PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data

Ying S Ting1, Jarrett D Egertson1, James G Bollinger1, Brian C Searle1, Samuel H Payne2, William Stafford Noble1,3 & Michael J MacCoss1

Data-independent acquisition (DIA) is an emerging mass spectrometry (MS)-based technique for unbiased and reproducible measurement of protein mixtures. DIA tandem mass spectrometry spectra are often highly multiplexed, containing product ions from multiple cofragmenting precursors. Detecting peptides directly from DIA data is therefore challenging; most DIA data analyses require spectral libraries. Here we present PECAN (http://pecan.maccosslab.org), a library-free, peptide-centric tool that robustly and accurately detects peptides directly from DIA data. PECAN reports evidence of detection based on product ion scoring, which enables detection of low-abundance analytes with poor precursor ion signal. We demonstrate the chromatographic peak picking accuracy and peptide detection capability of PECAN, and we further validate its detection with data-dependent acquisition and targeted analyses. Lastly, we used PECAN to build a plasma proteome library from DIA data and to query known sequence variants.

Recent improvements in instrumentation scan speed, resolution, mass accuracy, and dynamic range have positioned DIA1 as a viable strategy for analyzing complex peptide mixtures. DIA systematically selects mixtures of precursor ions for tandem mass spectrometry (MS/MS) analysis in an unbiased fashion.2 This unbiased MS/MS sampling distinguishes DIA from data-dependent acquisition (DDA)—which samples from precursors detected by MS analysis—and selected reaction monitoring (SRM)—which targets a predetermined set of precursors. To achieve unbiased sampling while providing comprehensive measurement, most DIA methods use wide isolation windows that sacrifice precursor selectivity. The resulting MS/MS spectra are usually highly chimeric and difficult to interpret for peptide identification by conventional database searching tools designed to identify one peptide per spectrum. Despite these challenges, DIA remains attractive because its unbiased measurements comprise a permanent, remineable digital record of the sample content3.

Analysis strategies tailored to DIA data are necessary. Library-based approaches, such as OpenSWATH3 and MSPLIT-DIA4, facilitate DIA data analysis by making the most of the rich knowledge accumulated from previous studies. These approaches are sensitive, but they limit data interrogation to only analytes present in the library. Thus, tools designed to detect peptides from DIA data without libraries are necessary to exploit the discovery potential of DIA data.

Library-free tools can be broken into two categories: spectrum-centric and peptide-centric tools.5 Spectrum-centric library-free tools, such as DIA-Umpire6 and Group-DIA7, typically generate pseudospectra from DIA data by detecting covarying precursor-product ion groups or deconvolving the multiplexed spectra. The quality of each pseudospectrum is often dependent on the quality and interpretability of the precursor signal in MS analysis. These pseudospectra are sent to conventional database searching pipelines that are designed for DDA identification and where precursor signal is a key filtering criterion for candidate peptides. As a result, pseudospectra with poor precursor signal are less likely to yield confident identifications. Such precursor dependency in database searching hinders the detection of analytes with detectable product signal but unresolved or detectable precursor signal in DIA data.8 The lack of detectable precursor signal for some detectable analytes is a common phenomenon resulting from limitations of intrascan dynamic range. The dynamic range of analytes in a single MS analysis may exceed the dynamic range of the mass analyzer, while the dynamic range of the product ions in an MS/MS analysis does not. This is most prevalent when analyzing complex samples, especially with limited chromatographic separation.

Unlike spectrum-centric approaches, peptide-centric approaches query the data for the best supporting evidence of detection for each query peptide. An early example is FT-ARM9, which queries simple theoretical spectra of peptides against high-mass-accuracy DIA data using a dot-product scoring function. FT-ARM was novel and straightforward, but with much to

1Department of Genome Sciences, University of Washington, Seattle, Washington, USA. 2Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA. 3Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to M.J.M. (maccoss@uw.edu).

RECEIVED 29 JUNE 2016; ACCEPTED 20 JUNE 2017; PUBLISHED ONLINE 7 AUGUST 2017; DOI:10.1038/NMETH.4390

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.
improve, specifically in its sensitivity and false discovery rate (FDR). We believe that peptide-centric approaches offer inherent analytical advantages over traditional spectrum-centric approaches when analyzing DIA data. Accordingly, we here report a new peptide-centric tool called peptide-centric analysis (PECAN) that detects peptides directly from DIA data without prerequisite spectral or retention time libraries.

The inputs to PECAN include centroided DIA data, a list of query (target) peptides, and a background proteome database (typically a species protein sequence database). PECAN outputs auxiliary scores describing the assigned evidence of detection with an associated retention time for every target peptide and PECAN-generated decoy peptide (Fig. 1 see Online Methods). These scores are then used by Percolator to estimate FDR and report confident peptide and protein detections. PECAN offers three primary advances relative to existing approaches. First, PECAN scoring weights theoretical fragment ions based on specificity to the query peptide relative to the background proteome to boost the score contribution of selective fragment ions even when they are low intensity. Second, PECAN incorporates a background score subtraction to correct for the scoring bias caused by uneven distribution of peptides in retention time and precursor space, thus reducing random matches in the peptide-dense regions. Last, PECAN scoring is primarily based on fragment ions, with precursor information as an auxiliary feature but not a requisite for generating evidence of detection. PECAN thus takes advantages of the common case where MS/MS analysis is more selective and sensitive than MS analysis. Here, we present in detail the PECAN algorithm (see Online Methods), validation and performance assessment, and applications to library building and proteogenomics.

