Spatially resolved top-down proteomics of tissue sections based on a microfluidic nanodroplet sample preparation platform

Yen-Chen Liao[1], James M. Fulcher[1], David J. Degnan[2], Sarah M. Williams[1], Lisa M. Bramer[2], Dušan Veličković[1], Kevin J. Zemaitis[1], Marija Veličković[1], Ryan Sontag[2], Ronald J. Moore[2], Ljiljana Paša-Tolić[1], Ying Zhu[1,*], and Mowei Zhou[1,*]

1. Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, 3335 Innovation Boulevard, Richland, Washington 99354, United States.

2. Biological Sciences Division, Pacific Northwest National Laboratories, 902 Battelle Boulevard, Richland, Washington 99354, United States.

[*]Correspondence: Dr. Mowei Zhou, mowei.zhou@pnnl.gov and Dr. Ying Zhu, ying.zhu@pnnl.gov

Highlights

1. Top-down proteomics (TDP) of small tissue sections was demonstrated.
2. Proteoforms specific to anatomical regions in rat brain were detected.
3. An integrated informatics workflow for quantitative TDP was presented.

Abbreviations:

ABC(ammonium bicarbonate); ACN(acetonitrile); BCA(bicinchoninic acid); BUP(bottom-up proteomics); CV(compensation voltages); DDM(n-dodecyl-beta-maltoside); DMSO(dimethyl sulfoxide); ETD(electron transfer dissociation); FA(formic acid); FAIMS(field asymmetric ion mobility spectrometry); GO(Gene ontology); HCD(higher-energy collisional dissociation); HEK293(human embryonic kidney 293 cell line); KEGG(Kyoto encyclopedia of genes and genomes); LCM(Laser Capture Microdisecction); LC-MS/MS(liquid chromotography with tandem mass spectrometry); MSI(mass spectrometry imaging); nanoPOTS(nanodroplet Processing in One pot for Trace Samples); PBS(phosphate buffer saline); PP (polypropylene); PPI(protein-protein interaction); ProMex(Protein Mass Extractor); PTM(post-translational modification); RT(retention time); SPE(solid phase extraction); STRING(Search Tool for the Retrieval of Interacting Gene/Proteins); TCEP(Tris(2-carboxyethyl)phosphine); TDP(top-down proteomics); TFA(tri-fluoroacetic acid); TopFD(Top-down mass spectrometry feature detection); TopPIC(Top-down mass spectrometry-based Proteoform Identification and Characterization)

Abstract

Conventional proteomic approaches measure the averaged signal from mixed cell populations or bulk tissues, leading to the dilution of signals arising from subpopulations of cells that might serve as important biomarkers. Recent developments in bottom-up proteomics have enabled spatial mapping of cellular heterogeneity in tissue microenvironments. However, bottom-up proteomics cannot unambiguously define and quantify proteoforms, which are intact (i.e. functional) forms of proteins capturing genetic variations, alternatively spliced transcripts and post-translational modifications. Herein, we described a spatially resolved top-down proteomics (TDP) platform for proteoform identification and quantitation directly from tissue sections. The spatial TDP platform consisted of a nanoPOTS (nanodroplet Processing in One pot for Trace Samples)-based sample preparation system and an LCM (laser capture microdissection)-based cell isolation system. We improved the nanoPOTS sample preparation by adding benzonase in the extraction buffer to enhance the coverage of nucleus proteins. Using ~200 cultured cells as test samples, this approach increased proteoform identifications from 493 to 700; with newly identified proteoforms primarily corresponding to nuclear proteins. To demonstrate the spatial TDP platform in tissue samples, we analyzed LCM-isolated tissue voxels from rat brain cortex and hypothalamus regions. We quantified 426 proteoforms by combining identifications from TopPIC and TDPortal with the quantitation from ProMex. Several proteoforms corresponding to the same gene exhibited mixed abundance profiles between two tissue regions, suggesting potential PTM-specific spatial distributions. The spatial TDP workflow has prospects for biomarker discovery at proteoform level from small tissue sections.

Keywords

Top-down proteomics, nanoPOTS, laser capture microdissection, proteoform, spatial proteomics, quantitation

Introduction

Top-down proteomics (TDP) is a mass spectrometry (MS) strategy for characterizing "proteoforms", which encompasses the combination of post-translational modifications (PTMs), splice-isoforms, and amino acid variants occurring on a protein sequence(1). These variations at the proteoform level are not directly encoded in the genes. Still, they are critical to regulating cellular functions, particularly in the case of histones where co-occurrence of PTMs is known to influence chromatin biology and epigenetic regulation of genes(2). Combinatorial PTMs present a significant challenge for bottom-up proteomics (BUP) or antibody-based methods(3-5). TDP avoids ambiguity associated with proteoform inference from peptides by bypassing proteolytic steps(6, 7). Achieving high-quality proteoform identification with TDP, however, is challenging as it requires sufficient protein sample amount, high MS performance, and efficient fragmentation for confident assignment of PTMs. Thus, TDP typically requires bulk-scale tissue or large quantities of cultured cells (~10^6) to obtain sufficient proteoform coverages. Encouragingly, recent developments in MS instrumentation, methods, and informatics have significantly improved attainable sensitivity and depth of coverage(8-12), and thus allowed for reduced sample requirement towards single-cell level (13, 14). These advances have enabled the characterization of cellular heterogeneity among isolated cell populations or tissue regions (e.g., functional tissue units) that contain specific morphological and functional biomarkers (15-17). However, most of these advances were made for BUP analysis, obscuring the critical information needed for proteoform characterization.

Several microsampling and MS detection methods have been developed to enable highly sensitive and spatially resolved TDP analysis. Most of these advances were achieved employing MS imaging (MSI) methods, including matrix-assisted laser desorption ionization (MALDI)(18), nanospray desorption electrospray ionization (nanoDESI),(19), (20) liquid extraction surface analysis (LESA)(21), and laser ablation electrospray ionization (LAESI)(22). However, directly identifying proteins with MS/MS fragmentation in MSI is not trivial due to overlapping signals, salt adducts, and low signal intensity(23). In MALDI MSI, an additional challenge is that ions typically have low charge states (≤3), which greatly reduces fragmentation efficieny(18). For this reason, intact protein databases or prior knowledge from MS profiles and fragmentation patterns are required for peak assignment(18). Additionally, because of the lack of separation, MSI methods are typically limited to detecting highly abundant or highly ionazible proteins. To address these challenges, liquid microjunction (LMJ) microextraction(24), parafilm-assisted microdissection (PAM)(24), and laser capture microdissection (LCM)(25) have been explored to isolate and characterize microstructures from tissue sections. For example, the integration of LCM and capillary electrophoresis with TDP has enabled identification of over 400 proteoforms from two different regions of zebrafish brain (25).

Herein, we describe an improved spatial TDP platform that integrates LCM-based sample isolation with our previously developed nanoPOTS (nanodroplet processing in one-pot

for trace samples) sample preparation. We have demonstrated that nanoPOTS-based TDP can significantly improve the recovery of low amounts of samples by minimizing protein absorption on container surfaces(26). Over 150 proteoforms were identified from ~70 cultured HeLa cells, and a variety of post-translational modifications and proteoforms assigned(26). In this work, we further improved the nanoPOTS protocol for enhanced proteoform coverage and extended the application from cultured cells to tissue sections. We added nuclease benzonase in the extraction buffer to reduce sample viscosity and improve protein extraction efficiency as reported previously for bulk analyses(27, 28). To achieve deeper proteome coverage and more confident identifications, we developed several scripts that integrate qualitative and quantitative results from ProMex, TopPIC, and TDPortal (available at https://github.com/PNNL-HubMAP-Proteoform-Suite/spatially-resolved-TDP). To demonstrate the spatial TDP analysis, we employed LCM to isolate neuron cells from the cortex and hypothalamus regions in a rat brain section, and detected differential proteoform profiles between the two regions. We found varying proteoform abundance profiles for the same protein (gene), highlighting the need for the the proteoform-centric measurements. Finally, we demonstrated the identified proteoforms from the LCM-nanoPOTS-TDP analyses can be used to annotate protein peaks in MALDI-MSI spectrum.

Experimental Procedure

Reagents and chemicals

Deionized water (18.2 MΩ) was purified using a Barnstead Nanopure Infinity system (Los Angeles, CA). Tris(2-carboxyethyl)phosphine (TCEP), n-dodecyl-beta-maltoside (DDM) detergent, and protease/phosphatase inhibitor cocktails (catalog 78430) are from ThermoFisher Scientific (St. Louis, MO, USA). Benzonase nuclease is purchased from EMD Millipore (Billerica, MA, USA). Magnesium chloride ($MgCl_2$), formic acid (FA), 1x phosphate buffer saline (PBS), dimethyl sulfoxide (DMSO), tri-fluoroacetic acid (TFA), ethanol (EtOH), FA, and ammonium bicarbonate (ABC) are from Sigma-Aldrich (St. Louis, MO, USA).

Cell culture

Human embryonic kidney 293 (HEK293) cells were cultured under Dulbecco's modified Eagle's medium (DMEM) with 10% fetal bovine serum and 1% penicillin streptomycin at 37°C and 5% CO_2 atmosphere.

Rat brain tissue sectioning

Frozen female rat brain, purchased from BioIVT, was mounted on cryomicrotome chuck and then sectioned (10 µm thickness; CryoStar NX70, Thermo Fisher) using temperature of -18 ºC and -20 ºC, for specimen and blade, respectively. Sections were thaw-mounted onto indium tin oxide (ITO)-coated glass slides (Bruker Daltonics) for MALDI imaging and

onto polyethylene naphthalate (PEN) membrane slides (Carl Zeiss Microscopy, Germany) for LCM coupled to nanoPOTS experiments.

MALDI-MSI analysis

Samples were desiccated for 30 minutes and then washed in fresh solutions of 70% ethanol for 30 seconds, 100% ethanol for 30 seconds, Carnoy's solution (6:3:1 v/v ethanol/chloroform/glacial acetic acid) for 2 minutes, 100% ethanol for 30 seconds, water with 0.2% TFA for 15 seconds, and immediately washed in 100% ethanol for 30 seconds. Samples were then dried by a stream of nitrogen gas prior to MALDI matrix application. HTX Technologies M5 Sprayer (Chapel Hill, NC) was used to deposit sonicated supernatant of 15 mg/mL 2,5-DHA (2,5-dihydroxyacetophenone) in 90% acetonitrile with 0.2% TFA. . The flow rate of matrix was 150 µL/min with a nozzle temperature of 30.0°C; velocity was set to 1300 mm/min with 10 PSI of nitrogen gas. Matrix was then recrystallized with 5% acetic acid solution in water at 38.5°C and dried for 3.5 minutes and then immediately analyzed using an elevated pressure MALDI source (Spectroglyph LLC, Kennewick, WA) coupled to a Thermo Scientific Q Exactive HF Orbitrap MS upgraded with ultra-high mass range (UHMR) boards(29). Spectra were acquired over the *m/z* range of 3,500 to 20,000 in positive polarity mode with a resolving power of 240k at *m/z* 200 (512 ms transient) and 250 laser shots per pixel. Scans in the .RAW file were summed as a single spectrum for proteoform assignment by accurate mass.

