Breaking the 2000-Protein Barrier in a Standard LC Run Using a New Orbitrap Benchtop Instrument and Multiple Search Engines

P1161

Carmen Paschke, Markus Kellmann, Yue Xuan, Eugen Damoc, Torsten Ueckert, Hans Grensemann, Ute Comberg, Bernard Delanghe
Thermo Fisher Scientific, Bremen, Germany

Overview

Purpose: Maximizing the number of true positive peptide spectrum matches.

Methods: Using multiple search engines and Percolator.

Results: More than 2600 protein groups have been identified in a single raw file.

Introduction

Getting a comprehensive view of the protein content of a complex sample is still a challenge in proteomics; accomplishing it in one analytical run even more so. A lot of efforts have been made the last few years to increase the speed and sensitivity of instruments . Here we describe a combination of a new instrument and data analysis using multiple search engines and proper statistical validation to maximize the number of true positive peptide-spectrum matches (PSMs), distinct peptides and protein groups.

Methods

Sample Analysis

A tryptic digest of a whole SILAC-labeled HeLa cell lysate was measured on a Thermo Scientific Q Exactive mass spectrometer—a new benchtop hybrid quadrupole-Orbitrap™ LC-MS/MS—coupled to a Thermo Scientific EASY-nLC liquid chromatograph. A top 10 HCD data-dependent tandem MS method was used for the experiments. The sample was run in duplicate.

Data Analysis

The data was analyzed using a pre-release version of Thermo Scientific Proteome Discoverer software version 1.3. Multiple search engines (SE) were used for identification: SEQUEST®, Mascot™ and X!Tandem. Validation based on separate target and decoy searches and subsequent calculation of classical score-based false discovery rates (classical FDR) or Percolator[1] determined q-values were used for assessing the statistical significance of the identifications.

The Percolator algorithm (available at http://per-colator.com) uses a set of features related to the quality of the peptide-spectrum matches (PSMs) to classify between correct and incorrect matches. In the Proteome Discoverer™ software, a set of more than 30 features (e.g. SE scores, mass deviation, etc.) which are descriptors of the match quality, are derived from PSMs and used for training the classifier. In many cases the use of Percolator increases the number of PSMs at a given FDR.

The workflow used is shown in Figure 1. All PSMs were filtered to 1% FDR. The filtered PSMs were grouped into distinct peptides from which protein groups were inferred in compliance with parsimony principles.

The Proteome Discoverer software default settings were used for quantification of the SILAC pairs.

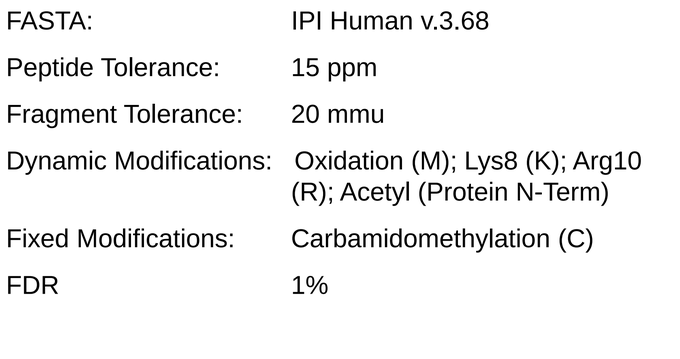
Table 1. Search conditions for all search engines, classical FDR and Percolator

Operating system:	Windows® 7, 64 bit
FASTA:	IPI Human v.3.68
Peptide Tolerance:	15 ppm
Fragment Tolerance:	20 mmu
Dynamic Modifications:	Oxidation (M); Lys8 (K); Arg10 (R); Acetyl (Protein N-Term)
Fixed Modifications:	Carbamidomethylation (C)
FDR	1%

Results

Several workflows for processing raw files have been used to access the optimum strategy. The number of distinct peptides and protein groups using one search engine (SEQUEST), two search engines (SEQUEST, Mascot), and three search engines (SEQUEST, Mascot and X!Tandem) was determined after validation with classical target/decoy based FDR and by using Percolator. The results for the duplicates are summarized in Figure 2. The average percentage increase compared to the SEQUEST search with classical target/decoy based FDR is displayed in Figure 3.

FIGURE 1. Overview of the workflows used. When using classical FDR instead of Percolator, the Percolator node was replaced by the Peptide Validator node.



Using SEQUEST Search Engine and Percolator

Number of distinct peptides increases by 65% to more than 11250 peptides on average for the two replicates.

Number of protein groups increases by more than 30% to more than 2500.

Using Multiple Search Engines and Percolator

Two search algorithms increase the number of peptides further to more than 11750 distinct peptides and to 2580 protein groups.

Employing three search engines the number of distinct peptides is still increasing to 12500. However the gain in protein groups is rather marginal, +2 % to more than 2600. As a result of the combination of search engines the distinct peptides have increased more then 80%, the protein groups more than 35%.

Merging Multiple Runs

Merging replicates distinct peptide identifications increase more than 100% and protein groups more than 50% compared to a single file with classical FDR.

