MSFragger-DDA+ Enhances Peptide Identification Sensitivity with Full Isolation Window Search

Fengchao Yu[1*], Yamei Deng[1], and Alexey I. Nesvizhskii[1,2*]

1. Department of Pathology, University of Michigan, Ann Arbor, MI, USA.

2. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.

*Corresponding Authors: yufe@umich.edu, nesvi@umich.edu

Abstract

Liquid chromatography-mass spectrometry (LC-MS) based proteomics, particularly in the bottom-up approach, relies on the digestion of proteins into peptides for subsequent separation and analysis. The most prevalent method for identifying peptides from data-dependent acquisition (DDA) mass spectrometry data is database search. Traditional tools typically focus on identifying a single peptide per tandem mass spectrum (MS2), often neglecting the frequent occurrence of peptide co-fragmentations leading to chimeric spectra. Here, we introduce MSFragger-DDA+, a novel database search algorithm that enhances peptide identification by detecting co-fragmented peptides with high sensitivity and speed. Utilizing MSFragger's fragment ion indexing algorithm, MSFragger-DDA+ performs a comprehensive search within the full isolation window for each MS2, followed by robust feature detection, filtering, and rescoring procedures to refine search results. Evaluation against established tools across diverse datasets demonstrated that, integrated within the FragPipe computational platform, MSFragger-DDA+ significantly increases identification sensitivity while maintaining stringent false discovery rate (FDR) control. It is also uniquely suited for wide-window acquisition (WWA) data. MSFragger-DDA+ provides an efficient and accurate solution for peptide identification, enhancing the detection of low-abundance co-fragmented peptides. Coupled with the FragPipe platform, MSFragger-DDA+ enables more comprehensive and accurate analysis of proteomics data.

Introduction

Liquid chromatography-mass spectrometry (LC-MS) based proteomics is a widely used, high-thorough method for studying proteins and endogenous peptides. In the bottom-up proteomics framework, proteins are first digested into shorter peptides to facilitate ionization and fragmentation. These peptides are then separated using liquid chromatography and subsequently analyzed using mass spectrometry. A mass spectrometer typically produces two types of spectra, including mass spectra (MS1) containing ions from intact peptides and tandem mass spectra (MS2) comprising fragmented ions from selected peptides. The fragmented ions can be used to infer the peptide sequence and post-translational modifications (PTMs). Two main strategies, data-independent acquisition (DIA) and data-dependent acquisition (DDA), have been developed based on how peptides are selected for fragmentation. In DIA, peptide ions within predefined mass-to-charge (m/z) windows are isolated and fragmented to generate MS2. The isolation windows are designed to cover the entire m/z range within a limited duty cycle. The resulting MS2 spectra are multiplexed, containing fragmented ions from multiple co-eluted peptides. In contrast, DDA isolates selected peptide ions to generate MS2. To ensure high specificity, isolation windows are narrower in DDA compared to DIA and typically range from 0.7 to 2.0 Th. However, with complex samples, co-eluting peptide with similar mass may still co-fragment and result in chimeric MS2 spectra[1, 2]. Recently, researchers studying single-cell proteomics noted advantages of using a wide-window acquisition[3, 4] (WWA) DDA method. The WWA method generates MS2 spectra similar to DDA but with wider isolation windows to reduce the duty cycle. As a result, more peptides are co-fragmented, producing MS2 spectra as multiplex as those generated by DIA.

Depending on the data acquisition strategy, DDA or DIA, different methods and software have been developed to identify peptides from the acquired MS2 spectra. For DDA, database search-based approaches[5-8] are commonly used to find the best scoring candidate peptide for each MS2. The proteins in the database are first *in silico* digested into peptides, and then the peptides are *in silico* fragmented to generate theoretical MS2. The theoretical MS2 are compared with the experimental MS2 to find the most likely peptide-to-spectrum (PSM) match. In doing so, only peptides with the theoretical mass matching (with a narrow tolerance, e.g. 20 ppm) the experimental MS2 precursor peptide mass are considered. The DDA peptide identification methods generally assume that MS2 is not multiplexed, and thus typically only a single peptide identification is reported for each spectrum. However, even with a narrow

isolation windows size, there are a substantial number of chimeric MS2 DDA spectra containing co-fragmented peptides[1, 2]. Strategies for improved identification of peptides from chimeric spectra have been discussed since the early days of proteomics and include iterative database search (with or without removal of fragments assigned to the top scoring peptide)[9-13] and more elaborated spectral deconvolution[14, 15]. Most of these strategies rely on the detection of MS1 precursor peptide features[16-18] of co-fragmented peptides. Although these methods increase the total number of identified peptides from DDA data, they require the co-fragmented peptides to have high-quality precursor peaks to facilitate their identification. Due to different sensitivity of MS1 and MS2, the co-fragmented peptides with high-quality MS2 might have low-quality or even no precursor peaks detected in MS1, hindering their identification.

In contrast, DIA MS2 spectra are assumed to be highly multiplexed, which makes peptide identification more challenging, but spectral library-based[19-21] and library-free[22-24] methods have been developed to tackle the problem. Compared to well-studied peptide identification methods for DDA and DIA data, there are few tools[25] that natively support WWA DDA data. Although WWA DDA MS2 are generated in a similar way to conventional DDA MS2, they are highly multiplexed, making them unsuitable for the existing DDA methods. At the same time, WWA MS2 do not allow extraction of fragment ion chromatograms (XIC), which also makes them unsuitable for most existing DIA software tools.

We have recently described a new computational strategy MSFragger-DIA[23] for direct peptide identification from DIA data that essentially blurs the boundary between DDA and DIA data by combining the initial spectrum-centric search of DIA MS2 with a subsequent, peptide-centric re-scoring. Here, we further extend this strategy and present MSFragger-DDA+, a new search mode of MSFragger which utilizes high-resolution DDA MS2 to detect co-fragmented peptides with high sensitivity and accuracy. Unlike other DDA tools, this method does not perform MS1 detection or spectral deconvolution before the database search. Instead, MSFragger-DDA+ performs a full isolation window search and then detects and utilizes MS1 precursor peaks to refine and rescore the results. We demonstrate that MSFragger-DDA+ is fast, has higher sensitivity than other DDA peptide identification tools, and fully supports peptide identification from WWA DDA data. We implemented MSFragger-DDA+ as a module in MSFragger, with DDA+ mode automatically triggered by annotating input DDA data as DDA+ type. Finally,

MSFragger-DDA+ is fully integrated in the widely used FragPipe computational platform to provide a complete solution for proteomics data analysis of DDA and WWA DDA data, from identification to quantification. MSFragger DDA+ has been publicly available as part of MSFragger since version 4.0 (released December 2023), and it has already been used by others in the proteomics community[25, 26].

