bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

High-throughput proteomic and phosphoproteomic analysis of formalin-fixed paraffin-embedded tissue

Moe Haines[1#]; John R. Thorup[1#]; Simone Gohsman[1]; Claudia Ctortecka[1.], Chelsea Newton[2]; Dan C. Rohrer[2]; Galen Hostetter[2]; D. R. Mani[1]; Michael A. Gillette[1]; Shankha Satpathy[1,3*]; Steven A. Carr[1*]

1. Broad Institute of MIT and Harvard, Cambridge, MA, USA
2. Van Andel Research Institute, Grand Rapids, MI, USA
3. Current address: AstraZeneca R&D, Waltham, MA, USA

#Co-first authors

*Corresponding authors:

(SAC) scarr@broad.mit.edu & (SS) shankha.satpathy@astrazeneca.com

Supplementary Figure Legends:

Supplemental Fig. 1: Method Optimization

A. Peptide summary from all workflows. Bar plots represent data from n=4 scrolls per block, showing means and standard deviation (SD) error bars.

B. Quantified TIC of all peptides from each experiment across n=4 replicates. The box represents the interquartile range (IQR), with the top and bottom edges indicating the 75th and 25th percentiles, respectively. The line inside the box denotes the median, and whiskers extend to capture data within 1.5x of the IQR.

C. Digestion efficiency of S-Trap and SP3 workflows. Bar plots represent data from n=4 scrolls per block, showing means and SD error bars. The box represents the IQR, with the top and bottom edges indicating the 75th and 25th percentiles, respectively. The line inside the box denotes the median, and whiskers extend to capture data within 1.5x of the IQR.

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplemental Fig. S2: Comparative Evaluation of Sample Collection in Wet vs. Dry Wells and Different Stage-Tipping Materials

A. Peptide yields remain consistent across storage conditions (n=2 scrolls per condition) for each block. The BRC2 block shows the highest yields, likely due to processing the full tissue homogenate, which may increase protein/peptide yields.

B. Protein identifications and overlap between whole tissue and clear lysate samples.

C. Density plot showing the protein abundance profile, which remains consistent between whole tissue and clear lysates. The peak shape represents the kernel density estimate of all proteins after log2 transformation and normalization.

D-E. Comparison of protein and peptide depth in samples desalted with tC18 and SDB-RPS. Each bar represents data from four independent 10 µm scrolls.

F. Density plot showing protein abundance for all identified proteins using tC18 and SDB-RPS. The peak shape represents the kernel density estimate of all proteins after log2 transformation.

Figure S3: Evaluation of Artifactual FFPE-Derived Peptide Modifications

A. Peptide digests from single replicates of CRC and BRC blocks were pooled equally and fractionated using SDB-XC stage tips into five fractions for data-dependent acquisition (DDA). Data were analyzed using the open-search workflow in FragPipe.

B. Open searching with MSFragger and PTM-Shepherd identified 99 different modifications, with most modifications centered around 0 Da. The data were generated from DDA runs of stage-tip fractionated samples pooled from CRC and BRC groups. MSFragger was used for searching, and PTM-Shepherd for modification characterization.

C. Peptide Spectrum Match (PSM) summary of all modifications. FFPE-related modifications make up less than 10% of all PSMs.

D. Enrichment scores by residue for the most common FFPE modifications. The enrichment score represents the PSM count for each localized residue associated with a specific

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

modification. These scores are representative of the first modified residue in each PSM for a relative enrichment estimate.

