ARTICLE

Mass Spectral Analysis of Synthetic Peptides: Implications in Proteomics

Medicharala Venkata Jagannadham, Pratap Gayatri, Taniya Mary Binny, Bathisaran Raman, Duvvuri Butchi Kameshwari, and Ramakrishnan Nagaraj**

CSIR-Centre for Cellular and Molecular Biology, Hyderabad 500007, India

Sequence determination of peptides is a crucial step in mass spectrometry–based proteomics. Peptide sequences are determined either by database search or by *de novo* sequencing using tandem mass spectrometry. Determination of all the theoretical expected peptide fragments and eliminating false discoveries remains a challenge in proteomics. Developing standards for evaluating the performance of mass spectrometers and algorithms used for identification of proteins is important for proteomics studies. The current study is focused on these aspects by using synthetic peptides. A total of 599 peptides were designed from *in silico* tryptic digest with 1 or 2 missed cleavages from 199 human proteins, and synthetic peptides corresponding to these sequences were obtained. The peptides were mixed together, and analysis was carried out using liquid chromatography–electrospray ionization tandem mass spectrometry on a Q-Exactive HF mass spectrometer. The peptides and proteins were identified with SEQUEST program. The analysis was carried out using the proteomics workflows. A total of 573 peptides representing 196 proteins could be identified, and a spectral library was created for these peptides. Analysis parameters such as "no enzyme selection" gave the maximum number of detected peptides as compared with trypsin in the selection. False discoveries could be identified. This study highlights the limitations of peptide detection and the need for developing powerful algorithms along with tools to evaluate mass spectrometers and algorithms. It also shows the limitations of peptide detection even with high-end mass spectrometers. The mass spectral data are available in ProteomeXchange with accession no. PXD017992.

Key Words: database search · false discovery · mass spectrometry · proteomics standards

INTRODUCTION

Detection and identification of proteins by proteomics involves the detection of tryptic peptides. The endeavor has been to optimize liquid chromatography (LC) conditions and design mass spectrometers to detect as many peptides as possible at low concentrations and thereby identify the proteins that they belong to. Detection of peptides depends on many factors, such as LC conditions, effective cleavage by trypsin, instrument capability, and power of the software used for analysis. The standards that are used to evaluate performance of LC systems and mass spectrometers are digests of total proteins. However, this does not ensure that all peptides that are obtained by enzymatic digestion will be detected in a sample. Use of synthetic peptides in a proteomics workflow would be a good alternative to examine LC and mass spectrometer performance because, theoretically, the exact number of peptides that should be detected would be known. Conditions can be optimized for detection of all peptides.

Identification of proteins is dependent on workflows practiced by different laboratories. The mass spectrometers and the algorithms obtained from different manufacturers with different technologies could also impact the detection of peptides. Therefore, for evaluating mass spectrometry (MS) instruments and algorithms, proper standards should be developed. Generally, short peptides are used as standards for calibrating MS instruments, and after satisfactorily passing the tests, MS data are acquired from the instruments. Several standards for proteomics, such as a 6-protein mixture, 48-protein mixtures with equimolar concentration, and dynamic concentration ranges, were developed. [1] Yeast proteome digests [2] have also been used. The 6-protein mixture after analysis provided a minimum of 15 proteins with confident identification. [3]

The Human Proteome Organization started in 2002 and has initiated the Proteomics Standards Initiative. These efforts have helped in generating guidelines for minimum information about proteomics experiments, data generation, and analysis. [4] These working groups generated and improved standards required for biomolecular interactions Protein Standards Initiative- Extended markup language

*Address correspondence to: Dr. M. V. Jagannadham, CSIR-Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500007, India (E-mail: medicharlavj@gmail.com). Dr. R. Nagaraj, CSIR-Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500007, India (E-mail: nraj@csirccmb.org).