Results

Overview of MSFragger-DDA+

In the database search framework, traditional peptide identification algorithms search each MS2 against candidate peptides within a narrow mass tolerance, which is normally 5-20 ppm (for high resolution MS2 data) around the precursor mass reported by a mass spectrometer for that scan, to identify the best match. In contrast, MSFragger-DDA+ searches each MS2 against all peptides within the full isolation window (**Figure 1**). During the database search, MSFragger-DDA+ uses hyperscore[5, 27] to measure the similarity between the MS2 and the candidate peptides in the isolation window. It is reasonable to assume that there will be many false matches since MSFragger-DDA+ does not constrain the search within the narrow mass tolerance of the peptide selected by the mass spectrometer for MS2. At the same time, many of the co-fragmented peptides do not have a high-quality precursor peak observed in the parent MS1 scan. Thus, we developed post-database search refinement steps, including precursor XIC detection, shared fragment removal, and rescoring. MSFragger-DDA+ detects and extracts XICs for each PSM after the database search. It extracts as many isotopic XICs as possible to compare the intensity distribution with the theoretical distribution[28] with the Kullback–Leibler divergence[29]. Because the theoretical m/z and charge are known after the database search, MSFragger-DDA+ extracts the XICs in the “targeted” manner. In contrast to the “untargeted” MS feature detection approach [16, 22, 30] that may not accurately detect and extract the low abundance features and their peak curves, the targeted approach always extracts the XICs given the theoretical m/z values and only within the retention time range determined based on the retention time of the corresponding MS2 scan. Moreover, the targeted approach does not suffer from the challenge of untargeted deconvolution of overlapped isotopic peak clusters which might result in incorrect charge and m/z determination. After the targeted extraction, PSMs with low-quality XICs are discarded. Then, MSFragger-DDA+ rescores the PSMs by removing the

fragments shared by multiple peptides using a greedy algorithm[23]. Finally, it generates output files that can be processed by downstream tools.

To make MSFragger-DDA+ easy to access and user-friendly, we have integrated it into FragPipe computational suite. The output can be seamlessly processed by FragPipe modules, including MSBooster[31] deep-learning-based rescoring, Percolator[32] PSM re-ranking, PeptideProphet[33] PSM rescoring, PTMProphet[34] modification localization, ProteinProphet[35] protein grouping, Philosopher[36] false discovery rate (FDR) filtering, IonQuant[29] quantification, EasyPQP spectral library generation, PDV[37] and Skyline[19] visualization. MSFragger-DIA+ mode can be triggered in any FragPipe workflow that supports DDA by simply annotating the input MS files as DDA+ type.

MSFragger-DDA+ improves the sensitivity of peptide identification from DDA data

First, we evaluated the performance of MSFragger-DDA+ using two DDA datasets. The first one was published by Searle et al[38]. There are five samples with different normalized collision energy (NCE): 22, 27, 32, 37, and 42. We used MaxQuant[30] (version 2.4.13), MetaMorpheus[18] (version 1.0.5), MSFragger[5] (version 4.1) in DDA mode, and MSFragger-DDA+ (version 4.1) for peptide identification. We also included Scribe's results from the original publication[38]. The second DDA dataset was published by Richards et al[39]. There are six samples with different types of enzymatic digestions (trypsin, AspN, and GluC) and fragmentations (CID and HCD). For MetaMorpheus, the precursor deconvolution was enabled by default to deconvolute co-fragment peptides separately. For MSFragger and MSFragger-DDA+, FragPipe (version 22.0) was used to perform MSBooster deep-learning-based rescoring, Percolator re-ranking, ProteinProphet protein grouping, and Philosopher FDR filtering (see **Methods** for detail).

The numbers of identified peptide sequences after filtering at a 1% false discovery rate (FDR) for the Searle et al. and Richards et al. datasets are shown in **Figures 2a and 2b**, respectively. We used the FDR reported by the tools to perform the filtering (**Methods**). **Figure 2a** shows that although MetaMorpheus, MSFragger (conventional DDA mode), and Scribe have similar performance, the sensitivity of MSFragger-DDA+ is much higher across all NCE. On average, MSFragger-DDA+ identified 57% more peptide sequences than MSFragger in the conventional

DDA mode. **Figure 2b** shows similar comparisons, with MSFragger-DDA+ identifying from 15% (AspN CID and GluC CID) to 45% (Trypsin HCD) more peptide sequences depending on the combination of enzymatic digestion and fragmentation. We also compared the runtime of MaxQuant, MetaMorpheus, MSFragger in DDA mode, and MSFragger-DDA+ (**Figure 2c).** The comparison shows that MaxQuant and MetaMorpheus have similar speed, whereas MSFragger-DDA+ coupled with FragPipe is four times faster.

MSFragger-DDA+ coupled with FragPipe has good FDR control

The previous section demonstrated that MSFragger-DDA+ identifies many more peptide sequences compared to MSFragger in the conventional DDA mode and the other tools tested. We performed an entrapment database search[40] to evaluate the FDR control. Given the original target database, we generated an entrapment sequence for each target protein sequence by random shuffling. During the shuffling, the peptide C-termini were fixed. The target and entrapment sequences were used as the new target database. We searched the Searle et al[38] dataset against the new database using MaxQuant, MetaMorpheus, MSFragger in DDA mode, and MSFragger-DDA+. For MSFragger and MSFragger-DDA+, FragPipe was used to perform downstream analysis as described in the previous section (see also **Methods**). **Figure 2d** shows the number of proteins, entrapment proteins, and two false discovery proportion (FDP) estimations proposed by Wen et al[40]. To make it more intuitive, we used the name “upper bound” to replace the name “combined method” used by Wen et al[40] since it is a method to calculate the upper bound of the FDP. The peptide level statistics are shown in **Supplementary Figure 1**. This experiment demonstrates that not only MSFragger-DDA+ identifies the largest number of target peptide sequences and proteins it also has a similarly low FDP as MSFragger and lower FDP than that for the other tools.

Application to DDA-PASEF data

We used a dataset published by Wang et al[41] to demonstrate that MSFragger-DDA+ performs well when analyzing DDA-PASEF data generated using the Bruker’s timsTOF platform. There are two cell lines, A549 and K562. Each cell line has three biological replicates, and each biological replicate has four technical replicates. We used MSFragger in DDA mode and MSFragger-DDA+ to perform the peptide identification. FragPipe was used to process the

peptide identification outputs as in the previous sections. For label-free quantification (LFQ), IonQuant[29, 42] with and without match-between-runs (MBR) was used. **Figure 3a** and **Supplementary Figure 2a** show the number of quantified proteins from the two cell lines, respectively. A protein is quantified when it has a non-zero intensity. For each cell line, three biological replicates with and without MBR, are listed separately. For each biological replicate, different color transparency levels indicate the number of proteins quantified in one to four technical replicates. These two figures show that MSFragger-DDA+ coupled with FragPipe has higher sensitivity than the DDA workflow. Without MBR, the DDA+ workflow quantified 16% to 19% more proteins among the six experimental replicates. The differences are smaller with MBR because it transfers identifications from one replicate to another to reduce missing values, however, the DDA+ workflow still quantified 11% to 13% more proteins.