E. Global modification profile represented by the count of identified peptides from the fractionated pooled sample.

Figure S4: Evaluation of Different DIA Methods and Description of Mouse FFPE Blocks

A. Proteome depth from 1 µg of Jurkat peptides analyzed using optimized DIA methods on the timsTOF HT and Exploris 480. Bars represent the average of (n=2 injections), with error bars showing the standard deviation. The Bruker timsTOF HT, equipped with a dual trapped ion mobility spectrometry (TIMS) tunnels, allows for parallel accumulation-serial fragmentation (PASEF), separating ions by their collisional cross-section (CCS), allowing precursor selection based on their ion mobility (IM) and mass to charge (m/z). Variable-window diaPASEF was optimized for FFPE proteomics using Jurkat peptide digests across four gradients (23, 30, 35, and 55 minutes) on a 25 cm PepSep column. The results were compared to data acquired on the Orbitrap Exploris 480 with 110-minute and 45-minute gradients using a 25 cm home-packed Reprosil C18 column. The wide-window DIA method on the Orbitrap Exploris480 utilized variable isolation windows ranging from 12 m/z to 24 m/z, adjusting to precursor density. Both methods aimed for six data points per peak (DPPP) for quantitative reproducibility. The timsTOF HT 35-minute gradient achieved ~8,000 unique proteins, comparable to the 110-minute Orbitrap gradient, with a balanced unique peptide depth (~135,000).

B. Peptide identifications across the four methods tested.

C. Four blocks from genetically engineered mouse models (GEMM) were used in the mouse FFPE experiments. Two blocks with total body Ncoa4 overexpression (OE) were compared to wild-type (WT) blocks. Blocks WT2 and OE2 included all organs, while WT1 lacked a few organs, and the OE1 block was missing the pancreas.

D. Slide representation showing all organs embedded into the GEMM FFPE blocks.

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S5: NMF Clustering of Proteome and Phosphoproteome

A. Cophenetic correlation and dispersion scores for each NMF proteome cluster count, with three clusters identified as optimal based on their intrinsic structure and high reproducibility.

B. Heatmap showing the clustering stability of LUAD samples into three distinct groups, as defined by NMF consensus and membership scores for each sample.

C. Silhouette scores for each LUAD sample based on NMF cluster assignment, used to evaluate the quality of assigned memberships.

D. Sankey diagram depicting NMF cluster memberships derived from either proteomic or phosphoproteomic datasets.

E. ssGSEA results from the three phosphoproteome NMF clusters.

F. Key phosphorylation events and their completeness across the LUAD 12 samples.

Supplemental Table Legends:

Supplemental Table S1: Global proteomics data associated with development and optimization of the workflow for analysis of the colorectal and breast cancer FFPE blocks

(A) DIA isolation scheme used for data generation on the Thermo Orbitrap Exploris 480. Method was used to analyze 500 ug of peptides, desalted by tC18 StageTips, derived from breast cancer (BRC) and colorectal cancer (CRC) FFPE blocks. Acquired raw files were searched using MSFragger-DIA Quant workflow.

(B) Protein-level abundance matrix of BRC and CRC blocks processed in 2% SDS and plate-based S-Trap.

(C) Protein-level abundance matrix of BRC and CRC blocks processed in 5% SDS and plate-based S-Trap.

(D) Protein-level abundance matrix of BRC and CRC blocks processed in 2% SDS and SP3 magnetic beads.

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(E) S1E: Stats file from the MSFragger-DIA output for BRC and CRC blocks processed by S-Trap (2% SDS or 5% SDS) and by SP3 beads (2% SDS)

Supplemental Table S2. Tables associated with the method application to macrodissected colorectal and breast cancer FFPE scrolls

(A) Dimensions of macrodissected tumor sections from breast cancer (BRC) and colorectal cancer (CRC) FFPE blocks, and their protein yields.

(B) DIA isolation scheme used for data generation on the Thermo Orbitrap Exploris 480. Method was used to analyze 1 ug of peptides, desalted by tC18 or SDB-RPS StageTips, derived from breast cancer (BRC) and colorectal cancer (CRC) tumor-rich FFPE sections.

(C) Protein-level abundance matrix from CRC and BRC peptides desalted with tC18StageTips.

(D) Protein-level abundance matrix from CRC and BRC peptides desalted with SDB-RPSStageTips.