RESULTS

PECAN peak picking performance

For every query peptide, PECAN reports the best evidence of detection and associated retention time in the data in a process analogous to selecting the best chromatographic peak from the peptide’s extracted ion chromatograms (XICs). To evaluate the ‘peak picking’ performance of PECAN, we analyzed a DIA data set containing 422 synthetic stable-isotope-labeled standard (SIS) peptides spiked into various background proteomes with ten dilution steps. The data set was published with a manually curated reference specifying the boundaries of chromatographic peaks of 387 detectable SIS peptides in each dilution step.

The percentage of correct to total SIS peaks reported by PECAN was calculated by determining if PECAN-reported evidence of detection fell within the manually curated reference peak boundaries (correct peak) or not (incorrect). Without FDR control, the percentage of correct SIS peaks reported decreased as the SIS spiked-in concentration decreased and as the sample matrix complexity increased (Fig. 2a,b). With percolator FDR control (q-value < 0.01), the percentage of correct SIS peaks reported greatly increased, even at low SIS spiked-in concentration and high interference from the background proteome (Fig. 2c). While PECAN reports the best evidence of detection for every query peptide, not all reported evidence is correct, just as not every query peptide is detectable from the data. Using PECAN-reported decoy evidence, Percolator rejected most of the incorrect evidence of detection (Fig. 2d) and greatly improved PECAN’s peak picking performance.

PECAN detection validation

We analyzed 90-min deep gas-phase fractionation (4×GPF, see Online Methods) HeLa data sets using Comet and PECAN in conjunction with a GST-fusion-protein database. At 1% FDR (Percolator), we compared peptides detected from DIA by PECAN with those from DDA by Comet; this yielded 12,767 and 6,221 unique peptides, respectively, with an overlap of 5,182 peptides (Fig. 3a). 83% of Comet DDA peptides were detected by PECAN directly from the DIA data. Of the 5,182 common detections, only 27 had contradicting retention times between the two methods (Supplementary Fig. 1). Of the 1,039 peptides only identified in Comet DDA, 179 had precursor ions at charge state 4+ that were not considered in this PECAN analysis; 428 had PECAN-reported evidence that did not pass the FDR control; 96 had precursor m/z that fell between adjacent DIA isolation windows; and 336 had no qualifying evidence (see Online Methods). The PECAN DIA and Comet DDA approaches detected 2,613 and 1,759 protein groups, respectively, with an overlap of 1,510 proteins (Fig. 3b); this indicates that many of the distinct peptides from two approaches were derived from the same proteins.

To verify the PECAN-DIA-specific detections, we randomly selected 16 GST fusion proteins and expressed them using in vitro transcription translation (IVTT; Fig. 3c). We measured the corresponding 91 peptides using SRM from individually tryptic-digested GST-enriched proteins. Of the 91 peptides monitored by SRM, we manually assigned chromatographic peak boundaries for 86 peptides without ambiguity of detection, and we created a normalized retention time library referenced to the spiked-in stable-isotope-labeled peptides. The correlation coefficient between the measured retention time of the 73 peptides detected in PECAN DIA to the SRM library was 0.999 (Fig. 3d). With a threshold of <0.1% difference in total normalized retention time, all 73 peptides were correct, which suggests that the large majority of the PECAN-DIA-specific peptide detections were correct.

Impact of precursor selectivity on PECAN detection

Current DIA methods often use isolation windows five to ten times wider than those used in conventional DDA (typically <2-m/z wide) to sample a desired precursor range. Using wide isolation windows (i.e., low precursor selectivity) dramatically increases the complexity of the resulting MS/MS spectra because of cofragmenting analytes. To test how precursor selectivity impacts PECAN’s performance in detection, we used GPF to
PECAN-reported evidence of detection were subjected to peptide-level FDR control per measurement at q-value < 0.01 by Percolator.

Figure 2 | PECAN peak picking performance on the SIS data set. 422 stable-isotope-labeled standard (SIS) peptides were diluted in water (blue), yeast lysate (orange), or HeLa lysate (black) and measured in three replicates. The combined percentage of correct SIS peaks (a) and the total number of SIS peaks (b) reported by PECAN before FDR control. The combined percentage of correct SIS (c) and the total number of reported SIS peaks (d) after the PECAN-reported evidence of detection were subjected to peptide-level FDR control per measurement at q-value < 0.01 by Percolator.

vary DIA precursor selectivity while holding the cycle time and sampled precursor m/z range constant. GPF DIA data on HeLa cells were acquired with 20 (1×GPF)-, 10 (2×GPF)-, and 5 (4×GPF)-m/z-wide isolation windows and interrogated using the human UniProt Swiss-Prot database. From the 1×GPF, 2×GPF, and 4×GPF DIA data sets, PECAN detected 14,135; 23,398; and 34,813 unique peptides and 1,834; 5,191; and 9,132 protein groups, respectively (Fig. 4a,b), which indicates that better precursor selectivity (i.e., narrower isolation windows) dramatically improves PECAN’s performance. Additionally, the majority of peptide and protein detections from DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Of the 12,952 peptides detected in all three data sets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1×GPF or 2×GPF compared with the 4×GPF data set (Fig. 4c), and this indicates robust peptide detection.