The quality of quantification was also evaluated by calculating the coefficient of variations (CVs) using the four technical replicates of each sample (see **Figure 3b** and **Supplementary Figure 2b**). Overall, the proteins identified uniquely with one method (MSFragger-DDA+ or MSFragger) have higher CVs than the proteins identified in common. The same set of common proteins quantified by both DDA and DDA+ workflows do not have significantly different CVs. The proteins unique to the DDA+ workflow have slightly higher CVs than the proteins unique to DDA. We reasoned that these proteins are of lower abundance and have low signal-to-noise ratios because their peptides were co-fragmented but not dominant in the MS2. There are also more DDA+ unique proteins than the DDA unique proteins. **Figure 3c** and **3d** show the proportion of missing values in the DDA and DDA+ workflows, respectively. **Figure 3c** is without MBR, and **Figure 3d** is with MBR. The results show that the DDA+ workflow quantified notably more proteins (6585 vs 5619 without MBR, and 6725 vs 5916 with MBR) and with a comparable missing value proportion (17% without MBR and 7% with MBR) as the conventional MSFragger DDA workflow. **Supplementary Figure 2c** and **2d** show the principal component analysis (PCA) plots from the DDA and DDA+ workflows, respectively, exhibiting comparable separation between the samples (and similar to that reported in the original publication).

MSFragger-DDA+ fully supports wide-window acquisition data

WWA[3, 4] is a data acquisition approach primarily designed for single-cell proteomics. It generates MS2 in the DDA mode, but with wider isolation windows. Multiple peptides are

intentionally co-fragmented to produce multiplexed spectra. As a result, traditional DDA database search tools are not well-suited for such data. Using the datasets published by Truong et al[3] and Matzinger et al[4], we demonstrate that MSFragger-DDA+ has excellent performance on such data.

The first dataset[3] contains 29 low-input samples with different isolation window sizes (1.6, 2, 4, 8, 12, 18, 24, and 48 Th), maximum injection time (54, 86, 118, and 246 ms), and MS2 resolutions (30K, 45K, 60K, and 120K). Each sample has two technical replicates. We used CHIMERYS[25] coupled with Proteome Discoverer, MetaMorpheus[18], and MSFragger-DDA+ for peptide identification. CHIMERYS natively supports the WWA data type. For MetaMorpheus, the precursor deconvolution was enabled to deconvolute co-fragmented peptides into separated spectra. **Figure 4a** shows the number of peptides identified from the samples with 54 ms maximum injection time, 30K MS2 resolution, and 2 - 48 Th isolation window sizes. **Figure 4b** shows the number of identified peptides from the samples with 118 ms maximum injection time, 60K MS2 resolution, and 2 - 48 Th isolation window sizes. To avoid redundancy, the figures for other samples are not shown, but they demonstrate a similar trend. The result files can be found as described in the **Data availability** section. CHIMERYS and MSFragger-DDA+ obtained similar numbers, and higher than that of MetaMorpheus. Since this dataset has another four high-input samples to be used as “library” runs in the MBR, we performed the same analysis with MBR enabled (**Supplementary Figure 3a** and **3b**). For MSFragger-DDA+, the MBR was performed using IonQuant as part of FragPipe. For CHIMERYS and MetaMorpheus, the MBR was performed by Proteome Discoverer and MetaMorpheus, respectively. The figures show that MetaMorpheus reported the highest number of peptides after applying MBR, followed by MSFragger-DDA+ coupled with FragPipe. CHIMERYS coupled with Proteome Discoverer reported the lowest number compared to the other two. However, considering that different tools have different parameters and criteria to perform and quality control the MBR results, it is difficult to make a fair comparison when MBR is enabled.

The second dataset is also from Truong et al[3]. There are two cell lines: K562 and HeLa. Each cell line has two biological replicates with different gradient lengths of 20 min and 40 min. Each biological replicate has eight technical replicates. We used CHIMERYS coupled with Proteome Discoverer 3.0, MetaMorpheus, and MSFragger-DDA+ coupled with FragPipe to analyze the

data. As there are eight technical replicates, we evaluated the quantification quality by calculating the CVs for each biological replicate. **Figure 4c** and **4d** show the number of quantified proteins. Similar to the first dataset, there are four high-input samples to be used as "library". **Supplementary Figure 3c** and **3d** show bar plots with the MBR enabled. CHIMERYS coupled with Proteome Discoverer and MSFragger-DDA+ coupled with FragPipe have similar sensitivity when the MBR is not enabled. Both have higher sensitivity than MetaMorpheus. When the MBR is enabled, MSFragger-DDA+ coupled with FragPipe and MetaMorpheus have similar sensitivity, which is higher than that of CHIMERYS coupled with Proteome Discoverer.

The third dataset is from Matzinger et al[4]. There are 43 samples with different isolation window sizes (1, 2, 3, 4, 5, 6, 7, 8, 12, 18, 24, 28, 56 Th) and sample amounts (250 pg, 1 ng, 10 ng, 200 ng, and 400 ng). Each sample has three technical replicates. We used MSFragger-DDA+ coupled with FragPipe, and MetaMorpheus to analyze the dataset. We also used the CHIMERYS/Proteome Discoverer 3.0 results as provided by the authors as part of the original publication. **Figure 4e**, **4f**, **Supplementary Figure 3e**, and **3f** show the number of identified peptides from samples with 250 pg, 1 ng, 200 ng, and 400 ng. The samples with 10 ng are not shown due to redundancy. The result files can be found as described in the **Data availability** section. CHIMERYS coupled with Proteome Discoverer and MSFragger-DDA+ coupled with FragPipe have similar sensitivity, and higher than that of MetaMorpheus. From the samples with smallest amount (250 pg), CHIMERYS, MSFragger-DDA+, and MetaMorpheus identified on average 2562, 2186, and 1713 peptides, respectively. From the samples with the largest amount (400 ng), CHIMERYS, MSFragger-DDA+, and MetaMorpheus identified on average 47100, 44879, and 25689 peptides, respectively.

The above datasets from two laboratories with different sample preparation and data acquisition configurations demonstrate that MSFragger-DDA+ performs well for identifying peptides from WWA data. Furthermore, MSFragger-DDA+ is faster than other tools when analyzing WWA dataset. For the first dataset with 62 mzML files as input, MSFragger-DDA+ coupled with FragPipe took 68.8 minutes on a Linux server with Intel Xeon Gold 6354 CPU (3.00 GHz, 36 physical cores) and 768 GB RAM (although 768 GB RAM was available, only a small proportion was used during the analysis). In contrast, CHIMERYS and Proteome Discoverer took 25 hours in total, of which most of the time (16.5 hours) was taken by CHIMERYS running on the

company's proprietary cloud computing platform. Although a fair runtime comparison is difficult, our experiment showed that MSFragger-DDA+ coupled with FragPipe was at least 20 times faster than CHIMERYS coupled with Proteome Discoverer in these data.

