Optimizing data processing parameters for samples acquired on instruments using Advanced Peak Determination

Bernard Delanghe[1]; Tabiwang N. Array[1]; Aaron Gajahar[2]; David Horn[2]; Pedro Navarro[1]; Eugen N. Damoc[1]; Andreas Huhmer[2]
[1]Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany; [2]Thermo Fisher Scientific, San Jose, CA

ABSTRACT

Purpose: Improving identification and quantification rates for instruments using Advanced Peak Detection (APD).

Methods: Optimizing the data processing workflow.

Results: Short synopsis of the results.

INTRODUCTION

Recently, the Advanced Peak Determination (APD) algorithm was introduced on two Thermo Fisher Scientific instruments: the Thermo Scientific™ Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer and the Thermo Scientific™ Q Exactive™ HF-X Hybrid Quadrupole-Orbitrap™ Mass Spectrometer. This algorithm dramatically improves the missing or erroneous peak assignments, due to overlapping isotopic envelopes resulting in more features available for tandem mass spectrometry (MS^2) selection. Using APD for Label Free quantification (LFQ) experiments results in almost complete feature detection. However, when those features are fragmented this produces mixed or chimeric spectra of two or more peptides. Here we will investigate different data analysis parameters and algorithms to maximize the number of identified and quantified peptides.

MATERIALS AND METHODS

Sample Preparation

The Thermo Scientific™ Pierce™ HeLa Protein Digest Standard was used.

MS Measurements

The samples were measured using an Evosep One coupled to an Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer.

The Evosep One was operated using the “100 samples per day” method, which is a 11.5 min. HPLC run with a 5 min. gradient. A total of 50 ng of the Hela Protein digest was loaded on the disposable trap column tip. 2 Replicates were measured.

The Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer was operated at 45.000 resolution in Full MS, AGC target was set to 3e6, maximum inject time (max IT) to 20 ms. MS2 set at 7500 resolution, AGC target at 5e4 and max IT at 20 ms. Isolation window was set at 1.4 m/z. Dynamic exclusion was set to 10 s.

Data Analysis

A beta version of Thermo Scientific™ Proteome Discoverer™ software version 2.3 was used.

The processing workflow is displayed in Figure 1.

The precursor masses were mass recalibrated using the Spectrum Files RC node.

Two cascading Sequest HT nodes were used. In the first Sequest HT node search using following modifications:
Dynamic Modifications: Oxidation / +15.995 Da (M); Carbamyl / +43.006 Da (K, R)
Dynamic Modifications (peptide terminus): Carbamyl / +43.006 Da (N-Terminus); Gln->pyro-Glu / -17.027 Da (Q); Glu->pyro-Glu / -18.011 Da (E)
Dynamic Modifications (protein terminus): Acetyl / +42.01 Da
Static Modifications: Carbamidomethyl / +57.021 Da (C)
Protein Database: Homo sapiens (SwissProt TaxID=9606) (v2017-07-05)

All medium and low confident peptides were resubmitted for a semi-tryptic Sequest HT search.

In parallel to Sequest a spectral library was search using MSPepSearch. The library was created from the data of the first ProteomeTools release.[1] The library used only includes the HCD spectra of about 200.000 unique synthetic peptides.
The reverse database was constructed by reversing the library’s peptide sequences except for the C-terminal residue.[2]

All spectra were searches with precursor tolerance of 10 ppm and fragment tolerance of 20 mmu.
Percolator was used for calculating q-values and PEPs.
Feature detection was performed using the Minora node.

The consensus workflow is displayed in Figure 2.
The peptides were filtered for 1% FDR.
The features of the replicate runs were mapped and quantified.

