A comprehensive evaluation of consensus spectrum generation methods in proteomics

Xiyang Luo [1, &,] Wout Bittremieux [2, &], Johannes Griss [3,4], Eric W Deutsch [5], Timo Sachsenberg [6], Lev I. Levitsky [7], Mark V. Ivanov [7], Julia A. Bubis [7], Ralf Gabriels [8,9], Henry Webel [10], Aniel Sanchez [11], Mingze Bai [1], Lukas Kall [12,*] and Yasset Perez-Riverol [3,*]

[1] Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China.

[2] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92093, United States.

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.

[4] Department of Dermatology, Medical University of Vienna, 1090 Vienna, Austria.

[5] Institute for Systems Biology (ISB), Seattle, Washington 98109, USA.

[6] Applied Bioinformatics, Department for Computer Science, University of Tuebingen, Sand 14, 72076 Tuebingen, Germany.

[7] V.L. Talrose Institute for Energy Problems of Chemical Physics, N.N. Semenov Federal Research Center for Chemical Physics, Russian Academy of Sciences, Moscow, Russia

[8] VIB-UGent Center for Medical Biotechnology, Ghent, Belgium

[9] Department of Biomolecular Medicine, Ghent University, Belgium

[10] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

[11] Section for Clinical Chemistry, Department of Translational Medicine, Lund University, Skåne University Hospital Malmö, 205 02 Malmö, Sweden

[12] Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Royal Institute of Technology - KTH, Box 1031, 17121 Solna, Sweden.

[&] Xiyang Luo and Wout Bittremieux: These authors contributed equally to this work. * Corresponding authors Yasset Perez-Riverol (yperez@ebi.ac.uk) and Lukas Kall (lukas.kall@scilifelab.se).

Abstract

Spectrum clustering is a powerful strategy to minimize redundant mass spectral data by grouping highly similar mass spectra corresponding to repeatedly measured analytes. Based on spectrum similarity, near-identical spectra are grouped in clusters, after which each cluster can be represented by its so-called consensus spectrum for downstream processing. Although several algorithms for spectrum clustering have been adequately benchmarked and tested, the influence of the consensus spectrum generation step is rarely evaluated.

Here, we present an implementation and benchmark of common consensus spectrum algorithms, including spectrum averaging, spectrum binning, the most similar spectrum, and the best-identified spectrum. We have analyzed diverse public datasets using two different clustering algorithms (spectra-cluster and MaRaCluster) to evaluate how the consensus spectrum generation procedure influences downstream peptide identification. The BEST and BIN methods were found the most reliable methods for consensus spectrum generation, including for datasets with post-translational modifications (PTM) such as phosphorylation. All source code and data of the present study are freely available on GitHub at https://github.com/statisticalbiotechnology/representative-spectra-benchmark.

Introduction

Spectrum clustering, i.e. the process of grouping similar spectra in a larger collection of MS2 spectra into smaller subsets, has multiple applications in mass spectrometry in general and in proteomics in particular (1), including the generation of spectral libraries (2) and spectral archives (3), quality assessment of peptide identifications in public repositories (2) and improvement of quantification results (4, 5). Spectrum clustering algorithms strive to group highly similar spectra so that each cluster contains spectra generated from the same analyte (peptidoforms with a specific charge in the case of proteomics). Differences between tools for spectrum clustering vary in their implementation of the various data processing steps, including the pre-processing of spectra (e.g., intensity normalization and peak picking), the clustering algorithm used, the metric used for determining similarity between spectra, and the optional optimizations to increase computational efficiency. Current tools for spectrum clustering include MS-Cluster (3), spectra-cluster (2), MaRaCluster (6), msCRUSH (7), and falcon (8).

While the most apparent output of the process of spectrum clustering is a grouping of spectra into clusters, the majority of use cases benefit from a condensed single spectrum representation for each cluster. This is, for instance, useful for the data-driven creation of spectral libraries (9), for reannotation and visualization of clustering results in public data repositories (2), and label-free quantification (4). The generation of high-quality representative spectra for each cluster is a key aspect of spectrum clustering, as the resulting consensus spectra form the starting point for downstream analyses. Although several spectrum clustering algorithms have been adequately benchmarked [8], the impact of the consensus spectrum generation procedure has so far not been properly evaluated. Several common approaches can be used to generate representative spectra, including spectrum binning, spectrum averaging (3), and selecting the most similar spectrum to all cluster members (medoid) (8). Additionally, although this strategy can only be used for clusters that contain one or more identified spectra, the “best-identified spectrum” method uses the most confidently identified spectrum as cluster representative (10, 11)

Here, we have performed a comprehensive evaluation of algorithms for the generation of consensus spectra to assess their performance for downstream processing of spectrum clustering results. We have used the spectra-cluster and MaRaCluster tools to generate clusters from diverse publicly available datasets and explore whether consensus spectrum generation algorithms perform differently between different tools. Additionally, we have evaluated the impact of consensus spectrum generation on downstream peptide and protein identification performance. All code and analyses are open-source and available at https://github.com/statisticalbiotechnology/representative-spectra-benchmark under the permissive Apache 2.0 license.