LCM-nanoPOTS-TDP sample preparation

NanoPOTS chips were fabricated on glass substrates using photolithography, followed by a wetting etching solution containing 1 M HF, 0.5 M NH_4F, and 0.75 M HNO_3 processed with procedures as described previously(12). Polypropylene (PP) chips were produced by an injection molding company (Proto Labs, Maple Plain, MN). Glass or PP chips with an array of 4 × 12 nanowells were used throughout the study. Cells were collected in 1x PBS with protease and phosphatase inhibitor. After cell deposition, 100-nL lysis buffer containing 2 mM $MgCl_2$, 10 mM TCEP, and 16 M urea with 0.4% DDM in 50 mM ABC was added into each well, followed by 1-hour incubation under room temperature. Next, 200 nL of 2 mM $MgCl_2$ with 2.5 unit/ µL of benzonase nuclease was added in each well and incubated at 37°C for 1 hour. Finally, the sample was acidified by adding 50 nL of 5% FA into each well and dried in a vacuum chamber.

For tissue samples, the sections were fixed in 70% EtOH for 1 min and dehydrated in 95% and 100% EtOH (1 min per wash). A PALM MicroBeam system (Carl Zeiss MicroImaging, Munich, Germany) was used to perform cell isolation from different regions of rat brain. For each replicate, tissue voxels with an area of 100,000 μm^2 were excised and collected in PP microPOTS chip preloaded with 2 µL DMSO as capture liquid. Before protein extraction, DMSO was evaporated by heating the chip to 70°C. Next, we added 2 µL lysis buffer in each well that contained 2.5 unit/ µL benzonase nuclease, 2 mM $MgCl_2$, 10 mM TCEP, 0.2% DDM, and 4M urea in 50 mM ABC, followed by 1-hour incubation at 37°C. The sample was acidified by adding 500 nL of 5% FA into each well and dried in a

vacuum chamber. Dried microPOTS chips were frozen at -20°C or directly used for LC-MS/MS analyses.

LC-MS analysis

SPE columns (150 µm i.d.,4 cm long) and the analytical columns (100 µm i.d., 50 cm long) were packed in-house using C2 particles (SMTC2MEB2-3-300) from Separation Methods Technologies (Newark, DE). A home-built autosampler system was used for direct sample injection from nanoPOTS chip (30). The injected samples were loaded and desalted on SPE column by infusing with 0.1% FA at 3 µL/min for 5 minutes. We used Dionex nanoUPLC pump (NCP-3200RS, Thermo Scientific, Waltham, MA) system with 0.1% FA in H_2O (buffer A) and 0.1% FA in acetonitrile (buffer B). The LC gradient was programmed as a 120 min gradient from 10% to 50% buffer B followed by a 5 min linear gradient to 80% solvent B. The column was then washed with 70% solvent B for 5 min and re-equilibrated with 5% solvent B for 15 min. The LC flow rates were set at 300 nL/min for the 100-µm column.

Data were collected using Orbitrap Lumos Tribrid and Eclipse mass spectrometers (Thermo Scientific, San Jose, CA) in data-dependent acquisition mode. We applied field asymmetric ion mobility spectrometry (FAIMS) with compensation voltages (CV) of -30 V, -40 V, and -50 V. Precursor ion mass spectra were acquired with a resolution of 120 000 (at m/ z 200), a maximum injection time of 250 ms, a scan range of $600 < m/z < 2000$, an AGC target of 5E5, and 5 microscans. Precursor ions with charges 5+ or higher and intensities above 2E4 were isolated using an isolation window of 2 m/z for MS/MS analysis. A single charge state was selected for each neutral mass (i.e., proteoform) within 120 s dynamic exclusion. Tandem mass spectra were acquired with a resolution of 120K (at m/z = 200), using higher-energy collisional dissociation (HCD) with stepped collision energy levels (20%, 30%, and 40%), an AGC target of $1x\ 10^6$, and a maximum injection time with 500 ms. MS raw data and search results were uploaded to MassIVE with accession MSV000089163 (with password: TDP4066).

Proteoform identification and quantitation.

The FAIMS datasets were separated into individual raw files by FreeStyle (Thermo Scientific) for each CV. All files were deconvoluted with TopFD (TOP-down mass spectrometry feature detection) and searched by TopPIC (Top-down mass spectrometry-based Proteoform Identification and Characterization)(31) (ver. 1.4.13). All spectra were processed with parameters set at 15 ppm mass error tolerance, only one unexpected modification, proteoform error tolerance with 3.2 Da (for merging proteoforms with similar masses), and combined target and decoy search with an FDR (false discovery rate) threshold of 1%. MS/MS spectra were searched against UniProtKB/Swiss-Prot rat database (downloaded on August, 2021, containing 8,131 reviewed, 21,803 TrEMBL, and 1628 VarSplic sequences) or the human database (downloaded on June 29, 2019, containing 20,352 reviewed sequences).

We performed FDR filtering at the protein level, resulting in a combined FDR of < 1%. To describe ambiguity in proteoform identifications, we implemented a custom R function that determined a proteoform's "level" of ambiguity, following the five-level classification system (from 1-5 and 1 being unambiguous and 5 being ambiguous in all metrics) defined by the Consortium for TDP.(32) Our function could account for all forms of ambiguity apart from amino acid sequence ambiguity. Open-modification searches, while useful, can sometimes provide erroneous mass shift assignments.(33) To address these issues, we performed retention time alignment (LOESS regression) and mass error recalibration for proteoform spectrum matches (PrSMs) using the dataset with the larger number of PrSMs as a reference. Retention times were aligned using the apex spectrum (most intense) for each proteoform. Aligned and recalibrated datasets were then clustered using retention time and precursor mass for all PrSMs. We refer to these clusters as "Proteoform Clusters" (PfCs). A minimum of 3 PrSMs were required per cluster, and PrSMs not meeting this criterion were pooled together as a "noise" cluster and ignored for quantitative analysis. Within each PfC, the proteoform with the highest number of PrSMs was selected to represent the entire cluster. A newer implementation of the workflows for TopPIC post-processing with additional functions are available on GitHub within the R package TopPICR.(34) In parallel, we also processed the same data (after splitting CVs) by TDPortal(35) with *Rattus norvegicus* protein data set (May 2016) and parameters, including high precursor resolution, filter by FDR, and TDPortal's code set of standard 4.0.0. The proteoform identifications were exported as tables using TDViewer for merging with TopPIC results. The script used to accomplish merging of the two search results can be found at https://github.com/PNNL-HubMAP-Proteoform-Suite/spatially-resolved-TDP.

For label-free quantitation of proteoforms, we relied on the feature abundances from Protein Mass Extractor (ProMex). Retention time alignment of ProMex features was performed with ProMexAlign, with each CV separately aligned and missing features replaced with "NA" in R. The accurate masses and retention times were then used to link the feature abundances to proteoform identifications. Redundant proteoforms were first collapsed by Pfc in TopPIC results, and by accession number and monoisotopic mass in TDPortal results. Only the top-scored (lowest E-value) proteoform was used to represent each unique feature. Next, collapsed TopPIC and TDPortal proteoforms were matched individually to the aligned ProMex tables within +/- 1 Da and +/- 4 minutes mass and retention time tolerances, respectively.

After concatenating all CVs together, we sorted low-high by mass, and assigned a mass group when each subsequent mass was within 1 Da of a previous mass. Within each mass group, we sorted by retention time and assigned an RT group when each subsequent retention time was within 4 minutes of the previous retention time. Mass and RT groups were then combined to generate a unique "feature group" in which we collapsed all detected features. The final table includes count, and values per feature group, including maximum mass, mean retention times, and median intensity, along with TopPIC and TDPortal proteoform annotation. The features were annotated with proteoforms and filtered for downstream analyses, where each proteoform had to be

identified in at least two samples. The proteoform abundances were normalized to the median of each sample (combined FAIMS CV), missing values were imputed randomly from a normal distribution with 0.3 widths and downshift 1.8 standard deviations of each sample's $\log_2$ intensity distribution and an unpaired t-test was performed.

Pathway and network analysis

Protein association networks for the identified proteins were analyzed by STRING database (version 11.5)(36) for high-confidence (score>0.7) and medium-confidence (0.4<score<0.7) protein-protein interaction networks. Functional enrichment analysis was performed by ClueGO plugin (version 2.5.8)(37) in Cytoscape (version 3.8.2)(38) against the gene ontology (GO)(39), tissue expression database (TISSUES)(40), and Kyoto encyclopedia of genes and genomes (KEGG) database(41, 42) with rat (*Rattus norvegicus*) protein database.

Experimental Design and Statistical Rationale

To compare the improvement of benzonase treatment, we identified proteoforms from ~ 100 HEK293 cells with and without benzonase treatment (n=5 each) after LC-MS/MS analysis. We depict a scatter plot with cell numbers versus identified proteoforms for performing the slope differences after benzonase treatment.

We applied the benzonase treatment to rat brain LCM tissue TDP analysis. We collect five spots from the rat cortex region and four from the region near the hypothalamus. After protein extraction and LC-MS/MS analysis, we use principal component analysis (PCA) to distinguish the protein characteristic from profile of proteoform abundance in each LCM section. PCA was performed by Perseus(43). All of the figures are plotted by GraphPad Prism 9 (GraphPad Software).

Results

Benzonase treatment improved proteoform identifications

One of the main challenges with top-down proteomics is the extraction of intact proteins under conditions compatible with downstream analysis. To address this, we evaluated the effect of benzonase, which has been shown to improve the recovery of nuclear proteins in proteomics preparation(44) by digesting nucleic acid polymers bound to these proteins. The benzonase was added to 100-200 HEK293 cells in nanoPOTS wells and analyzed by LC-MS/MS following previous methods.(26) Overall, benzonase addition improved nuclear protein recovery at higher cell counts (p-value = 0.08) (**Fig. 1**). We fit linear regression models with the number of identified proteoforms as the response variable and the number of cells as the predictor per sample type (all or nuclear) and treatment type (with or without benzonase) (**Fig. 1a**). At 100 cells or less, the effect of benzonase on proteoform recovery was not significant. Therefore, sensitivity at this level is likely restricted by LC-MS/MS and not the extraction step. Based on 95% confidence intervals, significantly more proteoforms were identified with benzonase treatment when starting with at least 165 cells. The slope for all samples increased from 1.5 to 3.2 following benzonase addition.

Based on gene ontology (GO) annotation, we separately counted the changes of nuclear proteoforms from total proteoforms. Digestion of DNA strands released more nuclear proteoforms. Benzonase treatment increased proteoforms from cell nucleus significantly (p-value = 0.005), with a slope increase from -0.15 to 1.7 (**Fig. 1b**). In addition, we observed the reduced viscosity of sample solution after benzonase treatment, which was consistent with the previous reports to improve protein extraction and sample handling during transfer(44).

We also investigated if the use of PP plastic chip could reduce non-specific binding-related protein losses. Our previous evaluation indicated PP surface can improve the recovery of peptide samples(45). As shown in **Fig. S1**, we found the PP chips yielded a modest increase in the number of identified proteoforms using ~100 HEK cells as a test sample. Next, we combined these improvements into our nanoPOTS protocol and applied them to small-scale tissue samples, which represent a more challenging test for protein extractions.

LCM-NanoPOTS-TDP enabled the quantitation of 426 proteoforms from two rat brain regions with an area of ~100,000 μm^2 each

We applied the improved nanoPOTS TDP protocol to study LCM-derived rat brain tissues from cortex and hypothalamus regions. In these analyses, we employed FAIMS, which has been previously shown to improve proteoform coverage from bulk brain tissues.(46) The top-down workflow, illustrated in **Fig. 2a**, involved proteoform identification using two software tools (TopPIC(31) and TDPortal(35)); proteoform clustering to minimize redundancy using TopPICR; proteoform quantitation with ProMex; and data integration using custom R scripts.