FIGURE 2. Comparison of the number of distinct peptides and protein groups identified by the different data analysis strategies.

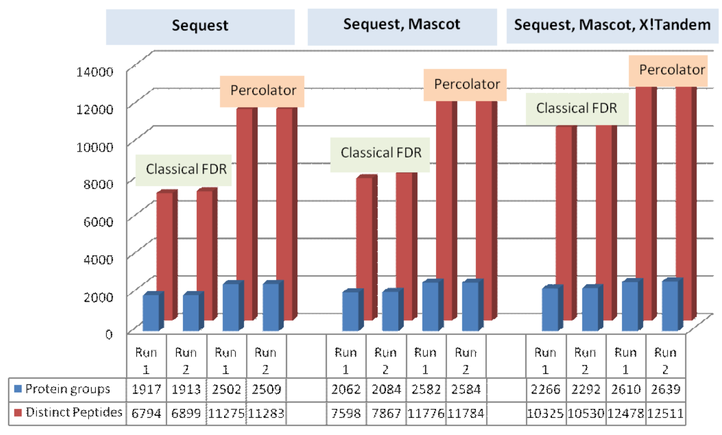


FIGURE 3. The average relative increase of the number of distinct peptides and protein groups obtained for different data analysis strategies.

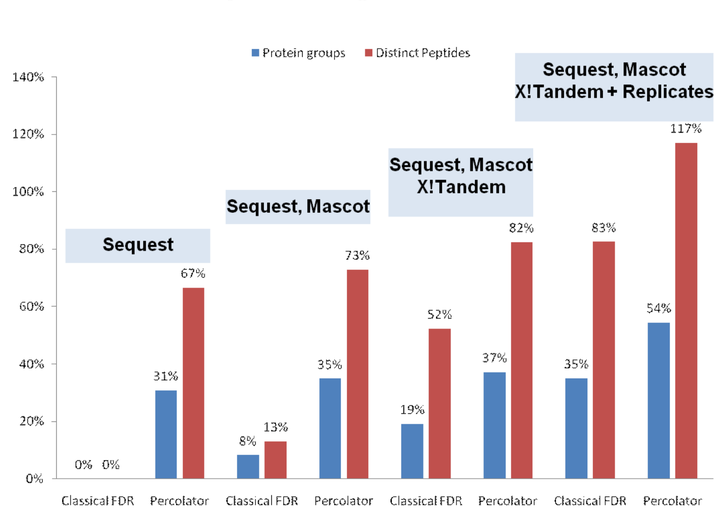
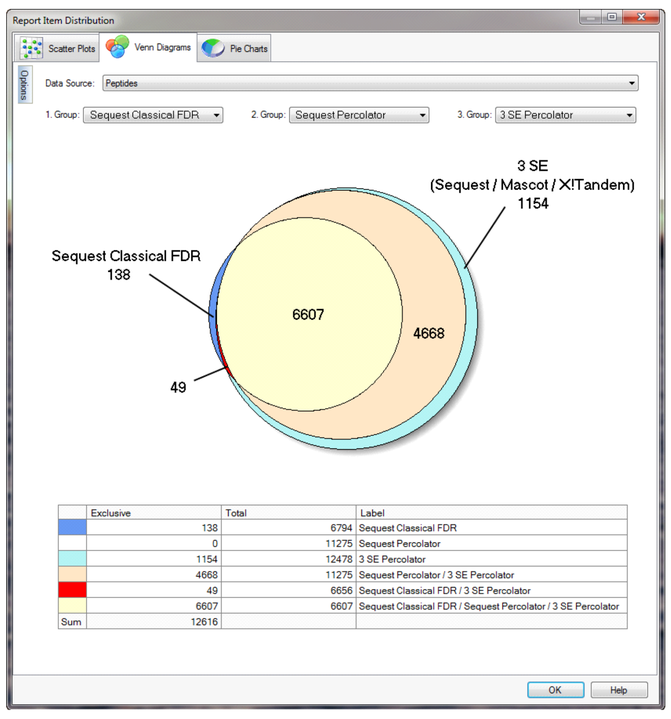


Figure 4. shows a Venn diagram comparing distinct peptide identifications from SEQUEST classical-FDR, SEQUEST Percolator, and 3 search engines with Percolator data analysis approaches. The number of exclusive, distinct peptides for SEQUEST using a classical FDR is small: 138 distinct peptides and 41 protein groups (data not shown).

FIGURE 4. Venn diagram comparison of SEQUEST result with classical score-based FDR, a SEQUEST and Percolator combination and a combination of three search engines with Percolator.



Recovering PSMs With Percolator

The SEQUEST Percolator combination identified a large number of additional distinct peptides (> 4650). Figure 5 shows an example of a PSM with XCorr of 1.46, below the cut-off threshold for the classical score-based FDR calculation (XCorr = 2.465 for 2+ peptides), but still classified as high confident with Percolator (FDR < 1%). Nevertheless the complete series of y-ion fragments is matched (Figure 5B).

