Published in final edited form as:

J Proteome Res. 2023 October 06; 22(10): 3149–3158. doi:10.1021/acs.jproteome.3c00177.

Data-Driven Optimization of DIA Mass Spectrometry by DO-MS

Georg Wallmann,

Departments of Bioengineering, Biology, Chemistry and Chemical Biology, Single Cell Proteomics Center, Northeastern University, Boston, Massachusetts 02115, United States

Andrew Leduc,

Departments of Bioengineering, Biology, Chemistry and Chemical Biology, Single Cell Proteomics Center, Northeastern University, Boston, Massachusetts 02115, United States

Nikolai Slavov

Departments of Bioengineering, Biology, Chemistry and Chemical Biology, Single Cell Proteomics Center, Northeastern University, Boston, Massachusetts 02115, United States; Parallel Squared Technology Institute, Watertown, Massachusetts 02472, United States

Abstract

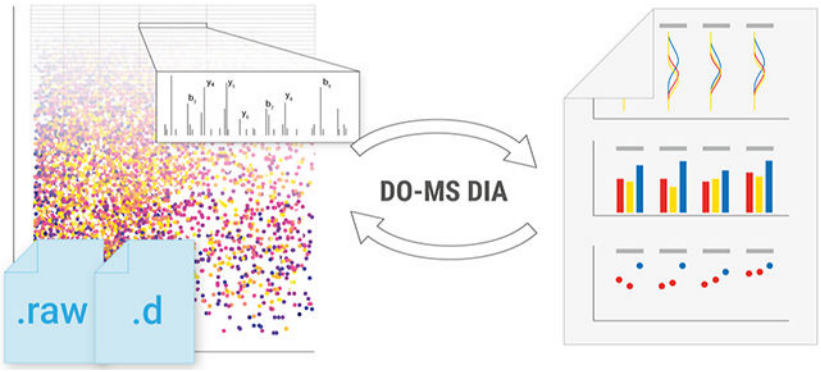
Mass spectrometry (MS) enables specific and accurate quantification of proteins with ever-increasing throughput and sensitivity. Maximizing this potential of MS requires optimizing data acquisition parameters and performing efficient quality control for large datasets. To facilitate these objectives for data-independent acquisition (DIA), we developed a second version of our framework for data-driven optimization of MS methods (DO-MS). The DO-MS app v2.0 (do-ms.slavovlab.net) allows one to optimize and evaluate results from both label-free and multiplexed DIA (plexDIA) and supports optimizations particularly relevant to single-cell proteomics. We demonstrate multiple use cases, including optimization of duty cycle methods, peptide separation, number of survey scans per duty cycle, and quality control of single-cell plexDIA data. DO-MS allows for interactive data display and generation of extensive reports, including publication of quality figures that can be easily shared. The source code is available at github.com/SlavovLab/DO-MS.

Graphical Abstract

Corresponding Author Nikolai Slavov – Departments of Bioengineering, Biology, Chemistry and Chemical Biology, Single Cell Proteomics Center, Northeastern University, Boston, Massachusetts 02115, United States; Parallel Squared Technology Institute, Watertown, Massachusetts 02472, United States; nslavov@northeastern.edu.

The authors declare the following competing financial interest(s): Nikolai Slavov is a founding director and CEO of Parallel Squared Technology Institute, which is a non-profit research institute.

Further documentation on the use of DO-MS is available at do-ms.slavovlab.net. The current version 2.0 is open source and freely available at github.com/SlavovLab/DO-MS. All data shown as example applications are available at do-ms.slavovlab.net/docs/DO-MS_examples. The 30 single-cell plexDIA dataset acquired on the timsTOF has been published as part of plexDIA and is available at https://scp.slavovlab.net/Derks_et_al_2022. All other data acquired for this study have been deposited on MassIVE under the accession MSV000091733.



Keywords

mass spectrometry; proteomics; MS; data; acquisition; quality; control; optimization; DO-MS; plexDIA; single-cell; visualization

INTRODUCTION

Mass spectrometry (MS) allows for comprehensive quantification and sequence identification of proteins from complex biological samples.[1] Reliable sequence identification of peptides by MS relies on the fragmentation of peptides.[2] This can be performed for one precursor at a time, as in the case of data-dependent acquisition (DDA), or for multiple precursors in parallel, as in the case of data-independent acquisition (DIA). Using real-time instrument control for DDA can achieve high sensitivity, depth, and data completeness[3,4] but remains limited to fragmenting only a subset of the available precursors. This limitation is relaxed by DIA, which systematically selects groups of precursors for fragmentation which cover the whole *m/z* range.[5,6] This parallel analysis of multiple precursors can have many benefits, including (1) consistent collection of data from all detectable peptides,[7] (2) high sensitivity due to long ion accumulation times,[8] and (3) high throughput due to the parallel data acquisition.[9] Despite these benefits, parallel fragmentation of all precursors within the isolation window results in highly complex spectra.

This complexity initially challenged the interpretation of DIA spectra, but advances in machine learning and computational power have gradually increased sequence identification from DIA spectra. Initial approaches were based on sample-specific spectral libraries, but newer methods have allowed for direct library-free DIA and deeper proteome coverage.[10-14] Many current approaches use computationally predicted peptide properties (libraries),[15] which remove the overhead of experimentally generated libraries. These improvements continue with new acquisition methods[16-18] and contribute to achieving high proteome depth, data completeness, reproducibility, and throughput.[19,20] This has enabled the quantitative analysis of proteomes down to the single-cell level[21-24] and can continue to increase the throughput and accuracy of single-cell proteomics toward its biological applications.[25]

Orthogonal to the acquisition method, performance can be further increased when labeling samples with non-isobaric mass tags and analyzing them with the plexDIA framework.[26-28] Multiple labeled samples can be combined and analyzed in a single acquisition, multiplicatively increasing the number of protein data points.[29] At the same time, quantitative accuracy and proteome coverage are preserved as identifications can be translated between different samples labeled by non-isobaric mass tags.[26]

To further empower these emerging capabilities, we sought to extend the data-driven optimization of the MS method (DO-MS) app to optimization and quality control of DIA experiments by developing and releasing its second major version, v2.0. Indeed, optimization of DIA workflows requires setting multiple acquisition method parameters, such as the number of MS1 survey scans and the placement of fragmentation windows.

These parameters must be simultaneously optimized for multiple objectives, including throughput, sensitivity, and coverage. Defining the optimal acquisition method therefore becomes a multi-objective, multi-parameter optimization.[30,31] Many tools already exist which cover some aspects of method optimization, like MS2 window placement.[18,32,33] Others focus on quality control.[34,35] DO-MS takes a different approach and offers a holistic view of the acquisition and data processing method specifically designed to diagnose analytical bottlenecks.[31] With this release, DO-MS v2.0 can be used with both DDA data like MaxQuant and DIA data from tools like DIA-NN while having an open interface allowing for adoption to other search engines.

