern of generally searching for any specific protein modifications at proteome level is the noise signal ratio. Less protein representation, protein contamination, insufficient MS/MS fractionation, and vast number of peptides analyzed might lead to false positive results. To further optimize global N-terminal methylome investigations, two methodologies might yield higher quality and quantity: (i) Combination of SILAC and N-terminal enrichment methods and (ii) Negative selection method removing internal peptides and N-terminal acetylated peptides. Supplementary methods may be necessary to verify potentially modified proteins and accurate modification site assignment, such as probing the protein of interest under either physiological levels or overexpression followed by mass spectrometry.

The vast number of datasets generated for various purposes such as protein identification, modification detection and investigating proteome perturbances are openly accessible on proteome consortiums sites such as ProteomeXchange (http://www.proteomexchange.org)[24] and databases such as iProx (https://www.iprox.org/)[25] and PRIDE (https://www.ebi.ac.uk/pride/)[23]. Reanalysis of public datasets has been reported previously to discover novel protein encoding genes [61], finding missing human proteins [62], identify novel PTMs [63] and reveal PTM sites [64]. One of the major barriers for reusing proteomic datasets is due to the complexity of MS data. Both sample preparation procedure and the mass spectrometry technique used in the datasets need to be closely examined to find suitable ones for repurposing. Efforts to mitigate these factors are underway including integrating experimental metadata and reanalysis tools in a platform such as the Reanalysis of Data User (ReDU) interface which is focused on chemicals and metabolite proteomics data.[65] With abundant MS datasets for probing modified protein N-terminal, repurposing those datasets for N-terminal methylation events will shorten the research cycle and reduce the expense of proteomic studies. By repurposing the current datasets, proteins that are potentially α-N-terminal methylated are generated with confidence depending on the type of instrument, complexity of the sample, protein abundance and the accessibility of N-termini. Notably, no previous reports have examined the N-terminal methylome by employing the reanalysis of public datasets. Of course, it would be necessary to verify α-N-terminal methylation events with purified proteins or orthologous methods to investigate specific proteins further. The establishment of the prevalence of this α-N-terminal modification and building a compendium of methylated proteins is a crucial step to discover the functions of this cryptic modification.

Supporting Information.

Table S1. 45 canonical N-terminal motif containing proteins in yeast.

Table S2. FDR information of each dataset.

Table S3. b/y ions of MS2 spectrum of α-N-methylated peptides.

Table S4. Mascot search results with alternative K/R methylation sites.

Table S5. Protein hits from each repurposed dataset.

Table S6. Protein hits containing canonical motif.

Table S7. List of sequences used for WebLogo and iceLogo analysis.

Figure S1. MS/MS spectra for all hits from three repurposed yeast datasets.

Figure S2. WebLogos and IceLogos for repurposed datasets.

Figure S3. N-Ac and N-Me distribution plots (iMet retained hits included).

Figure S4. Western blot of purified proteins.

Figure S5. Intact mass spectrometry of Vma1.

Figure S6. Protein coverage map of Vma1.

Figure S7. Methods description and MS analysis of Ssa3.

Figure S8. MS2 spectra of monomethylated and formylated Hsp31.

Table S6, Figure S2-8 are integrated into a single Word file. Table S1-5, Table S7, Figure S1 are individual Excel files. Figure S1 is attached as a separate individual PDF file.

Mascot output files for the four repurposed datasets are compressed into “Supporting information raw search output.zip” file.

Acknowledgment

Analysis of Ssa3, Vma1 and Hsp31 was performed at the Indiana University School of Medicine (IUSM) proteomics core, Purdue Proteomics Facility and BGI Americas Corp. We gratefully acknowledge the Purdue University Center for Cancer Research (NIH P30CA023168) for supporting the Purdue Proteomics Facility and Dr. Uma Aryal for assistance with the MS analysis. Work in the IUSM Proteomics Core was supported, in part, with support from the Indiana Clinical and Translational Sciences Institute which is funded by Award Number UL1TR002529 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. We thank Dr. Amber Mosley and Dr. Emma Doud at the IUSM Proteomics Core for assistance with the mass spectrometry analysis. Acquisition of the IUSM Proteomics core instrumentation used for this project, the Orbitrap Fusion Lumos, was provided by the Indiana University Precision Health Initiative.