As a benchmark, we processed the GPF DIA data sets with DIA-Umpire followed by Comet database searching. From the 1×GPF, 2×GPF, and 4×GPF DIA data sets, DIA-Umpire and Comet identified 13,978; 20,266; and 24,721 unique peptides and 1,834; 5,191; and 9,132 protein groups, respectively (Fig. 4a,b), which indicates that better precursor selectivity (i.e., narrower isolation windows) dramatically improves PECAN’s performance. Additionally, the majority of peptide and protein detections from DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Of the 12,952 peptides detected in all three data sets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1×GPF or 2×GPF compared with the 4×GPF data set (Fig. 4c), and this indicates robust peptide detection.

As a benchmark, we processed the GPF DIA data sets with DIA-Umpire followed by Comet database searching. From the 1×GPF, 2×GPF, and 4×GPF DIA data sets, DIA-Umpire and Comet identified 13,978; 20,266; and 24,721 unique peptides and 1,834; 5,191; and 9,132 protein groups, respectively (Fig. 4a,b), which indicates that better precursor selectivity (i.e., narrower isolation windows) dramatically improves PECAN’s performance. Additionally, the majority of peptide and protein detections from DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Of the 12,952 peptides detected in all three data sets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1×GPF or 2×GPF compared with the 4×GPF data set (Fig. 4c), and this indicates robust peptide detection.

As a benchmark, we processed the GPF DIA data sets with DIA-Umpire followed by Comet database searching. From the 1×GPF, 2×GPF, and 4×GPF DIA data sets, DIA-Umpire and Comet identified 13,978; 20,266; and 24,721 unique peptides and 1,834; 5,191; and 9,132 protein groups, respectively (Fig. 4a,b), which indicates that better precursor selectivity (i.e., narrower isolation windows) dramatically improves PECAN’s performance. Additionally, the majority of peptide and protein detections from DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Of the 12,952 peptides detected in all three data sets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1×GPF or 2×GPF compared with the 4×GPF data set (Fig. 4c), and this indicates robust peptide detection.

As a benchmark, we processed the GPF DIA data sets with DIA-Umpire followed by Comet database searching. From the 1×GPF, 2×GPF, and 4×GPF DIA data sets, DIA-Umpire and Comet identified 13,978; 20,266; and 24,721 unique peptides and 1,834; 5,191; and 9,132 protein groups, respectively (Fig. 4a,b), which indicates that better precursor selectivity (i.e., narrower isolation windows) dramatically improves PECAN’s performance. Additionally, the majority of peptide and protein detections from DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Of the 12,952 peptides detected in all three data sets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1×GPF or 2×GPF compared with the 4×GPF data set (Fig. 4c), and this indicates robust peptide detection.

As a benchmark, we processed the GPF DIA data sets with DIA-Umpire followed by Comet database searching. From the 1×GPF, 2×GPF, and 4×GPF DIA data sets, DIA-Umpire and Comet identified 13,978; 20,266; and 24,721 unique peptides and 1,834; 5,191; and 9,132 protein groups, respectively (Fig. 4a,b), which indicates that better precursor selectivity (i.e., narrower isolation windows) dramatically improves PECAN’s performance. Additionally, the majority of peptide and protein detections from DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Of the 12,952 peptides detected in all three data sets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1×GPF or 2×GPF compared with the 4×GPF data set (Fig. 4c), and this indicates robust peptide detection.

As a benchmark, we processed the GPF DIA data sets with DIA-Umpire followed by Comet database searching. From the 1×GPF, 2×GPF, and 4×GPF DIA data sets, DIA-Umpire and Comet identified 13,978; 20,266; and 24,721 unique peptides and 1,834; 5,191; and 9,132 protein groups, respectively (Fig. 4a,b), which indicates that better precursor selectivity (i.e., narrower isolation windows) dramatically improves PECAN’s performance. Additionally, the majority of peptide and protein detections from DIA data with lower precursor selectivity were successfully captured by data with higher precursor selectivity. Of the 12,952 peptides detected in all three data sets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1×GPF or 2×GPF compared with the 4×GPF data set (Fig. 4c), and this indicates robust peptide detection.
Querying sequence variants with PECAN

Large-scale genomics projects have greatly expanded the catalog of known sequence variants. PECAN can leverage this catalog by querying variant-containing peptides in the context of proteogenomics. Of the proteins detected by PECAN in the DIA plasma library, 342 are in the UniProt Swiss-Prot human natural variant database. These proteins collectively contain 4,264 single amino acid variants. Of the 4,264 variants, 3,714 result in at least one theoretical variant-specific tryptic peptide missing in the reference human UniProt Swiss-Prot database.

In some cases, PECAN detected multiple variant-specific peptides resulting from the same sequence variant. In serotonin (Fig. 5a), two variant-specific peptides were detected for the variant Ile448Val, while no canonical peptide spanning Ile448 was detected. In addition, three variant-specific peptides were detected for Pro589Ser, of which two resulted from the introduction of a new trypsin cleavage site by the variant. In some cases, PECAN detected multiple similar peptides from the same group of MS/MS spectra. For instance, in apolipoprotein A1 (Fig. 5b), the peptide spanning the Glu134Lys variant was detected with the same group of spectra as the canonical peptide spanning Glu134. This is a challenging case because these two peptides are so similar that they share most of their fragment ions and differ only by 0.94763 Da in intact masses. Even with the 2-m/z-wide isolation windows, the canonical and variant peptides were not resolved by precursor isolation, and the same group of spectra provided statistically significant evidence of detection for both peptides. Because both peptides were detected at q-value < 0.01, PECAN did not choose one detection over the other, even though both peptides were supported by the same MS/MS spectra. Among the three Glu to Lys variants in apolipoprotein A1, only Glu160Lys had a definitive peptide resulting from cleavage at the new tryptic site introduced by the variant. Of the 21 Glu to Lys variants in the plasma library data set (Supplementary Table 1), 8 were covered by at least one peptide generated from variant-specific trypsin cleavages.