DO-MS is particularly useful for optimizing single-cell proteomic and plexDIA analysis by displaying numerous features relevant to these workflows. These features include intensity distributions for each channel of n-plexDIA[27,29] and ion accumulation times, which are useful for optimizing single-cell analysis,[36,37] particularly when using isobaric and isotopologue carriers.[27,38] In addition to optimization, DO-MS also facilitates data quality control and experimental standardization with large sample cohorts, especially large-scale single-cell proteomic experiments.[39,40] Here, we demonstrated how DO-MS helps achieve these aims in concrete use cases.

METHODS

Data Acquisition

Apart from the 30 single cells acquired on the timsTOF as part of plexDIA, all samples consist of bulk cellular lysates diluted down to the respective number of single-cell equivalents by assuming a 250 pg of protein per cell. Melanoma cells (WM989-A6-G3, a kind gift from Arjun Raj, University of Pennsylvania), U-937 cells (monocytes), and HPAF-II cells (PDACs, ATCC, CRL-1997) were cultured as previously described by Derks et al.[26]—Methods—Cell culture. Cells were harvested, processed, and labeled with mTRAQ as described by Derks et al.[26]—Methods—Preparation of bulk plexDIA samples.

All bulk data were acquired on a Thermo Fisher Scientific *Q*-Exactive Classic Orbitrap mass spectrometer. Samples of 1 μL volume were injected with the Dionex UltiMate 3000 UHPLC using a 25 cm × 75 μm IonOpticks Aurora Series UHPLC column

(AUR2-25075C18A). Two buffers A and B were used with buffer A made of 0.1% formic acid (Pierce, 85178) in liquid chromatography (LC)–MS-grade water and buffer B made of 80% acetonitrile and 0.1% formic acid mixed with LC–MS-grade water.

Systematic Optimization of Precursor Isolation Windows—A combined sample consisting of one single-cell equivalent PDAC lysate labeled with mTRAQd0, one single-cell equivalent U937 lysate labeled with mTRAQd4, and one single-cell equivalent Melanoma lysate labeled with mTRAQd8 was injected in a volume of 1 μL. LC was performed with 200 nL/min flow rate for 30 min of active gradient starting with 4% Buffer B (min 0–2.5), 4–8% B (min 2.5–3), 8–32% B (min 3–33), 32–95% B (min 33–34), 95% B (min 34–35), 95–4% B (min 35–35.1), and 4% B (min 35.1–53). All acquisition methods had a single MS1 scan covering the range of 380–1400 mz followed by DIA MS2 scans: 2×MS2 starting at 380 mz: 240Th, and 780Th width; 4×MS2 starting at 380 mz: 120Th, 120Th, 200Th, and 580Th width; 6×MS2 starting at 380 mz: 80Th, 80Th, 80Th, 120Th, 240Th, and 420Th width; 8×MS2 starting at 380 mz: 60Th, 60Th, 60Th, 60Th, 100Th, 100Th, 290Th, and 290Th width; 10×MS2 starting at 380 mz: 50Th, 50Th, 50Th, 50Th, 50Th, 75Th, 75Th, 150Th, 150Th, and 320Th width; 12×MS2 starting at 380 mz: 40Th, 40Th, 40Th, 40Th, 40Th, 40Th, 60Th, 60Th, 120Th, 120Th, 210Th, and 210Th width; 16×MS2 starting at 380 mz: 30Th, 30Th, 30Th, 30Th, 30Th, 30Th, 30Th, 30Th, 50Th, 50Th, 50Th, 50Th, 145Th, 145Th, 145Th, and 145Th width. All MS1 and MS2 scans were performed with 70,000 resolving power, 3×10^6 AGC maximum, 300 ms maximum accumulation time, NCE at 27%, a default charge of 2, and RF S-lens was at 80%.

Data-Driven Optimization of Window Placement—A combined sample consisting of 100 single-cell equivalents of PDAC, U937, and Melanoma cells were labeled with mTRAQd0, mTRAQd4, and mTRAQd8, respectively. LC was performed with 200 nL/min flow rate for 30 min of active gradient starting with 4% Buffer B (min 0–2.5), 4–8% B (min 2.5–3), 8–32% B (min 3–33), 32–95% B (min 33–34), 95% B (min 34–35), 95–4% B (min 35–35.1), and 4% B (min 35.1–53). Both MS1 and MS2 scans covered the range of 380–1400 mz with a single MS1 scan and eight MS2 scans. The distribution of precursors was determined based on the DO-MS report using equal-sized windows, starting at 380 mz: 127.5Th, 127.5Th, 127.5Th, 127.5Th, 127.5Th, 127.5Th, 127.5Th, and 127.5Th width. MS2 windows were then distributed to have equal total ion current (TIC) based on the DO-MS output: starting at 380mz: 100Th, 64Th, 61Th, 66Th, 91Th, 100Th, 153Th, and 385Th width. For the equal number of precursors, the original sample was searched with DIA-NN as described, and MS2 windows were distributed to have an equal number of precursors: starting at 380mz: 84Th, 63Th, 49Th, 66Th, 59Th, 101Th, 176Th, and 422Th width. All MS1 and MS2 scans were performed with 70,000 resolving power, 3×10^6 AGC maximum, 251 ms maximum accumulation time, NCE at 27%, a default charge of 2, and RF S-lens was at 80%.

Optimizing the Gradient Profile and Length—A combined sample consisting of 100 single-cell equivalents of PDAC, Melanoma, and U937 were labeled with mTRAQd0, mTRAQd4, and mTRAQd8, respectively. LC was performed with 200 nL/min flow rate starting with 4% Buffer B (min 0–2.5) followed by 4–8% B (min 2.5–3). The active gradient

with 8% buffer B to 32% buffer B stretched across 15, 30, and 60 min followed by a 1 min 32–95% B ramp, 1 min at 95%, and 18 min at 4% B. All acquisition methods had a single MS1 scan covering the range of 478–1500 mz followed by 8 DIA MS2 scans: starting at 380 mz: 60Th, 60Th, 60Th, 60Th, 100Th, 100Th, 290Th, 290Th. All MS1 and MS2 scans were performed with 70,000 resolving power, 3×10^6 AGC maximum, 300 ms maximum accumulation time, NCE at 27%, a default charge of 2, and RF S-lens was at 80%.

Effect of Additional Survey Scans—A 100 single-cell equivalent of each, PDAC, U937, and Melanoma cells were labeled with mTRAQd0, mTRAQd4, and mTRAQd8, respectively, and injected in a volume of 1 μL. LC was performed with 200 nL/min for 30 min of active gradient starting with 4% buffer B (min 0–2.5), 4–8% B (min 2.5–3), 8–32% B (min 3–63), 32–95% B (min 63–64), 95% B (min 64–65), 95–4% B (min 65–65.1), and 4% B (min 65.1–83). A single MS1 scan with the range of 478–1500 mz was followed by MS2 scans starting at 380 mz with 60Th, 60Th, 60Th, 60Th, 100Th, 100Th, 290Th, and 290Th width. For the method with increased MS1 sampling, a second MS1 scan was incorporated after the fourth MS2 scan. All MS1 and MS2 scans were performed with 70,000 resolving power, 3×10^6 AGC maximum, 251 ms maximum accumulation time, NCE at 27%, a default charge of 2, and RF S-lens was at 80%.