We sectioned five spots in the cortex and four spots in the hypothalamus with an area of ~100,000 μm^2 each (**Fig. 2b**), corresponding to roughly 200 cells. In the raw data, we

observed a cluster of peaks with high intensities near 6.5 kDa in all analyses, which were not identified by the database search. With manual analysis of fragmentation data, we assigned these signatures to aprotinin, one of the ingredients from protease inhibitor cocktails we added in the lysis buffer. While these species did not directly interfere with the analysis, their high abundance suppressed endogenous proteoform signals and reduced MS/MS time available for their characterization, outweighing the benefit of protease addition. This finding corroborates a recent TDP study(46), which mentioned some protease inhibitor cocktails branded as MS-compatible contain small proteins and should be carefully considered for TDP applications.

The initial output from TopPIC and TDPortal listed 621 and 925 proteoforms, respectively. The two search engines have complementary algorithms but also feature different scoring and formatting, making it difficult to directly compare the results. To leverage complementarity and enhance proteoform coverage, we combined identifications from TDPortal and TopPIC that passed 1% FDR as defined by each tool. In parallel, ProMex was used to quantify proteoform features at the MS1 level independent of the identifications from the MS/MS data. Detected features were also aligned across all the samples using ProMexAlign algorithm. This alignment step, which is similar to the commonly used match-between-run(47, 48) or accurate mass and time tag (49)approach in bottom-up proteomics, was particularly important for filling the missing values in quantitative analysis. The aligned feature abundances were then attached to the combined proteoform identifications based on accurate mass and retention time matching. With this data integration approach, we obtained 426 quantifiable proteoforms (**Supplementary Table 1**). These included 183 proteoforms identified by both TopPIC and TDPortal, 102 identified only by TopPIC, and 141 identified only b TDPortal (**Fig. 2c**). This workflow can be applied to other TDP software outputs (such as identifications by pTop(50) or ProSightPC(51), and feature abundances from FLASHDeconv(52)) by adapting our R scripts. Combining the identifications from the two complementary algorithms resulted in a higher number of total proteoform counts. One challenge that proteoform quantitation faces is the split of proteoform abundance into multiple isotopologs for the same proteoform due to deisotoping error in the deconvolution step. To minimize redundancy, we chose to cluster LC-MS features within +/- 1 Da and +/- 4 minutes retention time, as described in Experimental section. Improved deisotoping algorithms(52) are needed to more effectively address this challenge.

The region-specific LC-MS/MS data can be used to generate spatially resolved proteoform databases for assigning peaks in MSI data(24), where MS/MS data are typically limited or absent. This approach significanly enhances spatially resolved proteoform measurement throughput. **Fig. 2d** shows an example of the highly abundant doubly charged peaks near *m*/*z* 5653.81 in MALDI-MSI, which can be assigned as H4c2[N-acetyl&dimethyl] (5650.69 monoisotopic*, charge 2*+) using the matching LCM-nanoPOTS-TDP data from the adjacent rat brain sections (**Fig. 2d** blue dots). The representative full MALDI spectrum annotated with proteoform identifications from nanoPOTS data is shown in **Fig. S2**. For MALDI-MSI in particular, the singly charged or doubly charged ions can be recalcitrant to fragmentation. Hence, proteoform assignments in MALDI-MSI often rely on global TDP data generated using bulk samples. Recent human proteoform atlas building efforts have been fruitful in generating tissue and cell-

type-specific proteoform databases,(53-55) but they may not fully represent the proteoform subpopulations in specific tissue regions. The proteoform profile may change in different microenvironments, and these differences can remain hidden in bulk analyses due to “signal dilution”, where bulk analyses average the response of entire tissues, obscuring region and cell-specific responses. Therefore, a spatially resolved proteoform database from nanoPOTS (or microPOTS) TDP could be highly valuable for accurate assignment of proteoforms in different tissue functional units and cells.

LCM-NanoPOTS-TDP captured PTM and isoform information

The majority of identified proteoforms were unmodified (not counting backbone cleavages), concurring with ~24% modified proteoforms (678 PTM modified in a total of 2843 proteoforms) in a recent TDP study of bulk human tissues(56). Nonetheless, several interesting modified proteoforms were confidently identified. For example, we identified Gng5 (guanidine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma) with S-geranylgeranyl modification at C64 (**Fig. 3a**), in agreement with previous reports(57) and the UniProt protein database. Other notable modified proteoforms included calmodulin with N-terminal acetylation (Calm2-N-acetyl, **Fig. S3a**) and phosphorylated calmodulin (Calm2-N-acetyl&-O-phospho, **Fig. S3b**). The unique benefit of TDP is the straightforward identification of proteoforms that can be challenging to differentiate using peptide-only data. In our results, myelin isoform 4 (P02688-4) was the only proteoform confidently assigned among the 5 recorded isoforms in UniProt. The other isoforms are results of alternative splicing and are only missing segments of the canonical sequence. Several myelin isoform 4 proteoforms with known PTMs were also detected with high confidence (proteoform level 1 or 2A). Distinct spatial distribution of myelin isoforms has been reported by nanoDESI measurements(58-60). We found that Mbp-o-phospho has higher abundance in the cortex than in the hypothalamus, which is consistent with a previous study(59). These findings demonstrate TDP could play important role in deciphering proteoform-specific information, which is critical for understanding the contributions of proteoforms to cellular heterogeneity and function.

LCM-NanoPOTS-TDP captured differential proteoform profiles in the cortex and hypothalamus regions of rat brain

LCM-nanoPOTS-TDP captured different proteoform compositions in the cortex and hypothalamus regions based on the principal components analysis (PCA) analysis as shown in **Fig. 4a**. Tissue samples from the cortex and hypothalamus were grouped in blue and pink clusters, respectively. Variances in the nearby spots of the same tissue region implied potential heterogeneity even within the same region. The score plot of PCA (**Fig. 4b**) showed the differentiating proteoforms for the two tissue regions. Calm2-(1-149)O-phospho, Snca(1-140)[Acetyl], Pcp4(2-62)[Acetyl], and Mbp(2-128)O-phospho were enriched with cortex region, while Sncb(84-134), Vgf(285-346), Gap43(188-226), and Gap43(48-90) were enriched with hypothalamus regions. To investigate possible connections between PTMs, proteoforms, and spatial abundance differences, we mapped the proteoforms to the protein-protein interaction (PPI) database with the network plot in STRING (**Fig. 5**). Because some of the truncated proteoforms may be result of sample degradation, we further filtered the identified proteoforms to include only proteoforms covering over 60% of the canonical sequence from Uniprot protein database.

In addition, only proteoforms from genes categorized as highly expressed in the brain were included. We selected one proteoform with the lowest p-value (i.e., most significantly changed in abundance between the two tissue regions) to represent each protein (**Fig. 5**). Several proteins (e.g., Pvalb, Mbp) were known to be highly expressed in the prefrontal cortex (highlighted by green dash lines) in the tissue expression database(TISSUES)(40). We observed significantly higher abundances of their proteoforms in the cortex (blue circles in **Fig. 5a**), validating that our method captured the expected proteome differences between the two tissue regions.

While many identified proteoforms derived from the same gene had similar abundance profiles, some proteoforms showed opposite patterns (e.g., yellow circles in **Fig. 5a**). For these genes, we selected two representative proteoforms with the lowest p-value in each direction of the abundance profile change (i.e., blue indicates enrichment in cortex, and red indicates enrichment in hypothalamus). For example, two most significantly differentiating calmodulin proteoforms (**Fig. 5b**) showed different abundance profiles, with Calm2[N-acetyl&O-phospho] being highly abundant in cortex ($p = 0.0175$) and Calm2[N-acetyl] being highly abundant in hypothalamus ($p = 0.194$). Calm2 is known to interact with both Gap43 and Mbp (myelin basic protein), whose major proteoforms also showed opposite abundance profiles. MbpN-methyl&O-phospho] showed significantly higher abundance in cortex ($p = 0.0044$), suggesting a positive correlation with Calm2[N-acetyl&O-phospho]. In contrast, Gap43[O-phospho] showed higher abundance in the hypothalamus ($p = 0.0055$). Both Calm2 and Gap43 are involved in filopodia growth in neurons(61). Phosphorylation of Ser41 on Gap43 eliminates calmodulin binding(62) and stabilizes the interaction of Gap43 with actin filaments,(61) leading to increased membrane tension and promotion of filopodia growth(63). Therefore, the higher abundance of Gap43[O-phospho] may be related to the enhanced filopodia in hypothalamus relative to cortex. Moreover, calmodulin is a Ca^{2+} sensor. When its calcium binding pocket is affected by phosphorylation at Thr29, Tyr99, and Ser101, the binding affinity of Ca^{2+} will reduce(64). The released calcium could stimulate phosphorylation on myelin protein(65) by calcium/calmodulin-dependent protein kinase(66). This is in agreement with the same abundance trends of Mbp[O-phospho] and Calm2[N-acetyl&O-phospho] in our data. The spatially different abundance of Calm2[N-acetyl&O-Phospho] and Calm2[N-acetyl] suggested the proteoforms derived from the same gene (protein) have different functional roles in the cortex and hypothalamus regions.

Several of the proteoforms and their interacting partners had unknown PTMs (i.e., not assignable within the scope of this study). They were simply annotated as mass shifts here (see representative spectrum for Tmsb4x in **Fig. S4**). Some of the unknown shifts may originate from noncovalent adducts or labile PTM (which was lost during fragmentation; e.g., **Fig. S5** describing Cox7c proteoforms), with their biological significance currently unknown. In addition, limited sequence coverage did not allow for confident localization of PTMs in several cases. Future experiments will aim at higher sequence coverage employing alternative fragmentation methods, such as electron transfer dissociation (ETD) or ultraviolet photodissociation (UVPD), to improve proteoform characterization. Larger number of datasets is also needed to better define the statistical significance. For example, the Tmsb4x(2-42) Acetyl&[-56.05] proteoform showed significant difference in abundance between the two tissue regions, while the

Tmsb4x(2-42) Acetyl proteoform showed a large variation in abundance within the cortex group (**Figure 5c**). While experimental variation can simply explain the lack of statistical significance, microheterogeneity within the same tissue region may also play a role. Nonetheless, our current study demonstrated the potential of integrated LCM-nanoPOTS-TDP and MALDI-MSI platforms for quantifying proteoforms in a spatially resolved manner. The distinct abundance profiles for several proteoforms originating from the same gene reinforce the importance of proteoform-specific measuerments to precisely define their functional roles.

Discussion

In this study, we improved our previous nanoPOTS-TDP protocol for small sample analysis and applied it to quantitative TDP study of LCM-derived rat brain tissue sections. The use of benzonase in the extraction step improved proteoform counts by efficiently digesting DNA polymers and releasing DNA binding proteins. We also streamlined the data analysis workflow by integrating several TDP software tools. The R scripts(34) combined and clustered proteoform identifications from TopPIC(31)and TDPortal(35) outputs to maximize proteoform coverage and minimize redundancy. Independently, proteoforms were quantified at the MS1 level using ProMex, and aligned across all datasets to reduce missing values. The proteoform identifications were then combined with their corresponding abundances for label-free quantitation. Our data analysis workflow is generic and can be readily adapted to other TDP software outputs. Overall, we obtained 426 quantifiable proteoforms across cortex and hypothalamus regions of rat brain. The abundance profiles facilitated elucidation of proteoforms' function connecting with protein-protein interaction network databases. Notably, we observed different abundance profiles among several proteoforms derived from the same gene, highlighting the need for the proteoform-aware mapping of tissues. Our future work will involve integration of LCM-TDP with MALDI-MSI for enhanced throughput and spatial resolution for proteoform imaging from tissues. We envision that spatially resolved TDP will become an essential tool for generating high confidence identifications and quantitation necessary for biomarker discovery, e.g. higher throughput MSI experiments for precision diagnosis.