Figure 1. Processing Workflow

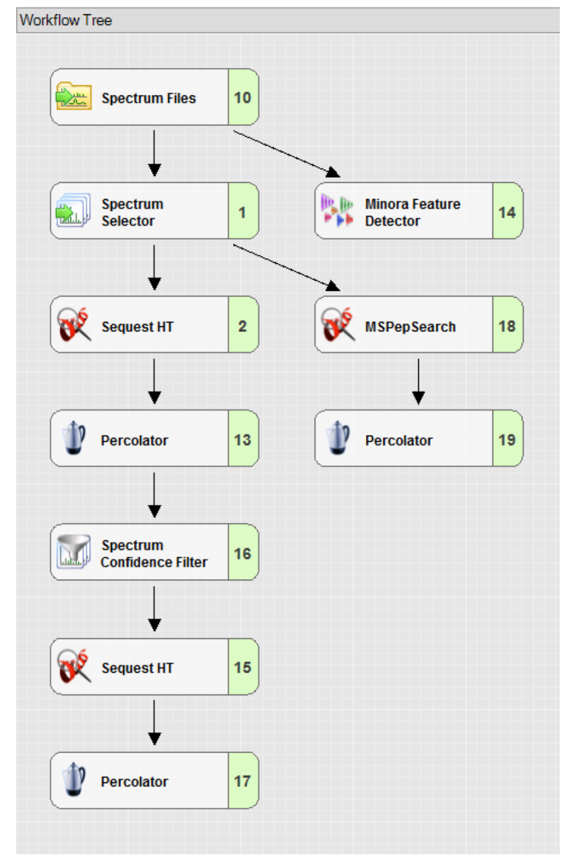
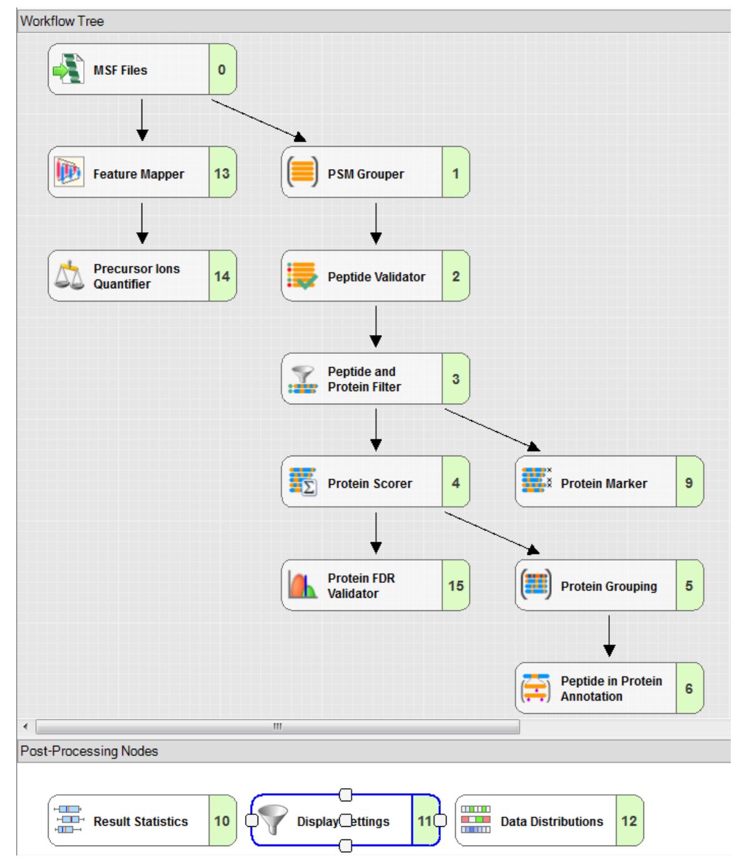


Figure 2. Consensus workflow



RESULTS

Identification

Although producing more unique peptide groups and protein groups,[1] the spectra acquired on the Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer (with APD) show relatively, higher isolation interference, compared to a similar run on a Thermo Scientific™ Q Exactive™ HF hybrid quadrupole-Orbitrap Mass Spectrometer without APD (Figure 3). The distribution of the isolation interference varies slightly with gradient length and isolation window but shows generally the same trend: approximately 2/3 of the spectra have more then 20% isolation interference.

