

Increased MS Protein Identification Rates Using 75 cm Long nano LC C18 Separation Columns: Pushing the Limits of Bottom-Up Proteomics

Daniel Lopez-Ferrer,¹ Michael Blank,¹ Stephan Meding,² Aran Paulus,¹ Romain Huguet,¹ Remco Swart,² Andreas FR Huhmer¹

¹Thermo Fisher Scientific, San Jose, USA; ²Thermo Fisher Scientific, Germering, Germany

Overview

Purpose: Bottom-up proteomics has always aimed to identify and quantify the complete proteome from a cell, tissue, or whole organism. Many advances have been made in the last 15 years. Still sample separation is one of the technological challenges. Separation columns have continuously increased in length. So far, 50 cm columns were the longest commercially available high performance nano LC columns. It was evaluated whether using newly available 75 cm separation columns will significantly increase peptide and protein identification rates.

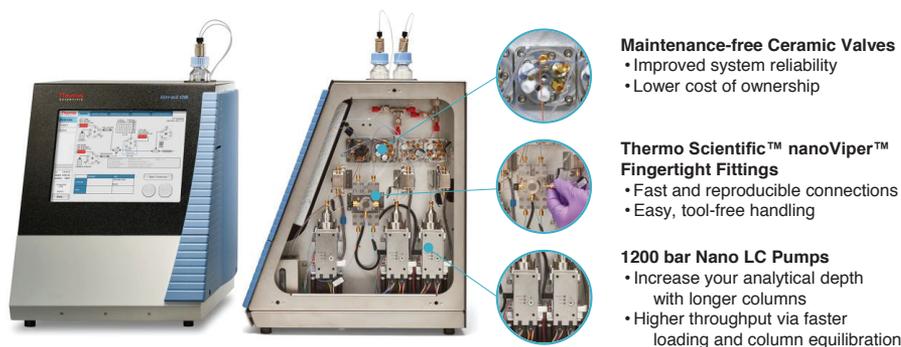
Methods: A Thermo Scientific™ EASY-nLC™ 1200 LC system with a Thermo Scientific™ Orbitrap™ Fusion™ Lumos™ Tribrid™ mass spectrometer were used to analyse a HeLa cell lysate with a 75 cm long 75 μ m ID Thermo Scientific™ Acclaim™ PepMap™ nano LC column using both 2 and 4 hour gradients. The results were compared with those obtained under the same conditions with a 50 cm column, which was until now the longest commercially available high performance nano LC column for bottom-up proteomics. In both cases, the columns were used in EASY-Spray™ column format.

Results: The length increase resulted in the separation and detection of 10% more unique peptides, and 7% more protein identification in a 4 hour gradient, with protein identifications exceeding 5700 proteins for a single injection of mammalian cell lysate. More importantly, longer columns showed better reproducibility as seen by increased correlation among technical replicates, higher numbers of quantifiable peptides, and a smaller coefficient of variance (CV), resulting in improved protein quantification for complex lysates by high resolution accurate mass (HRAM) LC-MS.

Introduction

Since its inception, bottom-up proteomics has aimed to identify and quantify the complete proteome from a cell, tissue, or whole organism¹. Although many advances have been made in the last 15 years, there are still three main challenges to overcome. The first is to obtain complete coverage of the proteome by identifying all the expressed proteins in a given time². The second is working with samples of limited amount like clinical biopsies³, and the third is achieving sufficient analytical throughput⁴. Peptide separation and their MS/MS identification are pillars of modern proteomic analysis and each has seen performance improvements with advances in instrumentation. Thermo Scientific Orbitrap mass spectrometers are now considered the gold standard for mass spectrometry-based proteomics⁵. The recently introduced Orbitrap Fusion Lumos MS is at the time of this writing the instrument with the best sensitivity, best mass resolution, and fastest scan rate. However, due to the complexity of the proteome, even the best mass spectrometers have limitations in dynamic range per spectrum. In order to get the most out of today's state-of-the-art Orbitrap mass spectrometers, an efficient sample separation method has to be coupled with the best peptide separation system to characterize as many unique peptides, and identify as many proteins in a given lysate as possible. In order to improve the sample separation efficiency longer columns and smaller stationary phase particles have been developed. Using the EASY-nLC 1000 system with a maximum pressure limit of 1000 bars, it was possible to run columns of up to 50 cm at elevated temperatures of 40 to 50°C at standard nano LC flow rates of approximately 300 nL/min. The newly introduced EASY-nLC 1200 system now allows for a maximum back pressure up to 1200 bar, enabling routine operation with columns of 50 cm and longer.