Data Analysis

Data were analyzed using DIA-NN 1.8.1, using the 5000 protein group human-only spectral library published previously by Derks et al.[26]—Methods—Spectral library generation. Data were then processed with DO-MS. For preprocessing of Orbitrap data, DO-MS used ThermoRawFileParser 1.4.0 to convert the proprietary raw format to the open mzML standard and Dinosaur 1.2.0 for feature detection. All other preprocessing steps were performed in the Python programming language version 3.10 and made use of its extensive ecosystem for scientific programing including Numpy, Pandas, pymzML, and scikit-learn. All plots were created in DO-MS, which utilized the R programing language version 4.3.1. Figure 5B was created using matplotlib.

Data completeness is shown for all pairwise comparisons in a plex DIA set. It is calculated as the Jaccard index between two sets of identifications A and B given by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

RESULTS

We developed DIA-specific modules of the DO-MS app[31] to enable monitoring and optimization of DIA experiments. The DO-MS v2.0 app consists of two parts: A post-processing step which collects additional metrics on the performance of the acquisition method in use, and an interactive application to visualize the metrics and results reported by DIA search engines, Figure 1. All components are built in a modular way, which allows creation of new visualization modules and extending the input source to other search engines (the default engine is DIA-NN[13]). The base functionality is available for all input formats

compatible with the respective search engine, which includes Thermo Fisher Scientific Orbitrap and Bruker TimsTOF data.

Further, instrument-specific information is collected in a post-processing step, which is only implemented for Thermo Fisher Scientific Orbitrap[42] raw files. However, the user has the flexibility to adapt the method to other vendors, given that they can be converted to the open mzML format[43] using tools like msConvert.[44] The current implementation uses a custom version of the ThermoRawFileParser,[41] which reports additional instrument-specific information like the noise level. It is implemented in Python[45] and can be called from the command line, which allows the search engine to automatically call post-processing after it has finished the search. General metrics like the TIC and the MS1 and MS2 accumulation times are extracted and reported in individual files. Precursor-specific metrics, such as the signal-to-noise level (S/N), are reported based on the search engine results. Peptide-like features are identified using the Dinosaur feature finder.[46] This step is independent of the amino acid sequence identification of a precursor and is only based on the shape of its elution profile and isotopic envelope distribution. The metrics are then visualized in an interactive R shiny[47,48] app, which allows the generation of portable html reports. All metrics shown in this article are accessible with DO-MS, and all figures resemble figures generated with DO-MS unless explicitly noted otherwise. An overview of all metrics available in DO-MS can be found in the Supporting Information, Table S2.

Systematic Optimization of Precursor Isolation Window Placement

In DIA experiments, fragmentation spectra are highly complex due to the parallel fragmentation of multiple precursors. To reduce complexity, the range of precursor masses is distributed across multiple MS2 windows, which need to be designed by the experimenter. While increasing the number of MS2 windows results in less complex spectra, it comes at the expense of an increased duty cycle length. The more MS2 scans are incorporated, the fewer data points are collected across each and every elution peak, impeding identification and optimal quantification. This trade-off needs to be optimized in a context-specific manner, depending on the sample complexity, abundance, choice of chromatography, and gradient length.

DO-MS helps optimize this trade-off by systematically assessing the impact of different parameters with respect to multiple performance metrics at the same time. This is exemplified by a plexDIA experiment consisting of a 3-plex bulk lysate diluted down to the single-cell level, Figure 2. The fastest duty cycle with a single MS1 and two MS2 scans has a duration of approximately 0.9 s, which allows for frequent sampling of the elution profile. This results in a higher chance to sample the elution apex and is reflected in the increased MS1 peak height compared to methods with more MS2 windows, Figure 2A,B. An acquisition method with 16 MS2 scans sample precursors only every 5.1 s and thus may fail to sample the elution peak apex (Supporting Information, Table S1). This becomes evident when the intensity of the same peptide is compared across runs. The median ratio between shared peptides is more than 2-fold lower for a method with more than 12 MS2 windows compared to 2 MS2 windows, Figure 2B. In contrast, optimal sampling of the elution apex requires more frequent sampling, which comes at the cost of fewer

MS2 isolation windows. Indeed, sampling the most intense precursor signal is achieved in our experiment when using only two isolation windows. At the same time, such an acquisition method distributes fragment ions across only two isolation windows, resulting in high co-isolation and reduced proteome coverage. DO-MS allows one to systematically and comprehensively explore this inherent trade-off between proteome coverage and sampling elution peak apexes.

For the chosen chromatography and specimen, the DO-MS report indicates that the largest number of precursors is identified with an acquisition method of 6, 8, or 10 MS2 windows, Figure 2C. Across all three channels, about 10,000 precursors are identified on the MS2 level and quantified on the MS1 level. As we required MS2 information for sequence identification, our identifications did not benefit from the higher temporal resolution of MS1 scans and these identifications cannot exceed the number of MS2 identifications. The results indicate that overall performance balancing quantification and coverage depth is best when using four or six MS2 scans, Figure 2. This trade-off may be mitigated by using multiple MS1 scans per duty cycle,[26,27] and such methods optimized by DO-MS using the metrics are displayed in Figure 2.

Data-Driven Optimization of Window Placement

DO-MS also allows for refinement of the precursor isolation window placement, Figure 3. The MS2 windows can be selected to utilize equal *m/z* ranges[49] or to optimize the distribution of ions across MS2 windows and thereby increase the proteome coverage.[18,50] Recently, even dynamic online optimization has been proposed.[51] The metrics provided by DO-MS allow users to implement previously suggested strategies or develop new ones and to continuously monitor the performance, including metrics which are often not easily accessible.

As the distribution of peptide masses is not uniform across the *m/z* range, equal-sized isolation windows will result in more precursors per window in the lower *m/z* range. Thus, placement of isolation windows across an equal *m/z* range is likely suboptimal, as manifested by lower proteome coverage shown in Figure 3A. One of the reasons for this is the associated suboptimal MS2 accumulation time, which is limited by the capacity of the ion trap. When analyzing a 3-plex experiment of 100 cell equivalent bulk lysate, the lowest *m/z* windows will fill up in a few milliseconds, while windows with higher *m/z* will accumulate ions for the maximum accumulation time of 251 ms, Figure 3B. This leads to complex fragmentation spectra, loss in sensitivity in lower mass ranges, and unused ion capacity in higher *m/z* ranges. The effect of accumulation times on the sensitivity is likewise reflected in the lower coverage of the proteome at the MS1- than at the MS2 level. The wider isolation windows at the MS1 level lead to shorter accumulation times before the maximum ion trap capacity is reached. This limits sensitivity and leads to fewer quantified peptides at the MS1 than the MS2 level (see also the Supporting Information, full DO-MS report).

Windows placed based on an equal TIC per window, determined in a previous experiment, or based on the precursor *m/z* can lead to improved proteome coverage. The metrics available in DO-MS, such as accumulation times, data completeness, and number of

identifications as a function of the false discovery rate (FDR), allow for evaluating different choices of window placement, detecting bottlenecks, and improving them.

