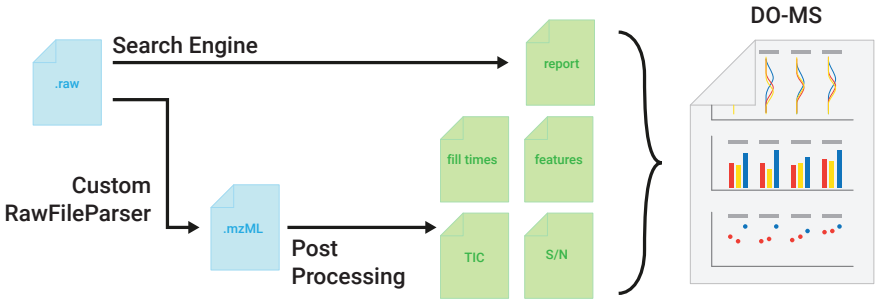
bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Georg Wallmann,[1] Andrew Leduc,[1] & Nikolai Slavov[1,✉]

[1]Departments of Bioengineering, Biology, Chemistry and Chemical Biology, Single Cell Proteomics Center, and Barnett Institute, Northeastern University, Boston, MA 02115, USA
✉ Correspondence: nslavov@northeastern.edu
∈ Data, code & protocols: do-ms.slavovlab.net

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, we extended the DO-MS app (do-ms.slavovlab.net) to optimize and evaluate results from data independent acquisition (DIA) MS. The extension works with both label free and multiplexed DIA (plexDIA) and supports optimizations particularly relevant for 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 quality figures, that can be easily shared. The source code is available at: github.com/SlavovLab/DO-MS.



bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Mass spectrometry (MS) allows for comprehensive quantification and sequence identification of proteins from complex biological samples. Reliable sequence identification of peptides by MS relies on the fragmentation of peptides[1], which can be both performed successively as well as in parallel. Parallel fragmentation in the form of data independent acquisition (DIA) systematically selects groups of precursors for fragmentation which cover the whole m/z range[2,3]. This parallel analysis of multiple precursors can have many benefits, including: (1) consistent collection of data from all detectable peptides[4], (2) high sensitivity due to long ion accumulation times[5], and (3) high-throughput due to the parallel data acquisition[6]. At the same time, 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[7–11]. Many current approaches use in-silico predicted peptide properties (libraries),[12] which removes the overhead of experimentally generated libraries. The improvements have allowed DIA to achieve high proteome depth, data completeness, reproducibility and throughput[13,14]. This has allowed for the quantitative analysis of proteomes down to the single-cell level[15–18]. Throughput can be further increased when labeling samples with non-isobaric mass tags and analyzing them with the plexDIA framework[19–21].

To further empower these emerging capabilities, we sought to extend the 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 per duty cycle and the number, width and placement of precursor ion isolation windows. These parameters must be simultaneously optimized for multiple objectives, including throughput, sensitivity and coverage. Defining the optimal acquisition method is therefore a multi-objective, multi-parameter optimization[22,23]. Numerous tools can be used to facilitate such optimization[24–26], and DO-MS v2.0 extends a distribution-based

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data-driven approach to specific diagnosis of analytical bottlenecks[23].

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[20,27] and ion accumulation times, which are useful for optimizing single-cell analysis[28,29], particularly when using isobaric and isotopologous carriers[20,30]. In addition to optimization, DO-MS also facilities data quality control and experimental standardization with large sample cohorts, especially large scale single-cell proteomic experiments[31,32]. Here we demonstrated how DO-MS helps achieve these aims in concrete use cases.

Results

We developed DIA specific modules of the DO-MS app[23] and updated it version 2.0 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 method and an interactive application to visualize the metrics and results reported by DIA search engines, Fig. 1. All components are built in a modular way, which allows creating new visualization modules and to extend the input source to other search engines (the default engine is DIA-NN[10]).

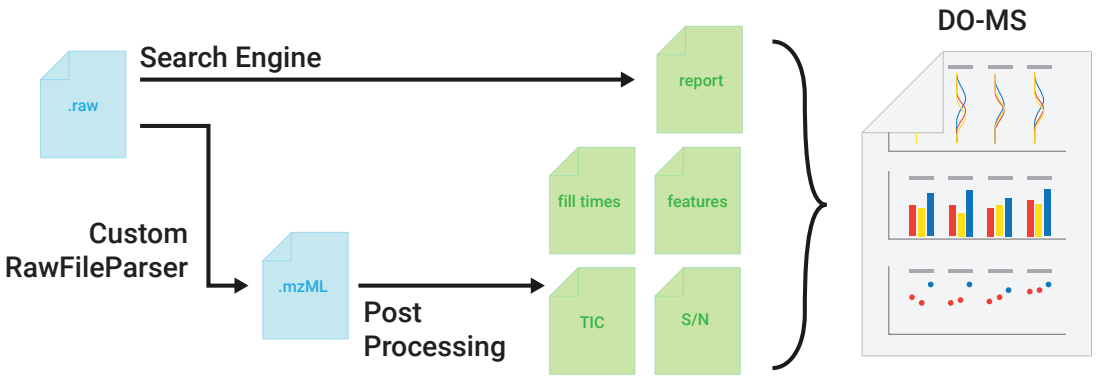


Figure 1 | Schematic of the DO-MS pipeline version 2.0. A schematic of the processing and intermediate steps of the updated DO-MS pipeline. Input files (blue) in the raw format are searched by a search engine (the default one is DIA-NN[10]) and converted to mzML using a custom version of the ThermoRawFile parser[33]. 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, total ion current (TIC) information, precursor-wise signal to noise levels and MS1 features.

