Title: ProteoSushi: a software tool to biologically annotate and quantify modification-specific, peptide-centric proteomics datasets

Authors: Robert W. Seymour¹, Sjoerd van der Post^{1,4}, Arshag D. Mooradian¹, and Jason M. Held^{1,2,3*}

¹Department of Medicine, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA

²Department of Anesthesiology, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA

³Siteman Cancer Center, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA

⁴Department of Medical Biochemistry, University of Gothenburg, Gothenburg, Sweden.

*Corresponding Author: Jason M. Held, Washington University School of Medicine, Campus Box 8076, 660 South Euclid Avenue, Saint Louis, MO 63110. Phone: 314-747-9738. Email: jheld@wustl.edu.

Keywords: post-translational modifications, PTMs, modification-specific proteomics, phosphorylation, bioinformatics, acetylation, proteomics, chemoproteomics, proteosushi

Abstract

Large-scale proteomic profiling of protein post-translational modifications has provided important insights into the regulation of cell signaling and disease. These modification-specific proteomics workflows nearly universally enrich modified peptides prior to mass spectrometry analysis, but protein-centric proteomic software tools have many limitations evaluating and

interpreting these peptide-centric datasets. We therefore developed ProteoSushi, a software tool tailored to the analysis of each modified site in peptide-centric proteomic datasets that is compatible with any post-translational modification or chemical label. ProteoSushi uses a unique approach to assign identified peptides to shared proteins and genes, minimizing redundancy by prioritizing shared assignments based on UniProt annotation score and optional user-supplied protein/gene lists. ProteoSushi simplifies quantitation by summing or averaging intensities, merging overlapping peptide charge states, missed cleavages, peptide spectral matches, and variable modifications into a single value for each modified site. ProteoSushi annotates each PTM site with the most up-to-date biological information available from UniProt, such as functional roles or known modifications, the protein domain in which the site resides, the protein's subcellular location and function and more. ProteoSushi has a graphical user interface for ease of use. ProteoSushi's flexibility and combination of features streamlines peptide-centric data processing and knowledge mining of large modification-specific proteomics datasets.

Introduction

Protein post-translational modifications (PTMs) expand the proteome's function by regulating protein structure, activity, localization, and interactions. Approximately 5% of human genes encode PTM modifying enzymes, and of the several million proteoforms found within a cell, the vast majority of proteome diversity stems from PTMs¹. There are hundreds of known PTMs, with some PTMs predominantly regulated by enzyme activity and others more often the result of chemical modification^{2,3}.

Large-scale, unbiased identification and quantitation of modified residues using proteomics has provided unparalleled insight into the regulation of PTM networks, typically quantifying many thousands of modified sites in parallel. In a typical modification-focused proteomic experiment, the PTM of interest is enriched via antibodies or affinity tags^{2,4,5}. Transient modifications can also be chemically labeled for chemoproteomics, potentially utilizing click chemistry to conjugate on a handle for enrichment such as biotin^{6,7}. Thus, a key feature of modification-specific proteomics datasets is a highly enriched set of peptides modified by the

target modification(s). There is typically a single target PTM or chemical label of interest per peptide, or several in the case of differential alkylation of cysteines or phosphorylation of serine, threonine, and tyrosine^{8,9}.

The data processing goals of modification-specific datasets coupled with their peptide-centric nature dictate several unique analytical challenges compared to protein-centric datasets. First, typical database search algorithms leverage the many peptides identified in a sample to infer which related protein isoforms or protein groups are present, and then assign peptides to these protein groups parsimoniously. However, protein coverage in modification-specific proteomic datasets is sparse, as only modified peptides are analyzed. Since each modified peptide identified is a unique entity, it must be treated as such for analysis. Assigning modified peptides using typical protein-based typical grouping algorithms is inconsistent and varies depending on the other peptides identified in a sample. Grouping modified peptides at the protein level may inappropriately minimize the actual level of ambiguity in protein assignment, especially since the amino acids proximal to regulatory PTMs are often highly conserved^{10–12}. Second, database search tools don't provide downstream annotation of the modified sites identified. Third, multiple peptide modifications detected on the modified peptides due to sample processing, such as deamidation or methionine oxidation, or missed cleavages, are of limited biological interest. Therefore, the typical goal of modification-specific analysis is to collapse these additional peptides to a single PTM site and quantitative value for simplification, a feature not typical in protein-centric tools that focus on grouping proteins rather than PTM sites.

