Published in final edited form as:
Nat Methods. 2017 September ; 14(9): 903–908. doi:10.1038/nmeth.4390.

PECAN: Library Free Peptide Detection for Data-Independent Acquisition Tandem Mass Spectrometry Data

Ying S. Ting[1], **Jarrett D. Egertson**[1], **James G. Bollinger**[1], **Brian C. Searle**[1], **Samuel H. Payne**[2], **William Stafford Noble**[1,3], and **Michael J. MacCoss**[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA

[2]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA

[3]Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA

Abstract

In mass spectrometry-based shogun proteomics, data-independent acquisition (DIA) is an emerging technique for unbiased and reproducible measurement of protein mixtures. Without targeting a specific precursor ion, DIA MS/MS spectra are often highly multiplexed, containing product ions from multiple co-fragmenting precursors. Thus, detecting peptides directly from DIA data is challenging; most DIA data analyses require spectral libraries. Here we present a new library-free, peptide-centric tool PECAN that detects peptides directly from DIA data. PECAN reports evidence of detection based on product ion scoring, enabling detection of low abundance analytes with poor precursor ion signal. We benchmarked PECAN with chromatographic peak picking accuracy and peptide detection capability. We further validated PECAN detection with data-dependent acquisition and targeted analyses. Last, we used PECAN to build a library from DIA data and to query sequence variants. Together, these results show that PECAN detects peptides robustly and accurately from DIA data without using a library.

INTRODUCTION

Mass spectrometry (MS)-based shotgun proteomics is a powerful tool in modern life sciences. In a typical shotgun proteomics experiment, a mixture of proteolytic peptides from sample digestion is separated by liquid chromatography, ionized with electrospray ionization, and then analyzed by tandem mass spectrometry. During tandem mass spectrometry, MS analysis surveys the intact precursor ions, and MS/MS analysis

Correspondence should be addressed to M. J. M. (maccoss@uw.edu).

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 input from J.D.E. and B.C.S. The manuscript 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 MacCoss Lab at the University of Washington has a sponsored research agreement with Thermo Fisher Scientific, the manufacturer of the instrumentation used in this research. Additionally, MJM is a paid consultant for Thermo Fisher Scientific.

characterizes the product ions generated from isolation and fragmentation of the selected precursor ions. Recent improvements in instrumentation scan speed, resolution, mass accuracy, and dynamic range, have positioned data-independent acquisition (DIA)[1] as a viable strategy for analyzing complex peptide mixtures. DIA systematically selects mixtures of precursor ions for MS/MS analysis in an unbiased fashion[2]. The 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 pre-determined 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 of its unbiased measurements comprise a permanent, re-mineable digital record of the sample content[3].

Analysis strategies tailored to DIA data are necessary and subject to intense interest. Library-based approaches, such as OpenSWATH[3] and MSPLIT-DIA[4], facilitate DIA data analysis by making the most of the rich knowledge accumulated from previous studies. These approaches are sensitive, but limited the data interrogation to only analytes present in the library. Thus, tools designed to detect peptides from DIA data without libraries are necessary to deliver on the premise of 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-Umpire[6] and Group-DIA[7], typically generate pseudo-spectra from DIA data by detecting covarying precursor-product ion groups or deconvolving the multiplexed spectra. The quality of each pseudo-spectrum is often dependent on the quality and interpretability of the precursor signal in MS analysis. These pseudo-spectra are sent to conventional database searching pipelines designed for DDA identification, where precursor signal is a key filtering criterion for candidate peptides. As a result, pseudo spectra 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 undetectable precursor signal in DIA data[8]. The lack of detectable precursor signal for some detectable analytes is a common phenomenon resulting from limitation of intra-scan 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-ARM[9], 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 improve, specifically in its sensitivity and false discovery rate. We believe that peptide-centric approaches offer inherent analytical advantages over traditional spectrum-centric approaches when analyzing DIA data[5]. Accordingly, we developed a new peptide-centric tool called PECAN (PEptide Centric ANalysis) that detects peptides directly from DIA data without prerequisite spectral or retention time libraries.

The inputs to PECAN are 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, Online Methods). These scores are used by Percolator[10] to estimate false discovery rate 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 (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 dataset containing 422 synthetic stable isotope-labeled standard (SIS) peptides spiked into various background proteomes with 10 dilution steps[3]. The dataset was published with a manual 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 improved, 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 every 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 (4xGPF, see online Methods) HeLa datasets 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, yielding 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 charge 4 precursors 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 (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), indicating that many of the distinct peptides from two approaches were in fact derived from the same proteins.

