Speeding up label-free quantitation of complex proteome samples using dia-PASEF

Bruker's timsTOF HT in combination with dia-PASEF® provides reproducible and accurate qualitative and quantitative results in complex proteomics samples with more than 11,400 protein groups identified in 25 minutes.

Abstract

Accurate relative protein quantitation is essential for the proper understanding of biological processes. The unique combination of speed, selectivity, sensitivity, and robustness delivered by Bruker's proteomics solutions enables users to obtain reliable quantitative information for more proteins while working with small sample amounts and/or short gradient times.

Data-independent acquisition using dia-PASEF [1] is both more sensitive and selective than traditional DIA approaches as it combines the advantages of DIA with the inherent ion-usage efficiency of PASEF (Parallel Accumulation Serial Fragmentation). Making use of the correlation of molecular weight and CCS coded information from the dual-TIMS funnel, dia-PASEF enables highly confident identification.

We utilized sample sets with known expected ratios to investigate the potential of dia-PASEF for high-throughput proteomics. Analyzing a complex hybrid proteome sample (human, yeast, *E.coli*) using dia-PASEF on the timsTOF HT reliably identified more than 11,400 protein groups from 116,600 peptides using a single 25-minute gradient applying library-free data processing (directDIA™, Spectronaut/Biognosys). When reducing the injected sample amounts to 10 ng, the approach still identifies more than 8000 protein groups. Further increasing throughput by combining a 7-minute gradient and a dia-PASEF method optimized via the py_diAID tool developed by the group of Matthias Mann (https://pypi.org/project/pydiaid) allowed identification and quantification of more than 7500 protein groups underlining the perfect suitability of dia-PASEF for high-throughput proteomics.

Keywords:
dia-PASEF, label-free quantitation, high-throughput proteomics, timsTOF HT

Romano Hebeler, Stephanie Kaspar-Schoenefeld, Gary Kruppa; Bruker Daltonics GmbH & Co. KG, Bremen, Germany.

Introduction

High-throughput proteomics is more and more commonly used to address increasing demands for large cohort analysis required in clinical research and drug development applications. Due to the extreme complexity of proteomes, optimized acquisition methods are of high importance. DIA approaches are very attractive when the aim is to analyze a large sample cohort due to the very high reproducibility and data completeness. In DIA methods, all ions in a given *m/z* window are fragmented in every run, in principle allowing full identification and quantitative comparison from run-to-run. dia-PASEF, which combines DIA with the PASEF principle, has already been proven to deliver reproducible identification and quantitation information in discovery proteomics studies [1].

The newly developed timsTOF HT instrument enhances the benefits of the existing platform with improved ion storage capacity. This results in an increase in dynamic range for the eluting peptides providing additional analytical depth for quantitative proteomics and retaining the additional specificity provided by the TIMS separation.

Besides improvements on the hardware and data-acquisition side, high-throughput proteomics is also dependent on sophisticated data processing software. DIA approaches are typically dependent on comprehensive spectral libraries created from fractionated samples. This approach not only adds additional steps in sample preparation and data acquisition but also for subsequent processing leading to a dramatic reduction in throughput, which is a bottleneck for large biological and clinical studies. Recent developments in the various software tools available for DIA-based proteomics, showed significant improvement of library-free data processing making DIA approaches even more attractive compared to classical DDA workflows.

In this study, library-free data processing was applied to investigate the performance of dia-PASEF for high-throughput proteomics using short gradients (25 and 7 minutes). To mimic complexity found in biological samples three model proteomes were mixed (human, yeast, *E. coli*, Figure 1).

