Mapping the Proteoform Landscape of Five Human Tissues

*Bryon S. Drown[1], Kevin Jooß[1], Rafael D. Melani[1], Cameron Lloyd-Jones[1], Jeannie M. Camarillo[1], Neil L. Kelleher[1]**

[1]Departments of Molecular Biosciences, Chemistry, and the Feinberg School of Medicine, Northwestern University, Evanston, IL, USA.

KEYWORDS

Top-down proteomics, proteomics, capillary zone electrophoresis, heart, small intestine, kidney, lung, spleen

ABSTRACT

A functional understanding of the human body requires structure-function studies of proteins at scale. The chemical structure of proteins is controlled at the transcriptional, translational, and post-translational levels, creating a variety of products with modulated functions within the cell. The term "proteoform" encapsulates this complexity at the level of chemical composition.

Comprehensive mapping of the proteoform landscape in human tissues necessitates analytical techniques with increased sensitivity and depth of coverage. Here, we took a top-down proteomics approach, combining data generated using capillary zone electrophoresis (CZE) and nanoflow reversed-phase liquid chromatography (RPLC) hyphenated to mass spectrometry to identify and characterize proteoforms from human lung, heart, spleen, small intestine, and kidney. CZE and RPLC provided complementary post-translational modification (PTM) and proteoform selectivity, thereby enhancing overall proteome coverage when used in combination. Of the 11,466 proteoforms identified in this study, 7,373 (64%) were not reported previously. Large differences in protein- and proteoform-level were readily quantified, with initial inferences about proteoform biology operative in the analyzed organs. Differential proteoform regulation of defensins, glutathione transferases, and sarcomeric proteins across tissues generate hypotheses about how they function and are regulated in human health and disease.

Introduction

Mapping the human body is critical to improving our understanding by setting definitive reference points for organs, tissues, and cells of diverse types. In proteomics, a complete understanding of proteoform[1] diversity requires measurements that systematically capture protein-level complexity. In projects like the Human Biomolecular Atlas Program (HuBMAP)[2] and Human Cell Atlas,[3] the resolution of mapping can handle single cells in tissues, with several highly multiplexed methods enabled by antibody-based affinity reagents: CODEX,[4] Immuno-SABER,[5] CyTOF,[6] and MIBI,[7, 8] among others. These methods measure the expression of particular epitopes on proteins, though they still fail to capture the full complexity of the proteoforms present. Proteoform-level measurements are more specific for a particular biological state compared to measurements on the gene or even protein level.[9, 10] While our long-term goal is to develop new technologies that deliver spatial proteoform analysis and build a comprehensive atlas of human proteoforms,[11] our goal here is to identify proteoforms present in primary human tissue and provide an initial assessment of their PTMs across tissue types.

Top-down proteomics (TDP), where intact proteins are isolated and fragmented by mass spectrometry (MS), is well suited for the identification and characterization of tissue-specific proteoforms. For the analysis of complex proteome samples, upfront separation and/or fractionation represents a crucial part in TDP workflows to reduce complexity prior to MS. Reversed-phase liquid chromatography (RPLC) is traditionally employed as the method of choice in TDP, *i.a.* due to its reproducibility, separation capacity, and MS compability, though capillary zone electrophoresis (CZE) represents an alternative for online MS. In particular, the separation principle of CZE is based on differences in electrophoretic mobilities (*charge-to-size* ratio) and is considered largely "orthogonal" to RPLC, where separation is driven by the hydrophobicity of analyte molecules. For this reason, the combination of information generated by both techniques is anticipated to increase the number of identified proteins and proteoforms.

Here, we report results from two workflows for mapping the proteoform landscape of solid tissues and present the first iteration with five commonly studied human tissues (heart, lung, kidney, small intestines, and spleen). Initially, the extracted proteoforms were pre-fractionated using Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE),[12] followed by subsequent CZE-MS and nano RPLC-MS analysis. This study contributes 7,373 proteoforms to the Human Proteoform Atlas (HPfA) a FAIR[13] knowledgebase that now contains approximately 60,000 unique proteoforms linked to their biological context.[14]

Experimental Procedures

Reagents

All reagents were purchased from Thermo Fisher Scientific at the highest available purity unless otherwise specified.

Tissue Lysate Preparation

Fresh-frozen tissue samples of human heart, lung, small intestine, and spleen were obtained from HuBMAP Tissue Mapping Centers (Table S1). Tissue samples were collected under IRB approved protocols at each institution. Kidney samples were received as 10 µm microtome scrolls embedded in methylcellulose (each ~5 mg). All other tissue types were cut into small pieces (~5 mm) by specimen preparer at Mapping Centers. Kidney scrolls were cryopulverized in 2 mL Eppendorf Protein Lo-Bind tubes containing a 5-mm stainless steel ball (Qiagen, cat. no. 69989) with a Cryomill (Retsch, cat. no. 20.749.001) equipped with a tube adaptor. Non-kidney tissue specimen (50-100 mg) were cryopulverized with the cryomill equipped with a 25 mL grinding jar containing a 1-inch stainless steel ball. Three cycles of precooling with liquid nitrogen at 1 Hz for 3 min and grinding at 30 Hz for 1 min were performed. Pulverized tissue was transferred to a 15 mL conical tube and resuspended in 2 mL cold RIPA lysis buffer (50 mM Tris, 150 mM NaCl, 1% NP-40 (v/v), 0.5% sodium deoxycholate (w/v), 0.1% sodium dodecyl sulfate (w/v), pH 7.4, 1X Halt Protease and Phosphotase Inhibitor Cocktail (Thermo Scientific)). The suspension was further disrupted by sonication on ice (40% power, cycle 2 s on, 3 s off, for 30 s total) with a probe sonicator (FisherBrand Model 120 with 1/8 inch probe) and then clarified by centrifugation (3234 × *g*, 30 min, 4 °C).

Sample Prefractionation and Preparation for Mass Spectrometry

Kidney lysates were studied with a 5x4x1x2 design: five biospecimen from separate donors were GELFrEE-fractionated into four fractions, analyzed by RPLC-MS/MS, and injected in duplicate. Lung lysates were studied in a 3x6x1x3 design: three samples from a single donor, six fractions, only RPLC, and three injections. Heart lysates were studied in a 2x6x2x3 design: two donors, six fractions, both CZE and RPLC, and three injections. Small intestine and spleen were studied in a 1x6x2x3 design: one sample, six fractions, both CZE and RPLC, and three injections. Lysates were fractionated and prepared for mass spectrometry as described previously.[15] Briefly, lysates were precipitated by adding four volumes of cold acetone and incubating at -80 °C for 1 hour. The precipitate was collected by centrifugation (20,000 × *g*, 30 min, 4 °C), and proteins were resolubilized in 1% sodium dodecyl sulfate (w/v). Total protein content was determined by BCA assay (Thermo Scientific). Samples were fractioned with the GELFrEE 8100 Fractionation Station (Expedeon). Protein samples (300 µg in 150 µL) were combined with 30 µL GELFrEE running buffer, and 8 µL 1 M DTT. The samples were incubated at 95 °C for 5 minutes, cooled to room temperature, and separated with a 10% GELFrEE cartridge following manufacturer's protocol. Six (four in the case of kidney samples) 150 µL fractions were collected and stored at -80 °C until immediately prior to analysis. On the day of analysis, fractions were thawed on ice and precipitated with methanol-chloroform-water as described.[16] Pellets were resuspended in 10 µL 0.3% acetic acid (HAc) (v/v) and subjected to CZE-MS/MS. When CZE-MS/MS analysis was completed, the samples were diluted with 20 µL of buffer A (5% acetonitrile, 94.8% water, 0.2% formic acid) and subjected to RPLC-MS/MS analysis. If only RPLC-MS/MS was conducted, the pellets were resuspended directly in 30 µL buffer A.

Capillary Zone Electrophoresis (CZE)

CZE was performed with a CESI 8000 Plus (Sciex) equipped with a Neutral OptiMS capillary cartridge (30 µm ID, L = 90 cm), neutrally coated. The cartridge was washed and conditioned according to the manufacturer's protocols. Separation conditions: Cartridge temperature: 15 ºC, Sample tray temperature: 4 ºC, background electrolyte: 3% HAc, conductive liquid: 3% HAc, hydrodynamic injection: 2.5 psi for 60 s (corresponds to ~20 nL). The individual separation method steps are listed in **Table S2**. Overnight, the capillary was rinsed alternating between high flow (100 psi, 2 min)and low flow (10 psi, 120 min) steps with water. For long-term storage, both separation and conductive lines were rinsed (100 psi) with water for 5 min, respectively, and the cartridge was stored at 4 ºC.

Reversed Phase Liquid Chromatography (RPLC)

RPLC was performed on an UltiMate 3000 RSLCnano system (Thermo Fisher Scientific) as described previous.[17] Briefly, a self-packed trap column (150 µm x 2.5 cm, PLRP-S 5 µm 1000-Å pore size) and analytical column (75 µm x 25 cm, PLRP-S 5µm 1000-Å pore size) were configured in a vented T setup. Trap and column were kept at 55 °C. Buffer A: 94.8% water, 5% acetonitrile, 0.2% formic acid, Buffer B: 94.8% acetonitrile, 5% water, 0.2% formic acid. Samples were injected (6 µL) onto the trap column and washed with 5% Buffer B at 3 µL/min for 10 min. Following a valve switch, proteins are separated on the analytical column according to the following gradient: 5% B at 10 min, 15% B at 13 min, 45% B at 70 min, 95% B at 72 min, 95% B at 76 min, 5% B at 80 min, 5% B from 80 to 90 min. For fractions 5 and 6 proteins were separated according to the following gradient: 5% B at 10 min, 15% B at 13 min, 50% B at 70 min, 95% B at 72 min, 95% B at 76 min, 5% B at 80 min, 5% B from 80 to 90 min. Eluted proteins were ionized in positive ion mode nanoelectrospray ionization using a pulled tip nanospray emitter (15 µm i.d. x 125 mm, New Objective) packed with 1mm of PLRP-S 5 µm 1000-Å pore size with a custom nano-source.

Top-down Mass Spectrometry

Mass spectrometry was performed either using a Thermo Scientific Orbitrap Eclipse Tribrid mass spectrometer or a Thermo Scientific Fusion Lumos Orbitrap Tribrid mass spectrometer. For analysis on Eclipse MS, data was acquired with the following global parameters spray voltage: 1600 V, sweep gas: 0, ion transfer tube temperature: 320 ºC, application mode: Intact Protein, pressure mode: Low Pressure (2 mTorr), Advanced Peak Determination: True, default charge state: 15, S-lens RF: 30%, source collision induced dissociation: 15 eV. Precursor spectra were acquired at 120,000 resolving power, detect type: Orbitrap, scan range: 600-2000 *m*/*z*, mass range: normal, AGC target 2E6, normalized AGC target: 500%, max injection time: 50 ms, microscans: 1. The mass spectrometer was operated using a TopN 3 s data-dependent acquisition mode. Precursor ions were filtered by intensity, charge state, and dynamic exclusion. Intensity minimum: 5E3, intensity maximum: 1E20, included charge states: 4-60, include underdetermined charge states: False, dynamic exclusion after n times: 1, dynamic exclusion duration: 60 s, mass tolerance: 0.5 *m*/*z*, exclude isotopes: True. Ions for fragmentation were isolated and fragmented via higher energy dissociation (HCD). Detector type: Orbitrap, isolation mode: quadrupole, resolving power: 60,000, scan range: 350-2000 *m*/*z*, AGC target: 1E6, normalized AGC target: 2000%, max injection time: 600 ms, microscans: 1, isolation window: 3 *m*/*z*, activation type: HCD, collision energy: 32, collision energy mode: fixed.

For analysis on Orbitrap Fusion Lumos mass spectrometer, data was acquired with the following global parameters: spray volage: 1600 V, sweep gas: 0, ion transfer tube temperature: 320 ºC, application mode: Intact Protein, pressure mode: Low Pressure (2 mTorr), Advanced Peak Determination: True, default charge state: 15, S-lens RF: 30%, source collision induced dissociation: 15 eV. Precursor spectra were acquired at 120,000 resolving power (at 200 *m*/*z*), mass range: normal, detector type: Orbitrap, scan range: 600-2000 *m*/*z*, AGC target: 1E6, normalized AGC target: 250%, max injection time: 100 ms, microscans: 4. The mass spectrometer was operated using a Top2 data-dependent acquisition mode. Precursor ions were filtered by intensity, charge state, and dynamic exclusion. Intensity minimum: 2E4, intensity maximum:1E20, included charge states: 6-60, include undetermined charge states: False, dynamic exclusion after n times: 1, dynamic exclusion duration: 60 s, mass tolerance: 1.5 *m*/*z*, exclude isotopes: True. Ions for fragmentation were isolated and fragmented via HCD. Detector type: Orbitrap, isolation mode: quadrupole, resolving power: 60,000 (at 200 *m*/*z*), scan range: 400-2000 *m*/*z*, AGC target: 1E6, normalized AGC target: 2000%, max injection time: 400 ms, microscans: 4, isolation window: 3 *m*/*z*, activation type: HCD, collision energy: 27, collision energy mode: fixed.

Protein and Proteoform Identification

The raw data files were processed with the publicly available workflow on TDPortal (https://portal.nrtdp.northwestern.edu, Code Set 4.0.0) that performed mass inference, searched a database of human proteoforms derived from Swiss-Prot (June 2020) with curated histones, and estimated conservative, context-dependent 1% FDR at the protein, isoform, and proteoform levels.[18] Each tissue type was searched separately with its own FDR context. Aggregated search results were used in further data analysis.

Code and Data Availability

Raw files, mzIdentML, and tdReport files were deposited in Massive (Accession MSV000088565). Search results in tdReport format are viewable using TDViewer – a freeware from Northwestern University (http://topdownviewer.northwestern.edu). Search results were further analyzed, and figures were generated with custom code written for R 4.1.0. Source code for data analysis is available at https://github.com/bdrown/rplc-cze-tissues.

Results and Discussion

Samples were obtained from HuBMAP Tissue Mapping Centers from ten human donors. Tissue was cryopulverized, lysed, and proteins precipitated (**Figure 1**). To increase the depth of proteome coverage, proteins were fractionated with GELFrEE prior to MS analysis. Since we intended to analyze each sample by both CZE and RPLC, we setup two Orbitrap tribrid MS instruments configured with either CZE or RPLC, acquired data for a sample on one system, and immediately acquired data for the same sample on the second one. CZE substantially benefits from a higher scan rate due to generally narrower peak widths. Consequently, the CESI 8000 Plus was hyphenated to the Orbitrap Eclipse while a Dionex nanoLC was coupled to the Orbitrap Fusion Lumos. Three tissue types (heart, small intestine, and spleen) were analyzed by this paired analysis while two tissues (lung and kidney) were analyzed solely by RPLC-MS on the Orbitrap Eclipse (**Table 1**).