This article includes supplemental data. Please visit *http://jbt.abrf.org/* to obtain this information.

https://doi.org/10.7171/jbt.2020-3201-001

(PSIMI-XML) and also created web sites for depositing data for public use, such as the ProteomeXchange consortium's web site and others. Generation of standard guidelines is a continuous process because of the improvements in the technologies and upgrading of the bioinformatics tools over time.

The Association for Biomolecular Research Facilities is also involved in generating experimental data from different laboratories for optimum utilization of the methods and standardizing the experimental protocols. Proteomics research groups have been developing various tools for improving proteomics technologies. [5–7] All these efforts are aimed at generating reproducible data from proteomics workflows and eliminating false discoveries in proteomics. In addition, several efforts are being carried out to identify the problematic areas in a large data analysis. A multilaboratory study has revealed complications in MS-based proteomics analysis. [8]

The protein standards currently used do not rely on the reproducibility of peptide sequences or number of peptides detected. There is more emphasis on identification of proteins. Protein identification may also be possible with a single unique peptide. In addition, in protein mixtures, the protein purity plays an important role. Even if the protein is 99% pure, the proteins present in the 1% contamination will also be detected with the powerful MS instruments that are currently available. In these cases, determination of false discoveries is a complicated issue. The protein digests from yeast, *Escherichia coli*, or human protein, and some of which that are used as standards, will have problems of reproducibility. A different peptide may be identified each time even though the proteins are the same.

Determining false discovery using different algorithms is an important issue for the identification of proteins. Evaluating the algorithms for identification of peptide sequences/protein identification and eliminating false discoveries is crucial for proteomics workflows.

In this paper, we describe mass spectral analysis of 599 synthetic tryptic peptides using a proteomics workflow. The sequences of these peptides were obtained from *in silico* digestion of human proteins. The peptides were synthesized and used in the study. We have observed that even with this relatively small set of peptides, it is not possible to detect all the peptides, even with multiple workflows. False discovery can be eliminated using such peptide standards, and conditions for peptide detection can be optimized.

MATERIALS AND METHODS

Synthetic peptides

We have designed 599 peptides from human proteins. The human proteins identified by different methods were selected. [9, 10] Trypsin digestion was carried out *in silico* with 1 or 2 missed cleavages, and peptides were selected from a mass range of 616–3169 Da. Few peptides that did not have Lys or Arg at the carboxy-terminus from some proteins were also included for synthesis. Peptides with Cys were omitted. Starting from a single peptide to 18 peptides per protein was selected from different proteins. Sequences of the synthetic peptides corresponding to the designed peptides (purchased from China peptides, Shanghai, China; more than 85% pure) are shown in Supplemental Table S1.

LC-MS/MS analysis

Each peptide was solubilized in the solvent in which it is freely soluble. The 599 peptides were mixed (500 fmol to 1 pmol) for analysis by liquid chromatography tandem mass spectrometry (LC-MS/MS) with different LC gradients. The peptides were separated on a PepMap RSLC C18 column (3 μm, 100 Å 75 μm × 15 cm) using 5% acetonitrile (ACN) in water containing 0.1% formic acid as solvent A and 95% ACN in water containing 0.1% formic acid in gradient runs. Q-Exactive HF Plus mass spectrometer from Thermo Fisher Scientific was used in the analysis of the peptides. Samples were directly injected for LC-MS/MS analysis using Thermo Scientific Easy–nLC 1200 equipment. Parameters used for acquiring mass spectra are shown in Supplemental Table S2. A total of 7 runs were made with different gradients on the LC, and peptides were injected. The mass range used for the precursor ion detection was m/z 200–2000. Data from all the runs were combined for analysis for identification of peptides and proteins. The MS proteomics data have been deposited to the ProteomeXchange consortium through the Proteomics Identifications Database PRIDE [11] repository with the data set identifier PXD017992.