DISCUSSION

We have demonstrated the ability of PECAN to detect peptides robustly and accurately from DIA data without using a library. Because the detection of peptides improves as the precursor isolation window decreases, PECAN can be used to build libraries directly from DIA data collected using narrow isolation windows for later application to wide isolation data. This approach can augment existing DDA-based libraries as is evidenced by our detection of hundreds of novel peptides from 12 LC-MS/MS runs; these peptides were either not detected or did not make it through the FDR or statistical cutoff required when validating peptide–spectrum matches from over 100 DDA experiments in plasma. These novel detections could arise from the ability to detect peptides with...
weak or undetectable MS1 signal from DIA data. Existing libraries may be extended even further by combining the DIA library approach with sample fractionation and/or depletion.

Because PECAN does not use a library, however, it may not be as sensitive as library-based tools for detecting some peptides. To further improve the sensitivity of PECAN, we expect that training the hyperparameters, $\alpha$ and $\beta$, with DIA data of various precursor selectivity will be effective. We also expect that incorporating a sequence-based retention time predictor to filter based on expected retention time will improve the sensitivity of PECAN detection.

As a peptide-centric, library-free tool, PECAN is well suited for proteogenomics studies. For decades, genetics and genomics have focused on studying sequence variation and its influence on phenotype. Modern large-scale exome and genome sequencing projects have done much to expand the catalog of known sequence variation. With PECAN, one can easily leverage this catalog of variation by directly querying for variant-specific peptides against DIA proteomics data.

However, it should be noted that mass spectrometry data itself may not be sufficient to conclusively demonstrate the presence of some sequence variants. For example, some variants, such as leucine to isoleucine, are identical in mass and indistinguishable by the method described here. Some variants, such as asparagine to aspartic acid, are difficult to differentiate from residue modification (e.g., deamination). Digestion with an alternative protease might be necessary to produce peptides that could differentiate sequence polymorphism from modification. Other variants, such as glutamic acid to lysine, shift the peptide mass so little that the canonical and variant peptide ions will likely be isolated and fragmented together, which would result in similar MS/MS spectra. In this case, depending on the variant position relative to the peptide N terminus, two peptide ions may share most of the y-ions. We have demonstrated detection of variant-specific peptides from DIA data with high precursor selectivity (i.e., 2-m/z-wide isolation windows). Because the canonical and variant peptides from SNPs often have very similar fragmentation patterns, high precursor selectivity may be necessary to adequately resolve variants with precursor isolation. In the case where these similar peptides are not resolved by precursor isolation, it is possible that the same group of spectra may provide statistically significant evidence to multiple similar peptides. This phenomenon is most likely
to happen with wide-isolation-window DIA data. Thus, extra caution is warranted when making PTM- or variant-specific detections from wide-isolation-window DIA data with PECAN or any other tool. Additional steps could include requiring robust precursor ion signal as a criterion for detection; incorporating additional scoring, similar to A-score\textsuperscript{13} for phosphorylation site localization; or further validating the detections with analytical standards as is common practice in targeted assay development. The incorporation of extended scoring to resolve site-specific modifications and variant peptides within the framework of PECAN warrants further study.

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.

ACKNOWLEDGMENTS

The authors thank L. Käll, A.I. Nesvizhskii, N. Bandeira, and J.K. Eng for insightful discussions. This work was supported by the National Institutes of Health Grants P30 AG013280, R21 CA192983, P41 GM103533, and U54 HG008097. S.H.P. was supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, and Early Career Research Program.

AUTHOR CONTRIBUTIONS

Y.S.T. and M.J.M. designed the experiments. Y.S.T. developed the algorithms with input from J.D.E., S.H.P., B.C.S., W.S.N., and M.J.M. Y.S.T. performed the analyses. Y.S.T. and J.G.B. acquired the data. Software was written by Y.S.T. with substantial input from J.D.E., S.H.P., W.S.N., and M.J.M.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Venable, J.D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J.R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Methods 1, 39–45 (2004).
2. Chapman, J.D., Goodlett, D.R. & Masselon, C.D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom. Rev. 33, 452–470 (2014).
3. Röst, H.L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. 32, 219–223 (2014).
4. Wang, J. et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nat. Methods 12, 1106–1108 (2015).
5. Ting, Y.S. et al. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. Mol. Cell. Proteomics 14, 2301–2307 (2015).
6. Tsou, C.-C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods 12, 258–264 (2015).
7. Li, Y. et al. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat. Methods 12, 1105–1106 (2015).
8. Panchaud, A. et al. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. Anal. Chem. 81, 6481–6488 (2009).
9. Weisbrod, C.R., Eng, J.K., Hoopmann, M.R., Baker, T. & Bruce, J.E. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. J. Proteome Res. 11, 1621–1632 (2012).
10. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods 4, 923–925 (2007).
11. Eng, J.K., Jahan, T.A. & Hoopmann, M.R. Comet: an open-source MS/MS sequence database search tool. Proteomics 13, 22–24 (2012).
12. Gillet, I.C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics 11, 0111.016717 (2012).
13. 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. 24, 1285–1292 (2006).
ONLINE METHODS

PECAN workflow. PECAN uses the open-source application programming interface pymzML\textsuperscript{14} and supports the HUPO Proteomics Standard Initiative standard file format mzML\textsuperscript{15}. PECAN search results can be imported into Skyline\textsuperscript{16}, an open-source platform for mass spectrometry data visualization, quantification, interactive analyses, and report generation.