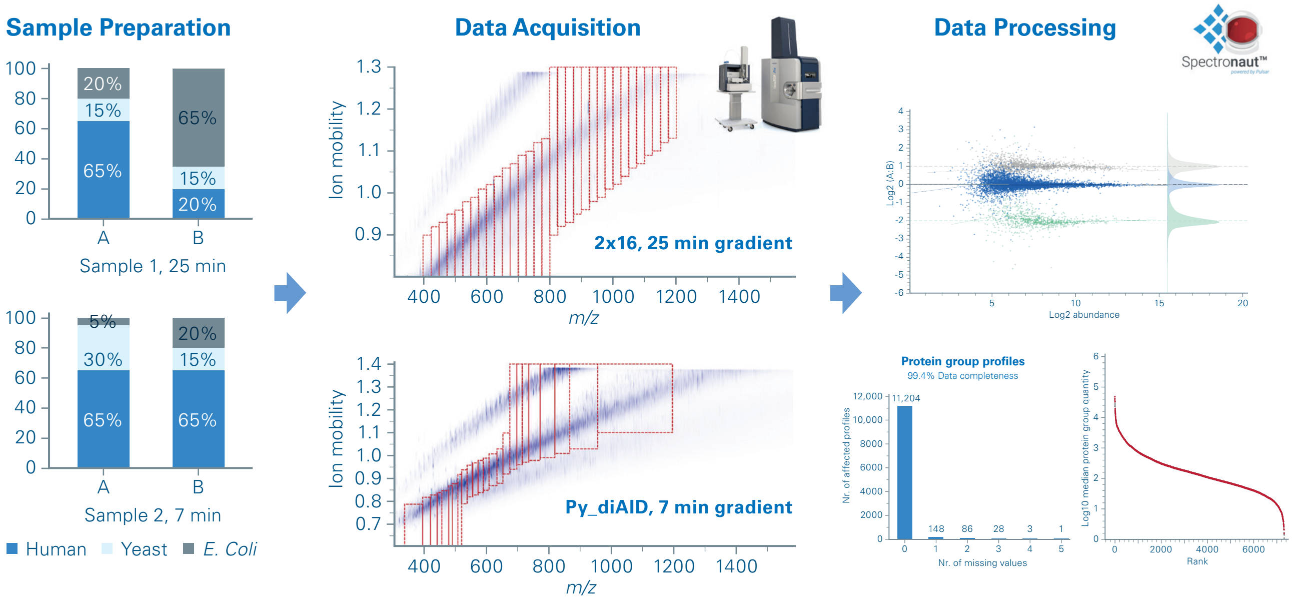


Figure 1

Workflow for in-depth proteomics analysis of complex proteomics samples. We used a three-proteome sample (human, yeast, *E. coli*) to mimic the complex biological samples typically found in proteomic studies. Samples were analyzed using nanoLC coupled to a timsTOF HT using dia-PASEF with either a 25- or 7-min gradient. Data processing was done using Spectronaut (unreleased version, Biognosys).

Methods

Sample set 1_25-minute gradient: Tryptic digests of a human cell lysate (Promega), Saccharomyces cerevisiae (Promega) and *E. coli* (Waters) were combined in following ratios:

Sample A: 65% human, 15% yeast, 20% *E. coli*
Sample B: 20% human, 15% yeast, 65% *E. coli*

10, 50 and 100 ng of the mixed samples were loaded directly on a 25 cm C18 column (75 µm inner diameter, 1.9 µm particle size, Aurora, IonOpticks) using a nanoElute coupled to a timsTOF HT mass spectrometer via a CaptiveSpray ionization source. Peptides were eluted with a 25-minute acetonitrile (ACN) gradient. For the dia-PASEF acquisition, a window placement scheme using two windows in each 100 ms dia-PASEF scan was used. Sixteen of these scans covered an *m/z* range of 100 – 1700 *m/z* (mass width per window: 26 Da width, 1 Da overlap) and mobility range from 0.9 – 1.30 1/K_0 with a cycle time of 1.8 seconds, including one
MS1 frame.

Sample set 2_7-minute gradient: Tryptic digests of a human cell lysate (in-house digest), Saccharomyces cerevisiae (Promega) and *E. coli* (Waters) were combined in following ratios:

