QuantumSi

Control More of Your Protein Research

Introducing Platinum™ – The World's First *Next-Generation* Protein Sequencer

The power of protein sequencing is now in your hands! Sequence proteins right in your lab with Platinum™, the NEW benchtop solution from Quantum-Si.

Our first-of-its-kind platform gives you the power to take more control of your protein research by conveniently delivering simple, affordable protein sequencing capabilities right to your bench, without the need for bioinformatics expertise or extensive infrastructure. Now you can get deeper insights into protein biology to complement your existing proteomics approaches...right in your lab!

- Conduct proteomics experiments in your lab *at your bench*
- Interrogate protein variants and modifications, and correlate with biological function
- Achieve deeper proteomics insights faster
- Perform analytics with no bioinformatics expertise required

Introducing Platinum™

The Protein Sequencing Company™

Received: 1 June 2023 | Revised: 21 June 2023 | Accepted: 22 June 2023

DOI: 10.1002/pmic.202300236

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Spectral library search for improved TMTpro labelled peptide assignment in human plasma proteomics

Nicolai B. Palstrøm[1] | Amanda J. Campbell[1] | Caroline A. Lindegaard[1] | Samir Cakar[1] | Rune Matthiesen[2] | Hans C. Beck[1]

[1]Department of Clinical Biochemistry, Odense University Hospital, Odense, Denmark

[2]Computational and Experimental Biology Group, CEDOC, Chronic Diseases Research Centre, NOVA Medical School, Faculdade de Ciências Médicas, Universidade NOVA de Lisboa, Lisbon, Portugal

Correspondence
Hans Christian Beck, Department of Clinical Biochemistry, Odense University Hospital, J.B. Winsløws Vej 4, DK-5000 Odense C, Denmark.
Email: hans.christian.beck@rsyd.dk

Funding information
Danish Cardiovascular Academy

Abstract

Clinical biomarker discovery is often based on the analysis of human plasma samples. However, the high dynamic range and complexity of plasma pose significant challenges to mass spectrometry-based proteomics. Current methods for improving protein identifications require laborious pre-analytical sample preparation. In this study, we developed and evaluated a TMTpro-specific spectral library for improved protein identification in human plasma proteomics. The library was constructed by LC-MS/MS analysis of highly fractionated TMTpro-tagged human plasma, human cell lysates, and relevant arterial tissues. The library was curated using several quality filters to ensure reliable peptide identifications. Our results show that spectral library searching using the TMTpro spectral library improves the identification of proteins in plasma samples compared to conventional sequence database searching. Protein identifications made by the spectral library search engine demonstrated a high degree of complementarity with the sequence database search engine, indicating the feasibility of increasing the number of protein identifications without additional pre-analytical sample preparation. The TMTpro-specific spectral library provides a resource for future plasma proteomics research and optimization of search algorithms for greater accuracy and speed in protein identifications in human plasma proteomics, and is made publicly available to the research community via ProteomeXchange with identifier PXD042546.

KEYWORDS
peptide identification, plasma proteomics, spectral library search, TMTpro

1 | INTRODUCTION

Protein identification in mass spectrometry-based proteomics is based on the identification of peptides containing amino acid sequences unique to the identified protein. Traditionally, sequence database search engines, which identify peptides from tandem mass spectra using a reference protein sequence database [1], have been used to

Abbreviations: AGC, automatic gain control; DDA, data dependent acquisition; DIA, data independent acquisition; FAIMS, field asymmetric Ion mobility spectrometry; HCD, higher-energy collisional dissociation; IAA, Iodoacetamide; PSM, peptide spectrum match; TEAB, triethylammonium bicarbonate; TMT, tandem mass tag.

process raw files generated from mass spectrometry analyses. The search engines work by generating in silico predicted theoretical peptide fragmentation spectra from the protein sequences in the database and matching those with the acquired spectra. Several sequence search engines exist, such as Mascot, MSFragger, Comet, and Sequest HT which differ in the algorithms applied to identify peptides from raw tandem mass (MS/MS) spectra [2–5]. Spectral search engines have emerged as a complementary approach to traditional sequence search engines. Spectral search engines utilize spectral libraries containing pre-acquired tandem mass spectra of peptides, to improve the speed and accuracy of peptide identifications. The libraries are created by acquiring mass spectra of either pure protein standards or protein digests and annotating them with the protein identifications [6]. These identifications are typically made by comparing the mass spectra to known protein sequences using traditional sequence search engines. Spectral library searching involves comparing the acquired spectra to the reference spectra contained in the library and scoring the acquired spectra based on similarity to the reference spectra. Spectral database searches exhibit greater sensitivities than traditional sequence search engines because spectral search engines can leverage more of the information contained in acquired spectra, including ion intensity and the presence of additional fragments to identify peptides [7]. Most commonly applied spectral library-based search engines include MSPepSearch and Spectra ST [8, 9].

One popular method for improving high-throughput analysis is isobaric labelling of peptides using tandem mass tags (TMT) for multiplexed relative protein quantification of multiple samples in a single proteomic experiment [10, 11]. Isobaric-tagged peptides display fragmentation patterns distinctly different from their label-free counterparts where TMT-tagged peptides produce more pronounced b-ion series compared to untagged peptides [12]. This phenomenon was also observed with the 4-and 8-plex isobaric tagging for relative and absolute quantification (iTRAQ) method [13]. Consequently, a genuine spectral library constructed from the analysis of isobaric-tagged peptides would improve the performance of the spectral database as also demonstrated for a genuine 11-plex TMT library [12]. Mass spectrometry-based proteomics has been the primary tool for analysing plasma proteins for many years [14]. Besides high-throughput, DIA-based methods for mass spectrometry-based plasma proteomics [15], DDA-based multiplexed methods are also among the methods of choice for the analysis of larger cohorts. Common for both approaches is that the depth of LC-MS/MS-based plasma proteomics is a balance between the depth of analysis and the number of individual samples, where the latter is crucial for clinical proteomics. Moreover, the discovery of potential protein biomarkers is hindered by the complexity and extreme concentration range of proteins present in plasma [16], and methods that enable the measurement of the increased number of proteins without compromising the throughput of patient samples in clinical plasma proteomics experiments are highly sought after. The recently introduced TMTpro multiplex method allows the multiplex relative quantification of up to 18 samples in a single proteomic experiment, which powers up the sample throughput but with fairly poor depth — even with extensive sample pre-fractionation or sample depletion for high abundant proteins.