Discovery of New Human Proteoforms

By searching the TDP data against a database of human proteoforms using TDPortal and 1% conservative false discovery rate (FDR), a total of 11,466 proteoforms from 740 proteins were identified (**Table 1**). Of these annotations, 8,784 proteoforms and 343 proteins were unique to a single tissue type (**Table 1**, **Figure 2A**). Lung tissue contained the highest number of proteoforms and proteins (overall and unique) while kidney tissue contained the fewest unique proteoforms (**Figure S1**). Despite having the lowest number of proteins identified, spleen tissue had a high number of proteoforms per protein (**Figure S1**). While histones and hemoglobin generated the highest number or proteoforms per protein in most tissues, several other proteins populated the top fifteen proteins (**Figure S2**). Overall, CZE-MS/MS resulted in a higher number of protein and proteoform identification than RPLC (**Figure 2B**). However, the difference in MS instrument performance likely contributed to the increased number of IDs characterized the CZE-MS/MS workflow.

We also sought to compare the proteoforms identified in this work to those reported in prior studies. The Human Proteoform Atlas (HPfA, http://human-proteoform-atlas.org/) is the most comprehensive collection of characterized proteoforms.[14] The HPfA consists of 48 datasets which include numerous studies on immortalized cell lines, one study on healthy human solid tissue,[19] two studies on human cancer tissue,[20, 21] and the Blood Proteoform Atlas.[22] Of the 11,466 proteoforms identified in this study, a substantial number of 7,373 (64.3%) were not previously reported in the HPfA while 4,093 (35.7%) proteoforms were present in this database (**Figure 2C**). The frequency of rediscovery was higher on the protein level with 198 (26.8%) proteins first reported here and 542 (73.2%) proteins included in the HPfA database (**Figure 2C**). Thus, while some proteins were identified for the first time in this study, the majority of new proteoforms are differently-modified forms of proteins which were previously detected by TDP. Presence and absence matrices showed clear clustering of tissue at the proteoform (**Figure 2D**) level demonstrating that proteoform identifications are more characteristic of the tissues under study.

A “bird’s-eye” view of the physicochemical properties of proteoforms identified in the five different tissue types, including hydrophobicity, monoisotopic mass, and pI value, can be found in **Figure 3A and S3**. While kidney, lung, and spleen tissue proteoforms show similar distributions in their violin plots regarding all three investigated characteristics, distinct differences for heart and especially small intestine tissue were detected. For example, in the case of the small intestine, a high number of proteoforms in the pI range of 10.5 to 12.0 was observed, which can be explained by a relative increase in histone proteoforms compared to the other analyzed tissue types. This is also supported by the negative GRAVY score, showing a large distribution at around -0.6. On the other hand, proteoforms observed in heart tissue exhibit a relatively broad distribution of pI values.

Influence of separation technique

While the performance of CZE and RPLC have been compared in numerous contexts,[23-27] the paired analysis of heart, small intestine, and spleen provides an opportunity to explore how proteoforms behave regarding these two separation techniques. Despite requiring similarly long acquisition times, the window of separation for CZE was smaller than RPLC. The difference in separation principle was evident in the relationship between proteoform retention/migration times and mass (**Figure 3B**) as well as time and hydrophobicity (**Figure 3C**). While there is a strong correlation between mass and retention time with RPLC, no significant correlation was observed between mass and migration time with CZE (**Table S3**). Both separation methods demonstrate a correlation between hydrophobicity and time, but RPLC has a stronger correlation. While CZE was performed with an acidic background electrolyte (pH 2.4), we observed a positive correlation between proteoform hydrophobicity and mass-to-charge ratio (**Figure S3I**), which helps to explain the increase in hydrophobicity with migration time (less number of "ionizable" amino acids available per size).

In addition to the physiochemical properties of proteoforms identified using CZE and RPLC differing, the distribution of post-translational modifications (PTMs) was similarly asymmetrical. Twelve PTM categories were identified (**Table 2**), and their identifications differed significantly (Pearson's Chi-squared test, $\chi^2 = 196$, p-value $<2\text{x}10^{-16}$) depending on the fractionation method. Two-by-two Chi-squared tests were performed to determine which PTMs had significant deviations in their identification rates (observed PTM / the sum of all other PTMs) as described previously.[28] Monomethylation, half cystines, and monohydroxylation were elevated on CZE-MS/MS, while on RPLC-MS/MS, detection of monoacetylated and trimethylation proteoforms was enhanced. PTM observation frequencies at the proteoform spectral match level followed the same trends in observation biases (**Table S4**). Summarized, these observations substantiate the benefit of the combination of CZE and RPLC derived data from increasing the coverage of the proteoform discovery workflow.

Tissue-Specific Proteoforms and Handling of PTM Ambiguity

Uncertainty in exact position of a PTM on a proteoform can arise in cases where SwissProt entries have many recorded modifications and amino acid variants and fragmentation data are incomplete to assert an umambiguous level 1 proteoform.[29] This phenomenon is exemplified by cardiac troponin C (cTnC), which was identified in its canonical form (full length, N-terminal acetylated, PFR55232) as a level 1 proteoform (**Figure 4A**). Nine additional proteoforms had sufficiently high proteoform-level Q-scores to pass FDR cutoffs due to excellent sequence coverage in regions without modifications and they were classified as level 3 proteoforms with some PTM site ambiguity (**Figure 4A**). The example of cTnC is not alone; the majority of proteoforms identified in this study are either chemically modified or bear a sequence variant, as only 33% are unmodified (**Figure 4B**). While filtering by C-score can help triage level 3 proteoforms for which PTM localization is ambiguous, the C-score does not help in cases where there is only one possible site of modification.[30]

To curate a core set of proteoforms uniquely expressed in the five individual tissue types, we implemented a conservative process to select those proteoforms with PTMs with direct fragment ion support (level 1 proteoforms[29]). To this end, the number of matching fragment ions that bear a PTM (or amino acid variant) was counted for each proteoform spectral match (PrSM). While many mutated and modified proteoforms have supporting fragment ions (level 1), a disproportionate number of modified proteoforms were level 3 with two or fewer (**Figure 4C, D**). Consequently, the requirement of having $\geq$3 supporting fragment ions for modified proteoforms was added in addition to a C-Score >30. This process culled the set of 8784 unique proteoforms in Table 1 down to 2843 level 1 tissue-specific proteoforms (**Figure 4E, Supplementary Data 1**).

More level 1 tissue-specific proteoforms were identified in a Subsequence search (previously called BioMarker search that identifies portions of full length proteoforms[31, 32]) than in Absolute Mass searches. Specifically, 2,548 proteoforms were identified in Subsequence searching compared to 295 proteoforms identified in Absolute Mass searches. Subsequence searches identify proteolytic fragments that often arise from endogenous proteolytic events and can serve as significant biomarkers.[21] While a portion of these proteoforms may be the product of non-specific proteolysis, the consensus sequence of cleavage sites varied across tissues (**Figure S4**). Truncated proteoforms from the heart, kidney, and small intestine showed enrichment of F, Y, W, and L at P1, which suggests chymotrypsin activity. Spleen proteoforms demonstrated enrichment of hydrophobic residues but no apparent sequence specificity. This lack of specificity combined with a high proteoform to protein ratio agrees well with the role of the spleen for scavenging senescent blood cells.[33] Lung proteoforms had a higher propensity of cysteine at P1, which is not commonly observed for specific proteases. This enrichment was driven by 24 of the 715 lung-specific proteoforms with N-terminal cleavage. Nine of these 24 proteoforms originate from collapsing response mediator protein 2 (CRMP-2, Q16555), with cleavage occurring at C439 (**Figure S5**). CRMP-2 has largely been studied in the context of neurological diseases due to its role in microtubule assembly and axon growth.[34] Indeed, C-terminal truncation of CRMP-2 has been linked to neurodegeneration,[35] and the cleavage site was later localized to S517.[36] As the function of CRMP-2 in lung tissue has only recently begun to be characterized,[37] this novel truncation at C439 may assist in elucidating its role.

Subsequence searching also identified a proteolytic cleavage site in CDGSH iron-sulfur domain-containing protein 1 (mitoNEET, Q9NZ45) at L47 (**Figure S6**). MitoNEET is a mitochondrial outermembrane protein that was initially discovered as an off-target interactor of the PPAR-γ agonist pioglitazone.[38] With its iron-sulfur cluster oriented toward the cytosol, mitoNEET acts as a redox sensor and regulator of mitochondrial iron.[39-41] Downregulation of mitNEET has been associated with aging and increased risk of heart failure.[42] The canonical proteoform of mitoNEET was observed in both small intestine and heart tissue, while both proteolytic products were observed solely in heart tissue (**Figure S6**). Cleavage at L47 does not disrupt the iron-sulfur cluster binding site but does separate this reactive center from the protein's transmembrane domain. Thus, proteolytic cleavage may act as a means of regulating mitoNEET or a mechanism by which full-length mitoNEET abundance declines in aging cardiomyocytes.

Unique Proteoforms Are Reflective of Tissue Central Function

Many of the tissue-specific proteoforms originate from genes involved in the core function of these tissues, as indicated by gene ontology enrichment (**Figure 2E**, **Figure S7**). The Subsequence proteoform search identified a series of proteoforms associated with defensins with distinct expression patterns (**Figure 4F**, **Figure S8**). Defensins are a family of small cationic host defense proteins characterized by three conserved intramolecular disulfide bonds.[43] Six human alpha-defensins have been identified to date and are subdivided into human neutrophil peptides 1 to 4 (HNP1-4) and human (enteric) defensins (HD5-6). HNPs are stored as mature peptides in granules of neutrophils and released upon activation by exocytosis.[44] HNP1 (PFR69106) was identified in both lung and spleen tissue as expected for tissues with high neutrophil content. HNP2 (PFR69109), HNP3 (PFR69079), HNP4 (PFR65983), and truncation products of HNP2 (PFR165182 and PFR165183) were observed exclusively in spleen tissue. No beta-defensin proteoforms were identified. HD5 and HD6 are produced in Paneth cells at the base of small intestinal crypts.[45] Accordingly, HD5 and HD6 were detected exclusively in small intestinal tissue. Unlike other defensins, HD5 is stored as a propeptide, and the fully mature peptides are thought to be produced by intracellular trypsin.[46] Consequently, the HD5 propeptide (PFR165815) and several truncated products were observed. Several of these truncated proteoforms (PFR5737351, PFR97759, and PFR97755) correspond to trypsin cleavage sites (R25, R55, and R62), while others (PFR5741069, PFR5737454, and PFR5737363) seem to correspond to other mechanisms of cleavage considering the residues at the P1 positions (D41, F46, and A61). Despite reducing samples with DTT prior to analysis, several proteoforms were observed with disulfide bridges intact (PFR4919881, PFR4919882, and PFR5026622). The disulfide linkages in these proteoforms are inconsistent with the canonical model of alpha-defensins that includes end-to-end disulfides (**Figure 4G**). Although these non-canonical disulfides might be biologically relevant, spontaneous reformation of disulfides in denatured samples is likely. Defensins are important components of the host innate immunity, so observing new proteoforms on mucosal surfaces is important in understanding their regulation and design of therapeutic mimetics.[47, 48] Furthermore, these findings are a good showcase for the capabilities of the presented setup to evaluate tissue-specific proteoform-related questions.

Glutathione S-transferases are a family of proteins involved in inflammation and the cellular defense against toxic and carcinogenic compounds.[49, 50] Proteoforms from this protein family were broadly observed but with distinct tissue distributions (**Figure S9**). Glutathione S-transferase A1 (P08263) and A2 (P09210) were observed primarily in the small intestine and kidney, respectively. The polymorphism E210A (rs6577) was observed in a single kidney sample (Biorep 3), which was derived from a 53-year-old African American male (**Table S1**). This coding SNP occurs with much higher frequency in Africa Americans (56.5%) compared to the global population (9.9%).[51] Microsomal glutathione S-transferase (MGST) 1, 2, and 3 were observed in the small intestine and lung (1), small intestine and kidney (2), and heart tissue (3), respectively (**Figure S9C & D**). These glutathione transferases are polytopic membrane proteins located in the endoplasmic reticulum membrane with both glutathione conjugation and peroxidase activity.[52, 53] A novel MGST3 proteoform (PFR5719232) that lacks the C-terminal cysteine necessary for *S*-palmitoylation was the predominant form observed in heart tissue.[54]

Enrichment of functionally relevant genes from the identified proteoforms was particularly notable for heart tissue, with terms associated with ATP synthesis and muscle contraction leading the list (**Figure 2E**). Six proteoforms of cardiac phospholamban (PLN), a key regulator of cardiac contraction via inhibition of the sarcoplasmic reticulum calcium pump (SERCA), were identified by RPLC-MS/MS (**Figure 5A**).[55] While unmodified PLN and palmitoylated PLN have both been reported previously,[56] this study is the first report of phosphorylated PLN and combined phosphorylation and palmitoylation. Phosphorylation and palmitoylation of PLN have both been shown to control the impact localization, complexation, and inhibition of SERCA, so accurate measurement of their combination will help clarify PLN's role in health and disease.[57]

We also present evidence for phosphorylation at ~30% stoichiometry of ventricle myosin regulatory light chain (RLC_V). Prior reports by the Ge group have established N-terminal trimethylation of RLC_V and phosphorylation of swine RLC_V, but phosphorylation of human RLC_V was unlocalized and observed at <10% stoichiometry.[58, 59] The removal of N-terminal methionine and trimethylation was confirmed by tandem HCD fragmentation, and the site of phosphorylation was localized to S15, which is analogous to the site identified on swine RLC_V (**Figure 5B**). On a last analytical note, phosproteoforms of cardiac troponin I (cTnI)[60] were not separated by RPLC but were at baseline by CZE (**Figure 5C**); proteoform quantitation by both techniques showed <10% coefficient of variation between them. Better separation of the CZE should translate into better on-the-fly sequence coverage and proteoform characterization with tandem MS scan speeds.

Conclusions

We have described the combination of TDP data collected with online separation by RPLC and CZE to expand the depth of human proteome coverage. All proteomics methods face the challenge of measuring low-abundance analytes, so identifying robust approaches that introduce new proteoform selectivity are highly sought. RPLC and CZE were shown to possess differential proteoform selectivity that manifests as different physiochemical properties and PTM profiles. In a TDP study of five human tissues, we dramatically expanded the number of proteoforms associated with these tissues by combining the two methods.

Confident assignment of proteoforms bearing PTMs or sequence variations becomes more challenging as query proteoforms get larger and the search databases contain more candidate PTM sites. Unambiguous level 1 proteoform assignments are particularly troublesome when seeking proteoforms specific to a particular biological context (e.g., tissue types), but this can be significantly mitigated with the inclusion of fragment-ion data quality standards. Even at current levels of proteoform characterization quality, organ-specific proteoforms achieve robust tissue type identification.

The genes from the tissue-specific proteoforms identified in this study were tied to the core function of the tissues as broadly indicated by GEO analysis. This is further supported by specific examples such as proteins that regulate muscle contractility (PLN, RLCV, cardiac troponins), host-pathogen interaction (defensins), cytoskeletal reorganization (CRMP-2), and metabolic detoxification (family of glutathione transferases). In many cases, these unique proteoforms were detected with only one of the upfront separation methods. Thus, proper exploration of our hypothesis that proteoform-level measurements more fully capture biological context than protein-level measurement requires an increased depth of proteome coverage.

ASSOCIATED CONTENT

Supporting Information.

The following files are available free of charge.