FIGURE 1. Front and side views of the EASY-nLC 1200 system, and detailed view of its technical features.



Methods

Reagents

All solvents were LC-MS grade and purchased from Fisher Scientific. Solvent A was 100% water with 0.1% formic acid. Solvent B was 80% acetonitrile, 20% water and 0.1% formic acid. Aliquots containing 500 ng/ μ L HELA protein digest (Pierce, PN 88328) and 50 fmol/ μ L of peptide retention time calibration (PRTC) standards (Pierce, PN 88320) in water with 0.1% formic acid were prepared for the study.

LC-MS/MS

All analyses were performed using an EASY-nLC 1200 system. HeLa cell digest sample was loaded directly onto the column using the one-column (direct injection) mode, with either 2 or 4 μ L injected onto the column, corresponding to 1 or 2 ug respectively. The analytical columns used were a 75 μ m ID Acclaim PepMap column with 2 μ m particles manufactured in EASY-Spray format being either 50 cm (ES803) or 75 cm in length (ES805). The column temperature was maintained at 55 °C. Gradient conditions are described below:

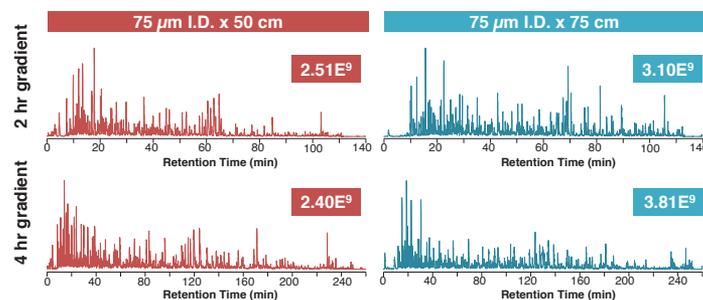
Composition	120min gradient	240min gradient
5-28%B	0-105 min	0-210 min
28%-40%B	105-120 min	210-240 min
40-95%B	120-130 min	240-250 min
95-95%B	130-140 min	250-260 min

An Orbitrap Fusion Lumos instrument was used for peptide MS/MS analysis. Survey scans of peptide precursors were performed from 375 to 1575 m/z at 120K FWHM resolution (at 200 m/z) with a 4×10^5 ion count target and a maximum injection time of 50 ms. The instrument was set to run in top speed mode with 3 second cycles for the survey and the MS/MS scans. After a survey scan, tandem MS was then performed on the most abundant precursors exhibiting a charge state from 2 to 7 of greater than 5×10^3 intensity by isolating them in the quadrupole at 1.2 Th. CID fragmentation was applied with 35% collision energy and resulting fragments detected using the rapid scan rate in the ion trap. The AGC target for MS/MS was set to 10^4 and the maximum injection time limited to 35 ms. The dynamic exclusion was set to 12 seconds with a 10 ppm mass tolerance around the precursor and its isotopes. Monoisotopic precursor selection was enabled.