(E) PTM-Shepherd summary

Supplemental Table S3. Global proteomics data associated with the quantitative evaluation across multiple platforms

(A) DIA isolation scheme used for data generation using the wide-window DIA method on the Thermo Orbitrap Exploris 480.

(B) DIA isolation scheme used for data generation using the diaPASEF method on the Bruker timsTOF HT.

(C) S4C:DIA isolation scheme used for data generation using the 4 m/z DIA method on the Thermo Orbitrap Astral.

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(D-G) Protein-level abundance matrix from FFPE blocks, embedding organs originating from genetically-engineered mouse models (GEMM), including wild-type (WT) and Ncoa4 over-expressing (OE) samples, acquired by wide-window DIA on the Thermo Orbitrap Exploris 480 (D), TIms TOF (E), Astral (F), TMT (G) used for data generation using the 4 m/z DIA method on the Thermo Orbitrap Astral.

Supplemental Table S4. Global proteomics and phosphoproteomics data associated with the application of this workflow to lung adenocarcinoma samples

(A) DIA isolation scheme used for data generation using the 4 m/z DIA method on the Thermo Orbitrap Astral. The method was used to analyze lung adenocarcinoma (LUAD) tumor rich FFPE sections

(B) Protein-level abundance matrix obtained from LUAD FFPE samples.

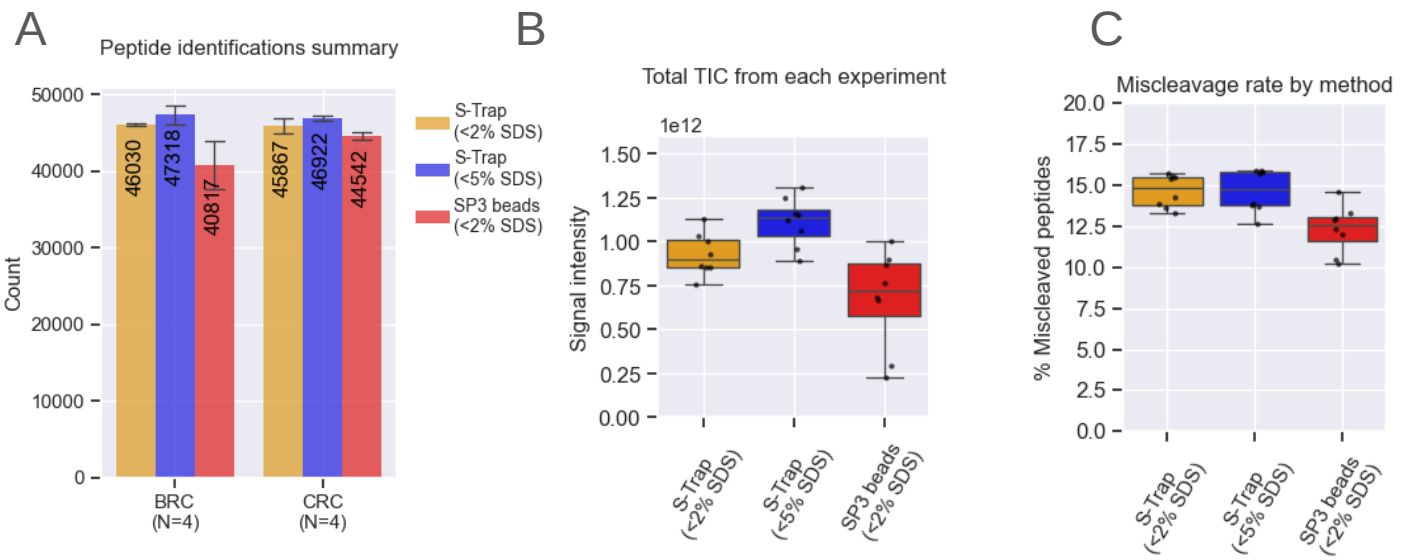
(C) Phosphosite-level abundance matrix obtained from LUAD FFPE samples.

