

# Informed-Proteomics: open-source software package for top-down proteomics

Jungkap Park<sup>1</sup>, Paul D Piehowski<sup>1</sup> , Christopher Wilkins<sup>1</sup>, Mowei Zhou<sup>2</sup> , Joshua Mendoza<sup>1</sup>, Grant M Fujimoto<sup>1</sup>, Bryson C Gibbons<sup>1</sup>, Jared B Shaw<sup>2</sup>, Yufeng Shen<sup>1</sup>, Anil K Shukla<sup>1</sup>, Ronald J Moore<sup>1</sup>, Tao Liu<sup>1</sup>, Vladislav A Petyuk<sup>1</sup> , Nikola Tolic<sup>2</sup>, Ljiljana Paša-Tolić<sup>2</sup>, Richard D Smith<sup>1</sup> , Samuel H Payne<sup>1</sup>  & Sangtae Kim<sup>1,3</sup> 

**Top-down proteomics, the analysis of intact proteins in their endogenous form, preserves valuable information about post-translation modifications, isoforms and proteolytic processing. The quality of top-down liquid chromatography–tandem MS (LC-MS/MS) data sets is rapidly increasing on account of advances in instrumentation and sample-processing protocols. However, top-down mass spectra are substantially more complex than conventional bottom-up data. New algorithms and software tools for confident proteoform identification and quantification are needed. Here we present Informed-Proteomics, an open-source software suite for top-down proteomics analysis that consists of an LC-MS feature-finding algorithm, a database search algorithm, and an interactive results viewer. We compare our tool with several other popular tools using human-in-mouse xenograft luminal and basal breast tumor samples that are known to have significant differences in protein abundance based on bottom-up analysis.**

While mass spectrometry (MS)-based proteomics has been successful for identifying and quantifying peptides and post-translational modifications (PTMs), the characterization of intact protein forms (i.e., proteoforms) remains challenging<sup>1–4</sup>. Intact protein (top-down) proteomics is more challenging at almost every stage of the analytical process—sample preparation, liquid chromatography (LC) separation, fragmentation, and data analysis<sup>5,6</sup>. The challenges and lack of confidence in data analysis are major factors preventing proteomics researchers from adopting top-down studies<sup>7</sup>. Unlike traditional bottom-up proteomics, where numerous software tools are available, only a handful of tools are available for top-down characterization, and data analysis often requires laborious manual interpretation. Here, we present an open-source software suite for top-down data analysis named Informed-Proteomics (available at <https://github.com/PNNL-Comp-Mass-Spec/Informed-Proteomics> and as **Supplementary Software**).

In general, the top-down data analysis workflow consists of three steps—feature deconvolution, protein characterization via database search of fragmentation data, and validation. In every step, there are challenges that make top-down data analysis

substantially more difficult than bottom-up data analysis. First, the size of intact proteins means that they typically have higher and more diverse charge states following electrospray ionization. This distributes the ion signal over a broader number of charge states with increasingly large isotopomer envelopes, which substantially reduces the signal-to-noise ratio. Detecting ion signals and accurately calculating precursor mass is essential to proteoform identification and quantification. Existing deconvolution tools such as THRASH<sup>8</sup>, Xtract<sup>9</sup>, and MS-Deconv<sup>10,11</sup> adopt spectrum-centric approaches and create a simplified spectrum of singly charged monoisotopic ion species.

The second challenge in top-down analysis is determining how to explore the search space of potential proteoforms. Because most proteins are post-translationally modified (e.g., through proteolytic cleavage, acetylation, etc.), the number of possible proteoforms is exponentially greater than the number of genes; for example, there are over a billion combinatorially possible proteoforms in humans. Popular top-down data analysis tools ProSightPC<sup>12,13</sup> and MS-Align+ (recently renamed to TopPIC)<sup>14,15</sup> address this challenge using different approaches. ProSightPC restricts the search space to a limited set of proteoforms in a ‘proteome warehouse’, a curated collection derived from known PTMs, splice variants, and single-nucleotide variants. While this approach has the advantages of being able to confirm known variants and accurately characterize proteoforms, it is effective only for organisms that have a well-annotated genome and a well-characterized proteome. In contrast, MS-Align+ allows ‘blind’ modifications accounting for any and all PTMs and mutations, and it uses the spectral alignment algorithm to efficiently score multiple proteoforms simultaneously. Although this blind search approach is valuable for discovering unknown PTMs and mutations, it may produce a substantial amount of false-positive proteoform spectrum matches (PrSMs). Recently, another top-down analysis software, pTop, has been developed<sup>16</sup>. In pTop, the search space is restricted by taking only expected modifications into account; however, the current version of this software cannot find cleaved or truncated proteoforms.

<sup>1</sup>Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA. <sup>2</sup>Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, USA. <sup>3</sup>Present address: Illumina Inc., San Diego, California, USA. Correspondence should be addressed to S.H.P. ([samuel.payne@pnnl.gov](mailto:samuel.payne@pnnl.gov)) or S.K. ([sangtae.kim@gmail.com](mailto:sangtae.kim@gmail.com)).

As a result of these two challenges, proteoform identifications are error prone and frequently mislocalize PTMs, identify false cleavages, or erroneously compute precursor mass. Thus, it is often necessary for users to manually validate and refine results. Additionally, quantification studies may require researchers to examine the features or extracted ion chromatogram (XIC) of precursor ions of different charge states, or even different regions of the isotopic peak distributions for protein ions. Therefore, there is high demand for top-down proteomics visualization tools to assist with such data curation<sup>12,13,16–18</sup>.

Informed-Proteomics contains a new LC-MS feature-finding algorithm (ProMex), a new database-search algorithm (MSPathFinder), and an interactive results viewer (LcMsSpectator) (Supplementary Fig. 1). We demonstrate both the identification and quantification capabilities of Informed-Proteomics and compare it with other existing tools. Key advantages of Informed-Proteomics over existing software include high accuracy in LC-MS features detection by ‘smart’ aggregation and summation of features from the same species (which, e.g., enhances measurement sensitivity); an efficient algorithm for high-throughput searching proteoforms with combinations of PTMs and truncations by reducing redundancies to minimize the search space; and an interactive visualization tool for easy and fast manual validating and refining the results.

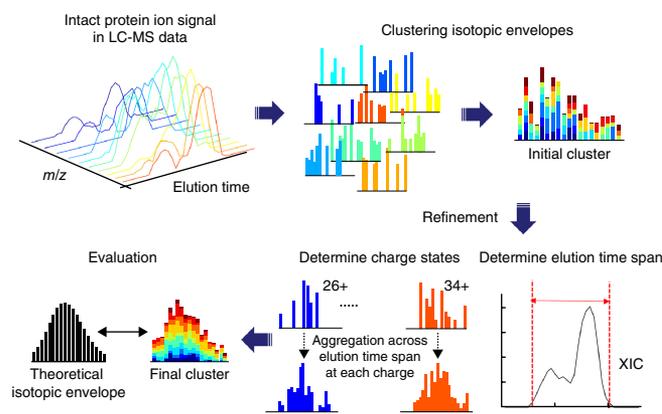
## RESULTS

### Informed-Proteomics workflow

The first component of the Informed-Proteomics software suite, ProMex, finds and characterizes putative proteoforms in LC-MS data. An LC-MS feature represents a group of isotopomer envelopes corresponding to the same putative proteoform ion across all charge states and LC elution times. Because ions are dispersed widely across LC times, charge states, and isotope species, individual isotope envelopes typically have poor shape compared with the shape of expected profiles (Supplementary Fig. 2). ProMex incorporates two key innovations to improve accuracy of feature detection (see Fig. 1). First, ProMex both aggregates signals across different charge states and explicitly uses the LC dimension to aggregate features over elution time. Some existing tools (e.g., Xtract in specific commercial implementations) also use the LC dimension by periodically averaging a fixed number of spectra at chromatographic peaks to increase the signal-to-noise ratio. ProMex explicitly looks for all isotopomer envelopes distributed over 3D LC-MS data and dynamically determines the elution time spans for every candidate mass.