Peptide/protein identification

The LC-MS/MS data were analyzed using SEQUEST provided by the manufacturer using Proteome Discoverer. Surprisingly, using trypsin as the enzyme for digestion did not yield maximum results. No enzyme was well suited for the analysis of these peptides. One missed cleavage, and oxidized methionine were set as variable modification. Precursor mass tolerance of 5 ppm and fragment ion mass tolerance of 0.2 Da were set for the identification of peptides/proteins. Human protein database from UNIPROT was used for the identification of peptides/proteins. Peptides with high confidence (1% False Discovery Rate) were selected for the identification. Medium confidence (5% False Discovery Rate) was also used to detect more peptides.

RESULTS

The synthetic peptides and the proteins they were a part of are summarized in Supplemental Table S1. For several proteins, more than 1 peptide was selected. All the synthetic peptides were more than 85% pure. Each peptide was made up to a concentration of 1 μg per microliter. A peptide mixture was obtained by mixing 2 μl of each peptide solution, resulting in a final volume of about 1.2 ml. For mass spectral analysis, 10–25 µl from this stock was used for injection. The conditions for runs and the mode of analysis are summarized in **Table 1** and Supplemental Table S2. The samples were run with 3 different gradients, run times, and volume of injection. The window for detection of MS was also varied.

A single run with a 60-min gradient, when analyzed with different MS and MS/MS parameters summarized in Supplemental Table S3, yielded 502 peptides. The list of 502 peptides along with those that were not detected is shown in Supplemental Table S3. In order to examine whether there is increased coverage, data from multiple runs were merged from different run conditions. The number of peptides detected went up to 535 (Table 2 and Supplemental Table S4). However, 64 peptides were not detected. Out of these, 10 peptides were assigned based on the detection of their C-terminal fragments, which were obtained as deletion peptides during synthesis. The intact peptides for these fragments were not detected. A mixture of 64 peptides that were not detected were added 10-fold more to the original 599 mix and were analyzed again. This resulted in the detection of 28 more peptides. By these methods, 573 peptides were detected, and 26 were not detected (**Table 2**). The results are shown in **Fig. 1**. Analysis of the raw data of only the 64-peptide mix manually indicated the presence of 13 peptides whose MS were identified (Supplemental Fig. S1). The doubly, triply, and multiply charged peaks were clearly identified. Further examination revealed that these peptides did not fragment to yield MS/MS spectra. The peptides that were not detected were analyzed separately by MS. These peptides could be identified when they were analyzed individually. The MS/MS of some representative peptides is shown in **Fig. 2**. The peptides detected by MS alone and the peptides not detected are shown in **Table 3**. They belong to a different mass range (800–2800 Da). Short or low-mass peptides with multiply charged species of peptides have not been detected. Isoelectric point or hydrophobic nature may have not played a major role in not detecting them. The peptides may have been missed in the selection of precursor ions.

TABLE 1

LC gradients used in different runs.

Gradient: Time [mm:ss]	Duration [mm:ss]	Flow [nl/min]	Mixture [%B]
00:00	00:00	300	3
35:00	35:00	300	25
45:00	10:00	300	40
53:00	08:00	300	80
55:00	02:00	300	80
60:00	05:00	300	3

Gradient: Time [mm:ss]	Duration [mm:ss]	Flow [nl/min]	Mixture [%B]
00:00	00:00	300	3
60:00	60:00	300	25
75:00	15:00	300	40
83:00	08:00	300	80
85:00	02:00	300	80
90:00	05:00	300	3

Gradient: Time [mm:ss]	Duration [mm:ss]	Flow [nl/min]	Mixture [%B]
00:00	00:00	300	3
80:00	80:00	300	25
100:00	20:00	300	40
108:00	08:00	300	80
114:00	06:00	300	80
120:00	06:00	300	3

Solvent A: 5% ACN in water with 0.1% formic acid, Solvent B: 95% ACN with 0.1% formic acid

TABLE 2

Detection of peptides using different conditions

Condition	No. of peptides detected	comment
single run 60min gradient	502	97 peptides not detected
Combining all runs (different gradients)	535	64 peptides not detected. 10 peptides assigned based on C-terminal fragments
10- fold more of 64 peptides added to 599	573	26 peptides not detected. 28 peptides detected after spiking.
Analysis of 26 peptides		Analysis of raw data showed MS of 13 peptides detected. 13 peptides not detected.