An array of software tools exists for peptide- and PTM-centric analysis including databases^{13–18}, PTM site assignment^{19–22}, and visualization^{23,24}. However, no software tool is able to quantify and annotate any site and type of modified residue(s) in proteomic datasets. Software tools focused on downstream processing of proteomic data include Perseus, an integrated tool for statistics and comparisons across samples²⁵. However, Perseus is in essence protein-centric and has limited capacity to annotate all but the most common PTM sites with up-to-date information. PTMsigDB is another PTM-focused annotation tool to identify biological signatures based on site-specific, rather than gene-level data, but it is limited to

phosphorylation and is also designed to act downstream of many key peptide-centric processing decisions and steps²⁶. Skyline²⁷, a software tool for label free quantitation and analysis of data-independent acquisition (DIA) data²⁸, is peptide-focused, though the resulting output requires reassignment of modified peptides to shared proteins and genes, merging of modified peptides and charge states, and biological annotation of the modified sites²⁹. Therefore, there is an unmet need for a flexible and easy to use pipeline to process, characterize, and interrogate modification-specific proteomic data downstream of both database search and quantitative results.

To meet this need we developed ProteoSushi, a purpose-built software framework tailored to the peptide centric quantification, protein assignment and biological annotation of modification-specific proteomics data (**Fig. 1**). ProteoSushi is highly flexible and can be applied to any PTM or chemical mass addition of interest, and can be used for any type of quantitation, including isotopic- and isobaric-labeling and data-independent acquisition (DIA) datasets. ProteoSushi integrates quantitative information from all forms of a modified peptide covering the same site including charge states, missed cleavages, and other variable modifications such as methionine oxidation and asparagine deamidation found in the peptide sequence, to a single value, greatly simplifying statistics and interpretation. ProteoSushi assigns each peptide to all shared proteins and genes based on the peptide spectral match search database output or peptide list, prioritizing those with higher UniProt annotation scores for the most comprehensive annotation to balance the inherent ambiguity of assignment to proteins and genes that share the peptide sequence. In addition, ProteoSushi can prioritize user-defined lists of proteins for situations such as enrichment of a specific organelle where a subset of homologous proteins may be expected to be present. For quantitation, ProteoSushi imports the quantitative results from search engine outputs or DIA-MS, and can either sum or average the intensities. Lastly, ProteoSushi annotates each modified peptide based on the most up-to-date UniProt knowledgebase³⁰ retrieved in real time via a web API. This includes residue-specific annotations such as known PTMs, metal binding, or mutagenesis; protein domain-level annotations such as secondary structure and the protein domain the modification resides in; and protein-level annotation such as function and subcellular location.

ProteoSushi is freely available at <http://github.com/HeldLab/ProteoSushi> and utilizes a graphical user interface (GUI) for ease of use (**Fig. 2**). A computer running ProteoSushi needs at least 8 gigabytes of RAM (12 gigabytes on a Windows-based system), Python, and a stable internet connection to retrieve UniProt information. A detailed online tutorial including step-by-step instructions is available to guide the user through the installation and use on the ProteoSushi github (<https://github.com/HeldLab/ProteoSushi>). Example files are included as well (<https://tinyurl.com/y5gtk8y9>). A flow chart detailing each assignment step and quantitation integration of modified peptides is included as **Supp. Fig. S1**.

Results and Discussion

Identification: Parsing the modifications of interest and assigning modified peptides to proteins, genes and sites

ProteoSushi requires two input files: 1) a FASTA file, typically the same file used for the database search such as a UniProt reference proteome, and 2) a list of peptides, both with and without modifications, such as the modified peptide sequences detected by a database search (**Fig. 2**). Alternatively, the peptide lists, from Skyline or from a supplementary table for example, can be used as input for annotation and quantification. Quantative values can be optionally included. The unrestricted nature of ProteoSushi inputs gives it the flexibility to be used for any peptide-centric dataset including any species and any database search tool. For convenience, ProteoSushi is designed to directly accept input from the output files from MaxQuant³¹ and Mascot³², two widely used database search engines, but also accepts any CSV file with columns for modified and unmodified peptides. Specific information about this format is provided on the GitHub page (<https://github.com/HeldLab/ProteoSushi>).

ProteoSushi can process any mass shift(s) and modified residue(s), and thus can be applied to any PTM or chemical modification of interest including cases where several modifications are of interest, such as phosphorylation which can occur on serine, threonine, or tyrosine⁹. ProteoSushi can also be used for chemoproteomic results, as in the case of differential alkylation of cysteines, where reduced and oxidized cysteines are labeled with distinct chemical reagents³³. Each modification present in the peptide list is retrieved by ProteoSushi and

presented as an option in the GUI to select as a modification target. Additional modifications that are not the focus, such as chemical artifacts or routine cysteine alkylation, are merged together for quantitation along with missed cleaved and multiply charged forms of the peptide.