SIGNIFICANCE STATEMENT

Discovery of clinical biomarkers is crucial for improving current diagnostics and treatment strategies. However, mass spectrometry-based plasma proteomics is hindered by the complexity and dynamic range of human plasma. Current approaches to reduce the complexity require time-consuming pre-analytical sample preparation, thereby limiting efficiency and throughput. In this study, we have developed and evaluated a TMTpro-specific spectral library for improving protein identification in human plasma proteomics. The results demonstrate that employing the TMTpro spectral library for spectral library searching improves the number of protein identifications compared to conventional sequence database searching, which indicates the feasibility of increasing identifications without additional sample preparation, while also providing a great resource for the research community to leverage its potential.

In this study, we first describe the difference in fragmentation patterns of both b and y ions between TMTpro and label-free MS data. We then construct a tailored human plasma protein TMTpro spectral library using TMTpro 2D-LC MS/MS data from the analysis of highly fractionated and depleted human plasma samples. The library was evaluated by the analysis of additional TMTpro-labelled human plasma samples. A decoy library was constructed based on the target-decoy strategy to estimate false discovery rate (FDR).

2 | MATERIALS AND METHODS

2.1 | Plasma and tissue samples

Human plasma pool was obtained from healthy blood donors at Odense University Hospital. Tissue slices of intraluminal thrombus and diseased arterial tissue from abdominal aortic aneurysm patients were procured as part of a previously published study involving protein biomarkers for abdominal aortic aneurysms (AAA) [17].

2.2 | Sample preparation for proteomics analysis

2.2.1 | Sample preparation

Plasma samples were prepared either as un-depleted, MARS14-depleted or Proteominer enriched identically as previously published [18]. Briefly described, un-depleted plasma proteins were reduced with

TABLE 1 Summary table of high pH fractionated tryptic digests.

Type	High pH fractions	Input amount (μg)	Amount analysed by LC-MS/MS (μg)	Source
Tissue extracts				
Artery walls	10	20	2	[17]
Intraluminal thrombus	10	20	2	
Plasma samples				
MARS14 depleted	10	20	2	[18]
Proteominer enriched	10	20	2	
Undepleted	10	20	2	
Cell digest				
HeLa	10	20	2	Commercially available: Thermo Scientific, Rockford, IL, USA

Note: TMTpro-tagged tryptic digest of protein extracts of tissues from the artery wall and intraluminal thrombus of abdominal aortic aneurysms as well as from plasma samples (depleted/enriched/undepleted), and a commercial human cell lysates were fractionated by high-pH chromatography. In total, 20 μg of each sample type was fractionated into 10 fractions of approximately 2 μg and analysed by LC-MS/MS.

50 mM dithiothreitol (DTT), alkylated with150 mM iodoacetamide (IAA), acetone precipitated, and vacuum centrifuged to dryness. Protein pellets were then redissolved in 8 M urea prior to the addition of a triethylammonium bicarbonate solution (TEAB) followed by tryptic protein digestion (protein:trypsin: 20:1 w/w) and incubation at 37°C overnight. Depletion procedures of 14 high-abundant proteins using the MARS14 spin column (Agilent Technologies, Palo Alto, CA, USA) were performed as previously described [18]. Depleted protein samples were re-dissolved in 0.2 M tetraethyl ammonium bicarbonate (TEAB) for further processing as described below. Protein enrichment using the ProteoMiner Enrichment Kit (BioRad Corporation, Hercules, CA, USA) was carried out as described by the manufacturer's instructions with a few minor adjustments, as previously described [18].

Tryptic peptides from artery walls and intraluminal thrombus from abdominal aortic aneurysms were prepared also as previously described [17] and further processed as described below. A vial of commercially available lyophilized tryptic Pierce HeLa Protein Digest standard (Thermo Scientific, Rockford, IL, USA) was redissolved in 80 μL of 0.2 M TEAB and further processed as described below.

2.2.2 | Isobaric labelling using TMTpro

For the creation of the spectral library, seven μg of each tryptic sample digest was tagged with TMTpro 16-tags (126-134N) in 1:1 ratio (w/w) using the manufacturers' standard procedure (Thermo Scientific, Rockford, IL, USA; Lot. Nr: WA314599).

For the evaluation of the spectral library, 10 sets of 16 samples containing 3 μg of tryptic plasma digest were randomly labelled with any of the available TMTpro 16-tags (126-134N), also in a 1:1 ratio (w/w).

2.3 | Mass spectrometry data generation

2.3.1 | High pH sample fractionation

Samples used for creating the spectra library were offline fractionated by high pH chromatography into 10 fractions prior to LC-MS/MS analysis essentially as previously described (Table 1) [19]. Briefly described, peptide samples were purified using custom-made Poros R2 and Oligo R3 (equal ratio w/w) microcolumns. Peptides were fractionated into 10 fractions using basic pH reversed-phase LC separated on an Acquity UPLC M-Class CSHTM C18 (130Å, 1.7 μm bead size, 300 μm ID × 100 MM length) using a linear gradient from 10% B (20 mM ammonium formate/80% acetonitrile (ACN), pH 9.3) to 55% B using a 25 min linear gradient at 6 μL/min flow rate on a Dionex Ultimate 3000 RSLCnano system (Thermo Scientific, Bremen, Germany).