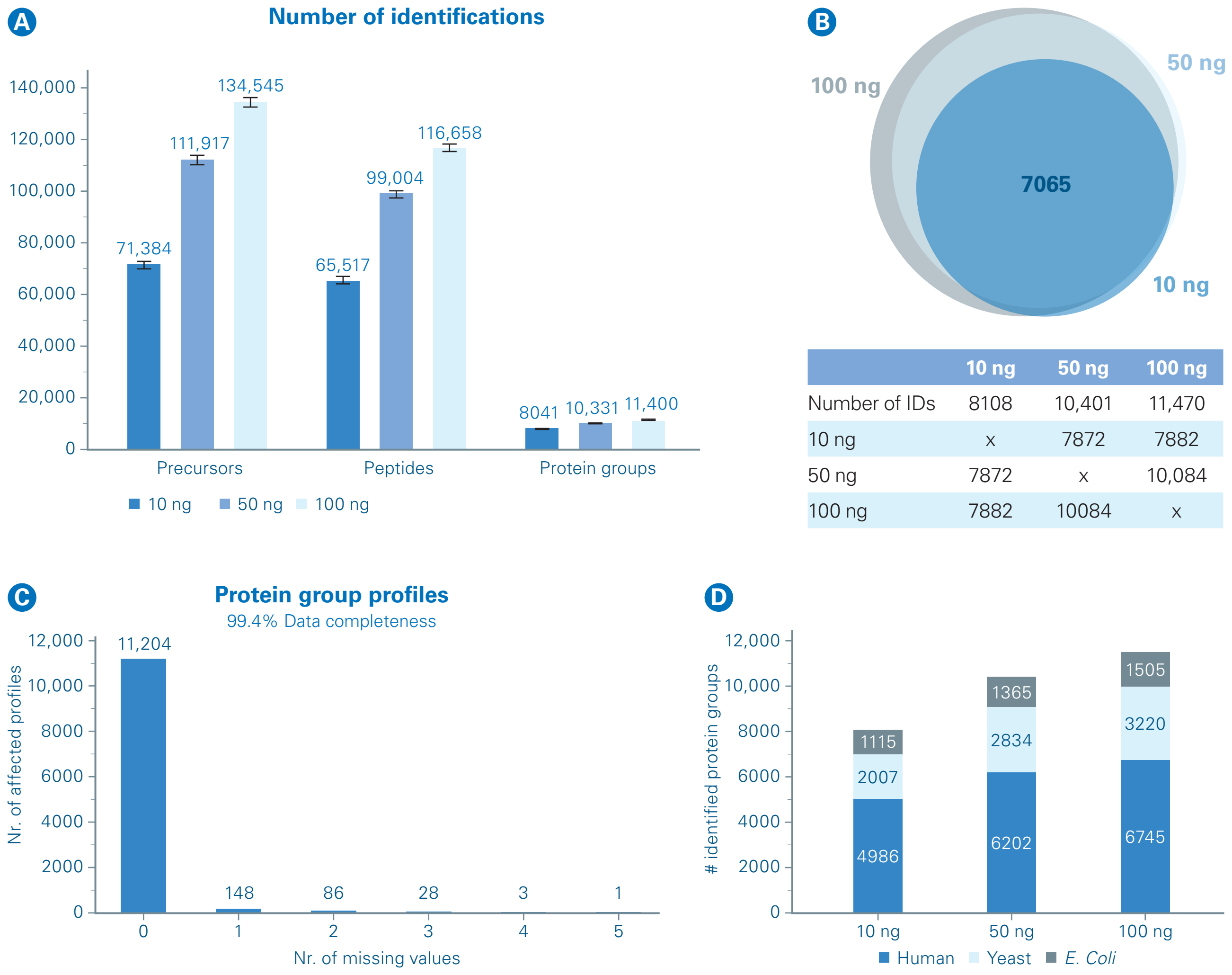
Sample A: 65% human, 30% yeast,   5% *E. coli*
Sample B: 65% human, 15% yeast, 20% *E. coli*

400 ng of the mixed sample was loaded directly on an 8 cm C18 column (150 µm inner diameter, 1.5 µm particle size, PepSep, Bruker Daltonics GmbH & Co KG) using a nanoElute coupled to a timsTOF HT mass spectrometer via a CaptiveSpray ionization source. Peptides were eluted with a 7-minute acetonitrile (ACN) gradient. The mass spectrometer was operated in dia-PASEF mode, for which we defined isolation windows in the *m/z* vs. ion mobility plane adjusted to expected precursor ion density using the py_diAID software (https://github.com/MannLabs/pydiaid, [2]). In short, a sample specific library created with Spectronaut software was loaded using the browser-based GUI of py_diAID according to the instructions. Automatic optimization was done applying following parameters: *m/z* range from 300 to 1200 Da, ion mobility range from 0.6 to 1.3 1/K0, number of dia-PASEF scans set to 12 and number of ion mobility windows per dia-PASEF scan set to 2. Isolation window overlap was set to 0 and shift of the final acquisition scheme in IM dimension to 0.022 1/K_0. The applied number of iterative optimization steps was 200 and number of starting points 20. The resulting method (Figure 1) has an average dia-PASEF window size of 35.66 Da (minimum window size: 11.51 Da, maximum window size: 240.04 Da) and consists of 8 frames with 3 mass windows per frame (75 ms) with a cycle time of 0.675 seconds (including one MS1 frame).