ProteoSushi assigns each peptide to all proteins with a matching sequence in the user-provided FASTA proteome using the amino acid sequence of each modified peptide. Leucines and isoleucines, which are isobaric and can't be distinguished through mass spectrometry, are considered equivalent. Proteins in UniProt reference proteome are not equally likely studied and characterized resulting in a wide range of annotation quality and depth. Therefore, ProteoSushi compares the UniProt annotation scores of all matched proteins and annotates the peptide with the available information of those with the highest score (**Supp. Fig. S1**). This approach minimizes over-annotation and prioritizes assignment to better characterized proteins. This is especially useful for species that are not annotated well. The primary gene name assigned to the UniProt ID is used to assign the gene, and the modified residue(s) of the peptide are assigned to the amino acid number in the UniProt amino acid sequence.

ProteoSushi provides an option for a user-input list of gene names to prioritize identification to a specific subset of candidate proteins. Potential applications of this include organelle or protein complex enrichment prior to PTM enrichment, such as mitochondrial fractionation followed by enrichment for lysine acetylation⁴. In this case, the user can input a list of mitochondrial genes and ProteoSushi will prioritize assignment of peptides to these genes instead of cytoplasmic proteins that may have sequence homology. The ProteoSushi output indicates that the assigned proteins are in the target list to simplify follow-up analysis.

To compare how proteins and genes are assigned by ProteoSushi versus typical protein-centric database search software, the data output from MaxQuant were compared before and after ProteoSushi processing. A publicly available proteomic dataset that enriched and quantified the reversible oxidation of ~4,000 cysteines²⁹ was used for evaluation. Focusing on the carbamidomethylated peptide YDGYTSC(ca)PLVTGYYNR, and its missed cleaved form KYDGYTSC(ca)PLVTGYYNR, the MaxQuant output has more than twelve rows of information for different peptide spectral matches (PSMs), including different charge states (**Fig. 3**). Typical database search results also include additional peptide forms with chemical modifications. The

peptide is assigned to two proteins with UniProt IDs H3BNP9 and Q9Y6N5, with UniProt annotation scores of 1 and 5, respectively, indicating that the level of annotation of Q9Y6N5 is better. However, since MaxQuant uses the identified proteins to prioritize protein assignment, the poorly annotated H3BNP9 is indicated as the leading protein. In contrast, ProteoSushi lists Q9Y6N5 as the assigned protein based on its higher UniProt annotation score. ProteoSushi also provides the gene name, site assignment of the modified residue and minimizes redundancy by collapsing missed cleaved peptides, multiple charge states, and chemical modifications down to a single row (**Fig. 3**). In addition, ProteoSushi biologically annotates the modified site and protein. Notably, C379 is only annotated in a disulfide bond in Q9Y6N5 as opposed to H3BNP9, highlighting the benefits of annotation score prioritization.

Site-specific data quantitation for each modified residue

ProteoSushi simplifies the quantitative results by integrating results from all peptide modified sequences that have the same PTM site(s) such as multiple charges, missed cleaved peptides, and additional variable modifications such as methionine oxidation or deamidation (**Fig. 3**). The user has an option to sum the areas, which would be appropriate for DIA results, isotopic labelling and label free quantification, or averaging the results, which is a useful option for the reporter ion ratios reported by isobaric tagging studies. Quantitative results are reported for each identified PTM site (**Fig. 3**).

Biological annotation of PTM sites: Mining knowledge at the protein-, domain-, and site-specific level

Database search engines do not provide annotation of the identified proteins, and downstream proteomic tools that are protein-centric do not provide site-specific biological annotations (**Fig. 3**). Up-to-date annotation allows researchers to make the best use of current knowledge for scientific discoveries. For example, it was demonstrated that in the majority of the biological pathway analysis studies the full biological impact was missed due to outdated gene annotation used by the analysis tools³⁴. ProteoSushi facilitates knowledge mining of modification-specific datasets by using UniProt's SPARQL endpoint to provide residue- and protein-specific

information for each identified site (**Fig. 4**). This approach enables ProteoSushi to provide up to date annotations on the fly for any of the nearly 200,000,000 entries in UniProt expressed in more than 1,000,000 species. Site-specific annotated features include whether the site is known to be a modified active site, binds known ligands, has functional consequences when mutated and others (**Fig. 4A**). Domain-level annotations include the domain type, secondary structure and topology that the modified site resides in. Protein-level annotations include subcellular location and function. UniProt includes known annotations for many different PTMs as well (**Fig. 4B**). Many these are rapidly changing, such as succinyl lysine annotation in human UniProt entries (**Fig. 4C**), despite no increase in total human protein entries in UniProt over time (**Fig. 4D**), underscoring the importance of utilizing updated annotations. Similar increases in UniProt annotations over time are observed for many residue and protein-level annotations (**Fig. 4E**), including protein secondary structure (**Fig. 4F**).