Acknowledgement

We thank Matthew Monroe at PNNL for helping with data upload; Ryan Tal Feller, and Joseph Brent Greer at Northwestern University for helping with TDPortal. This work was performed at the Environmental Molecular Science Laboratory (EMSL), a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830. This research was funded by the National Institutes of Health (NIH) Common Fund, Human Biomolecular Atlas Program (HuBMAP) grant UG3CA256959-01. This research was performed on EMSL project awards doi.org/10.46936/staf.proj.2020.51770/60000309.

Data availability

The MS proteomics data have been deposited to MassIVE with accession MSV000089163 (with password: TDP4066). It includes MS raw data files, TDPortal search results of rat brain (**Supplementary Table 2**), and TopPIC search results of rat brain (**Supplementary Table 3**) and benzonase experiment (**Supplementary Table 4**). To read the annotated MS/MS spectra, please use TDViewer 2.0 to open TDPortal results and the "_html.zip" files of TopPIC results deposited on MSV000089163(with password: TDP4066).

Author contribution

Yen-Chen Liao: Methodology, Formal analysis, Investigation, Writing – Original Draft, Visualization. **James M. Fulcher:** Software, Formal analysis, Writing - Review & Editing. **David J. Degnan:** Software, Formal analysis, Writing - Review & Editing. **Sarah M. Williams:** Investigation. **Lisa M. Bramer:** Formal analysis, Writing - Review & Editing. **Dušan Veličković:** Investigation, Writing - Review & Editing. **Kevin J. Zemaitis:** Investigation, Writing - Review & Editing. **Marija Veličković:** Resources. **Ryan Sontag:** Resources. **Ronald J. Moore:** Resources. **Ljiljana Paša-Tolić:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Funding acquisition. **Ying Zhu:** Conceptualization, Methodology, Formal analysis, Writing - Review & Editing. **Mowei Zhou:** Conceptualization, Methodology, Formal analysis, Writing - Review & Editing, Project administration.

Figure legends

Figure 1. Benzonase treatment enhanced both total (a) and nucleus (b) proteoform identifications at high cell counts. The scatter plots show the relationship of cell number to the number of identified proteoforms with benzonase (black dots) and without benzonase (gray triangles) treatment, where each point represents one experiment.

Figure 2. (a) Workflow of processing LCM-derived tissue samples with nanoPOTS-TDP platform. (b) Optical image of rat brain tissue section showing where the small LCM punches were taken in the cortex and hypothalamus regions. (c) Venn diagram showing the overlap of quantifiable proteoforms across all samples by TopPIC and TDPortal. (d) Zoom-in view of the MALDI intact protein MSI spectrum for the histone H4 proteoform, which was assigned based on identification by nanoPOTS LC-MS/MS.

Figure 3. Representative tandem mass spectrum of modified Gng5 proteoform with (a) annotated fragments and (b) fragment error map. Despite the relatively low sequence coverage, the b/y ions supported assignment of N-terminal acetylation and S-geranylgeranyl modification at the cysteine near the C-terminus (scan #3185 in Hubmap_Intact_Brain_C1_CV40.raw).

Figure 4. Principal component analysis (PCA) of proteoform abundances yields (a) two distinct clusters of cortex (blue) and hypothalamus (pink) samples, and (b) candidate proteoforms for differentiating brain tissue types. Proteoforms are named as gene name, followed by starting and ending residue numbers in parentheses, and PTM (if any).

Figure 5. (a) Several proteoform clusters revealed significant differences in the protein-protein interaction network between the cortex and hypothalamus region. Proteins either had higher abundance in the cortex (light blue), hypothalamus (pink), or had mixed abundance profiles between the two regions (yellow). The box next to the circle corresponds to one representative proteoform for the protein with lowest p-value, which is colored with $\log_2$(C/H) with dark blue for higher expression in the cortex and red with higher abundance in the hypothalamus. In the case of proteins with mixed abundance profiles (yellow circles), two proteoforms with lowest p-value and enriched in cortex and hypothalamus were shown. Each line between proteins has interaction evidence in the String database. (b) Box plots showing the abundances of Calm2-N-acetyl and Calm2-N-acetyl&O-phospho, as well as (c) Tmsb4x N-acetyl and Tmsb4x N-acetyl & [-56.05] in cortex (c) and hypothalamus (h) regions.

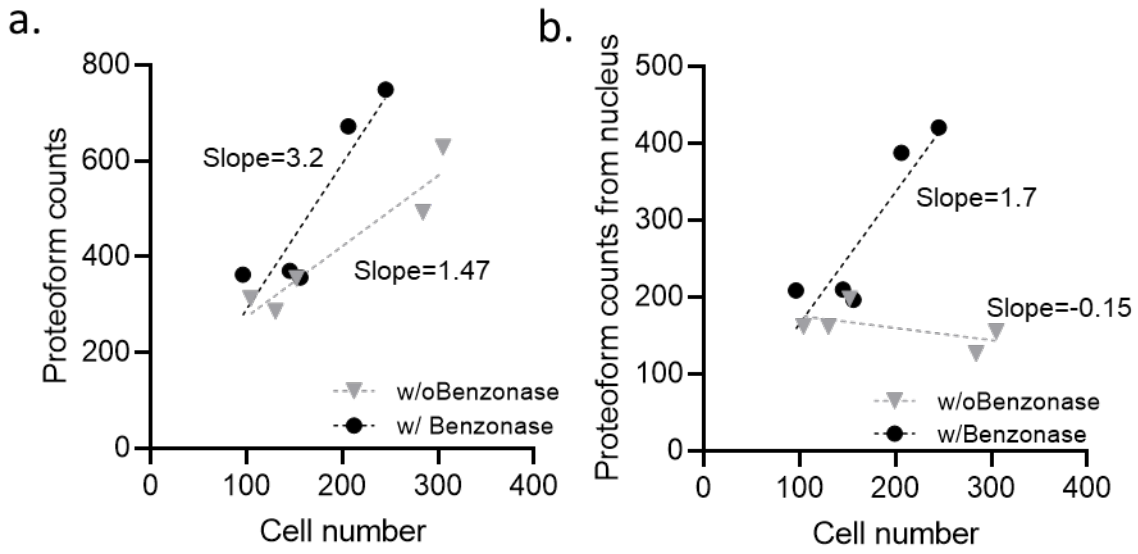


Figure 1. Benzonase treatment enhanced both total (a) and nucleus (b) proteoform identifications at high cell counts. The scatter plots show the relationship of cell number to the number of identified proteoforms with benzonase (black dots) and without benzonase (gray triangles) treatment, where each point represents one experiment.

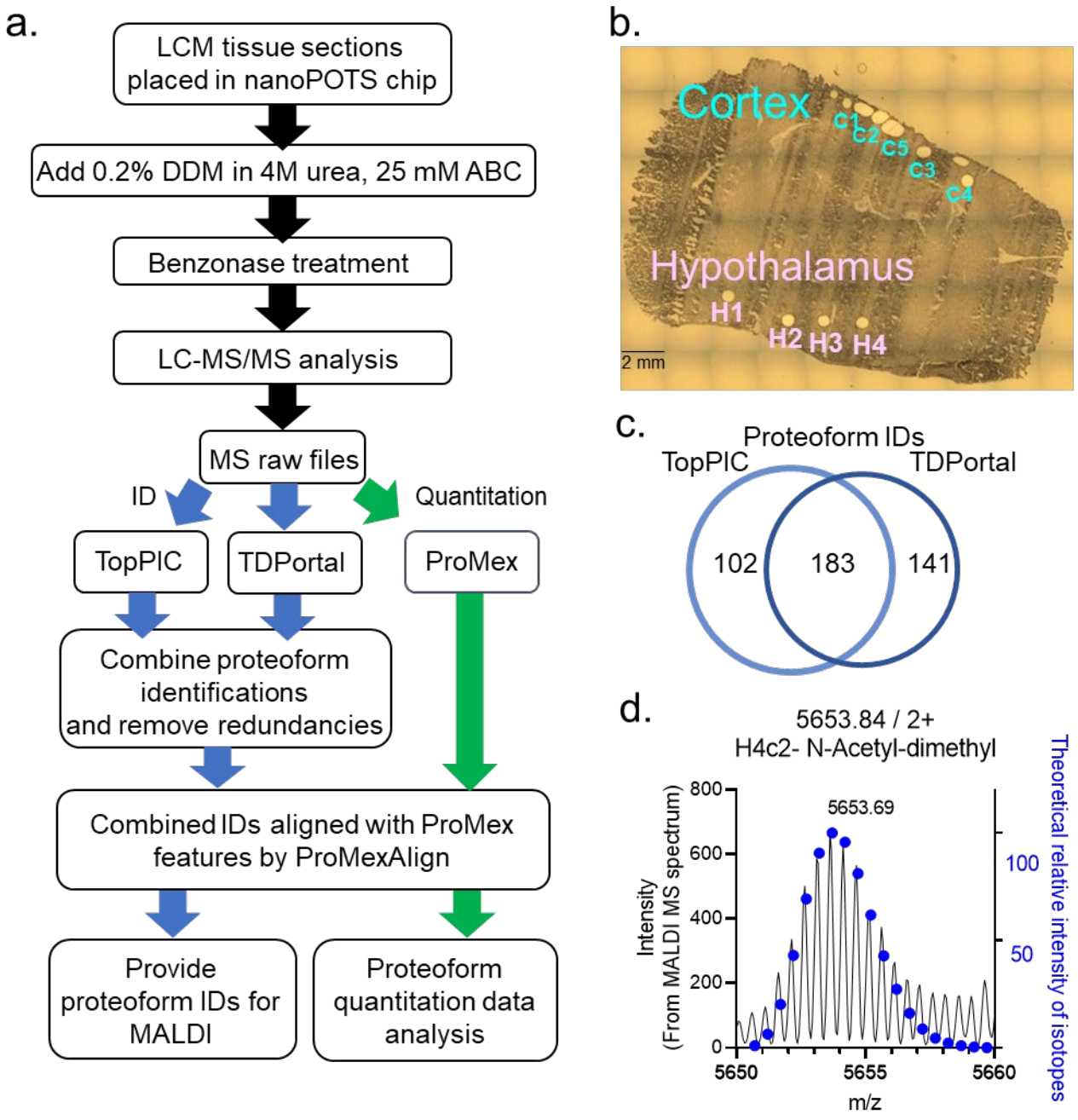


Figure 2. (a) Workflow of processing LCM-derived tissue samples with nanoPOTS-TDP platform. (b) Optical image of rat brain tissue section showing where the small LCM punches were taken in the cortex and hypothalamus regions. (c) Venn diagram showing the overlap of quantifiable proteoforms across all samples by TopPIC and TDPortal. (d) Zoom-in view of the MALDI intact protein MSI spectrum for the histone H4 proteoform, which was assigned based on identification by nanoPOTS LC-MS/MS.