The post-processing step is currently implemented for Thermo Fisher Scientific Orbitrap[34] raw

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

files but can be generalized to all files that can be converted to mzML format[35]. The current implementation uses a custom version of the ThermoRawFileParser[33] to convert the vendor specific file format to the open mzML format[35]. It is implemented in Python[36] 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 total ion current (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 and peptide like features are extracted using the Dinosaur feature finder[37]. The metrics are then visualized in an interactive R shiny[38,39] app which allows the generate 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.

Systematic Optimization of Precursor Isolation Window Placement

In DIA experiments fragmentation spectra are highly complex due to 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 for a plexDIA experiment consisting of a 3-plex bulk lysate diluted down to the single cell level, Fig. 2. The fastest duty cycle with a single MS1 and two MS2 scans has a duration of approx 0.9 seconds 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, Fig. 2A,B. An acquisition method with 16 MS2 scans samples precursors less frequently and thus may fail to sample the elution peak apex. As a result, we observed a more than

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

two-fold lower median intersected precursor intensity, Fig. 2B. Optimal sampling of the elution apex requires more frequent sampling, which comes at the cost of fewer MS2 isolation windows. Indeed, the best precursor sampling in our experiments is achieved when using only 2 isolation windows, which distributed fragment ions across only two isolation windows, resulted in high co-isolation, and reduced proteome coverage. DO-MS allows 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 Fig. 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 this identifications cannot exceed the number of MS2 identifications. The results indicate the that overall performance balancing quantification and coverage depth is best when using 4 or 6 MS2 scans, Fig. 2. This trade-off may be mitigated by using multiple MS1 scans per duty cycle[19,20], and such methods optimized by DO-MS using the metrics displayed in Fig. 2.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