Methods

Consensus spectrum generation algorithms and evaluation

For the benchmark, we implemented four consensus spectrum generation algorithms:

- Spectrum averaging (AVERAGE): The representative spectrum is an average of all the spectra in the cluster (9, 12, 13). In this algorithm, for every m/z value, the corresponding intensities on each spectrum in the cluster are averaged.
- Spectrum binning (BIN): In this method, for each cluster, a consensus spectrum vector with bin width 0.02 *m*/*z* was first constructed (13). For all spectra in the cluster, peak m/z and intensity values were assigned to the corresponding bin in the

consensus spectrum vector. Bins that contained values from fewer than 25% of the cluster members were discarded. Next, the vector was converted to a consensus spectrum by averaging all peak m/z and intensity values per bin (2).

- Most similar spectrum (MOST): For each cluster, the spectrum that is on average most similar to all cluster members was selected as representative (14, 15). This was determined by first calculating the dot product of all pairwise similarities between spectra in the cluster. Next, the spectrum with the maximal summed dot product to all other spectra was selected as the representative for that cluster.
- Best identified spectrum (BEST): For each cluster that contained at least one identified spectrum, the spectrum with the maximal peptide-spectrum match score was chosen as the representative for that cluster. Note that this approach is not valid if all spectra in the cluster are unmatched.

Data manipulation steps were implemented as reproducible Nextflow workflows (**Figure 1**). The spectra-cluster (version 1.1.2) (2) and MaRaCluster (version 1.0) (6) spectrum clustering tools were used to cluster the mass spectrum data, and the MS-GF+ sequence database search engine (version v2021.03.22) (16) was used to perform peptide identification. For each cluster, representative (consensus) spectra were directly generated from the clustering output using the first three consensus generation procedures described above. For the best-identified method, the spectra were additionally identified using MS-GF+, after which the PSMs with the maximum scores were selected as representatives for each cluster. To ensure a fair comparison between all consensus spectrum generation procedures, clusters that only contained unidentified spectra were ignored, as no valid representative spectrum could be obtained using the best-identified method. To evaluate downstream peptide identification performance, the consensus spectra obtained for both spectrum clustering tools with each consensus spectrum generation method were searched using MS-GF+ (16), after which the number of peptide identifications was compared between all combinations of clustering and consensus generation methods, and with the original data without clustering.

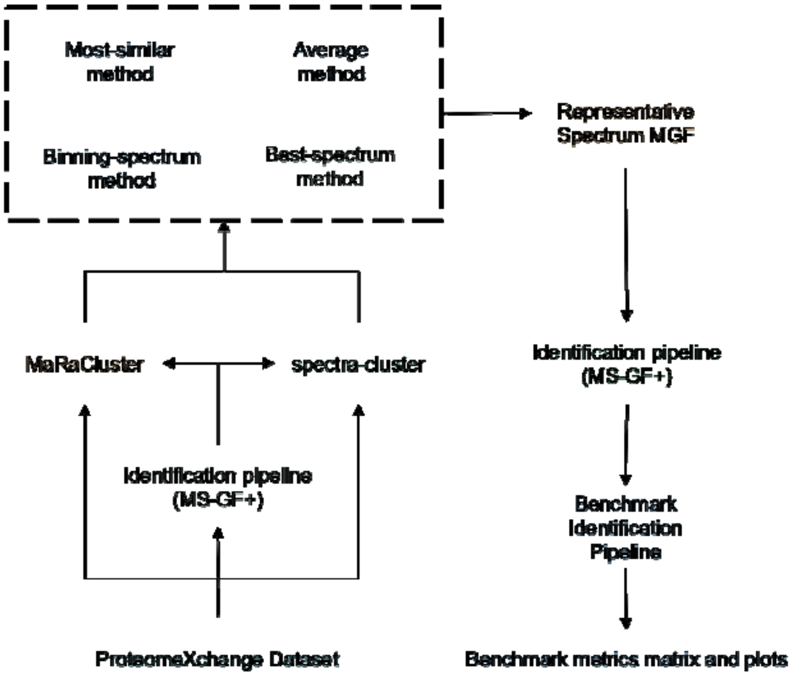
Most-similar method
Average method
Binning-spectrum method
Best-spectrum method
Representative Spectrum MGF
MaRaCluster
spectra-cluster
Identification pipeline (MS-GF+)
Identification pipeline (MS-GF+)
Benchmark Identification Pipeline
ProteomeXchange Dataset
Benchmark metrics matrix and plots

Figure 1: Study workflow, including clustering and peptide identification of publicly available ProteomeXchange datasets, consensus spectrum generation using alternative procedures, and evaluation of cluster representatives using an identification benchmark.

