Comparison and Combination of Search Engines to Discover and Characterize Identifications and PTM Signatures in Biology

Xiaoyue Jiang[1], Keith Waddell[1], David Horn[1], Bernard Delanghe[2], Michael Blank[1], Devin Drew[1], Rosa Viner[1], Andreas FR Huhmer[1]

[1]Thermo Fisher Scientific, San Jose, CA, USA; [2]Thermo Fisher Scientific, Bremen, Germany

Poster Note 64480

Overview

Purpose: Compare search engine platforms for peptide identification and complex PTM analysis.

Methods: Data for HeLa and histone lysates were run in Thermo Scientific™ Proteome Discoverer™ software and MaxQuant to evaluate the different search engines.

Results: Search engines perform differently for the same dataset for both, numbers of identifications and analysis times, especially for the histone samples. A combination of search engines is recommended to maximize number of identifications while keeping search time as minimal as possible.

Introduction

New advances in mass spectrometry enable comprehensive characterization and accurate quantitation of complete proteomes. However, complex biological questions can only be answered through sophisticated data processing using proteomics search engines. The list of identified peptides and proteins returned from such search engines ultimately determine the conclusions for the whole experiment and thus high confidence in the results is critical. The identification of biological post-translational modifications (PTMs) is even more challenging than standard protein database searches that simply look to identify and quantify proteins. As the number, types and combinatorial variations of PTMs expand, the analysis time significantly increases with a concomitant increase in incorrect assignments and missed identifications. This fact hinders the broader application of LC/MS/MS for disease studies related to PTM signatures. Several database search strategies have emerged for handling such complex PTM schemes, but the results have been difficult to confirm by simple comparison as the different software applications are not easily accessible. In this study, we established a workflow-based database searching method for a HeLa lysate and histone PTMs using the Proteome Discoverer software platform. Multiple search algorithms were compared and used in combination to provide an increase in the number of confidently identified peptides with PTMs.

Methods

Sample Preparation

Thermo Scientific™ Pierce™ HeLa Protein Digest Standard was used as the simple standard for peptide and protein identification. Histone samples prepared according to Ref. 1 was used as the mixture with complex PTM signatures.

Liquid Chromatography and Mass Spectrometry

The HeLa sample was analyzed on a Thermo Scientific™ Q Exactive™ Plus mass spectrometer coupled to a Thermo Scientific™ Easy-nLC™ 1000 chromatography with a 50 cm Thermo Scientific™ EASY-Spray™ Column. The histone sample was analyzed on a Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer coupled to the Easy-nLC™ 1000 chromatograph on the same 50cm column. The gradient for both samples was 120min.

Data Analysis

The data were analyzed using Proteome Discoverer and MaxQuant software. The search algorithms used in the study were Sequest HT, Mascot, Byonic and MS Amanda as part of the Proteome Discoverer software platform, and MaxQuant with the Andromeda search engine. For the HeLa searches, carbamidomethylation was set as a fixed modification while oxidation (M) and acetylation (N-terminus) were set as dynamic modifications. For the histone study, propionylation modifications were considered as fixed, while acetylation, methylation, dimethylation, trimethylation and phosphorylation were used as variable modifications. A FDR of 1% at the peptide level was used to filter the results. ptmRS was used to calculate the site localization probabilities of all the PTMs. All search engines were used the same FASTA database and all searches were performed using the same 2.9 GHz processing PC. The identifications generated from Proteome Discoverer software and MaxQuant were imported into Thermo Scientific™ ProteinCenter™ software for comparison.