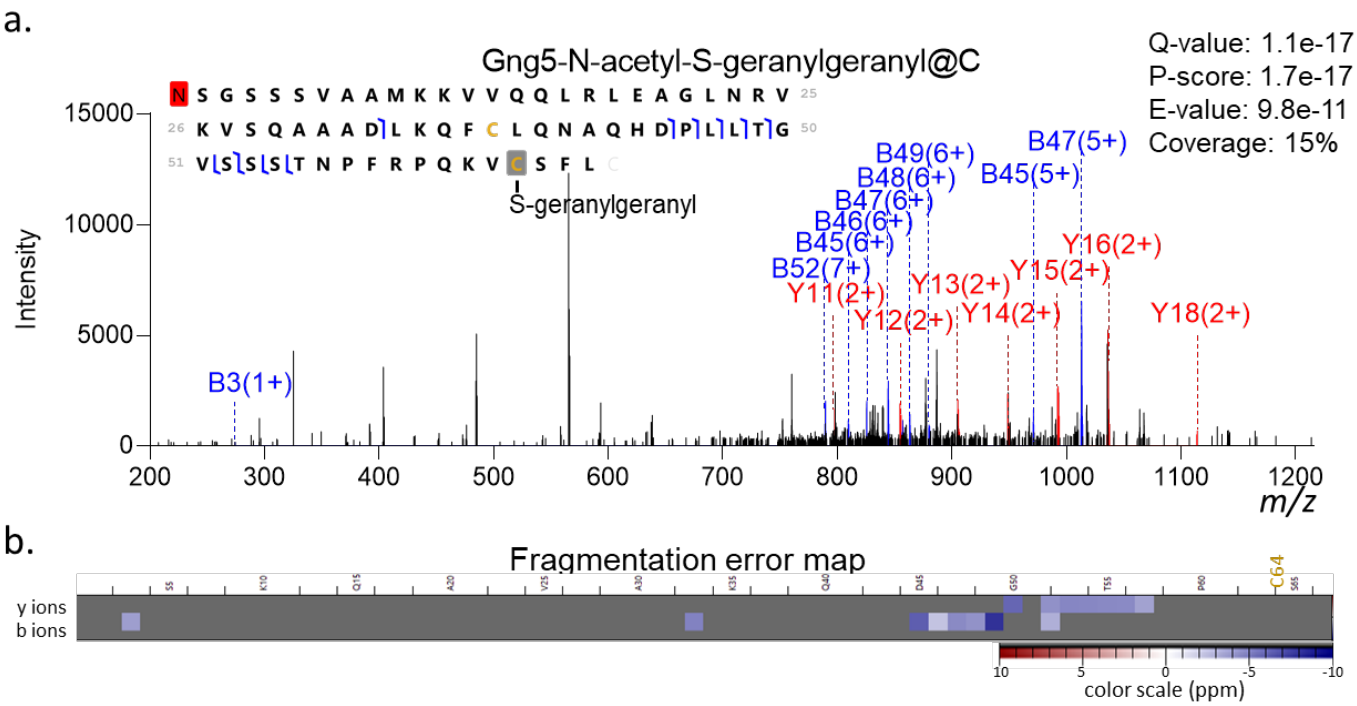


Figure 3. Representative tandem mass spectrum of modified Gng5 proteoform with (a) annotated fragments and (b) fragment error map. Despite the relatively low sequence coverage, the b/y ions supported assignment of N-terminal acetylation and S-geranylgeranyl modification at the cysteine near the C-terminus (scan #3185 in Hubmap_Intact_Brain_C1_CV40.raw).

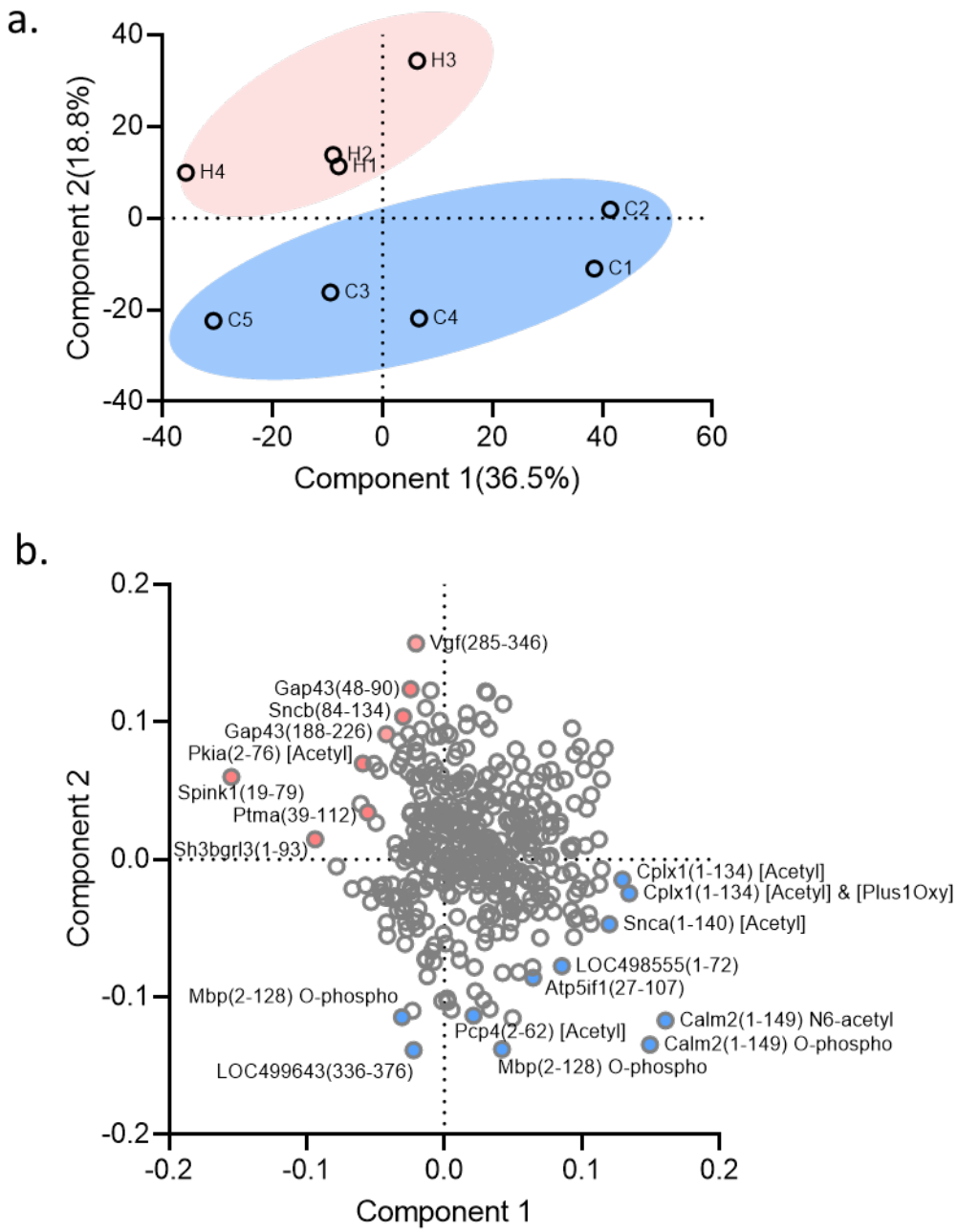


Figure 4. Principal component analysis (PCA) of proteoform abundances yields (a) two distinct clusters of cortex (blue) and hypothalamus (pink) samples, and (b) candidate proteoforms for differentiating brain tissue types. Proteoforms are named as gene name, followed by starting and ending residue numbers in parentheses, and PTM (if any).

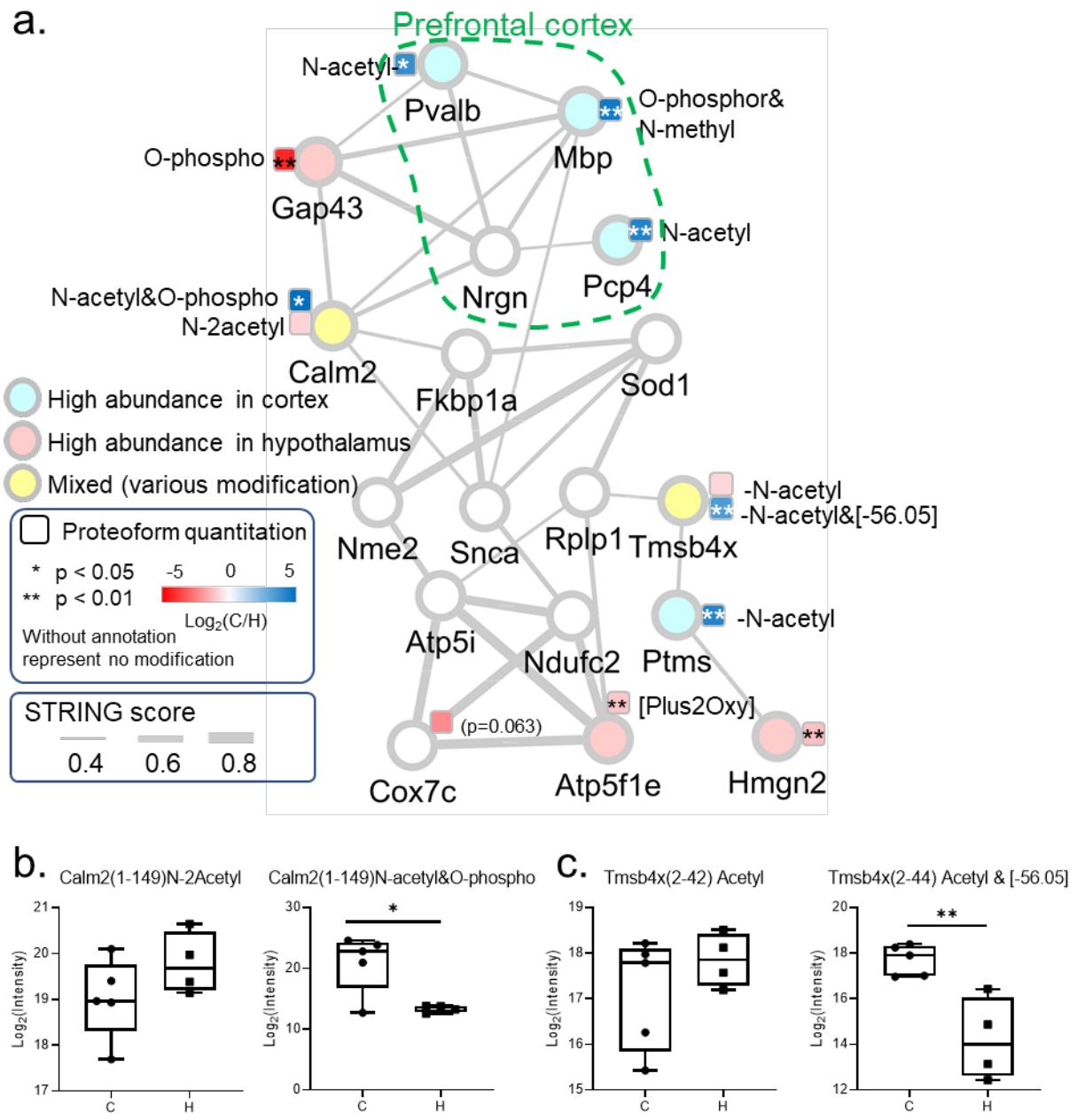


Figure 5. (a) Several proteoform clusters revealed significant differences in the protein-protein interaction network between the cortex and hypothalamus region. Proteins either had higher abundance in the cortex (light blue), hypothalamus (pink), or had mixed abundance profiles between the two regions (yellow). The box next to the circle corresponds to one representative proteoform for the protein with lowest p-value, which is colored with $\log_2$(C/H) with dark blue for higher expression in the cortex and red with higher abundance in the hypothalamus. In the case of proteins with mixed abundance profiles (yellow circles), two proteoforms with lowest p-value and enriched in cortex and

hypothalamus were shown. Each line between proteins has interaction evidence in the String database. (b) Box plots showing the abundances of Calm2-N-acetyl and Calm2-N-acetyl&O-phospho, as well as (c) Tmsb4x N-acetyl and Tmsb4x N-acetyl & [-56.05] in cortex (c) and hypothalamus (h) regions.