Optimizing the Chromatographic Profile and Length

To reduce the complexity of peptide sample mixtures, dimensions of separation including LC or gas phase fractionation like trapped ion mobility spectrometry are used. Separation by LC has been the default separation method for MS proteomics. The improved separation with longer gradients comes at the cost of increased measurement time. DO-MS allows for balancing this trade-off and for performing routine quality control on peptide separation.

Longer LC gradients improve proteome coverage in DIA in two different ways. First, longer gradients lead to better separation of different peptide species reducing coelution of interfering species and improving spectral quality. Second, they lead to elongation of elution profiles, resulting in precursors being sampled for a longer duration. This allows for sampling each ion species less frequently and gives room for more specific isolation, improving spectral quality. Thereby, while identifying fewer peptides per unit time, longer gradients facilitate identifying more peptides per sample. The general trend is shown by the DO-MS output for a 3-plex 100-cell equivalent bulk dilution analyzed with 15, 30, and 60 min of the active gradient using the same duty cycle, Figure 4. One benefit of the longer gradients can be seen when the ion accumulation time of the Orbitrap instrument is plotted as a function of the retention time, Figure 4A. Longer gradients distribute the analytes and lead to a longer accumulation of ions before the maximum capacity is reached. Individual spectra therefore contain fewer ion species and sample sufficient ions even from low abundant peptides. This improves not only the absolute numbers of identifications but also the fraction of precursors quantified at the MS1 level, Figure 4B.

DO-MS also allows for optimizing the slope and profile of the gradient to evenly distribute ions across a gradient while keeping its duration constant. Depending on the sample, peptides might not elute evenly across the gradient. This information becomes accessible in three different ways. DO-MS reports the accumulation time of the ion trap (Figure 4A), peptide identifications across the gradient (Figure 4C), and peptide-like features or potential contaminants assembled by Dinosaur across the gradient (Figure 4D).

Having access to gradient-specific parameters facilitates effective quality control and problem identification. Identified MS1 features provide useful information for ion clusters not assigned to a peptide sequence including singly charged species and peptide-like ions not mapped to a sequence, Figure 4D. This can be useful to identify contaminants[31] and estimate the ions accessible to MS analysis that may be interpreted by improved algorithms.[8,52] The binned TIC output allows for identifying errors in the method setup and gives a quick overview of the sampled mass range, Figure 4E.

Improving Sampling Using Additional Survey Scans

The conflict between reducing spectral complexity and increasing the number of data points per peak mentioned in Figure 2 can be partially alleviated by increasing the number of survey scans.[27] When duty cycles are long, more frequent sampling on the MS1 level can increase the fraction of precursors with MS1 information and the probability of sampling

close to the elution apex.[19,26] The DO-MS framework can be used to assess the contribution of such additional MS1 scans for improving precursor sampling.

The effect can be exemplified based on a 3-plexDIA set whose samples correspond to 100 cells per channel analyzed, analyzed with 60 min of active gradient. A method with a single survey scan is compared to a method with two survey scans evenly distributed between the eight MS2 scans, Figure 5A. The additional survey scan increases the duty cycle length only marginally while increasing the frequency of precursor sampling almost 2-fold. Thus, the adapted method increases the probability that precursors are sampled close to their elution apex and that peptides with a shorter elution profile and potentially lower intensity can be quantified on the MS1 level, which would be otherwise missed. These expectations are supported by the results shown in Figure 5B-D.

More survey scans lead to almost doubling the number of identified peptide-like features, with the increase being particularly pronounced for features with short elution lengths, Figure 5B. The improvements also result in higher MS1 intensity estimates by the search engine for intersected precursors since more precursors are sampled close to their apexes. Furthermore, a larger fraction of precursors is quantified at the MS1 level, Figure 5C,D. These improvements are observed without associated negative effects due to the longer overall duty cycle. These results indicate that the duty cycle with two MS1 survey scans outperforms the one with a single MS1 survey scan.

Quality Control for Routine Sample Acquisition

When acquiring large datasets, it is important to continuously monitor the performance of the acquisition method and identify potential failed experiments.[37] This monitoring for plexDIA experiments should include metrics for each labeled sample, i.e., channel-level metrics.

DO-MS provides a convenient way to perform such quality control, exemplified by the single-cell plexDIA set by Derks et al.,[26] as shown in Figure 6. Using nPOP sample preparation,[53] 10 sets with 3 single cells each were prepared and measured on a timsTOF instrument, resulting in about 1000 quantified proteins per single cell on average, Figure 6A. As plexDIA can benefit from translating precursor identifications between channels,[26,27] the impact of translation on identifications and data completeness is reported by DO-MS. With single cells, it is vital to identify potential dropouts where sample preparation might have failed and exclude them from processing. One useful metric for this is the precursor intensity distribution for every single cell, which is displayed by DO-MS, Figure 6B. Another metric to assess the single-cell proteome quality is the quantification variability between peptides originating from the same protein, which has been proposed as a metric for single-proteome quality,[54] Figure 6C. In this dataset, the cells in channel Δ0, set 06, and Δ8, set 10, show both a lower number of proteins before translation and a higher quantification variability and should potentially be excluded from further analysis.

Conclusions

The DO-MS framework provides a systematic approach to benchmarking, optimizing, and reporting results from label-free and multiplexed DIA-MS. We exemplified how key method parameters such as the number of precursor scans or isolation window placement can be benchmarked and optimized. DO-MS aims to foster understanding from first-principles calculations, considering fundamental trade-offs such as spectral complexity and sampling frequency. By adopting this approach, it becomes possible to design methods tailored to specific application needs, such as emphasizing data completeness, quantitative accuracy, or proteome depth. DO-MS should enable broader adoption of cutting-edge methods, such as DIA and plexDIA methods for driving biological research.[55]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Luke Khoury for support with sample processing and acquisition and Jason Derks for sample preparation. The work was funded by an Allen Distinguished Investigator award through The Paul G. Allen Frontiers Group to N.S., a Seed Networks Award from CZI CZF2019-002424 to N.S., an NIGMS award R01GM144967 to N.S., and an NCI award UG3CA268117 to N.S.