Because many proteins had multiple peptides in the library, they were detected even though several peptides were not detected. The data are presented in Supplemental Table S5. The synthetic peptides were ∼85% pure. Hence, a large

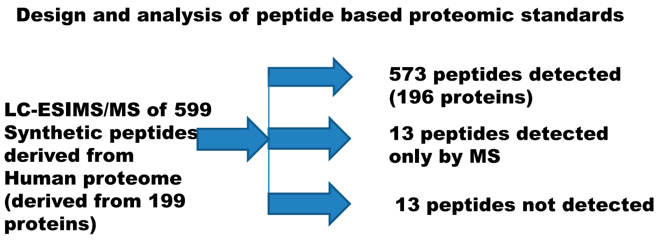
Design and analysis of peptide based proteomic standards
LC-ESIMS/MS of 599 Synthetic peptides derived from Human proteome (derived from 199 proteins)
573 peptides detected (196 proteins)
13 peptides detected only by MS
13 peptides not detected

FIGURE 1

Total number of peptides used in the analyses. The number of peptides detected and peptides not detected is shown. ESI is Electrospray Spray Ionization.

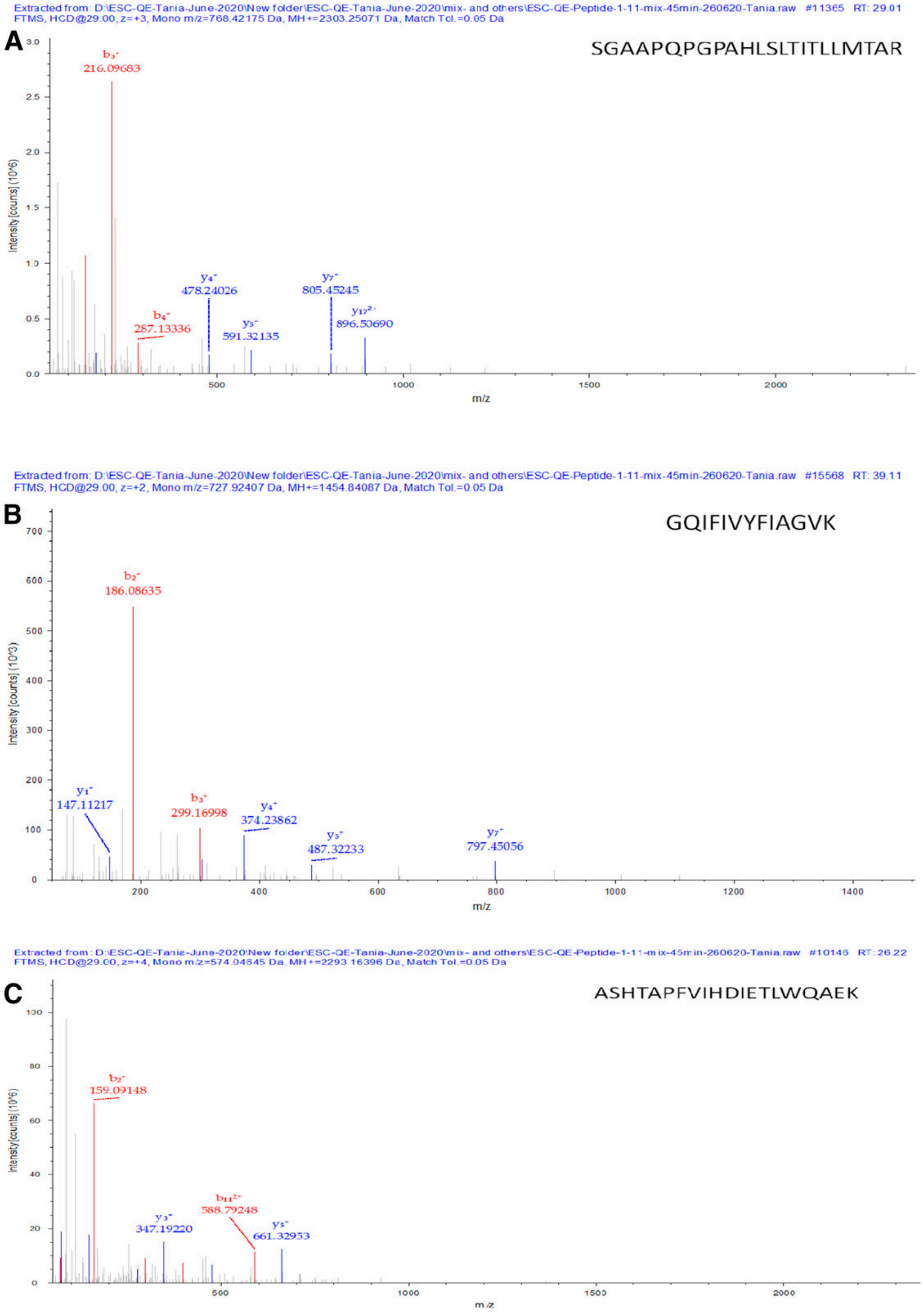

FIGURE 2

The MS/MS spectra of some representative peptides not detected in the mixture of peptides were analyzed individually, and the spectra are shown. The mass spectrum of a peptide from urokinase plasminogen receptor (*A*), vitamin K–dependent γ-carboxylase (*B*), and peroxisome proliferator–activated receptor-δ (*C*).

TABLE 3

Peptides not detected in MS/MS

Peptides detected only in MS spectra	Mol. Weight (Da)	*pI	Hydro phobicity	Peptides not detected at all	Mol. Weight (Da)	*pI	Hydro phobicity
APPTELLARPER	1349.75	6.19	-0.858	AAAAAEQQQFYLLLGNLLSPDNVVR	2701.43	4.37	0.244