2.3.2 | Mass spectrometry

Nano-LC-MS/MS analysis of samples for spectral library construction was performed on an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nano-HPLC interface (Dionex UltiMate 3000 nano-HPLC, Thermo Scientific, Bremen, Germany). Mass spectra was acquired using data-dependent acquisition (DDA) with a Orbitrap MS1 scan (400–1200 m/z, 120,000 resolution, 100% AGC, auto injection time) followed by a Orbitrap MS2 scan (30,000 resolution, TurboTMT enabled, HCD, 35% normalized collision energy, 0.7 m/z isolation width, 30 s dynamic exclusion, 200% AGC target, 86 ms injection time). Field asymmetric waveform ion mobility spectrometry (FAIMS) was used with standard resolution, 3.8 L/min gas flow and two different compensation voltages (CV): -50 and -70. A cycle time of 1.5 s was allotted to each CV value. Samples for spectral library construction were loaded on

custom-made fused capillary pre-columns and separated on custom-made columns essentially as previously described (19) using a linear gradient ranging from 90% to 86% solvent A (0.1% formic acid, Fluka, Seetze, Germany) to 30%–34% B (80% acetonitrile (J.T. Baker, Gliwice, Poland) in 0.1% formic acid) over 106 min followed by 5 min at 90% B and 5 min at 98% A at a flow rate of 250 nL/min, whereas samples for evaluation of the spectral library were separated on a double nanoViper PepMap Neo UHPLC column (50 cm length, 75 μm ID, 2 μm C18) using a linear gradient from 90% solvent A to 25% B over 171 min.

Nano-LC-MS/MS analysis of samples for evaluation of the spectral library was performed on an Orbitrap Tribrid Eclipse (Thermo Fisher Scientific, San Jose, CA, USA). MS settings for samples used for evaluation were unchanged except for dynamic exclusion was increased to 60 s and an isolation width of 0.4 m/z was used instead.

2.4 | Spectral library generation

Orbitrap Exploris 480 raw files acquired for spectral library construction were converted to mzXML files using MSConvertGUI from ProteoWizard (v. 3.0.23052) [20] with peakPicking enabled and zlib compression disabled. The resulting mzML files were then processed using the FragPipe (v. 19.1) suite (https://fragpipe.nesvilab.org/) applying the 'TMT16' template in which files were searched using MSFragger (v. 3.7) and, peptide validation using PeptideProphet (v. 4.8.1), both applying default settings. The generated .pepXML files were then imported into SpectraST (v. 5.0) as part of the Trans Proteomic Pipeline (v. 6.2.0) [21], which then generated the initial raw spectral library from the .pepXML files. SpectraST was then used to generate the consensus spectral library. The consensus spectral library was finally filtered for impure, low-quality spectra and spectra of non-TMTpro labelled peptides.

2.5 | Data analysis

All 10 raw data files (10 TMTpro sets) were processed using MSPepSearch and Sequest HT in parallel integrated into the Proteome Discoverer version 2.4 (Thermo Scientific, San Jose, CA, USA). All searches were performed against the SwissProt database restricted to human (downloaded 23 January 20P23, 20,330 entries) [22]. Sequest HT searches were performed using trypsin with two missed cleavages allowed, an 8 ppm precursor mass tolerance and a 0.05 Da fragment mass tolerance. Static modification was limited to carbamidomethylation at cysteines and TMTpro at lysine and N-terminal amines, while methionine oxidation and deamidation of asparagine and glutamine were set as dynamic. Percolator was used for FDR calculation and filter out non-confident peptides. We chose to use the generated spectral library to process data using MSPepSearch in Proteome Discoverer. The generated spectral library is imported into Proteome Discoverer, which also handles the creation of decoy spectral library for FDR estimations using a version of the reverse sequence decoy spectrum algorithm implemented in Proteome Discoverer [23]. Searches using MSPepSearch were used with a precursor mass tolerance and fragment mass tolerance of 15 ppm.

3 | RESULTS

In this study, we aimed to develop and validate a tailored human plasma protein spectral library based on tandem mass spectra from the analysis of TMTpro-labelled plasma samples for improved analysis of plasma samples. A TMTpro consensus spectral library was constructed by searching raw files against a sequence database followed by the creation and validation of a decoy spectral library for FDR estimations. Finally, we compared the performance of the constructed spectral library with searches against a sequence database.

3.1 | Fragmentation difference in MS/MS analysis of TMTpro-tagged and untagged spectra

MS/MS fragmentation patterns of isobaric-tagged peptides are known to produce fragmentation patterns that are markedly different from their un-tagged counterparts [12, 13]. Chemical tagging of the peptide amino termini and lysine side chains alters peptide proton affinities and gas phase basicities that again may lead to an alteration of peptide fragmentation pattern [24], as frequently observed for both K- and R-containing peptides in the present study (results not shown). A systematic comparison of the distribution of b- and y-ion intensities of un-tagged and TMTpro-tagged peptides (Figure 1) actually showed a notable increase in the abundance of b-ions for both K- and R-containing TMTpro-tagged peptides, whereas the impact of TMTpro-tagging on the decrease in intensity of y-ions is more pronounced for K-containing peptides. This indicates that TMTpro-tagging changes peptide fragmentation pattern during MS/MS analysis, in all arguing for the construction of a TMTpro-specific spectral library when searching TMTpro-tagged MS/MS.