MSFragger-DDA+ leads to detection of more differentially expressed proteins and genes

We further demonstrate the performance of MSFragger-DDA+ combined with the FragPipe suite by utilizing a dataset from IDH-mutated glioma patients[43]. The dataset includes three groups of samples: IDH-mutant group (11 1p/19q-codeleted samples and 10 astrocytoma samples), one IDH wild-type group (11 samples), and one non-neoplastic CNS tissue control group (10 samples). Further details about sample preparation and data acquisition can be found in the original publication. We utilized MSFragger-DDA+ for peptide identification as part of the FragPipe's LFQ-MBR built-in workflow. Following the database search, FragPipe performed the PSM rescoring[31], protein grouping[35], FDR filtering[36], and LFQ[29] (**Methods**). FragPipe with MSFragger-DDA+ identified 10738 proteins, 935 more proteins than MSFragger (and 1231 more proteins than MaxQuant reported in the original publication). Proteins with non-zero MaxLFQ intensities were used to perform downstream analysis using FragPipe-Analyst[44] - a recent addition to the FragPipe family of tools that enables seamless analysis of FragPipe generated quantification data. Among the FragPipe-Analyst's output files, the heatmap (**Figure 5a**) successfully recovered distinct protein expression patterns for the three sample groups (IDH wild-type, IDH mutant, and control). The PCA (**Figure 5b**) plot demonstrated that the IDH wild-type gliomas were most distinct from the control group and were separated from the IDH mutant cases, as expected. We then used FragPipe-Analyst to generate a volcano plot (**Figure 5c**) of the IDH mutant versus wild-type. After filtering the genes with 0.05 Benjamini-Hochberg adjusted p-value and 1 log2 fold change, the differentially expressed genes highly overlapped with those reported in the original publication. We also compared the results of the DDA and DDA+ workflows. The DDA+ workflow quantified more proteins (counting unique gene symbols) (**Supplementary Figure 4a)** and with fewer missing values (**Figure 5d** and **Supplementary Figure 4b**), leading to the detection of more differentially expressed proteins after Limma analysis (**Figure 5e**). Finally, we took advantage of the Enrichr[45] enrichment analysis tool integrated into FragPipe-Analyst. We performed Gene Ontology (GO) analysis using proteins differentially expressed between the IDH wild-type and the mutant samples (**Supplementary Figure 4c**). Three levels of GO analysis show that the DDA+ workflow identified more differential proteins corresponding to each of the enriched categories. Overall, our analysis

suggests that MSFragger-DDA+ coupled with the FragPipe leads to more sensitive detection of differentially expressed proteins and pathways in cancer proteomics profiling experiments.

Discussion

We presented a new peptide identification method, MSFragger-DDA+, that has a higher sensitivity compared to the conventional DDA peptide identification strategy. Unlike conventional tools developed for DDA data, it starts by searching MS2 spectra against all peptides within the full isolation window to enable identification of all possible co-fragmentated ions. Importantly, many of those co-fragmented peptides do not have a strong precursor MS1 signal, making them difficult to identify from chimeric DDA spectra using strategies that rely on the knowledge of the masses of co-fragmented peptides prior to the search. Instead, in MSFragger-DDA+, precursor signals are considered only at the second, targeted rescoring step. At that stage, a small list of the top scoring candidate peptides for each MS2 spectrum has been established, and the XICs for these peptide ions can be more easily extracted from MS1 data in a targeted way. Thus, by reversing the order of the two key steps, database search and MS1 feature detection, MSFragger-DDA+ can detect more of low abundant, co-fragmented peptides.

MSFragger-DDA+ is the latest tool in our long-going efforts to develop a computational peptide identification workflow that provides a uniform treatment of MS2 data across the entire spectrum of data acquisition strategies, from conventional DDA to narrow-window DIA, to wide-window DDA, and wide-window DIA. The key to these efforts has been our development of the fragment ion indexing algorithm that enabled spectrum-centric search of MS2 data against protein sequence databases in essentially unrestricted way. In the original MSFragger[5] manuscript we demonstrated the application of fragment ion indexing to open (also known as mass tolerant or unrestricted) searches of DDA data to identify modified peptides. Later we have extended this work in MSFragger-LOS[46] for localization-aware open search, in MSFragger-Glyco[47] for glycopeptide identification, and in MSFragger-Labile[48] for labile modifications. In parallel, we have applied our strategy to enable direct, spectrum-centric search of DIA data with MSFragger-DIA[23]. With MSFragger-DDA+, we now return to the analysis of conventional DDA data and enable the identification of co-fragmented peptides from chimeric DDA spectra using full isolation window search. Regardless of the MS2 data type, we start with the spectrum-centric search of MS2 spectra against the theoretical spectra generated from the protein sequence

database, without the need for a spectral library or any predictions of peptide properties or MS2 spectra prior to the search. It is only after the MSFragger database search that we leverage deep learning-based fragment intensity and retention time predictions (and only when such predictions are likely to be accurate) in MSBooster to refine and rescore the peptide candidates to boost the identification sensitivity[31]. Furthermore, regardless of the MS2 data type, all MSFragger search results are processed using the same downstream tools (MSBooster with Percolator for rescoring; protein inference with ProteinProphet; FDR filtering with Philosopher). To summarize, with the spectrum-centric database search framework at its core, the MSFragger family of tools unifies peptide identification across different MS2 data acquisition modalities (DDA, wide window DDA, DIA) and different search modes (closed, open, and mass-offset).

There are still important differences between the full isolation window search described in this work and the open search for modified peptides using DDA data. Given an MS2, MSFragger-LOS performs an open search of conventional DDA spectra using a very wide mass window (e.g. -150 to 500 Da, much larger than the isolation window) around the precursor mass associated with the corresponding DDA MS2 scan to identify modified (mass shifted) peptides. In doing so, it matches the fragment peaks shifted by unknown modifications, in addition to matching the unshifted peaks. The open modification search normally reports one peptide for each MS2 and does not consider co-fragmented peptides. On the contrary, MSFragger-DDA+ searches for peptides that fall within the isolation window, without allowing any unexpected mass shifts (larger than the isolation window width) or using shifted peaks in scoring. MSFragger-DDA+ then detects the peptide precursor signals after searching the database to refine the matches and report the observed precursor m/z and charge that may be different from that listed for the DDA spectrum. Enabling the full isolation window search for co-fragmented peptides in parallel with the open search for modified peptides is the next computational challenge that we plan to address in future work.

With the development of MSFragger-DDA+, all workflows developed in FragPipe for the analysis of DDA data can now be run using DDA+ mode. In our experience, MSFragger-DDA+ provides a higher sensitivity boost with unfractionated single-shot LC-MS runs compared to fractionated ones. This is expected given that an unfractionated run, keeping the LC gradient constant, contains more co-eluting peptides, and thus more chimeric spectra, than an MS run

on a fractionated peptide sample. Furthermore, our experiments have shown that MSFragger-DDA+ significantly reduces the number of missing intensity values across all samples in multi-sample analyses typical to label-free quantification workflows. However, when LFQ analysis is performed with MBR enabled, the differences became smaller because MBR helps to achieve more complete quantification matrix via the transfer of peptide identifications between the runs. Nevertheless, MSFragger-DDA+ still results in more peptides and proteins in total, which results in better downstream analysis results. Finally, not all DDA-based workflow, most notably quantitative workflows based on isobaric labeling such as tandem mass tag (TMT), would benefit from an increase in the number of peptide identifications afforded by MSFragger-DDA+. In TMT workflows, co-fragmented peptides can introduce interference in the isobaric quantification because they share the same set of reporter ions. To address this issue, most computational workflows for TMT data (including TTM-Integrator in FragPipe) by default discard MS2 spectra with low isolation purity scores[49]. Therefore, unlike LFQ, it may not be beneficial to detect co-fragmented peptides using the MSFragger-DDA+ approach in TMT-based quantitative proteomics datasets.