To verify the PECAN-DIA specific detections, we randomly selected 16 GST-fusion proteins and expressed them using the *in vitro* transcription translation (IVTT, Fig. 3c). We measured the corresponding 91 peptides using SRM from individually trypsin 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 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, suggesting 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 5–10 times wider isolation windows compared to 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 co-fragmenting analytes[11]. To test how precursor selectivity impacts PECAN's performance in detection, we used gas-phase fractionation (GPF) to vary DIA precursor selectivity while holding the cycle time and sampled precursor *m/z* range constant. GPF DIA data on HeLa were acquired with 20 (1xGPF), 10 (2xGPF), and 5 (4xGPF) *m/z*-wide isolation windows and interrogated using the human UniProt Swiss-Prot database. From the 1xGPF (20 *m/z*), 2xGPF (10 *m/z*), and 4xGPF (5 *m/z*) DIA datasets, 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), indicating 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 datasets, only 30 (0.2%) peptides showed a discrepancy in retention time in either the 1xGPF or 2xGPF compared to the 4xGPF dataset (Fig. 4c), indicating robust peptide detection.

As a benchmark, we processed the GPF DIA datasets with DIA-Umpire followed by Comet database searching. From the 1xGPF, 2xGPF and 4xGPF DIA datasets, DIA-Umpire-Comet identified 13,978, 20,266, and 24,721 unique peptides at Percolator peptide-level *q*-value <0.01. Compared to the results from this DIA-Umpire workflow, PECAN detected 157, 3,132, and 10,092 more unique peptides from 1xGPF, 2xGPF and 4xGPF, respectively, while 9,919, 15,369, and 20,015 peptides were detected by both tools (Fig. 4d). Overall, PECAN outperformed DIA-Umpire more in data with higher precursor selectivity.

Both PECAN and DIA-Umpire showed significant improvements in peptide detection as precursor selectivity increased. Compared to 1xGPF, the 4xGPF setting not only improves the precursor sensitivity for MS analysis by reducing the isolated range of precursor ions in each fractionation, but also improves the precursor selectivity for MS/MS analysis. For DIA-Umpire, the “pre-database searching” process that groups covarying product signals to precursor signals, the 4xGPF improvement in the precursor sensitivity was particularly beneficial as it revealed more precursor ions that were undetectable in 1xGPF. However, 4xGPF improves sensitivity but not resolving power in MS analysis. Thus, if a precursor ion was interfered by chemical noise in 1xGPF, it is likely to be interfered in the 4xGPF. Moreover, the low intensity precursors that are only detectable in 4xGPF are more likely to have interference and more susceptible to stochastic sources of noise such as spray instability, both hinder the detection of precursor/product covariation required by DIA-Umpire. Nonetheless, peptide detection by DIA-Umpire workflow improved by 77% from 1xGPF to 4xGPF. On the other hand, PECAN reports evidence of detection based on product scoring, taking advantage of the fact that improved precursor selectivity generates less product interference. PECAN also benefits from the improvement in precursor sensitivity reflected in the precursor auxiliary scores. Together, PECAN detected 146% more peptides in 4xGPF than 1xGPF.

Building libraries from DIA data with PECAN

Libraries are commonly used to improve sensitivity of peptide detection in DIA data. Typically, these libraries are generated from DDA data, which inherently depends on detectable precursors for identification. With PECAN, a library can be generated directly from DIA data. As a demonstration, we acquired a library dataset using 12 gas phase fractionation DIA runs, each with twenty-five 2-*m/z*-wide isolation windows, on a non-depleted, pooled plasma sample. We queried the data with the human UniProt Swiss-Prot database using PECAN, thereby generating a detection library. From the library data, PECAN detected 3,689 peptides and 520 protein groups, spanning > 5 orders of magnitude of protein concentration (Supplementary Fig. 2). Of the 3,689 detected peptides, 379 were not present in the PeptideAtlas Human Plasma spectral library (2012-08 release) constructed from 177 public datasets. All 379 peptides were from proteins with other peptides detected by PECAN. The fragmentation patterns and retention time information for the peptides detected by PECAN are available on Panorama Public (Online Methods) and can be incorporated into existing libraries.

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 the mass range of 600 to 4000 Da. We used PECAN to query these variant-specific, tryptic peptides against the plasma library data. PECAN detected 133 variant-specific peptides, corresponding to 115 variants (Supplementary Table 1).