3.2 | Construction of a tailored human plasma protein TMTpro spectral library

The construction of a TMTpro-specific spectral library was based on data from LC-MS/MS analysis of highly fractionated TMTpro-tagged human plasma samples as well as human cell extracts (HeLa cells), and TMTpro-tagged protein extracts from aortic tissue in order to create a comprehensive spectral library for improved peptide identification tailored to plasma proteomics (Figure 2). First, raw data were searched against the target-decoy human database by using MSFragger. Then peptides assigned to MS/MS spectra were validated using PeptideProphet. Next, SpectraST was used to generate the initial raw spectral library. The initial raw spectral library contained 319,301 spectra. Replicate spectra assigned to the same peptide identification were merged into a single consensus spectrum. The spectral library was

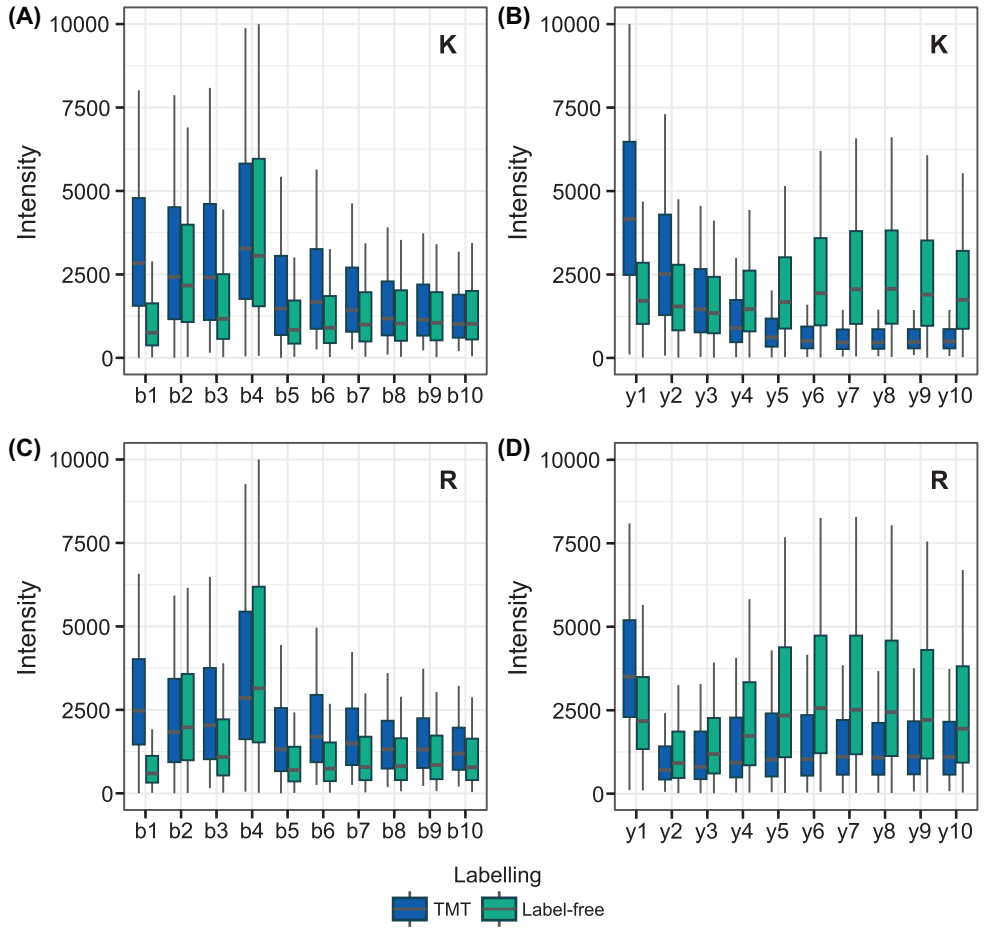

FIGURE 1 Distribution of fragment ions for TMTpro (blue) and label-free data (green). Boxplot of intensity distribution of b1-b10 [A, C] and y1-y10 ions [B, D] of K-and R-containing peptides. Data are from 64623 TMTpro-tagged peptides and 44998 un-tagged peptides.

then carefully curated by the application of several quality filters, using SpectraST, to ensure the reliability of peptide identifications. The quality filters included removing un-tagged spectra, noisy low quality spectra, as well as spectra from peptides that were only observed once. The final spectral library contained 52,599 spectra.

3.3 | Evaluation of the TMTpro decoy spectral library for FDR estimation

As for a sequence database search, an estimated FDR is also a requirement for spectral database searches. We used a target-decoy search for FDR estimation, and a decoy spectral library was created by Proteome Discoverer (v. 2.4), using a reverse decoy library algorithm during spectral library import into Proteome Discoverer. To evaluate the decoy library created by Proteome Discoverer, we examined the distributions of the scores of peptide spectrum matches (PSMs) for both the target and decoy hits as calculated by the MSPepSearch search engine (Figure 3). The score distribution of the target and decoy library is remarkably similar, which demonstrate the suitability of the decoy library for FDR estimation.

3.4 | Evaluation of the TMTpro spectral library

We then compared the efficacy of spectral library searching with a TMTpro-specific library to sequence database searching. To accomplish this, we examined data from the LC-MS/MS analysis of the 10 replicate TMTpro-tagged datasets of human plasma samples. First, we compared the number of peptides and proteins found using each search engine. As shown in Figure 4, sequence database searching using Sequest HT identified the greatest number of unique peptides, with an average of unique 2750 peptides across the 10 datasets, while spectral library searching using MSPepSearch identified an average of 2375 unique peptides. Furthermore, the number of proteins identified by each search method differed, with the sequence database search identifying, on average, 272 proteins and the spectral library search identifying 340 proteins. In summary, the spectral library search using MSPepSearch identified, on average, 25% more proteins than the sequence database search using Sequest HT across the individual sets. The discrepancy between the higher numbers of peptides identified by sequence database searches, but lower number of protein identifications is related to the fact that more peptides per protein are identified by sequence database searches (Figure S1). In addition, spectral library