Second, rather than examining individual isotopomer envelopes separately, ProMex measures the likelihood of detected LC-MS features based on the aggregated isotopomer envelope. The score is calculated by a likelihood scoring function which takes into account the aggregated isotopomer envelope shape, intensity, charge distribution, and the correlation of elution profile at different charge states (see details in Online Methods). The output of ProMex is a list of LC-MS features defined by monoisotopic mass, range of charge states, elution time span, abundance, and likelihood scores.

Detected LC-MS features are fed into the database search tool, MSPathFinder, to characterize proteoforms from MS/MS spectra. MSPathFinder operates much like bottom-up proteomics tools; it allows users to specify a set of post-translational modifications

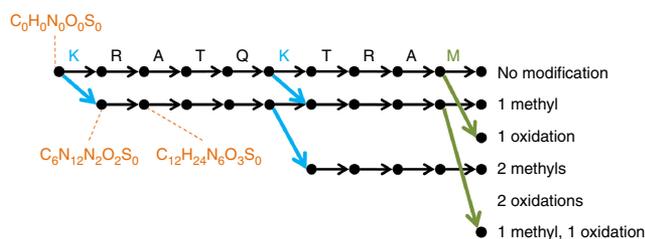


**Figure 1** | LC-MS feature finding in ProMex. An LC-MS feature refers to a group of isotopomer envelopes corresponding to the same proteoform species across all charge states and LC elution times. The ProMex algorithm begins with clustering isotopomer envelopes across adjacent time and charge state. The initial cluster is refined to accurately determine its elution time span and range of charge states. After refinement, ProMex calculates the likelihood that the final cluster is a true LC-MS feature.

and the maximum number of allowable modifications in a sequence. MSPathFinder also provides the statistical significance of PrSMs with E-values computed by the generating function approach<sup>19,20</sup>; and it provides the false discovery rate (FDR), which is estimated using the target–decoy approach<sup>21</sup>.

MSPathFinder efficiently explores the combinatorial proteoform space using a graph-based approach called the sequence graph (see Fig. 2 for illustration), which allows quick exploration of the vast number of possible proteoforms when considering variable PTMs. There are two important motivations behind the sequence graph. First, because many proteoforms differ only by the location of PTMs, the number of unique elemental compositions is much smaller than the number of possible proteoforms. Using histone H4 as an extreme example, the number of proteoforms possible when applying five modifications (acetylation, methylation, and dimethylation of Lys and Arg; trimethylation of Arg; and phosphorylation of Ser, Thr, and Tyr) is about 50 trillion; but the number of their unique elemental compositions is only 2,344. Second, many fragment compositions are shared by proteoforms with the same composition. Therefore, it is inefficient to score these proteoforms independently. The goal of the sequence-graph approach is to effectively remove such redundancies. MSPathFinder, pTop, and MS-Align+ all use similar spectral alignment algorithms<sup>14,22–25</sup> based on a parametric dynamic programming algorithm to find the best scoring proteoform in a sequence. However, MSPathFinder uniquely uses node in the sequence graph to represent a composition of atoms (mostly C, H, N, O, S) rather than a combination of modifications. Since some combinations of modifications have exactly the same atomic compositions (e.g. tri-methylation vs methylation + di-methylation), atom-centric graphs tend to be smaller, which leads to a faster running time in general.

MSPathFinder uses a second technique to efficiently explore the vast search space of intact proteoforms. As many proteoforms are enzymatically cleaved or truncated forms of proteins, allowing both N-terminal and C-terminal truncations are necessary to



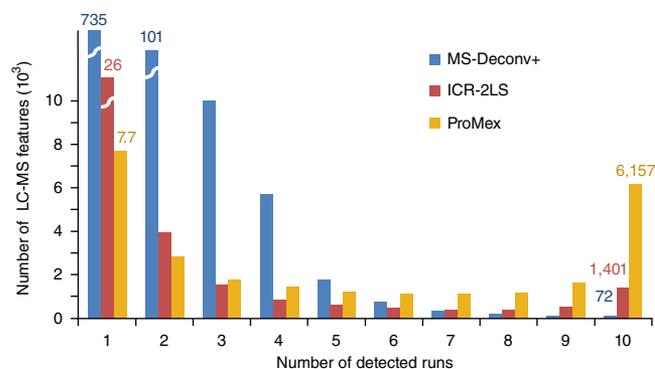
**Figure 2** | Illustration of the sequence graph for 'KRATQKTRAM'. The sequence graph compactly represents all possible proteoforms of a single atomic composition and facilitates efficient scoring of the search space. In this example, oxidized methionine and methylated lysine are considered as dynamic modifications, and up to two modifications are allowed per sequence. The graph is constructed from left to right, with the leftmost vertex (source) corresponding to  $C_0H_0N_0O_0S_0$ . The vertically aligned vertices correspond to fragments created by cleaving  $j^{\text{th}}$  and  $(j+1)^{\text{th}}$  amino acids. The horizontally aligned vertices represent the fragments with the same modifications. The black, green, and blue edges correspond to unmodified amino acids, oxidized methionine, and methylated lysine, respectively. Each vertex corresponds to a composition, and several compositions are shown for illustration. Each of the rightmost vertices (sink) is called a precursor vertex and represents a unique elemental composition of proteoforms with the specified combination of modifications. Thus a path from source to sink represents a proteoform.

identify the mature processed proteoform, but this substantially increases database search time. In order to reduce the number of possible sequence candidates which serve as query sequences in the search mode of multiple internal cleavages, we implemented a *de novo* sequencing algorithm to find short amino acid sequences, called sequence tags, as similar to a previous approach<sup>26</sup>. Once a protein matches to a sequence tag, MSPathFinder searches multiply cleaved proteoforms of the protein using two sequence graphs toward opposite terminals (**Supplementary Fig. 3**). While this tag-based approach is helpful for restricting the search space significantly, it may fail to find correct proteoforms when a sufficient number of consecutive fragment ions are not detected in MS/MS spectra.

For visualizing and analyzing top-down proteomics data, we created LcMsSpectator as a stand-alone desktop application that is fully integrated with both ProMex and MSPathFinder. This allows maximum data exploration by interacting with both the LC-MS features and MS/MS identifications. The spectral and chromatographic evidence for the search results are delivered instantly upon completion of the search for comparison with the original identifications. Sequences can also be edited in the application and scored on the fly, which makes it easy to find evidence for proteoforms that were not found in the original database search. LcMsSpectator utilizes a floatable and dockable tabbed-document interface that lets users customize various data grids and spectrum and chromatogram views (see **Supplementary Fig. 4**). It supports both automatic and assisted revision of results and identifications (see an example in **Supplementary Fig. 5**). All of the views and data plots can be exported to high-resolution, publication-ready images.