DLDNAIEAVDEFAFLEGTLD	2197.02	3.22	0.1	ASHTAPFVIHDIETLWQAEK	2293.56	5.28	-0.19
DVLSLSGLSSDPADLDP	1700.82	3.32	0.024	DDQSIQK	833.4	4.21	-2.029
ESLDTAAVVQVGISR	1544.83	4.37	0.453	FHGGALPAYVVSNILLAYR	2061.13	8.6	0.816
LELTTYLFGQDP	1396.7	3.67	-0.083	GQIFIVYFIAGVK	1454.84	8.59	1.523
LQFLQLSQR	1132.64	9.75	-0.178	IFFGQWTLVQFNFLK	1888.02	8.75	0.74
LWGGTLLWT	1046.57	5.52	0.822	IILQAFSLSLVSSFLLIFLGK	2309.39	8.75	1.89
QFTLALGTTQDENG	1494.7	3.67	-0.586	LASLNQILDPWVYLLLR	2027.17	5.84	0.806
RPEEGEK (+3 detectd)	844.41	4.79	-2.986	MDLGVYQLR	1094.57	5.59	0.056
SQLVVTSPAPASEK	1413.75	5.72	-0.1	NLPTVSALR	970.568	9.75	0.278
THEEHHAAK	1059.49	6.19	-1.956	NWVDLAWAVSYYIR	1755.89	5.83	0.257
TYVISRTEPAMATTK	1668.86	8.26	-0.28	SGAAPQPGPAHLSLTITLLMTAR	2303.25	9.49	0.361
VTDATETTITISWR	1593.81	4.37	-0.121	WLYSLYDAETLMDR	1775.83	4.03	-0.35

Monoisotopic mass of the peptides are shown *pI was calculated using ExPASy tools.

number of N-terminal deletion fragments were detected. This facilitated the identification of proteins with confidence. Several methionine oxidized peptides were also detected along with their native peptides. The total number of peptides detected was 1223. Peptides numbering 72 that were N-, C-terminal deletions were detected. These are normally not obtained as a result of incomplete synthesis. Their detection could arise as a result of incomplete fragmentation.