Data was processed in Spectronaut (unreleased version, Biognosys) using library-free mode (directDIA). For direct database identifications from dia-PASEF runs, we used human, *E. coli* and yeast Uniprot fasta files. False discovery rate (FDR) was controlled at 1% for peptide and protein group level.

Results and Discussion

We investigated the performance of dia-PASEF on the timsTOF HT for high-throughput label-free quantitation of complex proteomics mixtures using short gradients. For evaluation of the presented approach using a 25-minute gradient, three proteomes were mixed in defined ratios (3.25:1 for human, 1:3.25 for *E. coli* and 1:1 for yeast) with yeast proteins representing the background proteome.



	10 ng	50 ng	100 ng
Number of IDs	8108	10,401	11,470
10 ng	x	7872	7882
50 ng	7872	x	10,084
100 ng	7882	10084	x

Figure 2

Reproducible and in-depth peptide and protein identification using library-free data processing (directDIA, Spectronaut [unreleased version]). (A) Different amounts of proteome mix HYE (10, 50, 100 ng) were separated using a 25-minute gradient in triplicate injections. Number of precursors, peptides and protein groups are shown. (B) Overlap in protein groups between different amounts injected on column. In total, 7065 protein groups were reproducibly found in all three sample sets. (C) Example showing data completeness of the presented study. In total, 99.4% /11,204 protein groups were found in all 6 runs (2 samples, 3 technical replicates) from 100 ng injections. Same degree of reproducibility was found for the 50 ng (99.3%/10,111 protein groups) and 10 ng (99.2%/7810 protein groups) sample sets. (D) Number of identified protein groups per spiked-in proteome for the different amounts injected on column.

High-throughput proteomics screening is not only dependent on the latest hardware but also on fast processing software. Traditionally, DIA approaches required generation of spectral libraries created from time-consuming measurement of fractionated samples based on DDA approaches, significantly increasing the instrument run time and data processing time. Currently library-free approaches show improving performance and are of great interest. Here, we have applied directDIA, implemented in the Spectronaut software from Biognosys, enabling analysis of DIA data without the need for DDA-based spectral libraries. Using an improved directDIA approach, we identified and quantified on average 11,400 protein groups / 116,658 peptides at 1% FDR using a 25-minute gradient (resulting in 35-minute run time) when injecting 100 ng of the mixed proteome sample (Figure 2). Injection of just 10 ng sample still led to the identification of an average of 8041 protein groups. DIA approaches benefit from extremely good reproducibility as they don't rely on stochastic precursor selection approaches. Notably, the overlap of identified protein groups between the three different amounts injected was extremely high, with 7065 protein groups identified in all three sample sets (Figure 2 B). Data completeness on protein group level was above 99% for all sample sets (considering 6 runs per sample: 2 conditions with triplicate injections each), representing very few missing values.

Confidence in the applied proteomics approach is not only determined by the reproducibility of identification but also by the peptide and protein quantitation. The great advantage of DIA approaches in general and very specifically of dia-PASEF lies not only in the very reproducible identification of 1000s of proteins but also in a high quantitative consistency and accuracy.

The median coefficient of variation for the replicate runs was around 10% for all three sample sets (13.1, 12.2 and 12.0% for 10, 50, and 100 ng, respectively), illustrating the excellent reproducibility of the timsTOF HT's MS/MS level quantitation. A very high sample correlation with average correlation values of 0.99 between technical replicates was observed.