Methods

DDA+ enhanced peptide identification

MSFragger-DDA+ supports both DDA and WWA data types. In contrast to the traditional database search approach that matches a spectrum against peptides within a narrow mass window, MSFragger-DDA+ matches all peptides in the isolation window to detect all possible co-fragmented peptides. Hyperscore[5, 27] is used to measure the peptide-spectrum similarity during this process. After the database search, MSFragger-DDA+ detects and extracts the precursor XICs for each matched peptide. Fragment indexing[42] is used to accelerate this procedure. MSFragger-DDA+ extracts as many isotopic XICs as possible, and compares the intensity distribution with theoretical distribution[28] using Kullback-Leibler divergence[29]. Peptides with low-quality XICs are discarded. Last, all peptides are rescored using the greedy algorithm used by Yu et al[23].

Traditional DDA data

Two DDA datasets published by Searle et al[38] and Richards et al[39] were used to evaluate the performance of MSFragger-DDA+. The first dataset was generated by a Thermo Exploris 480 mass spectrometer. There are five HeLa samples with different NCEs: 22, 27, 32, 37, and 42. The second dataset was generated by a Thermo Orbitrap Fusion Lumos mass spectrometer. There are six HEK 293T samples with different enzymatic digestions (trypsin, AspN, and GluC) and fragmentations (CID and HCD). Details of sample preparation and data acquisition can be found in the original publications. The data was analyzed using MaxQuant[30] (version 2.4.13), MetaMorpheus[18] (version 1.0.5), Scribe[38], MSFragger[5] (version 4.1) in DDA mode, and MSFragger-DDA+ (version 4.1). The FASTA database is the *Homo Sapiens* reference proteome used by Searle et al[38] (20361 proteins). MaxQuant was run using default settings without including built-in contaminants. The "evidence.txt" file with decoys removed was used to count the peptide sequences. For MetaMorpheus, the precursor deconvolution was enabled by default to deconvolute co-fragmented peptides into separated PSMs. We used the "*_Peptides.psmtsv" files in the "Individual File Results" folder to count the peptide sequences identified from each input file. The decoys were removed, and the peptides were filtered with "PEP_QValue < 0.01". For MSFragger and MSFragger-DDA+, the reversed decoy sequences were generated and appended to the target database. FragPipe (version 22.0) was used to process the outputs of MSFragger and MSFragger-DDA+ to perform MSBooster deep-learning-based rescoring, Percolator PSM re-ranking, ProteinProphet protein grouping, and Philosopher FDR filtering. The "Default" workflow was applied with adjusted maximum allowed missed cleavage and enzymatic rules. We used the "peptide.tsv" files, which were filtered at 1% peptide- and protein-level FDR, to count the peptide sequences. For all tools, the maximum allowed missed cleavage was set to 1. Acetylation of the protein N-terminus and oxidation of methionine were set as variable modifications, and carboxymethylation of cysteine was set as a fixed modification. The results of Scribe were downloaded from Searle et al[38]. The parameter, log, and result files are available as described in the **Data availability** section. The scripts to summarize the results and generate the figures are available as described in **Code availability**.

False-discovery rate evaluation using entrapment database search

An entrapment database was used to evaluate the FDR of MSFragger-DDA+ and compare with that from the existing tools. Each target protein was randomly shuffled to generate one entrapment protein[40]. Peptide C-termini were fixed during the shuffling. The code to generate the entrapment database and to calculate the false discovery proportion (FDP) can be found in

Code availability. The dataset published by Searle et al[38] was used to search against the entrapment database. MaxQuant (version 2.4.13), MetaMorpheus (version 1.0.5), MSFragger (version 4.1) in DDA mode, and MSFragger-DDA+ (version 4.1) were used. The parameters are the same as those used in the previous section. For MaxQuant, the "evidence.txt" and "proteinGroups.txt" files with decoys removed were used to count the peptide sequences and proteins, respectively. For MetaMorpheus, the "*_Peptides.psmtsv" and "*_ProteinGroups.tsv" files in the "Individual File Results" folder were used to count the peptide sequences and proteins, respectively. The decoys were removed, and peptides were filtered with "PEP_QValue < 0.01", and the proteins were filtered with "Protein QValue < 0.01". For MSFragger and MSFragger-DDA+, the "peptide.tsv" and "combined_protein.tsv" files were used to count the peptide sequences and proteins, respectively. The "peptide.tsv" files contain the peptides filtered with 1% peptide- and protein-level FDR, and the "combined_protein.tsv" file contains the proteins filtered with 1% protein-level FDR. The raw parameter files, result files are available as described in **Data availability**. The scripts to summarize the results and generate the figures are available as described in **Code availability**.

timsTOF DDA-PASEF data analysis

The dataset published by Wang et al[41] was used to evaluate the performance of MSFragger-DDA+ when analyzing timsTOF DDA-PASEF data. There are two cell lines, A549 and K562. Each cell line has three biological replicates, and each biological replicate has four technical replicates. A Bruker timsTOF Pro mass spectrometer was used to generate the data. Details of the sample preparation and data acquisition can be found in the original publication. We used MSFragger (version 4.1) in DDA mode and MSFragger-DDA+ (version 4.1) to analyze the dataset. As in the previous section, FragPipe (version 22.0) was used to process the outputs of MSFragger and MSFragger-DDA+. In addition, IonQuant[29, 42] was used to perform the label-free quantification with and without MBR, respectively. The "LFQ-MBR" workflow with MBR enabled and disabled was applied, respectively. The FASTA database is the *Homo Sapiens* proteome provided by the original publication[41] (20437 proteins including common contaminants from https://www.thegpm.org/crap/). Reversed decoy sequences were generated and appended to the target database. The maximum allowed missed cleavages were set to 2. The other parameters were the same as those described in the previous section. The "combined_protein.tsv" files were used to summarize the results and generate the figures. The

"MaxLFQ Intensity" was used as the protein intensity. Detailed parameter, log, and result files can be found as described in **Data availability**.