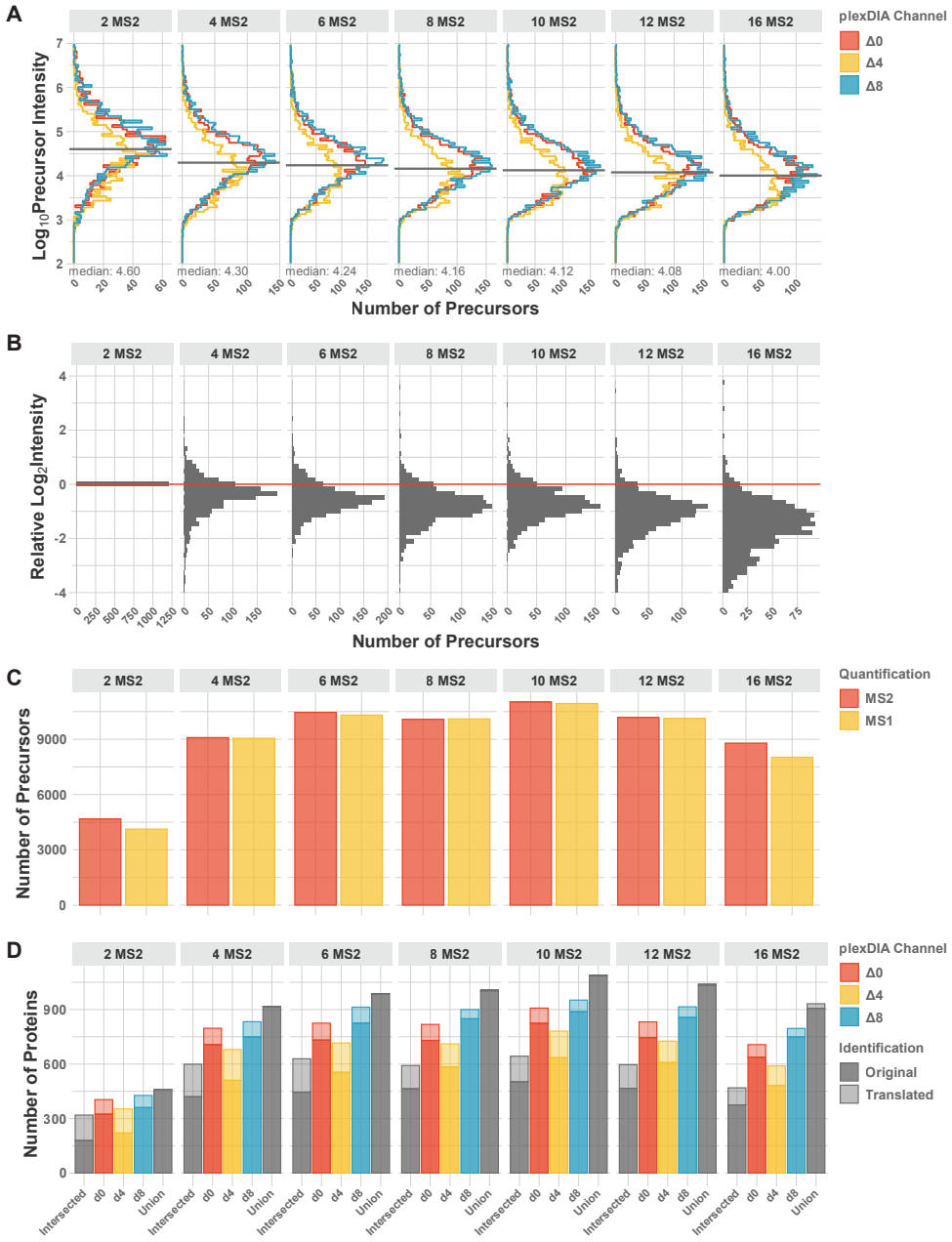


Figure 2 | Optimizing the number of MS2 windows in the duty cycle of plexDIA methods. Example DO-MS output for a plexDIA experiment using 3-plex bulk lysate diluted down to the single-cell level with different numbers of MS2 windows. All intensities were extracted as peak heights. **A** Histogram of precursor (MS1) intensities for each plexDIA channel shown separately. **B** Distributions pf ratios between precursor intensities for precursors identified across all conditions. All ratios are displayed on Log2 scale relative to the first condition. **C** Total number of identified precursors per run, which were quantified on the MS1 and MS2 level. **D** Number of protein identifications per channel which, were identified without or with translation (sequence propagation) within a plexDIA set. The number of proteins shared across all channels (intersected) and for the whole run (union) are also shown for every run.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data Driven Optimization of Window Placement

DO-MS also allows for refinement of the precursor isolation window placement, Fig. 3. The MS2 windows cam be selected to encamps equal m/z range[40] or to optimize the distribution of ions across MS2 windows and thereby increase the proteome coverage[26,41]. Recently, even dynamic online optimization has been proposed[42]. 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.

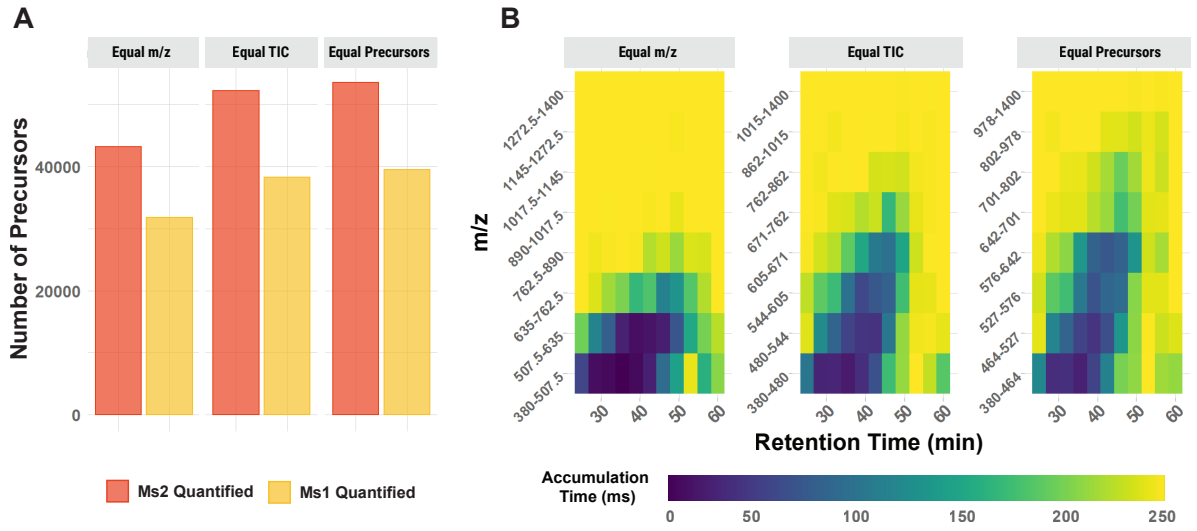


Figure 3 | Optimizing MS2 window placement A 3-plex experiment of 100 cell equivalent bulk lysate was analysed with 8 MS2 windows whose ranges were chosen to achieve equal distribution of (i) m/z range, (ii) ion current per window or (iii) number of precursors. **A** The total number of precursors identified 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.

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 window across an equal m/z range is likely suboptimal, as manifested by lower proteome coverage shown in Fig. 3A. One of the reasons for this are the associated suboptimal MS2 accumulation times, which are limited by the capacity of the ion trap. When analysing 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 250 ms, Fig. 3B. This leads to complex fragment spectra, loss in sensitivity in lower mass

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ranges and unused ion capacity in higher m/z ranges.

Placement based on an equal total ion current (TIC) per window, determined in a previous experiment, or based on the precursors m/z in a library can lead to improved proteome coverage. The metrics available in DO-MS, such as accumulation times, data completeness and identification confidence, allow evaluating different choices of window placement, detecting bottlenecks and improving them.

Optimizing Chromatographic Profile and Length

To reduce the complexity of peptide sample mixtures, dimensions of separation including liquid chromatography or gas phase fractionation like trapped ion mobility spectrometry are used. Separation by liquid chromatography 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 to balance this trade-off and to perform routine quality control on peptide separation.