Reference

1. Smith, L. M., Agar, J. N., Chamot-Rooke, J., Danis, P. O., Ge, Y., Loo, J. A., Paša-Tolić, L., Tsybin, Y. O., and Kelleher, N. L. (2021) The Human Proteoform Project: Defining the human proteome. *Science Advances* 7, eabk0734
2. Bannister, A. J., and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Research* 21, 381-395
3. Chen, L., and Kashina, A. (2021) Post-translational Modifications of the Protein Termini. *Frontiers in Cell and Developmental Biology* 9
4. Rape, M. (2018) Ubiquitylation at the crossroads of development and disease. *Nature Reviews Molecular Cell Biology* 19, 59-70
5. Michalak, E. M., Burr, M. L., Bannister, A. J., and Dawson, M. A. (2019) The roles of DNA, RNA and histone methylation in ageing and cancer. *Nature Reviews Molecular Cell Biology* 20, 573-589
6. Bludau, I., Frank, M., Dörig, C., Cai, Y., Heusel, M., Rosenberger, G., Picotti, P., Collins, B. C., Röst, H., and Aebersold, R. (2021) Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nature Communications* 12, 3810
7. Liu, Y. (2022) A peptidoform based proteomic strategy for studying functions of post-translational modifications. *PROTEOMICS* 22, 2100316
8. Kafader, J. O., Melani, R. D., Durbin, K. R., Ikwuagwu, B., Early, B. P., Fellers, R. T., Beu, S. C., Zabrouskov, V., Makarov, A. A., Maze, J. T., Shinholt, D. L., Yip, P. F., Tullman-Ercek, D., Senko, M. W., Compton, P. D., and Kelleher, N. L. (2020) Multiplexed mass spectrometry of individual ions improves measurement of proteoforms and their complexes. *Nature Methods* 17, 391-394
9. Fornelli, L., and Toby, T. K. (2022) Characterization of large intact protein ions by mass spectrometry: What directions should we follow? *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1870, 140758
10. Fornelli, L., Toby, T. K., Schachner, L. F., Doubleday, P. F., Srzentić, K., DeHart, C. J., and Kelleher, N. L. (2018) Top-down proteomics: Where we are, where we are going? *J Proteomics* 175, 3-4
11. Melby, J. A., Roberts, D. S., Larson, E. J., Brown, K. A., Bayne, E. F., Jin, S., and Ge, Y. (2021) Novel Strategies to Address the Challenges in Top-Down Proteomics. *J Am Soc Mass Spectr* 32, 1278-1294
12. Zhu, Y., Piehowski, P. D., Zhao, R., Chen, J., Shen, Y., Moore, R. J., Shukla, A. K., Petyuk, V. A., Campbell-Thompson, M., Mathews, C. E., Smith, R. D., Qian, W.-J., and Kelly, R. T. (2018) Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat Commun* 9, 882
13. Woo, J., Williams, S. M., Markillie, L. M., Feng, S., Tsai, C. F., Aguilera-Vazquez, V., Sontag, R. L., Moore, R. J., Hu, D., Mehta, H. S., Cantlon-Bruce, J., Liu, T., Adkins, J. N., Smith, R. D., Clair, G. C., Pasa-Tolic, L., and Zhu, Y. (2021) High-throughput and high-efficiency sample preparation for single-cell proteomics using a nested nanowell chip. *Nat Commun* 12, 6246

14. Petelski, A. A., Emmott, E., Leduc, A., Huffman, R. G., Specht, H., Perlman, D. H., and Slavov, N. (2021) Multiplexed single-cell proteomics using SCoPE2. *Nature Protocols* 16, 5398-5425
15. Lubeckyj, R. A., and Sun, L. (2022) Laser capture microdissection-capillary zone electrophoresis-tandem mass spectrometry (LCM-CZE-MS/MS) for spatially resolved top-down proteomics: a pilot study of zebrafish brain. *Molecular Omics*
16. Piehowski, P. D., Zhu, Y., Bramer, L. M., Stratton, K. G., Zhao, R., Orton, D. J., Moore, R. J., Yuan, J., Mitchell, H. D., Gao, Y., Webb-Robertson, B.-J. M., Dey, S. K., Kelly, R. T., and Burnum-Johnson, K. E. (2020) Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-μm spatial resolution. *Nature Communications* 11, 8
17. Martinez-Val, A., Bekker-Jensen, D. B., Steigerwald, S., Koenig, C., Østergaard, O., Mehta, A., Tran, T., Sikorski, K., Torres-Vega, E., Kwasniewicz, E., Brynjólfsdóttir, S. H., Frankel, L. B., Kjøbsted, R., Krogh, N., Lundby, A., Bekker-Jensen, S., Lund-Johansen, F., and Olsen, J. V. (2021) Spatial-proteomics reveals phospho-signaling dynamics at subcellular resolution. *Nature Communications* 12, 7113
18. Ryan, D. J., Spraggins, J. M., and Caprioli, R. M. (2019) Protein identification strategies in MALDI imaging mass spectrometry: a brief review. *Curr Opin Chem Biol* 48, 64-72
19. Yang, M., Hu, H., Su, P., Thomas, P. M., Camarillo, J. M., Greer, J. B., Early, B. P., Fellers, R. T., Kelleher, N. L., and Laskin, J. (2022) Proteoform-Selective Imaging of Tissues Using Mass Spectrometry. *Angewandte Chemie International Edition*, e202200721
20. Hale, O. J., and Cooper, H. J. (2021) Native Mass Spectrometry Imaging of Proteins and Protein Complexes by Nano-DESI. *Analytical chemistry* 93, 4619-4627
21. Sarsby, J., Griffiths, R. L., Race, A. M., Bunch, J., Randall, E. C., Creese, A. J., and Cooper, H. J. (2015) Liquid Extraction Surface Analysis Mass Spectrometry Coupled with Field Asymmetric Waveform Ion Mobility Spectrometry for Analysis of Intact Proteins from Biological Substrates. *Anal Chem* 87, 6794-6800
22. Kiss, A., Smith, D. F., Reschke, B. R., Powell, M. J., and Heeren, R. M. (2014) Top-down mass spectrometry imaging of intact proteins by laser ablation ESI FT-ICR MS. *Proteomics* 14, 1283-1289
23. Hale, O. J., and Cooper, H. J. (2020) Native Mass Spectrometry Imaging and In Situ Top-Down Identification of Intact Proteins Directly from Tissue. *Journal of the American Society for Mass Spectrometry* 31, 2531-2537
24. Delcourt, V., Franck, J., Quanico, J., Gimeno, J. P., Wisztorski, M., Raffo-Romero, A., Kobeissy, F., Roucou, X., Salzet, M., and Fournier, I. (2018) Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions. *Mol Cell Proteomics* 17, 357-372
25. Lubeckyj, R. A., and Sun, L. (2022) Laser capture microdissection-capillary zone electrophoresis-tandem mass spectrometry (LCM-CZE-MS/MS) for spatially resolved top-down proteomics: a pilot study of zebrafish brain. *Mol Omics* 18, 112-122
26. Zhou, M., Uwugiaren, N., Williams, S. M., Moore, R. J., Zhao, R., Goodlett, D., Dapic, I., Paša-Tolić, L., and Zhu, Y. (2020) Sensitive Top-Down Proteomics Analysis of a Low Number of Mammalian Cells Using a Nanodroplet Sample Processing Platform. *Analytical Chemistry* 92, 7087-7095
27. Benedik, M. J., and Strych, U. (1998) Serratia marcescens and its extracellular nuclease. *FEMS Microbiology Letters* 165, 1-13
28. Franke, I., Meiss, G., and Pingoud, A. (1999) On the Advantage of Being a Dimer, a Case Study Using the DimericSerratia Nuclease and the Monomeric Nuclease fromAnabaena sp. Strain PCC 7120*. *Journal of Biological Chemistry* 274, 825-832
29. Zemaitis, K. V., Dusan; Kew, William; Fort, Kyle; Reinhardt-Szyba, Maria; Pamreddy, Annapurna; Ding, Yani; Kaushik, Dharam; Sharma, Kumar; Makarov, Alexander; Zhou, Mowei;

Paša-Tolić, Ljiljana (2022) Enhanced Spatial Mapping of Histone Proteoforms in Human Kidney Through MALDI-MSI by High-Field UHMR Orbitrap Detection. *ChemRxiv*
30. Williams, S. M., Liyu, A. V., Tsai, C. F., Moore, R. J., Orton, D. J., Chrisler, W. B., Gaffrey, M. J., Liu, T., Smith, R. D., Kelly, R. T., Pasa-Tolic, L., and Zhu, Y. (2020) Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics. *Anal Chem* 92, 10588-10596
31. Kou, Q., Xun, L., and Liu, X. (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* 32, 3495-3497
32. Smith, L. M., Thomas, P. M., Shortreed, M. R., Schaffer, L. V., Fellers, R. T., LeDuc, R. D., Tucholski, T., Ge, Y., Agar, J. N., Anderson, L. C., Chamot-Rooke, J., Gault, J., Loo, J. A., Paša-Tolić, L., Robinson, C. V., Schlüter, H., Tsybin, Y. O., Vilaseca, M., Vizcaíno, J. A., Danis, P. O., and Kelleher, N. L. (2019) A five-level classification system for proteoform identifications. *Nature Methods* 16, 939-940
33. Lysiak, A., Fertin, G., Jean, G., and Tessier, D. (2021) Evaluation of open search methods based on theoretical mass spectra comparison. *BMC Bioinformatics* 22, 65
34. Martin, E. A. (2022) evanamartin/TopPICR: AMP-AD pilot(v0.0.1). zenodo
35. Toby, T. K., Fornelli, L., Srzentić, K., DeHart, C. J., Levitsky, J., Friedewald, J., and Kelleher, N. L. (2019) A comprehensive pipeline for translational top-down proteomics from a single blood draw. *Nature Protocols* 14, 119-152
36. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and Mering, C. V. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47, D607-d613
37. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W. H., Pagès, F., Trajanoski, Z., and Galon, J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091-1093
38. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498-2504
39. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25-29
40. Palasca, O., Santos, A., Stolte, C., Gorodkin, J., and Jensen, L. J. (2018) TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database* 2018
41. Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28, 27-30
42. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 49, D545-d551
43. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* 13, 731-740
44. Li, Q., Jain, M. R., Chen, W., and Li, H. (2013) A multidimensional approach to an in-depth proteomics analysis of transcriptional regulators in neuroblastoma cells. *J Neurosci Meth* 216, 118-127
45. Dou, M., Tsai, C. F., Piehowski, P. D., Wang, Y., Fillmore, T. L., Zhao, R., Moore, R. J., Zhang, P., Qian, W. J., Smith, R. D., Liu, T., Kelly, R. T., Shi, T., and Zhu, Y. (2019) Automated Nanoflow Two-Dimensional Reversed-Phase Liquid Chromatography System Enables In-Depth Proteome and Phosphoproteome Profiling of Nanoscale Samples. *Anal Chem* 91, 9707-9715