(D) Protein-level matrix obtained from non-negative matrix factorization (NMF) clustering of LUAD samples.

(E) Phosphosite-level matrix obtained from NMF clustering of LUAD samples.

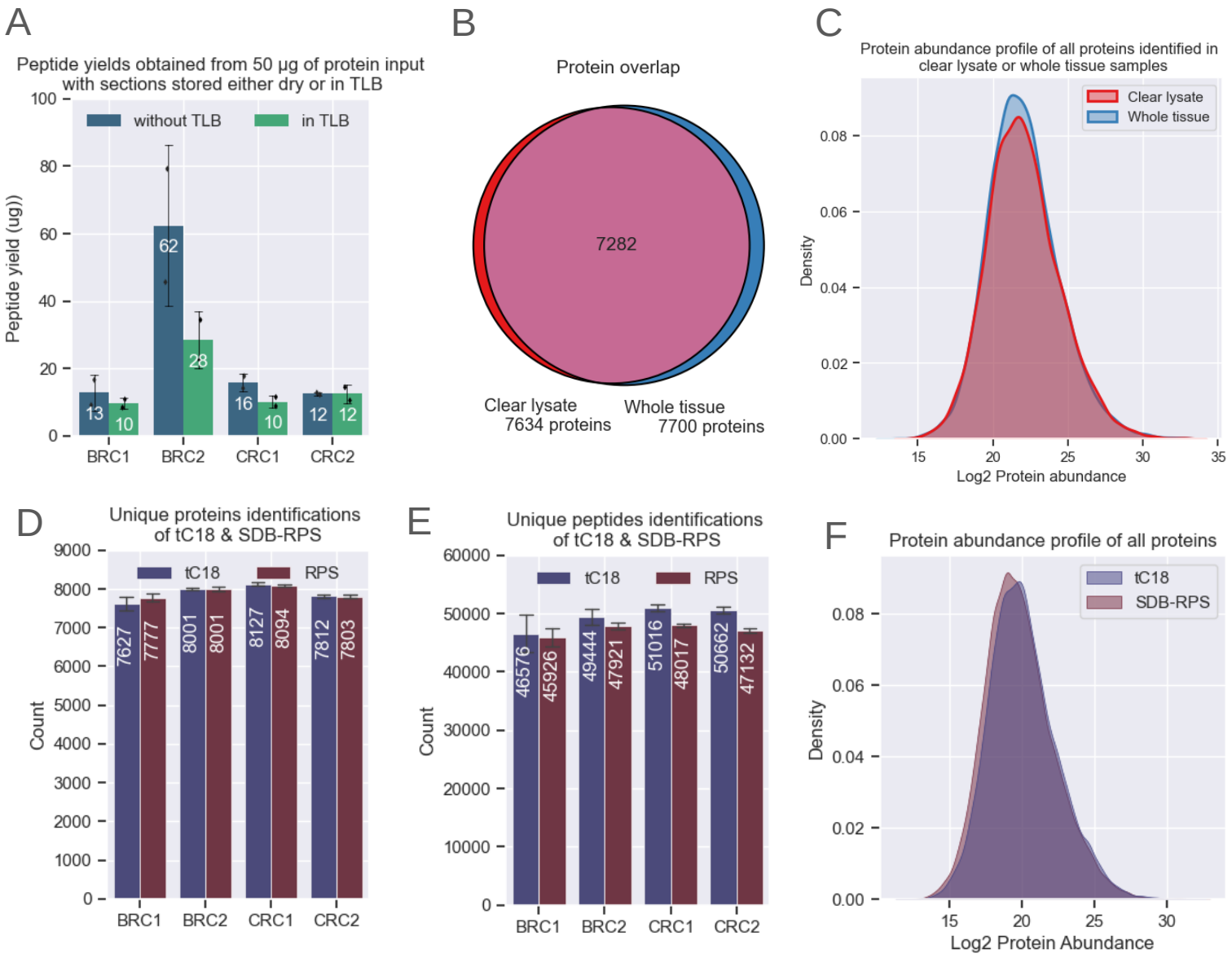
bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S1