Figure 3. Histogram of the isolation interference distribution of spectra in a raw file, (A) acquired on a Q Exactive HF hybrid quadrupole-Orbitrap Mass Spectrometer (without APD) and on a Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer

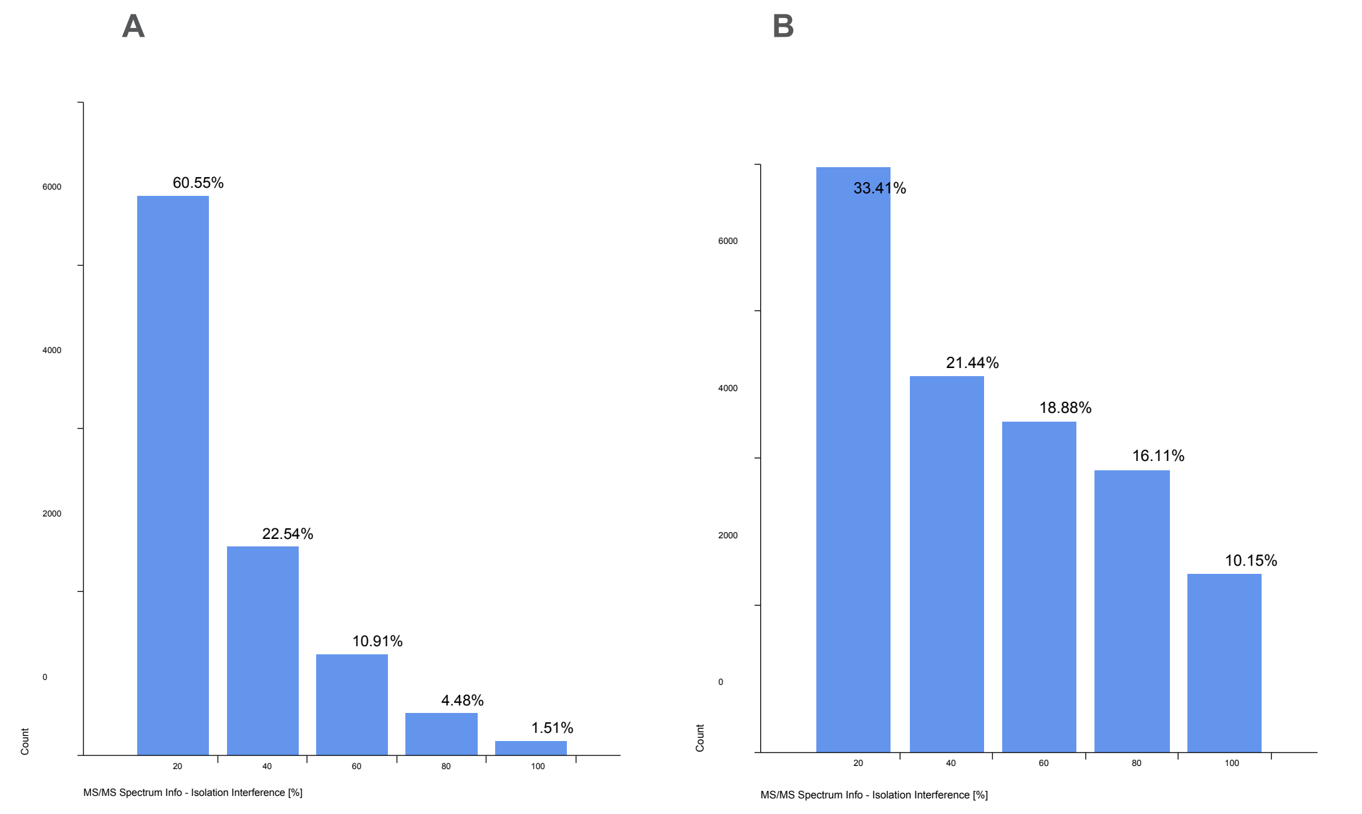


Figure 3.A shows the interference distribution of spectra in a typical run for an Q Exactive HF hybrid quadrupole-Orbitrap Mass Spectrometer with the classical peak detection.