### LC-MS feature detection using ProMex

We assessed the accuracy of our feature-detection algorithm by benchmarking ProMex against other MS1 feature-detection algorithms including ICR-2LS (<http://omics.pnl.gov/software/icr-2ls>)



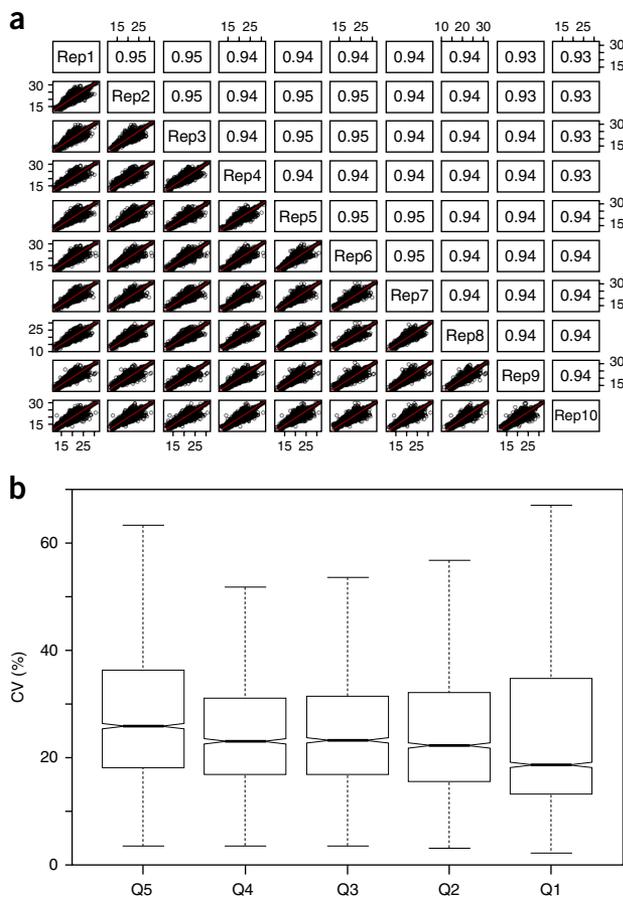
**Figure 3** | Consistency of LC-MS feature detection in ten technical replicate analyses of an ovarian tumor sample. Three deconvolution algorithms (MS-Deconv+, ICR-2LS, and ProMex) were used to identify LC-MS features present in the sample. For each tool, features were aligned across replicate data sets and merged. The chart shows how often each LC-MS feature was observed within the set of replicates. True features are expected to be present in a majority of the replicate data sets.

and MSDeconv+<sup>10,11</sup> (see details in Online Methods). For this benchmarking test, we created ten replicate LC-MS/MS data files from an ovarian tumor sample. The average running times of ICR-2LS, MSDeconv+, and ProMex were 180, 23, and 35 minutes, respectively. The metric for accuracy in this test was two-fold. Because they are replicates runs, we anticipated that true features would be present in most files and at similar retention times with similar intensities. As shown in **Figure 3**, ProMex had a significantly higher number of features detected in all ten replicates. MSDeconv+ had an overwhelming number of detected features present in only one or two data sets, and this pointed to a high variability in the data deconvolution; only 0.04% of features were found in eight or more data sets. ICR-2LS is an early implementation of the THRASH deconvolution algorithm<sup>8</sup>. Although it performed substantially better than MSDeconv+, it still had only 6% of features appear in eight or more data sets. ProMex showed the best performance in reproducible detection of LS-MS features in these replicate data sets, with 34% of features identified in eight or more data sets.

The second metric for determining the accuracy of LC-MS feature detection is quantitative reproducibility, as this ultimately defines the utility of the methodology for interrogating changes in a system of interest. **Figure 4a** shows the abundance correlation plots for ProMex identified features for ten replicate analyses. The high reproducibility of the platform is demonstrated by Pearson correlation coefficients that vary from 0.93–0.95 across all runs. Furthermore, when we applied our workflow to the ovarian tumor replicates, we were able to achieve coefficients of variation similar to those obtained in label-free bottom-up proteomics<sup>27–30</sup> (**Fig. 4b**).

### Proteoform identification using MSPathFinder

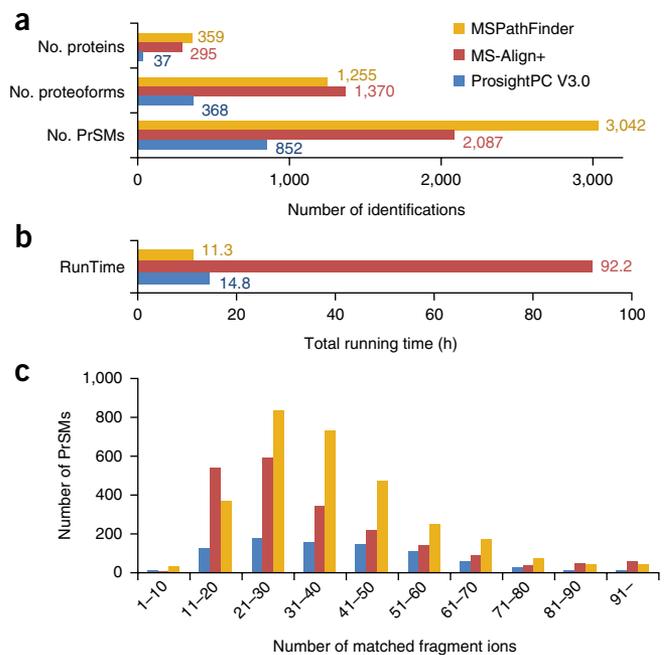
Next, we compared the performance of MSPathFinder to that of other top-down database search tools MS-Align+ (i.e., TopPIC v0.9.1)<sup>14,15</sup>, pTop v1.2 (ref. 16), and ProSightPC v3.0 (ref. 13) (see details in Online Methods). We ran each program on the same computer against an ovarian tumor replicate run containing 3,696 MS1 spectra and 4329 MS2 spectra. A human proteome sequence database (UniProt Release 2015\_10) which contains 20,209 protein



**Figure 4** | Quantitative reproducibility for LC-MS features detected in ten technical replicate LC-MS analyses. **(a)** Correlation plots for ten replicate injections of ovarian tumor tissue lysate. Each plot compares the quantitative measurement of LC-MS features between two replicates. The R-squared value summarizes the similarity between data sets, and this value was 0.93 or greater for all pairs of replicates. **(b)** Coefficient of variance (CV) for measured proteoform abundances divided by abundance quintile. Overall median CV for feature abundance was 22.3%. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5 $\times$  interquartile range; points, outliers.

sequences was used for MS-Align+, pTop, and MSPathFinder; while ProSightPC was run against the annotated human proteoform databases (2014\_07 version). PrSMs identified by MS-Align+, pTop, and MSPathFinder were controlled at FDR 1% using the same target–decoy databases. Since pTop v1.2 is not able to search cleaved proteoforms, we compared pTop and MSPathFinder separately, and we disabled internal cleavages in MSPathFinder. There was no option to run ProSightPC against user-provided target–decoy databases; we therefore used an E-value cutoff of  $1 \times 10^{-4}$ , which is the default cutoff to distinguish good and bad matches in the software.

Since each tool explores different regions of the proteoform space, it is difficult to directly compare the results. Moreover, we believe the searches are complimentary and can be used in combination to achieve the best results. MS-Align+ has the greatest search space and consequently identified the greatest number of unique proteoforms (Fig. 5a). ProSightPC has the most restrictive search space and therefore identified the fewest; this indicated that even for human samples, the annotation of known proteoforms is often incomplete. MSPathFinder showed dramatically faster run time (11.3 h) than