Benchmark Datasets

We used four public ProteomeXchange datasets: PXD008355, PXD023047, PXD021518, and PXD023361 (**Table 1**). RAW data from each dataset was converted to MGF using the ThermoRawFileParser (version: 1.2.3) tool (17) with default parameters. Among them, PXD008355, PXD023047, and PXD021518 are from Arabidopsis thaliana (mouse-ear cress), and PXD023361 is from Saccharomyces cerevisiae (baker's yeast). The datasets have been acquired using three different instrument models: Q Exactive, Q Exactive HF, and Q Exactive HF-X. The description of the samples, instrument configuration, sample processing steps, and analytical method can be read in the original publications: PXD008355 (18), PXD023047 (19), PXD021518 (20), and PXD023361 (21).

Table 1. Datasets were reanalysed to evaluate the performance of each consensus spectrum generation algorithm. The number of peptide identifications and peptide-spectrum matches can be found in the Supplementary Notes. In addition, the description of each dataset can be found in the original publication and PRIDE Archive (22).

Project accession	Instrument	No. MS/MS
PXD008355 (18)	Q Exactive	1,477,567
PXD023047 (19)	Q Exactive HF	109,333
PXD021518 (20)	Q Exactive HF-X	286,410
PXD023361 (21)	Q Exactive	38,286

For datasets PXD008355, PXD023047, and PXD021518, the Arabidopsis Thaliana protein database was downloaded from http://ftp.ebi.ac.uk/pride-archive/2019/07/PXD008355/TAIR10.fasta, while for dataset PXD023361 the Saccharomyces cerevisiae database was downloaded from http://ftp.pride.ebi.ac.uk/pride/data/archive/2021/04/PXD023361/uniprot-S_yeast.fasta.

For datasets PXD008355, PXD021518, and PXD023361, the precursor error tolerance was set to 10 ppm; while for dataset PXD023047, it was set to 20ppm. Target-decoy was performed using MS-GF+ (parameter -tda). For datasets PXD023047 and PXD021518 two modifications were allowed (NumMods=2): fixed carbamidomethyl cysteine modification, and variable methionine oxidation; while for datasets PXD008355 and PXD023361 Phosphorylation was also considered as variable modification.

Code availability

All code and analyses are freely available as open source under the Apache 2.0 license at https://github.com/statisticalbiotechnology/representative-spectra-benchmark. The consensus generation procedures were implemented in Python 3.6. Software dependencies that were used include Matplotlib (version 3.1.2) (23), Numba (version 0.47.0) (24), NumPy (version 1.17.3) (25), Pandas (version 0.25.3), pyOpenMS (version 2.4.0) (26), Pyteomics (version 4.1.2) (27), and spectrum_utils (version 0.3.3) (28).

Results

Figure 2 shows the number of PSMs (FDR=1%) identified with MS-GF+ (datasets PXD023047, PXD021528, PXD008355, and PXD023361) for spectrum clustering using MaRaCluster and spectra-cluster followed by consensus spectrum generation using the MOST, AVERAGE, BIN, and BEST procedures. Among the four public proteomics datasets, whether using spectrum clustering results from MaRaCluster or spectra-cluster, the identification rate for the MOST method is lower compared to the other methods, while the BIN and BEST methods achieve a higher spectrum identification rate (**Figure 2**).