While identifying fewer peptides per unit time, longer gradients facilitate identifying more peptides per sample. This general trend is shown by the DO-MS output for a 3-plex 100-cell equivalent bulk dilution analyzed with 15, 30 and 60 minutes of active gradient using the same duty cycle, Fig. 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, Fig. 4A. Longer gradients distribute the analytes and lead to longer accumulation of ions, before the maximum capacity is reached. Individual spectra therefore contain fewer ion species and sample sufficient ions even from lowly abundant peptides. This improves not only the absolute number of identification but also the fraction of precursors quantified at MS1 level, Fig. 4B.

DO-MS also allows to optimize 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 (Fig. 4A), peptide identifications across the gradient (Fig. 4C), and peptide like features and potential contaminants across the gradient (Fig. 4D).

Having access to gradient specific parameters is facilitates effective quality control and problem

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

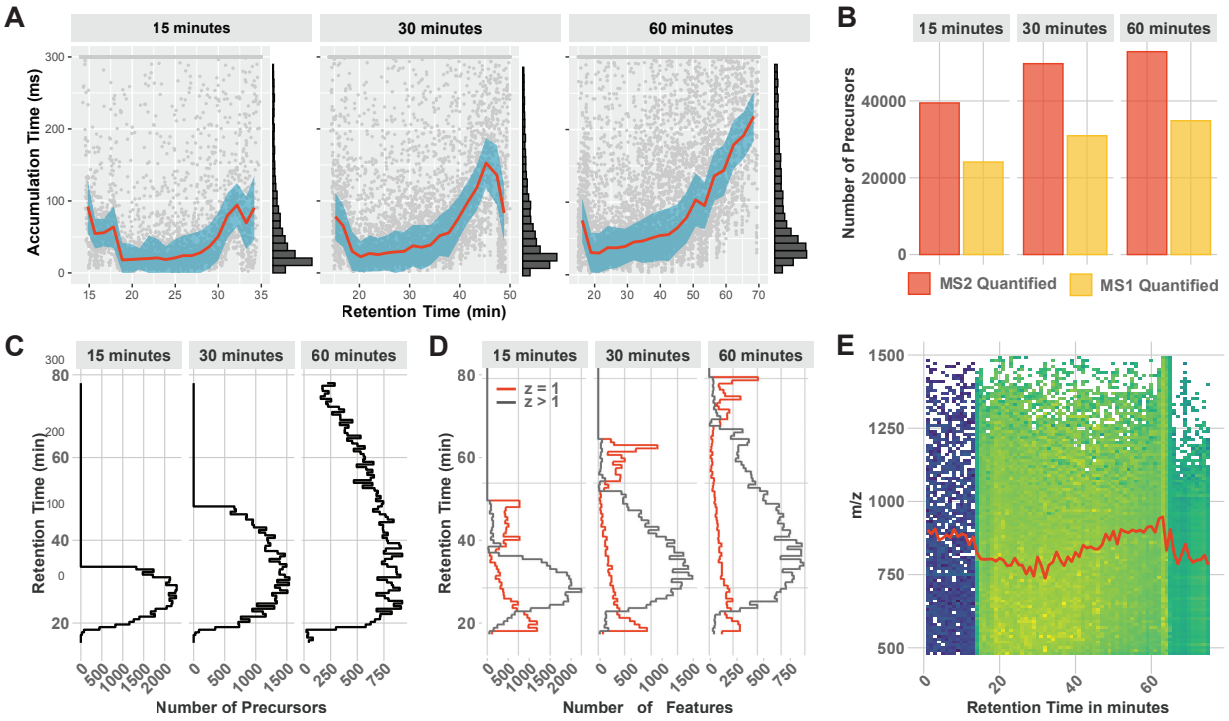


Figure 4 | Optimizing gradient profile and length DO-MS allows to optimize the LC gradient of experiments based on metrics capturing the whole LC-MS workflow. **A** The distribution of MS1 accumulation times across the LC gradient. **B** Number of quantified precursors in relation to the gradient length. **C** Number of identified precursors by the search engine across the gradients and **D** ion features identified by Dinosaur. **E** Ion map desiplaying the total ion current and mean m/z (red curve) as a function of the retention time. All data are from 100x 3-plexDIA samples as described in the methods.

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, Fig. 4D. This can be useful to identify contaminants[23] and estimate the ions accessible to MS analysis that may be interpreted by improved algorithms[5,43]. The binned total ion current output allows to identify errors in the method setup and gives a quick overview of the sampled mass range, Fig. 4E.

Improving Sampling Using Additional Survey Scans

The conflict between reducing spectral complexity and increasing the number of data points per peak mentioned in Fig. 2 can be partially alleviated by increasing the number of survey scans[20]. When duty cycles are long, more frequent sampling on the MS1 level can increase the fraction

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of precursors with MS1 information and the probability of sampling close to the elution apex[13,19]. The DO-MS framework can be used to asses the contribution of such additional MS1 scans to improved precursor sampling.