Conclusions

We report ProteoSushi, an open source, platform independent software tool that assigns, quantifies, and annotates peptide-centric proteomics data in a single tool. ProteoSushi fills the void between database search engines and discovery. It is focused on user flexibility, accepting generic peptide lists as input and thus is compatible with any database search engine. It can be applied to all peptide modifications and species, and can process data from all common quantitative mass spectrometry platforms. ProteoSushi complements the many PTM databases and PTM-focused tools available for functional prediction³⁵ and text mining^{36,37}. ProteoSushi can be implemented in any lab and operating system, and has a GUI for ease of use. By facilitating knowledge extraction from modification-specific datasets, ProteoSushi enables better understanding of the function of PTMs and the regulation of signal transduction networks.

Figure Legends

Figure 1. ProteoSushi overview. ProteoSushi processes modification-specific proteomics datasets typically enriched for peptides with a single, or a small number, of PTMs or chemical

modifications. This includes phosphorylation (P), acetylation (Ac), ubiquitylation (Ub), as well as chemical trapping of cysteine modifications. ProteoSushi requires several input files, including a list of peptides with modifications, and a proteome FASTA file. ProteoSushi then assigns PTMs to shared proteins and genes, reducing complexity by combining forms of the peptides sharing the same PTMs with additional chemical modifications, cleavage sites, or charge states. The UniProt SPARQL endpoint is used to provide site-, domain-, and protein-specific biological annotations.

Figure 2. ProteoSushi graphical user interface (GUI). The GUI includes options to analyze generic peptide lists, or directly process the output files from MaxQuant or Mascot. A prompt enables loading of the proteome FASTA file and flexible options for the user to tailor processing based on species, quantitation, peptide false discovery rate filtering and other options.

Figure 3. Quantitative proteomics data processing with ProteoSushi. A) A typical search result from a protein-centric database search engine such as MaxQuant for the carbamidomethylated (ca) peptide YDGYTSC(ca)PLVTGYYNR, and its missed cleaved form KYDGYTSC(ca)PLVTGYYNR. Each peptide spectral match is listed in a new row. **B)** ProteoSushi output that collapses peptide forms with different charge states and missed cleavages, biological annotation from UniProt, and intensity values from each PSM combined into a single value.

Figure 4. ProteoSushi retrieves UniProt annotations for PTM-sites and modified proteins. A) ProteoSushi retrieves UniProt Annotations on the fly via the SPARQL endpoint for the PTM site-, protein domain-, and protein-level biological annotations indicated. **B)** A subset of the PTMs annotated in UniProt. **C)** Number of annotated human UniProt (Swiss-Prot) entries for succinyl lysine, **D)** total human protein entries in UniProt (Swiss-Prot), **E)** miscellaneous UniProt (Swiss-Prot) annotations for proteins and sites, including annotation of mutagenesis, and **F)** protein secondary structure.

References

- (1) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; Maccoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How Many Human Proteoforms Are There? *Nat. Chem. Biol.* **2018**, *14* (3), 206–214. <https://doi.org/10.1038/nchembio.2576>.

- (2) Zhao, Y.; Jensen, O. N. Modification-Specific Proteomics: Strategies for Characterization of Post-Translational Modifications Using Enrichment Techniques. *Proteomics* **2009**, *9* (20), 4632–4641. <https://doi.org/10.1002/pmic.200900398>.

- (3) Held, J. M. Redox Systems Biology: Harnessing the Sentinels of the Cysteine Redoxome. *Antioxid. Redox Signal.* **2019**, *32* (10), 659–676. <https://doi.org/10.1089/ars.2019.7725>.