In some cases, PECAN detected multiple variant-specific peptides resulting from the same sequence variant. In Serotransferrin (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 were 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 A-1 (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 they were supported by the same MS/MS spectra. Among the three Glu to Lys variants in Apolipoprotein A-1, only Glu160Lys had a definitive peptide resulting from cleavage at the new tryptic site introduced by the variant. Of the twenty-one Glu to Lys variants in the plasma library dataset (Supplementary Table 1), eight were covered by at least one peptides 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 is decreases, PECAN can be used to build libraries directly from DIA data collected using narrow isolation windows and applied to wide isolation data later. This approach can augment existing DDA-based libraries as evidenced by detection of hundreds of novel peptides from 12 LC-MS/MS runs that were either not detected or did not make it through the FDR/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, 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, α and β, with DIA data of various precursor selectivity will be effective. We also expect that incorporating a retention time predictor, such as SSRCalc[12] or BioLCCC[12], 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. PECAN intuitively tests for the detection of each peptide directly, an intuitive approach of hypothesis testing.

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, resulting 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[12] 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.

ONLINE METHODS

PECAN workflow

PECAN uses the open source application programming interface pymzML[13] and supports the HUPO Proteomics Standard Initiative standard file format—mzML[14]. PECAN search results can be imported into Skyline[15], 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)[16]. For a DIA dataset 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

$$XIC_p = \begin{bmatrix} I_{b_1,t_1} & I_{b_1,t_2} & \dots & I_{b_1,t_m} \\ \vdots & \vdots & & \vdots \\ I_{b_{n-1},t_1} & I_{b_{n-1},t_2} & \dots & I_{b_{n-1},t_m} \\ I_{y_1,t_1} & I_{y_1,t_2} & \dots & I_{y_1,t_m} \\ \vdots & \vdots & & \vdots \\ I_{y_{n-1},t_1} & I_{y_{n-1},t_2} & \dots & I_{y_{n-1},t_m} \end{bmatrix}$$

where $I_{x,t}$ 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 ±10 ppm) of *x*. Let the peptide vector (see definition below) corresponding to peptide *p* be V_p. Then the peptide score matrix is calculated as

$$S_p = V_p \cdot XIC_p = [s_{t_1}, s_{t_2} \dots, s_{t_m}]$$

where each s_t is mathematically equivalent to the scalar projection of O_t, the observed MS/MS spectrum at retention time *t*, onto the peptide scoring vector V_p. Because the scalar value s_t 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 $p = [b_2, \dots, b_{n-1}, y_1, \dots, y_{n-1}]$ 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 includes +2 fragment ions for precursor ions with +3 charges and above. The peptide vector for peptide *p* is then

$$V_p = \frac{[w_{b_1}, \dots, w_{b_{n-1}}, w_{y_1}, \dots, w_{y_{n-1}}]}{\|[w_{b_1}, \dots, w_{b_{n-1}}, w_{y_1}, \dots, w_{y_{n-1}}]\|} = [w'_{b_1}, \dots, w'_{b_{n-1}}, w'_{y_1}, \dots, w'_{y_{n-1}}],$$

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 ppm), generated by *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, resulting 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 non-zero. 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 to 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 score each decoy against the data. Let *z* be a charge state of interest. The background score $B_{y,z}$ 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:

$$E_p(t)=\frac{1}{u}\sum_{k=t-\frac{u}{2}}^{t+\frac{u}{2}-1} {S'}_{p_k}$$

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 (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 one hundred 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 seconds at FWHM. If a DIA method has a cycle time of 2 seconds, then a peptide would be measured by MS/MS at least 6 times. In this case, PECAN would then use $u = 6$ for the moving average calculation.

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, suggesting 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, α and β, 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

$$T_p(t)=\frac{1}{N^{\alpha}}\sum_{k=t-\frac{u}{2}}^{t+\frac{u}{2}-1} S_{p_k}$$

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

$$ionS_p(x,t)={w'}_x\cdot\sum_{k=t-\frac{u}{2}}^{t+\frac{u}{2}-1} I_{x,k}$$

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 $C_p = \beta N$, 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 *S. cerevisiae* DIA dataset with 1,224 known boundaries of chromatographic peaks to optimize

α and β for the evidence qualifying procedure (Supplementary Note 2). The resulting values of $\alpha = 1.8$ and $\beta = 0.4$ were used throughout this study.