All data were deposited to the PRIDE Archive (http://www.ebi.ac.uk/pride/archive/) via PRIDE partner repository and are public accessible with the data set identifier PXD022833.

Notes

The authors declare no competing financial interest.

Reference

(1) Stock, A.; Clarke, S.; Clarke, C.; Stock, J. N-Terminal Methylation of Proteins: Structure, Function and Specificity. *FEBS Letters* **1987**, *220* (1), 8–14. https://doi.org/10.1016/0014-5793(87)80866-9.

(2) Huang, R. Chemical Biology of Protein N-Terminal Methyltransferases. *ChemBioChem* **2019**, *20* (8), 976–984. https://doi.org/10.1002/cbic.201800615.

(3) Dong, C.; Mao, Y.; Tempel, W.; Qin, S.; Li, L.; Loppnau, P.; Huang, R.; Min, J. Structural Basis for Substrate Recognition by the Human N-Terminal Methyltransferase 1. *Genes Dev.* **2015**, *29* (22), 2343–2348. https://doi.org/10.1101/gad.270611.115.

(4) Webb, K. J.; Lipson, R. S.; Al-Hadid, Q.; Whitelegge, J. P.; Clarke, S. G. Identification of Protein N-Terminal Methyltransferases in Yeast and Humans. *Biochemistry* **2010**, *49* (25), 5225–5235. https://doi.org/10.1021/bi100428x.

(5) Kimura, Y.; Kurata, Y.; Ishikawa, A.; Okayama, A.; Kamita, M.; Hirano, H. N-Terminal Methylation of Proteasome Subunit Rpt1 in Yeast. *PROTEOMICS 13* (21), 3167–3174. https://doi.org/10.1002/pmic.201300207.

(6) Alamgir, M.; Eroukova, V.; Jessulat, M.; Xu, J.; Golshani, A. Chemical-Genetic Profile Analysis in Yeast Suggests That a Previously Uncharacterized Open Reading Frame, YBR261C, Affects Protein Synthesis. *BMC Genomics* **2008**, *9*, 583. https://doi.org/10.1186/1471-2164-9-583.

(7) Hamey, J. J.; Winter, D. L.; Yagoub, D.; Overall, C. M.; Hart-Smith, G.; Wilkins, M. R. Novel N-Terminal and Lysine Methyltransferases That Target Translation Elongation Factor 1A in Yeast and Human. *Mol Cell Proteomics* **2016**, *15* (1), 164–176. https://doi.org/10.1074/mcp.M115.052449.

(8) Tooley, C. E. S.; Petkowski, J. J.; Muratore-Schroeder, T. L.; Balsbaugh, J. L.; Shabanowitz, J.; Sabat, M.; Minor, W.; Hunt, D. F.; Macara, I. G. NRMT Is an α-N-Methyltransferase That Methylates RCC1 and Retinoblastoma Protein. *Nature* **2010**, *466* (7310), 1125–1128. https://doi.org/10.1038/nature09343.

(9) Bailey, A. O.; Panchenko, T.; Sathyan, K. M.; Petkowski, J. J.; Pai, P.-J.; Bai, D. L.; Russell, D. H.; Macara, I. G.; Shabanowitz, J.; Hunt, D. F.; Black, B. E.; Foltz, D. R. Posttranslational Modification of CENP-A Influences the Conformation of Centromeric Chromatin. *Proc Natl Acad Sci U S A* **2013**, *110* (29), 11827–11832. https://doi.org/10.1073/pnas.1300325110.

(10) Sathyan, K. M.; Fachinetti, D.; Foltz, D. R. α-Amino Trimethylation of CENP-A by NRMT Is Required for Full Recruitment of the Centromere. *Nat Commun* **2017**, *8*. https://doi.org/10.1038/ncomms14678.

(11) Dai, X.; Otake, K.; You, C.; Cai, Q.; Wang, Z.; Masumoto, H.; Wang, Y. Identification of Novel α-n-Methylation of CENP-B That Regulates Its Binding to the Centromeric DNA. *J. Proteome Res.* **2013**, *12* (9), 4167–4175. https://doi.org/10.1021/pr400498y.