Data Analysis

Raw data was processed using Thermo Scientific™ Proteome Discoverer™ 2.1.0.80 software. MS2 spectra were searched with the SEQUEST® HT engine against a database of 42085 human proteins including proteoforms (UniProt, July 14th, 2015). Peptides were generated from a tryptic digestion allowing for up to two missed cleavages, carbamidomethylation (+57.021 Da) of cysteine residues was set as fixed modification, and oxidation of methionine residues (+15.9949 Da), acetylation of the protein N-terminus (+42.0106) and deamidation of asparagine and glutamine (+0.984) were treated as variable modifications. Precursor mass tolerance was 10 ppm and product ions were searched at 0.8 Da tolerances. Peptide spectral matches (PSM) were validated using the Percolator algorithm⁶, based on q-values at a 1% FDR. With Proteome Discoverer software, peptide identifications were grouped into proteins according to the law of parsimony and filtered to 1% FDR. The area of the precursor ion from the identified peptides was extracted using the Precursor Ions Area Detector plug-in. For further analysis PSMs and Peptide Groups passing the FDR were exported to a text file and processed using Dante RDN⁷. In addition, Skyline 3.1 software⁸ was used to extract ion chromatograms of the PRTC standards to calculate full width at half maximum (FWHM), coefficients of variation, retention time variation and peptide peak capacity.

FIGURE 2. Representative chromatograms obtained for 2 and 4 hr gradients, and 50 cm and 75 cm columns, respectively.

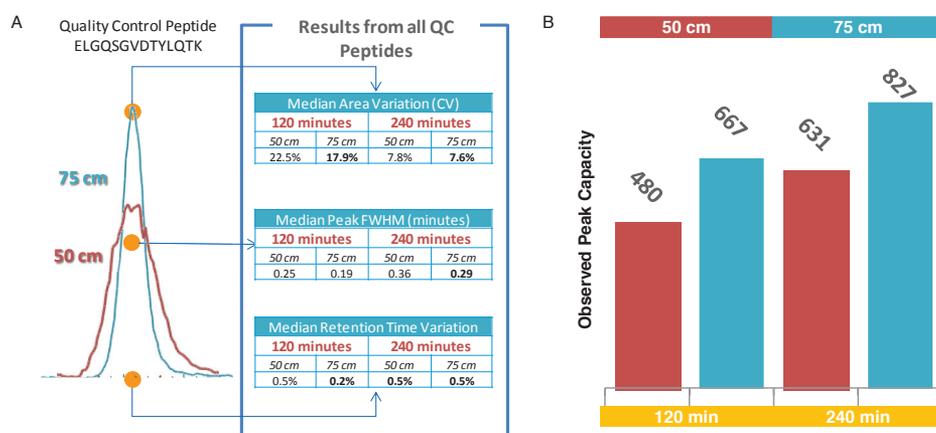


Results & Discussion

Chromatographic Performance

Reproducibility of the chromatographic separation is the number one requisite for a reliable comparison among different runs and ultimately obtaining quantitative information about the proteome under analysis. Figure 2 shows representative chromatograms for each of the columns and gradients. As it can be seen, base peak chromatograms are very consistent among all the analyses with the highest degree of similarity among replicates for each of the setups. At the beginning of the chromatogram a small shift in the retention time (RT) is observable due to the increased volume of the 75 cm column. The peak profiles among replicates were almost identical and peptide retention time shifts of less than 1 minute were observed even when employing a 240 minute long gradient. Figure 3 shows the significant chromatographic performance parameters for the 15 PRTC standards spiked in as a quality control in all the runs. The 75 cm column performs significantly better than its shorter counterpart. Coefficients of Variation for peptide peak areas, median full width half maximum values, and RT variation are always less for the longer column, independent of the gradient length. Furthermore peak capacity was used to evaluate the performance of each chromatography configuration. The 75 cm column achieves a peak capacity of over 800 employing a 240 minute gradient, which almost doubles previous data reported recently by MacCoss and colleagues⁹. Interestingly, the 75 cm column achieves a higher peak capacity in 2 hours than that of the 50 cm column with a 4 hour gradient. With careful optimization of the LC and MS parameters the 75 cm column could achieve very similar results with a shorter gradient than the 50 cm column with the 4 hour gradient. Since the 75 cm column does not approach the maximum pressure rated for the EASY-nLC 1200 system, the chromatography could potentially be optimized even further to maximize separation.