Estimating detection FDR

PECAN employs Percolator[10], a semi-supervised 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 *E. coli* proteome against HeLa DIA datasets with various DIA isolation schemes. By design, no query peptides were supposed to be detected from the datasets, and thus the target *p*-values should be uniformly distributed[17]. 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, indicating that the auxiliary scores from PECAN did not introduce undesired separation of targets from decoys. Furthermore, tests of the same dataset 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 prior to 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 prior to 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[18]. Finally, these enriched GST-fusion proteins were reduced, alkylated, and digested for 2 h with trypsin individually.

Ninety-one 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 was acquired using a scan width of 0.002 mass to charge ratio (*m/z*) and a dwell time of 10 ms.

HeLa datasets

HeLa protein digest (Thermo) spiked-in with stable-isotope labeled peptides (PRTC, Thermo) was analyzed on a Q-Exactive HF mass spectrometer (Thermo). One 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)[19] settings were used to cover the precursor *m/z* range of 500 to 900: one injection (1xGPF), two injections covering 500–700 and 700–900 *m/z* (2xGPF), and four injections covering 500–600, 600–700, 700–800, and 800–900 *m/z* (4xGPF). 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 4xGPF 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 1xGPF for 2xGPF, and quadruple for 4xGPF.

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 1xGPF DDA, 2xGPF DDA, and 4xGPF DDA (Supplementary Note 5). A standard (one MS analysis with 60,000 resolution followed by twenty MS/MS with 30,000 resolution) DIA method with 20, 10, or 5 *m/z*-wide isolation windows was used to acquire 1xGPF DIA, 2xGPF DIA, or 4xGPF DIA respectively. For FDR control, data from multiple injections were analyzed together as if they were from one instrument run.

Plasma library data

Non-depleted plasma samples from five deidentified donors and a normal female plasma standard (Lampire Biological Laboratories) were individually digested. Plasma samples were diluted 200-fold prior to digestion with a diluent containing heavy labeled protein and peptide standards and PPS silent surfactant in 50 mM ammonium bicarbonate. Post dilution, the sample contained 1 ng/uL 15N-labeled human Apolipoprotein A-1 (Cambridge Isotope Laboratories), 2.5nM heavy lysine labeled GST peptides, and 0.1% PPS silent surfactant (Protein Discovery). Each diluted plasma sample was boiled at 95°C for 5 minutes 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 minutes 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 10mM added. Sequencing grade trypsin (Pierce #1862748) was added to a 1:10 trypsin to protein ratio followed by sample incubation at 37 °C / 1200 RPM for 4 hours 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.

Twelve gas phase fractionations were used in acquiring the DIA plasma library data. Together, the precursor range of 400–1000 *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 950–1000. One 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 non-overlapping 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 400–1000 *m/z* (at 15,000 resolution). The MS spectra with 400–1000 *m/z* precursor range were stripped from mzML files prior to 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 can 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 to 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 (v.0.9.9) 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 database[21], allowing for up to +4 precursor ions. For DIA-Umpire workflow, DIA-Umpire (v.1.4) was used to extract signal and generate pseudo spectra from DIA data, allowing for 2+ to 4+ precursor ions. The resulting pseudo spectra was searched by Comet (v.2016.01 rev.0) with corresponding charge states. A ±10 ppm 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 Percolator[10] (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 algorithm[20] with *q*-value < 0.01, unless indicated otherwise. For protein comparison, protein groups reported by Fido are considered identical if all protein members of the group are identical.

Data and software access

PECAN is open-source, 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://panoramaweb.org/labkey/pecan-manuscript.url).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank L. Käll, A. I. Nesvizhskii, N. Bandeira, 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, Early Career Research Program.

References

1. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods. 2004; 1:39–45. [PubMed: 15782151]
2. Chapman JD, Goodlett DR, Masselon CD. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom Rev. 2014; 33:452–470. [PubMed: 24281846]
3. Röst HL, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 2014; 32:219–223. [PubMed: 24727770]
4. Wang J, et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nat Methods. 2015 advance online publication.