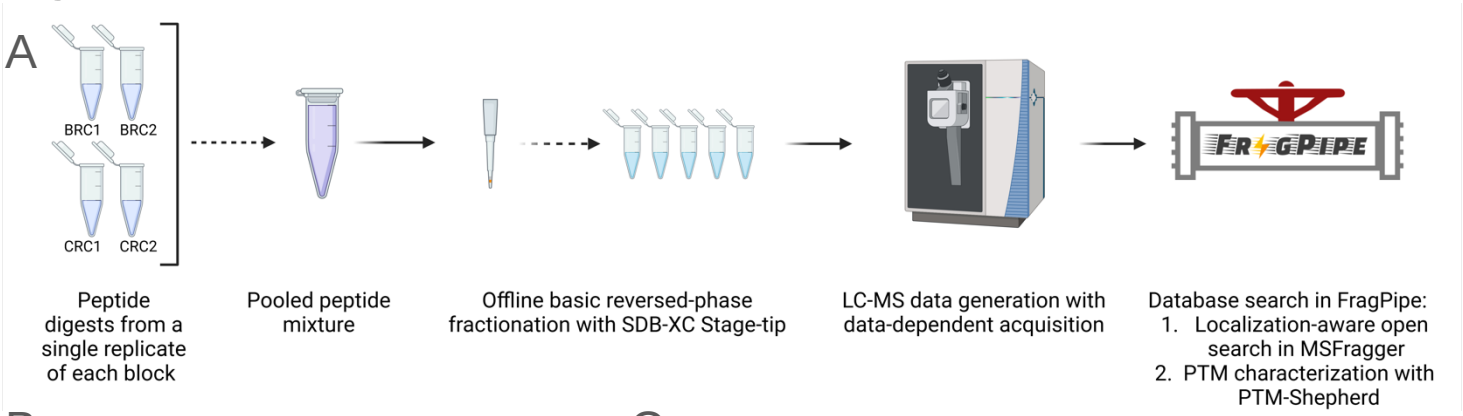
bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