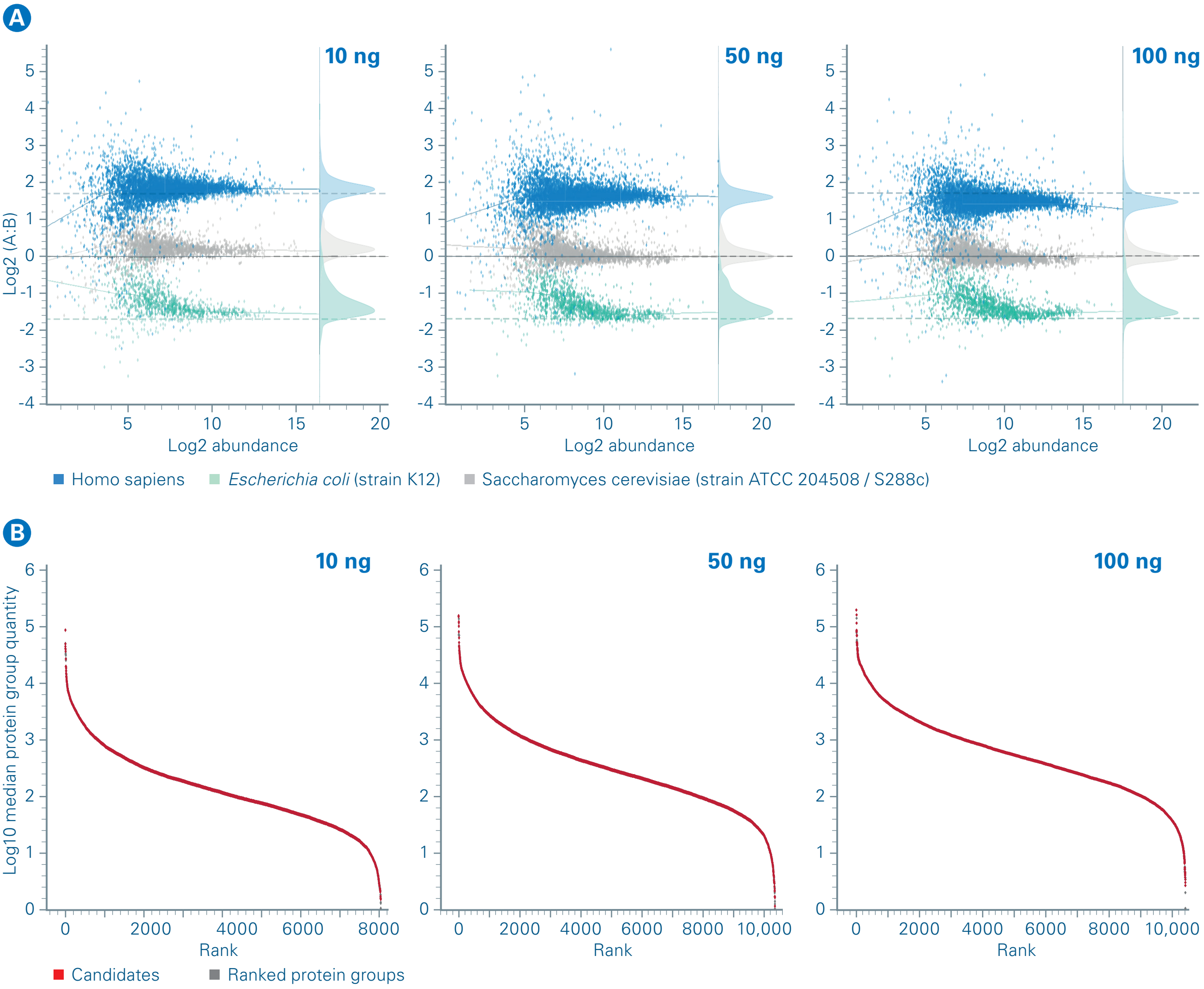


Figure 3

Accurate quantitation of three-proteome mixture independent of sample amount injected. Ⓐ LFQ Bench plot showing measured Log2 fold changes for human (blue), *E. coli* (green) and yeast (grey) for the three different amounts being close to the theoretical ratios (dotted lines). Ⓑ Ranked protein groups are shown, referring to the covered dynamic range.