(12) Cai, Q.; Fu, L.; Wang, Z.; Gan, N.; Dai, X.; Wang, Y. α-N-Methylation of Damaged DNA-Binding Protein 2 (DDB2) and Its Function in Nucleotide Excision Repair. *J Biol Chem* **2014**, *289* (23), 16046–16056. https://doi.org/10.1074/jbc.M114.558510.

(13) Dai, X.; Rulten, S. L.; You, C.; Caldecott, K. W.; Wang, Y. Identification and Functional Characterizations of N-Terminal α-N-Methylation and Phosphorylation of Serine 461 in Human Poly(ADP-Ribose) Polymerase 3. *J Proteome Res* **2015**, *14* (6), 2575–2582. https://doi.org/10.1021/acs.jproteome.5b00126.

(14) Nevitt, C.; Tooley, J. G.; Schaner Tooley, C. E. N-Terminal Acetylation and Methylation Differentially Affect the Function of MYL9. *Biochemical Journal* **2018**, *475* (20), 3201–3219. https://doi.org/10.1042/BCJ20180638.

(15) Jia, K.; Huang, G.; Wu, W.; Shrestha, R.; Wu, B.; Xiong, Y.; Li, P. In Vivo Methylation of OLA1 Revealed by Activity-Based Target Profiling of NTMT1. *Chemical Science* **2019**, *10* (35), 8094–8099. https://doi.org/10.1039/C9SC02550B.

(16) Liu, S.; Hausmann, S.; Carlson, S. M.; Fuentes, M. E.; Francis, J. W.; Pillai, R.; Lofgren, S. M.; Hulea, L.; Tandoc, K.; Lu, J.; Li, A.; Nguyen, N. D.; Caporicci, M.; Kim, M. P.; Maitra, A.; Wang, H.; Wistuba, I. I.; Porco, J. A.; Bassik, M. C.; Elias, J. E.; Song, J.; Topisirovic, I.; Van Rechem, C.; Mazur, P. K.; Gozani, O. METTL13 Methylation of EEF1A Increases Translational Output to Promote Tumorigenesis. *Cell* **2019**, *176* (3), 491-504.e21. https://doi.org/10.1016/j.cell.2018.11.038.

(17) Bade, D.; Cai, Q.; Li, L.; Yu, K.; Dai, X.; Miao, W.; Wang, Y. Modulation of N-Terminal Methyltransferase 1 by an N6-Methyladenosine-Based Epitranscriptomic Mechanism. *Biochemical and Biophysical Research Communications* **2021**, *546*, 54–58. https://doi.org/10.1016/j.bbrc.2021.01.088.

(18) Hao, Y.; Macara, I. G. Regulation of Chromatin Binding by a Conformational Switch in the Tail of the Ran Exchange Factor RCC1. *J Cell Biol* **2008**, *182* (5), 827–836. https://doi.org/10.1083/jcb.200803110.

(19) Jakobsson, M. E.; Małecki, J. M.; Halabelian, L.; Nilges, B. S.; Pinto, R.; Kudithipudi, S.; Munk, S.; Davydova, E.; Zuhairi, F. R.; Arrowsmith, C. H.; Jeltsch, A.; Leidel, S. A.; Olsen, J. V.; Falnes, P. Ø. The Dual Methyltransferase METTL13 Targets N Terminus and Lys55 of EEF1A and Modulates Codon-Specific Translation Rates. *Nature Communications* **2018**, *9* (1), 3411. https://doi.org/10.1038/s41467-018-05646-y.

(20) Kleifeld, O.; Doucet, A.; Prudova, A.; auf dem Keller, U.; Gioia, M.; Kizhakkedathu, J. N.; Overall, C. M. Identifying and Quantifying Proteolytic Events and the Natural N Terminome by Terminal Amine Isotopic Labeling of Substrates. *Nature Protocols* **2011**, *6* (10), 1578–1611. https://doi.org/10.1038/nprot.2011.382.