Wide-window acquisition data analysis

Datasets published by Truong et al[3] and Matzinger et al[4] were used to demonstrate the performance of MSFragger-DDA+ on WWA data. The first dataset contains 29 low-input samples with different isolation windows sizes (1.6, 2, 4, 8, 12, 18, 24, 48 Th), maximum injection time (54, 86, 118, and 246 ms), and MS/MS resolutions (30K, 45K, 60K, and 120K) from Truong et al[3]. Each sample has two technical replicates. There are also four high-input samples to be used as "library" during the MBR. The second dataset[3] contains four samples with different cell lines (K562 and HeLa) and gradient lengths (20 min and 40 min). Each sample contains eight technical replicates. Similar to the first dataset, there are also four high-input samples to be used as "library" runs. The third dataset[4] contains 43 samples from different isolation window sizes (1, 2, 3, 4, 5, 6, 7, 8, 12, 18, 24, 28, and 56 Th) and sample amounts (250 pg, 1 ng, 10 ng, 200 ng, and 400 ng). Each sample has three replicates. Thermo Orbitrap Exploris 480 mass spectrometers were used to generate these three datasets. Details of sample preparation and data acquisition can be found in the original publications. The data were analyzed using CHIMERYS[25] coupled with Proteome Discoverer 3.0, MetaMorpheus (version 1.0.5), and MSFragger-DDA+ (version 4.1) coupled with FragPipe (version 22.0). For MetaMorpheus and MSFragger, the maximum allowed missed cleavages were set to 2. For MetaMorpheus, the precursor deconvolution was enabled to deconvolute co-fragmented peptides into separated spectra. The FASTA databases are from the original publications[3, 4]. For FragPipe, the built-in "WWA" workflow was applied with the MBR settings adjusted accordingly. Detailed parameter, log, and result files can be found as described in **Data availability**.

Glioma data analysis

A Glioma study[43] dataset was used to demonstrate the performance of MSFragger-DDA+ on data from cancer proteomics experiments. There are 42 samples analyzed using the Thermo Q Exactive HF Orbitrap mass spectrometer. Details of the sample preparation and data acquisition can be found in the original publication. MSFragger (version 4.1) and MSFragger-DDA+ (version 4.1) were used to analyze the data. The outputs were processed using FragPipe

(version 22.0). The FASTA database contains human reviewed proteins and common contaminants downloaded from UniProt (20468 proteins). The "LFQ-MBR" workflow was applied with "MBR top runs" set to 100. Details of the parameters and result files can be found as described in the **Data availability**. Section. FragPipe-Analyst[44] was used to perform downstream analysis using the output files generated by FragPipe.

Runtime comparison

The dataset published by Searle et al[38] was used to evaluate the speed of MSFragger-DDA+. MaxQuant (version 2.4.13), MetaMorpheus (version 1.0.5), and MSFragger (version 4.1) in DDA mode were also used. FragPipe (version 22.0) was used to process the outputs of MSFragger and MSFragger-DDA+. The mzML files converted[50] from the raw files were used in MetaMorpheus, MSFragger, and MSFragger-DDA+. MaxQuant was run using the raw files. All tools were run on a computer with Intel Xeon W-2235 CPU (3.80 GHz, 6 physical cores, 12 threads) and 128 GB RAM. The detailed log files can be found as described in the **Data availability** section.

Acknowledgements

This work was supported in part by National Institutes of Health grants R01-GM-094231 and U24-CA271037. We thank Ryan Kelly and Thy Truong for the discussions regarding WWA data.

Abbreviations

LC-MS: liquid chromatography-mass spectrometry

MS1: mass spectra

MS2: tandem mass spectra

MS3: triple-stage mass spectrometry

PTM: post-translational modification

DIA: data-independent acquisition

DDA: data-dependent acquisition

WWA: wide-window acquisition

XIC: extracted ion chromatogram

PSM: peptide-spectrum match

FDR: false discovery rate

NCE: normalized collision energy

CID: collision-induced dissociation

HCD: higher-energy collisional dissociation

PASEF: parallel accumulation serial fragmentation

LFQ: label-free quantification

MBR: match-between-run

CV: coefficient of variation

GO: gene ontology

TMT: tandem mass tag

PCA: principal component analysis

Author contributions

F.Y. and A.I.N developed the MSFragger-DDA+ algorithm. F.Y. implemented the algorithm in the software. F.Y., Y.D., and A.I.N. analyzed the results. F.Y. and A.I.N. wrote the manuscript with input from Y.D. A.I.N. and F.Y. conceived the study.

Data availability

The raw MS/MS files used in this study can be found at the ProteomeXchange Consortium and PRIDE[51] partner repository or at the MassIVE[52] repository with the following accession codes: PXD027242, MSV000090552, PXD041421, PXD037527, PXD045500, PXD024427. The parameter, log, and result files generated in this study can be found at XXX.

Code availability

The standalone version of MSFragger-DDA+ can be downloaded as part of MSFragger at https://msfragger.nesvilab.org/. FragPipe code is available at

https://github.com/Nesvilab/FragPipe. The program to generate the entrapment database is available at https://github.com/Nesvilab/EntrapBench. The Python and R scripts for processing the results and generating the figures are available at https://github.com/Nesvilab/MSFragger-DDAPlus-Manuscript.

Competing interests

A.I.N. and F.Y. receive royalties from the University of Michigan for the sale of MSFragger and IonQuant software licenses to commercial entities. All license transactions are managed by the University of Michigan Innovation Partnerships office, and all proceeds are subject to university technology transfer policy. Other authors declare no other competing interests.

References

1. Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M., Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J Proteome Res* **2010,** *9* (8), 4152-60.
2. Michalski, A.; Cox, J.; Mann, M., More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res* **2011,** *10* (4), 1785-93.
3. Truong, T.; Webber, K. G. I.; Madisyn Johnston, S.; Boekweg, H.; Lindgren, C. M.; Liang, Y.; Nydegger, A.; Xie, X.; Tsang, T. M.; Jayatunge, D.; Andersen, J. L.; Payne, S. H.; Kelly, R. T., Data-Dependent Acquisition with Precursor Coisolation Improves Proteome Coverage and Measurement Throughput for Label-Free Single-Cell Proteomics. *Angew Chem Int Ed Engl* **2023,** *62* (34), e202303415.
4. Matzinger, M.; Schmucker, A.; Yelagandula, R.; Stejskal, K.; Krssakova, G.; Berger, F.; Mechtler, K.; Mayer, R. L., Micropillar arrays, wide window acquisition and AI-based data analysis improve comprehensiveness in multiple proteomic applications. *Nat Commun* **2024,** *15* (1), 1019.
5. Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I., MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **2017,** *14* (5), 513-520.
6. 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. *J Am Soc Mass Spectrom* **1994,** *5* (11), 976-89.
7. Kim, S.; Pevzner, P. A., MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **2014,** *5*, 5277.
8. Eng, J. K.; Jahan, T. A.; Hoopmann, M. R., Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013,** *13* (1), 22-4.
9. Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R., ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005,** *5* (16), 4096-106.
10. Shteynberg, D.; Mendoza, L.; Hoopmann, M. R.; Sun, Z.; Schmidt, F.; Deutsch, E. W.; Moritz, R. L., reSpect: Software for Identification of High and Low Abundance Ion Species in Chimeric Tandem Mass Spectra. *Journal of the American Society for Mass Spectrometry* **2015,** *26* (11), 1837-1847.
11. Dorfer, V.; Maltsev, S.; Winkler, S.; Mechtler, K., CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research* **2018,** *17* (8), 2581-2589.
12. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* **2011,** *10* (4), 1794-1805.
13. Zhang, B.; Pirmoradian, M.; Chernobrovkin, A.; Zubarev, R. A., DeMix Workflow for Efficient Identification of Cofragmented Peptides in High Resolution Data-dependent Tandem Mass Spectrometry [>]. *Molecular & Cellular Proteomics* **2014,** *13* (11), 3211-3223.
14. Wang, J.; Bourne, P. E.; Bandeira, N., Peptide identification by database search of mixture tandem mass spectra. *Molecular & cellular proteomics : MCP* **2011,** *10* (12), M111.010017.
15. Kryuchkov, F.; Verano-Braga, T.; Hansen, T. A.; Sprenger, R. R.; Kjeldsen, F., Deconvolution of Mixture Spectra and Increased Throughput of Peptide Identification by