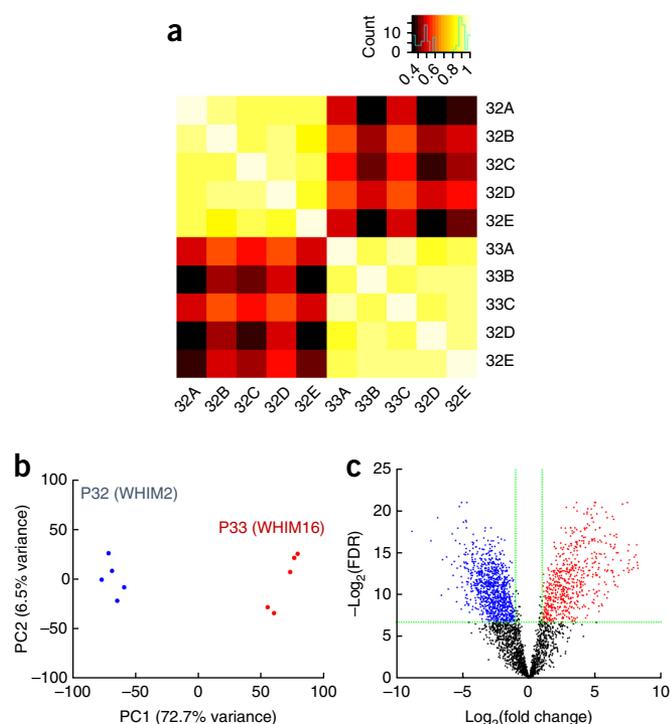


**Figure 5** | Protein identification and characterization results for a human ovarian tumor. **(a)** The number of proteins, proteoforms, and protein–spectrum matches (PrSMs) identified by ProSightPC V3.0 (E-value  $\leq 10^{-4}$ ), MS-Align+ (1% FDR), and MSPathFinder (1% FDR). **(b)** Total running time for deconvolution and database search. **(c)** Histogram of the number of matched fragment ions.

MS-Align+ (92.2 h) and comparable run time with that of ProSightPC (14.8 h) (Fig. 5b). The slower run time for MS-Align+ was expected because it has a larger search space. In the comparison with pTop, MSPathFinder found 10–20% more proteins, proteoforms, and protein–spectrum matches (PrSMs) than pTop (Supplementary Table 1). While the total running time of MSPathFinder was longer than that of pTop because of the running time of ProMex, MSPathFinder showed a faster running time in database search than pTop. Finally, when we look at the number of annotated peaks in an identified spectrum as a proxy measure for the quality of identifications, MSPathFinder annotates significantly more peaks per spectrum than either TopPIC or ProSightPC (Fig. 5c).

### Label-free quantification

Lastly, we applied our top-down proteomics workflow for label-free quantification of human-in-mouse xenograft breast tumor samples<sup>31</sup> previously characterized by the Clinical Proteomic Tumor Analysis Consortium<sup>32</sup>. Two subtypes of breast cancer tumors, basal like (WHIM2-P32) and luminal B (WHIM16-P33), were analyzed. We created five technical replicate analyses for each subtype. First, the LC-MS features detected across all ten replicates runs were quantified and aligned, and then statistical significance tests were performed to find differentially expressed LC-MS features in the two breast tumor subtypes, WHIM2 and WHIM16 (see Online Methods). There were a total of 7,300 differentially expressed LC-MS features at adjusted  $P$  value of  $<0.01$  and a fold change of  $>1$  (one-way ANOVA, Benjamini-Hochberg adjusted,  $n = 17,870$ ) (Supplementary Fig. 6). Next, we quantified the differential expression of identified proteoforms (Fig. 6). Among a total of 3,207 proteoforms identified in WHIM2



**Figure 6** | Differentially expressed proteoforms in CompRef breast tumor sample. Two breast cancer xenograft tumors representing basal-like (P32) and luminal B (P33) subtypes were compared to identify differentially expressed proteoforms. Five technical replicate LC-MS/MS analyses were performed for each tumor. (a) Pearson correlation plot, (b) PCA plot, and (c) volcano plot. 622 proteoforms are upregulated in P32 (WHIM2), while 1,014 proteoforms are upregulated in P33 (WHIM16) at 1% FDR and fold change  $>2$ .

and WHIM6 samples, 1,636 proteoforms were found to be differentially expressed with adjusted  $P$  value of  $<0.01$  and fold change of  $>2$ .

Recently, an integrated approach of bottom-up and top-down proteomics to detect differentially expressed protein and proteoforms was reported for this same tumor comparison<sup>33</sup>. In both LC-MS feature- and proteoform-level analysis, we found ten times more differentially expressed entities than were found by the approach described in ref. 33. Furthermore, we achieved this characterization using only 30 h of instrument time as compared to the 200 h reported in ref. 33.

We reanalyzed the same data set used in one of the studies in ref. 33 using Informed-Proteomics. Using the same statistical model, we found 412 differentially expressed proteoforms mapping to 280 proteins with adjusted  $P$  value of  $<0.01$  and absolute  $\log_2$  fold change of  $>1$ . Our analysis pipeline found 2.7 and 2.4 times more differentially expressed proteoforms and proteins, respectively, compared to those found in ref. 33 (Supplementary Fig. 7).

## DISCUSSION

Informed-Proteomics includes an LC-MS feature-finding algorithm (ProMex), a database search algorithm (MSPathFinder), and an interactive results viewer (LcMsSpectator). Our open-source software suite is designed for sensitive and comprehensive high-throughput analysis of complex mixtures of intact proteins. We demonstrate how our tools can identify

differently expressed LC-MS features and proteoforms from breast cancer samples.

ProMex aggregates signals not only across different charge states, but also over LC time, such that it measures the likelihood of detected LC-MS features based on the aggregated isotopomer envelope. It relies on isotopically resolved peaks and is designed for high-resolution LC-MS data. Therefore, it does not currently work for data with only a charge envelope or for data without an LC separation. We demonstrated that ProMex accurately recovers more common features and less uncommon (irreproducible) features across multiple replicate runs than do other existing algorithms. We also showed that our database search algorithm efficiently explores the combinatorial proteoform space using a graph-based approach called the sequence graph. MSPathFinder operates in a manner similar to that of most common bottom-up proteomics algorithms, requiring users to enumerate specific post-translational modifications of interest. It does not discover unknown modifications, which can be done with complementary algorithms for open PTM search.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGEMENTS

Portions of this work were supported by the NIH National Institute of General Medical Sciences grant GM103493 (R.D.S.), the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) grant U24CA160019 (R.D.S.), the National Institute of Allergy and Infectious Diseases NIH/DHHS through interagency agreement Y1-A1-8401-01 (J. Adkins, PNNL), and the U.S. Department of Energy (DOE) Office of Science and Office of Biological and Environmental Research, under the Pan-omics program (R.D.S.). L.P.T., N.T., M.Z., and J.B.S. were supported as part of the “High Resolution and Mass Accuracy Capability” development project at the Environmental Molecular Science Laboratory (EMSL), a U.S. DOE national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. Battelle operates PNNL for the DOE under contract DE-AC05-76RL001830.

## AUTHOR CONTRIBUTIONS

J.P., P.D.P., S.H.P., and S.K. designed and executed the study. J.P., C.W., J.M., G.M.F., B.C.G., and S.K. implemented algorithms in software. T.L. contributed samples. P.D.P., Y.S., A.K.S., R.J.M. performed LC-MS/MS experiments. J.P., P.D.P., J.B.S., V.A.P., M.Z., T.L., and N.T. analyzed data. L.P.-T. and R.D.S. provided technical leadership and oversight. J.P., P.D.P., and S.K. contributed to writing the manuscript with input from all authors.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Garcia, B.A. What does the future hold for top down mass spectrometry? *J. Am. Soc. Mass Spectrom.* **21**, 193–202 (2010).
- Siuti, N. & Kelleher, N.L. Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **4**, 817–821 (2007).
- Smith, L.M. & Kelleher, N.L. Proteoform: a single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).
- Zhang, Z., Wu, S., Stenoien, D.L. & Paša-Tolić, L. High-throughput proteomics. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **7**, 427–454 (2014).