46. Fulcher, J. M., Makaju, A., Moore, R. J., Zhou, M., Bennett, D. A., De Jager, P. L., Qian, W.-J., Paša-Tolić, L., and Petyuk, V. A. (2021) Enhancing Top-Down Proteomics of Brain Tissue with FAIMS. *Journal of Proteome Research* 20, 2780-2795
47. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10, 1794-1805
48. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 13, 2513-2526
49. Paša-Tolić, L., Masselon, C., Barry, R. C., Shen, Y., and Smith, R. D. (2004) Proteomic analyses using an accurate mass and time tag strategy. *BioTechniques* 37, 621-639
50. Sun, R.-X., Luo, L., Wu, L., Wang, R.-M., Zeng, W.-F., Chi, H., Liu, C., and He, S.-M. (2016) pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Analytical Chemistry* 88, 3082-3090
51. Zamdborg, L., LeDuc, R. D., Glowacz, K. J., Kim, Y.-B., Viswanathan, V., Spaulding, I. T., Early, B. P., Bluhm, E. J., Babai, S., and Kelleher, N. L. (2007) ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Research* 35, W701-W706
52. Jeong, K., Kim, J., Gaikwad, M., Hidayah, S. N., Heikaus, L., Schlüter, H., and Kohlbacher, O. (2020) FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. *Cell Systems* 10, 213-218.e216
53. Melani, R. D., Gerbasi, V. R., Anderson, L. C., Sikora, J. W., Toby, T. K., Hutton, J. E., Butcher, D. S., Negrão, F., Seckler, H. S., Srzentić, K., Fornelli, L., Camarillo, J. M., LeDuc, R. D., Cesnik, A. J., Lundberg, E., Greer, J. B., Fellers, R. T., Robey, M. T., DeHart, C. J., Forte, E., Hendrickson, C. L., Abbatiello, S. E., Thomas, P. M., Kokaji, A. I., Levitsky, J., and Kelleher, N. L. (2022) The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* 375, 411-418
54. Hollas, M. A. R., Robey, Matthew T., Fellers, Ryan T., LeDuc, Richard D., Thomas, Paul M., and Kelleher, Neil L. (2021) The Human Proteoform Atlas: a FAIR community resource for experimentally derived proteoforms. *Nucleic Acids Research* 50, D526-D533
55. Drown, B. J., K.; Melani, R.; Lloyd-Jones, C.; Camarillo, J.; Kelleher, N. (2022) Mapping the Proteoform Landscape of Five Human Tissues. *ChemRxiv*
56. Drown, B. S., Jooß, K., Melani, R. D., Lloyd-Jones, C., Camarillo, J. M., and Kelleher, N. L. (2022) Mapping the Proteoform Landscape of Five Human Tissues. *Journal of Proteome Research* 21, 1299-1310
57. Schwindinger, W. F., and Robishaw, J. D. (2001) Heterotrimeric G-protein betagamma-dimers in growth and differentiation. *Oncogene* 20, 1653-1660
58. Yang, M. H., H.; Su, P.; Thomas, P. M.; Camarillo, J. M.; Greer, J. B.; Early, B. P.; Fellers, R. T.; Kelleher, N. L.; Laskin, J. (2022) Proteoform-Selective Imaging of Tissues Using Mass Spectrometry. *ChemRxiv*
59. Hsu, C.-C., Chou, P.-T., and Zare, R. N. (2015) Imaging of Proteins in Tissue Samples Using Nanospray Desorption Electrospray Ionization Mass Spectrometry. *Analytical Chemistry* 87, 11171-11175
60. Anderson, D. M., Van de Plas, R., Rose, K. L., Hill, S., Schey, K. L., Solga, A. C., Gutmann, D. H., and Caprioli, R. M. (2016) 3-D imaging mass spectrometry of protein distributions in mouse Neurofibromatosis 1 (NF1)-associated optic glioma. *J Proteomics* 149, 77-84
61. He, Q., Dent, E. W., and Meiri, K. F. (1997) Modulation of actin filament behavior by GAP-43 (neuromodulin) is dependent on the phosphorylation status of serine 41, the protein kinase C site. *J Neurosci* 17, 3515-3524

62. Chapman, E. R., Au, D., Alexander, K. A., Nicolson, T. A., and Storm, D. R. (1991) Characterization of the calmodulin binding domain of neuromodulin. Functional significance of serine 41 and phenylalanine 42. *Journal of Biological Chemistry* 266, 207-213
63. Denny, J. B. (2006) Molecular mechanisms, biological actions, and neuropharmacology of the growth-associated protein GAP-43. *Curr Neuropharmacol* 4, 293-304
64. Benaim, G., and Villalobo, A. (2002) Phosphorylation of calmodulin. *European Journal of Biochemistry* 269, 3619-3631
65. Sulakhe, P. V., Petrali, E. H., Thiessen, B. J., and Davis, E. R. (1980) Calcium ion-stimulated phosphorylation of myelin proteins. *Biochemical Journal* 186, 469-473
66. Atkins, C. M., Yon, M., Groome, N. P., and Sweatt, J. D. (1999) Regulation of Myelin Basic Protein Phosphorylation by Mitogen-Activated Protein Kinase During Increased Action Potential Firing in the Hippocampus. *Journal of Neurochemistry* 73, 1090-1097

Supporting Information

Spatially resolved top-down proteomics of tissue sections based on a microfluidic nanodroplet sample preparation platform

Yen-Chen Liao[1], James M. Fulcher[1], David J. Degnan[2], Sarah M. Williams[1], Lisa M. Bramer[2], Dušan Veličković[1], Kevin J. Zemaitis[1], Marija Veličković[1], Ryan Sontag[2], Ronald J. Moore[2], Ljiljana Paša-Tolić[1], Ying Zhu[1,*], and Mowei Zhou[1,*]

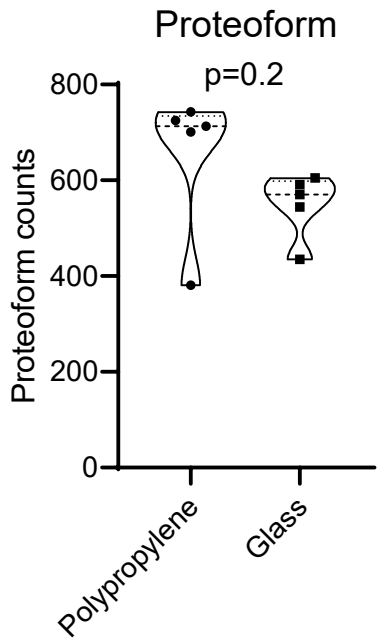
1. Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, 3335 Innovation Boulevard, Richland, Washington 99354, United States.

2. Biological Sciences Division, Pacific Northwest National Laboratories, 902 Battelle Boulevard, Richland, Washington 99354, United States.

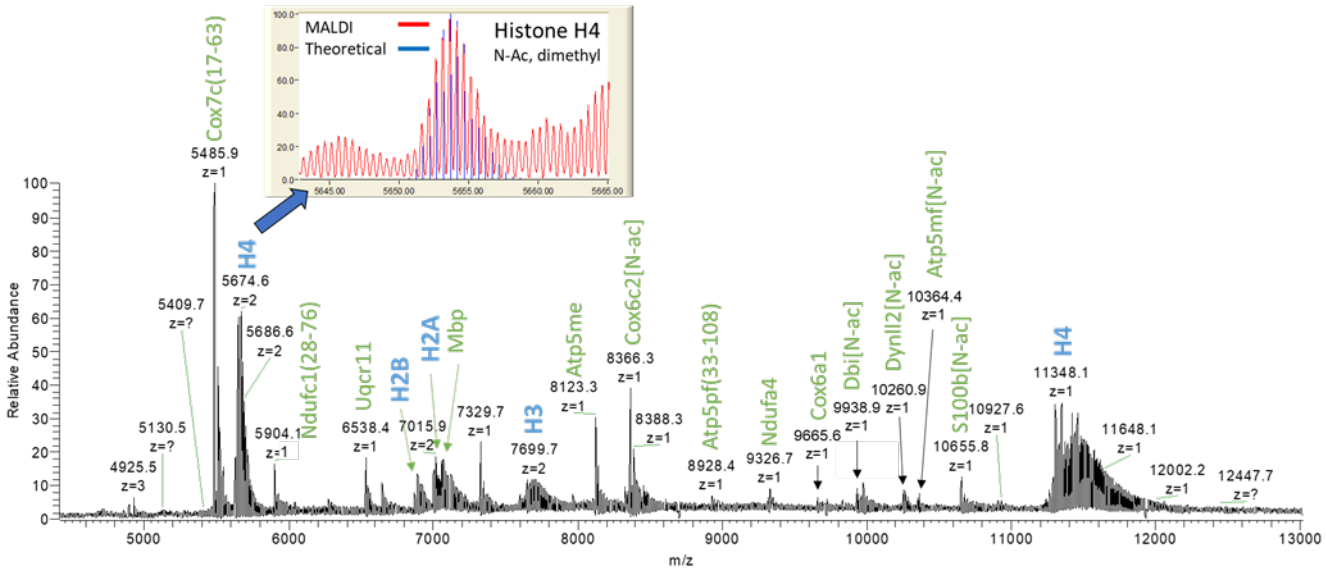
[*]Correspondence: Dr. Mowei Zhou, mowei.zhou@pnnl.gov

Dr. Ying Zhu, ying.zhu@pnnl.gov

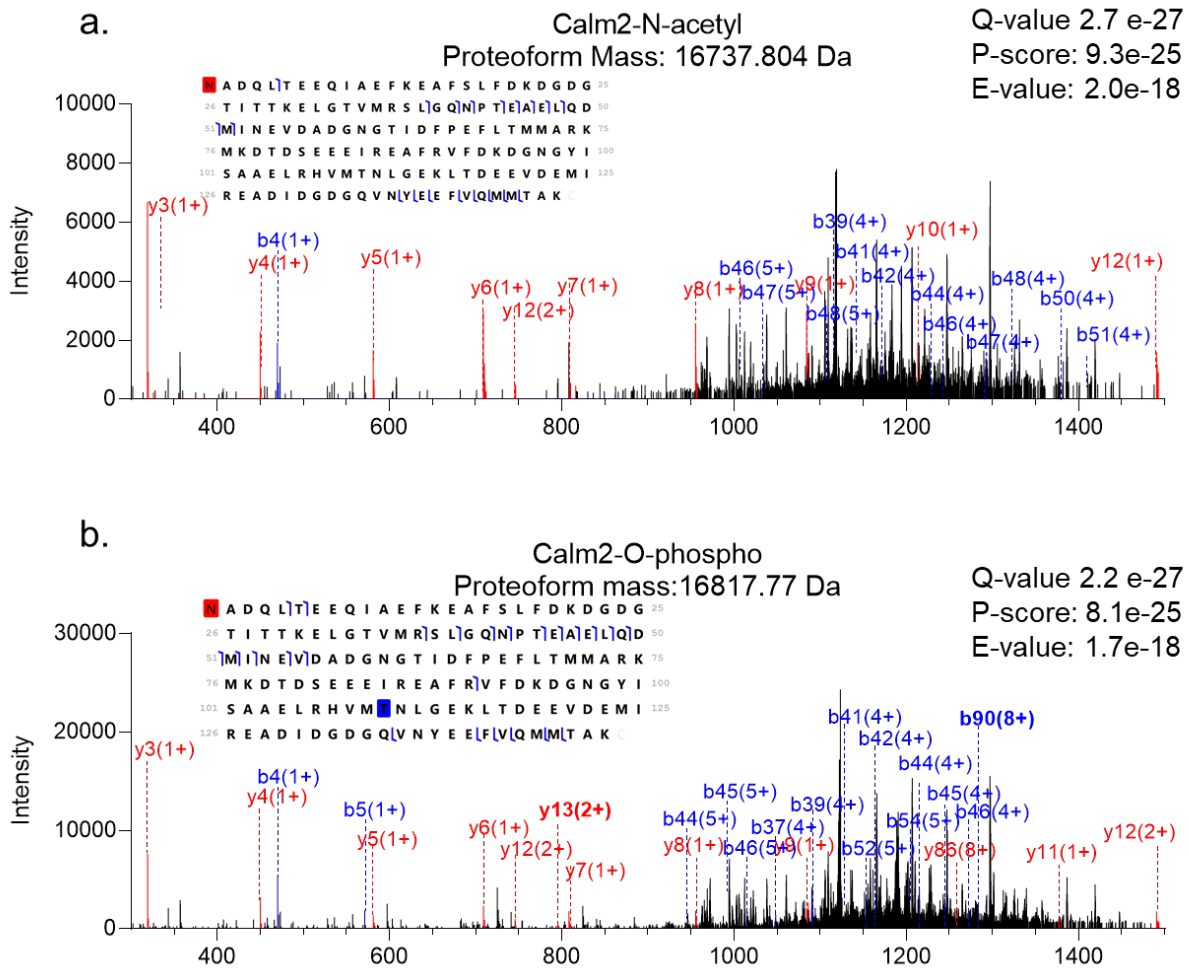
Includes Figure S1-S5.