FIGURE 3. A) Extracted ion chromatogram for one of the 15 representative QC peptides and average chromatographic metrics of all 15 QC peptides obtained for different experimental configurations. B) Histogram comparing the peak capacity obtained for each of the experiments.



Peptide and Protein Identifications

From a proteomics perspective, researchers in the field are commonly interested in the number of peptides identified, either in terms of peptide spectral matches, unique peptides or protein groups. As shown in Figure 4, the 75 cm column consistently results in the highest total number of peptides and protein identifications by at least 7% margin. Whereas in the past, reproducibility among replicates was typically around 80%, in this study, the results are highly reproducible with less than 5% of the peptide/protein identifications for a given dataset not shared with any of the other replicates in all 4 replicates.

We further investigated if the better performance in the peptide identification occurred only in certain parts of the LC-MS analysis or consistently across the whole gradient. As shown in Figure 4D, the 75 cm column provides better identifications across the whole gradient. This can be explained by the improved separation in the 75 cm column setup which will cause a given peptide to elute at higher concentration and thus more likely to yield a higher quality MS/MS spectrum, which in turn results in a positive identification. Figure 5A shows the rank of the identified and quantifiable proteins for the 4 hour gradient, as expected the longer column goes deeper into the proteome coverage.

It can be claimed that 5 to 10% increase in peptide and protein identifications is not substantial. However compared with other studies, these experiments represent breakthrough new levels of both peptide and protein identification and depth of coverage. Pathway analysis was then performed using Thermo Fisher Cloud. Results from both columns yielded the same profile of overrepresented pathways, but with different degrees of coverage, demonstrating that the overall study was unbiased. Figure 5B shows that the data obtained for the 75 cm column with the 4 hour long gradient provides direct quantitation of almost 50 % percent of the proteins in any of the 23 overrepresented pathways.

FIGURE 4. A) Venn diagrams showing the overlap among technical replicates for the identified proteins from varying column and gradient lengths. B) Number of identified peptide and protein groups. C) Venn diagram showing the total number of overlapped proteins for both column lengths. D) Line graph displaying the trend of identified peptides versus retention time during the LC-MS analysis.