Reasons for the low SEQUEST score:

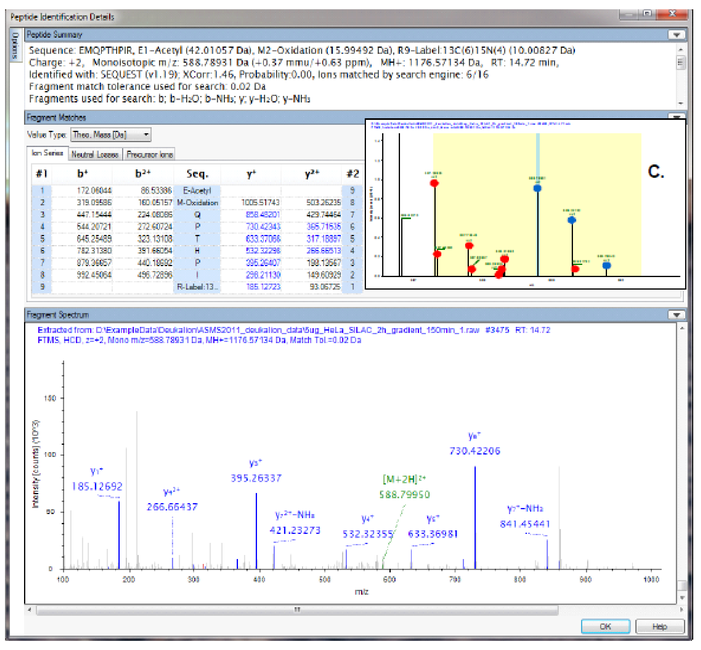
- Large number of fragments not explained by the proposed peptide sequence.
- More peptides are isolated and fragmented in that spectrum (Figure 5C), isolation interference of 56% (Figure 5A).
- Low peptide abundance: the area is low and the maximum inject time has been reached (Figure 5A).

FIGURE 5. (A) Details of a PSM with low XCorr and still classified confident with Percolator. (B) Annotated Spectrum. (C) Insert showing the precursor isolation window (yellow); peaks from the peptide precursor (blue) and interfering ions (red) are marked.

A.

Sequence	q-Value	PEP	XCorr	Isolation Interference [%]	Ion Inject Time [ms]	Area
emQPTHPIr	0.00023	0.00436	1.46	56	100	1.477e7

B.



Quantification

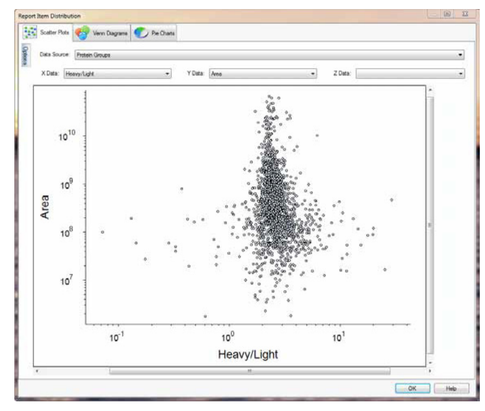
The same increase in identification is also observed for protein groups in quantification experiments

- Using three search engines and Percolator more than 2000 protein groups and more then 10000 PSMs are identified and quantified.
- Merging replicates more than 2300 protein groups and more then 10000 PSMs are identified and quantified.

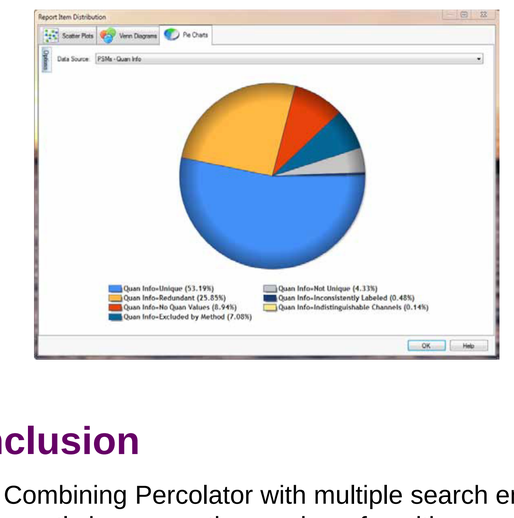
The distribution of the ratios of the protein groups is displayed in Figure 6A. Figure 6B shows the percentage of quantified PSMs.

FIGURE 6. (A) Volcano plot of the area vs. the light/heavy ratio of the protein groups. (B) Percentages of quantified PSMs.

A.



B.



Conclusion

- Combining Percolator with multiple search engines greatly increases the number of positive identifications, in the example described here to more than 35% for protein groups and to more than 80% for distinct peptides.
- Using a combination of 3 search engines with Percolator scoring the workflow identified more than 2600 protein groups in a single LC run. A large proportion of those proteins identified, over 2000 protein groups, were also quantified.

References

1. Lukas Käll et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods,* **2007** *4.11: 923-925.*

Acknowledgements

We would like to thank Annette Michalski for supplying the samples.

Thermo SCIENTIFIC