(21) Staes, A.; Van Damme, P.; Helsens, K.; Demol, H.; Vandekerckhove, J.; Gevaert, K. Improved Recovery of Proteome-Informative, Protein N-Terminal Peptides by Combined Fractional Diagonal Chromatography (COFRADIC). *PROTEOMICS* **2008**, *8* (7), 1362–1370. https://doi.org/10.1002/pmic.200700950.

(22) Venne, A. S.; Vögtle, F.-N.; Meisinger, C.; Sickmann, A.; Zahedi, R. P. Novel Highly Sensitive, Specific, and Straightforward Strategy for Comprehensive N-Terminal Proteomics Reveals Unknown Substrates of the Mitochondrial Peptidase Icp55. *J. Proteome Res.* **2013**, *12* (9), 3823–3830. https://doi.org/10.1021/pr400435d.

(23) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yılmaz, Ş.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Res* **2019**, *47* (D1), D442–D450. https://doi.org/10.1093/nar/gky1106.

(24) Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; Moritz, R. L.; Carver, J. J.; Wang, M.; Ishihama, Y.; Bandeira, N.; Hermjakob, H.; Vizcaíno, J. A. The ProteomeXchange Consortium in 2017: Supporting the Cultural Change in Proteomics Public Data Deposition. *Nucleic Acids Res* **2017**, *45* (Database issue), D1100–D1106. https://doi.org/10.1093/nar/gkw936.

(25) Ma, J.; Chen, T.; Wu, S.; Yang, C.; Bai, M.; Shu, K.; Li, K.; Zhang, G.; Jin, Z.; He, F.; Hermjakob, H.; Zhu, Y. IProX: An Integrated Proteome Resource. *Nucleic Acids Res* **2019**, *47* (D1), D1211–D1217. https://doi.org/10.1093/nar/gky869.

(26) Shteynberg, D. D.; Deutsch, E. W.; Campbell, D. S.; Hoopmann, M. R.; Kusebauch, U.; Lee, D.; Mendoza, L.; Midha, M. K.; Sun, Z.; Whetton, A. D.; Moritz, R. L. PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. *J Proteome Res* **2019**, *18* (12), 4262–4272. https://doi.org/10.1021/acs.jproteome.9b00205.

(27) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a Standardized Data Processing Pipeline for Large-Scale Reproducible Proteomics Informatics. *Prot. Clin. Appl.* **2015**, *9* (7–8), 745–754. https://doi.org/10.1002/prca.201400164.

(28) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker Enables Reanalysis of MS-Derived Proteomics Data Sets. *Nature Biotechnology* **2015**, *33* (1), 22–24. https://doi.org/10.1038/nbt.3109.

(29) Gelperin, D. M.; White, M. A.; Wilkinson, M. L.; Kon, Y.; Kung, L. A.; Wise, K. J.; Lopez-Hoyo, N.; Jiang, L.; Piccirillo, S.; Yu, H.; Gerstein, M.; Dumont, M. E.; Phizicky, E. M.; Snyder, M.; Grayhack, E. J. Biochemical and Genetic Analysis of the Yeast Proteome with a Movable ORF Collection. *Genes Dev.* **2005**, *19* (23), 2816–2826. https://doi.org/10.1101/gad.1362105.

(30) Hedrick, V. E.; LaLand, M. N.; Nakayasu, E. S.; Paul, L. N. Digestion, Purification, and Enrichment of Protein Samples for Mass Spectrometry. *Current Protocols in Chemical Biology* **2015**, *7* (3), 201–222. https://doi.org/10.1002/9780470559277.ch140272.

(31) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat Methods* **2019**, *16* (1), 63–66. https://doi.org/10.1038/s41592-018-0260-3.

(32) Mosley, A. L.; Sardiu, M. E.; Pattenden, S. G.; Workman, J. L.; Florens, L.; Washburn, M. P. Highly Reproducible Label Free Quantitative Proteomic Analysis of RNA Polymerase Complexes. *Mol Cell Proteomics* **2011**, *10* (2), M110.000687. https://doi.org/10.1074/mcp.M110.000687.

(33) Mosley, A. L.; Florens, L.; Wen, Z.; Washburn, M. P. A Label Free Quantitative Proteomic Analysis of the Saccharomyces Cerevisiae Nucleus. *J Proteomics* **2009**, *72* (1), 110–120. https://doi.org/10.1016/j.jprot.2008.10.008.