The PECAN workflow comprises four steps: generate peptide vectors, subtract background scores, report evidence of detection, and estimate detection FDRs. Here, we first describe PECAN’s primary score function and then each of the four steps.

PECAN primary scoring. PECAN uses matrix multiplication to score each peptide relative to its fragment extracted ion chromatograms (XICs)\textsuperscript{17}. For a DIA data set where m MS/MS spectra are generated from the isolation window that contains the precursor ion of peptide \(p\), the fragment XICs of peptide \(p\) can be represented as

\[
\text{XIC}_p = \begin{bmatrix}
I_{b_1,t_1} & I_{b_1,t_2} & \ldots & I_{b_1,t_m} \\
\vdots & \vdots & \ddots & \vdots \\
I_{b_{n-1},t_1} & I_{b_{n-1},t_2} & \ldots & I_{b_{n-1},t_m} \\
I_{y_1,t_1} & I_{y_1,t_2} & \ldots & I_{y_1,t_m} \\
\vdots & \vdots & \ddots & \vdots \\
I_{y_{n-1},t_1} & I_{y_{n-1},t_2} & \ldots & I_{y_{n-1},t_m}
\end{bmatrix}
\]

where \(I_{i,j}\) is the extracted intensity of an expected fragment ion with \(m/z\) value \(x\) at retention time \(t\). The extracted intensity is the sum of the square root of the intensities of ions with \(m/z\) values within the extraction mass error tolerance (default \(\pm 10\) p.p.m.) of \(x\). Let the peptide vector (see definition below) corresponding to peptide \(p\) be \(V_p\). Then the peptide score matrix is calculated as

\[
S_p = V_p \cdot \text{XIC}_p = [s_{t_1}, s_{t_2}, \ldots, s_{t_m}]
\]

where each \(S_i\) is mathematically equivalent to the scalar projection of \(O_p\), the observed MS/MS spectrum at retention time \(t\), onto the peptide scoring vector \(V_p\). Because the scalar value \(S_i\) represents the magnitude of the spectrum at retention time \(t\) supporting a peptide with \(V_p\), the vector \(S_p\) represents the evidence of detection for peptide \(p\) over time.

Generate peptide vectors. For each query peptide, PECAN generates a normalized scoring vector called a peptide vector. A peptide vector is a unit vector that represents the theoretical fragmentation pattern of the peptide. For a peptide \(p\) with \(n\) amino acids, let \(p = [b_2, \ldots, b_{n-1}, y_1, \ldots, y_{n-1}]\), where \(b_i\) and \(y_i\) are the theoretical \(m/z\) values of the corresponding fragment ions at position \(i\). By default, PECAN considers only +1 fragment ions for precursor ions with less than or equal to +2 charges, and it includes +2 fragment ions for precursor ions with +3 charges and above. The peptide vector for peptide \(p\) is then

\[
V_p = \frac{[w_1, \ldots, w_{b_{n-1}}, w_1, \ldots, w_{y_{n-1}}]}{[w_1, \ldots, w_{b_{n-1}}, w_1, \ldots, w_{y_{n-1}}]} = [w_1', \ldots, w_{b_{n-1}}', w_1', \ldots, w_{y_{n-1}}']
\]

where \(w_x\) is the ‘raw weight’ of a fragment ion with \(m/z\) value \(x\), and \(w_x'\) is the weight normalized to the magnitude of the vector containing raw weights. The raw weight \(w_x\) is calculated as the multiplicative inverse of the frequency of observing fragment ions with \(m/z\) value \(x\) (plus or minus a given mass accuracy, such as 10 p.p.m.), generated by \textit{in silico} fragmentation of proteolytic (e.g., tryptic) peptides from the background proteome database.

The \(w_x\) is calculated in a window-by-window fashion. For each distinct isolation window in a DIA experiment, only proteolytic peptides with precursor ions falling in the \(m/z\) range of the isolation window and therefore could contribute to product ion interference for the query peptide are used to calculate the \(w_x\) for the window. As a result, fragment ions with high-frequency \(m/z\) values, such as 147.113 (\(y_1\)-Lysine) and 175.119 (\(y_1\)-Arginine) for trypsin digestion, are weighted less than those with low-frequency \(m/z\) values. While \(w_x\) represents the specificity of observing a fragment ion with \(m/z\) value \(x\) in an isolation window with a given species database, \(w_x'\) represents the relative specificity for such observation to the peptide \(p\).

Subtract background scores. In DIA, multiple precursor ions within an isolation window are fragmented together, and this results in highly multiplexed MS/MS spectra. Because these spectra typically contain so many fragment ions, the expected score for a typical peptide against such spectra is nonzero. To estimate how high a peptide score can be achieved by chance, PECAN calculates ‘background scores’ represented by the means of thousands of decoy peptides (Supplementary Note 1). In addition, within the same isolation windows, higher charged precursor ions are assigned more fragment ions and hence exhibit a different score distribution compared with lower charged precursor ions. To account for these differences, the background scores are calculated in a window-by-window and charge-by-charge fashion. Peptides with precursor ions in different isolation windows, or in the same window but of different charge states, have different calibrating backgrounds (Supplementary Fig. 3).