5. Ting YS, et al. Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. Mol Cell Proteomics. 2015; 14:2301–2307. [PubMed: 26217018]
6. Tsou CC, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods. 2015; 12:258–264. [PubMed: 25599550]
7. Li Y, et al. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat Methods. 2015 advance online publication.
8. Panchaud A, et al. Precursor Acquisition Independent from Ion Count: How to Dive Deeper into the Proteomics Ocean. Anal Chem. 2009; 81:6481–6488. [PubMed: 19572557]
9. Weisbrod CR, Eng JK, Hoopmann MR, Baker T, Bruce JE. Accurate Peptide Fragment Mass Analysis: Multiplexed Peptide Identification and Quantification. J Proteome Res. 2012; 11:1621–1632. [PubMed: 22288382]
10. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007; 4:923–925. [PubMed: 17952086]
11. Gillet LC, 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. 2012; 11:O111.016717.
12. Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006; 24:1285–1292. [PubMed: 16964243]
13. Bald T, et al. pymzML - Python module for high throughput bioinformatics on mass spectrometry data. Bioinformatics. 2012; 28:1052–1053. [PubMed: 22302572]
14. Martens L, et al. mzML—a Community Standard for Mass Spectrometry Data. Mol Cell Proteomics. 2011; 10:R110.000133.
15. MacLean B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010; 26:966–968. [PubMed: 20147306]
16. Murray KK, et al. Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). Pure Appl Chem. 2013; 85:1515–1609.
17. Granholm V, Navarro JCF, Noble WS, Käll L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. J Proteomics. 2013; 0:123–131.
18. Stergachis AB, MacLean B, Lee K, Stamatoyannopoulos JA, MacCoss MJ. Rapid empirical discovery of optimal peptides for targeted proteomics. Nat Methods. 2011; 8:1041–1043. [PubMed: 22056677]
19. Davis MT, et al. Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry II. Limitations of complex mixture analyses. Proteomics. 2001; 1:108–117. [PubMed: 11680890]
20. Serang O, MacCoss MJ, Noble WS. Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data. J Proteome Res. 2010; 9:5346–5357. [PubMed: 20712337]

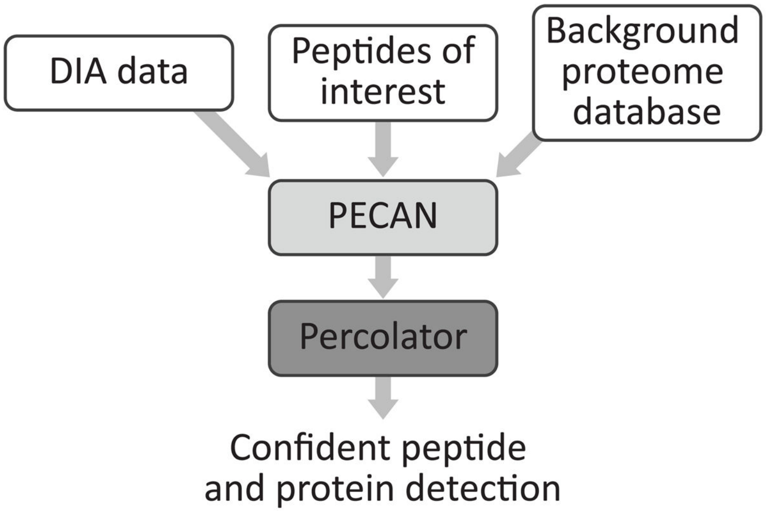
DIA data
Peptides of interest
Background proteome database
PECAN
Percolator
Confident peptide and protein detection

Figure 1. Overview of PECAN workflow

PECAN takes DIA data, peptides of interest, and a background proteome database as inputs, and outputs evidence of detection with auxiliary scores for every query peptide and PECAN generated decoy peptide. Percolator uses PECAN output to train a classifier to distinguish correct and incorrect evidence, and then outputs confident peptide and protein detection with estimated FDR.