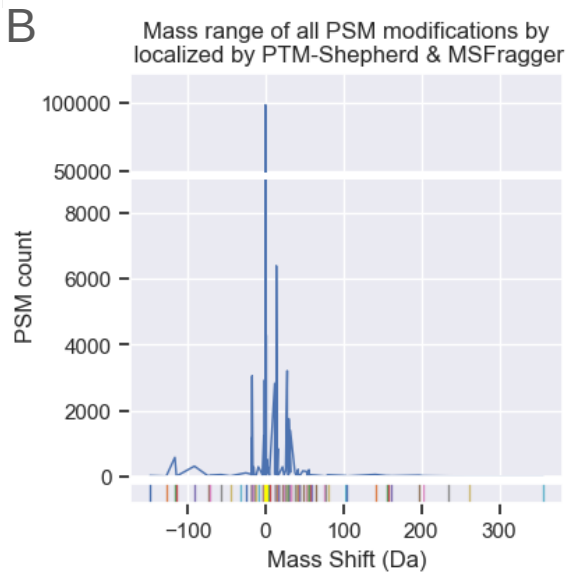
Figure S2

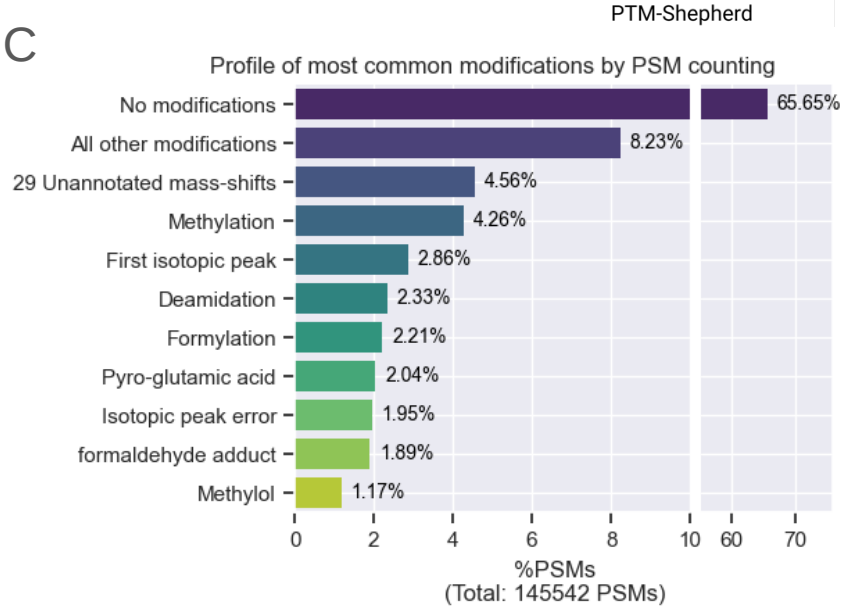


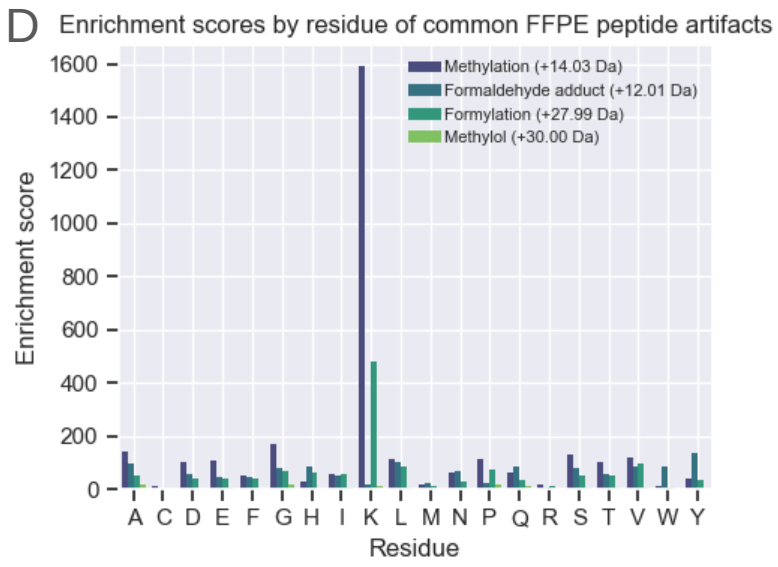
bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S3









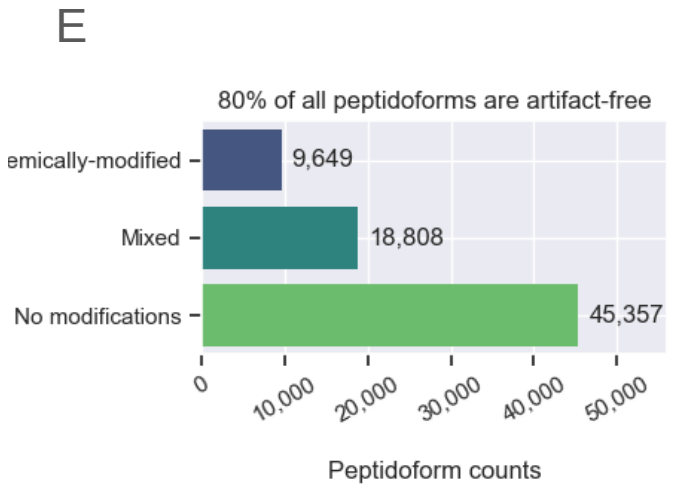
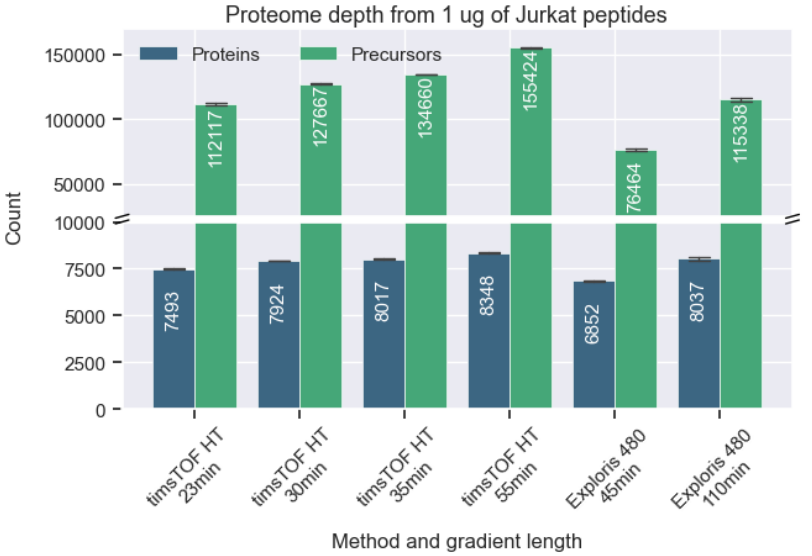