- (4) Rardin, M. J.; Newman, J. C.; Held, J. M.; Cusack, M. P.; Sorensen, D. J.; Li, B.; Schilling, B.; Mooney, S. D.; Kahn, C. R.; Verdin, E.; Gibson, B. W. Label-Free Quantitative Proteomics of the Lysine Acetylome in Mitochondria Identifies Substrates of SIRT3 in Metabolic Pathways. *Proc Natl Acad Sci U S A* **2013**, *110* (16), 6601–6606. <https://doi.org/10.1073/pnas.1302961110>.

- (5) Drake, P. M.; Schilling, B.; Niles, R. K.; Prakobphol, A.; Li, B.; Jung, K.; Cho, W.; Braten, M.; Inerowicz, H. D.; Williams, K.; Albertolle, M.; Held, J. M.; Iacovides, D.; Sorensen, D. J.; Griffith, O. L.; Johansen, E.; Zawadzka, A. M.; Cusack, M. P.; Allen, S.; Gormley, M.; Hall, S. C.; Witkowska, H. E.; Gray, J. W.; Regnier, F.; Gibson, B. W.; Fisher, S. J. Lectin Chromatography/Mass Spectrometry Discovery Workflow Identifies Putative Biomarkers

- of Aggressive Breast Cancers. *J. Proteome Res.* **2012**, *11* (4), 2508–2520.
<https://doi.org/10.1021/pr201206w>.
- (6) Clements, J. L.; Pohl, F.; Muthupandi, P.; Rogers, S. C.; Mao, J.; Doctor, A.; Birman, V. B.; Held, J. M. A Clickable Probe for Versatile Characterization of S-Nitrosothiols. *Redox Biol.* **2020**, *37* (August), 101707. <https://doi.org/10.1016/j.redox.2020.101707>.
- (7) Yang, J.; Gupta, V.; Carroll, K. S.; Liebler, D. C. Site-Specific Mapping and Quantification of Protein S-Sulphenylation in Cells. *Nat. Commun.* **2014**, *5* (1), 4776.
<https://doi.org/10.1038/ncomms5776>.
- (8) Held, J. M.; Gibson, B. W. Regulatory Control or Oxidative Damage? Proteomic Approaches to Interrogate the Role of Cysteine Oxidation Status in Biological Processes. *Mol. Cell. Proteomics* **2012**, *11* (4), R111.013037.
<https://doi.org/10.1074/mcp.R111.013037>.
- (9) Zawadzka, A. M.; Schilling, B.; Held, J. M.; Sahu, A. K.; Cusack, M. P.; Drake, P. M.; Fisher, S. J.; Gibson, B. W. Variation and Quantification among a Target Set of Phosphopeptides in Human Plasma by Multiple Reaction Monitoring and SWATH-MS2 Data-Independent Acquisition. *Electrophoresis* **2014**, *35* (24), 3487–3497.
<https://doi.org/10.1002/elps.201400167>.
- (10) Studer, R. A.; Rodriguez-Mias, R. A.; Haas, K. M.; Hsu, J. I.; Viéitez, C.; Solé, C.; Swaney, D. L.; Stanford, L. B.; Liachko, I.; Böttcher, R.; Dunham, M. J.; De Nadal, E.; Posas, F.; Beltrao, P.; Villén, J. Evolution of Protein Phosphorylation across 18 Fungal Species. *Science* (80-.). **2016**, *354* (6309), 229–232. <https://doi.org/10.1126/science.aaf2144>.
- (11) Narasumani, M.; Harrison, P. M. Discerning Evolutionary Trends in Post-Translational Modification and the Effect of Intrinsic Disorder: Analysis of Methylation, Acetylation and