To calculate background scores, PECAN generates thousands of decoys by shuffling proteolytic peptides from the background proteome database and scores each decoy against the database. Let \(z\) be a charge state of interest. The background score \(B_{p,z}\) for isolation window \(y\) at charge state \(z\) is calculated as the average score of the thousands of decoys generated within window \(y\) with charge state \(z\). With the background scores, PECAN calibrates each peptide score by

\[
S'_p = S_p - B_{y,z}
\]

Here, the isolation window \(y\) and charge state \(z\) are selected by the precursor ions of query peptide \(p\). The calibrated score \(S'_p\) is then subjected to a simple moving average smoothing with a factor \(u\). One of the strengths of DIA is the systematic measurement of the product ions. Depending on the liquid chromatography separation and DIA cycle time, PECAN uses the smoothing to capture the continuous scoring patterns and smooth out the noise contributed from sources of stochastic variation such as spray instability. PECAN considers the average score at every time point as an ‘evidence of detection’ centered at this time point. The evidence of detection \(E\) for peptide \(p\) at center time \(t\) is

\[
E_p(t) = \frac{1}{u} \sum_{k=t-u/2}^{t+u/2} S'_p[k]
\]

The smoothing factor \(u\) is an estimate of the number of times a peptide is analyzed by MS/MS at its full width at half maximum.
PECAN employs a semisupervised support vector machine algorithm, to estimate FDR of the reported evidence of detection. PECAN generates one decoy peptide for every query (target) peptide by shuffling the target sequence (Supplementary Note 3). These decoys undergo the same scoring processes as the targets, including subtraction with the same background scores. For each reported evidence of detection, whether for a target or a decoy peptide, PECAN calculates auxiliary scores (Supplementary Table 2). These auxiliary scores are used by Percolator to train a classifier from the target–decoy paradigm to distinguish between correct and incorrect matches and then estimate FDRs. In this target–decoy paradigm, the set of targets contains a mixture of detectable and undetectable peptides, whereas decoys by design consist only of undetectable peptides. Thus, PECAN-reported evidence of detection for targets are a mixture of correct and incorrect, whereas all evidence for decoys are incorrect by design. We combined all PECAN-reported evidence of detection from different isolation windows of one experiment so that Percolator could use the auxiliary scores to separate correct from incorrect evidence. We refer to the PECAN-reported evidence of detection with q-value < 0.01 after Percolator as 'PECAN detection'.

To test if the auxiliary scores, single or combined, incorrectly differentiated targets from decoys when used by Percolator, we queried ~100,000 tryptic peptides from the Escherichia coli proteome against HeLa DIA data sets with various DIA isolation schemes. By design, no query peptides were supposed to be detected from the data sets, and thus the target P values should be uniformly distributed. We generated quantile–quantile (Q–Q) plots to compare the P values reported by Percolator with the normalized rank P values that represent the uniform distribution (Supplementary Note 3). The results showed that Percolator could not differentiate the targets from decoys in this test, and this indicated that the auxiliary scores from PECAN did not introduce undesired separation of targets from decoys. Furthermore, tests of the same data set with peptide vectors generated from either an E. coli or a human protein sequence database showed that different origins of peptide vectors did not introduce undesired separation of targeted from decoys.

Liquid chromatography. All chromatography was performed using a nanoACQUITY (Waters) system set to a flow rate of 250 nl/min during linear gradient. Buffer A was 2% ACN, 0.1% formic acid, and 97.9% water. Buffer B was 99.9% ACN and 0.1% formic acid.

Homemade 3-cm-long 100-μm inner diameter (I.D.) trapping columns were used before the homemade 75-μm I.D. resolving column, that is, either 15- or 30-cm-long for a 27.5-min or 90-min linear gradient from 2% to 32% buffer B, respectively. For the plasma library sample, a homemade 2-cm-long 150-μm I.D. trapping column was used before a self-packed 30-cm-long 75-μm I.D. PicoFrit resolving column (New Objective) for a 90-min linear gradient from 2% to 35% buffer B. Both trapping and resolving columns were packed with 3-μm ReproSil-Pur C18 AQ (Dr. Maisch GmbH). The gradient was followed by a wash at 80% buffer B and a column re-equilibration at 2% buffer B.

SRM validation of IVTT proteins. Full-length cDNA clones for the 16 selected proteins were obtained from the pANT7_cGST clone collection distributed by the Arizona State University Biodesign Institute plasmid repository. Each bacterial stock clone was grown independently overnight in 5 ml of Luria–Bertani broth with 100 μg ml⁻¹ ampicillin (LB-amp). Plasmid DNA was extracted using the manufacturer’s spin mini-prep protocol (QIAGEN). Proteins were then synthesized from plasmid DNA using the Pierce Human in vitro Protein Expression kit (Thermo) according to the manufacturer’s protocol with GFP control. We then enriched the GST fusion proteins using glutathione sepharose 4B beads (GE) with a published method. Finally, these enriched GST fusion proteins were released, alkylated, and digested for 2 h with trypsin individually.