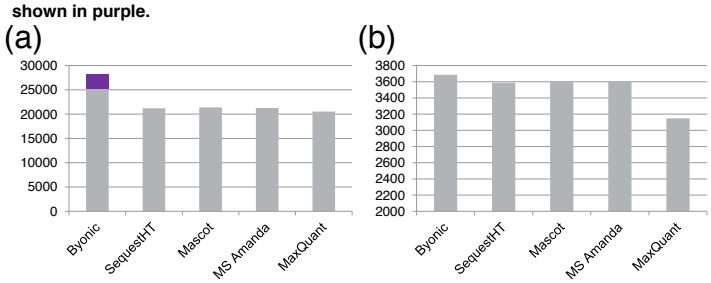
Thermo SCIENTIFIC

Results

Search engine performance comparison on HeLa digest study

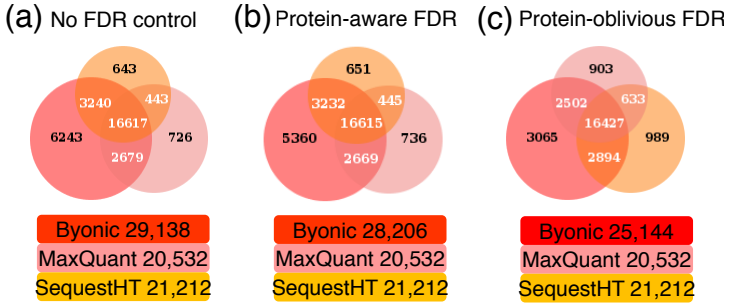
Different search engines will generate different numbers of peptides and proteins due to their different preprocessing and scoring algorithms. 1% peptide FDR was set for all search engines (1% PSM FDR for MaxQuant), and 2% protein FDR for Byonic and MaxQuant. The comparison of results are shown in Figure 1. SequestHT, Mascot, MS Amanda and MaxQuant generated similar numbers of peptide groups. Byonic has two types of peptide group FDR control, one using protein-oblivious FDR without considering the protein origin. The other is to use protein-aware FDR, also called 2D FDR, which gives bonus for PSMs from proteins almost sure to be true [2]. The protein-oblivious FDR generated about 4,000 more peptide groups compared to other search engines, and protein-aware FDR increased the number of IDs by another 3,000 additional peptides. For the protein group identifications, Byonic again outperformed all other searches engines (Figure 1b).

FIGURE 1. Numbers of (a) peptide groups and (b) protein groups identified by different search engines for 200 ng HeLa sample. The number of peptide groups by the protein-oblivious FDR control in Byonic is shown in grey. The number increment from protein-oblivious FDR to protein-aware FDR control is shown in purple.



In order to confirm if the extra identifications generated in Byonic were valid, we imported the peptide group identifications from Byonic, SequestHT and MaxQuant into Protein Center for a Venn Diagram comparison. If the same peptide was identified by more than one search engine, it was assumed that the confidence is high. When not using any peptide FDR error control (only 2% protein FDR), the Byonic search engine reported 29,138 peptides, with over 6,000 peptides unique to Byonic engine (Figure 2a). Applying the protein-aware FDR on peptides decreased the identification number to 28,206. The loss of ~1,000 peptides was mainly from the Byonic unique IDs, suggesting these identifications were of low confidence. However, if protein-oblivious FDR was applied, we observed loss of not only Byonic unique IDs, but also some of common identifications, indicating some high confident peptide identifications were lost in this step. Therefore, we conclude applying the protein-aware FDR is a good balance for identification sensitivity and accuracy.

FIGURE 2. Venn Diagrams to compare peptide identifications by Byonic, MaxQuant and SequestHT using different FDR settings for Byonic results.



The search time differs significantly between search engines as well (Table 1). SequestHT with multithreaded capability consumed the shortest time, only 21 minutes for a 2 hour gradient high resolution experiment (55,000 MS2 spectra). It took Byonic 34 minutes to finish the search, followed by MS Amanda at 73 minutes. Finally, Mascot was the slowest with over 2 hours for the search. Therefore for the analysis of the typical proteomic standard such HeLa digest, we recommend using SequestHT for quick sample overview or Byonic for a more comprehensive identification.

Table 1. Searching time for HeLa digest sample on different search engines.

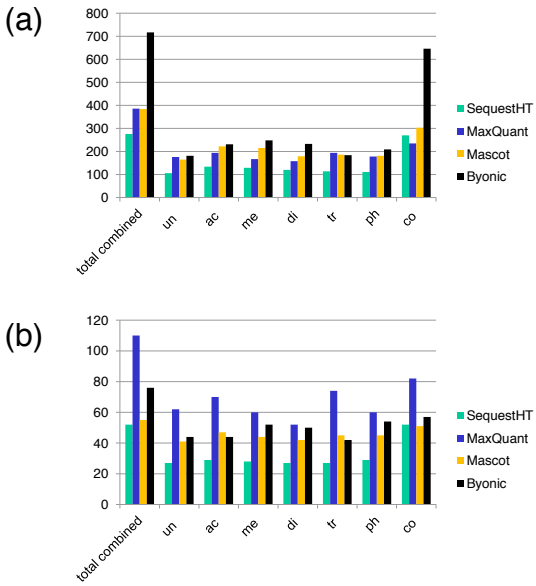
	Byonic	SequestHT	Mascot	MS Amanda	MaxQuant
Time (hour)	0.5	0.3	2	1.2	1.5