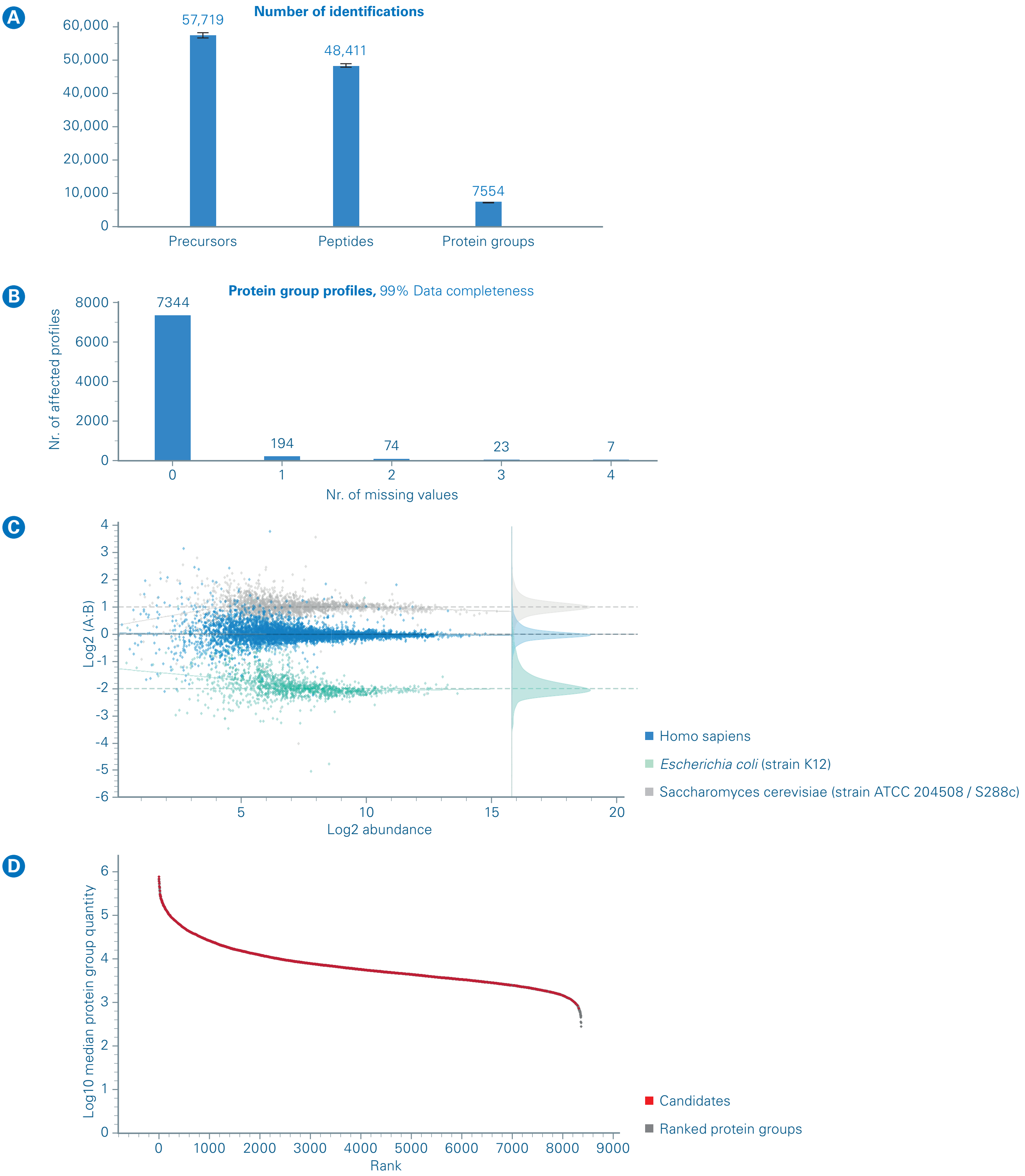


Figure 4

Application of dia-PASEF for high-throughput proteomics reveals more than 7600 protein groups identified using 7-minute gradient.
Ⓐ Average number of identified precursors, peptides and protein groups using library-free data processing (directDIA). Ⓑ Data completeness on protein group level. 98.9% (7344) protein groups were identified in all runs. Ⓒ LFQ Bench plot showing measured Log2 fold changes for human (blue), *E. coli* (green) and yeast (grey) for the three different amounts being close to the theoretical ratios (dotted lines). Ⓓ Ranked protein groups are shown, referring to the covered dynamic range.