Report evidence of detection. For every peptide, PECAN default reports the best scoring evidence of detection and its associated center time \( t \) from all evidence that pass empirical criteria of the evidence-qualifying procedure (Supplementary Fig. 4). The goal of these empirical criteria is to disqualify evidence whose scores are predominantly contributed by a small number of fragment ions, which suggests that the score could be resulting from interference of a few high-abundance ions rather than a collaboration of multiple fragment ions. To this end, two hyperparameters, \( \alpha \) and \( \beta \), are used to set the criteria. Let peptide \( p \) contain \( N \) components (i.e., number of theoretical fragment ions) in the peptide vector \( V_p \). For the candidate \( E_p(t) \), the evidence of detection for peptide \( p \) at time \( t \), the component score threshold is set as

\[
T_p(t) = \frac{1}{N^\alpha} \sum_{k=-u}^{u} S_p k
\]

The score contribution of a fragment ion component with \( m/z \) value \( x \) to the \( E_p(t) \) is

\[
ionS_p(x,t) = w \sum_{k=-\frac{u}{2}}^{\frac{u}{2}} I_{x,k}
\]

We call the fragment ion components that score no less than the threshold \( T_p(t) \) ‘contributing ions’. Let the number of contributing ions (NCI) of the evidence \( E_p(t) \) be the number of ion components with score contribution at time \( t \) no less than the threshold \( T_p(t) \). If the number of contributing ions of \( E_p(t) \) is larger than the threshold \( \sum_{k=-\frac{u}{2}}^{\frac{u}{2}} C_p = BN \), the evidence of detection \( E_p(t) \) is marked qualified and will be reported. If the candidate evidence of detection is disqualified, the next-highest scoring evidence will be considered. We used a Saccharomyces cerevisiae DIA data set with 1,224 known boundaries of chromatographic peaks to optimize \( \alpha \) and \( \beta \) for the evidence-qualifying procedure (Supplementary Note 2). The resulting values of \( \alpha = 1.8 \) and \( \beta = 1.4 \) were used throughout this study.

Estimating detection false discovery rate. PECAN employs Percolator, a semisupervised support vector machine algorithm, to estimate FDR of the reported evidence of detection. PECAN generates one decoy peptide for every query (target) peptide by shuffling the target sequence (Supplementary Note 3). These decoys undergo the same scoring processes as the targets, including subtraction with the same background scores. For each reported evidence of detection, whether for a target or a decoy peptide, PECAN calculates auxiliary scores (Supplementary Table 2). These auxiliary scores are used by Percolator to train a classifier from the target–decoy paradigm to distinguish between correct and incorrect matches and then estimate FDRs. (FWHM) on average. This factor is calculated by dividing the user input minimum peptide elution time (in seconds) to the averaged cycle time of the first 100 cycles. For example, with a 90 min linear gradient liquid chromatography on a 30 cm 3 μm C18 column, most peptides elute for 12–20 s at FWHM. If a DIA method has a cycle time of 2 s, then a peptide would be measured by MS/MS at least six times. In this case, PECAN would then use \( u = 6 \) for the moving average calculation.
91 peptides were selected for the 16 proteins based on a preliminary analysis of PECAN during its early development (Supplementary Note 4). Each protein digestion was injected separately and analyzed with a TSQ-Vantage triple-quadrupole instrument (Thermo) using a nanoACQUITY UPLC (Waters). A 3-µl aliquot of sample was loaded for a 27.5-min LC setting. Ions were isolated in both Q1 and Q3 using 0.7 FWHM resolution. Peptide fragmentation was performed at 1.5 mTorr in Q2 without peptide specific collision energies. Data were acquired using a scan width of 0.002 mass to charge ratio (m/z) and a dwell time of 10 ms.

**HeLa data sets.** HeLa protein digest (Thermo) spiked in with stable-isotope-labeled peptides (PTC, Thermo) was analyzed on a Q-Exactive HF mass spectrometer (Thermo). 1 µg of HeLa peptides and 40 fmol of PRTC were loaded in each injection and separated with a 90-min linear gradient LC. Three gas-phase fractionation (GPF)20 settings were used to cover the precursor m/z range of 500 to 900—one injection (1×GPF); two injections covering 500–700 and 700–900 m/z (2×GPF); and four injections covering 500–600, 600–700, 700–800, and 800–900 m/z (4×GPF). The isolation ranges of MS analysis (SIM scans) for all GPF settings correspond to the precursor range covered in each injection. For example, the third injection of 4×GPF contains MS analysis with scanning ranges of 700 to 800 m/z, and MS/MS analysis of selected (either by DDA or DIA) precursor ions within precursor ranges of 700 to 800 m/z. Thus, the costs of sample amount and instrument time are double of the costs in 1×GPF for 2×GPF, and quadruple for 4×GPF.

Both DDA and DIA data were acquired with three GPF settings. A standard, top-20 DDA method (MS analysis with 120,000 resolution and MS/MS analysis with 15,000 resolution) with 1.5-m/z-wide isolation windows was used in data collection of 1×GPF DDA, 2×GPF DDA, and 4×GPF DDA (Supplementary Note 5). A standard (one MS analysis with 60,000 resolution followed by 20 MS/MS with 30,000 resolution) DIA method with 20-, 10-, or 5-m/z-wide isolation windows was used to acquire 1×GPF DIA, 2×GPF DIA, or 4×GPF DIA, respectively. For FDR control, data from multiple injections were analyzed together as if they were from one instrument run.