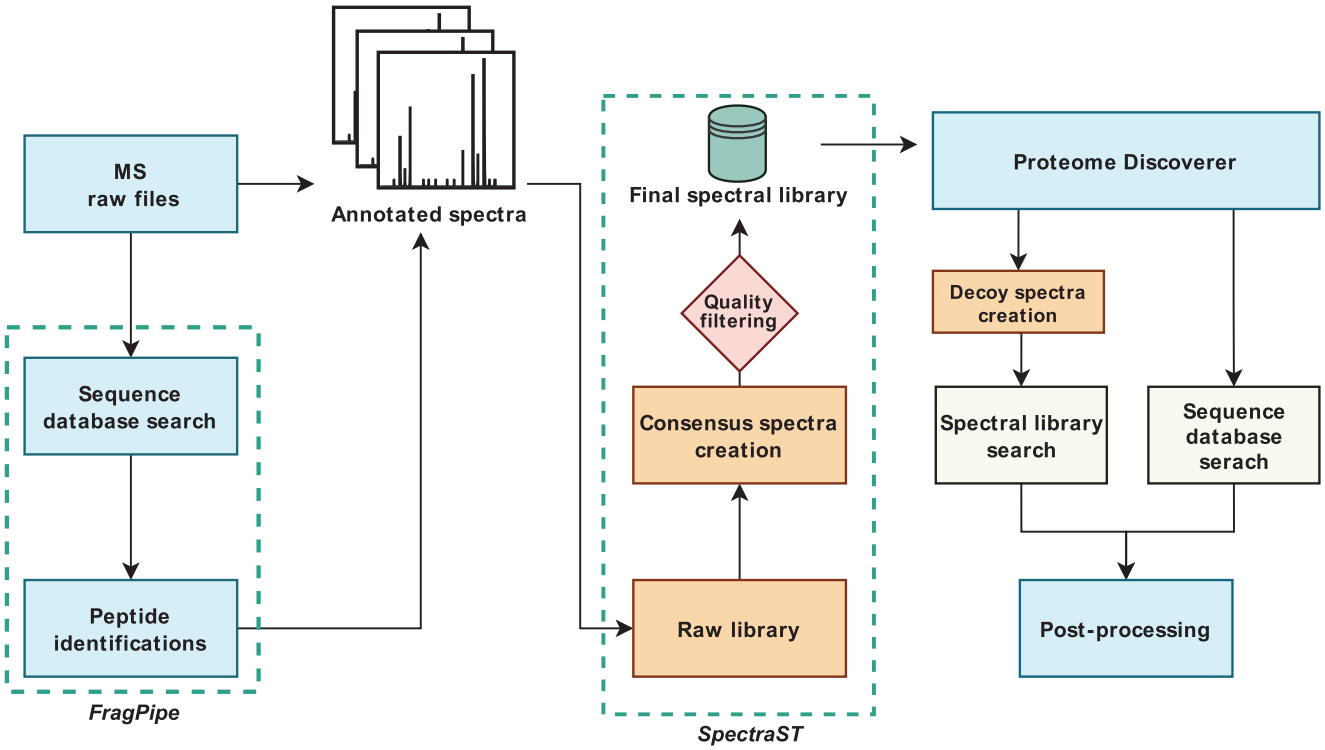

FIGURE 2 Flow chart of TMTpro spectral library generation. Raw files were first converted to mzXML files and then searched by MSFragger and peptide identifications was validated by PeptideProphet using the FragPipe suite. An initial raw version of the spectral library was created from the annotated spectra, using SpectraST, prior to the creation of consensus spectra. In the consensus step, spectra from identical peptides are merged together into one consensus spectrum. Then, spectra with no labelling, low quality or spectra from peptides only observed once was removed as part of the quality filtering. The final spectral library was imported into Proteome Discoverer and used in combination with the sequence database search engine to evaluate the spectral library.

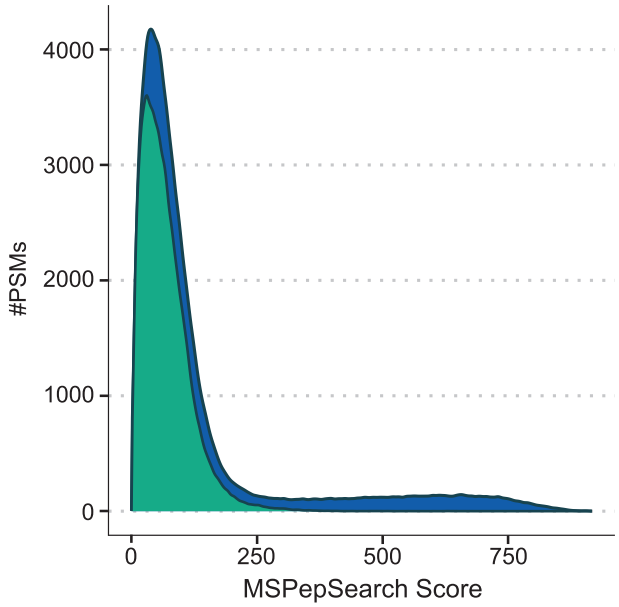

FIGURE 3 Evaluation of score distributions of the reverse decoy library. The distribution of the MSPepSearch scores for hits in the decoy library (green) generated using Proteome Discoverer greatly resembles the distribution of the scores for hits in the target library (blue). Data from 10 replicate TMTpro-sets are presented.

searches seem to identify more proteins with only one unique peptide than sequence database searches (Figure S1).

Then we explored the overlap between the total number of unique peptide and protein identifications obtained by each search method. With an 80% overlap, the majority of the protein groups identified by sequence database search were also identified by spectral library searching (Figure 5A). The overlap between the peptides identified by the sequence database search and the spectral library search, on the other hand, was lower, at 67% (Figure 5B). These findings suggest that spectral library searching with the TMTpro spectral library can identify a large number of peptides and protein groups compared to sequence database search. This suggests that the spectral library-based approach can be used in addition to traditional sequence database search engines to identify peptides in complex samples like plasma.