Additional experimental results and figures (PDF)

List of tissue-specific proteoform identified in this study (XLSX)

AUTHOR INFORMATION

Corresponding Author

*Neil L. Kelleher – Department of Molecular Biosciences, Chemistry, and the Feinberg School of Medicine, Northwestern University, Evanston, IL, USA; Email: n-kelleher@northwestern.edu.

Author Contributions

Data acquisition was performed by B.S.D., K.J., and R.D.M with support from C.L.J. Data analysis and visualization was performed by B.S.D. with additional input from K.J. and N.L.K. J.M.C and N.L.K. collected funding support. B.S.D., K.J., R.D.M., and N.L.K. wrote and edited the manuscript. All authors critically reviewed and given approval to the final version of the manuscript.

Funding Sources

This material is based upon the work supported by Human Biomolecular Atlas Program (HuBMAP) grant UH3 CA246635-02 (N.L.K.), National Institute of General Medical Sciences of the National Institutes of Health grant P41 GM108569 (N.L.K.), National Institute of Cancer of the National Institutes of Health grant F32 CA246894 (B.S.D.).

Notes

Competing interests: NLK is involved in entrepreneurial activities in top-down proteomics and consults for Thermo Fisher Scientific.

ACKNOWLEDGMENT

The authors thank the principal investigators at HuBMAP Tissue Mapping Centers Mark Atkinson (University of Florida), Shin Lin (University of Washington), Mike Snyder (Stanford University), Jeff Spraggins (Vanderbilt University), and Gloria Pryhuber (University of Rochester) for contributing tissue samples. We thank SCIEX for their support throughout this research project.

ABBREVIATIONS

BCA, bicinchoninic acid; CZE, capillary zone electrophoresis; CRMP-2, collapsing response mediator protein 2; cTnC, cardiac troponin C; cTnI, cardiac troponin I; cTnT, cardiac troponin T; DTT, dithiothreitol; FDR, false-discovery rate; GELFrEE, Gel-Eluted Liquid FRaction Entrapment Electrophoresis; HAc, acetic acid; HCD, higher-energy collisional dissociation; HD, human enteric defensin; HNP, human neutrophil defensin peptide; HPfA, Human Proteoform Atlas; HuBMAP, Human BioMolecular Atlas Program; MGST, micrisimal glutathione S-transferase; MS, mass spectrometry; PLN, phospholamban; PPARγ, peroxisome proliferator-activated receptor gamma; PTM, post-translational modification; RLC_V, ventricle myosin regulatory light chain; RPLC, reversed phase liquid chromatography; SERCA, sarcoplasmic reticulum calcium pump; SNP, single nucleotide polymorphism; TDMS, top-down mass spectrometry; TDP, top-down proteomics; TMC, tissue mapping center.

REFERENCES

1. Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P., Proteoform: a single term describing protein complexity. *Nat Methods* **2013,** *10* (3), 186-7.
2. Consortium, H., The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **2019,** *574* (7777), 187-192.
3. Regev, A.; Teichmann, S. A.; Lander, E. S.; Amit, I.; Benoist, C.; Birney, E.; Bodenmiller, B.; Campbell, P.; Carninci, P.; Clatworthy, M.; Clevers, H.; Deplancke, B.; Dunham, I.; Eberwine, J.; Eils, R.; Enard, W.; Farmer, A.; Fugger, L.; Gottgens, B.; Hacohen, N.; Haniffa, M.; Hemberg, M.; Kim, S.; Klenerman, P.; Kriegstein, A.; Lein, E.; Linnarsson, S.; Lundberg, E.; Lundeberg, J.; Majumder, P.; Marioni, J. C.; Merad, M.; Mhlanga, M.; Nawijn, M.; Netea, M.; Nolan, G.; Pe'er, D.; Phillipakis, A.; Ponting, C. P.; Quake, S.; Reik, W.; Rozenblatt-Rosen, O.; Sanes, J.; Satija, R.; Schumacher, T. N.; Shalek, A.; Shapiro, E.; Sharma, P.; Shin, J. W.; Stegle, O.; Stratton, M.; Stubbington, M. J. T.; Theis, F. J.; Uhlen, M.; van Oudenaarden, A.; Wagner, A.; Watt, F.; Weissman, J.; Wold, B.; Xavier, R.; Yosef, N.; Human Cell Atlas Meeting, P., The Human Cell Atlas. *Elife* **2017,** *6*.
4. Neumann, E. K.; Patterson, N. H.; Allen, J. L.; Migas, L. G.; Yang, H.; Brewer, M.; Anderson, D. M.; Harvey, J.; Gutierrez, D. B.; Harris, R. C.; deCaestecker, M. P.; Fogo, A. B.; Van de Plas, R.; Caprioli, R. M.; Spraggins, J. M., Protocol for multimodal analysis of human kidney tissue by imaging mass spectrometry and CODEX multiplexed immunofluorescence. *STAR Protoc* **2021,** *2* (3), 100747.
5. Saka, S. K.; Wang, Y.; Kishi, J. Y.; Zhu, A.; Zeng, Y.; Xie, W.; Kirli, K.; Yapp, C.; Cicconet, M.; Beliveau, B. J.; Lapan, S. W.; Yin, S.; Lin, M.; Boyden, E. S.; Kaeser, P. S.; Pihan, G.; Church, G. M.; Yin, P., Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. *Nat Biotechnol* **2019,** *37* (9), 1080-1090.
6. Giesen, C.; Wang, H. A.; Schapiro, D.; Zivanovic, N.; Jacobs, A.; Hattendorf, B.; Schuffler, P. J.; Grolimund, D.; Buhmann, J. M.; Brandt, S.; Varga, Z.; Wild, P. J.; Gunther, D.; Bodenmiller, B., Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* **2014,** *11* (4), 417-22.
7. Keren, L.; Bosse, M.; Thompson, S.; Risom, T.; Vijayaragavan, K.; McCaffrey, E.; Marquez, D.; Angoshtari, R.; Greenwald, N. F.; Fienberg, H.; Wang, J.; Kambham, N.; Kirkwood, D.; Nolan, G.; Montine, T. J.; Galli, S. J.; West, R.; Bendall, S. C.; Angelo, M., MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci Adv* **2019,** *5* (10), eaax5851.
8. Ptacek, J.; Locke, D.; Finck, R.; Cvijic, M. E.; Li, Z.; Tarolli, J. G.; Aksoy, M.; Sigal, Y.; Zhang, Y.; Newgren, M.; Finn, J., Multiplexed ion beam imaging (MIBI) for characterization of the tumor microenvironment across tumor types. *Lab Invest* **2020,** *100* (8), 1111-1123.
9. Toby, T. K.; Abecassis, M.; Kim, K.; Thomas, P. M.; Fellers, R. T.; LeDuc, R. D.; Kelleher, N. L.; Demetris, J.; Levitsky, J., Proteoforms in Peripheral Blood Mononuclear Cells

as Novel Rejection Biomarkers in Liver Transplant Recipients. *Am J Transplant* **2017,** *17* (9), 2458-2467.

10. Seckler, H. D. S.; Fornelli, L.; Mutharasan, R. K.; Thaxton, C. S.; Fellers, R.; Daviglus, M.; Sniderman, A.; Rader, D.; Kelleher, N. L.; Lloyd-Jones, D. M.; Compton, P. D.; Wilkins, J. T., A Targeted, Differential Top-Down Proteomic Methodology for Comparison of ApoA-I Proteoforms in Individuals with High and Low HDL Efflux Capacity. *J Proteome Res* **2018,** *17* (6), 2156-2164.

11. Smith, L. M.; Agar, J. N.; Chamot-Rooke, J.; Danis, P. O.; Ge, Y.; Loo, J. A.; Pasa-Tolic, L.; Tsybin, Y. O.; Kelleher, N. L.; Consortium for Top-Down, P., The Human Proteoform Project: Defining the human proteome. *Sci Adv* **2021,** *7* (46), eabk0734.

12. Lee, J. E.; Kellie, J. F.; Tran, J. C.; Tipton, J. D.; Catherman, A. D.; Thomas, H. M.; Ahlf, D. R.; Durbin, K. R.; Vellaichamy, A.; Ntai, I.; Marshall, A. G.; Kelleher, N. L., A robust two-dimensional separation for top-down tandem mass spectrometry of the low-mass proteome. *J Am Soc Mass Spectrom* **2009,** *20* (12), 2183-91.

13. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; t Hoen, P. A.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S. A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B., The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **2016,** *3*, 160018.

14. Hollas, M. A. R.; Robey, M. T.; Fellers, R.; LeDuc, R. D.; Thomas, P. M.; Kelleher, N. L., The Human Proteoform Atlas: a FAIR community resource for experimentally derived proteoforms. *Nucleic Acids Research* **2021**, gkab1086.

15. Toby, T. K.; Fornelli, L.; Srzentic, K.; DeHart, C. J.; Levitsky, J.; Friedewald, J.; Kelleher, N. L., A comprehensive pipeline for translational top-down proteomics from a single blood draw. *Nat Protoc* **2019,** *14* (1), 119-152.

16. Wessel, D.; Flugge, U. I., A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **1984,** *138* (1), 141-3.

17. Fornelli, L.; Durbin, K. R.; Fellers, R. T.; Early, B. P.; Greer, J. B.; LeDuc, R. D.; Compton, P. D.; Kelleher, N. L., Advancing Top-down Analysis of the Human Proteome Using a Benchtop Quadrupole-Orbitrap Mass Spectrometer. *J Proteome Res* **2017,** *16* (2), 609-618.

18. LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Shams, D. P.; Thomas, P. M.; Kelleher, N. L., Accurate Estimation of Context-Dependent False Discovery Rates in Top-Down Proteomics. *Mol Cell Proteomics* **2019,** *18* (4), 796-805.

19. Chen, Y. C.; Sumandea, M. P.; Larsson, L.; Moss, R. L.; Ge, Y., Dissecting human skeletal muscle troponin proteoforms by top-down mass spectrometry. *J Muscle Res Cell Motil* **2015,** *36* (2), 169-81.

20. Ntai, I.; Fornelli, L.; DeHart, C. J.; Hutton, J. E.; Doubleday, P. F.; LeDuc, R. D.; van Nispen, A. J.; Fellers, R. T.; Whiteley, G.; Boja, E. S.; Rodriguez, H.; Kelleher, N. L., Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk. *Proc Natl Acad Sci U S A* **2018,** *115* (16), 4140-4145.

21. Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D.; Ellis, M. J.; Ruggles, K. V.; Fenyo, D.; Boja, E. S.; Rodriguez, H.; Townsend, R. R.; Kelleher, N. L., Integrated Bottom-Up and Top-Down Proteomics of Patient-Derived Breast Tumor Xenografts. *Mol Cell Proteomics* **2016,** *15* (1), 45-56.

22. Melani, R. D.; Gerbasi, V. R.; Anderson, L. C.; Sikora, J. W.; Toby, T. K.; Hutton, J. E.; Butcher, D. S.; Negrao, F.; Seckler, H. D. S.; Srzentic, K.; Fornelli, L.; Camarillo, J. M.; LeDuc, R. D.; Cesnik, A. J.; Lundberg, E.; Greer, J. B.; Fellers, R.; Robey, M. T.; DeHart, C. J.; Forte, E.; Hendrickson, C. L.; Abbatiello, S. E.; Thomas, P. M.; Kokaji, A. I.; Levitsky, J.; Kelleher, N. L., The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* **2022**.

23. Faserl, K.; Sarg, B.; Kremser, L.; Lindner, H., Optimization and evaluation of a sheathless capillary electrophoresis-electrospray ionization mass spectrometry platform for peptide analysis: comparison to liquid chromatography-electrospray ionization mass spectrometry. *Anal Chem* **2011,** *83* (19), 7297-305.

24. Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J., Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry as an alternative proteomics platform to ultraperformance liquid chromatography-electrospray ionization-tandem mass spectrometry for samples of intermediate complexity. *Anal Chem* **2012,** *84* (3), 1617-22.

25. McCool, E. N.; Liangliang, S., Comparing nanoflow reversed-phase liquid chromatography-tandem mass spectrometry and capillary zone electrophoresis-tandem mass spectrometry for top-down proteomics. *Se Pu* **2019,** *37* (8), 878-886.

26. Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates, J. R., 3rd, In-line separation by capillary electrophoresis prior to analysis by top-down mass spectrometry enables sensitive characterization of protein complexes. *J Proteome Res* **2014,** *13* (12), 6078-86.

27. Nowak, P. M.; Sekuła, E.; Kościelniak, P., Assessment and Comparison of the Overall Analytical Potential of Capillary Electrophoresis and High-Performance Liquid Chromatography Using the RGB Model: How Much Can We Find Out? *Chromatographia* **2020,** *83* (9), 1133-1144.

28. Latta, S. C.; Howell, C. A.; Dettling, M. D.; Cormier, R. L., Use of data on avian demographics and site persistence during overwintering to assess quality of restored riparian habitat. *Conserv Biol* **2012,** *26* (3), 482-92.

29. 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.; Pasa-Tolic, L.; Robinson, C. V.; Schluter, H.; Tsybin, Y. O.; Vilaseca, M.; Vizcaino, J. A.; Danis, P. O.; Kelleher, N. L., A five-level classification system for proteoform identifications. *Nat Methods* **2019,** *16* (10), 939-940.

30. LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Thomas, P. M.; Kelleher, N. L., The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. *J Proteome Res* **2014,** *13* (7), 3231-40.

31. Leduc, R. D.; Kelleher, N. L., Using ProSight PTM and related tools for targeted protein identification and characterization with high mass accuracy tandem MS data. *Curr Protoc Bioinformatics* **2007,** *Chapter 13*, Unit 13 6.

32. Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L., ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W701-6.
33. Klei, T. R.; Meinderts, S. M.; van den Berg, T. K.; van Bruggen, R., From the Cradle to the Grave: The Role of Macrophages in Erythropoiesis and Erythrophagocytosis. *Front Immunol* **2017,** *8*, 73.
34. Zhang, J. N.; Michel, U.; Lenz, C.; Friedel, C. C.; Koster, S.; d'Hedouville, Z.; Tonges, L.; Urlaub, H.; Bahr, M.; Lingor, P.; Koch, J. C., Calpain-mediated cleavage of collapsin response mediator protein-2 drives acute axonal degeneration. *Sci Rep* **2016,** *6*, 37050.
35. Taghian, K.; Lee, J. Y.; Petratos, S., Phosphorylation and cleavage of the family of collapsin response mediator proteins may play a central role in neurodegeneration after CNS trauma. *J Neurotrauma* **2012,** *29* (9), 1728-35.
36. Shinkai-Ouchi, F.; Yamakawa, Y.; Hara, H.; Tobiume, M.; Nishijima, M.; Hanada, K.; Hagiwara, K., Identification and structural analysis of C-terminally truncated collapsin response mediator protein-2 in a murine model of prion diseases. *Proteome Sci* **2010,** *8*, 53.
37. Morales, X.; Pelaez, R.; Garasa, S.; Ortiz de Solorzano, C.; Rouzaut, A., CRMP2 as a Candidate Target to Interfere with Lung Cancer Cell Migration. *Biomolecules* **2021,** *11* (10).
38. Colca, J. R.; McDonald, W. G.; Waldon, D. J.; Leone, J. W.; Lull, J. M.; Bannow, C. A.; Lund, E. T.; Mathews, W. R., Identification of a novel mitochondrial protein ("mitoNEET") cross-linked specifically by a thiazolidinedione photoprobe. *Am J Physiol Endocrinol Metab* **2004,** *286* (2), E252-60.
39. Kusminski, C. M.; Holland, W. L.; Sun, K.; Park, J.; Spurgin, S. B.; Lin, Y.; Askew, G. R.; Simcox, J. A.; McClain, D. A.; Li, C.; Scherer, P. E., MitoNEET-driven alterations in adipocyte mitochondrial activity reveal a crucial adaptive process that preserves insulin sensitivity in obesity. *Nat Med* **2012,** *18* (10), 1539-49.
40. Habener, A.; Chowdhury, A.; Echtermeyer, F.; Lichtinghagen, R.; Theilmeier, G.; Herzog, C., MitoNEET Protects HL-1 Cardiomyocytes from Oxidative Stress Mediated Apoptosis in an In Vitro Model of Hypoxia and Reoxygenation. *PLoS One* **2016,** *11* (5), e0156054.
41. Wiley, S. E.; Paddock, M. L.; Abresch, E. C.; Gross, L.; van der Geer, P.; Nechushtai, R.; Murphy, A. N.; Jennings, P. A.; Dixon, J. E., The outer mitochondrial membrane protein mitoNEET contains a novel redox-active 2Fe-2S cluster. *J Biol Chem* **2007,** *282* (33), 23745-9.
42. Furihata, T.; Takada, S.; Kakutani, N.; Maekawa, S.; Tsuda, M.; Matsumoto, J.; Mizushima, W.; Fukushima, A.; Yokota, T.; Enzan, N.; Matsushima, S.; Handa, H.; Fumoto, Y.; Nio-Kobayashi, J.; Iwanaga, T.; Tanaka, S.; Tsutsui, H.; Sabe, H.; Kinugawa, S., Cardiac-specific loss of mitoNEET expression is linked with age-related heart failure. *Commun Biol* **2021,** *4* (1), 138.
43. Xu, D.; Lu, W., Defensins: A Double-Edged Sword in Host Immunity. *Front Immunol* **2020,** *11*, 764.
44. Faurschou, M.; Sorensen, O. E.; Johnsen, A. H.; Askaa, J.; Borregaard, N., Defensin-rich granules of human neutrophils: characterization of secretory properties. *Biochim Biophys Acta* **2002,** *1591* (1-3), 29-35.
45. Sankaran-Walters, S.; Hart, R.; Dills, C., Guardians of the Gut: Enteric Defensins. *Front Microbiol* **2017,** *8*, 647.

46. Ghosh, D.; Porter, E.; Shen, B.; Lee, S. K.; Wilk, D.; Drazba, J.; Yadav, S. P.; Crabb, J. W.; Ganz, T.; Bevins, C. L., Paneth cell trypsin is the processing enzyme for human defensin-5. *Nat Immunol* **2002,** *3* (6), 583-90.
47. Varney, K. M.; Bonvin, A. M.; Pazgier, M.; Malin, J.; Yu, W.; Ateh, E.; Oashi, T.; Lu, W.; Huang, J.; Diepeveen-de Buin, M.; Bryant, J.; Breukink, E.; Mackerell, A. D., Jr.; de Leeuw, E. P., Turning defense into offense: defensin mimetics as novel antibiotics targeting lipid II. *PLoS Pathog* **2013,** *9* (11), e1003732.
48. Pachon-Ibanez, M. E.; Smani, Y.; Pachon, J.; Sanchez-Cespedes, J., Perspectives for clinical use of engineered human host defense antimicrobial peptides. *FEMS Microbiol Rev* **2017,** *41* (3), 323-342.
49. Mannervik, B.; Awasthi, Y. C.; Board, P. G.; Hayes, J. D.; Di Ilio, C.; Ketterer, B.; Listowsky, I.; Morgenstern, R.; Muramatsu, M.; Pearson, W. R.; et al., Nomenclature for human glutathione transferases. *Biochem J* **1992,** *282 (Pt 1)*, 305-6.
50. Oakley, A., Glutathione transferases: a structural perspective. *Drug Metab Rev* **2011,** *43* (2), 138-51.
51. Phan, L.; Jin, Y.; Zhang, H.; Qiang, W.; Shekhtman, E.; Shao, D.; Revoe, D.; Villamarin, R.; Ivanchenko, E.; Kimura, M.; Wang, Z. Y.; Hao, L.; Sharopova, N.; Bihan, M.; Sturcke, A.; Lee, M.; Popova, N.; Wu, W.; Bastiani, C.; Ward, M.; Holmes, J. B.; Lyoshin, V.; Kaur, K.; Moyer, E.; Feolo, M.; Kattman, B. L. ALFA: Allele Frequency Aggregator. www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/ (accessed 12/13/2021).
52. Jakobsson, P. J.; Mancini, J. A.; Riendeau, D.; Ford-Hutchinson, A. W., Identification and characterization of a novel microsomal enzyme with glutathione-dependent transferase and peroxidase activities. *J Biol Chem* **1997,** *272* (36), 22934-9.
53. Morgenstern, R.; Zhang, J.; Johansson, K., Microsomal glutathione transferase 1: mechanism and functional roles. *Drug Metab Rev* **2011,** *43* (2), 300-6.
54. Forrester, M. T.; Hess, D. T.; Thompson, J. W.; Hultman, R.; Moseley, M. A.; Stamler, J. S.; Casey, P. J., Site-specific analysis of protein S-acylation by resin-assisted capture. *J Lipid Res* **2011,** *52* (2), 393-8.
55. Frank, K.; Kranias, E. G., Phospholamban and cardiac contractility. *Ann Med* **2000,** *32* (8), 572-8.
56. Brown, K. A.; Chen, B.; Guardado-Alvarez, T. M.; Lin, Z.; Hwang, L.; Ayaz-Guner, S.; Jin, S.; Ge, Y., A photocleavable surfactant for top-down proteomics. *Nat Methods* **2019,** *16* (5), 417-420.
57. Zhou, T.; Li, J.; Zhao, P.; Liu, H.; Jia, D.; Jia, H.; He, L.; Cang, Y.; Boast, S.; Chen, Y. H.; Thibault, H.; Scherrer-Crosbie, M.; Goff, S. P.; Li, B., Palmitoyl acyltransferase Aph2 in cardiac function and the development of cardiomyopathy. *Proc Natl Acad Sci U S A* **2015,** *112* (51), 15666-71.
58. Gregorich, Z. R.; Cai, W.; Lin, Z.; Chen, A. J.; Peng, Y.; Kohmoto, T.; Ge, Y., Distinct sequences and post-translational modifications in cardiac atrial and ventricular myosin light chains revealed by top-down mass spectrometry. *J Mol Cell Cardiol* **2017,** *107*, 13-21.
59. Cai, W.; Hite, Z. L.; Lyu, B.; Wu, Z.; Lin, Z.; Gregorich, Z. R.; Messer, A. E.; McIlwain, S. J.; Marston, S. B.; Kohmoto, T.; Ge, Y., Temperature-sensitive sarcomeric protein post-translational modifications revealed by top-down proteomics. *J Mol Cell Cardiol* **2018,** *122*, 11-22.
60. Zhang, J.; Guy, M. J.; Norman, H. S.; Chen, Y. C.; Xu, Q.; Dong, X.; Guner, H.; Wang, S.; Kohmoto, T.; Young, K. H.; Moss, R. L.; Ge, Y., Top-down quantitative proteomics

identified phosphorylation of cardiac troponin I as a candidate biomarker for chronic heart failure. *J Proteome Res* **2011,** *10* (9), 4054-65.

Table 1. Proteins and proteoforms identified from sampling 5 human tissue types.

Tissue Type	Biological Replicates[a]	Separation	MS/MS runs	Proteins 1% FDR[b]	Unique proteins 1% FDR[c]	Proteoforms 1% FDR (C-score >30)	Unique proteoforms (C-score >30)
Lung	3	RPLC	49	437	132	5,566 (2,940)	3,601 (1,462)
Kidney	5	RPLC	42	307	62	2,278 (988)	641 (306)
Heart	2	CZE, RPLC	72	305	70	2,897 (1,346)	1,623 (772)
Small intestine	1	CZE, RPLC	36	305	43	3,101 (1,214)	2,049 (643)
Spleen	1	CZE, RPLC	35	213	36	1,869 (972)	870 (589)
Total	12	-	234	1,567	343	15,711 (7,460)	8,784 (3,772)
Total non-redundant[d]	12	-	234	740	343	11,466 (4,906)	8,784 (3,772)

[a] Biological replicate refers to a sample from a single human being. Sample descriptions and metadata are shown in Table S1.

[b] The term ‘protein’ refers to that SwissProt entry mapping to a single human gene

[c] Unique identifications refer to proteins or proteoforms that were only identified in the tissue type indicated.

[d] Proteins and proteoforms that were observed in more than one human tissue type are counted once in non-redundant totals.

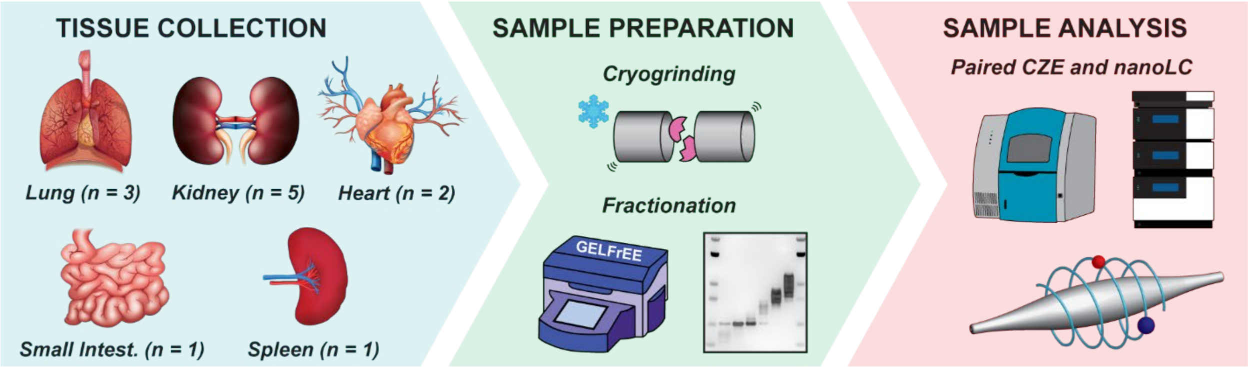


Figure 1. Top-down proteomics of healthy human tissues. Tissues were obtained from HuBMAP Tissue Mapping Centers. Fresh-frozen tissue was cryogenically pulverized, lysed and precipitated. Intact proteins were pre-fractioned using GELFrEE. Each sample was analyzed by CZE-MS/MS and RPLC-MS/MS, respectively.

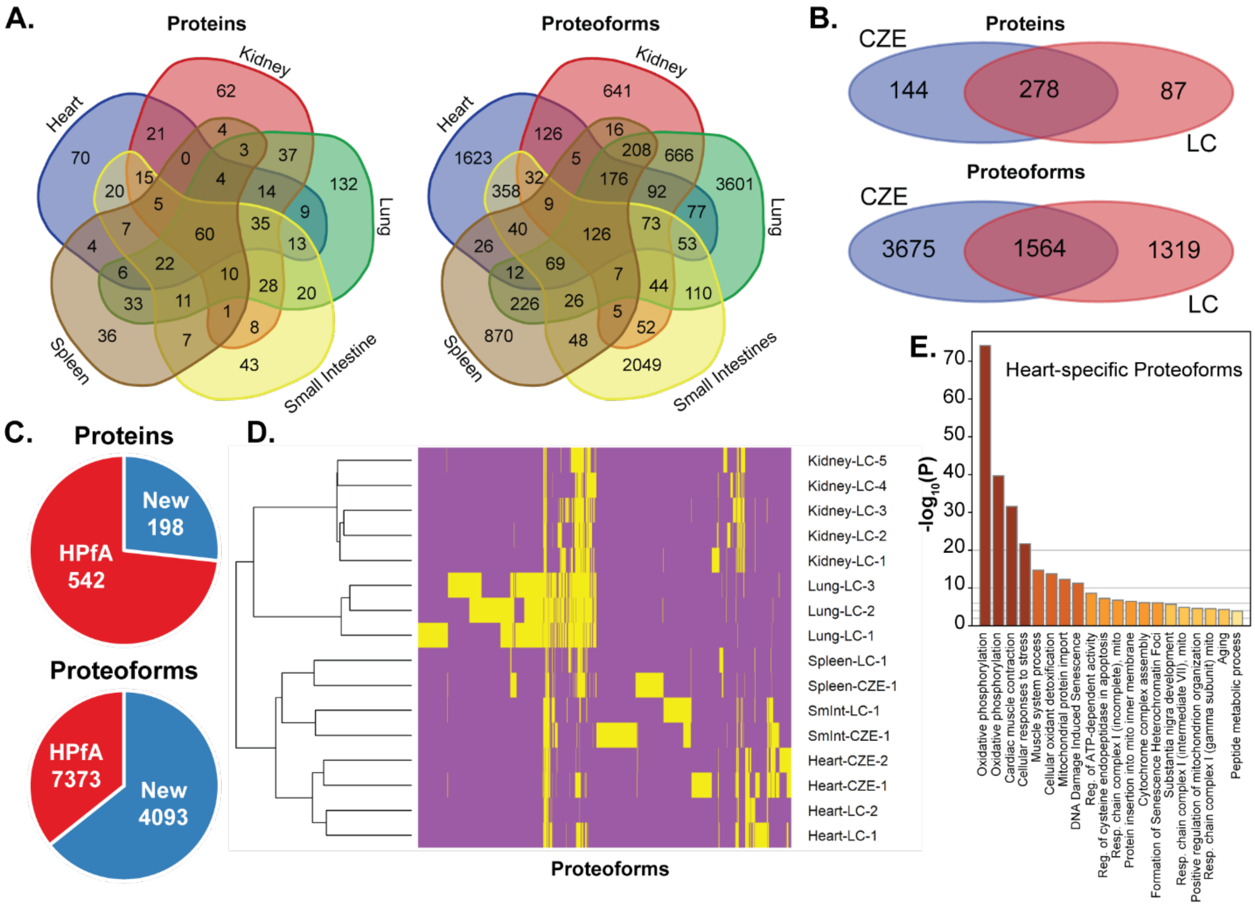


Figure 2. Systematic discovery of unique proteoforms across human tissues. **A.** Venn diagrams of shared and unique proteins and proteoforms identified in each tissue. 1% FDR filtering was applied at the PrSM, proteoform, and protein levels for each tissue. **B.** Venn diagrams of shared and unique proteins and proteoforms identified in heart, small intestine, and/or spleen tissues by either capillary zone electrophoresis or reverse-phase liquid chromatography. **C.** Pie charts representing the rediscovery of proteoforms and proteins previously deposited in the Human Proteoform Atlas (HPfA, red) or only this study (New, blue). HPfA was accessed on 8/18/2021. **D.** Heatmap showing presence (yellow) and absence (purple) of proteoforms in each tissue sample with hierarchical clustering. **E.** Bar graph of top twenty enriched terms from genes associated with proteoforms uniquely identified in heart tissue using Metascape.