(34) Smith-Kinnaman, W. R.; Berna, M. J.; Hunter, G. O.; True, J. D.; Hsu, P.; Cabello, G. I.; Fox, M. J.; Varani, G.; Mosley, A. L. The Interactome of the Atypical Phosphatase Rtr1 in Saccharomyces Cerevisiae. *Mol Biosyst* **2014**, *10* (7), 1730–1741. https://doi.org/10.1039/c4mb00109e.

(35) Crooks, G. E. WebLogo: A Sequence Logo Generator. *Genome Research* **2004**, *14* (6), 1188–1190. https://doi.org/10.1101/gr.849004.

(36) Colaert, N.; Helsens, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K. Improved Visualization of Protein Consensus Sequences by IceLogo. *Nat Methods* **2009**, *6* (11), 786–787. https://doi.org/10.1038/nmeth1109-786.

(37) Oliveros, J.C. An Interactive Tool for Comparing Lists with Venn's Diagrams. *Venny*.

(38) Chen, T.; Muratore, T. L.; Schaner-Tooley, C. E.; Shabanowitz, J.; Hunt, D. F.; Macara, I. G. N-Terminal α-Methylation of RCC1 Is Necessary for Stable Chromatin Association and Normal Mitosis. *Nat Cell Biol* **2007**, *9* (5), 596–603. https://doi.org/10.1038/ncb1572.

(39) Swaney, D. L.; Wenger, C. D.; Coon, J. J. The Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *J Proteome Res* **2010**, *9* (3), 1323–1329. https://doi.org/10.1021/pr900863u.

(40) Zhang, M.; Xu, J.-Y.; Hu, H.; Ye, B.-C.; Tan, M. Systematic Proteomic Analysis of Protein Methylation in Prokaryotes and Eukaryotes Revealed Distinct Substrate Specificity. *PROTEOMICS* **2018**, *18* (1), 1700300. https://doi.org/10.1002/pmic.201700300.

(41) Wang, K.; Zhou, Y. J.; Liu, H.; Cheng, K.; Mao, J.; Wang, F.; Liu, W.; Ye, M.; Zhao, Z. K.; Zou, H. Proteomic Analysis of Protein Methylation in the Yeast Saccharomyces Cerevisiae. *Journal of Proteomics* **2015**, *114*, 226–233. https://doi.org/10.1016/j.jprot.2014.07.032.

(42) Yeom, J.; Ju, S.; Choi, Y.; Paek, E.; Lee, C. Comprehensive Analysis of Human Protein N-Termini Enables Assessment of Various Protein Forms. *Scientific Reports* **2017**, *7* (1). https://doi.org/10.1038/s41598-017-06314-9.

(43) Chalkley, R. J.; Clauser, K. R. Modification Site Localization Scoring: Strategies and Performance. *Mol Cell Proteomics* **2012**, *11* (5), 3–14. https://doi.org/10.1074/mcp.R111.015305.

(44) Jentoft, N.; Dearborn, D. G. Labeling of Proteins by Reductive Methylation Using Sodium Cyanoborohydride. *J. Biol. Chem.* **1979**, *254* (11), 4359–4365.

(45) Olsen, J. V.; Ong, S.-E.; Mann, M. Trypsin Cleaves Exclusively C-Terminal to Arginine and Lysine Residues *. *Molecular & Cellular Proteomics* **2004**, *3* (6), 608–614. https://doi.org/10.1074/mcp.T400003-MCP200.

(46) Jakobsson, M. E.; Davydova, E.; Małecki, J.; Moen, A.; Falnes, P. Ø. Saccharomyces Cerevisiae Eukaryotic Elongation Factor 1A (EEF1A) Is Methylated at Lys-390 by a METTL21-Like Methyltransferase. *PLOS ONE* **2015**, *10* (6), e0131426. https://doi.org/10.1371/journal.pone.0131426.

(47) Porras-Yakushi, T. R.; Whitelegge, J. P.; Clarke, S. A Novel SET Domain Methyltransferase in Yeast Rkm2-Dependent Trimethylation of Ribosomal Protein L12ab at Lysine 10. *J. Biol. Chem.* **2006**, *281* (47), 35835–35845. https://doi.org/10.1074/jbc.M606578200.