Figure 3.B shows the interference distribution of spectra in a typical run for an Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer with APD. A larger number of spectra has higher isolation interference compared to a similar run on an instrument with no APD.

Table 1. Summary of peptide groups identified with Sequest HT

Sequest HT	Replicate 1	Replicate 2	Total	Total %
Unmodified Peptide groups	4482	4596	5802	74%
Modified peptide Groups	1301	1305	1705	22%
Semi-tryptic peptide groups	234	231	318	4%
Total	6014	6132	7825	

An overview of the identified peptide groups by Sequest HT can be found in Table 1. More then 7600 peptide groups have been identified in the duplicate 11.5 min. runs. The most abundant and common modifications represent 22% of the identifications. Additionally 4% more peptides are identified by a semi-tryptic search.

To increase further the number of identifications the spectra were searched in parallel with MSPepSearch using the first release of the ProteomeTools library. The summary is displayed in Figure 5 for the peptides and Figure 6 for the proteins.

Figure 4. Histogram of the isolation interference distribution of spectra and psms of the 2 replicates.

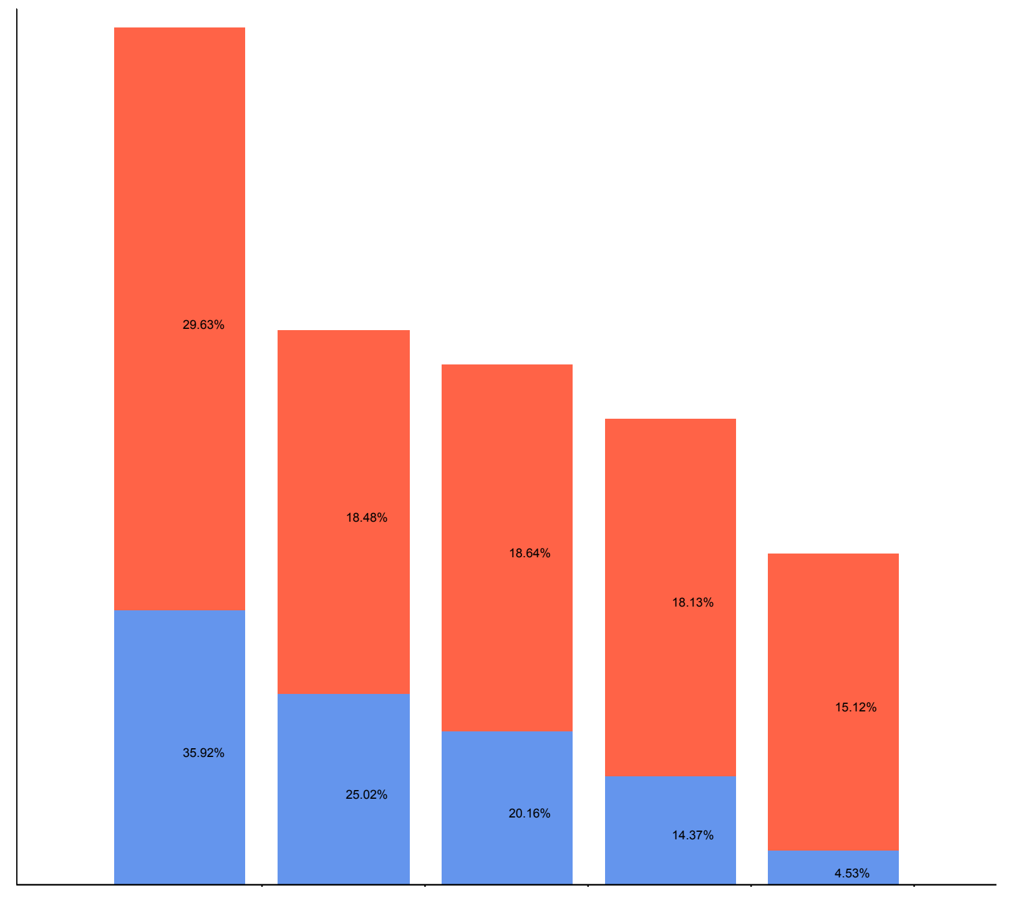
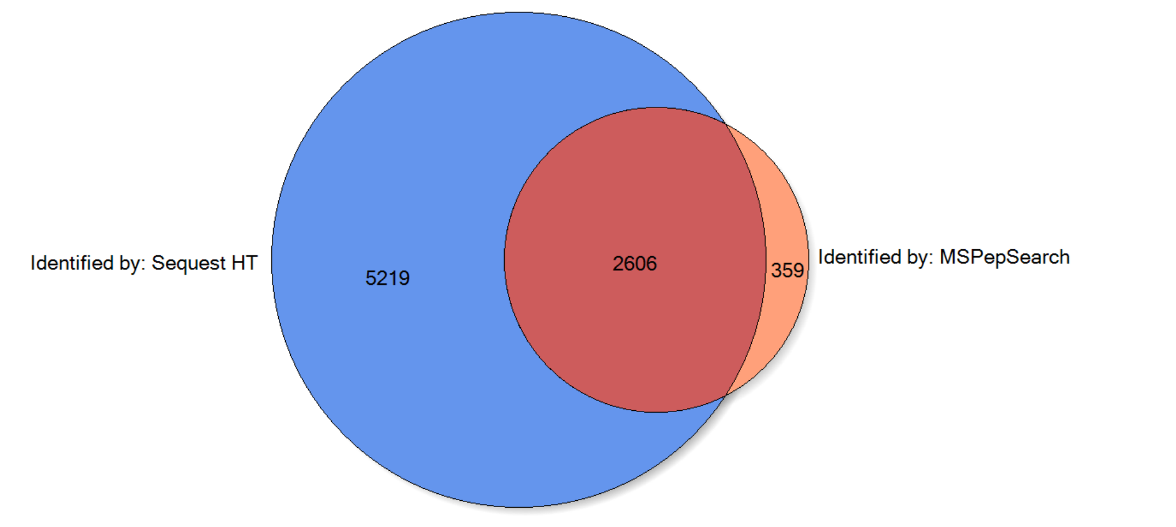