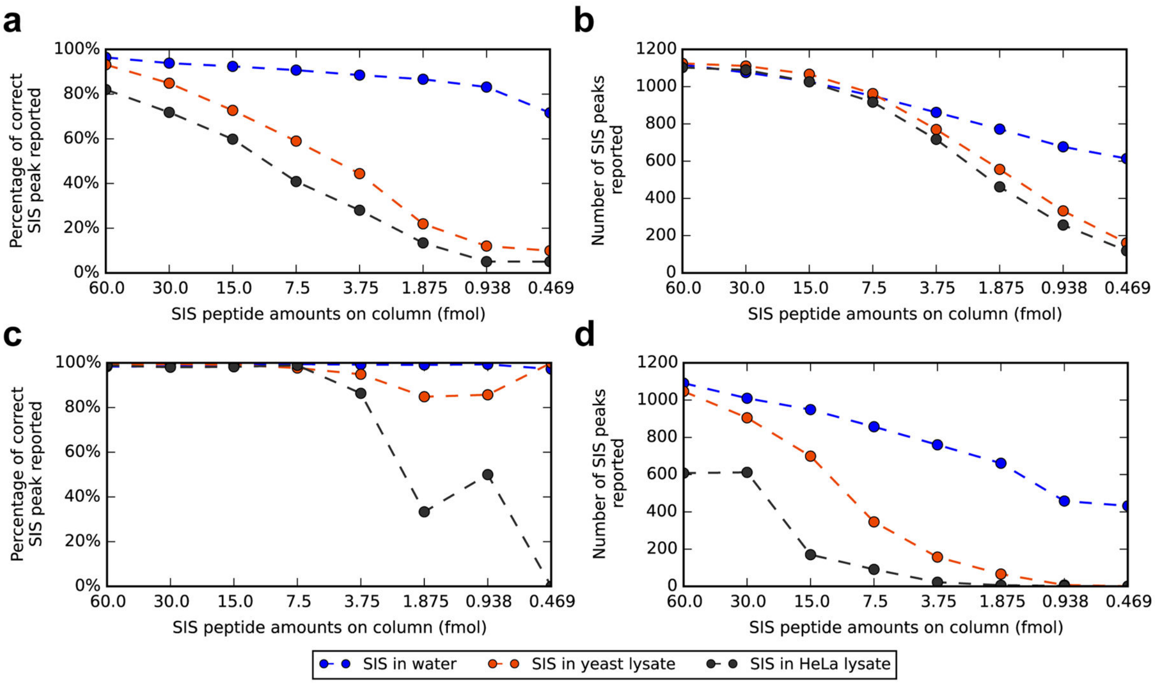

Figure 2. PECAN peak picking performance on SIS dataset

The percentage of total correct SIS peaks (a) and the number of SIS peaks (b) reported by PECAN prior to FDR control from three replicates combined. Same figures (c) and (d) respectively after the PECAN reported evidence of detection were subjected to peptide level FDR control per measurement at *q*-value < 0.01 by Percolator.