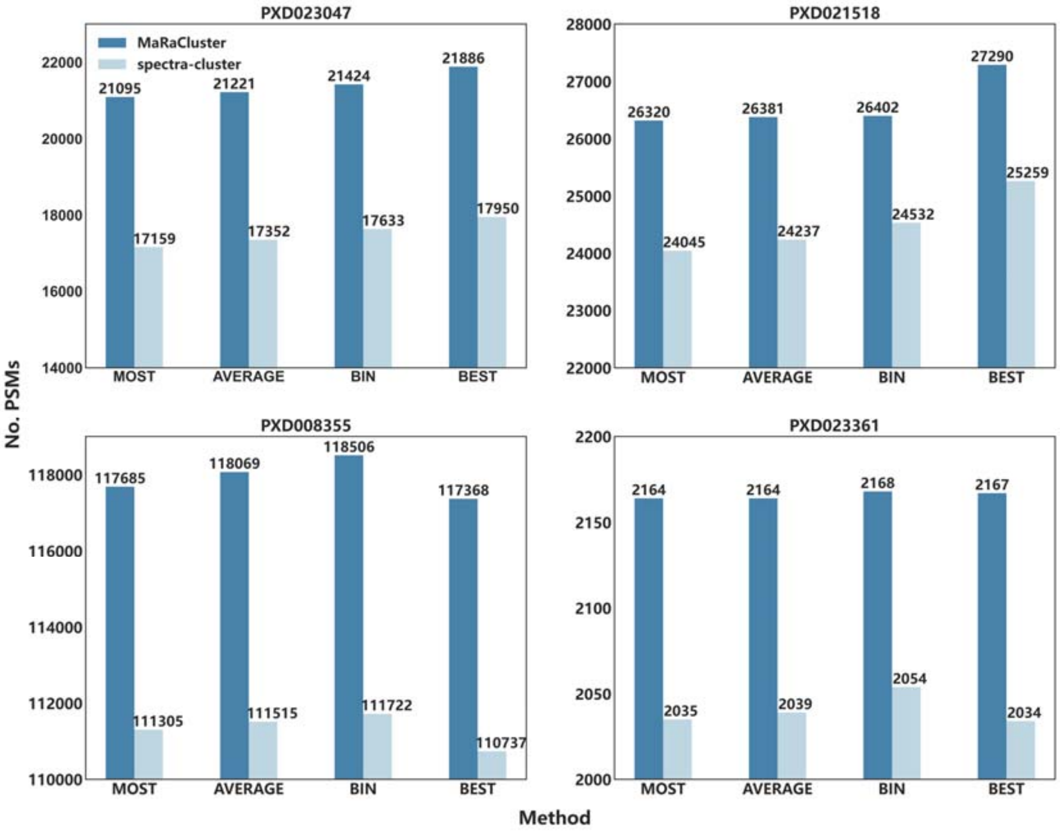

Figure 2: The number of PSMs obtained by MS-GF+ when searching consensus spectra produced by the MOST, AVERAGE, BIN, and BEST representative cluster generation methods for public proteomics datasets PXD023047, PXD021528, PXD008355, and PXD023361. Note that the bar plots are truncated past 0 to highlight relevant performance differences.

While the number of identified spectra only differs by a small amount between the various consensus spectrum generation procedures, when analyzing big public proteomics databases (billions of spectra) (29) these differences can be translated into millions of spectrum identifications. Among the methods that transform the original spectra, the BIN method is the one that performs best. The BIN method divides the m/z values into small bins and then overlaps multiple spectra within those bins. If there are multiple intensities in a bin, the algorithm will superimpose intensities in the same bin, favoring the most intensive peaks, which could improve peptide identification. However, in some cases, can also remove important peaks from the MS/MS spectra.