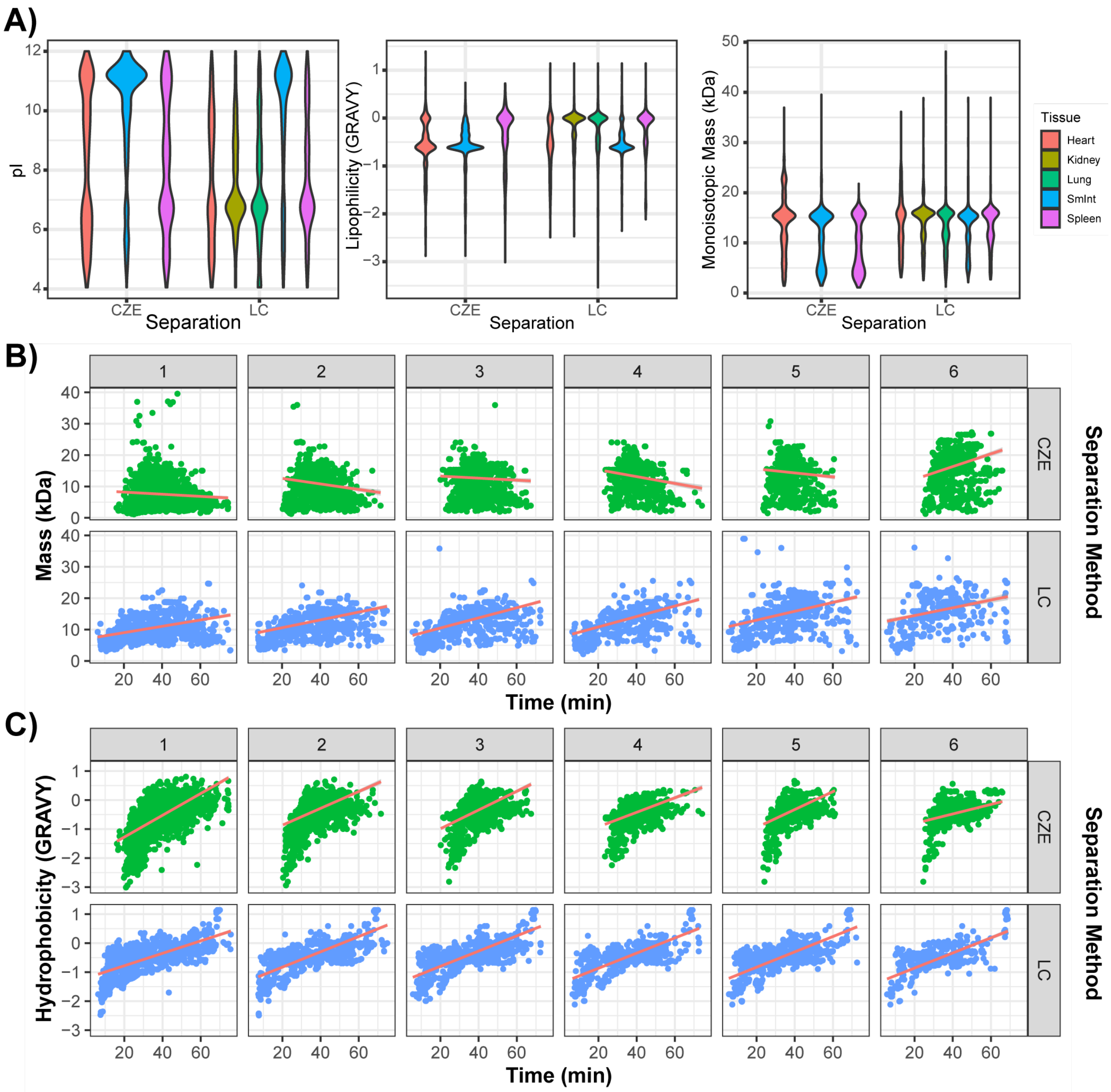


Figure 3. Complimentary separation of intact proteins by CZE and RP-nanoLC. **A**. Violin plots of proteoform physiochemical properties by tissue and separation technique. **B.** Scatterplots relating migration/retention time to monoisotopic mass of proteoforms from heart, small intestine, and spleen samples subdivided by separation method and GELFrEE fraction. **C.** Scatterplots relating migration/retention time to GRAVY score of proteoforms from heart, small intestine, and spleen samples subdivided by separation method and GELFrEE fraction. Corresponding correlation coefficients of data presented in panels B and C are listed in **Table S3**.

Table 2. Frequency of observation for different types of post-translational modifications on identified proteoforms categorized by separation technique used in top-down proteomics.

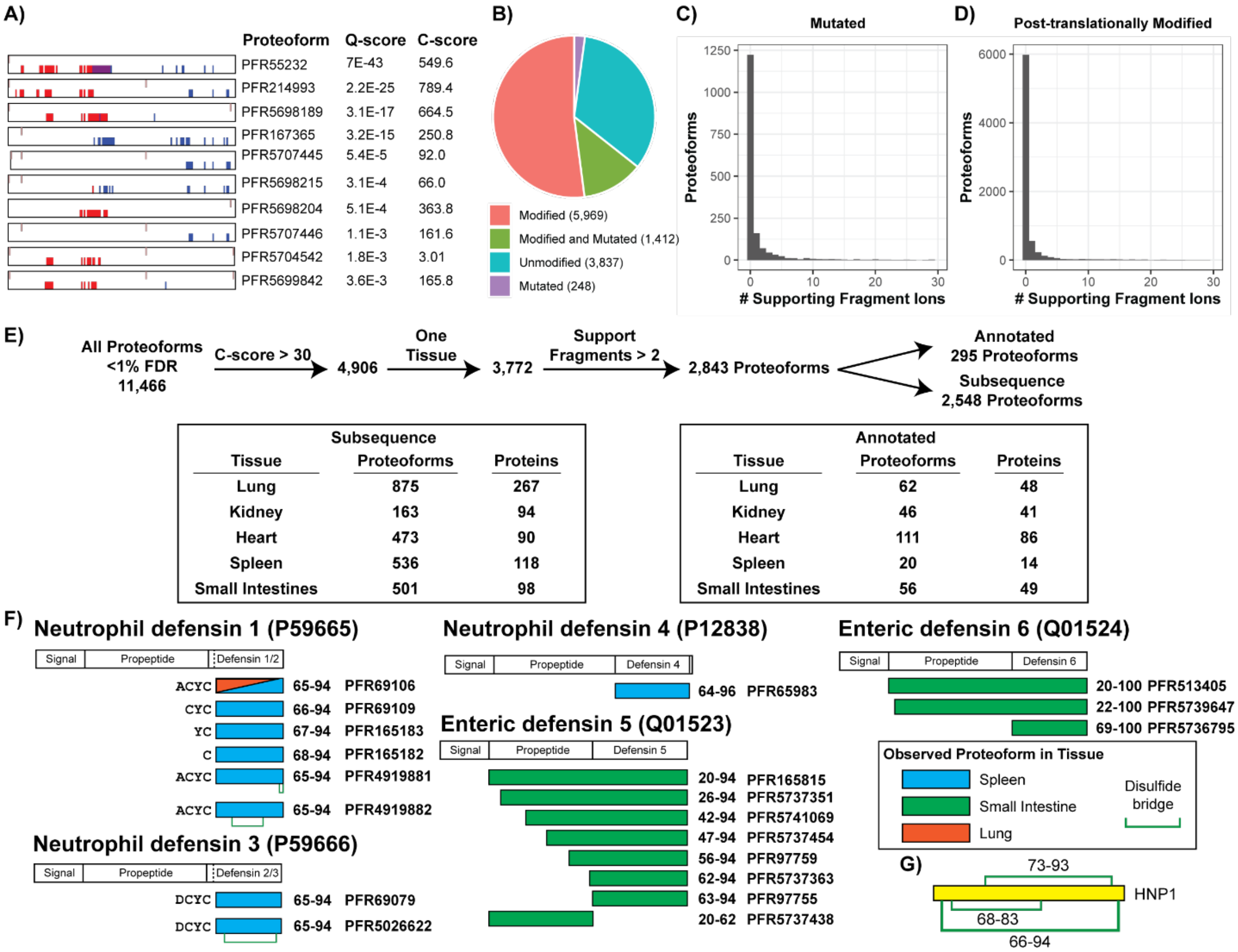
	CZE		RPLC			
PTM type	Observed[a]	Freq.[b]	Observed[a]	Freq.[b]	χ^2	p-value[c]
Monoacetylation[d]	2,723	0.26	1,984	0.31	54	2.6×10^{-12}
Unmodified[d]	2,298	0.22	1,123	0.18	44	4.3×10^{-10}
Phosphorylation	1,644	0.16	1,006	0.16	0.057	9.7
Monomethylation[d]	1,201	0.11	556	0.088	31	3.6×10^{-7}
Trimethylation[d]	920	0.088	667	0.11	14	2.8×10^{-3}
Dimethylation	919	0.088	642	0.10	8.3	4.9×10^{-2}
Half cystine[d]	360	0.034	118	0.019	35	3.8×10^{-8}
Nitrosylation	239	0.023	165	0.026	1.6	2.5
Monohydroxylation[d]	72	0.0069	5	7.9×10^{-4}	31	3.4×10^{-7}
Pyruvic acid iminylated residue	48	0.0046	41	0.0065	2.3	1.6
Deamidated L-asparagine	42	0.0040	38	0.0060	2.9	1.1
S-palmitoylation	14	0.0013	7	0.0011	0.037	10
Total	10,480		6,352			

[a] Number of modifications observed on proteoforms at 1%; count does not include N-terminal and C-terminal modifications; multiple PTMs on the same proteoform are counted multiple times.

[b] Number of observations/sum of PTM observations for each separation technique.

[c] Bonferroni corrected p-value (n = 12)

[d] Statistically significant difference (alpha <0.01) in frequency of observation.



Proteoform	Q-score	C-score
PFR55232	7E-43	549.6
PFR214993	2.2E-25	789.4
PFR5698189	3.1E-17	664.5
PFR167365	3.2E-15	250.8
PFR5707445	5.4E-5	92.0
PFR5698215	3.1E-4	66.0
PFR5698204	5.1E-4	363.8
PFR5707446	1.1E-3	161.6
PFR5704542	1.8E-3	3.01
PFR5699842	3.6E-3	165.8

Tissue	Subsequence Proteoforms	Proteins
Lung	875	267
Kidney	163	94
Heart	473	90
Spleen	536	118
Small Intestines	501	98

Tissue	Annotated Proteoforms	Proteins
Lung	62	48
Kidney	46	41
Heart	111	86
Spleen	20	14
Small Intestines	56	49

Figure 4. Selection of tissue-specific proteoforms. **A.** Cigar depiction of cardiac troponin C proteoforms identified by in human heart tissue. Red, blue, and purple marks on the bottom of cigars indicate b, y, and both b and y fragment ions. Tan marks on top of cigars indicate the presence of PTM or sequence variant. **B.** Distribution of proteoforms identified with PTMs or sequence variance. Proteolytic cleavage and N-terminal acetylation are excluded from consideration as PTMs in this panel. **C.** Histogram of proteoforms and the number of matching fragment ions that support the presence of a sequence variant (*e.g.*, a polymorphism). **D.** Histogram of proteoforms and the number of matching fragment ions that support the presence of a PTM. **E.** Sequential filtering of proteoforms to identify high-confidence tissue-specific proteoforms. **F.**

Identification of tissue-specific defensin proteoforms. **G.** Canonical disulfide bridge structure for alpha defensins.

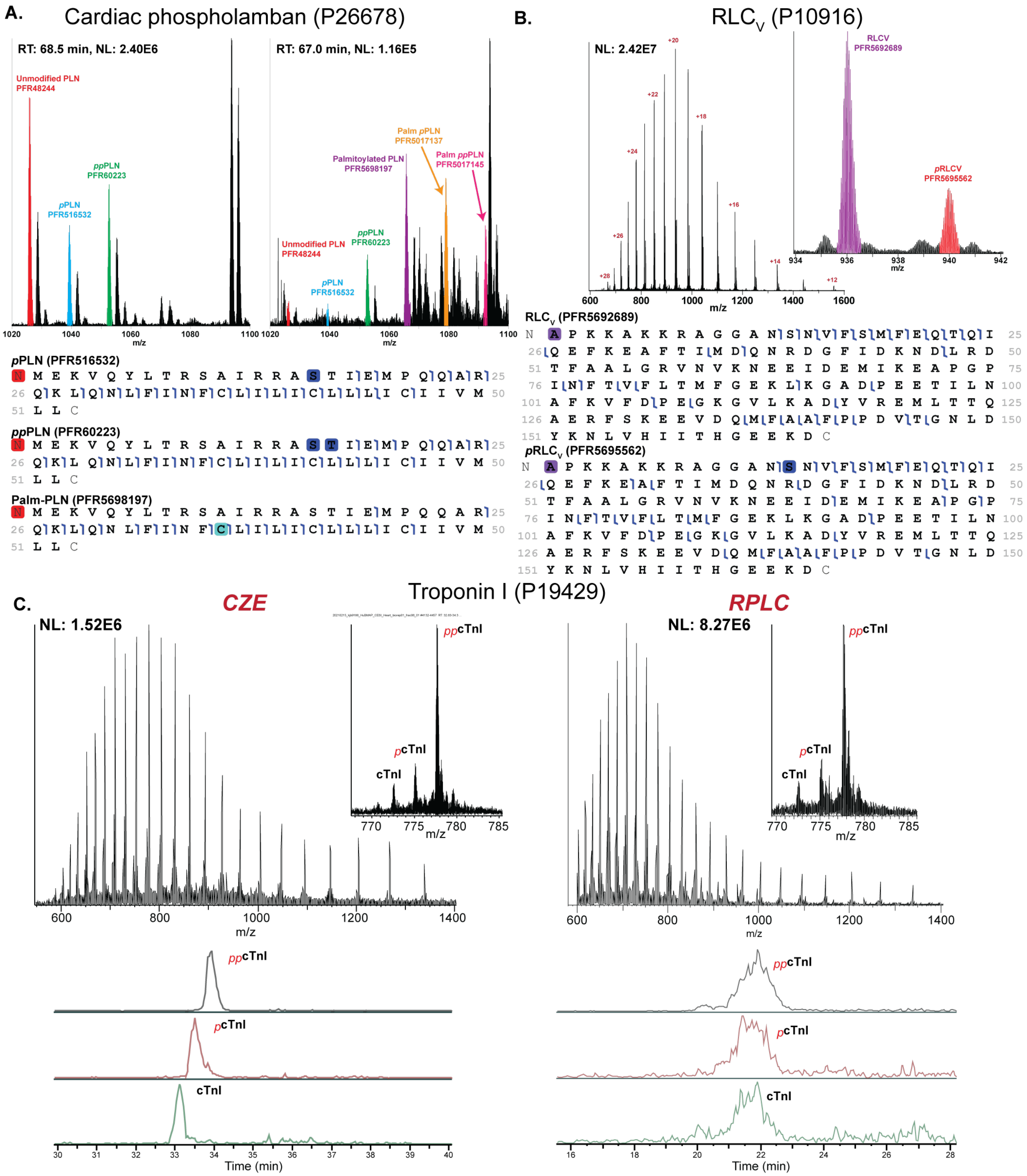


Figure 5. Unique cardio-proteoforms identified in paired RPLC/CZE-MS/MS analysis. **A.** Phosphorylated and palmitoylated proteoforms of phospholamban (PLN, P26678) were observed by RPLC-MS/MS late in the chromatogram. **B.** Phosphorylation of ventricular myosin regulatory light chain (RLCV, P10916). HCD fragmentation precisely localized the phosphorylation to S15.