- Ubiquitination Sites in Human Proteins. *PLoS Comput. Biol.* **2018**, 14 (8), 1–26.
<https://doi.org/10.1371/journal.pcbi.1006349>.
- (12) Hu, Y.; Sopko, R.; Chung, V.; Foos, M.; Studer, R. A.; Landry, S. D.; Liu, D.; Rabinow, L.; Gnäd, F.; Beltrao, P.; Perrimon, N. IProteinDB: An Integrative Database of Drosophila Post-Translational Modifications. *G3 Genes, Genomes, Genet.* **2019**, 9 (1), 1–11.
<https://doi.org/10.1534/g3.118.200637>.
- (13) Cruz, E. R.; Nguyen, H.; Nguyen, T.; Wallace, I. S. Functional Analysis Tools for Post-Translational Modification: A Post-Translational Modification Database for Analysis of Proteins and Metabolic Pathways. *Plant J.* **2019**, 99 (5), 1003–1013.
<https://doi.org/10.1111/tpj.14372>.
- (14) Xu, H.; Wang, Y.; Lin, S.; Deng, W.; Peng, D.; Cui, Q.; Xue, Y. PTMD: A Database of Human Disease-Associated Post-Translational Modifications. *Genomics, Proteomics Bioinforma.* **2018**, 16 (4), 244–251. <https://doi.org/10.1016/j.gpb.2018.06.004>.
- (15) Mooradian, A. D.; Held, J. M.; Naegle, K. M. Using ProteomeScout: A Resource of Post-Translational Modifications, Their Experiments, and the Proteins That They Annotate. *Curr. Protoc. Bioinforma.* **2017**, 59 (1), 13.32.1-13.32.27. <https://doi.org/10.1002/cpbi.31>.
- (16) Huang, H.; Arighi, C. N.; Ross, K. E.; Ren, J.; Li, G.; Chen, S. C.; Wang, Q.; Cowart, J.; Vijay-Shanker, K.; Wu, C. H. IPTMnet: An Integrated Resource for Protein Post-Translational Modification Network Discovery. *Nucleic Acids Res.* **2018**, 46 (D1), D542–D550.
<https://doi.org/10.1093/nar/gkx1104>.
- (17) Minguez, P.; Letunic, I.; Parca, L.; Garcia-Alonso, L.; Dopazo, J.; Huerta-Cepas, J.; Bork, P. PTMcode v2: A Resource for Functional Associations of Post-Translational Modifications within and between Proteins. *Nucleic Acids Res.* **2015**, 43 (D1), D494–D502.

<https://doi.org/10.1093/nar/gku1081>.

- (18) Hornbeck, P. V.; Zhang, B.; Murray, B.; Kornhauser, J. M.; Latham, V.; Skrzypek, E. PhosphoSitePlus, 2014: Mutations, PTMs and Recalibrations. *Nucleic Acids Res.* **2015**, *43* (D1), D512–D520. <https://doi.org/10.1093/nar/gku1267>.
- (19) Chalkley, R. J.; Clauser, K. R. Modification Site Localization Scoring: Strategies and Performance. *Mol. Cell. Proteomics* **2012**, *11* (5), 3–14. <https://doi.org/10.1074/mcp.R111.015305>.
- (20) Devabhaktuni, A.; Lin, S.; Zhang, L.; Swaminathan, K.; Gonzalez, C. G.; Olsson, N.; Pearlman, S. M.; Rawson, K.; Elias, J. E. TagGraph Reveals Vast Protein Modification Landscapes from Large Tandem Mass Spectrometry Datasets. *Nat. Biotechnol.* **2019**, *37* (4), 469–479. <https://doi.org/10.1038/s41587-019-0067-5>.
- (21) Yu, F.; Teo, G. C.; Kong, A. T.; Haynes, S. E.; Avtonomov, D. M.; Geiszler, D. J.; Nesvizhskii, A. I. Identification of Modified Peptides Using Localization-Aware Open Search. *Nat. Commun.* **2020**, *11* (1), 1–9. <https://doi.org/10.1038/s41467-020-17921-y>.
- (22) Beausoleil, S. A.; Villén, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285–1292. <https://doi.org/10.1038/nbt1240>.
- (23) Preston, G. W.; Yang, L.; Phillips, D. H.; Maier, C. S. Visualisation Tools for Dependent Peptide Searches to Support the Exploration of in Vitro Protein Modifications. *PLoS One* **2020**, *15* (7), 1–16. <https://doi.org/10.1371/journal.pone.0235263>.
- (24) Raaijmakers, L. M.; Giansanti, P.; Possik, P. A.; Mueller, J.; Peeper, D. S.; Heck, A. J. R.; Altelaar, A. F. M. PhosphoPath: Visualization of Phosphosite-Centric Dynamics in