Finally, we examined the number of PSMs identified at different false discovery rate thresholds (Figure 6). The *q*-values, which represent the minimal FDR threshold observed for a PSM, were compared for both search engines. The analysis revealed that the spectral search engine produces considerably more PSMs at lower *q*-values compared to the sequence database search engine (Figure 6B, blue graph). Therefore, our findings suggest that the choice of the search engine has a notable influence on the results, since a substantial increase in the number of PSMs at the same *q*-value is found using the spectral search

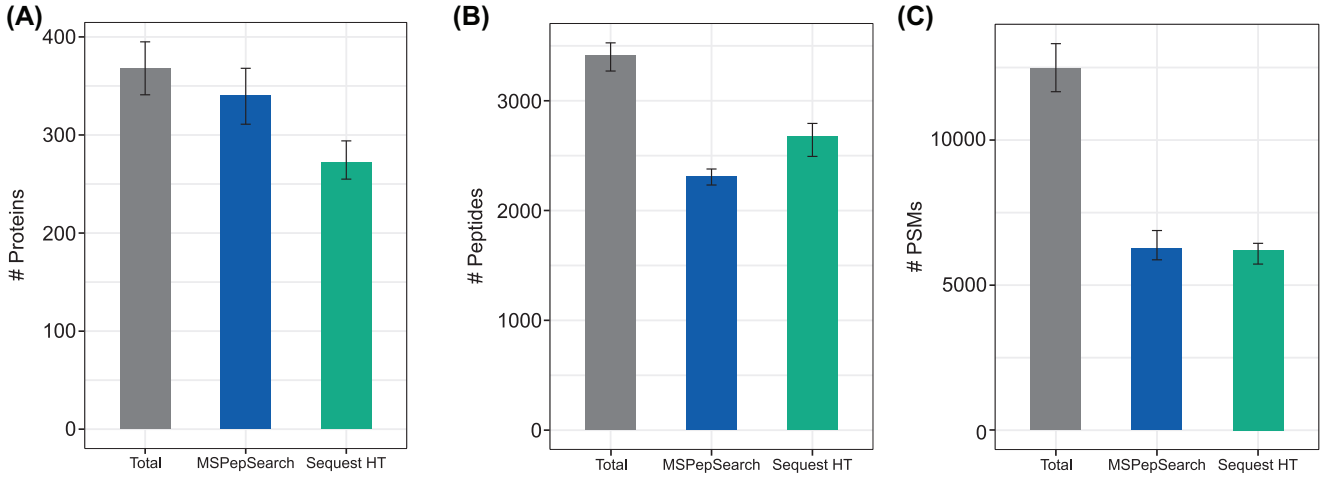

FIGURE 4 Comparison of spectral library searches with sequence database searches. Spectral library searches using MSPepSearch with the TMTpro spectral library increase the number of proteins (A), but not the number of peptides (B) identified across the 10 TMTpro sets, compared to sequence database searches using Sequest HT. No major difference in the number of PSMs (C) identified between the two search methods was observed. Data is presented as mean ± minimum/maximum value.

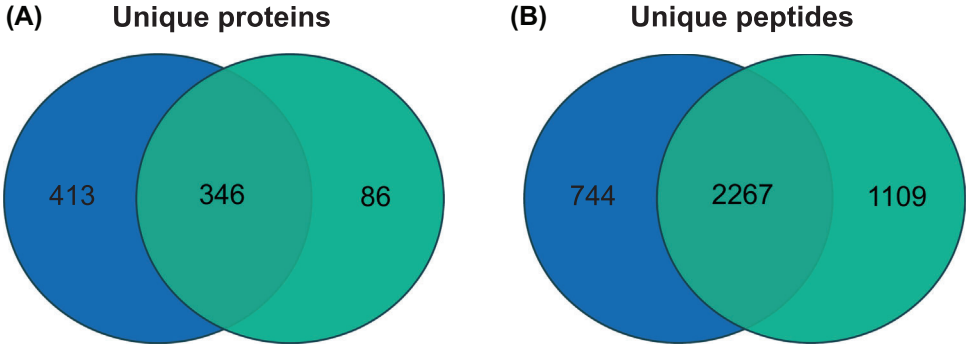

FIGURE 5 Venn diagrams of the total number of overlapping identified (A) proteins, or (B) peptides. Searches performed using MSPepSearch (blue) with the TMTpro spectral library returns considerably more unique protein identifications compared to sequence database searches using Sequest HT (green).

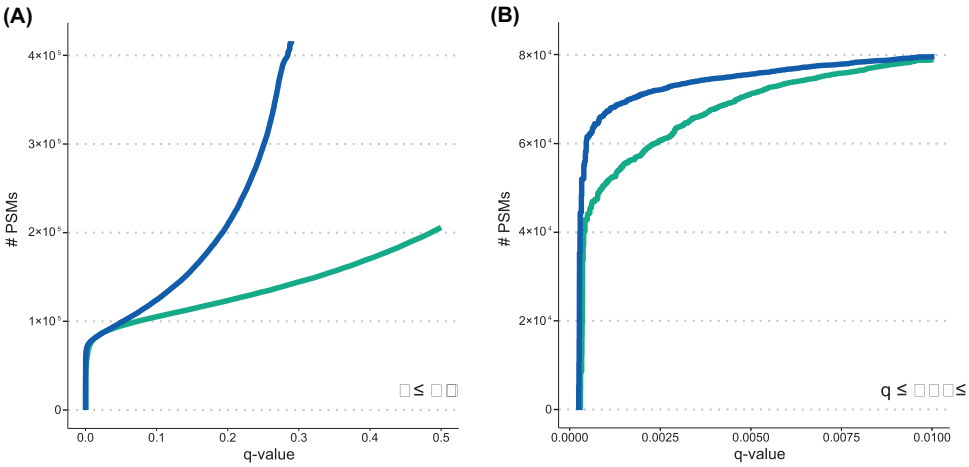

FIGURE 6 Comparison of the number of PSMs for each search engine at different *q*-values. Spectral library searches generate considerably higher numbers of PSMs at lower *q*-values than the sequence database searches. This is most clearly demonstrated for higher *q*-values (A: $q \leq 0.5$), but the same tendency can be seen for *q*-value considerably lower than the conventional 1% FDR threshold (B: $q \leq 0.01$). Blue: MSPepSearch; Green: Sequest HT.

engine, as both search engines select candidates from identical MS/MS spectra (Figure 6).