The experimental design enabled the evaluation of the quantitative accuracy of dia-PASEF in a complex proteomics mixture with pre-determined theoretical ratios. Background yeast proteins were spiked in equal amounts resulting in a theoretical ratio of 1:1. We found the relative quantitation of the corresponding yeast proteins to be centered at the expected log2 ratio of sample A to sample B of 0 across the complete dynamic range (Figure 3) for all three amounts injected on column. For human and *E. coli* proteins we achieved a very good global accuracy on the timsTOF HT with regulation ratios measured close to the expected ones (median ratio sample A / B for human for 10, 50, and 100 ng, respectively: 3.51; 3.10; 2.80 (expected value: 3.25) and for *E. coli*: 0.40; 0.39; 0.39 (expected value: 0.31). The identified and quantified proteins covered a dynamic range of around 5 orders of magnitude (Figure 3B).

For clinical research or high-throughput drug development proteomics analysis of several hundreds of samples per day is required, which necessitates fast and robust instrumentation. The recently introduced timsTOF HT with 4th generation high capacity TIMS-XR analyzer and advanced digitizer technology (ADT) offers higher dynamic range for unmatched analytical depth in high-throughput proteomics experiments. We made use of the speed and sensitivity of the instrument by testing performance on complex proteome samples using a very short 7-minute gradient (12-minute run time). Working with optimal placement of the dia-PASEF windows is important to make sure that all theoretical precursors are covered even within these very short gradients without compromising coverage of the chromatographic peak. We used py_diAID, a freely available tool developed by the group of Matthias Mann (MPI Martinsried), which automatically adjusts the isolation window width to the precursor density, and optimally positions the isolation design in the *m/z*-IM space. We used our in-depth sample specific library as the basis for optimal dia-PASEF window placement. The resulting method has an average dia-PASEF window size of 35.66 Da (minimum window size: 11.51 Da, maximum window size: 240.04 Da) and consists of 8 frames with 3 mass windows per frame.

As a benchmarking sample for this extremely short gradient method, we used again the hybrid proteome sample of tryptic digests of human, yeast, and *E.coli* proteins, mixed in defined ratios (1:1 for human, 2:1 for yeast, 1:4 for *E.coli* proteins). Notably, on average 7554 protein groups from 48,411 peptides were reproducibly identified and quantified with 99% of the identified protein groups being present in all six runs (2 samples with 3 replicates per sample). Quantitation accuracy was not influenced by the short gradient resulting in very dense peptide distributions within the 7-minute gradient as human background proteins were still centered around log2 ratio of 0 across the complete dynamic range. For yeast and *E. coli* proteins excellent global accuracy was achieved with determined ratios close to the theoretical ones (median ratios for yeast: 1.95 (expected value: 2.0) and for *E. coli*: 0.25 (expected value: 0.25)). The median coefficient of variation for the replicate runs was at 12.8%, illustrating the excellent reproducibility of the timsTOF HT's MS/MS level quantitation even for very short gradients.

Conclusion

The timsTOF HT with dia-PASEF technology delivers not only outstanding numbers of identified proteins and peptides using library-free data processing for short gradients, but notably also excellent quantitation accuracy is achieved for highly complex proteomics samples.

- More than 11,400 protein groups and 116,600 peptides could be reproducibly identified and quantified from a mixed proteome sample (100 ng) using a 25-minute gradient. Using sample loads of 10 ng still reveals more than 8040 protein groups identified and quantified underlining excellent sensitivity of the method.
- Even for very short 7-minute gradients more than 7500 protein groups were detected when applying a dia-PASEF scheme that takes precursor density distribution into account for optimal window placement.
- These results show a high degree of reproducibility and data completeness for short gradients as well as for low sample amounts, making dia-PASEF ideally suited for application to large sample cohorts as for example required in clinical applications.

References

[1] Meier et al. 2020. Nat Methods. **17**(12):1229-1236

[2] Skowronek et al. MCP, https://doi.org/10.1016/j.mcpro.2022.100279

For Research Use Only. Not for use in clinical diagnostic procedures.

Bruker Switzerland AG

Fällanden · Switzerland
Phone +41 44 825 91 11

Bruker Scientific LLC

Billerica, MA · USA
Phone +1 (978) 663-3660

ms.sales.bdal@bruker.com – www.bruker.com

Learn more at **www.bruker.com/timstofpro**