REFERENCES

(1). MacCoss MJ; Alfaro JA; Faivre DA; Wu CC; Wanunu M; Slavov N Sampling the proteome by emerging single-molecule and mass spectrometry methods. Nat. Methods 2023, 20, 339–346. [PubMed: 36899164]

(2). Eng JK; McCormack AL; Yates JR An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom 1994, 5, 976–989. [PubMed: 24226387]

(3). Huffman RG; et al. Prioritized mass spectrometry increases the depth, sensitivity and data completeness of single-cell proteomics. Nat. Methods 2023, 20, 714. [PubMed: 37012480]

(4). Nikolai S. Extending the sensitivity, consistency and depth of single-cell proteomics. Nat. Methods 2023, 20, 649–650. [PubMed: 37012482]

(5). Venable JD; Dong M-Q; Wohlschlegel J; Dillin A; Yates JR Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Methods 2004, 1, 39–45. [PubMed: 15782151]

(6). Dong M-Q; Venable JD; Au N; Xu T; Park SK; Cociorva D; Johnson JR; Dillin A; Yates JR Quantitative Mass Spectrometry Identifies Insulin Signaling Targets in C. elegans. Science 2007, 317, 660–663. [PubMed: 17673661]

(7). Ludwig C.; et al. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol. Syst. Biol 2018, 14, No. e8126. [PubMed: 30104418]

(8). Slavov N. Driving Single Cell Proteomics Forward with Innovation. J. Proteome Res 2021, 20, 4915–4918. [PubMed: 34597050]

(9). Slavov N. Increasing proteomics throughput. Nat. Biotechnol 2021, 39, 809–810. [PubMed: 33767394]

(10). Tsou C-C; Avtonomov D; Larsen B; Tucholska M; Choi H; Gingras AC; Nesvizhskii AI DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods 2015, 12, 258–264. [PubMed: 25599550]

(11). Bruderer R; Bernhardt OM; Gandhi T; Miladinović SM; Cheng LY; Messner S; Ehrenberger T; Zanotelli V; Butscheid Y; Escher C; et al. Extending the Limits of Quantitative Proteome

Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. Mol. Cell. Proteomics 2015, 14, 1400–1410. [PubMed: 25724911]

(12). Egertson JD; MacLean B;Johnson R; Xuan Y; MacCoss MJMultiplexed peptide analysis using data-independent acquisition and Skyline. Nat. Protoc 2015, 10, 887–903. [PubMed: 25996789]

(13). Demichev V; Messner CB; Vernardis SI; Lilley KS; Ralser M DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat. Methods 2020, 17, 41–44. [PubMed: 31768060]

(14). Sinitcyn P; Hamzeiy H; Salinas Soto F; Itzhak D; McCarthy F; Wichmann C; Steger M; Ohmayer U; Distler U; Kaspar-Schoenefeld S; et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. Nat. Biotechnol 2021, 39, 1563–1573. [PubMed: 34239088]

(15). Cox J. Prediction of peptide mass spectral libraries with machine learning. Nat. Biotechnol 2023, 41, 33–43. [PubMed: 36008611]

(16). Distler U.; et al. midiaPASEF Maximizes Information Content in Data-independent Acquisition Proteomics. bioRxiv 2023, DOI: 10.1101/2023.01.30.526204.

(17). Szyrwiel L; Sinn L; Ralser M; Demichev V Slice-PASEF: fragmenting all ions for maximum sensitivity in proteomics. bioRxiv 2022, DOI: 10.1101/2022.10.31.514544.

(18). Skowronek P; Thielert M; Voytik E; Tanzer MC; Hansen FM; Willems S; Karayel O; Brunner AD; Meier F; Mann M Rapid and In-Depth Coverage of the (Phospho-)Proteome With Deep Libraries and Optimal Window Design for dia-PASEF. Mol. Cell. Proteomics 2022, 21, 100279. [PubMed: 35944843]

(19). Xuan Y; Bateman NW; Gallien S; Goetze S; Zhou Y; Navarro P; Hu M; Parikh N; Hood BL; Conrads KA; et al. Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies. Nat. Commun 2020, 11, 5248. [PubMed: 33067419]

(20). Demichev V; Szyrwiel L; Yu F; Teo GC; Rosenberger G; Niewienda A; Ludwig D; Decker J; Kaspar-Schoenefeld S; Lilley KS; et al. dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. Nat. Commun 2022, 13, 3944. [PubMed: 35803928]

(21). Li Y.; et al. An integrated strategy for mass spectrometry-based multiomics analysis of single cells. Anal. Chem 2021, 93, 14059–14067. [PubMed: 34643370]

(22). Gebreyesus ST; Siyal AA; Kitata RB; Chen ESW; Enkhbayar B; Angata T; Lin KI; Chen YJ; Tu HL Streamlined single-cell proteomics by an integrated microfluidic chip and data-independent acquisition mass spectrometry. Nat. Commun 2022, 13, 37. [PubMed: 35013269]

(23). Brunner A-D; et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. Mol. Syst. Biol 2022, 18, No. e10798. [PubMed: 35226415]

(24). Phlairaharn T; Grégoire S; Woltereck LR; Petrosius V; Furtwängler B; Searle BC; Schoof EM High Sensitivity Limited Material Proteomics Empowered by Data-Independent Acquisition on Linear Ion Traps. J. Proteome Res 2022, 21, 2815–2826. [PubMed: 36287219]

(25). Slavov N Learning from natural variation across the proteomes of single cells. PLoS Biol. 2022, 20, No. e3001512. [PubMed: 34986167]

(26). Derks J; Leduc A; Wallmann G; Huffman RG; Willetts M; Khan S; Specht H; Ralser M; Demichev V; Slavov N Increasing the throughput of sensitive proteomics by plexDIA. Nat. Biotechnol 2022, 41, 50–59. [PubMed: 35835881]

(27). Derks J; Slavov N Strategies for increasing the depth and throughput of protein analysis by plexDIA. J. Proteome Res 2023, 22, 697–705. [PubMed: 36735898]

(28). Singh A. Sensitive protein analysis with plexDIA. Nat. Methods 2022, 19, 1032. [PubMed: 36068318]

(29). Nikolai S. Framework for multiplicative scaling of single-cell proteomics. Nat. Biotechnol 2022, 41, 23.

(30). Ludwig C; Gillet L; et al. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol. Syst. Biol 2018, 14, No. e8126. [PubMed: 30104418]

(31). Huffman G; Chen AT; Specht H; Slavov N DO-MS: Data-Driven Optimization of Mass Spectrometry Methods. J. Proteome Res 2019, 18, 2493–2500. [PubMed: 31081635]

(32). Bittremieux W; Valkenborg D; Martens L; Laukens K Computational quality control tools for mass spectrometry proteomics. Proteonomics 2017, 17, 1600159.

(33). Trachsel C; Panse C; et al. rawDiag: An R Package Supporting Rational LC–MS Method Optimization for Bottom-up Proteomics. J. Proteome Res 2018, 17, 2908–2914. [PubMed: 29978702]

(34). Bielow C; Mastrobuoni G; Kempa S Proteomics Quality Control: Quality Control Software for MaxQuant Results. J. Proteome Res 2016, 15, 777–787. [PubMed: 26653327]

(35). Soneson C; Iesmantavicius V; Hess D; Stadler MB; Seebacher J einprot: flexible, easy-to-use, reproducible workflows for statistical analysis of quantitative proteomics data. bioRxiv 2023, DOI: 10.1101/2023.07.27.550821.