The effect can be exemplified based on 3-plexDIA set whose samples correspond to 100-cells per channel samples analyzed with 60 minutes 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, Fig. 5A. The additional survey scan increases the duty cycle length only marginally while increasing the frequence of precursor sampling almost 2-fold. Thus, the adapted method increases the probability that precursors are sampled closed 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 expectation are supported by the results shown in Fig. 5B-D. While all data were processed and output by DO-MS, panels they were plotted outside of DO-MS. More survey scans lead to an increase in peptide like features with short elution length while maintaining the same intensity distribution, Fig. 5B. The improvements also materialize in the search engine results leading to higher MS1 intensity between intersected precursors and an increased fraction of MS1 quantified precursors without having negative effects due to the longer overall duty cycle, Fig. 5C,D. These results indicate that the duty cycle with 2 MS1 survey scans outperforms the one with a single MS1 survey scans.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

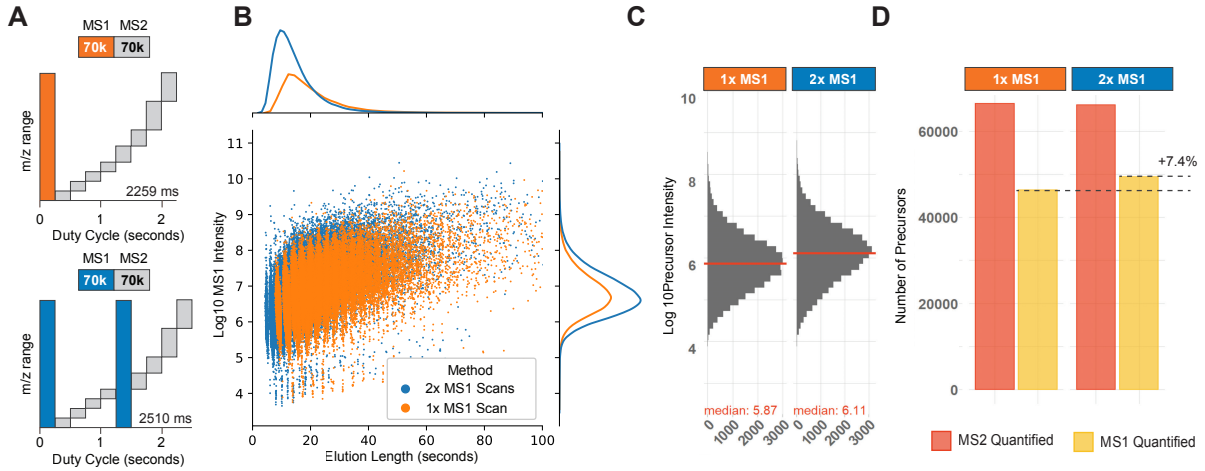


Figure 5 | **Effect of additional survey scans per duty cycle** Data acquisition methods can employ multiple survey scans to improve precursor sampling and reduce stochastic sampling effect. **A** Diagrams of a duty cycle with a single survey scan (orange) and a duty cycle with two survey scans (blue). **B** All peptide like features identified by Dinosaur[37] are displayed with their elution length at base and MS1 intensity. The absolute marginal distributions are shown. The additional survey scan allows to detect many additional peptide-like features with shorter elution profile. **C** The MS1 intensity of intersected precursors is increased upon introduction of an additional survey scan. **D** The 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. All data were acquired from a 100x 3-plexDIA samples analyzed on 60 min active gradient as described in the methods.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Quality Control for Routine Sample Acquisition

When acquiring large data sets, it is important to continuously monitor the performance of the method and identify potential failed experiments[29]. 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.*[19] shown in Fig. 6. Using nPOP sample preparation[44], 10 sets with 3 single cells each were prepared and measured on the Bruker timsTOF platform, resulting in about 1,000 quantified proteins per single cell on average, Fig. 6A. As plexDIA can benefit from translating precursor identifications between channels[19,20], the impact of translation on identifications and data completeness is reported by DO-MS. With single cells, it is vital to identify potential dropouts 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, Fig. 6B. Another metric to asses the single-cell proteome quality is the quantification variability between peptides originating from the same protein, which has been proposed as metric for single-proteome quality[45], Fig. 6C. In this data set, the cells in channel $\Delta 0$, set 06 and $\Delta 8$, set 10 show both lower number of proteins before translation and a higher quantification variability and should potentially be excluded from further analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