5. Tran, J.C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).
6. Chait, B.T. Chemistry. Mass spectrometry: bottom-up or top-down? *Science* **314**, 65–66 (2006).
7. Lanucara, F. & Eyers, C.E. Top-down mass spectrometry for the analysis of combinatorial post-translational modifications. *Mass Spectrom. Rev.* **32**, 27–42 (2013).
8. Horn, D.M., Zubarev, R.A. & McLafferty, F.W. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **11**, 320–332 (2000).
9. Zabrouskov, V., Senko, M.W., Du, Y., Leduc, R.D. & Kelleher, N.L. New and automated MSn approaches for top-down identification of modified proteins. *J. Am. Soc. Mass Spectrom.* **16**, 2027–2038 (2005).
10. Liu, X. *et al.* Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol. Cell. Proteomics* **9**, 2772–2782 (2010).
11. Kou, Q., Wu, S. & Liu, X. A new scoring function for top-down spectral deconvolution. *BMC Genomics* **15**, 1140 (2014).
12. LeDuc, R.D. *et al.* ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res.* **32**, W340–W345 (2004).
13. Zamdborg, L. *et al.* ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **35**, W701–W706 (2007).
14. Liu, X. *et al.* Protein identification using top-down. *Mol. Cell. Proteomics* **11**, M111.008524 (2012).
15. Kou, Q., Xun, L. & Liu, X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* **32**, 3495–3497 (2016).
16. Sun, R.-X. *et al.* pTop 1.0: a high-accuracy and high-efficiency search engine for intact protein identification. *Anal. Chem.* **88**, 3082–3090 (2016).
17. Cai, W. *et al.* MASH Suite Pro: a comprehensive software tool for top-down proteomics. *Mol. Cell. Proteomics* **15**, 703–714 (2016).
18. Guner, H. *et al.* MASH Suite: a user-friendly and versatile software interface for high-resolution mass spectrometry data interpretation and visualization. *J. Am. Soc. Mass Spectrom.* **25**, 464–470 (2014).
19. Kim, S., Gupta, N. & Pevzner, P.A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **7**, 3354–3363 (2008).
20. Kim, S. & Pevzner, P.A.M.S.-G.F. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
21. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
22. Pevzner, P.A., Dancík, V. & Tang, C.L. Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.* **7**, 777–787 (2000).
23. Liu, X. *et al.* Identification of ultramodified proteins using top-down tandem mass spectra. *J. Proteome Res.* **12**, 5830–5838 (2013).
24. Frank, A.M., Pesavento, J.J., Mizzen, C.A., Kelleher, N.L. & Pevzner, P.A. Interpreting top-down mass spectra using spectral alignment. *Anal. Chem.* **80**, 2499–2505 (2008).
25. Tsur, D., Tanner, S., Zandi, E., Bafna, V. & Pevzner, P.A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562–1567 (2005).
26. Frank, A., Tanner, S., Bafna, V. & Pevzner, P. Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* **4**, 1287–1295 (2005).
27. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **28**, 710–721 (2010).
28. Nagaraj, N. & Mann, M. Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. *J. Proteome Res.* **10**, 637–645 (2011).
29. Zhu, W., Smith, J.W. & Huang, C.-M. Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. Biotechnol.* **2010**, 840518 (2010).
30. Qian, W.-J., Jacobs, J.M., Liu, T., Camp, D.G. II & Smith, R.D. Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Mol. Cell. Proteomics* **5**, 1727–1744 (2006).
31. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **4**, 1116–1130 (2013).
32. Tabb, D.L. *et al.* Reproducibility of differential proteomic technologies in CPTAC fractionated xenografts. *J. Proteome Res.* **15**, 691–706 (2016).
33. Ntai, I. *et al.* Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Mol. Cell. Proteomics* **15**, 45–56 (2016).

## ONLINE METHODS

### Intact protein extraction and preparation for LC-MS/MS analysis.

**Ovarian tumor sample.** Ovarian tissue used in this manuscript was from a pool of five female patients; the samples were collected with the oversight of the Institutional Review Board at Oregon Health and Science University and the patients gave informed consent. An approximately 20 mg portion of fresh frozen tissue was taken from each patient to create a pooled sample. Tissue aliquots were homogenized with a pellet pestle in 1 mL of homogenization buffer (8 M urea, 50 mM ammonium bicarbonate, 1 mM PMSF, and 1% Sigma phosphatase inhibitor cocktail II and III). The resulting homogenate was then incubated at 37 °C for 30 min to facilitate protein extraction, and it was spun for 10 min to pellet insoluble debris. All centrifugation steps were carried out at 4 °C, to further limit potential enzymatic activity, and 15,000 r.p.m. The supernatant was then transferred to an Amicon ultra 100K MWCO filter (EMD Millipore) prerinsed with 500  $\mu$ L homogenization buffer. Samples were then centrifuged for 30 min to obtain minimum volume. A 450  $\mu$ L aliquot of homogenization buffer was added to the filter and spun for an additional 30 min to maximize protein recovery. The filtrate was then transferred to an Amicon ultra 10K MWCO filter and spun for 40 min to obtain minimum volume. Buffer exchange was achieved using three washes with 450  $\mu$ L of buffer A (3% acetonitrile, 0.2% formic acid in MilliQ water). The protein concentration was determined using a Coomassie assay (Thermo Fisher). The final protein concentration was adjusted to 0.5  $\mu$ g/ $\mu$ L for analysis. Ten replicate LC-MS/MS data sets were acquired.

**Breast tumor xenograft sample.** We created top-down LC-MS/MS data sets for two subtypes of breast cancer tumors, basal like (WHIM2-P32) and luminal B (WHIM16-P33). The source and provenance of the xenograft material is described in ref 32. For each subtype, six process replicates LC-MS/MS runs were created. Tumors were cryopulverized, distributed into six aliquots for each tumor, and stored  $-80$  °C until use. The six aliquots were processed independently for each tumor as described above. All samples were block-order randomized. Each sample was then analyzed by a single 180 min LC-MS/MS run. The first injection from each tumor was used to passivate the new LC column, and these files were not included in later analyses.

**LC-MS/MS analysis.** A dual-pump Waters nanoACQUITY UPLC system (Milliford, Massachusetts) in combination with a Velos Orbitrap Elite mass spectrometer (Thermo Fisher, San Jose, California) was used for these analyses. A 5  $\mu$ L sample injection was loaded on a solid phase extraction (SPE) column for rapid trapping and desalting prior to separation. The analytical column was prepared in house by slurry packing 3  $\mu$ m diameter C2 stationary phase (Separation Methods Technology, Newark, Delaware) into a 50 cm length of 360  $\mu$ m o.d.  $\times$  100  $\mu$ m i.d. fused silica capillary column (Polymicro Technologies Inc., Phoenix, Arizona). The SPE column (360  $\mu$ m outer diameter  $\times$  150  $\mu$ m internal diameter) of 5 cm length was similarly prepared. Mobile phases consisted of 0.2% formic acid in water (phase A) and 0.2% formic acid in acetonitrile (phase B). Sample was loaded for 30 min on the SPE column and then separated by the analytical column using a 190 min gradient from 99% A to 35% A in 180 min at a flow rate of 0.3  $\mu$ L/min. The LC column was interfaced with the mass spectrometer using a home-made nano-electrospray

ionization source with a chemically etched 150  $\mu$ m o.d.  $\times$  20  $\mu$ m i.d. fused silica emitter. A spray voltage of 2.3 kV and an ion transfer tube temperature of 325 °C were used for ionization and desolvation. Precursor spectra were acquired from  $m/z$  500 to 2,000 at a resolution of 240,000. Data-dependent product spectra of the top four ions were isolated in a 4 Da window and subjected to CID and HCD fragmentation modes at normalized collision energies of 35% and 30%, respectively. All product ions were detected in the Orbitrap at a resolution of 120,000.