**Plasma library data.** Nondepleted plasma samples from five identified donors and a normal female plasma standard (Lampire Biological Laboratories) were individually digested. Plasma samples were diluted 200-fold before digestion with a diluent containing heavy-labeled protein and peptide standards and PPS silent surfactant in 50 mM ammonium bicarbonate. Postdilution, the sample contained 1 ng/µL 15N-labeled human apolipoprotein A1 (Cambridge Isotope Laboratories), 2.5 mM heavy lysine labeled GST peptides, and 0.1% PPS silent surfactant (Protein Discovery). Each diluted plasma sample was boiled at 95 °C for 5 min to denature proteins. After denaturing, dithiothreitol (DTT, Sigma Aldrich #D0632) was added to a final concentration of 5 mM and samples incubated at 60 °C for 30 min to reduce disulfide bonds. Iodoacetamide (Sigma Aldrich #I1149) was then added to a concentration of 15 mM followed by a 30-min room-temperature incubation in the dark to alkylate reduced cysteine. The alkylation reaction was quenched by addition of DTT to a final concentration of 10 mM added. Sequencing grade trypsin (Pierce #1862748) was added to a 1:10 trypsin to protein ratio followed by sample incubation at 37 °C, 1,200 r.p.m. for 4 h to digest proteins. The digestion reaction was quenched by addition of hydrochloric acid to a final concentration of 9.4 mM. The resulting digests of equal volume were pooled to make the plasma library sample.

12 gas-phase fractionations were used in acquiring the DIA plasma library data. Together, the precursor range of 400–1,000 m/z was analyzed, where each fractionation covered a 50-m/z-wide portion of the precursor range 400–450, 450–500, 500–550, 550–600, 600–650, 650–700, 700–750, 750–800, 800–850, 850–900, 900–950, or 950–1,000. 1 µg of plasma sample, 50 fmol of PRTC, and 2.8 ng N15-APO-A1 were loaded in each injection, separated with a 90 min linear gradient LC, and analyzed on a Q-Exactive HF. For each fractionation, DIA method cycled with 25 nonoverlapping 2-m/z-wide isolation MS/MS scans (at 30,000 resolution), one 50-m/z-wide MS scan (at 30,000 resolution), and one 600-m/z-wide MS scan that covers 400–1,000 m/z (at 15,000 resolution). The MS spectra with 400–1,000 m/z precursor range were stripped from mzML files before PECAN analysis.

**Databases and data analysis.** Three sequence databases were used in this manuscript—the GST-fusion-protein database containing 8,207 protein sequences translated from the DNASU human cDNA plasmid library, the human UniProt Swiss-Prot database containing 42,128 protein isoforms, and the UniProt Swiss-Prot human natural variant database containing 74,733 single amino acid variants. For validation of PECAN detection, we targeted the GST-fusion-protein database because we could validate the detection with IVTT-SRM (Supplementary Note 6). For the comparison with DIA-Umpire workflow, we targeted the human UniProt Swiss-Prot database. For building DIA plasma library, we targeted the human UniProt Swiss-Prot database. For querying variant-specific peptides, we targeted the variants from the human natural variant database, which were associated with the proteins detected by PECAN from the plasma library data. In all cases, the human UniProt Swiss-Prot database was used as the background proteome for PECAN.

In all analysis, only fully tryptic peptides with up to one missed cleavage sites were considered, and only a fixed modification of carbamidomethyl cysteine was considered. For PECAN workflow, PECAN (http://pecan.maccosslab.org, v.0.9.9, git commit b4445a3) was used to query peptides from the target database, allowing for 2+ or 3+ precursor charge states. All PECAN analysis was done in y-ion mode where only product y-ion series were considered. For DDA data, Comet (v.2016.01 rev.0) was used to search the MS/MS spectra against the target database21, allowing for up to +4 precursor ions. For DIA-Umpire workflow, DIA-Umpire (v.1.4) was used to extract signal and generate pseudospectra from DIA data, allowing for 2+ to 4+ precursor ions. The resulting pseudo spectra was searched by Comet (http://comet-ms.sourceforge.net, v.2016.01 rev.0) with corresponding charge states. A ±10 p.p.m. mass error tolerance is used for precursor ions (in PECAN and Comet) and fragment ions (in PECAN only). A 0.02 m/z bin width for fragment ions is used in Comet. Both PECAN and Comet results are processed by Percolator10 (v.2.08.01) to separate targets and decoys. All peptides are reported by Percolator at the peptide level with q-value < 0.01, and proteins are reported by Percolator’s built-in Fido algorithm21 with q-value < 0.01 unless otherwise
indicated. For protein comparison, protein groups reported by Fido are considered identical if all protein members of the group are identical.

A Life Sciences Reporting Summary for this paper is available.

Data availability statement. PECAN is open source and freely available at http://pecan.maccosslab.org. All raw data acquired for this manuscript are publicly available at Chorus Project, project number 1105 (Supplementary Table 3). Skyline documents and libraries are publicly available at Panorama Public (https://panorama.web/labkey/pecan-manuscript.url).

14. Bald, T. et al. pymzML—Python module for high-throughput bioinformatics on mass spectrometry data. Bioinformatics 28, 1052–1053 (2012).

15. Martens, L. et al. mzML—a community standard for mass spectrometry data. Mol. Cell. Proteomics 10, R110.000133 (2011).

16. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26, 966–968 (2010).

17. Murray, K.K. et al. Definitions of terms relating to mass spectrometry (IUPAC recommendations 2013). Pure Appl. Chem. 85, 1515–1609 (2013).

18. Granholm, V., Navarro, J.F., Noble, W.S. & Käll, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. J. Proteomics 80, 123–131 (2013).

19. Stergachis, A.B., MacLean, B., Lee, K., Stamatoyannopoulos, J.A. & MacCoss, M.J. Rapid empirical discovery of optimal peptides for targeted proteomics. Nat. Methods 8, 1041–1043 (2011).

20. Davis, M.T. et al. Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. II. Limitations of complex mixture analyses. Proteomics 1, 108–117 (2001).

21. Serang, O., MacCoss, M.J. & Noble, W.S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J. Proteome Res. 9, 5346–5357 (2010).