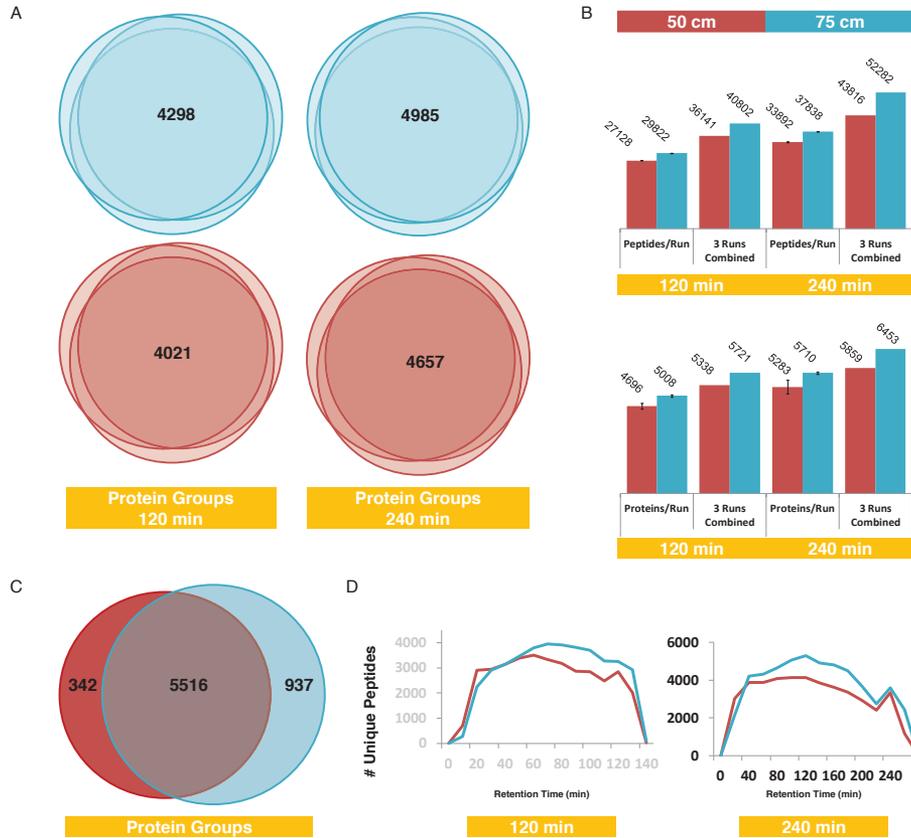
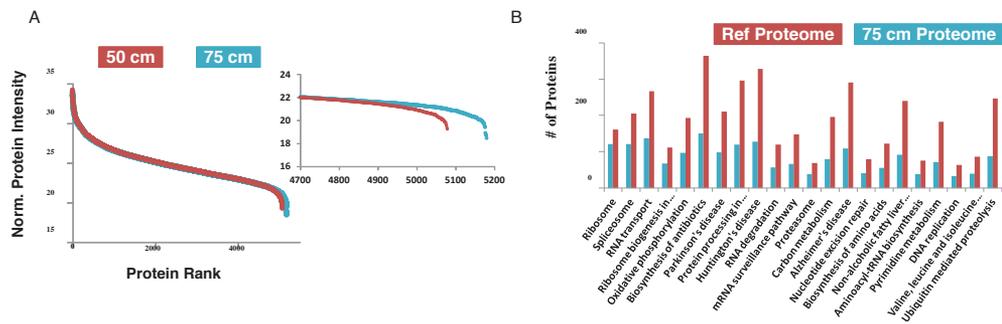
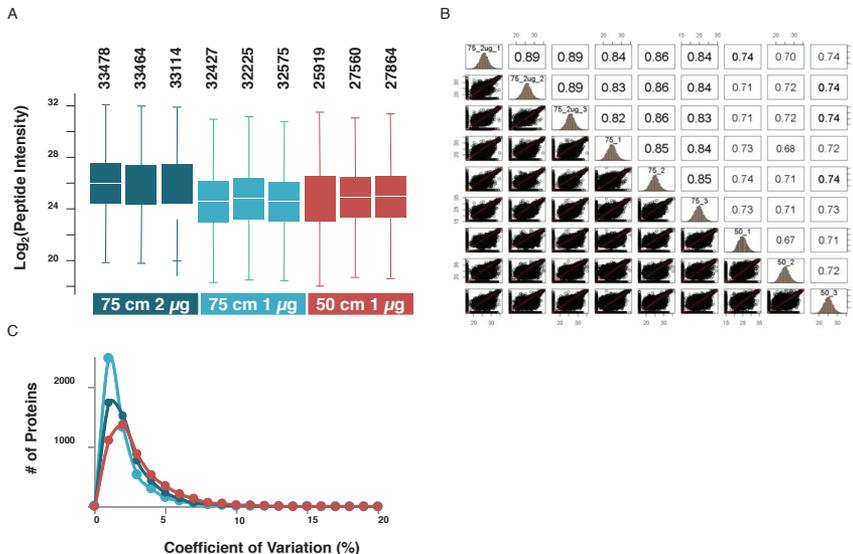


FIGURE 5. A) Protein rank of proteins over their normalized protein intensity for both column lengths. B) Overrepresented pathways for the 75 cm column length and 4 hour long gradient dataset.