**Data format for LC-MS/MS spectrometry.** For fast data access, MSPathFinder and ProMex use an internal file format for LC-MS data called a pbf file. This file stores LC-MS data as a collection of 3D data points, called peaks, defined as: scan number,  $m/z$ , and intensity. To support quick retrieval of both spectra and chromatograms, the PBF format indexes peaks in two ways: (i) a spectrum-centric way—get all the peaks for a certain scan number, and (ii) chromatogram-centric way—get all the peaks within a specified  $m/z$  range.

**ProMex, LC-MS feature extraction.** ProMex was developed to detect isotopomer envelopes of intact protein ions and determine their monoisotopic masses and abundances. The ProMex algorithm takes a range of monoisotopic mass and a mass tolerance as an input; it then outputs a collection of LC-MS features, each of which is specified by monoisotopic mass, charge states, elution time span, and abundance. The basic idea is that an individual isotopomer envelope of one charge state in one spectrum has poor ions statistics, especially as molecular weight increases (see **Supplementary Fig. 2**). Therefore, we evaluate an isotopic profile grouped across time and charge. The set of peaks attributed to a single proteoform species (across time and charge state) is referred to as an LC-MS feature. The process of determining which peaks belong to the same LC-MS feature is shown graphically in **Figure 1** and described in the subsequent paragraphs.

To identify which peaks in LC-MS data should be grouped, a list of potential masses is created using the user-specified mass range and tolerance. For each potential monoisotopic mass  $M$  (Da), a theoretical isotopomer envelope  $E_M$  is generated from Averagine model<sup>34</sup>; then, using the input charge range, various  $m/z$  values are calculated for  $E_M$ . ProMex then scans all MS1 spectra to identify peaks corresponding to these isotopomer envelopes. The collected peaks are grouped by their charge states and elution times, and thus an observed isotopomer envelope at charge state  $c_i$  and elution time  $t_j$  is denoted as  $E_{ij}$ .

The second step is to cluster isotopomer envelopes indicating the same proteoform species across charge states and LC elution time. ProMex gathers envelope peaks in adjacent charges and elution times using a greedy algorithm. It starts with seed isotopomer envelopes selected based on their similarity scores against  $E_M$  and statistical significances. Here the similarity score  $S(E_1, E_2)$  between two envelopes  $E_1$  and  $E_2$  is computed by the Pearson correlation. The statistical significance is determined by Wilcoxon rank-sum test and hypergeometric test as previously described<sup>35</sup>. Both tests are performed within a local range (5  $m/z$ ) of the spectrum encompassing the seed envelope. Seed envelopes must have  $P$  values that are less than 0.01 for both tests and similarity scores larger than specified thresholds. The increasing number of charge states and the increasing size of isotopomer envelope lower the

chance of observing isotopomer envelopes highly similar to the theoretical one. Thus, we use different thresholds depending on the mass  $M$ —0.7 for  $M < 10,000$  Da; 0.6 for  $10,000 < M < 15,000$ ; 0.5 for  $15,000 < M < 30,000$ ; 0.4 for  $30,000 < M$ .

The clustering process starts with the seed envelope having the highest similarity score among the seed set. The greedy algorithm iteratively explores observed envelopes in adjacent charge states and elution times. Adjacent envelopes are added to the cluster if they enhance the similarity between the aggregated isotopomer envelope in the cluster against  $E_M$ . When there is no adjacent isotopomer envelope improving the cluster, the algorithm stops exploration. This process continues until all seed envelopes are assigned to clusters.

Detected clusters are refined to accurately determine their elution time spans and ranges of charge states. The elution time span is determined based on elution profile (EP). The EP is constructed by peaks in the clustered isotopomer envelopes and smoothed by Savitzky–Golay filter using nine adjacent points with quadratic polynomial. The first and last elution times having intensities equal or greater than 1% of the apex intensity are set to elution start ( $t_{\min}$ ) and end time ( $t_{\max}$ ), respectively. To determine the range of charge states, at each possible charge state  $c_i$ , the algorithm examines not only individual isotope envelopes  $E_{i,j}$  in the elution time span ( $t_{\min} \leq j \leq t_{\max}$ ) but also aggregated isotopic envelop  $E_i$  over the span. If either any single  $E_{i,j}$  or  $E_i$  has similarity score higher than 0.7, the charge state is included into the cluster. Thus, the minimum ( $c_{\min}$ ) and maximum ( $c_{\max}$ ) charge states define the range of charge states. The final monoisotopic mass of LC-MS feature is determined by selecting the median value from all the clustered isotopomer envelopes.

The abundance of LC-MS features is measured by the area under EP. In order to avoid outlier peaks caused by signal interference or noise, it only includes peaks in isotopomer envelope  $E_{i,j}$  where

$$S(E_{ij}, E_M) > \min(0.7, \text{median}(\{S(E_{ij}, E_M) \mid c_{\min} \leq i \leq c_{\max}, t_{\min} \leq j \leq t_{\max}\}))$$

The area under smoothed EP is calculated and set to the abundance of LC-MS feature.

The quality of each feature is evaluated by a likelihood ratio scoring function. Features that fail the likelihood test are rejected and deleted. We devised a Bayesian network that models LC-MS features to determine the probability of observing aggregated isotopomer envelopes  $E_i$  given mass  $M$  (**Supplementary Fig. 8**). A series of isotopomer envelopes detected in the elution time span at a charge are described by four parameters— $A_i$ ,  $S_i$ ,  $I_i$ , and  $X_i$ . Here  $A_i$  is the ratio of abundance at charge  $c_i$  to total abundance, and  $S_i$  is the similarity score  $S(E_i, E_M)$  of aggregated isotopomer envelope  $E_i$ . At each spectrum, the intensity of isotopic peaks is scaled by dividing them the highest intensity in a window of width  $5 m/z$  around the isotope envelope.  $I_i$  is the sum of scaled intensities of the most abundant isotopic peaks within the elution time span.  $X_i$  is elution profile score, which is the average Pearson's correlation coefficient of EP at charge  $c_i$  against EPs at other charge states. Thus, the likelihood scoring function can be represented as

$$\text{Likelihoodscore} = \sum_{c_i=c_{\min}}^{c_{\max}} \log \frac{p^{\text{obs}}(C_i, A_i, S_i, I_i, X_i \mid M)}{p^{\text{null}}(C_i, A_i, S_i, I_i, X_i \mid M)}$$

where  $P^{\text{obs}}$  is the probability of a particular state ( $C_i, A_i, S_i, I_i, X_i \mid M$ ) observed in a sample of known LC-MS features, and  $P^{\text{null}}$  is the probability of the same state in a null hypothesis model where peaks are randomly shuffled over 3D LC-MS space.

Considering conditional dependencies of parameters as defined in **Supplementary Figure 8** and applying Bayes' theorem, the likelihood scoring function can be rewritten as

ProMex's likelihood scoring function

$$= \sum_{c_i=c_{\min}}^{c_{\max}} \left\{ \log \frac{P^{\text{obs}}(A_i \mid C_i, M)}{P^{\text{null}}(A_i \mid C_i, M)} + \log \frac{P^{\text{obs}}(S_i \mid C_i, M)}{P^{\text{null}}(S_i \mid C_i, M)} + \log \frac{P^{\text{obs}}(I_i \mid C_i, M)}{P^{\text{null}}(I_i \mid C_i, M)} + \log \frac{P^{\text{obs}}(X_i \mid C_i, M)}{P^{\text{null}}(X_i \mid C_i, M)} + \log \frac{P^{\text{obs}}(C_i \mid M)}{P^{\text{null}}(C_i \mid M)} \right\}$$

The detected LC-MS features often overlap and share peaks with each other, because a cluster of observed peaks can be well matched to different theoretical isotopomer envelopes (see **Supplementary Fig. 9**). To eliminate redundant, false LC-MS features, ProMex selects only the best scoring features iteratively and removes them from the LC-MS data. For this, ProMex constructs an undirected acyclic graph where each vertex represents an LC-MS feature. Two vertices are connected by an edge if they share peaks in their collected isotopomer envelopes. Vertices are grouped such that two vertices in a group are connected to each other by paths. In each group, the best scoring LC-MS feature is selected, and peaks associated with the feature are removed from the LC-MS data. If there are LC-MS features with a difference of  $\pm 1$  Da from the best scoring feature, they are selected together to maximize the chance of identifying correct proteoforms. Whenever peaks are removed from the LC-MS data, other remaining features are rescored. This process is repeated until the best score in the group is less than a certain likelihood score cutoff.