(48) Maddelein, D.; Colaert, N.; Buchanan, I.; Hulstaert, N.; Gevaert, K.; Martens, L. The IceLogo Web Server and SOAP Service for Determining Protein Consensus Sequences. *Nucleic Acids Research* **2015**, *43* (W1), W543–W546. https://doi.org/10.1093/nar/gkv385.

(49) Helbig, A. O.; Gauci, S.; Raijmakers, R.; van Breukelen, B.; Slijper, M.; Mohammed, S.; Heck, A. J. R. Profiling of N-Acetylated Protein Termini Provides In-Depth Insights into the N-Terminal Nature of the Proteome. *Mol Cell Proteomics* **2010**, *9* (5), 928–939. https://doi.org/10.1074/mcp.M900463-MCP200.

(50) Bonissone, S.; Gupta, N.; Romine, M.; Bradshaw, R. A.; Pevzner, P. A. N-Terminal Protein Processing: A Comparative Proteogenomic Analysis. *Mol Cell Proteomics* **2013**, *12* (1), 14–28. https://doi.org/10.1074/mcp.M112.019075.

(51) Drazic, A.; Myklebust, L. M.; Ree, R.; Arnesen, T. The World of Protein Acetylation. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2016**, *1864* (10), 1372–1401. https://doi.org/10.1016/j.bbapap.2016.06.007.

(52) Oishi, K.; Yamayoshi, S.; Kozuka-Hata, H.; Oyama, M.; Kawaoka, Y. N-Terminal Acetylation by NatB Is Required for the Shutoff Activity of Influenza A Virus PA-X. *Cell Rep* **2018**, *24* (4), 851–860. https://doi.org/10.1016/j.celrep.2018.06.078.

(53) Elleuche, S.; Pöggeler, S. Inteins, Valuable Genetic Elements in Molecular Biology and Biotechnology. *Appl Microbiol Biotechnol* **2010**, *87* (2), 479–489. https://doi.org/10.1007/s00253-010-2628-x.

(54) Holmes, W. M.; Mannakee, B. K.; Gutenkunst, R. N.; Serio, T. R. Loss of Amino-Terminal Acetylation Suppresses a Prion Phenotype by Modulating Global Protein Folding. *Nature Communications* **2014**, *5* (1), 4383. https://doi.org/10.1038/ncomms5383.

(55) Griffith, A. A.; Holmes, W. Fine Tuning: Effects of Post-Translational Modification on Hsp70 Chaperones. *International Journal of Molecular Sciences* **2019**, *20* (17), 4207. https://doi.org/10.3390/ijms20174207.

(56) Perrot, M.; Massoni, A.; Boucherie, H. Sequence Requirements for Nα-Terminal Acetylation of Yeast Proteins by NatA. *Yeast* **2008**, *25* (7), 513–527. https://doi.org/10.1002/yea.1602.

(57) Vanet, A.; Plumbridge, J. A.; Guérin, M.-F.; Alix, J.-H. Ribosomal Protein Methylation in Escherichia Coli: The Gene PrmA, Encoding the Ribosomal Protein L11 Methyltransferase, Is Dispensable. *Molecular Microbiology* **1994**, *14* (5), 947–958. https://doi.org/10.1111/j.1365-2958.1994.tb01330.x.

(58) Demirci, H.; Gregory, S. T.; Dahlberg, A. E.; Jogl, G. Recognition of Ribosomal Protein L11 by the Protein Trimethyltransferase PrmA. *EMBO J* **2007**, *26* (2), 567–577. https://doi.org/10.1038/sj.emboj.7601508.

(59) Cameron, D. M.; Gregory, S. T.; Thompson, J.; Suh, M.-J.; Limbach, P. A.; Dahlberg, A. E. Thermus Thermophilus L11 Methyltransferase, PrmA, Is Dispensable for Growth and Preferentially Modifies Free Ribosomal Protein L11 Prior to Ribosome Assembly. *J Bacteriol* **2004**, *186* (17), 5819–5825. https://doi.org/10.1128/JB.186.17.5819-5825.2004.