- Temporal Molecular Networks. *J. Proteome Res.* **2015**, *14* (10), 4332–4341.
<https://doi.org/10.1021/acs.jproteome.5b00529>.
- (25) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus Computational Platform for Comprehensive Analysis of (Prote)Omics Data. *Nat. Methods* **2016**, *13* (9), 731–740. <https://doi.org/10.1038/nmeth.3901>.
- (26) Krug, K.; Mertins, P.; Zhang, B.; Hornbeck, P.; Raju, R.; Ahmad, R.; Szucs, M.; Mundt, F.; Forestier, D.; Jane-Valbuena, J.; Keshishian, H.; Gillette, M. A.; Tamayo, P.; Mesirov, J. P.; Jaffe, J. D.; Carr, S. A.; Mani, D. R. A Curated Resource for Phosphosite-Specific Signature Analysis. *Mol. Cell. Proteomics* **2019**, *18* (3), 576–593.
<https://doi.org/10.1074/mcp.TIR118.000943>.
- (27) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**.
<https://doi.org/10.1093/bioinformatics/btq054>.
- (28) Held, J. M.; Schilling, B.; D’Souza, A. K.; Srinivasan, T.; Behring, J. B.; Sorensen, D. J.; Benz, C. C.; Gibson, B. W. Label-Free Quantitation and Mapping of the ErbB2 Tumor Receptor by Multiple Protease Digestion with Data-Dependent (MS1) and Data-Independent (MS2) Acquisitions. *Int. J. Proteomics* **2013**, *2013*, 1–11. <https://doi.org/10.1155/2013/791985>.
- (29) Behring, J. B.; van der Post, S.; Mooradian, A. D.; Egan, M. J.; Zimmerman, M. I.; Clements, J. L.; Bowman, G. R.; Held, J. M. Spatial and Temporal Alterations in Protein Structure by EGF Regulate Cryptic Cysteine Oxidation. *Sci. Signal.* **2020**, *13* (615), eaay7315. <https://doi.org/10.1126/scisignal.aay7315>.
- (30) The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids*

Res. **2019**, 47 (D1), D506–D515. <https://doi.org/10.1093/nar/gky1049>.

- (31) Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, 26 (12), 1367–1372. <https://doi.org/10.1038/nbt.1511>.
- (32) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999**, 20 (18), 3551–3567. [https://doi.org/10.1002/\(SICI\)1522-2683\(19991201\)20:18<3551::AID-ELPS3551>3.0.CO;2-2](https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2).
- (33) Held, J. M.; Danielson, S. R.; Behring, J. B.; Atsriku, C.; Britton, D. J.; Puckett, R. L.; Schilling, B.; Campisi, J.; Benz, C. C.; Gibson, B. W. Targeted Quantitation of Site-Specific Cysteine Oxidation in Endogenous Proteins Using a Differential Alkylation and Multiple Reaction Monitoring Mass Spectrometry Approach. *Mol. Cell. Proteomics* **2010**, 9 (7), 1400–1410. <https://doi.org/10.1074/mcp.m900643-mcp200>.
- (34) Wadi, L.; Meyer, M.; Weiser, J.; Stein, L. D.; Reimand, J. Impact of Outdated Gene Annotations on Pathway Enrichment Analysis. *Nat. Methods* **2016**, 13 (9), 705–706. <https://doi.org/10.1038/nmeth.3963>.
- (35) Torres, M. P.; Dewhurst, H.; Sundararaman, N. Proteome-Wide Structural Analysis of PTM Hotspots Reveals Regulatory Elements Predicted to Impact Biological Function and Disease. *Mol. Cell. Proteomics* **2016**, 15 (11), 3513–3528. <https://doi.org/10.1074/mcp.M116.062331>.
- (36) Torii, M.; Arighi, C. N.; Li, G.; Wang, Q.; Wu, C. H.; Vijay-Shanker, K. RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2015**, 12 (1),

17–29. <https://doi.org/10.1109/TCBB.2014.2372765>.

- (37) Tudor, C. O.; Ross, K. E.; Li, G.; Vijay-Shanker, K.; Wu, C. H.; Arighi, C. N. Construction of Phosphorylation Interaction Networks by Text Mining of Full-Length Articles Using the EFIP System. *Database* **2015**, *2015*, 1–16. <https://doi.org/10.1093/database/bav020>.

Acknowledgements

We acknowledge funding from R01 CA200893 (J.H). SVDP is supported by the SSMF foundation (P17-0060) We acknowledge Tristaan Haddad for evaluating ProteoSushi output files.

Supporting Info

Supplemental Figure S1. Rules for peptide assignment and merging multiple peptide forms.

Flow chart detailing how ProteoSushi assigns peptides to shared proteins and genes as well as combines multiple peptide forms.