Search engine performance comparison on the histone sample

The PTM analysis of histones using multiple search engines are more challenging compared to HeLa sample. We performed the modification search based on [1]. In short, we set seven combinations of modifications: (1) propionyl-only for unmodified peptides (un); (2) Propionyl and acetyl for acetylated peptides (ac); (3) Propionyl and methyl propionyl for monomethylated peptides (me); (4) Propionyl and dimethyl for dimethylated peptides (di); (5) Propionyl and trimethyl for trimethylated peptides (tr); (6) Propionyl and phospho for phosphorylated peptides (ph); and (7) All of the above modifications for multi modified peptides (co). All seven sets of modifications were run using each search engine separately and the results were combined. The identifications of peptide groups and protein groups on each modification are shown in Figure 3.

We found that similar to the HeLa study, Byonic identified the highest number of modified peptides for histones, especially for the methylation and dimethylation searches. Mascot produced the second most IDs. MaxQuant was comparable to Mascot but provided a higher number of unmodified and trimethylated forms. SequestHT identified the fewest number of modified peptides and proteins. One interesting observation is that MaxQuant reported highest number of protein groups, even though it did not identify the most peptides. MaxQuant uses a different method for protein grouping than Proteome Discoverer software and this led to the large difference in the number of protein groups.

FIGURE3. Number of modified (a) peptide and (b) protein identifications by different search engines.



We compared the search results for the acetylated histone forms identified by Byonic, Mascot and SequestHT in ProteinCenter (Figure 4). Unlike the HeLa study which had 80-90% overlapped identifications among the three search engines, the identification differences for histone sample were more pronounced. For example, both Byonic and Mascot have ~220 acetylated peptides identified, but they only overlap ~60%, with more than 80 unique peptide sequences for each engine.

It was found that some sequences uniquely identified by Byonic were high quality matched spectra, with an example shown in Figure 5a. Mascot identified the same spectrum as a different sequence with quite a few mismatched peaks (Figure 5b) and SequestHT did not produce a match to this spectrum. Similarly, there were also high quality PSMs identified only by Mascot (Figure 5c) that were missed by other two engines. Therefore, we conclude that the PTM searching capability from current engines still has room to improve. We recommend using several search engines in combination to improve the coverage for each modification, a unique capability within Proteome Discoverer 2.0. The workflow is shown in Figure 6 and results will be equivalent to what's shown in Figure 4.

FIGURE4. Comparison of acetylated histone (a) protein groups and (b) peptides by Byonic, Mascot and SequestHT.

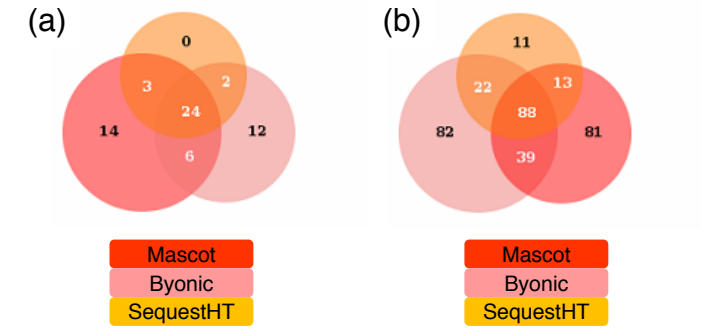


FIGURE 5. (a) An example of confident peptide identifications unique to Byonic (b) The same spectrum was identified as a different peptide with lower confidence by Mascot. (c) An example of peptide uniquely identified by Mascot