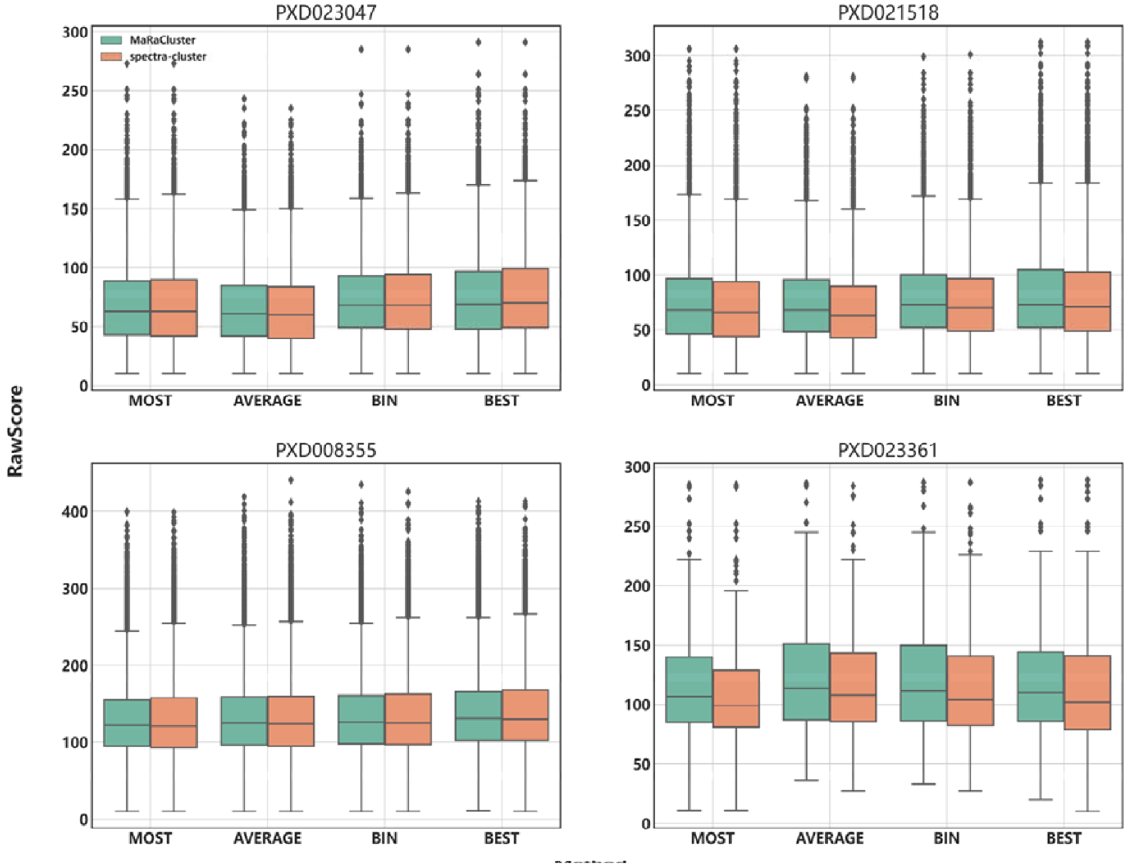

Figure 3: Distribution of MS-GF+ RawScores for MOST, AVERAGE, BIN, and BEST representative spectra from the public proteomics datasets PXD023047, PXD021528, PXD008355, and PXD023361.

Most of the consensus generation methods modify the original spectra, not only by removing or keeping some of the spectrum peaks, but also by modifying the corresponding intensity of each peak. We have used the distributions of the MS-GF+ RawScore to explore the relationship between the final spectra and the quality of the peptide identifications. Figure 3 shows the distribution of MS-GF+ RawScore for the four consensus generation methods (MOST, AVERAGE, BIN, and BEST) after clustering with MaRaCluster and spectra-cluster. For both clustering tools, the BIN and BEST method generate consensus spectra with higher average RawScore values (**Figure 3**), and similar to the previous metric (number of PSMs), the BEST algorithm achieves the highest average RawScore (**Supplementary Note 1**). The representative consensus spectra generated by the MOST method have the lowest average RawScore (**Figure 3**). The distribution of RawScore values (**Figure 3**) shows that the RawScores are more homogenous for the BIN method (lower standard deviation) than for all the other methods, including the BEST algorithm (**Supplementary Note 1**).

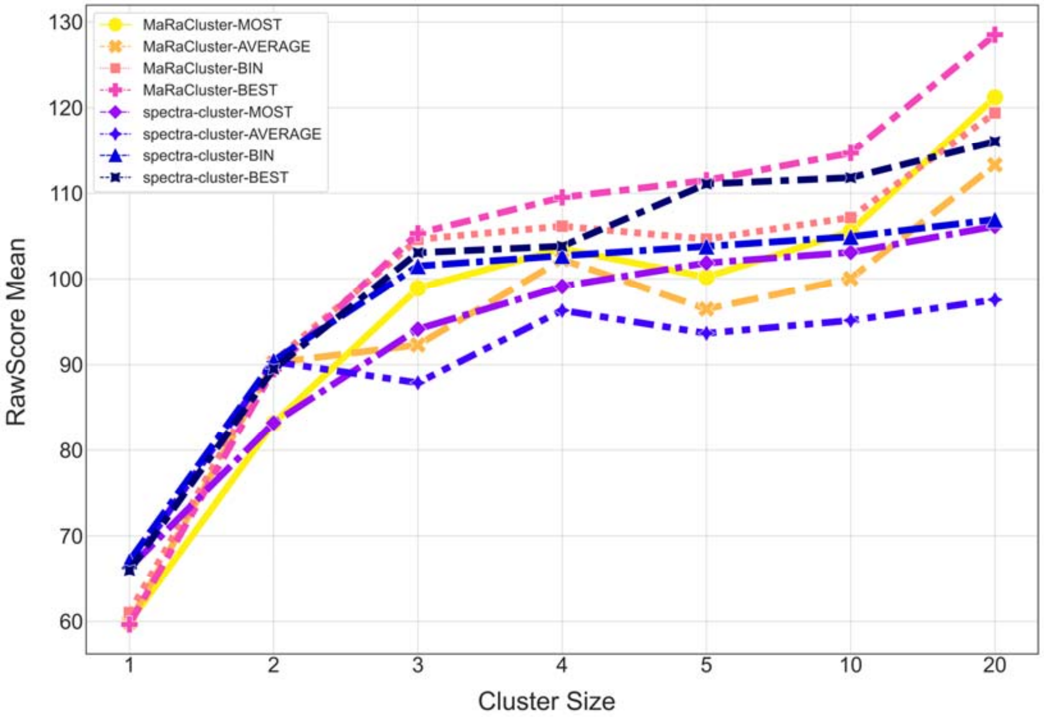

Figure 4: RawScore mean of the four different evaluated methods under different cluster sizes (1, 2, 3, 4, 5, 5-10, 10-20, 20 or higher).