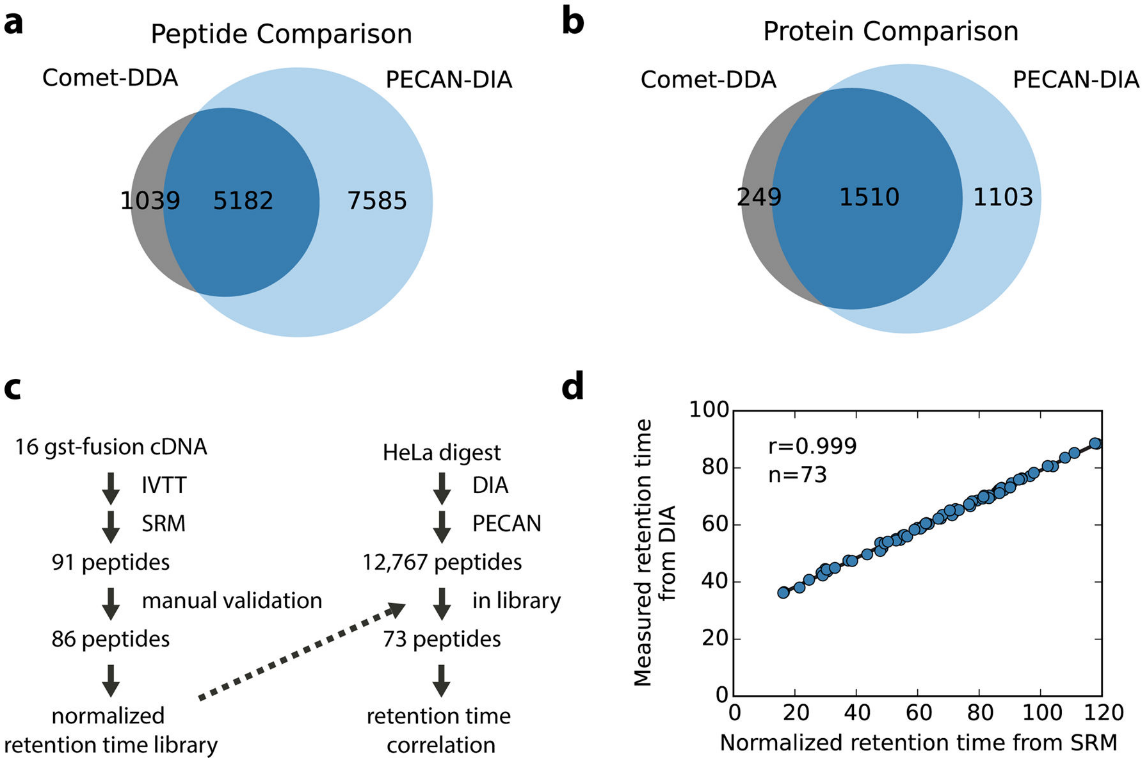

Figure 3. Validate PECAN detection with GST-fusion proteins

Comparative analysis of peptide detection from DIA and DDA data from HeLa protein digest. Peptide (a) and protein (b) comparison of PECAN-DIA detection and Comet-DDA identification. (c) SRM validation workflow for a set of analytical standards synthesized using *in vitro* transcription translation (IVTT). (d) Comparative analysis of retention time of HeLa peptides detected by PECAN from DIA data and IVTT peptides detected from SRM.

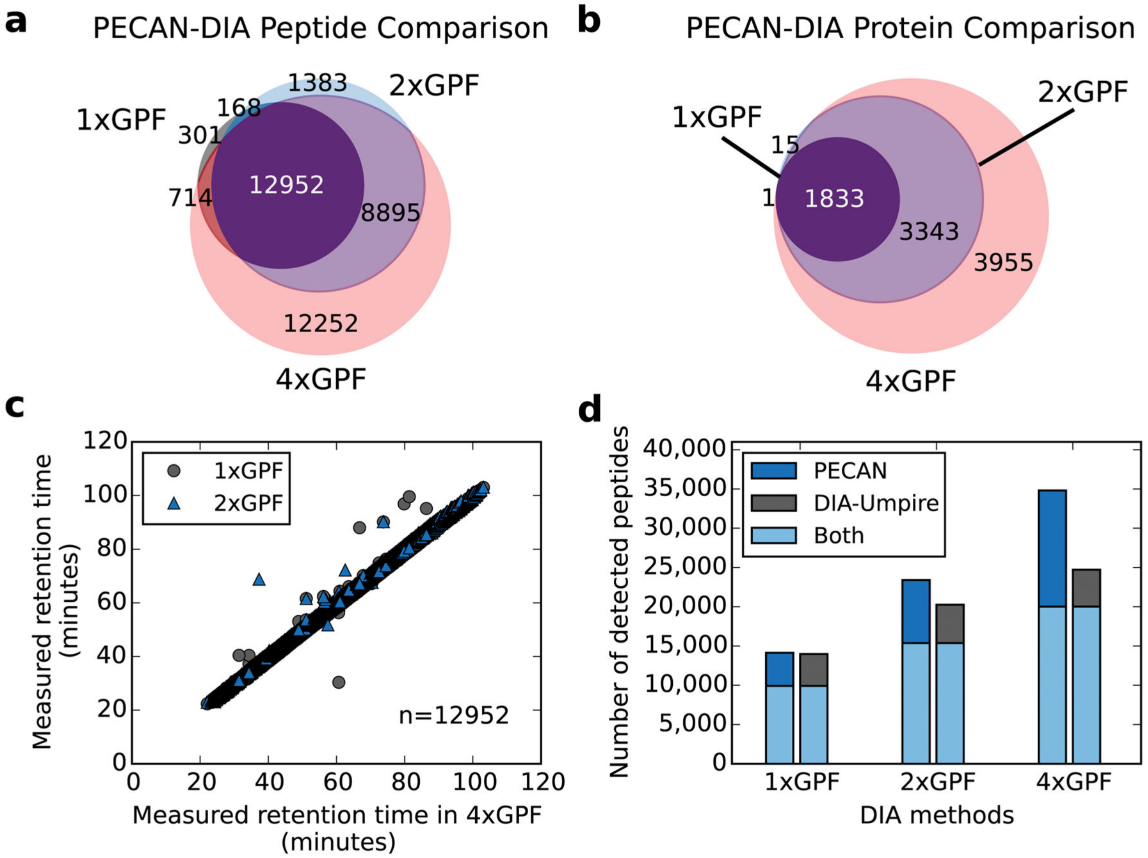

Figure 4. Deep proteome measurement with gas phase fractionation

Comparison of peptides (a) and proteins (b) detected by PECAN from 1xGPF, 2xGPF, and 4xGPF DIA data when queried with the human UniProt Swiss-Prot database. (c) Retention time comparison of 12,952 PECAN detected peptides form 1xGPF and 2xGPF relative to 4xGPF. (d) Number of peptides detected by either, or both PECAN and DIA-Umpire from the three GPF DIA datasets.