4 | DISCUSSION

Plasma is one of the most commonly used samples for disease biomarker discovery in clinical research. The complexity and high dynamic range of plasma represent significant challenges that limit the number of protein identifications in mass spectrometry-based proteomics. Current methods for increasing the depth of protein identifications involve laborious pre-analytical sample preparation methods such as depletion of high-abundant proteins, enrichment of low-abundant proteins and/or extensive sample pre-fractionation employing chromatographic methods orthogonal to reversed-phased chromatography normally applied in MS-based proteomics [18]. We proposed that a tailored plasma proteome spectral library could contribute to meet the demand for improved protein identification in plasma when combining a spectral library search with a conventional database search without the application of the above-mentioned pre-analytical procedures. In this study, we generated a TMTpro-specific spectral library using a shotgun proteomics approach and we analysed LC-MS/MS data from the analysis of 10 replicate TMTpro-labelled human plasma samples and compared the performance of sequence database searching (Sequest HT) and spectral library searching (MSPepSearch) for plasma proteomics.

During labelling with TMTpro tags, the amine reactive group binds to the N-terminus of a peptide or a lysine residue. While this process has limited impact on the LC separation, others have demonstrated that iTRAQ- and TMT-labelling alter the intensities of the fragment ions produced in the mass spectrometer [12, 13]. We have demonstrated that the TMTpro-labelling also alters the intensities of the fragment ions, which ultimately validates the need for TMTpro-specific spectral libraries, due to the vastly different intensity distribution of fragment ions in TMTpro-labelled peptides compared to label-free peptides.

Ideally, the features of a decoy library should be comparable to a target library in terms of the distribution of different peptide lengths, precursor masses, and size of the library. However, the decoy library should still be different enough in order to be statistically relevant for estimating the false discovery rate. In other words, the scores assigned to decoy spectra should resemble those of target spectra. We examined the decoy spectral library generated by Proteome Discoverer and found that the distribution of MSPepSearch scores resemble those of the target library, thereby demonstrating the usability of the decoy spectral library for FDR estimation. Proteome Discoverer was utilized in order to streamline data processing and ensure greater consistency in important steps, such as FDR calculation and protein grouping and ensure fair comparison between sequence database and spectral library searches.

We hypothesized that a spectral library search would increase the number of protein identifications when combined with a sequence database search as previous studies have demonstrated [25]. For example, in a recent study, Dorl et al. demonstrated the effectiveness of using a public spectral library with a spectral library search engine they had developed, which was 4.3% more than the conventional sequence database search using Sequest HT and Inferys rescoring [26]. Additionally, Shen et al. analysed TMT 10-plex data from 56 distinct LC-MS/MS experiments [12]. Although the authors did not report the number of proteins identified, the authors noted that spectral library searching with a TMT 10-plex spectral library identified 30% more PSMs than a sequence database search engine they had developed [12]. We have also demonstrated that the use of spectral libraries adds additional protein identifications to the sequence database searches, but also has the capability to outperform sequence database searches in terms of the number of protein identifications. Additionally, the spectral library searches resulted in more identifications across the analysis of 10 replicate TMTpro-tagged plasma samples. The spectral library searches identified up to 25% more proteins compared to traditional sequence database searches. In contrast to previous studies [12, 26], we did not find an increase in the number of identified PSMs. This may, however, be attributed to the use of human plasma to investigate the differences between spectral library search and sequence database search in this study, compared to previous studies that have used human cell lysates instead. The number of identified PSMs may be limited by the complex dynamic range of human plasma compared to human cell lysates in which proteins are more evenly distributed in terms of concentration. Likewise, the application of different spectral libraries between this study and previous studies will affect the number of identified PSMs. We further showed that the number of PSMs with low q-values was higher using spectral library searches compared to sequence database searches. Similar increases in the number of PSMs identified using spectral library search have also previously been demonstrated for label-free data, which also demonstrated the effect of using different spectral libraries on the number of identified PSMs [26]. Also worth mentioning is that the performance of spectral databases may be instrument specific, that is, optimal results are achieved when the searched data originate from the MS instrument that acquired data for the construction of the spectral library [27]. In our study, data for spectral library construction and data for testing the spectral library was acquired on different MS instruments, which further underlines the performance of the spectral library as the performance of spectral library searches may be instrument-dependent [27].

The superior performance of spectral library searches over sequence database searches is likely due to the fact that spectral libraries contain a vast amount of additional spectral data that can be used to identify peptides and proteins with greater confidence [7]. In contrast, sequence databases searches rely on the theoretical sequence data, which lacks additional information relating to the intensities of the fragment ions as well as non-canonical fragment ions that might be present in the empirical spectra. However, the process of generating a spectral library is also limited by the necessity of sequence database searches for spectra annotations. Aside from large, centralized efforts to generate comprehensive spectral libraries, such as PeptideAtlas [28] or NIST, spectral libraries are often limited in

proteome and sequence coverage. Nevertheless, the TMTpro-specific spectral library we generated has proven to be a comprehensive alternative to traditional sequence database searching for plasma proteomics and will likely also assist in improving other areas of proteomics that utilize TMTpro labelling.

In conclusion, the results of this study demonstrate that the use of a TMTpro-specific spectral library significantly improves the identification of proteins in plasma samples compared to traditional sequence database searches. Our spectral library represents a resource for future proteomics research, particularly in the context of plasma proteomics. Future research can build on these findings to develop more specialized spectral libraries and optimize search algorithms for even greater accuracy and speed in protein identifications.

ACKNOWLEDGEMENTS

This work was supported by a research grant from the Danish Cardiovascular Academy, which is funded by the Novo Nordisk Foundation, grant number NNF20SA0067242 and the Danish Heart Foundation.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [29] partner repository with the dataset identifier PXD042546.