Figure 4 shows the changes in mean RawScore of the identified spectra generated with the four evaluated methods (MOST, AVERAGE, BIN, and BEST) for clusters of different sizes (cluster sizes 1, 2, 3, 4, 5, 5-10, 10-20, 20 or higher). As expected, for clusters of one spectrum, no differences have been seen between different consensus methods, but minor differences are observed between clustering algorithms. For other small clusters containing three or fewer spectra, consensus spectra derived from the spectra-cluster results, in combinations with all the consensus spectrum generation methods, provide higher mean RawScores than consensus spectra derived from MaRaCluster results. In contrast, for larger clusters, MaRaCluster consensus spectra lead to higher mean RawScores. For both spectra-cluster and MaRaCluster, the mean RawScore increases with increasing cluster size. The BEST and BIN algorithms are stable for both clustering algorithms and all datasets (**Supplementary Note 1**), and the scores of these two algorithms are generally higher than MOST and AVERAGE. In combination with MaRaCluster, the AVERAGE algorithm shows instability and the score of the AVERAGE algorithm is generally lower than the other three algorithms.

In addition to peptide identification, we explored how using consensus spectra instead of the original spectra affects phospho-peptide identification and phosphorylation site localization.

We analyzed the number of phosphorylation sites identified in dataset PXD008355 after clustering with both tools (MaRaCluster and spectra-cluster) and the four different consensus spectrum generation methods (MOST, AVERAGE, BIN, and BEST). We have evaluated two metrics, (i) the number of phosphorylated PSMs identified and (ii) the phosphorylation sites identified.

Figure 5 shows the intersection of the phosphorylated PSMs among the four representative cluster methods after spectrum clustering with MaRaCluster and spectra-cluster. Most of the PSMs (91.2% for MaRaCluster and 96.4% for spectra-cluster) for the four representative cluster methods produce the same phosphorylated PSMs. The BIN method produces the largest number of unique PSMs, which is about double the number of other methods, followed by the BEST, MOST, and AVERAGE methods (**Figure 5**).

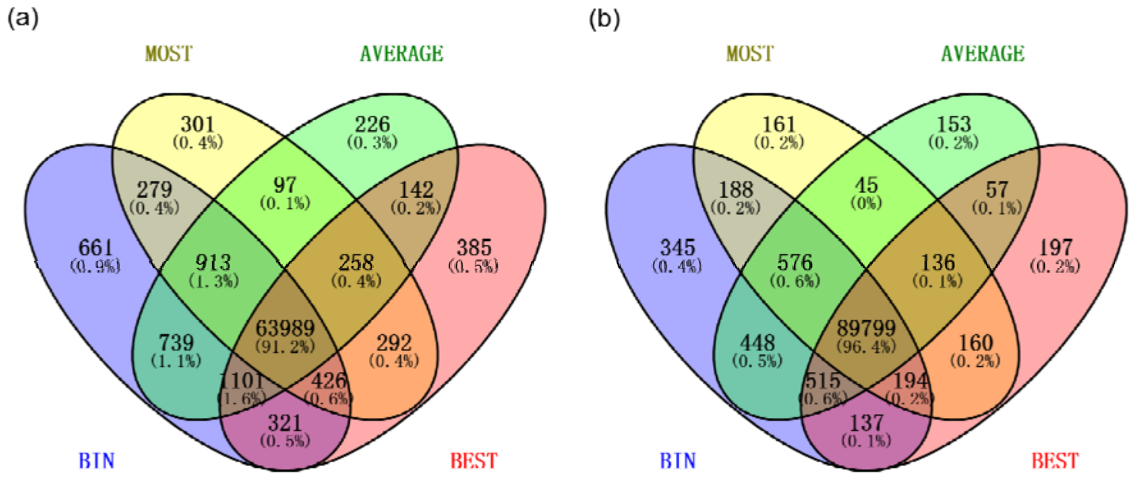

Figure 5: The intersection of the total phosphorylated PSMs among the four representative cluster methods in **(a)** MaRaCluster, and **(b)** spectra-cluster.

While the majority of phosphorylated PSMs are aggregated among all methods, around ~1% are different and we also observed differences in terms of phosphorylation sites. Table 2 shows the difference in phosphorylation sites between BIN and BEST representative spectra from MaRaCluster and spectra-cluster (extended table, **Supplementary Notes 3**). Because the BEST and BIN methods were the best performing consensus generation (13) options in terms of peptide identification, we focus the discussion on these two methods (extended table, Supplementary Notes 3). Most phosphorylated PSMs (63,165 for MaRaCluster and 89,161 for spectra-cluster) for the BEST and BIN methods are identical. However, a small number of phospho sites are different (2683 in MaRaCluster and 1494 in spectra-cluster), some of them due to different peptide identifications, some of them due to differences in the localization accuracy after clustering and the change of the spectra. These small differences can be attributed to the fact that the BIN method modifies the ion peak intensity and m/z of the spectrum through the binning algorithm.

Table 2: Analysis of phosphorylation sites identification of dataset PXD008355, after clustering with MaRaCluster and spectra-cluster, and generation of the consensus spectra using two different methods (BEST, BIN). We quantified the number of total phosphorylated PSMs and phosphorylation sites for each combination of clustering method and consensus generation method. In addition, we added the number of identical and different phospho-sites between the BEST and BIN methods for each clustering algorithm.