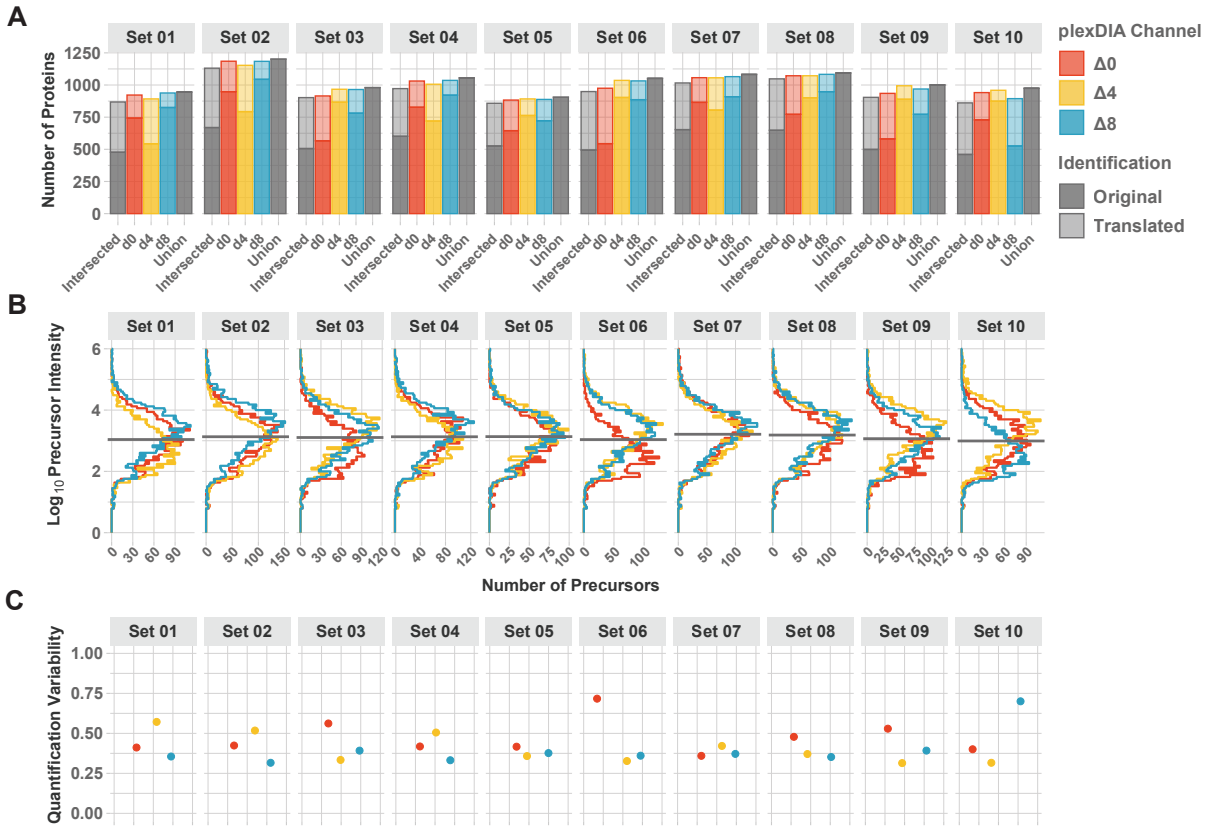


Figure 6 | **Routine quality control** When acquiring data from a large number of single cells, DO-MS can be used to get a quick overview of the quality of the processing results. **A** Number of protein identifications per single cell before and after translating identifications between channels. Only 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.*[19] for 10 single-cell 3-plex sets analyzed on the Bruker timsTOF SCP.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusion

The examples presented here illustrate some of the use cases in which DO-MS can help diagnose bottlenecks in DIA analysis and suggest simple changes in data acquisition parameters for optimizing proteomics analyses.

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 250pg 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, American Type Culture Collection (ATCC), CRL-1997) were cultured as previously described in *plexDIA - methods - cell culture*. Cells were harvested, processed, and labeled with mTRAQ as described in *plexDIA - methods - Preparation of bulk plexDIA samples*.

All bulk data was acquired on the Thermo Fisher Scientific Q-Exactive Classic Orbitrap mass spectrometer. Samples consisting of 1-μl were injected with the Dionex UltiMate 3000 UHPLC using 25 cm×75µm 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 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 with 1ul volume. Liquid chromatography was performed with 200nl/min for 30 minutes of active gradient starting with 4% Buffer B (minutes 0–2.5), 4–8% B (minutes 2.5–3), 8–32% B (minutes 3–33), 32–95% B (minutes 33–34), 95% B (minutes 34–35), 95-4% B (minutes 35–35.1), 4% B (minutes 35.1-53). All acquisition methods had a single MS1 scan cov-

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ering the range of 380mz-1400mz followed by DIA MS2 scans: 2xMS2 starting at 380mz: 240Th, 780Th width; 4xMS2 starting at 380mz: 120Th, 120Th, 200Th, 580Th width; 6xMS2 starting at 380mz: 80Th, 80Th, 80Th, 120Th, 240Th, 420Th width; 8xMS2 starting at 380mz: 60Th, 60Th, 60Th, 60Th, 100Th, 100Th, 290Th, 290Th width; 10xMS2 starting at 380mz: 50Th, 50Th, 50Th, 50Th, 50Th, 75Th, 75Th, 150Th, 150Th, 320Th width; 12xMS2 starting at 380mz: 40Th, 40Th, 40Th, 40Th, 40Th, 40Th, 60Th, 60Th, 120Th, 120Th, 210Th, 210Th width; 16xMS2 starting at 380mz: 30Th, 30Th, 30Th, 30Th, 30Th, 30Th, 30Th, 30Th, 50Th, 50Th, 50Th, 50Th, 145Th, 145Th, 145Th, 145Th width. All MS1 and MS2 scans were performed with 70,000 resolving power, 3×106 AGC maximum, 300-ms maximum accumulation time, NCE at 27%, 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 labled with mTRAQd0, mTRAQd4 and mTRAQd8 respectively. Liquid chromatography was performed with 200nl/min for 30 minutes of active gradient starting with 4% Buffer B (minutes 0–2.5), 4–8% B (minutes 2.5–3), 8–32% B (minutes 3–33), 32–95% B (minutes 33–34), 95% B (minutes 34–35), 95-4% B (minutes 35–35.1), 4% B (minutes 35.1-53). Both MS1 and MS2 scans covered a range of 380mz to 1400mz with a single MS1 scan and 8 MS2 scans. The distribution of precursors was determined based on DO-MS report using equal sized windows, starting at 380mz: 127.5Th, 127.5Th, 127.5Th, 127.5Th, 127.5Th, 127.5Th, 127.5Th, 127.5Th width. MS2 windows where 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, 385Th. 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, 422Th. All MS1 and MS2 scans were performed with 70,000 resolving power, 3×106 AGC maximum, 250-ms maximum accumulation time, NCE at 27%, default charge of 2, and RF S-lens was at 80%.