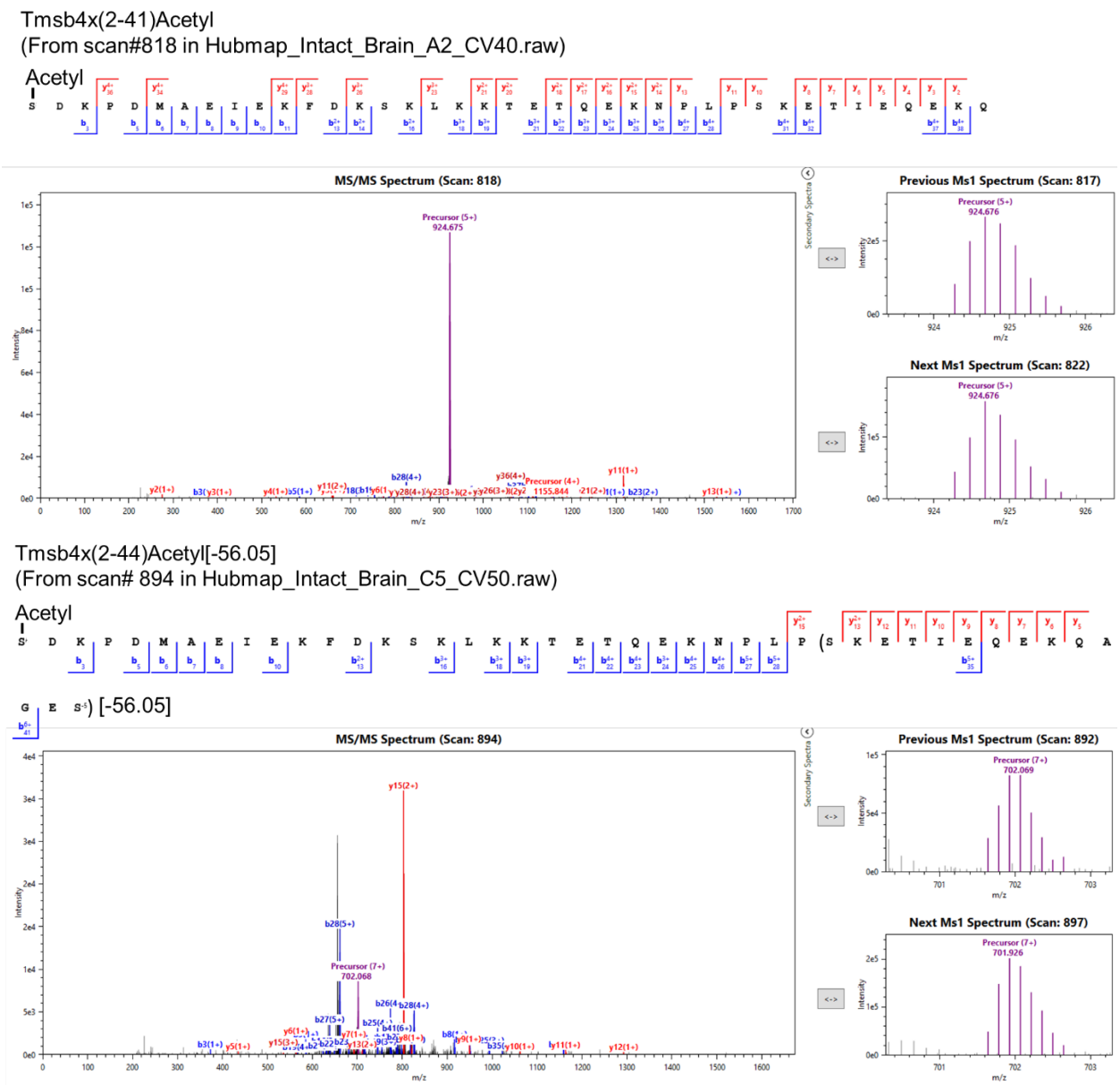
Supporting Figure S1. Polypropylene (PP) chips delivered higher number of proteoform identifications than glass chips from samples containing ~100 HEK cells. The improvement is due to reduced absorptive losses on PP surface in comparison to glass surface.



Supporting Figure S2. Representative MALDI spectrum of intact proteins from rat brain section. The data were acquired in MSI mode as described in the Experimetal section, and all pixels were summed to yield the spectrum shown. Major peaks were annotated with gene names (green text) based on the proteoforms identified in LCM-nanoPOTS. Truncated forms were labeled with the starting – ending residues in parentheses. PTMs were noted in brackets. Because multiple histones proteoforms corresponding to one or more histone genes were detected with similar masses, only the family names were labeled (blue text) for simplicity in this demonstration. The inset shows the zoom-in region of the MALDI spectrum (red trace) overlapping with the theoretical isotopic distribution of histone H4 N-ac, dimethyl proteoform (blue trace).

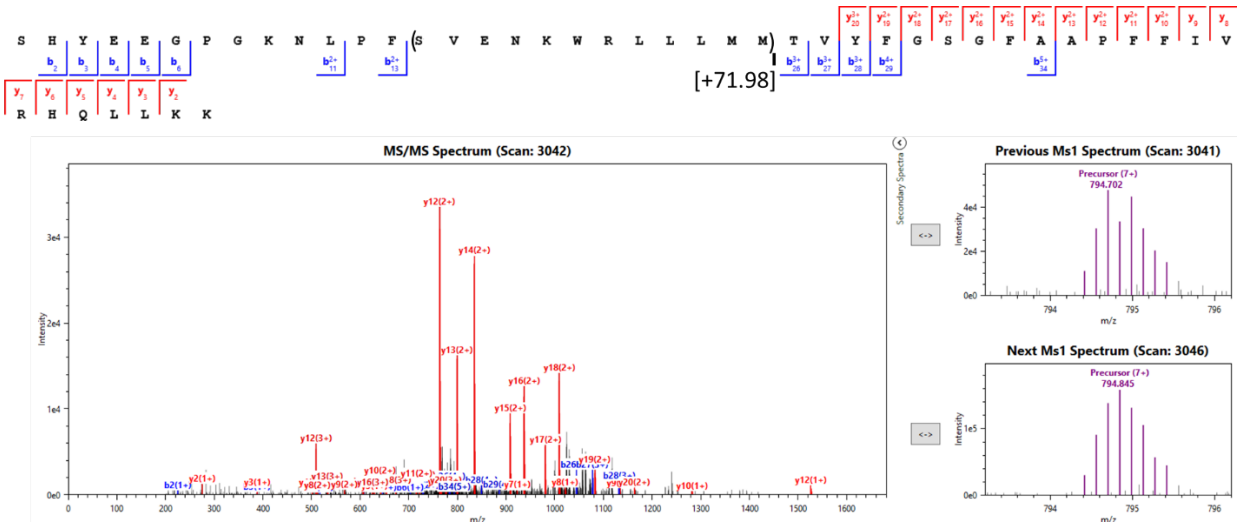


Supporting Figure S3. MS/MS data for (a) Calm2-N-acetyl (scan 2816 in cortex C3 CV40.raw file), and (b) Calm2-N-acetyl-O-phospho (scan 2714 in cortex C3 CV40.raw file). With additional phosphorylation modification between y13(2+) and b90(8+), one phosphorylation was localized between V91 and Q135. Y100, S102, and T111 have been recorded with phosphorylation modification in the uniport database.

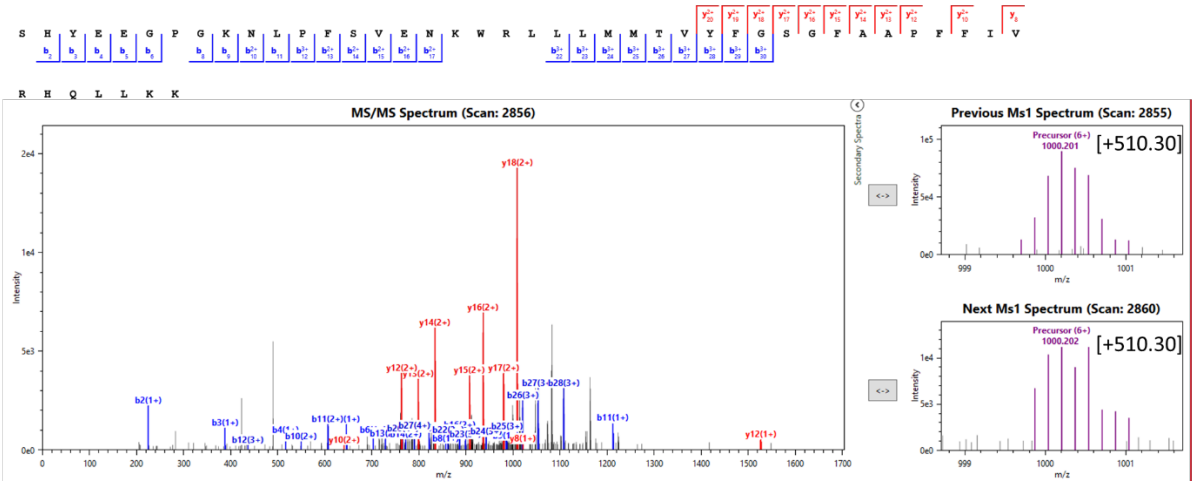


Supporting Figure S4. Annotated spectra for Tmsbx4 N-acetyl (top) and N-acetyl-[-56.0498] (bottom) proteoforms. Both spectra showed good sequence coverage, large number of matched fragments, and good signal for precursor ions. For the proteoform with mass shift of -56.05 at the C-terminus, truncation of residues alone cannot explain the observed mass difference.

Cox7c(17-63)-[+71.98]
From scan# 3042 in Hubmap_Intact_Brain_C1_CV50.raw)



Cox7c(17-63) [+510.30]
From scan# 2856 in Hubmap_Intact_Brain_A4_CV50.raw)



Supporting Figure S5. Annotated spectra for two Cox7c proteoforms with unknown mass shifts. Both spectra had high sequence coverage and good isotope fit for precursor matches. Mass shift of 71.98 Da may represent a combination of PTMs in the middle of the protein. Mass shift of 510.30 Da likely represents a noncovalent adduct or a labile PTM. The fragment spectra matched well to the unmodified protein, but the precursor ion contained an extra mass of 510.30 Da, implying the PTM was lost during fragmentation.