**MSPathFinder, proteoform identification.** MSPathFinder takes an LC-MS feature file generated by ProMex, a protein FASTA database, and a set of search modifications as an input and outputs PrSMs with E-values. A search modification is defined as a pair consisting of a PTM and a target amino acid. The maximum number of allowable modifications is also given as input. For each sequence present in the protein database, MSPathFinder constructs a sequence graph (described below) and scores proteoforms against MS/MS spectra through graph searching. The statistical significance of individual PrSMs (e.g., E-values) is also evaluated. Lastly, MSPathFinder estimates the false discovery rate.

**Enumerating protein substrings.** MSPathFinder supports three search modes, depending on the number of internal cleavages allowed. Search mode 2, similar to the nontolerable termini (NTT) 2 in bottom-up proteomics, does not allow any internal cleavage except the single amino acid cleavage at the N-terminus. Search mode 1, similar to NTT 1, additionally allows single internal cleavage; and search mode 0, similar to NTT 0, allows multiple internal cleavages. The numbers of sequences to be searched are different depending on the search mode. Also, the lowest and highest masses of detected LC-MS features also provide lower and upper bounds in sequence lengths.

**MS/MS spectra deconvolution.** MSPathFinder uses a fitting method similar to THRASH algorithm<sup>8</sup> to deconvolute MS/MS spectra. The deconvolution algorithm moves a window of a certain  $m/z$  width along the peaks (here, 2.2  $m/z$  was used). The most intense peak in the window is selected, and a few average masses are calculated using the observed  $m/z$  for a range of charge states (here, charge states of 1–20+ were used). For each average mass, a theoretical isotopomer envelope is generated from Averagine model<sup>34</sup>. Then, it identifies observed isotope peaks corresponding to the theoretical isotope envelope. Pearson's correlation coefficient between observed and theoretical isotopomer envelopes is computed. If the correlation coefficient is higher than a certain threshold (here, 0.7 was used), the observed isotopomer envelope is converted to a deconvoluted peak defined by monoisotopic mass, charge state, and intensity.

**Sequence graph.** The sequence graph is a directed acyclic graph (DAG) that represents all possible PTM-modified forms of a protein sequence (see Fig. 2). Each vertex of the sequence graph represents a unique fragment and corresponds to an elemental composition. Two vertices are connected by an edge if their difference in compositions equals to a composition of an amino acid and optionally a PTM. For example,  $C_6H_{12}N_2O_2S_0$  and  $C_{12}H_{24}N_6O_3S_0$  are connected by an edge representing Arg ( $C_6H_{12}N_4O_1S_0$ ). In the sequence graph, the leftmost vertex (called the source) represents the 'zero' fragment with a composition  $C_0H_0N_0O_0S_0$ , and each of the rightmost vertices (called precursor vertices) represents proteoforms with the same modifications. Each path in the graph corresponds to a proteoform.

Once the sequence graph is constructed, MSPathFinder selects a precursor vertex  $v_p$  one by one, and repeats the following procedure. It considers a subgraph from the source to  $v_p$ . This subgraph represents the proteoforms with the same composition, and each internal vertex represents a fragment of these proteoforms. For the precursor mass of  $v_p$ , MS/MS spectra are retrieved from the LC-MS look-up table. For each recruited MS/MS spectrum and each internal node  $v$ , MSPathFinder finds evidence of the ions generated by a fragment with a composition  $v$  and assigns a score to  $v$ . Edge scores are also assigned as necessary (e.g., consecutive fragment ion score). The score of a path is defined as the sum of scores of vertices and edges in the corresponding path. MSPathFinder finds the best scoring proteoform by backtracking the sequence graph.

**MSPathFinder scoring.** We designed a scoring function,  $MSPathScore(P, S)$  to evaluate a PrSM of a proteoform  $P$  and a spectrum  $S$ . The  $MSPathScore$  utilizes five characteristics of matched fragment ion peaks—intensity, isotopomer envelope shape, mass measurement error, existence of complementary fragment ion peak, and existence of consecutive fragment ion peaks, which can be written as:

$$\begin{aligned} MSPathScore(S, P) &= \sum_{i \in \alpha} \left[ W_{match}^P + W_{intensity}^P I_i + W_{dist}^P D_i + W_{error}^P E_i \right] \\ &+ \sum_{i \in \hat{\alpha}} \left[ W_{match}^S + W_{intensity}^S I_i + W_{dist}^S D_i + W_{error}^S E_i \right] \\ &+ \sum_{\substack{(i, j) \in (\alpha \cup \hat{\alpha}) \\ i \neq j}} W_{comp1} ISComplement(i, j) \\ &+ W_{consecutive} ISConsecutive(i, j) \end{aligned}$$

where  $\alpha'$  and  $\hat{\alpha}$  are sets of prefix and suffix fragment ion peak matches, respectively.  $I_i$ ,  $D_i$ , and  $E_i$  are normalized intensity, isotopomer envelope similarity score, and mass error of matched fragment ion  $i$ . The normalized intensity is calculated by dividing the peak's intensity by the highest intensity in the spectrum. The envelope similarity is determined by Pearson's correlation coefficient between observed and theoretical isotopomer envelopes. The mass error is measured in p.p.m.  $ISComplement(i, j)$  is an indicator function to denote whether the fragment ion pair  $(i, j)$  is a complementary fragment ion pair.  $ISConsecutive(i, j)$  is also an indicator function describing whether the ions  $(i, j)$  are consecutive fragment ions. The weight parameters in  $MSPathScore$  were determined by a logistic regression method with a training set of 30,000 PrSMs. The training set was obtained by scoring PrSMs as the number of matched fragment ion peaks.

**Sequence-tag-based search.** MSPathFinder uses sequence tags to filter the search space of multiply cleaved protein sequences. Sequence tags are short amino acid sequences that are found by combining consecutive fragment ions in protein sequences<sup>26</sup>. MSPathFinder generates all possible sequence tags with a minimum length and finds multiply cleaved protein sequences containing the sequence tags. Here a minimum length of five residues is chosen, as it gives a good balance between the number of identifications and the size of search space. Given a protein sequence matched to a sequence tag, two sequence graphs originating from the both ends of sequence tags are generated toward opposite directions as shown in Supplementary Figure 3. The flanking mass of sequence tags and the mass of LC-MS features are used to constrain candidate proteoforms to be searched.

**Statistical significance of protein–spectrum Match.** The statistical significance such as a  $P$  value or  $E$ -value is estimated by the generating function approach as previously described<sup>19,20</sup>. The implementation of the generating function for bottom-up proteomics is not directly applicable to top-down proteomics, because the underlying dynamic programming table grows to be exceptionally large with the large masses of intact proteins and the increased mass accuracy and resolution of typical top-down data. In order to minimize the number of integer masses to be considered in the generating function while accommodating the high mass accuracy, we discretize the real mass space with a window of a constant mass tolerance (e.g., 8 p.p.m.). In addition, masses within mass regions that cannot be reached by combinations of amino acid with allowable PTMs masses are removed.