Optimizing gradient profile and length A combined sample consisting of 100 single-cell equivalents of PDAC, Monocytes and U937 were labled with mTRAQd0, mTRAQd4 and mTRAQd8

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respectively. Liquid chromatography was performed with 200nl/min flow rate starting with 4% Buffer B (minutes 0–2.5) followed by 4–8% B (minutes 2.5–3). The active gradient with 8% buffer B to 32% buffer B stretched across 15, 30 and 60 minutes followed by a 1 minute 32–95% B ramp, 1 minute at 95% and 18 minutes at 4% B. All acquisition methods had a single MS1 scan covering the range of 478mz-1500mz followed by 8 DIA MS2 scans: starting at 380mz: 60Th, 60Th, 60Th, 60Th, 100Th, 100Th, 290Th, 290Th. All MS1 and MS2 scans were performed with 70,000 resolving power, 3×106 AGC maximum, 300-ms maximum accumulation time, NCE at 27%, 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 1ul. Liquid chromatography was performed with 200nl/min for 30 minutes of active gradient starting with 4% Buffer B (minutes 0–2.5), 4–8% B (minutes 2.5–3), 8–32% B (minutes 3–63), 32–95% B (minutes 63–64), 95% B (minutes 64–65), 95-4% B (minutes 65–65.1), 4% B (minutes 65.1-83). A single MS1 scan with a range of 478mz-1500mz was followed by MS2 scans starting at 380mz with 60Th, 60Th, 60Th, 60Th, 100Th, 100Th, 290Th, 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, 3×106 AGC maximum, 300-ms maximum accumulation time, NCE at 27%, default charge of 2, and RF S-lens was at 80%.

Data Analysis

Data was analysed using DIA-NN 1.8.1 using the 5,000 protein group human-only spectral library published previously *plexDIA - methods - Spectral library generation*. Data was then processed with DO-MS. For preprocessing of orbitrap data DO-MS uses 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 other preprocessing steps are performed in the Python programming language version 3.10 and makes use of its extensive ecosystem for scientific programming including Numpy, Pandas, Matplotlib, pymzML and scikit-learn.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Availability

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 application is available upon request. The 30 single cells plexDIA data set acquired on the timsTOF has been published as part of plexDIA and is available at `http://scp.slavovlab.net/Derks_et_al_2022`.

Acknowledgments

We Acknowledge 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., and an R01 by NIGMS 5R01GM144967 to N.S.

References

1. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the american society for mass spectrometry* **5,** 976–989 (1994).

2. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. en. *Nature Methods* **1,** 39–45. ISSN: 1548-7105. (2020) (Oct. 2004).

3. Dong, M.-Q. *et al.* Quantitative Mass Spectrometry Identifies Insulin Signaling Targets in *C. elegans*. *Science* **317,** 660–663 (2007).

4. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. en. *Mol. Syst. Biol.* **14,** e8126 (Aug. 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Slavov, N. Driving Single Cell Proteomics Forward with Innovation. *Journal of Proteome Research* **20,** 4915–4918. https://doi.org/10.1021/acs.jproteome.1c00639 (2021).

6. Slavov, N. Increasing proteomics throughput. *Nature Biotechnology* **39,** 809–810. https://doi.org/10.1038/s41587-021-00881-z (2021).

7. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. en. *Nat. Methods* **12,** 258–64, 7 p following 264 (Mar. 2015).

8. Bruderer, R. *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. en. *Mol. Cell. Proteomics* **14,** 1400–1410 (May 2015).

9. Egertson, J. D., MacLean, B., Johnson, R., Xuan, Y. & MacCoss, M. J. Multiplexed peptide analysis using data-independent acquisition and Skyline. en. *Nat. Protoc.* **10,** 887–903 (June 2015).

10. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature methods* **17,** 41–44 (2020).

11. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature Biotechnology,* 1–11 (2021).

12. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nature Biotechnology.* ISSN: 1546-1696. https://doi.org/10.1038/s41587-022-01424-w (Aug. 2022).

13. Xuan, Y. *et al.* Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies. *Nature Communications* **11,** 5248 (2020).

14. Demichev, V. *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. en. *Nat. Commun.* **13,** 3944 (July 2022).

15. Li, Y. *et al.* An integrated strategy for mass spectrometry-based multiomics analysis of single cells. *Analytical Chemistry* **93,** 14059–14067 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

16. Gebreyesus, S. T. *et al.* Streamlined single-cell proteomics by an integrated microfluidic chip and data-independent acquisition mass spectrometry. *Nature Communications* **13,** 37 (2022).

17. Brunner, A.-D. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. en. *Mol. Syst. Biol.* **18,** e10798 (2022).

18. Phlairaharn, T. *et al.* High Sensitivity Limited Material Proteomics Empowered by Data-Independent Acquisition on Linear Ion Traps. *J. Proteome Res.* **21,** 2815–2826 (Nov. 2022).