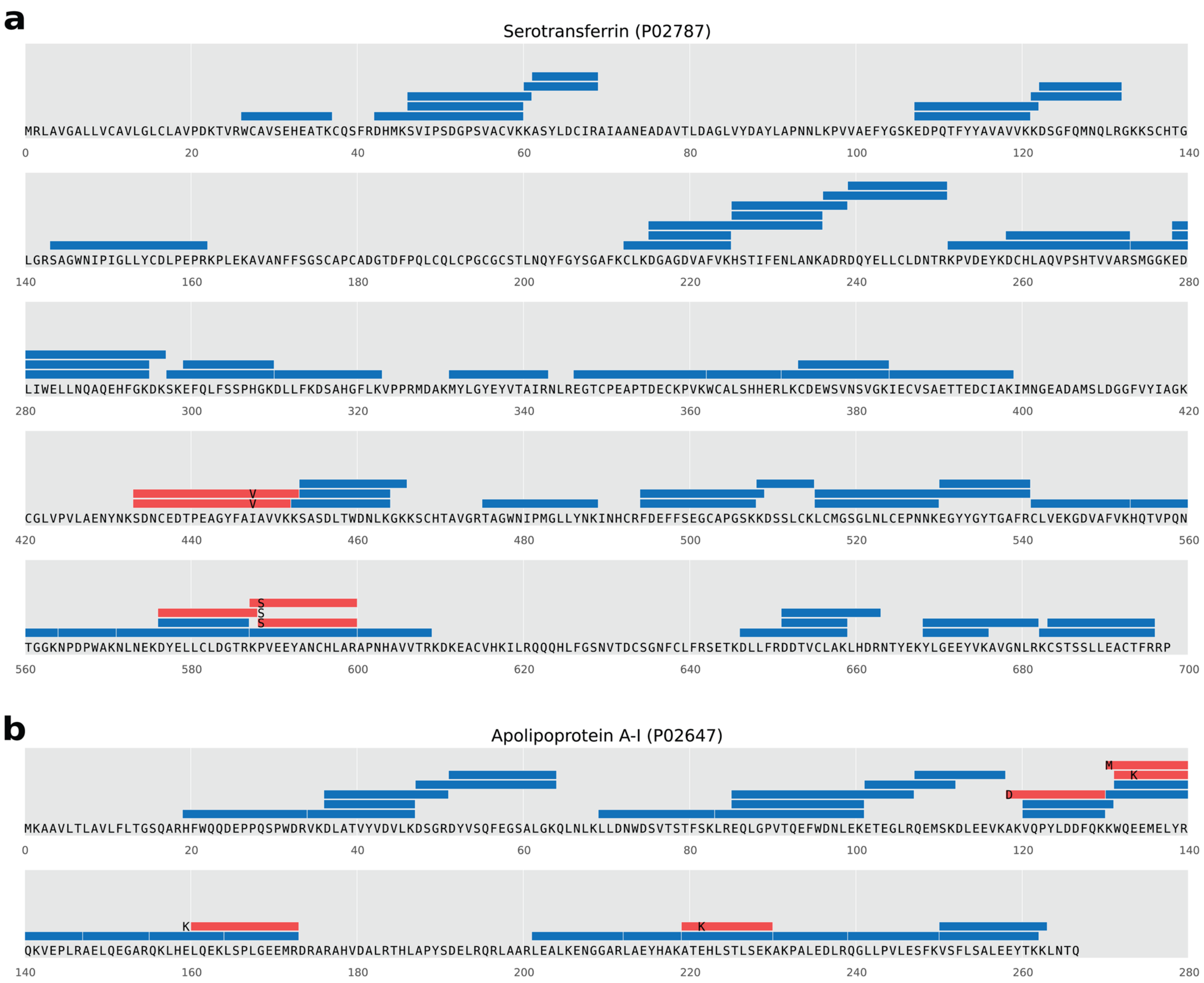

Figure 5. Natural variants in the plasma library data

Full-length canonical sequences of Serotransferrin (a) and Apolipoprotein A-1 (b) are obtained from the human UniProt Swiss-Prot database, accession number P02647 and P02787, respectively. Blue boxes represent PECAN detected peptides from the plasma library data when queried with canonical sequences. Red boxes represent PECAN detected variant-specific peptides from the plasma library data when queried with variant-specific tryptic peptides from 3,714 variants.