(36). Slavov N. Single-cell protein analysis by mass spectrometry. Curr. Opin. Chem. Biol 2021, 60, 1–9. [PubMed: 32599342]

(37). Gatto L; Aebersold R; et al. Initial recommendations for performing, benchmarking, and reporting single-cell proteomics experiments. Nat. Methods 2023, 20, 375–386. [PubMed: 36864200]

(38). Specht H; Slavov N Optimizing Accuracy and Depth of Protein Quantification in Experiments Using Isobaric Carriers. J. Proteome Res 2021, 20, 880–887. [PubMed: 33190502]

(39). Petelski AA; Emmott E; et al. Multiplexed single-cell proteomics using SCoPE2. Nat. Protoc 2021, 16, 5398–5425. [PubMed: 34716448]

(40). Slavov N. Scaling Up Single-Cell Proteomics. Mol. Cell. Proteomics 2022, 21, 100179. [PubMed: 34808355]

(41). Hulstaert N; Shofstahl J; et al. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. J. Proteome Res 2020, 19, 537–542. [PubMed: 31755270]

(42). Zubarev RA; Makarov A Orbitrap mass spectrometry. Anal. Chem 2013, 85, 5288–5296. [PubMed: 23590404]

(43). Martens L; Chambers M; et al. mzML a community standard for mass spectrometry data. Mol. Cell. Proteomics 2011, 10, R110.000133.

(44). Adusumilli R; Mallick P Proteomics. In Methods and Protocols; Comai L, Katz JE, Mallick P, Eds.; Springer New York: New York, NY, 2017, pp 339–368.

(45). Van Rossum G Python Tutorial. Technical Report CS-R9526, Entrum Voor Wiskunde En Informatica; CWI: Amsterdam, 1995.

(46). Teleman J; Chawade A; Sandin M; Levander F; Malmström J Dinosaur: A Refined Open-Source Peptide MS Feature Detector. J. Proteome Res 2016, 15, 2143–2151. [PubMed: 27224449]

(47). R Core Team. R A Language and Environment for Statistical Computing R Foundation for Statistical Computing: Vienna, Austria, 2022. https://www.R-project.org. Access date Aug 2023.

(48). Chang W.; et al. Shiny: Web Application Framework for R package version 1.7.2.9000, 2022. https://shiny.rstudio.com. Access date Aug 2023.

(49). Gillet LC; Navarro P; et al. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Mol. Cell. Proteomics 2012, 11, O111.016717.

(50). Kawashima Y; Watanabe E; et al. Optimization of Data-Independent Acquisition Mass Spectrometry for Deep and Highly Sensitive Proteomic Analysis. Int. J. Mol. Sci 2019, 20, 5932–0067. [PubMed: 31779068]

(51). Heil LR; Remes PM; Canterbury JD; Yip P; Barshop WD; Wu CC; MacCoss MJ; et al. Dynamic Data Independent Acquisition Mass Spectrometry with Real-Time Retrospective Alignment. Analytical Chemistry 2023, 95, 11854–11858. [PubMed: 37527417]

(52). Chen AT; Franks A; Slavov N DART-ID increases single-cell proteome coverage. PLoS Comput. Biol 2019, 15, 10070822–e1007130.

(53). Leduc A; Huffman RG; Cantlon J; Khan S; Slavov N Exploring functional protein covariation across single cells using nPOP. Genome Biol. 2022, 23, 261. [PubMed: 36527135]

(54). Specht H; Emmott E; et al. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. Genome Biol. 2021, 22, 50. [PubMed: 33504367]

(55). Slavov N. Single-cell proteomics: quantifying post-transcriptional regulation during development with mass-spectrometry. Development 2023, 150, dev201492. [PubMed: 37387573]

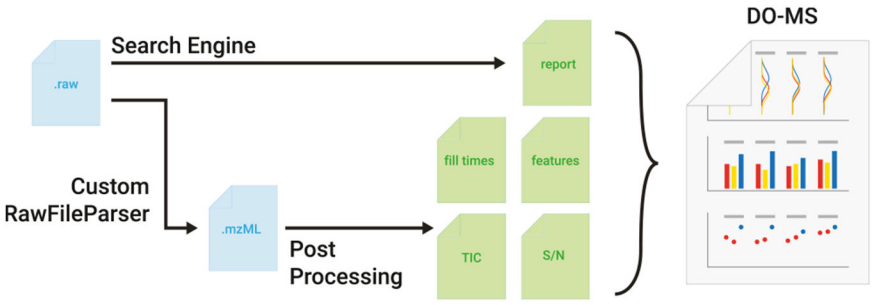


Figure 1.

Schematic of the DO-MS pipeline version 2.0. A schematic of the processing and intermediate steps of the updated DO-MS pipeline. Input files (blue) in the raw format are searched by a search engine (the default one is DIA-NN[13]) and converted to mzML using a custom version of the ThermoRawFile parser.[41] The search report from DIA-NN and the mzML are then used by the post-processing step to analyze and display data about MS1 and MS2 accumulation times, TIC information, precursor-wise signal-to-noise levels, and MS1 features.