Abbreviations

PTM: post-translational modification

GUI: graphical user interface

DIA: data-independent acquisition

API: application programming interface

FASTA: fast all

PSM: peptide spectral match

SPARQL: SPARQL Protocol and RDF Query Language

P: phosphorylation

Ac: acetylation

Ub: ubiquitylation

FDR: false discovery rate

Required
Input Files

Types of ProteoSushi Files

☒ MaxQuant

MaxQuant txt Folder

/home/rob/Documents/Held_Lab/ProteoSushi/teosushi/examples/txt

☐ Mascot

☐ Generic

PTMs for Analysis

☐ Acetyl (Protein N-term)

☐ Carbamidomethyl (C)

☐ NEM (C)

☐ Oxidation (M)

Proteome File

Uniprot Proteome FASTA

[Filepath]

Output Location

Choose Output Location

[Filepath]

Required Output Location

Options

☐ Use Target Genes

☐ Use Quantitation Values

☐ Annotate with Uniprot

Species ID

9606

Homo sapiens

Max Missed Cleavages

2

FDR Threshold

Protease used in sample digestion

trypsin/p

Rollup!

Optional
Analysis
Settings

Optional
FDR

Required Analysis Info

A) Typical MS2 data search output

YDGYTSC(ca)PLVTGYYNR

Multiple peptide forms with the same modified residue

Sequence	Modified sequence	# Missed Cleavages	z	m/z	PEP	Gene name(s)	Protein(s)	Leading razor protein	Protein name(s)	Intensity
YDGYTSCPLVTGYNR	YDGYTSC(ca)PLVTGYNR	0	2	883.40	0	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	8733
YDGYTSCPLVTGYNR	YDGYTSC(ca)PLVTGYNR	0	2	883.40	4.5E-70	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	4789
YDGYTSCPLVTGYNR	YDGYTSC(ca)PLVTGYNR	0	2	883.40	1.0E-51	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	5003
YDGYTSCPLVTGYNR	YDGYTSC(ca)PLVTGYNR	0	2	883.40	9.4E-245	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	7415
YDGYTSCPLVTGYNR	YDGYTSC(ca)PLVTGYNR	0	2	883.40	1.0E-273	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	6816
YDGYTSCPLVTGYNR	YDGYTSC(ca)PLVTGYNR	0	2	883.40	2.1E-130	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	7910
KYDGYTSCPLVTGYNR	KYDGYTSC(ca)PLVTGYNR	1	3	631.97	-1.7E-03	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	1822
KYDGYTSCPLVTGYNR	KYDGYTSC(ca)PLVTGYNR	1	3	631.97	-2.4E-03	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	1917
KYDGYTSCPLVTGYNR	KYDGYTSC(ca)PLVTGYNR	1	3	631.97	-1.4E-03	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	3341
KYDGYTSCPLVTGYNR	KYDGYTSC(ca)PLVTGYNR	1	3	631.97	1.4E-04	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	1519
KYDGYTSCPLVTGYNR	KYDGYTSC(ca)PLVTGYNR	1	3	631.97	4.1E-05	SQRDL	H3BNP9;Q9Y6N5	H3BNP9	Sulfide:quinone oxidoreductase	2250

No site assignment

multiple missed cleavages

multiple charge states

Peptide is shared between Uniprot IDs with different annotation scores

Leading protein UniProt annotation score is 1

Multiple intensity values need to be combined for analysis

B) ProteoSushi output

YDGYTSC(ca)PLVTGYYNR

Quantitation from all forms of the modified site in a single row

Peptide Sequence	Peptide Modified Sequence	Gene	Protein Name	Uniprot ID	Site	Shared Genes	UniProt Annotation Score	Active Site Annotation	Disulfide Bond Annotation	Secondary Structure	Subcellular Location	Enzyme Class	Rhea	Intensity (summed)
YDGYTSCPLVTGYNR	YDGYTSC(ca)PLVTGYNR	SQRDL	Sulfide:quinone oxidoreductase	Q9Y6N5	379	-	5	Cysteine persulfide intermediate	Disulfide Bond	Beta Strand	Mitochondrion	Eukaryotic sulfide quinone	55156 38359 62608	41248

All gene/protein/site assignments

Uniprot annotations for the protein and PTM sites

Peak areas summed or averaged for each PTM site(s)

a**PTM site-level**

Active site

Known PTMs

Metal or small molecule binding

Function of residue mutation

Natural variants

Zinc finger

Protein domain-level

Protein domain

Chain annotation

Secondary structure

Composition bias

Topology

Protein-level

Subcellular location

Catalytic activity

Enzyme class

Rhea reaction annotation

Alternative sequences

b Annotated PTMs

Acetylation

Amidation

Citrullination

Crotonylation

Disulfide bond

Glycosylation

Hydroxylation

Malonylation

Methylation

Nitrosation

Oxidation

Phosphorylation

Propionylation

Succinlation

Sulfation

Sulfenation

Ubiquitylation

c**Succinyl (K)** **e****Annotations** **d****Protein Entries** **f****Secondary Structure**