Figure 4 shows the interference distribution of spectra and psms for the 2 replicates. The spectra in red and the psms in blue. The relative identification rate of the spectra is decreasing with increasing isolation interference.

Although only a small amount (358 peptides) of peptide groups have been identified exclusively with MSPepSearch (Figure 5), this result in a relatively high number (381) of protein groups exclusively identified using the spectral library (Figure 6).

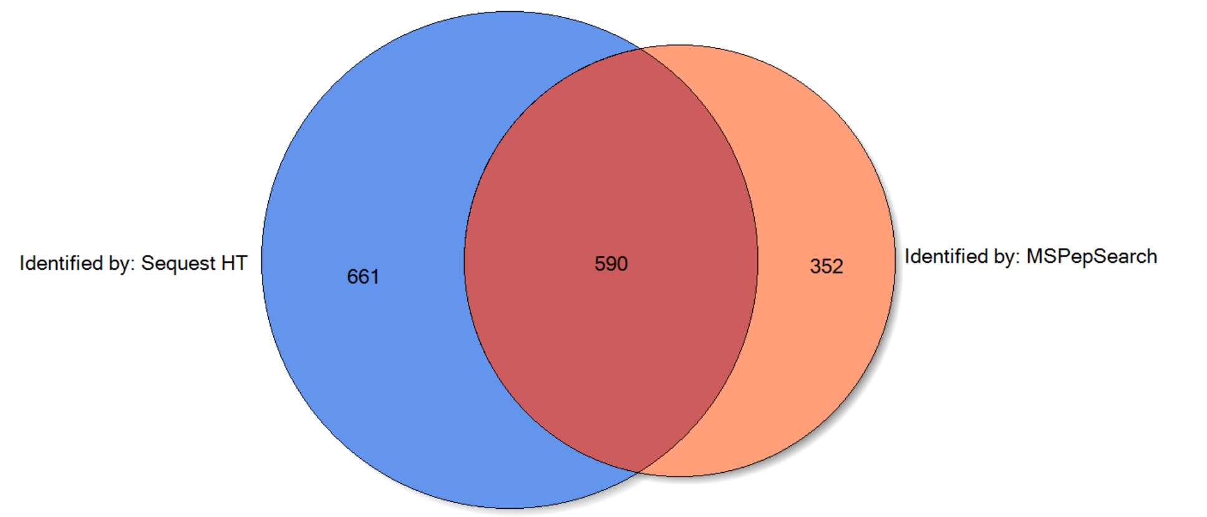
Nevertheless, there are still a high number of spectra with high isolation interference that are not identified (Figure 4).

Figure 5. Venn diagram of peptide groups identified with Sequest HT and MSPepSearch



	Exclusive	Total	Label	Description
🟦	5219	7825	Identified by: Sequest HT	Identified by: Sequest HT
🟧	359	2965	Identified by: MSPepSearch	Identified by: MSPepSearch
🟥	2606	2606	Identified by: Sequest HT / Identified by: ...	Identified by: Sequest HT / Identified by: ...
Sum	8184			

Figure 6. Venn diagram of protein groups identified with Sequest HT and MSPepSearch



	Exclusive	Total	Label	Description
🟦	661	1251	Identified by: Sequest HT	Identified by: Sequest HT
🟧	352	942	Identified by: MSPepSearch	Identified by: MSPepSearch
🟥	590	590	Identified by: Sequest HT / Identified by: ...	Identified by: Sequest HT / Identified by: ...
Sum	1603			

The low number of exclusive peptide groups identified by MSPepSearch can be explained by the fact that the number of peptides per protein is low in the spectral library, compared to the theoretical number in FASTA file, used for the Sequest HT database search. This number will increase when the ProteomeTools library will be completed (additional 1.100.000 synthetic peptides).

However as the peptide sequences, selected for the synthetic spectral library, are the best (most observed, unique, producing the highest score, etc.) for MS-MS analysis (2) those result in relatively more exclusive protein groups.