The peptides that were not detected in the proteomics workflow did give the Electrospray Ionization Mass Spectrometry (ESI-MS) spectra, as indicated in the data sheet supplied by the manufacturer. The possible reasons these peptides were not detected could be that they were missed out in the top 15 selection for MS/MS due to low intensity. It is possible that after mixing all the peptides, some peptides could have precipitated in the final mixture of solvent. Several peptides (Supplemental Table S6) were detected that are not in any way related to the 599 peptides. They are false positives.

DISCUSSION

Because of the dynamic nature of the proteome in cells or tissues, reproducibility is context dependent in proteomics. But obtaining consistent data for the same sample in multiple runs is crucial for reliable identification of proteins. MS-based proteomics is basically dependent on sequence determination of peptides. The interpretation of MS/MS spectra to the sequence of peptides is crucial in the identification of proteins. Synthetic peptides will help in evaluating the performance of mass spectrometers and the algorithms used for the identification of peptide sequences because the number of peptides that are injected and their sequences are known. In the present study, we used 599 synthetic peptides whose sequences were derived from human proteins and analyzed them using proteomics workflows. Our analysis indicates that even in a mix of 599 peptides run on a Q-Exactive HF plus instrument, all the expected peptides are not detected. In the protein workflow, in which the number of peptides could be large, the chances of missed peptides are substantial. These peptides could be crucial, particularly for peptide biomarker discovery. It would be necessary to work out conditions in which the entire library of synthetic peptides would be detected. A standard comprising synthetic peptides would be more reliable as compared with a digest in which one has no idea of the number of peptides present. Also, generation of the same set of peptides from enzymatic digests in different batches may not be possible.

Using different analytical conditions (different gradients in the LC), a total of 573 peptides could be detected in the present study. Earlier reports [11] have shown that analysis of MS data with 2 search engines increases the robustness of MS data analysis, and 2 or 3 technical replicates can expand protein identification. We have also used PEAKS analysis to examine whether there is any improvement in identification. However, this did not have the desired result (data not shown). In addition, our study used 7 different analysis conditions to increase the separation ability of the peptides. This change in conditions assisted in detecting more peptides.

Even after using several search engines, the correctness of peptide identification depends on scoring functions. This study uses synthetic peptides, which bypasses the scoring limitations and helps to evaluate both the MS instrument and the algorithm in the detection of peptides. Our study, along with other studies using multiple search engines, [12, 13] highlights that the development of strong search engines is imperative for the analysis of MS data.

Even from 515 peptides (Supplemental Table S4), 196 out of 199 proteins in Table 1 could be identified. Out of these, a large number of N-terminal deletion peptides and peptides that were from the central portion of the peptides in the library were identified (*i.e.*, 708). Few studies are aimed at evaluating the algorithms to identify proteins using MS data. [14–16] It is important to overcome some of these limitations for the identification of the entire set of expected peptides in the proteomics workflows because synthetic peptides are increasingly used in proteomics. [17] Curated databases will provide better results for the identification of proteins. [18] Earlier studies have shown that by using different algorithms and combining the results, the confidence and sequence coverage could be improved. [19] In the present study, involving a mixture with a small number of peptides, a majority of the peptides could be detected, and reasons for not detecting the remaining peptides could be recognized.