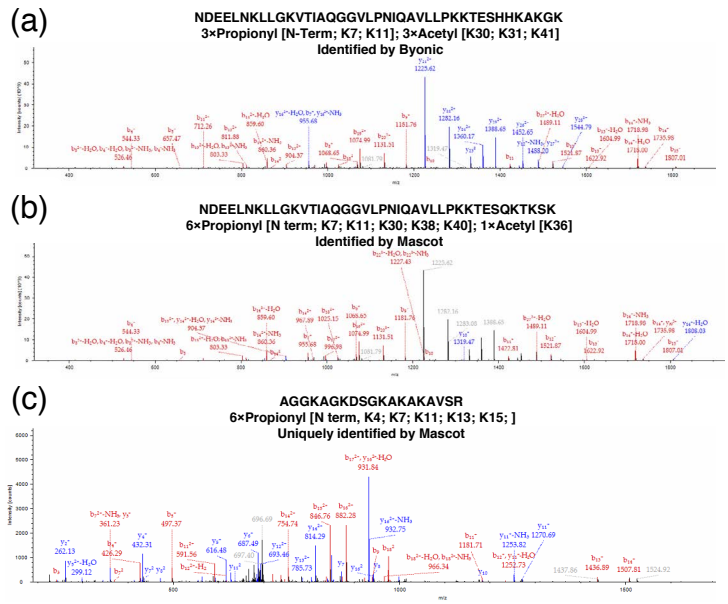


FIGURE 6. Proteome Discoverer workflow combining Mascot, Byonic and SequestHT for PTM analysis.

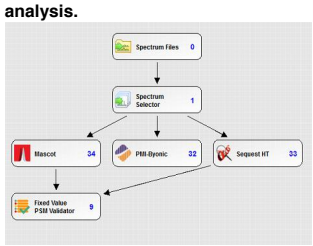
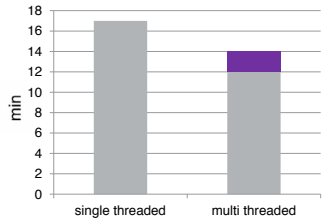


FIGURE 7. Comparison of search times of single vs. multi threaded searches available in PD 2.0. The timing for final consensus workflow is shown in purple.



The searching time for acetylated modified peptides and proteins was 17 minutes using the above workflow (Figure 6). The new Proteome Discoverer 2.0 has the capability to perform multi-threaded searches in parallel, and then generate a consensus report on all three processing results. The total time for paralleled searches will be equivalent to the longest single search, saving the extra time spent on other two engines in the traditional workflow. The time consumed for acetylated peptides with the new workflow was reduced to 14 minutes in total (Figure 7).

While the search for single modified form (Figure 3) could be achieved within reasonable time, the search for multiple PTMs was extremely slow in both Mascot and Byonic (Table 2). This suggested another bottleneck for the current search engines, which could be partially compensated by the parallel searching.

Table 2. Total searching time for 7 sets of modifications for Histone study using different search engines.

	Byonic	SequestHT	Mascot	Maxquant
Time (hour)	44	0.5	21	3

Conclusion

- Sequest HT is the fastest search engine on simple data set like HeLa digest and provides results of decent quality.
- Byonic is superior for peptide and protein identifications in part due to the 2D FDR capability.
- For heavily modified sample like histones, it is recommended to use a combination of different search engines in Proteome Discoverer for better coverage.
- New parallel searching capability in Proteome Discoverer enables a faster search compared to traditional search.

References

1. Yuan Zuo-Fei, Lin Shu, Molden Rosalynn, Garcia Benjamin, 2014, *J Proteome Res.*, **13(10):** 4470
2. Bern Marshall, Kil Yong, 2012, *J Proteome Res.*, **10(12):** 5296

www.thermoscientific.com

Thermo Fisher Scientific, San Jose, CA USA is ISO 13485 Certified.

Africa +43 1 333 50 34 0
Australia +61 3 9757 4300
Austria +43 810 282 206
Belgium +32 53 73 42 41
Canada +1 800 530 8447
China 800 810 5118 (free call domestic) 400 650 5118
Denmark +45 70 23 62 60
Europe-Other +43 1 333 50 34 0
Finland +358 10 3292 200
France +33 1 60 92 48 00
Germany +49 6103 408 1014
India +91 22 6742 9494
Italy +39 02 950 591
Japan +81 45 453 9100
Korea +82 2 3420 8600
Latin America +1 561 688 8700
Middle East +43 1 333 50 34 0
Netherlands +31 76 579 55 55
New Zealand +64 9 980 6700
Norway +46 8 556 468 00
Russia/CIS +43 1 333 50 34 0
Singapore +65 6289 1190
Spain +34 914 845 965
Sweden +46 8 556 468 00
Switzerland +41 61 716 77 00
UK +44 1442 233555
USA +1 800 532 4752

PN64480-EN 0615S

Thermo SCIENTIFIC
A Thermo Fisher Scientific Brand