(60) Truscott, R. J. W.; Mizdrak, J.; Friedrich, M. G.; Hooi, M. Y.; Lyons, B.; Jamie, J. F.; Davies, M. J.; Wilmarth, P. A.; David, L. L. Is Protein Methylation in the Human Lens a Result of Non-Enzymatic Methylation by S-Adenosylmethionine? *Experimental Eye Research* **2012**, *99*, 48–54. https://doi.org/10.1016/j.exer.2012.04.002.

(61) Brosch, M.; Saunders, G. I.; Frankish, A.; Collins, M. O.; Yu, L.; Wright, J.; Verstraten, R.; Adams, D. J.; Harrow, J.; Choudhary, J. S.; Hubbard, T. Shotgun Proteomics Aids Discovery of Novel Protein-Coding Genes, Alternative Splicing, and “Resurrected” Pseudogenes in the Mouse Genome. *Genome Res.* **2011**, *21* (5), 756–767. https://doi.org/10.1101/gr.114272.110.

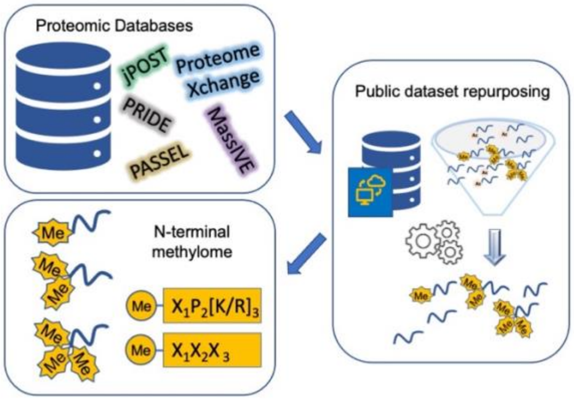
(62) Wang, M.; Wang, J.; Carver, J.; Pullman, B. S.; Cha, S. W.; Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems* **2018**, *7* (4), 412-421.e5. https://doi.org/10.1016/j.cels.2018.08.004.

(63) Hahne, H.; Kuster, B. Discovery of O-GlcNAc-6-Phosphate Modified Proteins in Large-Scale Phosphoproteomics Data. *Molecular & Cellular Proteomics* **2012**, *11* (10), 1063–1069. https://doi.org/10.1074/mcp.M112.019760.

(64) Matic, I.; Ahel, I.; Hay, R. T. Reanalysis of Phosphoproteomics Data Uncovers ADP-Ribosylation Sites. *Nature Methods* **2012**, *9* (8), 771–772. https://doi.org/10.1038/nmeth.2106.

(65) Jarmusch, A. K.; Wang, M.; Aceves, C. M.; Advani, R. S.; Aguirre, S.; Aksenov, A. A.; Aleti, G.; Aron, A. T.; Bauermeister, A.; Bolleddu, S.; Bouslimani, A.; Caraballo Rodriguez, A. M.; Chaar, R.; Coras, R.; Elijah, E. O.; Ernst, M.; Gauglitz, J. M.; Gentry, E. C.; Husband, M.; Jarmusch, S. A.; Jones, K. L.; Kamenik, Z.; Le Gouellec, A.; Lu, A.; McCall, L.-I.; McPhail, K. L.; Meehan, M. J.; Melnik, A. V.; Menezes, R. C.; Montoya Giraldo, Y. A.; Nguyen, N. H.; Nothias, L. F.; Nothias-Esposito, M.; Panitchpakdi, M.; Petras, D.; Quinn, R. A.; Sikora, N.; van der Hooft, J. J. J.; Vargas, F.; Vrbanac, A.; Weldon, K. C.; Knight, R.; Bandeira, N.; Dorrestein, P. C. ReDU: A Framework to Find and Reanalyze Public Mass Spectrometry Data. *Nature Methods* **2020**, *17* (9), 901–904. https://doi.org/10.1038/s41592-020-0916-7.

Graphical Abstract (For TOC only)


Proteomic Databases
jPOST
Proteome Xchange
PRIDE
MassIVE
PASSEL
Public dataset repurposing
N-terminal methylome
Me
$X_1P_2[K/R]_3$
$X_1X_2X_3$