In addition, major improvements were achieved with regard to quantitation. Figure 6 shows that the 75 cm column increased the number of quantifiable peptides by 20%. This results not only in more peptides to be quantified, but in a higher correlation among replicates (>85%) while at the same time those peptides and proteins have lower CVs allowing for more accurate quantitation. Finally, we examined the effect of the peptide loading amount on protein identification, quantitation and impact in the retention time. Doubling the amount of peptide digest loaded onto the column did not significantly increase the number of protein identifications or quantifiable peptides. However, increasing the loading amount dramatically improved the correlation among runs up to 89% and doubled the number of proteins with CVs below 5%, allowing for more accurate proteome quantitation. The chromatography was not substantially affected, with observed retention time shifts between 1 μ g and 2 μ g loads of less than 1 min.

FIGURE 6. A) Box and whisker-plot indicating the median peptide group intensity and distribution for increasing column length and load on column. B) Correlation plot showing the peptide peak area similarity between replicates and among other experimental conditions. C) Distribution of coefficients of variation (CV) for protein area quantitation among the different experimental conditions.



Conclusion

The EASY-nLC 1200 system coupled with a high performance Orbitrap mass spectrometer represents a very powerful platform for carrying out high performance proteomic experiments. By systematically evaluating the most common gradients in the proteomic field, we have demonstrated that the use of longer columns employing 2 or 4 hour gradients represents a valuable alternative to perform quantitative proteomics compared to current methods. Moreover, we have identified ~6500 proteins without fractionation, and reproducibly quantified over 5000 proteins based only on three technical replicate injections. These results clearly surpass the current standards in the proteomics paradigm and rival quantitation results derived from DIA methods in terms of reproducibility and depth of analysis, but with greater efficiency, as there is no need to first generate a spectral library.

- Increased number of peptide and protein identifications
- Increased identification rate with shorter gradients compared to 50 cm column analyses
- High sample loading capacity
- Increased number of proteins quantified
- Higher correlation of quantifiable peptides between injections and better run-to-run reproducibility

References

1. Wilhelm M et al. Mass-spectrometry-based draft of the human proteome. *Nature*. 2014 May 29;509(7502):582-7.
2. Hebert AS, et al. The one hour yeast proteome. *Mol Cell Proteomics*. 2014 Jan;13(1):339-47.
3. Wu X, et al. Oncotarget. 2015 Sep 3 Global phosphotyrosine survey in triple-negative breast cancer reveals activation of multiple tyrosine kinase signaling pathways.
4. Livesay EA, et al. Fully automated four-column capillary LC-MS system for maximizing throughput in proteomic analyses. *Anal Chem*. 2008 Jan 1;80(1):294-302.
5. Scigelova M, Hornshaw M, Giannakopoulos A, Makarov A. Fourier transform mass spectrometry. *Mol Cell Proteomics*. 2011 Jul;10(7):M111.009431.
6. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*. 2007 Nov;4(11):923-5.
7. Polpitiya AD, et al. DANTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*. 2008 Jul 1;24(13):1556-8.
8. MacLean B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010 Apr 1;26(7):966-8.
9. Hsieh EJ, Bereman MS, Durand S, Valaskovic GA, MacCoss MJ. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *J Am Soc Mass Spectrom*. 2013 Jan;24(1):148-53.

www.thermofisher.com

©2016 Thermo Fisher Scientific Inc. All rights reserved. SEQUEST is a registered trademark of the University of Washington. All other trademarks are the property of Thermo Fisher Scientific and its subsidiaries. This information is presented as an example of the capabilities of Thermo Fisher Scientific products. It is not intended to encourage use of these products in any manner that might infringe the intellectual property rights of others. Specifications, terms and pricing are subject to change. Not all products are available in all countries. Please consult your local sales representative for details.

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

Thermo
 SCIENTIFIC

A Thermo Fisher Scientific Brand