C. Cardiac troponin I (cTnI, P19429) was observed by CZE- and RPLC-MS/MS as three phosphoproteoforms, which correlate to enlargement of the heart in a model of hypertrophic cardiomyopathy (ref. 59). Both CZE- and RPLC-TDPs successfully resolved and quantified all three proteoforms.

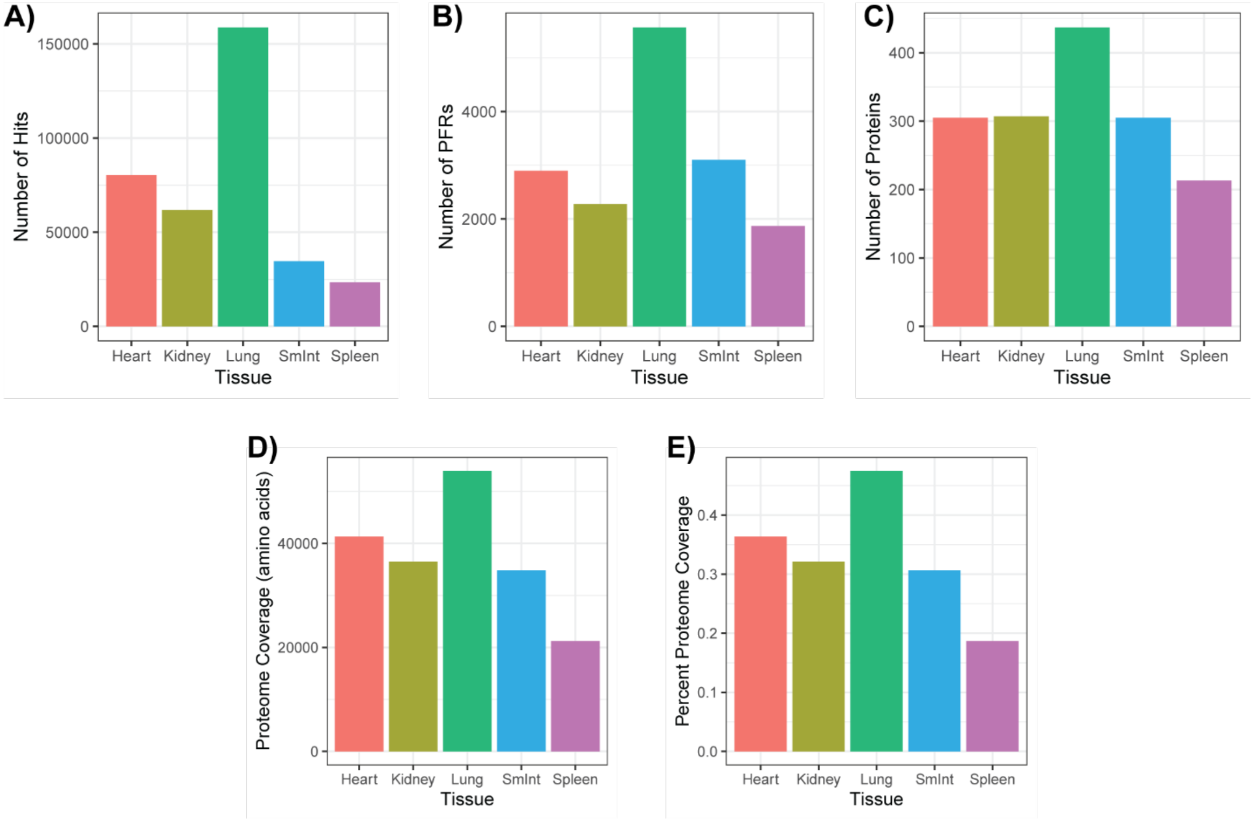
Supporting Information for:

Mapping the Proteoform Landscape of Five Human Tissues

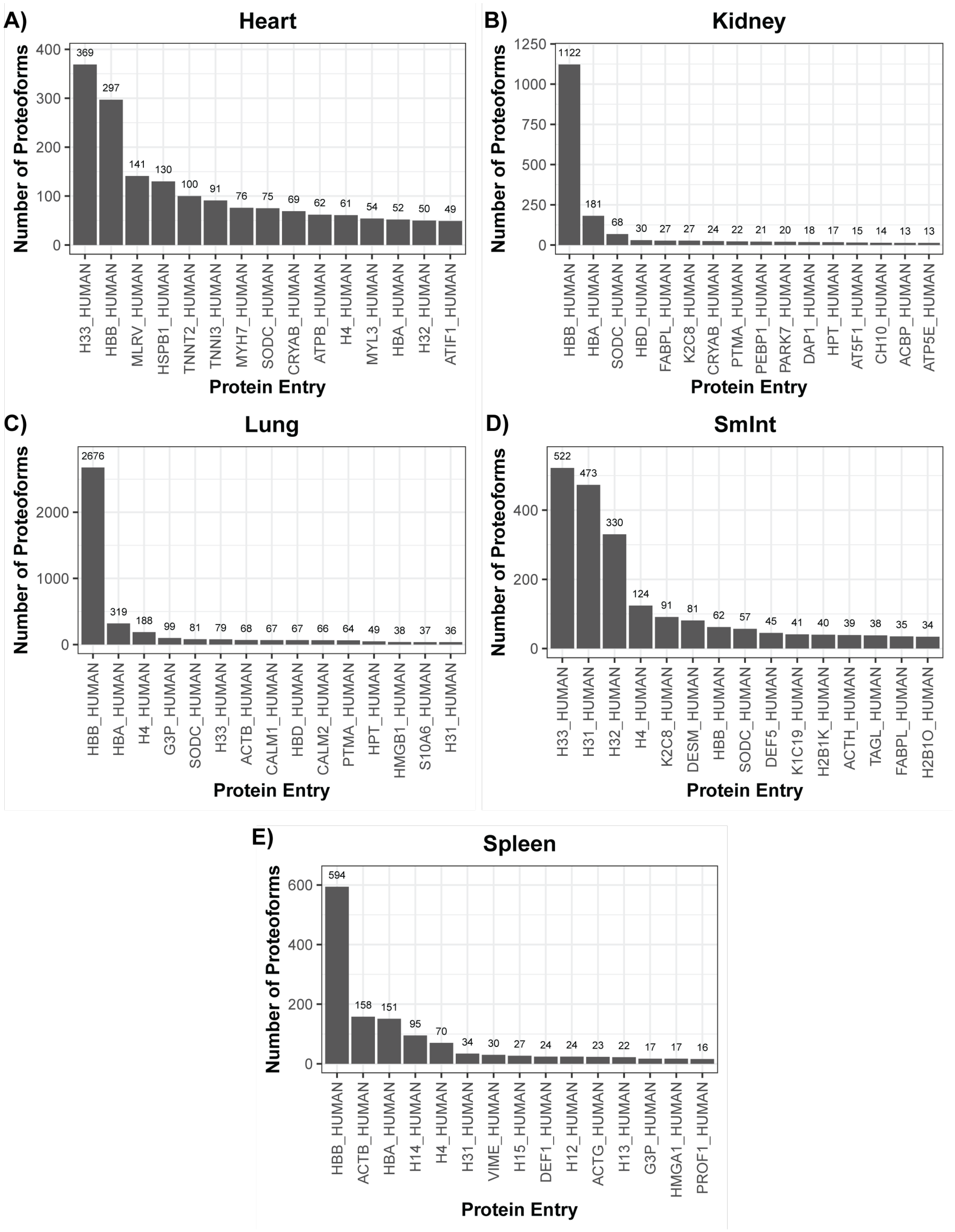
Bryon S. Drown, Kevin Jooß, Rafael D. Melani, Cameron Lloyd-Jones, Jeannie M. Camarillo, Neil L. Kelleher*

*Correspondence to: Neil L Kelleher, n-kelleher@northwestern.edu 2145 Sheridan Rd, Evanston, IL 60208.

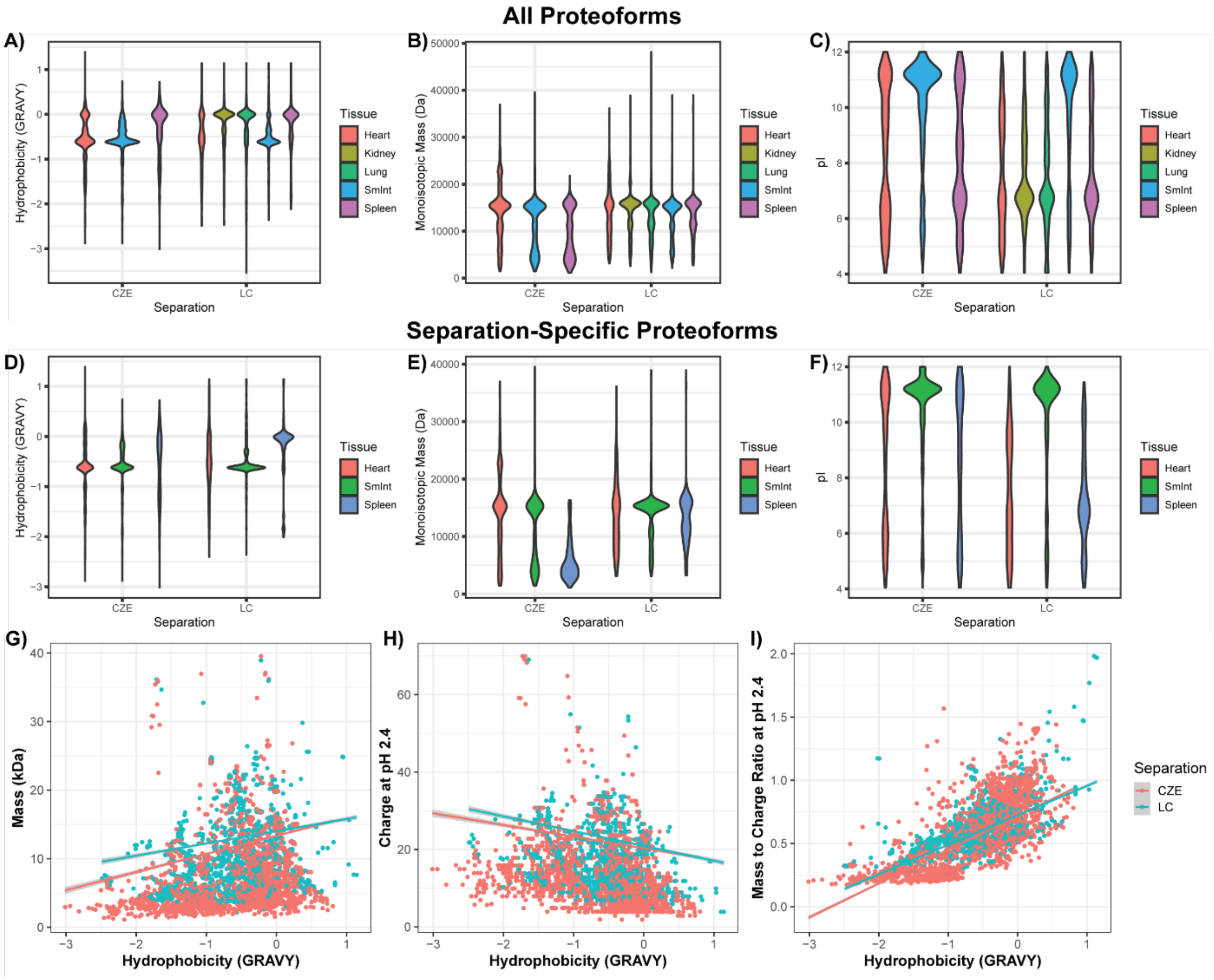
Table of Contents



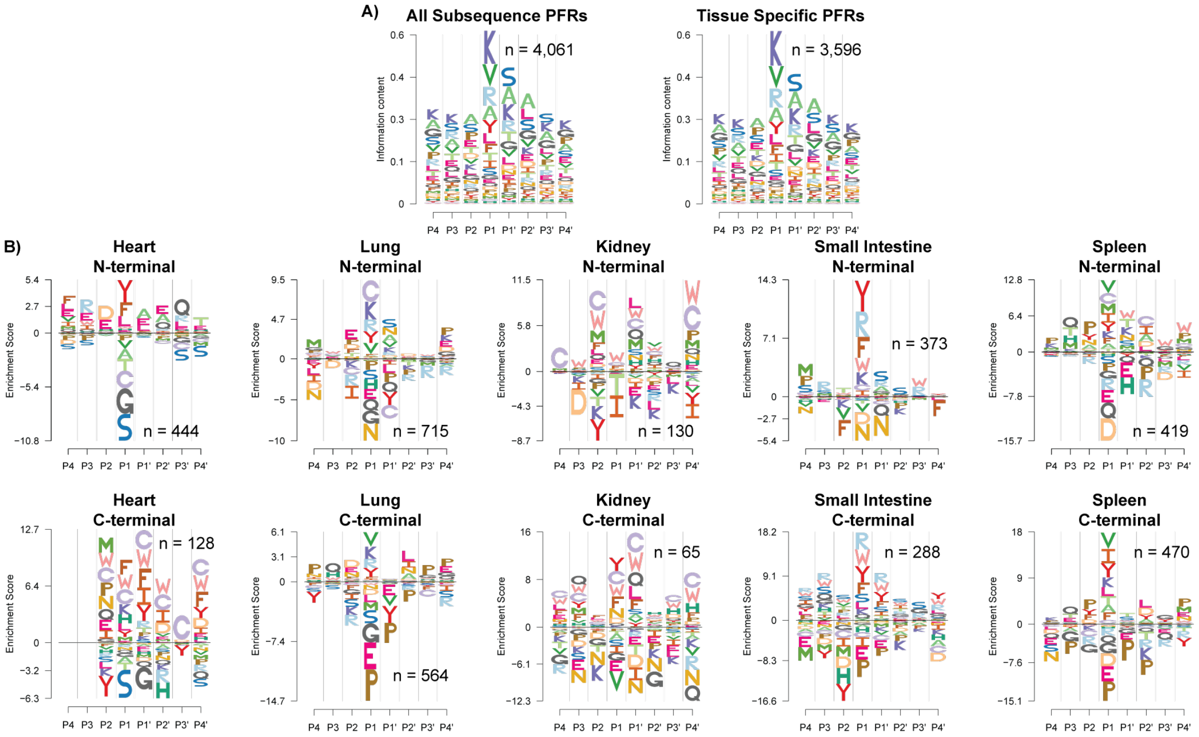
Supplementary Figure 1. Performance metrics of proteoform searches across tissues. **A)** Number of proteoform spectral matches (PrSMs or hits) in each tissue. **B)** Number of proteoforms identified in each tissue. **C)** Number of proteins identified in each tissue. **D)** Coverage of human proteome by non-redundant proteoform sequences. **E)** Percent coverage of the theoretical human proteome (calculated from the sequences of the canonical isoforms as prodcuts from the 20,300 human genes).