ORCID

Nicolai B. Palstrøm https://orcid.org/0000-0003-1480-6550

REFERENCES

1. UniProt Consortium (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*, D523–d531.
2. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*, 3551–3567.
3. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods*, *14*, 513–520.
4. Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *Proteomics*, *13*, 22–24.
5. Tabb, D. L. (2015). The SEQUEST family tree. *Journal of the American Society for Mass Spectrometry*, *26*, 1814–1819.
6. Deutsch, E. W., Perez-Riverol, Y., Chalkley, R. J., Wilhelm, M., Tate, S., Sachsenberg, T., Walzer, M., Käll, L., Delanghe, B., Böcker, S., Schymanski, E. L., Wilmes, P., Dorfer, V., Kuster, B., Volders, P.-J., Jehmlich, N., Vissers, J. P. C., Wolan, D. W., Wang, A. Y., & Röst, H. (2018). Expanding the use of spectral libraries in proteomics. *Journal of Proteome Research*, *17*, 4051–4060.
7. Zhang, X., Li, Y., Shao, W., & Lam, H. (2011). Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics*, *11*, 1075–1085.
8. Stein, S. E., & Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, *5*, 859–866.
9. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., & Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, *7*, 655–667.
10. Thompson, A., Wölmer, N., Koncarevic, S., Selzer, S., Böhm, G., Legner, H., Schmid, P., Kienle, S., Penning, P., Höhle, C., Berfelde, A., Martinez-Pinna, R., Farztdinov, V., Jung, S., Kuhn, K., & Pike, I. (2019). TMTpro: Design, synthesis, and initial evaluation of a proline-based isobaric 16-plex tandem mass tag reagent set. *Analytical Chemistry*, *91*, 15941–15950.
11. Li, J., Cai, Z., Bomgarden, R. D., Pike, I., Kuhn, K., Rogers, J. C., Roberts, T. M., Gygi, S. P., & Paulo, J. A. (2021). TMTpro-18plex: The expanded and complete set of TMTpro reagents for sample multiplexing. *Journal of Proteome Research*, *20*, 2964–2972.
12. Shen, J., Pagala, V. R., Breuer, A. M., Peng, J., Bin Ma, & Wang, X. (2018). Spectral library search improves assignment of TMT labeled MS/MS spectra. *Journal of Proteome Research*, *17*, 3325–3331.
13. Gabriel, W., Giurcoiu, V., Lautenbacher, L., & Wilhelm, M. (2022). Predicting fragment intensities and retention time of iTRAQ- and TMTPro-labeled peptides with Prosit-TMT. *Proteomics*, *22*, 2100257.
14. Palstrøm, N. B., Matthiesen, R., Rasmussen, L. M., & Beck, H. C. (2022). Recent developments in clinical plasma proteomics-applied to cardiovascular research. Biomedicines, *10*.
15. Zhou, Y., Tan, Z., Xue, P., Wang, Y., Li, X., & Guan, F. (2021). High-throughput, in-depth and estimated absolute quantification of plasma proteome using data-independent acquisition/mass spectrometry ("HIAP-DIA"). *Proteomics*, *21*, 2000264.
16. Anderson, N. L, & Anderson, N. G. (2002). The human plasma proteome. *Molecular & Cellular Proteomics*, *1*, 845–867.
17. Behr Andersen, C., Lindholt, J. S., Urbonavicius, S., Halekoh, U., Jensen, P. S., Stubbe, J., Rasmussen, L. M., & Beck, H. C. (2018). Abdominal aortic aneurysms growth is associated with high concentrations of plasma proteins in the intraluminal thrombus and diseased arterial tissue. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *38*, 2254–2267.
18. Palstrøm, N. B., Rasmussen, L. M., & Beck, H. C. (2020). Affinity capture enrichment versus affinity depletion: A comparison of strategies for increasing coverage of low-abundant human plasma proteins. International Journal of Molecular Sciences, *21*.
19. Andersen, L. C., Palstrøm, N. B., Diederichsen, A., Lindholt, J. S., Rasmussen, L. M., & Beck, H. C. (2021). Determining plasma protein variation parameters as a prerequisite for biomarker studies-A TMT-based LC-MSMS proteome investigation. Proteomes, *9*.
20. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., & Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, *30*, 918–920.
21. Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., & Aebersold, R. (2010). A guided tour of the trans-proteomic pipeline. *Proteomics*, *10*, 1150–1159.
22. Magrane, M. (2011). UniProt Knowledgebase: A hub of integrated protein data. *Database (Oxford)*, bar009.
23. Zhang, Z., Burke, M., Mirokhin, Y. A., Tchekhovskoi, D. V., Markey, S. P., Yu, W., Chaerkady, R., Hess, S., & Stein, S. E. (2018). Reverse and random decoy methods for false discovery rate estimation in high mass accuracy peptide spectral library searches. *Journal of Proteome Research*, *17*, 846–857.
24. Dongré, A. R., Jones, J. L., Somogyi, Á., & Wysocki, V. H. (1996). Influence of Peptide composition, gas-phase basicity, and chemical modi-

fication on fragmentation efficiency: Evidence for the Mobile Proton Model. *Journal of the American Chemical Society*, *118*, 8365–8374.
25. Dai, Y., Millikin, R. J., Rolfs, Z., Shortreed, M. R., & Smith, L. M. (2022). A hybrid spectral library and protein sequence database search strategy for bottom-up and top-down proteomic data analysis. *Journal of Proteome Research*, *21*, 2609–2618.
26. Dorl, S., Winkler, S., Mechtler, K., & Dorfer, V. (2023). MS Ana: Improving sensitivity in peptide identification with spectral library search. *Journal of Proteome Research*, *22*, 462–470.
27. Yang, Y., Liu, X., Shen, C., Lin, Y., Yang, P., & Qiao, L. (2020). In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications*, *11*, 146.
28. Desiere, F. (2006). The PeptideAtlas project. *Nucleic Acids Research*, *34*, D655–D658.
29. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D. J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., Walzer, M., Wang, S., Brazma, A., & Vizcaíno, J. A. (2022). The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, *50*, D543–d552.

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202300236 in the Supporting Information section at the end of the article.

How to cite this article: Palstrøm, N. B., Campbell, A. J., Lindegaard, C. A., Cakar, S., Matthiesen, R., & Beck, H. C. (2023). Spectral library search for improved TMTpro labelled peptide assignment in human plasma proteomics. *Proteomics*, e2300236. https://doi.org/10.1002/pmic.202300236