Utilization of Intensified Complementary Ions Formed in Tandem Mass Spectrometry. *Journal of Proteome Research* **2013,** *12* (7), 3362-3371.
16. Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J., High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal Chem* **2007,** *79* (15), 5620-32.
17. Yuan, Z. F.; Liu, C.; Wang, H. P.; Sun, R. X.; Fu, Y.; Zhang, J. F.; Wang, L. H.; Chi, H.; Li, Y.; Xiu, L. Y.; Wang, W. P.; He, S. M., pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *Proteomics* **2012,** *12* (2), 226-35.
18. Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M., Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *J Proteome Res* **2018,** *17* (5), 1844-1851.
19. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-8.
20. 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. *Nat Methods* **2020,** *17* (1), 41-44.
21. Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villen, J.; MacCoss, M. J., Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat Commun* **2018,** *9* (1), 5128.
22. Tsou, C. C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A. C.; Nesvizhskii, A. I., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **2015,** *12* (3), 258-64, 7 p following 264.
23. Yu, F.; Teo, G. C.; Kong, A. T.; Frohlich, K.; Li, G. X.; Demichev, V.; Nesvizhskii, A. I., Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat Commun* **2023,** *14* (1), 4154.
24. Li, K.; Teo, G. C.; Yang, K. L.; Yu, F.; Nesvizhskii, A. I., diaTracer enables spectrum-centric analysis of diaPASEF proteomics data. *bioRxiv* **2024**.
25. Frejno, M.; Berger, M. T.; Tüshaus, J.; Hogrebe, A.; Seefried, F.; Graber, M.; Samaras, P.; Fredj, S. B.; Sukumar, V.; Eljagh, L.; Brohnshtein, I.; Mamisashvili, L.; Schneider, M.; Gessulat, S.; Schmidt, T.; Kuster, B.; Zolg, D. P.; Wilhelm, M., Unifying the analysis of bottom-up proteomics data with CHIMERYS. *bioRxiv* **2024**, 2024.05.27.596040.
26. Tüshaus, J.; Eckert, S.; Fraefel, M.; Zhou, Y.; Pfeiffer, P.; Halves, C.; Fusco, F.; Weigl, J.; Hönikl, L.; Butenschön, V.; Todorova, R.; Rauert-Wunderlich, H.; The, M.; Rosenwald, A.; Heinemann, V.; Holch, J.; Meyer, B.; Weichert, W.; Mogler, C.; Hendrik-Kuhn, P.; Kuster, B., Towards routine proteome profiling of FFPE tissue: Insights from a 1,200 case pan-cancer study. *bioRxiv* **2024**, 2024.06.21.600043.
27. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** *20* (9), 1466-7.
28. Rockwood, A. L.; Van Orden, S. L., Ultrahigh-speed calculation of isotope distributions. *Anal Chem* **1996,** *68* (13), 2027-30.
29. Yu, F.; Haynes, S. E.; Nesvizhskii, A. I., IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Molecular & cellular proteomics : MCP* **2021,** *20*, 100077.
30. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **2008,** *26* (12), 1367–1372.
31. Yang, K. L.; Yu, F.; Teo, G. C.; Li, K.; Demichev, V.; Ralser, M.; Nesvizhskii, A. I., MSBooster: improving peptide identification rates using deep learning-based features. *Nat Commun* **2023,** *14* (1), 4539.

32. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **2007,** *4* (11), 923-5.
33. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **2002,** *74* (20), 5383-92.
34. Shteynberg, D. D.; Deutsch, E. W.; Campbell, D. S.; Hoopmann, M. R.; Kusebauch, U.; Lee, D.; Mendoza, L.; Midha, M. K.; Sun, Z.; Whetton, A. D.; Moritz, R. L., PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. *J Proteome Res* **2019,** *18* (12), 4262-4272.
35. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003,** *75* (17), 4646-58.
36. da Veiga Leprevost, F.; Haynes, S. E.; Avtonomov, D. M.; Chang, H. Y.; Shanmugam, A. K.; Mellacheruvu, D.; Kong, A. T.; Nesvizhskii, A. I., Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods* **2020,** *17* (9), 869-870.
37. Li, K.; Vaudel, M.; Zhang, B.; Ren, Y.; Wen, B., PDV: an integrative proteomics data viewer. *Bioinformatics* **2019,** *35* (7), 1249-1251.
38. Searle, B. C.; Shannon, A. E.; Wilburn, D. B., Scribe: Next Generation Library Searching for DDA Experiments. *J Proteome Res* **2023,** *22* (2), 482-490.
39. Richards, A. L.; Chen, K. H.; Wilburn, D. B.; Stevenson, E.; Polacco, B. J.; Searle, B. C.; Swaney, D. L., Data-Independent Acquisition Protease-Multiplexing Enables Increased Proteome Sequence Coverage Across Multiple Fragmentation Modes. *J Proteome Res* **2022,** *21* (4), 1124-1136.
40. Wen, B.; Freestone, J.; Riffle, M.; MacCoss, M. J.; Noble, W. S.; Keich, U., Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *bioRxiv* **2024**.
41. Wang, H.; Lim, K. P.; Kong, W.; Gao, H.; Wong, B. J. H.; Phua, S. X.; Guo, T.; Goh, W. W. B., MultiPro: DDA-PASEF and diaPASEF acquired cell line proteomic datasets with deliberate batch effects. *Sci Data* **2023,** *10* (1), 858.
42. Yu, F.; Haynes, S. E.; Teo, G. C.; Avtonomov, D. M.; Polasky, D. A.; Nesvizhskii, A. I., Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Molecular & cellular proteomics : MCP* **2020,** *19* (9), 1575-1585.
43. Bader, J. M.; Deigendesch, N.; Misch, M.; Mann, M.; Koch, A.; Meissner, F., Proteomics separates adult-type diffuse high-grade gliomas in metabolic subgroups independent of 1p/19q codeletion and across IDH mutational status. *Cell Rep Med* **2023,** *4* (1), 100877.
44. Hsiao, Y.; Zhang, H.; Li, G. X.; Deng, Y.; Yu, F.; Kahrood, H. V.; Steele, J. R.; Schittenhelm, R. B.; Nesvizhskii, A. I., Analysis and visualization of quantitative proteomics data using FragPipe-Analyst. *bioRxiv* **2024**, 2024.03.05.583643.
45. Kuleshov, M. V.; Jones, M. R.; Rouillard, A. D.; Fernandez, N. F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S. L.; Jagodnik, K. M.; Lachmann, A.; McDermott, M. G.; Monteiro, C. D.; Gundersen, G. W.; Ma'ayan, A., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **2016,** *44* (W1), W90-7.
46. Yu, F.; Teo, G. C.; Kong, A. T.; Haynes, S. E.; Avtonomov, D. M.; Geiszler, D. J.; Nesvizhskii, A. I., Identification of modified peptides using localization-aware open search. *Nat Commun* **2020,** *11* (1), 4065.
47. Polasky, D. A.; Yu, F.; Teo, G. C.; Nesvizhskii, A. I., Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat Methods* **2020,** *17* (11), 1125-1132.
48. Polasky, D. A.; Geiszler, D. J.; Yu, F.; Li, K.; Teo, G. C.; Nesvizhskii, A. I., MSFragger-Labile: A Flexible Method to Improve Labile PTM Analysis in Proteomics. *Molecular & cellular proteomics : MCP* **2023,** *22* (5), 100538.