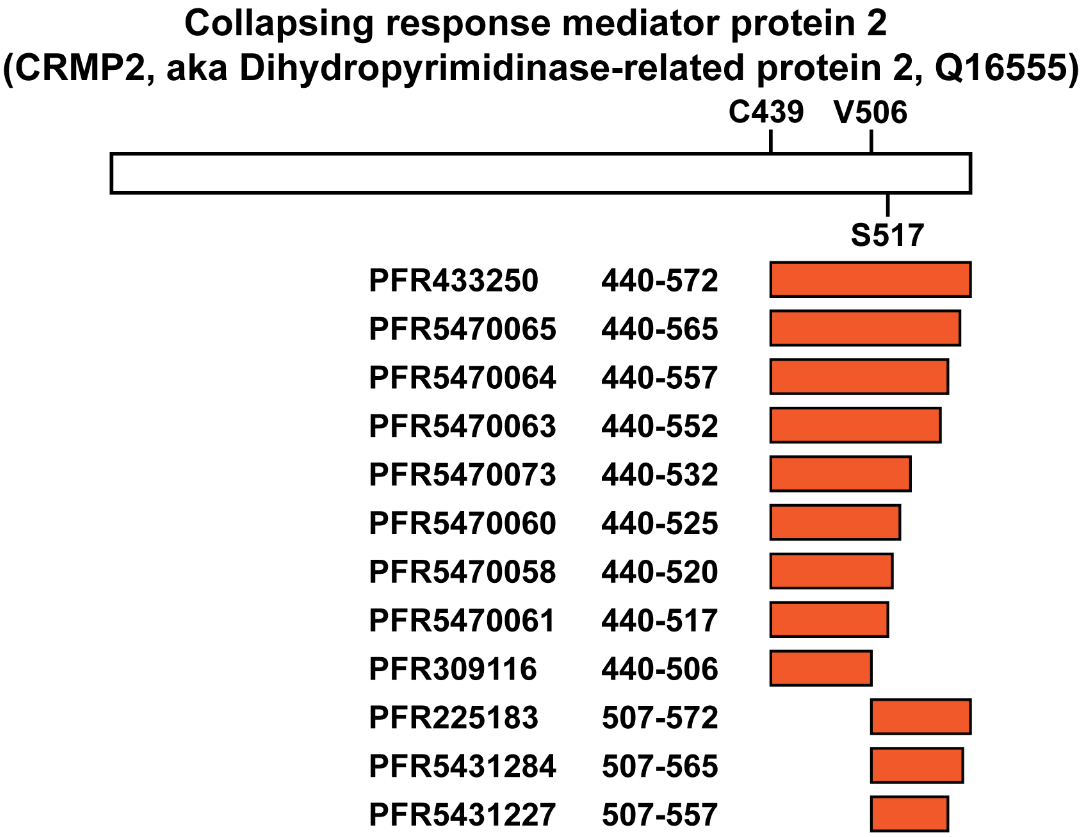
Supplementary Figure 2. Number of proteoforms identified per protein (SwissProt entry) for top fifteen proteins in heart (A), kidney (B), lung (C), small intestines (D), and spleen (E).



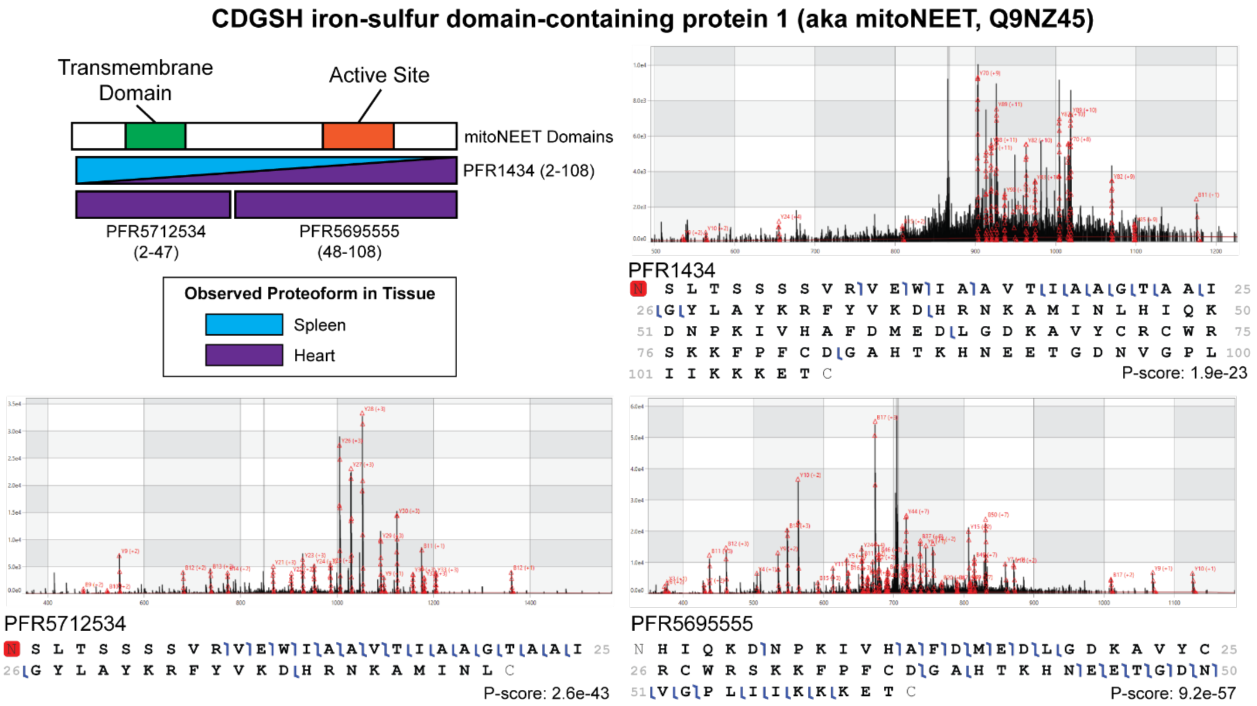
Supplementary Figure 3. Distribution of physiochemical properties of proteoforms, including hydrophobicity, monoisotopic mass, and pI-value, identified in human tissues by either CZE-MS/MS or RPLC-MS/MS. Violin plots depict the density of proteoforms at a given property value for all proteoforms (A-C) and proteoforms unique to a single separation method (D-F). Relationship between hydrophobicity and mass (G), charge at pH 2.4 (H), and mass to charge ratio at pH 2.4 (I). Hydrophobicity and isoelectric point (pI) were calculated from the base sequence of the proteoform and do not account for the presence of PTMs.



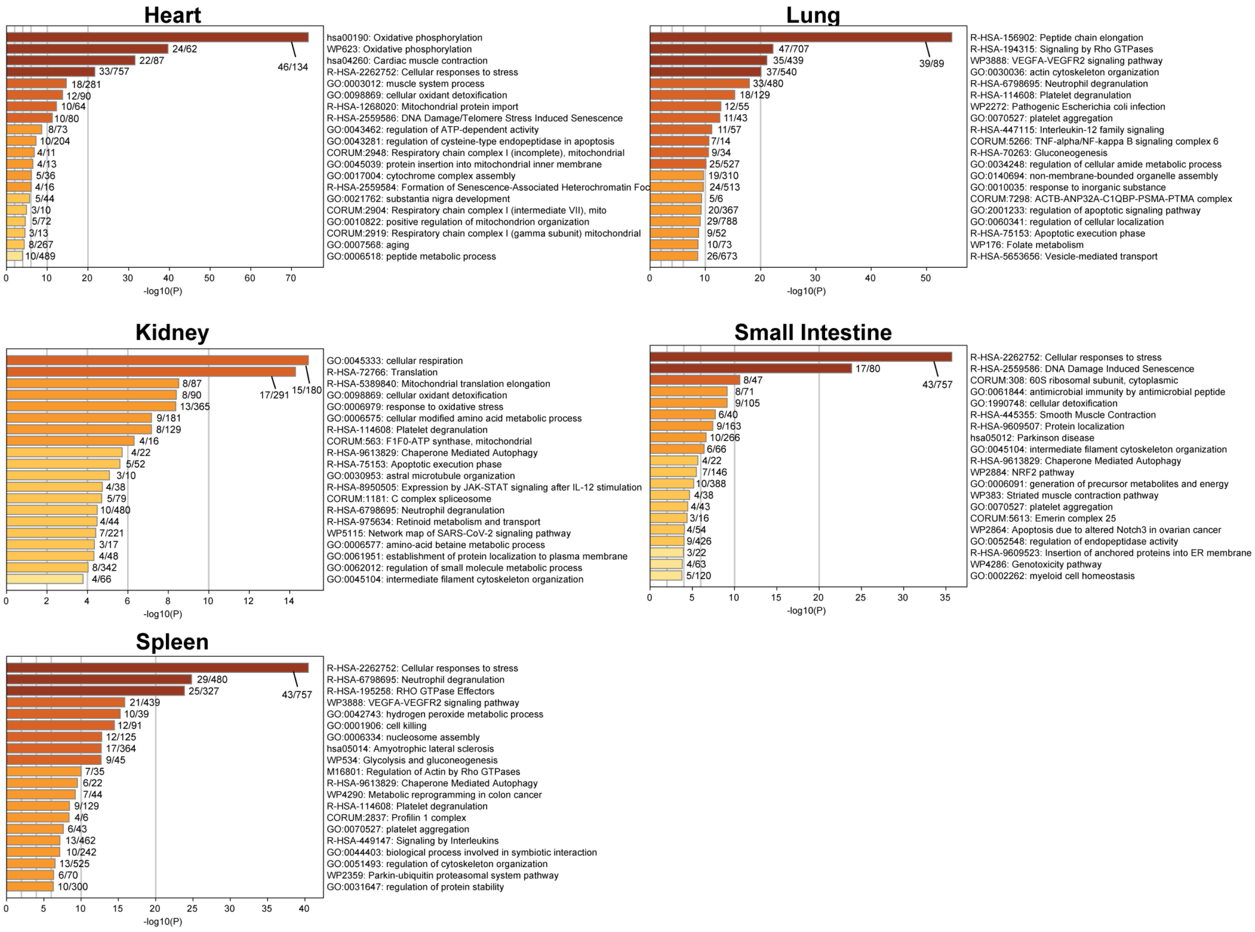
Supplementary Figure 4. Analysis of cleavage sites on proteoforms discovered by Subsequence search. **A)** Logo plots of all cleavage sites of all subsequence proteoforms (n = 4,061) and tissue-specific proteoforms (n = 3,596). **B)** Logo plots of cleavage sites subdivided by tissue type and termini for tissue-specific proteoforms.



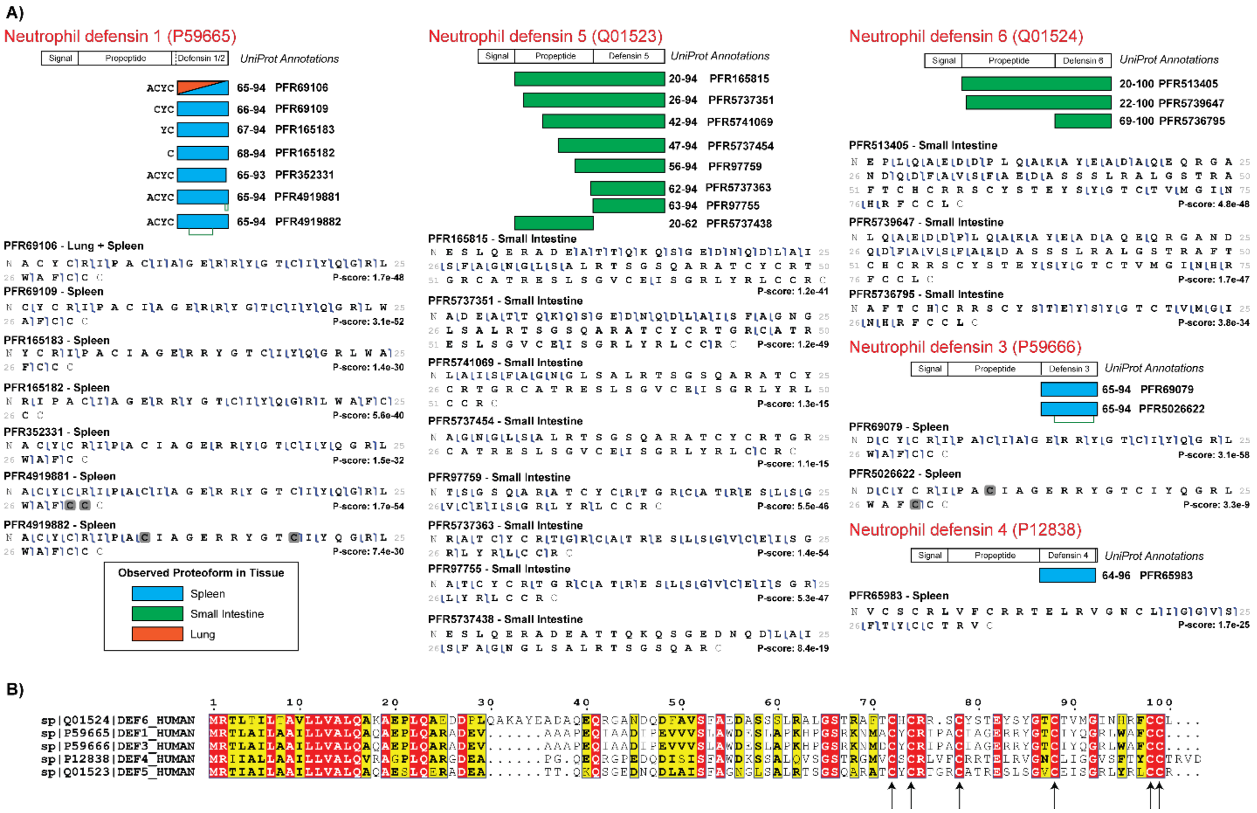
Supplementary Figure 5. Proteoforms of CRMP2 identified in human lung tissue with Subsequence search. Multiple proteoforms arising from cleavage at C439 or V506 were observed.