Cluster methods	Methods	No. phospho PSMs	No. phosphosites	No. PSMs with identical sites	No. PSMs with different sites
MaRaCluster	BEST	66914	81238	63165	2683
	BIN	68429	83091		
spectra-cluster	BEST	91195	109877	89161	1494
	BIN	92202	111230		

Conclusions

Representative spectra from clusters have typically been generated using four different algorithms: spectrum averaging, spectrum binning, the most similar spectrum, and the best-identified spectrum. Most tools and resources, including SpectraST (9), MassIVE (11, 30) spectral libraries, or spectra-cluster and PRIDE Cluster (2) use one of these methods. However, to our knowledge, no systematic analysis has been performed to compare multiple algorithms to generate consensus spectra. We implemented a Python framework to benchmark existing algorithms to generate representative spectra from clustering results from two different popular clustering tools—MaRaCluster and spectra-cluster.

The BEST and BIN methods were found to be the most reliable methods for consensus spectrum generation, including for datasets with post-translational modifications such as phosphorylation. The BEST method generates representative consensus spectra based on existing spectrum identification results, which requires that all clusters contain identified spectra. Therefore, the BEST method cannot be used on spectral archives (clusters of non-identified spectra) or if clustering is performed before the identification step. The BIN method is based on the original spectrum file and binning algorithm to generate representative consensus spectra and performed best in all benchmarks and comparisons after the BEST method. While the BIN algorithm modifies the original spectra, we do not observe major differences in identifying phosphorylated peptides and phosphorylation sites compared to the results of the BEST method to generate representative spectra. The fact that the BEST method is performing so well, compared to existing methods, suggests that better algorithms could be developed in the future to generate consensus spectra from clustering results.

Acknowledgment

The authors would like to acknowledge the EuBIC-MS community that organized the EuBIC-MS Developer Meeting in January 2020 (31), triggering the original discussions and implementations of this work.

References

1. Perez-Riverol, Y.; Vizcaino, J. A.; Griss, J., Future Prospects of Spectral Clustering Approaches in Proteomics. *Proteomics* **2018,** 18, (14), e1700454.
2. Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; Del-Toro, N.; Rurik, M.; Walzer, M. W.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcaino, J. A., Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods* **2016,** 13, (8), 651-656.
3. Frank, A. M.; Monroe, M. E.; Shah, A. R.; Carver, J. J.; Bandeira, N.; Moore, R. J.; Anderson, G. A.; Smith, R. D.; Pevzner, P. A., Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat Methods* **2011,** 8, (7), 587-91.
4. The, M.; Kall, L., Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. *Nat Commun* **2020,** 11, (1), 3234.
5. Griss, J.; Stanek, F.; Hudecz, O.; Durnberger, G.; Perez-Riverol, Y.; Vizcaino, J. A.; Mechtler, K., Spectral Clustering Improves Label-Free Quantification of Low-Abundant Proteins. *J Proteome Res* **2019,** 18, (4), 1477-1485.
6. The, M.; Kall, L., MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics. *J Proteome Res* **2016,** 15, (3), 713-20.
7. Wang, L.; Li, S.; Tang, H., msCRUSH: Fast Tandem Mass Spectral Clustering Using Locality Sensitive Hashing. *J Proteome Res* **2019,** 18, (1), 147-158.
8. Bittremieux, W.; Laukens, K.; Noble, W. S.; Dorrestein, P. C., Large-scale tandem mass spectrum clustering using fast nearest neighbor searching. *Rapid Commun Mass Spectrom* **2021**, e9153.
9. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007,** 7, (5), 655-67.
10. Griss, J.; Perez-Riverol, Y.; The, M.; Kall, L.; Vizcaino, J. A., Response to "Comparison and Evaluation of Clustering Algorithms for Tandem Mass Spectra". *J Proteome Res* **2018,** 17, (5), 1993-1996.

11. Wang, M.; Wang, J.; Carver, J.; Pullman, B. S.; Cha, S. W.; Bandeira, N., Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst* **2018,** 7, (4), 412-421 e5.

12. Tabb, D. L.; Thompson, M. R.; Khalsa-Moyers, G.; VerBerkmoes, N. C.; McDonald, W. H., MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J Am Soc Mass Spectrom* **2005,** 16, (8), 1250-61.

13. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R., Building consensus spectral libraries for peptide identification in proteomics. *Nat Methods* **2008,** 5, (10), 873-5.

14. Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., 3rd, Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* **2003,** 75, (10), 2470-7.

15. Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J., Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem* **2006,** 78, (16), 5678-84.

16. Kim, S.; Pevzner, P. A., MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **2014,** 5, 5277.

17. Hulstaert, N.; Shofstahl, J.; Sachsenberg, T.; Walzer, M.; Barsnes, H.; Martens, L.; Perez-Riverol, Y., ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J Proteome Res* **2020,** 19, (1), 537-542.