19. Derks, J. *et al.* Increasing the throughput of sensitive proteomics by plexDIA. *Nature Biotechnology.* https://doi.org/10.1038/s41587-022-01389-w (2022).

20. Derks, J. & Slavov, N. Strategies for increasing the depth and throughput of protein analysis by plexDIA. *bioRxiv 2022.11.05.515287.* https://doi.org/10.1101/2022.11.05.515287 (2022).

21. Singh, A. Sensitive protein analysis with plexDIA. en. *Nat. Methods* **19,** 1032 (Sept. 2022).

22. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology* **14,** e8126 (2018).

23. Huffman, G., Chen, A. T., Specht, H. & Slavov, N. DO-MS: Data-Driven Optimization of Mass Spectrometry Methods. *J. of Proteome Res.* **18,** 2493–2500. https://doi.org/10.1021/acs.jproteome.9b00039 (6 2019).

24. Bittremieux, W., Valkenborg, D., Martens, L. & Laukens, K. Computational quality control tools for mass spectrometry proteomics. en. *PROTEOMICS* **17,** 1600159. (2019) (2017).

25. Trachsel, C. *et al.* rawDiag: An R Package Supporting Rational LC–MS Method Optimization for Bottom-up Proteomics. *Journal of Proteome Research.* ISSN: 1535-3893. https://doi.org/10.1021/acs.jproteome.8b00173 (July 2018).

26. Skowronek, P. *et al.* Rapid and In-Depth Coverage of the (Phospho-)Proteome With Deep Libraries and Optimal Window Design for dia-PASEF. *Molecular & Cellular Proteomics* **21,** 100279. ISSN: 1535-9476 (2022).

27. Framework for multiplicative scaling of single-cell proteomics. en. *Nat. Biotechnol.,* 1–2. https://www.nature.com/articles/s41587-022-01411-1 (July 2022).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Slavov, N. Single-cell protein analysis by mass spectrometry. *Current Opinion in Chemical Biology* **60,** 1–9. ISSN: 1367-5931. https://doi.org/10.1016/j.cbpa.2020.04.018 (2020).

29. Gatto, L. *et al.* Initial recommendations for performing, benchmarking, and reporting single-cell proteomics experiments. https://doi.org/10.1038/s41592-023-01785-3 (2023).

30. Specht, H. & Slavov, N. Optimizing Accuracy and Depth of Protein Quantification in Experiments Using Isobaric Carriers. *Journal of Proteome Research* **20.** PMID: 33190502, 880–887. https://doi.org/10.1021/acs.jproteome.0c00675 (2021).

31. Petelski, A. A. *et al.* Multiplexed single-cell proteomics using SCoPE2. *Nature Protocols* **16,** 5398–5425. https://doi.org/10.1038/s41596-021-00616-z (2021).

32. Slavov, N. Scaling Up Single-Cell Proteomics. *Molecular & Cellular Proteomics* **21,** 100179. ISSN: 1535-9476. https://doi.org/10.1016/j.mcpro.2021.100179 (2022).

33. Hulstaert, N. *et al.* ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J Proteome Res* **19,** 537–542 (2020).

34. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal Chem* **85,** 5288–96. ISSN: 1520-6882 (Electronic) 0003-2700 (Linking) (2013).

35. Martens, L. *et al.* mzML a community standard for mass spectrometry data. *Mol Cell Proteomics* **10,** R110 000133 (2011).

36. Rossum, G. v. Python tutorial. *technical Report CS-R9526, entrum voor Wiskunde en Informatica (CWI), Amsterdam,* (1995).

37. Teleman, J., Chawade, A., Sandin, M., Levander, F. & Malmström, J. Dinosaur: A Refined Open-Source Peptide MS Feature Detector. *Journal of Proteome Research* **15,** 2143–2151 (2016).

38. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2022). https://www.R-project.org/.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.02.526809; this version posted February 3, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39. Chang, W. *et al. shiny: Web Application Framework for R* R package version 1.7.2.9000 (2022). https://shiny.rstudio.com/.

40. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11,** O111.016717. ISSN: 1535-9476 (2012).

41. Kawashima, Y. *et al.* Optimization of Data-Independent Acquisition Mass Spectrometry for Deep and Highly Sensitive Proteomic Analysis. *International Journal of Molecular Sciences* **20.** ISSN: 1422-0067. https://www.mdpi.com/1422-0067/20/23/5932 (2019).

42. Heil, L. R. *et al.* Dynamic Data Independent Acquisition Mass Spectrometry with Real-Time Retrospective Alignment. *bioRxiv* (2022).

43. Chen, A. T., Franks, A. & Slavov, N. DART-ID increases single-cell proteome coverage. *PLOS Computational Biology* **15,** 1–30. https://doi.org/10.1371/journal.pcbi.1007082 (July 2019).

44. Leduc, A., Huffman, R. G., Cantlon, J., Khan, S. & Slavov, N. Exploring functional protein covariation across single cells using nPOP. *Genome Biology* **23,** 261. https://doi.org/10.1186/s13059-022-02817-5 (2022).

45. Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology* **22** (2021).