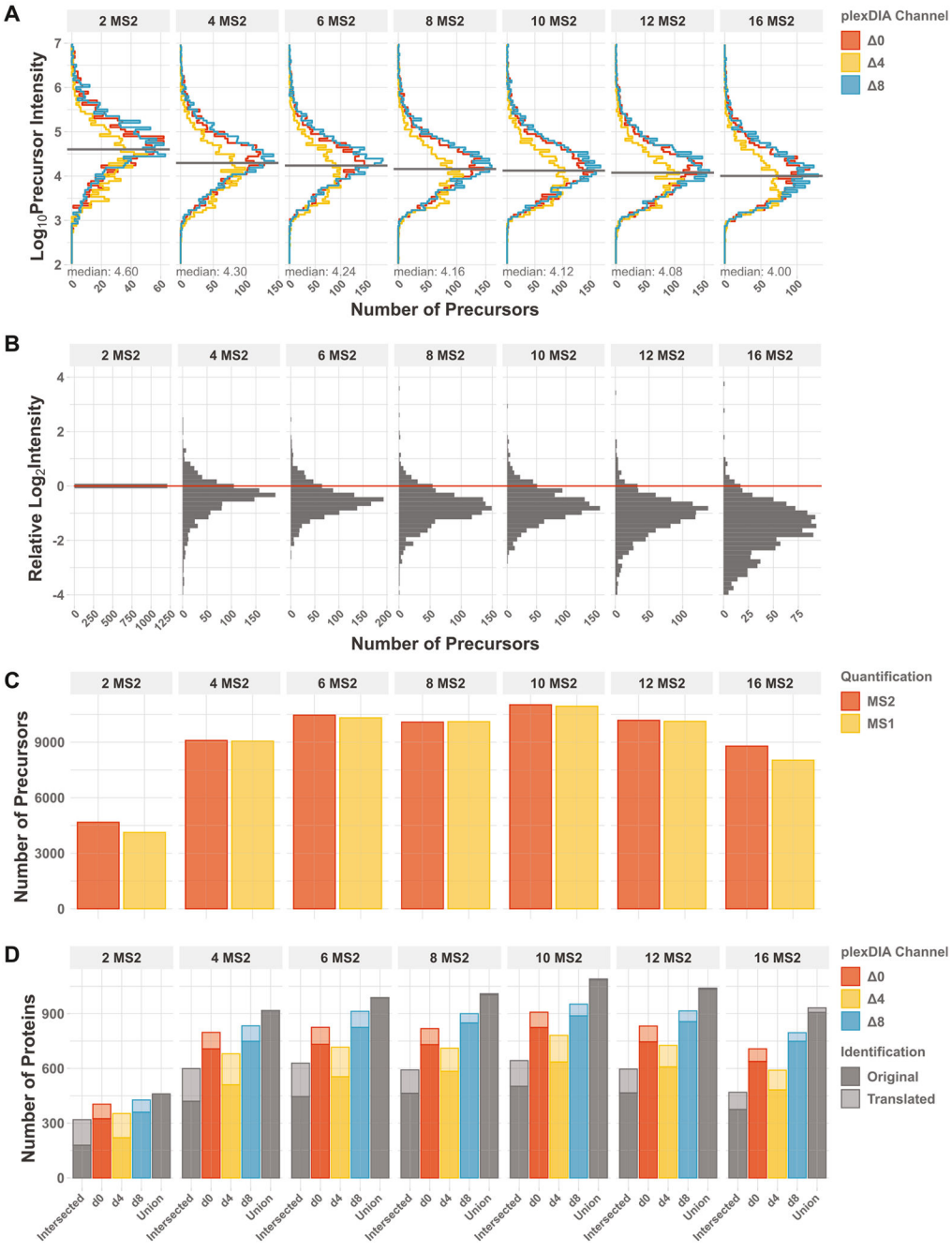


Figure 2.

Optimizing the number of MS2 windows in the duty cycle of plexDIA methods. Example DO-MS output for a plexDIA experiment using 3-plex bulk lysate diluted down to the single-cell level with different numbers of MS2 windows. All intensities were extracted as peak heights. (A) Histogram of precursor (MS1) intensities for each plexDIA channel shown separately. (B) Distributions of ratios between precursor intensities for precursors identified across all conditions. All ratios are displayed on the log2 scale relative to the first condition. (C) The total number of identified precursors per run is shown. Numbers are shown for precursors with MS1 (yellow) and MS2 (red) level quantification. (D) The number of protein

identifications in a plexDIA set is shown for each non-isobarically labeled sample (channel). Proteins shared across all three sets and the entirety of all proteins across sets is shown in gray. Identifications which were propagated within the set are highlighted with lighter colors.

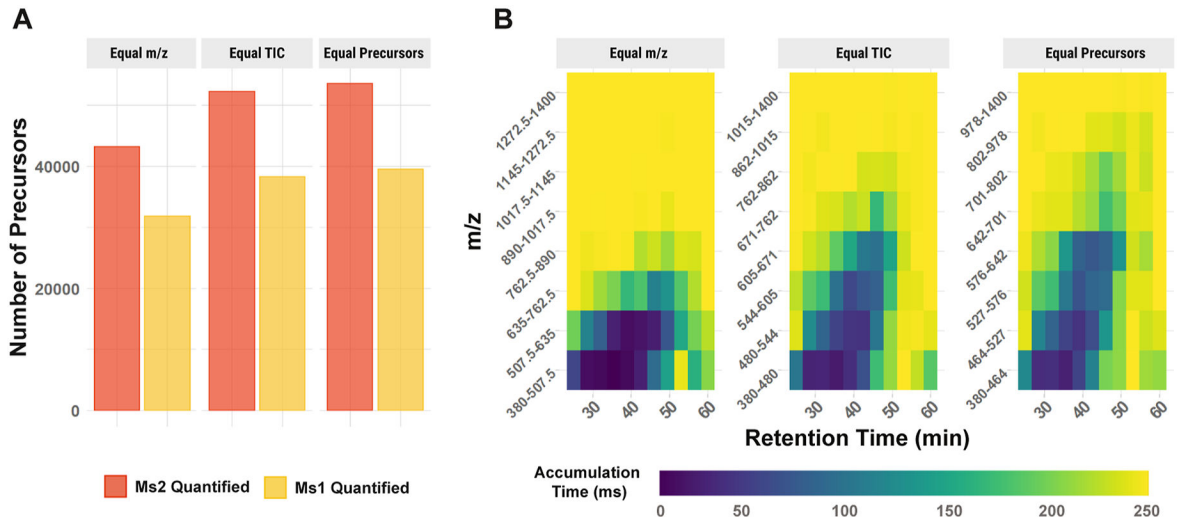


Figure 3.

Optimizing MS2 window placement A 3-plex experiment of 100 cell equivalent bulk lysate was analyzed with eight MS2 windows whose ranges were chosen to achieve equal distribution of (i) *m/z* range, (ii) ion current per window, or (iii) number of precursors. (A) Total number of precursors identified on the MS2 level and quantified on the MS1 level is shown for the three different strategies. (B) The average MS2 accumulation time is shown for every MS2 window across the retention time.

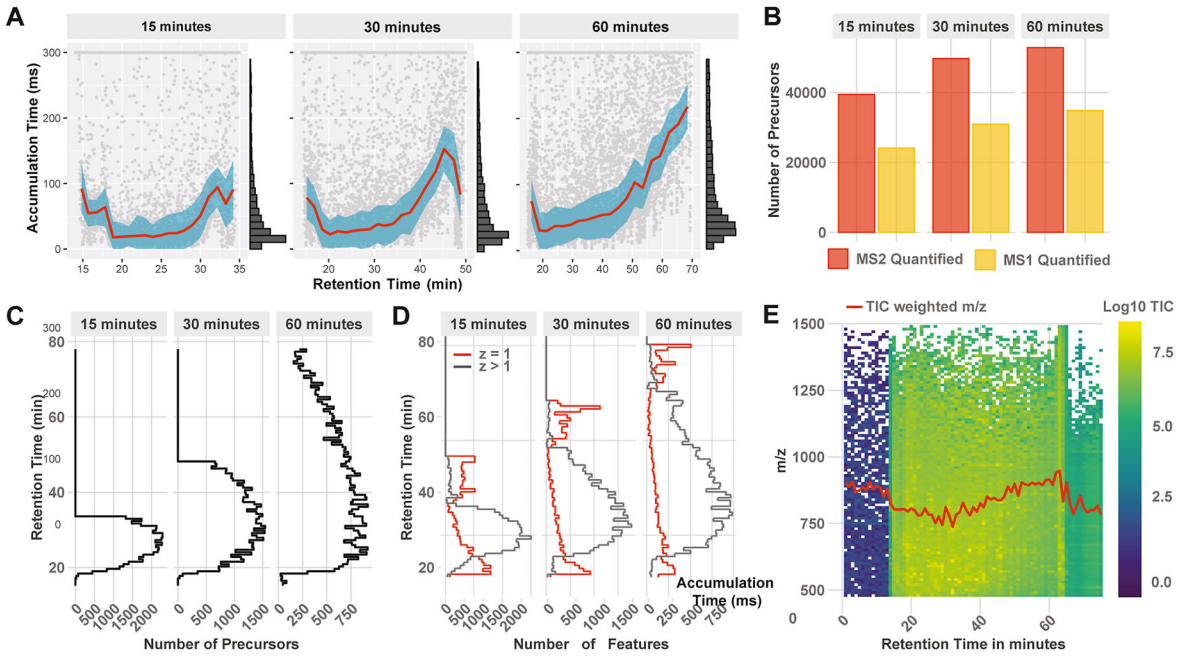


Figure 4.