18. Van Leene, J.; Han, C.; Gadeyne, A.; Eeckhout, D.; Matthijs, C.; Cannoot, B.; De Winne, N.; Persiau, G.; Van De Slijke, E.; Van de Cotte, B.; Stes, E.; Van Bel, M.; Storme, V.; Impens, F.; Gevaert, K.; Vandepoele, K.; De Smet, I.; De Jaeger, G., Capturing the phosphorylation and protein interaction landscape of the plant TOR kinase. *Nat Plants* **2019,** 5, (3), 316-327.

19. Doner, N. M.; Seay, D.; Mehling, M.; Sun, S.; Gidda, S. K.; Schmitt, K.; Braus, G. H.; Ischebeck, T.; Chapman, K. D.; Dyer, J. M.; Mullen, R. T., Arabidopsis thaliana EARLY RESPONSIVE TO DEHYDRATION 7 Localizes to Lipid Droplets via Its Senescence Domain. *Front Plant Sci* **2021,** 12, 658961.

20. Pipitone, R.; Eicke, S.; Pfister, B.; Glauser, G.; Falconet, D.; Uwizeye, C.; Pralon, T.; Zeeman, S. C.; Kessler, F.; Demarsy, E., A multifaceted analysis reveals two distinct phases of chloroplast biogenesis during de-etiolation in Arabidopsis. *Elife* **2021,** 10.

21. Osman, S.; Mohammad, E.; Lidschreiber, M.; Stuetzer, A.; Bazso, F. L.; Maier, K. C.; Urlaub, H.; Cramer, P., The Cdk8 kinase module regulates interaction of the mediator complex with RNA polymerase II. *J Biol Chem* **2021,** 296, 100734.

22. Perez-Riverol, Y.; Bai, J.; Bandla, C.; Garcia-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaino, J. A., The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* **2021**.

23. Hunter, J. D., Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007,** 9, (3), 90-95.

24. Lam, S. K.; Pitrou, A.; Seibert, S., Numba: a LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, Association for Computing Machinery: Austin, Texas, 2015; p Article 7.

25. Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; Del Rio, J. F.; Wiebe, M.; Peterson, P.; Gerard-

Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E., Array programming with NumPy. *Nature* **2020,** 585, (7825), 357-362.

26. Rost, H. L.; Schmitt, U.; Aebersold, R.; Malmstrom, L., pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. *Proteomics* **2014,** 14, (1), 74-7.

27. Levitsky, L. I.; Klein, J. A.; Ivanov, M. V.; Gorshkov, M. V., Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J Proteome Res* **2019,** 18, (2), 709-714.

28. Bittremieux, W., spectrum_utils: A Python Package for Mass Spectrometry Data Processing and Visualization. *Anal Chem* **2020,** 92, (1), 659-661.

29. Deutsch, E. W.; Perez-Riverol, Y.; Carver, J.; Kawano, S.; Mendoza, L.; Van Den Bossche, T.; Gabriels, R.; Binz, P. A.; Pullman, B.; Sun, Z.; Shofstahl, J.; Bittremieux, W.; Mak, T. D.; Klein, J.; Zhu, Y.; Lam, H.; Vizcaino, J. A.; Bandeira, N., Universal Spectrum Identifier for mass spectra. *Nat Methods* **2021,** 18, (7), 768-770.

30. Choi, M.; Carver, J.; Chiva, C.; Tzouros, M.; Huang, T.; Tsai, T. H.; Pullman, B.; Bernhardt, O. M.; Huttenhain, R.; Teo, G. C.; Perez-Riverol, Y.; Muntel, J.; Muller, M.; Goetze, S.; Pavlou, M.; Verschueren, E.; Wollscheid, B.; Nesvizhskii, A. I.; Reiter, L.; Dunkley, T.; Sabido, E.; Bandeira, N.; Vitek, O., MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. *Nat Methods* **2020,** 17, (10), 981-984.

31. Ashwood, C.; Bittremieux, W.; Deutsch, E. W.; Doncheva, N. T.; Dorfer, V.; Gabriels, R.; Gorshkov, V.; Gupta, S.; Jones, A. R.; Käll, L.; Kopczynski, D.; Lane, L.; Lautenbacher, L.; Legeay, M.; Locard-Paulet, M.; Mesuere, B.; Perez-Riverol, Y.; Netz, E.; Pfeuffer, J.; Sachsenberg, T.; Salz, R.; Samaras, P.; Schiebenhoefer, H.; Schmidt, T.; Schwämmle, V.; Soggiu, A.; Uszkoreit, J.; Van Den Bossche, T.; Van Puyvelde, B.; Van Strien, J.; Verschaffelt, P.; Webel, H.; Willems, S., Proceedings of the EuBIC-MS 2020 Developers' Meeting. *EuPA Open Proteomics* **2020,** 24, 1-6.