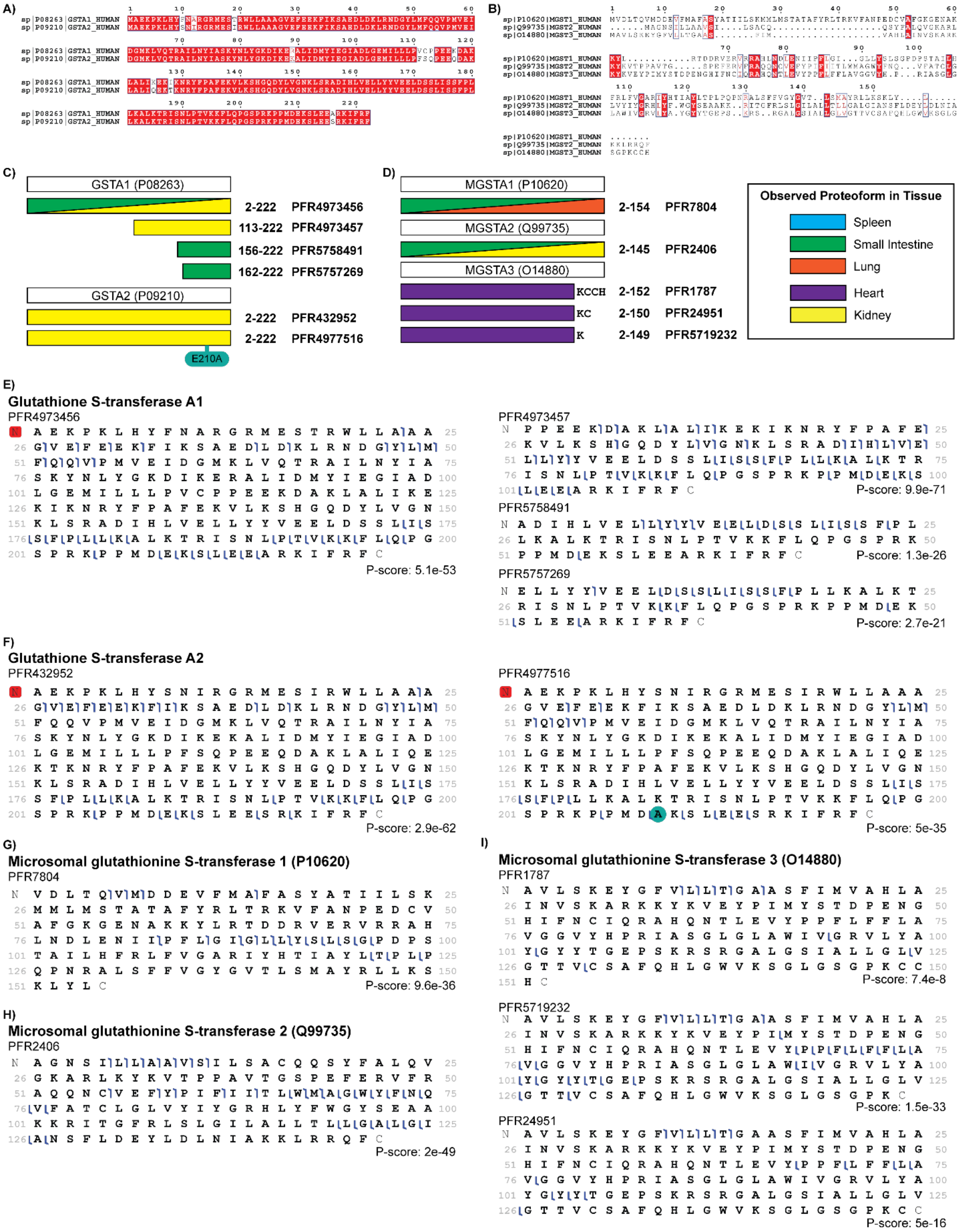
Supplementary Figure 6. Identification of mitoNEET proteoforms following proteolytic cleavage at L47. MS2 spectra following HCD fragmentation with matching fragment ions annotated.



Supplementary Figure 7. Gene ontology enrichment of genes associated with unique proteoforms. Each ontology term is labeled with the number of matching genes for the tissue type and total number of genes annotated with that term.



Supplementary Figure 8. **A)** Characterization of neutrophil defensin proteoforms with their supporting graphical fragment maps. **B)** Multi-sequence alignment of human alpha defensins with Clustal Omega visualized with ESPript 3. Arrow indicate conserved cysteines for disulfide bridges.

A)
sp|P08263|GSTA1_HUMAN
sp|P09210|GSTA2_HUMAN
B)
sp|P10620|MGST1_HUMAN
sp|Q99735|MGST2_HUMAN
sp|O14880|MGST3_HUMAN
C)
GSTA1 (P08263)
2-222 PFR4973456
113-222 PFR4973457
156-222 PFR5758491
162-222 PFR5757269
GSTA2 (P09210)
2-222 PFR432952
2-222 PFR4977516
E210A
D)
MGSTA1 (P10620)
2-154 PFR7804
MGSTA2 (Q99735)
2-145 PFR2406
MGSTA3 (O14880)
KCCH 2-152 PFR1787
KC 2-150 PFR24951
K 2-149 PFR5719232
Observed Proteoform in Tissue
Spleen
Small Intestine
Lung
Heart
Kidney
E)
Glutathione S-transferase A1
PFR4973456
N AEKPKLHYFNARGRMESTRWLLAAA 25
26 GVEFEEKFIKSAEDLDKLRNDGYLM 50
51 FQQVPMVEIDGMKLVQTRAILNYIA 75
76 SKYNLYGKDIKERALIDMYIEGIAD 100
101 LGEMILLLPVCPPEEKDAKLALIKE 125
126 KIKNRYFPAFEKVLKSHGQDYLVGN 150
151 KLSRADIHLVELLYYVEELDSSLIS 175
176 SFPLLKALKTRISNLPTVKKFLQPG 200
201 SPRKPPMDEKSLEEARKIFRF C
P-score: 5.1e-53
PFR4973457
N PPEEKDAKLALIKEKIKNRYFPAFE 25
26 KVLKSHGQDYLVGNKLSRADIHLVE 50
51 LLYYVEELDSSLISSFPLLKALKTR 75
76 ISNLPTVKKFLQPGSPRKPPMDEKS 100
101 LEEARKIFRF C
P-score: 9.9e-71
PFR5758491
N ADIHLVELLYYVEELDSSLISSFPL 25
26 LKALKTRISNLPTVKKFLQPGSPRK 50
51 PPMDEKSLEEARKIFRF C
P-score: 1.3e-26
PFR5757269
N ELLYYVEELDSSLISSFPLLKALKT 25
26 RISNLPTVKKFLQPGSPRKPPMDEK 50
51 SLEEARKIFRF C
P-score: 2.7e-21
F)
Glutathione S-transferase A2
PFR432952
N AEKPKLHYSNIRGRMESIRWLLAAA 25
26 GVEFEEKFIKSAEDLDKLRNDGYLM 50
51 FQQVPMVEIDGMKLVQTRAILNYIA 75
76 SKYNLYGKDIKEKALIDMYIEGIAD 100
101 LGEMILLLPFSQPEEQDAKLALIQE 125
126 KTKNRYFPAFEKVLKSHGQDYLVGN 150
151 KLSRADIHLVELLYYVEELDSSLIS 175
176 SFPLLKALKTRISNLPTVKKFLQPG 200
201 SPRKPPMDEKSLEESRKIFRF C
P-score: 2.9e-62
PFR4977516
N AEKPKLHYSNIRGRMESIRWLLAAA 25
26 GVEFEEKFIKSAEDLDKLRNDGYLM 50
51 FQQVPMVEIDGMKLVQTRAILNYIA 75
76 SKYNLYGKDIKEKALIDMYIEGIAD 100
101 LGEMILLLPFSQPEEQDAKLALIQE 125
126 KTKNRYFPAFEKVLKSHGQDYLVGN 150
151 KLSRADIHLVELLYYVEELDSSLIS 175
176 SFPLLKALKTRISNLPTVKKFLQPG 200
201 SPRKPPMDAKSLEESRKIFRF C
P-score: 5e-35
G)
Microsomal glutathione S-transferase 1 (P10620)
PFR7804
N VDLTQVMDDEVFMAFASYATIILSK 25
26 MMLMSTATAFYRLTRKVFANPEDCV 50
51 AFGKGENAKKYLRTDDRVERVRRAH 75
76 LNDLENIIPFLGIGLLYSLSGPDPS 100
101 TAILHFRLFVGARIYHTIAYLTPLP 125
126 QPNRALSFFVGYGVTLSMAYRLLKS 150
151 KLYL C
P-score: 9.6e-36
H)
Microsomal glutathione S-transferase 2 (Q99735)
PFR2406
N AGNSILLAAVSILSACQQSYFALQV 25
26 GKARLKYKVTPPAVTGSPEFERVFR 50
51 AQQNCVEFYPIFIITLWMAGWYFNQ 75
76 VFATCLGLVYIYGRHLYFWGYSEAA 100
101 KKRITGFRLSLGILALLTLLGALGI 125
126 ANSFLDEYLDLNIAKKLRRQF C
P-score: 2e-49
I)
Microsomal glutathione S-transferase 3 (O14880)
PFR1787
N AVLSKEYGFVLLTGAASFIMVAHLA 25
26 INVSKARKKYKVEYPIMYSTDPENG 50
51 HIFNCIQRAHQNTLEVYPPFLFFLA 75
76 VGGVYHPRIASGLGLAWIVGRVLYA 100
101 YGYYTGEPSKRSRGALGSIALLGLV 125
126 GTTVCSAFQHLGWVKSGLGSGPKCC 150
151 H C
P-score: 7.4e-8
PFR5719232
N AVLSKEYGFVLLTGAASFIMVAHLA 25
26 INVSKARKKYKVEYPIMYSTDPENG 50
51 HIFNCIQRAHQNTLEVYPPFLFFLA 75
76 VGGVYHPRIASGLGLAWIVGRVLYA 100
101 YGYYTGEPSKRSRGALGSIALLGLV 125
126 GTTVCSAFQHLGWVKSGLGSGPK C
P-score: 1.5e-33
PFR24951
N AVLSKEYGFVLLTGAASFIMVAHLA 25
26 INVSKARKKYKVEYPIMYSTDPENG 50
51 HIFNCIQRAHQNTLEVYPPFLFFLA 75
76 VGGVYHPRIASGLGLAWIVGRVLYA 100
101 YGYYTGEPSKRSRGALGSIALLGLV 125
126 GTTVCSAFQHLGWVKSGLGSGPKCC
P-score: 5e-16

Supplementary Figure 9. Identification and tissue distribution of glutathione transferase proteoforms. **A)** Sequence alignment of glutathione S-transferase A1 and A2. **B)** Sequence alignment of microsomal glutathione S-transferases. **C)** Overview of characterized glutathione S-transferase A1 and A2 proteoforms and their tissue distribution. **D)** Overview of characterized microsomal glutathione S-transferase proteoforms and their tissue distribution. **E)** Fragmentation maps of GSTA1 proteoforms. **F)** Fragmentation maps of GSTA2 proteoforms. **G)** Fragmentation map of MGSTA1 proteoform. **H)** Fragmentation map of MGSTA2 proteoform. **I)** Fragmentation maps of MGSTA3 proteoforms.

Supplemental Table 1. Descriptions and metdata for tissue samples analyzed by top-down proteomics.

Tissue Type	**HuBMAP Identifier**	**BioRep**[a]	**Demographics**
Lung	D2390RML-BPS-8	1	37 years, Male, African American
Lung	D2390RML-BPS-9	2	37 years, Male, African American
Lung	D2390RML-BPS-10	3	37 years, Male, African American
Kidney	VAN0003-LK-32	1	73 years, Female, White
Kidney	VAN0005-RK-4	2	58 years, Female, White
Kidney	VAN0009-LK-102	3	53 years, Male, African American
Kidney	VAN0011-RK-3	4	31 years, Male, White
Kidney	VAN0029-RK-1	5	62 years, Male, White
Heart	W146	1	25 years, Female, White
Heart	W158	2	61 years, Male, White
Small Intestine	B005-A-406	1	24 years, Female, White
Spleen	19-004-02	1	18 years, Male, White

[a]Biological Replicate defined as the number of tissue samples received from a particular donor.

Supplemental Table 2. Program used in the CZE separation method.

Time [min.]	Event	Value	Duration	Inlet	Outlet	Summary
	Rinse - Pressure	100 psi	5 min	BI: C1	BO: A1	forward
	Rinse - Pressure	100 psi	3 min	BI: A1	BO: A1	reverse
	Rinse - Pressure	100 psi	5 min	BI: A1	BO: A1	forward
	Inject - Pressure	2.5 psi	60 s	SI: A1	BO: A1	forward
	Wait		0 min	BI: D1	BO: A1	dipping
	Inject – Pressure	2.5 psi	10 s	BI: B1	BO: A1	forward
0	Separation Voltage	15 kV 0.5 psi	70 min	BI: B1	BO: A1	1 min ramp, normal polarity, both
1	Relay On					
70	Separation Voltage	1 kV 5 psi	5 min	BI: B1	BO: A1	5.0 min ramp, normal polarity, both
75	End					

BI: A1, B1 = BGE; C1 = 0.1 M HCl, D1 = H_2O SI: A1 = sample BO: A1: CL; B1 = H_2O

Supplemental Table 3. Correlation coefficients of retention/migration time and proteoform mass/lipophilicty by separation method and GELFrEE fraction.

	Proteoform Mass				Proteoform Hydrophobicity (GRAVY)			
	CZE		RPLC		CZE		RPLC	
Fraction	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
1	-0.081	0.061	0.41	0.50	0.63	0.69	0.68	0.67
2	-0.15	-0.12	0.46	0.49	0.45	0.43	0.74	0.65
3	0.10	-0.054	0.53	0.61	0.50	0.60	0.72	0.63
4	-0.20	-0.031	0.56	0.61	0.50	0.61	0.72	0.60
5	-0.11	0.23	0.40	0.50	0.51	0.63	0.70	0.55
6	0.33	0.44	0.30	0.35	0.40	0.44	0.71	0.47

Supplemental Table 4. Frequencies of observation for post-translational modifications on proteoforms counted at the level of proteoform spectral matches (PrSMs) categorized by separation technique.

	CZE		RPLC			
PTM type	Observed[a]	Freq.[b]	Observed[a]	Freq.[b]	χ^2	p-value[c]
Unmodified[d]	61,294	0.56	33,518	0.55	45	2.6×10^{-10}
Monoacetylation[d]	14,628	0.13	8,937	0.15	42	1.7×10^{-9}
Phosphorylation[d]	8,943	0.082	6,039	0.098	129	1.0×10^{-28}
Trimethylation[d]	7,141	0.066	4,793	0.078	94	4.9×10^{-21}
Monomethylation[d]	5,422	0.050	1,837	0.030	379	2.8×10^{-83}
Dimethylation[d]	4,948	0.045	1,995	0.032	168	3.4×10^{-37}
Half cystine[d]	2,822	0.026	1,033	0.017	146	1.8×10^{-32}
Nitrosylation[d]	2,081	0.019	1,980	0.032	291	5.1×10^{-64}
Deaminated L-asparagine[d]	804	0.0074	931	0.015	235	7.3×10^{-52}
Monohydroxylation[d]	344	0.0036	5	8.15×10^{-5}	180	6.3×10^{-40}
Pyruvic acid iminylated residue[d]	168	0.0015	195	0.0032	48	5.0×10^{-11}
L-cysteine sulfinic acid[d]	159	0.0015	7	1.1×10^{-4}	72	3.8×10^{-16}
S-myristoylation	76	7.0×10^{-4}	59	9.6×10^{-4}	3.1	1.2
S-palmitoylation[d]	28	2.6×10^{-4}	52	8.5×10^{-4}	28	2.0×10^{-6}
Nitration	34	3.1×10^{-4}	6	9.8×10^{-5}	6.8	0.14
Total	108,892		61,387			

[a] Number of proteoform spectral matches with specific PTMs at 1% FDR; count does not include N-terminal and C-terminal modifications; multiple PTMs on the same proteoform are counted multiple times.

[b] Number of observations/sum of PTM observations for each separation technique.

[c] Bonferroni corrected p-value (n = 15)
[d] Statistically significant difference (alpha <0.01) in frequency of observation.