Optimizing the gradient profile and length. DO-MS allows for optimizing the LC gradient of experiments based on metrics, capturing the whole LC–MS workflow. (A) Distribution of MS1 accumulation times across the LC gradient. (B) Number of quantified precursors in relation to the gradient length. (C) Number of identified precursors by the search engine across the gradients and (D) ion features identified by Dinosaur. (E) Ion map displaying the TIC and mean *m/z* (red curve) as a function of the retention time. All data are from 100× 3-plexDIA samples as described in the methods.

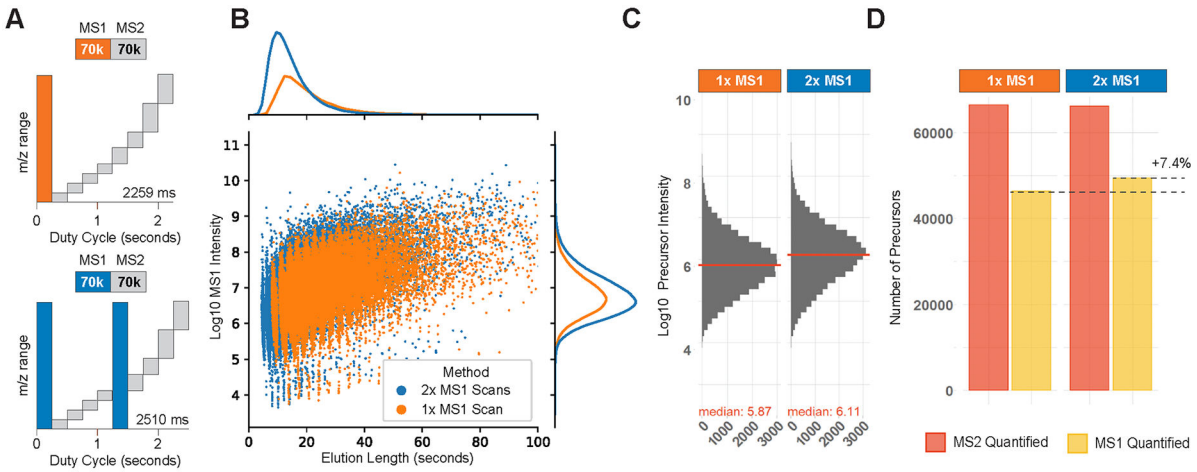


Figure 5.

Effect of additional survey scans per duty cycle. Data acquisition methods can employ multiple survey scans to improve precursor sampling and reduce the stochastic sampling effect. (A) Diagrams of a duty cycle with a single survey scan (orange) and a duty cycle with two survey scans (blue). (B) All peptide-like features identified by Dinosaur[46] are displayed with their elution length at the base and MS1 intensity. The associated marginal distributions are shown. The additional survey scan allows for detecting many additional peptide-like features with a shorter elution profile. (C) MS1 intensity of intersected precursors is increased upon introduction of an additional survey scan. (D) Fraction of MS1 quantified precursors is increased with additional survey scans while maintaining the total number of identifications, independent of the slightly increased duty cycle time. The data shows a 100-cell equivalent 3-plex dataset acquired on 60 min active gradient as described in the methods. Panel B was plotted outside of DO-MS using the peptide-like feature information as stated in the methods.

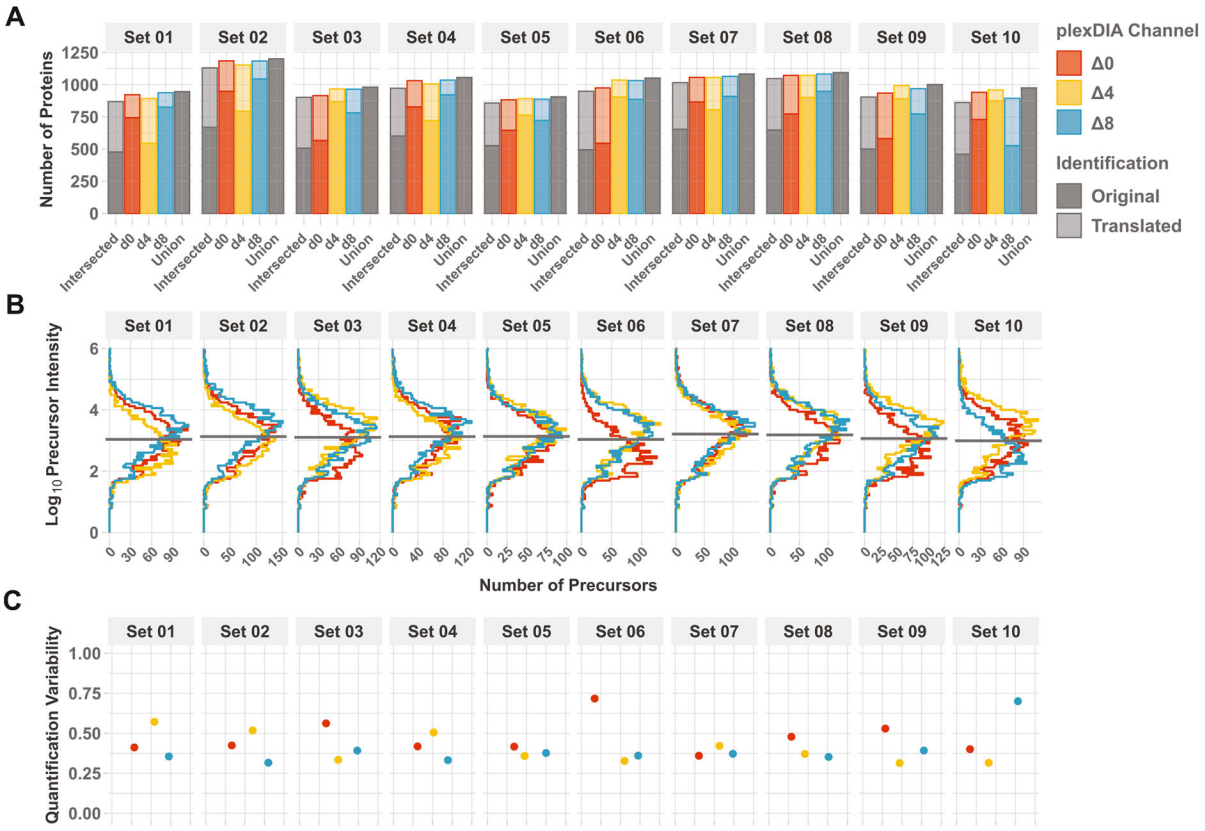


Figure 6.

Routine quality control when acquiring data from a large number of single cells. DO-MS can be used to get a quick overview of the quality of the processing results. (A) Number of protein identifications per single cell before and after translating identifications between channels. Only identifications quantified on the MS1 level are shown. (B) Channel-wise intensity distribution of identified precursors. (C) Quantification variability calculated as the coefficient of variation between peptides of the same protein. The report was generated from the data published by Derks et al.[26] for 10 single-cell 3-plex sets analyzed on a timsTOF instrument.