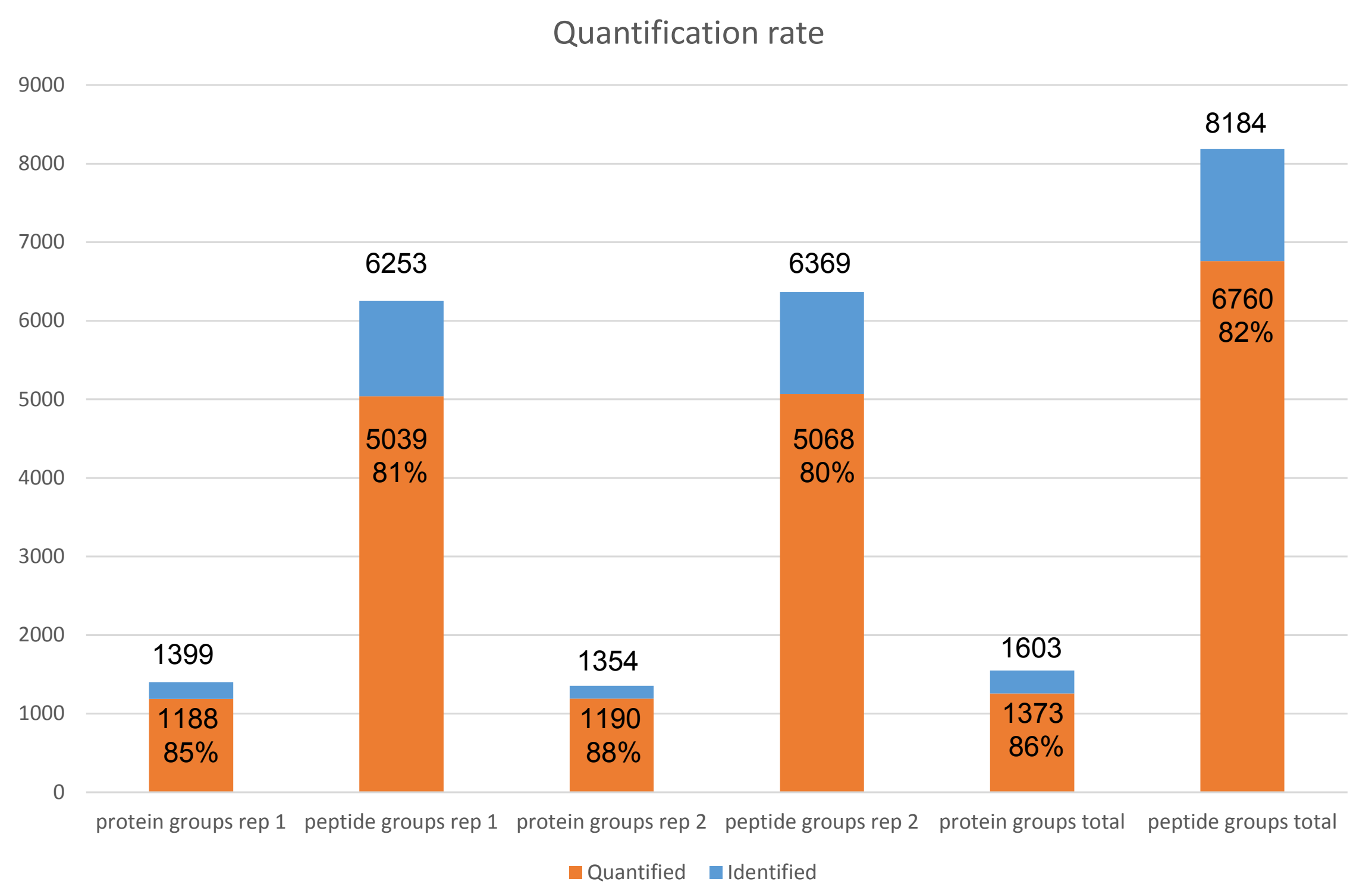
More then of 8100 peptide groups and 1600 protein groups have been identified in 11.5 min.

Quantification

The quantification result is displayed in Figure 7. The number of identified proteins is rather irrelevant if those cannot be quantified. Here more than 1370 proteins were quantified in a total run time of 11.5 min.

Similar results were obtained using a 5 min gradient @1.2uL/min with a Thermo Scientific™ Dionex™ UltiMate™ 3000 RSLC Nano System Flow Meter coupled to a Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer (data not shown).

Figure 7. Summary of the quantification



CONCLUSIONS

- More then 1370 protein groups have been quantified in 11.5 minutes, which allows to analyze more then 100 samples a day.
- Spectral library searching improves the identification rate of proteins. This will further improve upon completion of the spectral database generated by the ProteomeTools project.

REFERENCES

1. Tabiwang N. Arrey, et al, ASMS 2017: “New innovations implemented on the Q Exactive HF mass spectrometer”.
2. D. Zolg et al, Nature Methods, 14, 259 (2017): “Building ProteomeTools based on a complete synthetic human proteome”.
3. Zheng Zhang et al, J. Proteome Res. 2018, 17, 846−857: “Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches”.

TRADEMARKS/LICENSING

PO65241

ThermoFisher
SCIENTIFIC