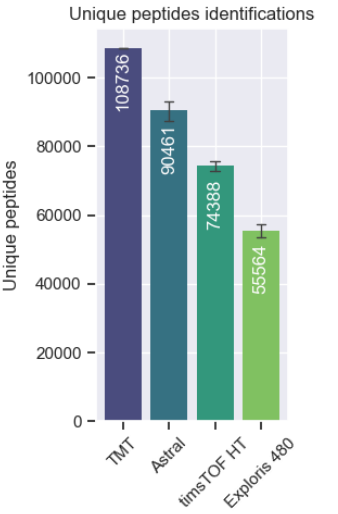
Figure S4

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

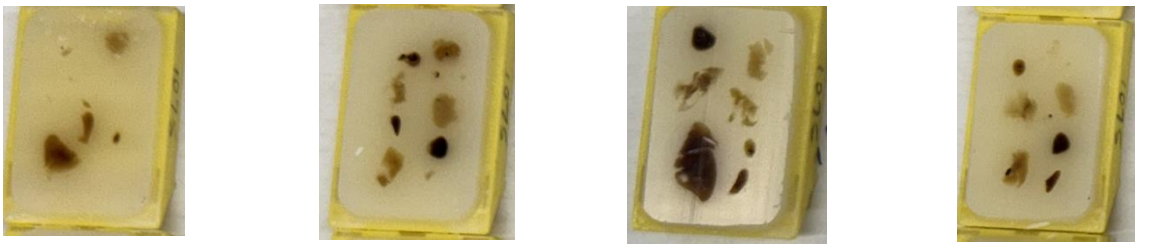
A.



B.



C.



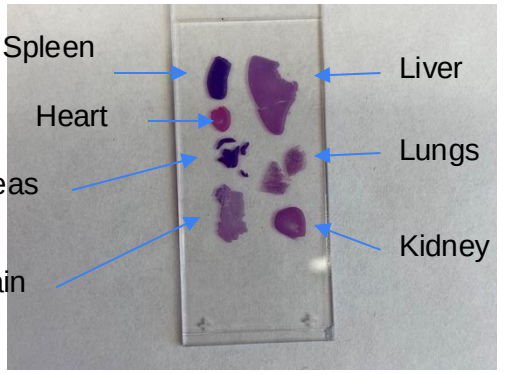
WT1
Wild-type 1
Lost few organs

WT2
Wild-type 2
All organs embedded

OE1
NCOA4-overexpressing
Missing pancreas

OE2
NCOA4-overexpressing 2
All organs embedded

D.



bioRxiv preprint doi: https://doi.org/10.1101/2024.11.17.624038; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S5