**Estimating false discovery rates.** The false discovery rate (FDR) is estimated using the target–decoy approach<sup>21</sup>, where a decoy database was constructed by reversing the protein sequences and applying three amino acid mutations at random positions.

**LcMsSpectator, visualization and refinement tool.** LcMsSpectator is a Windows desktop application that facilitates visualization and refinement of top-down proteomics analysis results reported by ProMex and MSPathFinder and features interactive spectral and chromatographic data plots as shown Supplementary Figure 4. All of the views and data plots can be exported to high-resolution, publication-ready images. LcMsSpectator supports the community standards for both spectrum data (mzML<sup>36</sup>) and spectrum

annotation (mzIdentML<sup>37</sup>). See <https://github.com/PNNL-Comp-Mass-Spec/LCMS-Spectator> for details.

**Software evaluation.** All the experiments were performed on a Windows computer with a 3.5 GHz CPU (Intel Xeon E3-1270) and 32 GB memory.

**Comparison of LC-MS feature-detection algorithms.** We ran ProMex, MS-Deconv+, and ICR-2LS against a total of ten replicate runs. MS-Deconv+ and ICR-2LS reported monoisotopic masses, charges, elution times, and intensities for each MS1 spectrum. We clustered these deconvolution results into LC-MS features if two monoisotopic masses were within the mass tolerance and elution time window. The abundance was computed by summing intensities of cluster members. We repeated this clustering procedure to group LC-MS features detected from multiple replicate runs. Here we used 10 p.p.m. mass tolerance and 1 min elution time window. Parameter settings used in ProMex, MS-Deconv+, and ICR-2LS are described in **Supplementary Table 2**.

**Comparison of proteoform identification algorithms.** We benchmarked MSPathFinder against TopPIC v0.9.1 (available at <http://proteomics.informatics.iupui.edu/software/toppic/>)<sup>14</sup>, ProSightPC v3.0 (ref. 13), and pTop v1.2 (available at <http://pfind.net/software/pTop/index.html>)<sup>16</sup>. For MSPathFinder, pTop, and MS-Align+, we used a fasta file of human and mouse from UniProt database for target database (2011\_12 version). The same decoy database was used for three tools, and PrSMs were collected at 1% FDR. For MS-Align+, the raw spectra files were converted to .msalign file format using msconvert tool in ProteoWizard software package<sup>38</sup> and MS-Deconv<sup>10</sup> as specified in TopPIC manual. Since pTop v1.2 is not able to search cleaved proteoforms, we compared pTop and MSPathFinder separately, and we disabled internal cleavages in MSPathFinder. For ProSightPC, we used the annotated mouse and human proteoform databases (2014\_07 version). We tested ProSightPC based on the absolute mass search mode with mass tolerance of 10 p.p.m. for precursor and fragment ions. We tried to use shuffled database search option to calculate FDR at ProSightPC. But there were very few PrSMs for shuffled sequences even though E-value cutoff increased to 10. Also, as the downloaded databases are encoded in binary format, we were not able to apply our target–decoy approach. Thus PrSMs were collected with E-value cutoff of  $1 \times 10^{-4}$  (default cutoff for good matches in ProSightPC). Parameter settings used in MSPathFinder, MS-Align+, ProSightPC, and pTop are described in **Supplementary Table 3**.

**Statistical analysis for label-free quantification.** LC-MS features are not always detected in all replicate runs. For missing LC-MS features, we integrated background signal intensities in the appropriate elution time spectra. Here we assumed that median

intensity in each spectrum is equal to background–signal intensity. Grouped LC-MS features were associated with proteoform identification results by MSPathFinder. Each group of LC-MS features is assumed to be a single proteoform species. If there are multiple different proteoforms matched to a LC-MS feature group, the best scoring proteoform (i.e., the lowest E-value PrSM) was selected as a representative proteoform of the LC-MS feature group. Informed-Proteomics does not provide a separate tool for this postprocessing step, but implemented source codes for LC-MS feature grouping were included in the package.

As a preprocessing step for normalization, we normalized the abundance values by equalizing the median of abundances across replicate runs. Conventional ANOVA analysis was performed and *P* values were adjusted by Benjamini–Hochberg (BH) procedure. All statistical analyses including ANOVA analysis and principal component analysis (PCA) were performed within MATLAB 2014b (The MathWorks, Inc., Natick, Massachusetts).

**Software availability.** MSPathFinder's home page is <https://omics.pnl.gov/software/mspathfinder>. All source codes were written in Microsoft C# with .NET framework 4.5 and are available at GitHub, <https://github.com/PNNL-Comp-Mass-Spec/Informed-Proteomics> and <https://github.com/PNNL-Comp-Mass-Spec/LCMS-Spectator> and as **Supplementary Software**. Each repository has a readme and wiki to describe installation and usage. Binary executables and installers are available at: <https://github.com/PNNL-Comp-Mass-Spec/Informed-Proteomics/releases> and <https://github.com/PNNL-Comp-Mass-Spec/LCMS-Spectator/releases>. A tutorial is available at <https://github.com/PNNL-Comp-Mass-Spec/Informed-Proteomics/wiki/MSPathFinder-Tutorial> and as a **Supplementary Protocol**.

A **Life Sciences Reporting Summary** for this paper is available.

**Data availability statement.** All data sets are available in the MassIVE proteomics repository under identifier: [MSV000080257](https://massive.ucsd.edu/MSV000080257). Source data files are available for **Figures 3–6**.

- Senko, M.W., Beu, S.C. & McLaffertycor, F.W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **6**, 229–233 (1995).
- Wang, X. *et al.* JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. *Mol. Cell. Proteomics* **13**, 3663–3673 (2014).
- Martens, L. *et al.* mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, R110.000133 (2011).
- Jones, A.R. *et al.* The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **11**, M111.014381 (2012).
- Chambers, M.C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

## Author Correction: Informed-Proteomics: open-source software package for top-down proteomics

Jungkap Park, Paul D Piehowski , Christopher Wilkins, Mowei Zhou , Joshua Mendoza, Grant M Fujimoto, Bryson C Gibbons, Jared B Shaw, Yufeng Shen, Anil K Shukla, Ronald J Moore, Tao Liu, Vladislav A Petyuk , Nikola Tolić, Ljiljana Paša-Tolić, Richard D Smith , Samuel H Payne  and Sangtae Kim 

Correction to: *Nature Methods* **14**, 909–914, <https://doi.org/10.1038/nmeth.4388> (2017), published online 7 August 2017.

In the version of this article initially published, the authors erroneously reported the search mode that was used for ProSightPC 3.0 in the Online Methods and in Supplementary Table 3.

The results presented in Fig. 5 were obtained with ‘absolute mass’ search mode, not ‘biomarker discovery’ search mode. The ‘biomarker discovery’ search mode of ProSightPC 3.0 looks for subsequences of those contained in the annotated proteoform database (e.g., truncated forms from degradation and/or cleavage). This search mode is expected to generate similar numbers of identifications as Informed-Proteomics, but is also expected to take dramatically longer (~480 CPU hours). Unfortunately, because of these heavy computational requirements, the authors were unable to complete an analysis using this search mode. They chose to use ‘absolute mass’ mode to illustrate the effect of search mode and database choice on the results. ‘Absolute mass’ mode is the most restrictive of the search modes illustrated in Fig. 5, as it searches only for proteoforms explicitly listed in the proteoform database within a user-defined mass tolerance.

In addition, in the supplementary information originally published online, Supplementary Table 3 incorrectly stated that ProSightPC v3.0 was used in ‘biomarker discovery’ mode. ‘Absolute mass’ mode was the mode actually used in this comparison. These errors have been corrected in the HTML and PDF versions of this article and in the associated supplementary information.

Published online: 13 June 2018

<https://doi.org/10.1038/s41592-018-0040-0>