Conclusions

This study shows the usefulness of synthetic tryptic peptides in evaluating the performance of an LC-based mass spectrometer and analysis software. Multiple LC gradients will help in improving the peptide detection and help in the identification of more proteins in the proteomics workflows. Moreover, the study of synthetic peptides helps in estimating false discoveries. The synthetic peptides and MS/MS spectra generated from the study can be used to improve the efficiency of algorithms for a more comprehensive coverage.

ACKNOWLEDGMENTS

Funding from the Council for Scientific and Industrial Research, India, in the network project BSC0111 is gratefully acknowledged. R.N. acknowledges National Academy of Sciences, (NASI) India, for NASI Platinum Jubilee Senior Scientist Fellowship. R.N. is the recipient of JC Bose Fellowship from the Department of Science and Technology, India.

REFERENCES

1. Ivanov AR, Colangelo CM, Dufresne CP, et al. Interlaboratory studies and initiatives developing standards for proteomics. *Proteomics*. 2013;13:904–909.
2. Jung S, Danziger SA, Panchaud A, von Haller P, Aitchison JD, Goodlett DR. Systematic analysis of yeast proteome reveals peptide detectability factors for mass spectrometry. *J Proteomics Bioinform*. 2015;8:231–239.
3. Kuchibhotla B, Kola SR, Medicherla JV, Cherukuvada SV, Dhople VM, Nalam MR. Combinatorial labeling method for improving peptide fragmentation in mass spectrometry. *J Am Soc Mass Spectrom*. 2017;28:1216–1226.
4. Deutsch EW, Orchard S, Alain Binz P, et al. Proteomics standards initiative: fifteen years of progress and future work. *J Proteome Res*. 2017;16:4288–4298.
5. Falick AM, Kowalak JA, Lane WS, et al. ABRF-PRG05: *de novo* peptide sequence determination. *J Biomol Tech*. 2008;19: 251–257.
6. Arnott DP, Gawinowicz M, Grant RA, et al. Proteomics in mixtures: study results of ABRF-PRG02. *J Biomol Tech*. 2002; 13:179–186.
7. Arnott D, Gawinowicz MA, Kowalak JA, et al. ABRF-PRG04: differentiation of protein isoforms. *J Biomol Tech*. 2007;18: 124–134.
8. Bell AW, Deutsch EW, Au CE, et al; HUPO Test Sample Working Group. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat Methods*. 2009;6:423–430.
9. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347: 1260419.
10. Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. *Nature*. 2014;509:575–581.
11. Paulo JA. Practical and efficient searching in proteomics: a cross engine comparison. *Webmedcentral*. 2013;4:WMCPLS0052.
12. Yuan ZF, Lin S, Molden RC, Garcia BA. Evaluation of proteomic search engines for the analysis of histone modifications. *J Proteome Res*. 2014;13:4470–4478.
13. Chen C, Hou J, Tanner JJ, Cheng J. Bioinformatics methods for mass spectrometry-based proteomics data analysis. *Int J Mol Sci*. 2020;21:2873.
14. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res*. 2019;47(D1):D442–D450.
15. Yates III JR, Park SK, Delahunty CM, et al. Toward objective evaluation of proteomic algorithms. *Nat Methods*. 2012;9:455–456.
16. Chamrad DC, Körting G, Stühler K, Meyer HE, Klose J, Blüggel M. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*. 2004;4:619–628.
17. Beveridge R, Stadlmann J, Penninger JM, Mechtler K. A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes. *Nat Commun*. 2020;11:742.
18. Ijaq J, Bethi N. Mass spectrometry based identification and characterization of human hypothetical proteins highlighting the inconsistency across the protein databases. *J Proteome Proteomics*. 2020;11:17–25.
19. Zhao P, Zhong J, Liu W, Zhao J, Zhang G. Protein-level integration strategy of multiengine MS spectra search results for higher confidence and sequence coverage. *J Proteome Res*. 2017; 16:4446–4454.