49. Ting, L.; Rad, R.; Gygi, S. P.; Haas, W., MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **2011,** *8* (11), 937-40.
50. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012,** *30* (10), 918-20.
51. Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, Ş.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A., The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* **2019,** *47* (Database issue), D442-D442.
52. Choi, M.; Carver, J.; Chiva, C.; Tzouros, M.; Huang, T.; Tsai, T. H.; Pullman, B.; Bernhardt, O. M.; Huttenhain, R.; Teo, G. C.; Perez-Riverol, Y.; Muntel, J.; Muller, M.; Goetze, S.; Pavlou, M.; Verschueren, E.; Wollscheid, B.; Nesvizhskii, A. I.; Reiter, L.; Dunkley, T.; Sabido, E.; Bandeira, N.; Vitek, O., MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. *Nat Methods* **2020,** *17* (10), 981-984.

Figure 1. Overview of MSFragger-DDA+ and FragPipe. (a) MSFragger-DDA+ algorithm. Each tandem mass spectrum is searched against all peptides within the isolation window. Subsequently, MSFragger-DDA+ detects the precursor signals for each of the matched PSMs. After filtering out the PSMs with low-quality precursor signals, MSFragger-DDA+ rescores the PSMs using a greedy algorithm. **(b)** MSFragger-DDA+ workflow in FragPipe software. The workflow contains MSFragger-DDA+ for database searching, MSBooster for deep-learning-based rescoring, Percolator for PSM ranking, PorteinProphet for protein inference, FDR filtering, IonQuant for label-free quantification (optional), and EasyPQP for spectral library generation (optional).

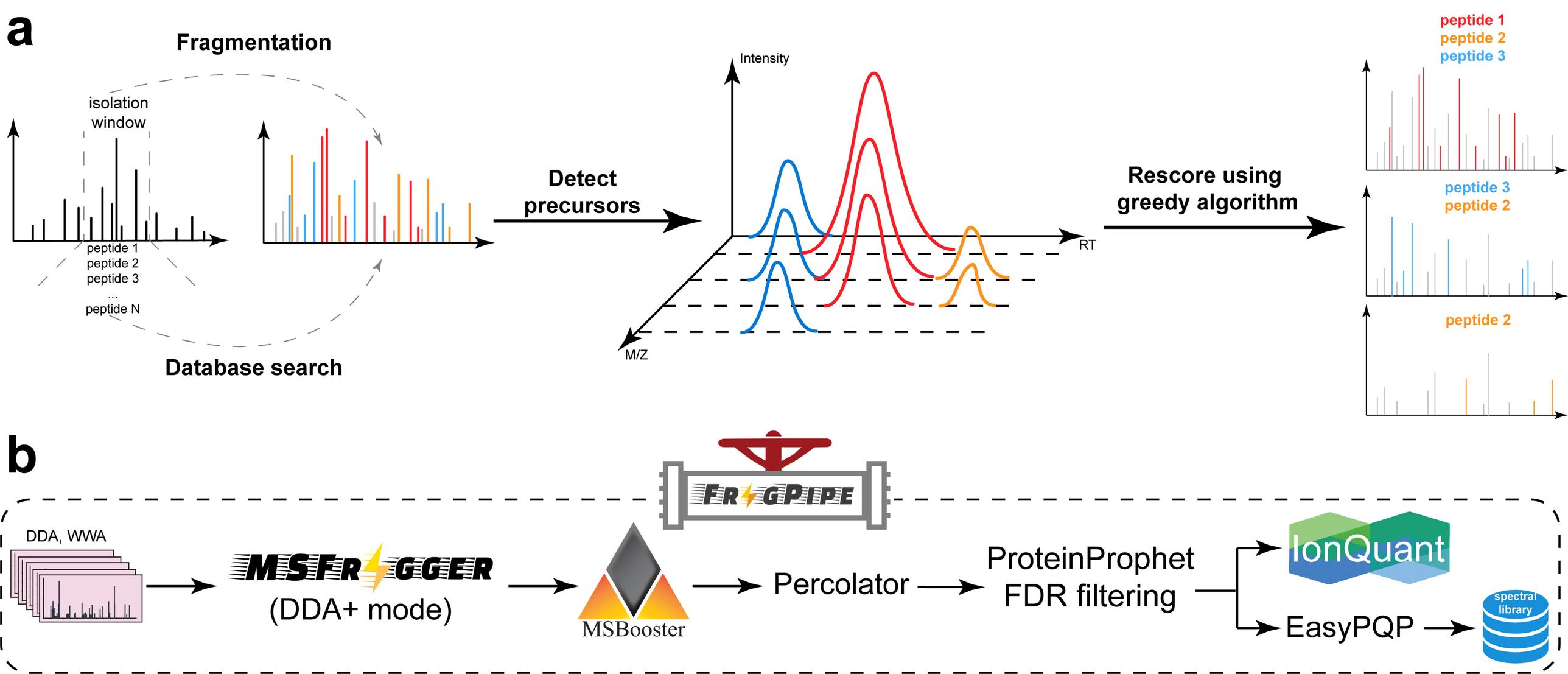
Figure 2. Sensitivity, speed, and false discovery proportion assessment using traditional DDA datasets with different NCE and enzymatic digestions. (a) and **(b)** Number of peptide sequences identified by MaxQuant, MetaMorpheus, MSFragger, MSFragger-DDA+, and Scribe. **(a)** Data with different NCE. **(b)** Data with different enzymatic digestions. **(c)** Runtime of MaxQuant, MetaMorpheus, MSFragger, and MSFragger-DDA+. **(d)** Protein-level FDP evaluation for MaxQuant, MetaMorpheus, MSFragger, and MSFragger-DDA+. Two calculation methods, including the upper bound and lower bound, were applied.

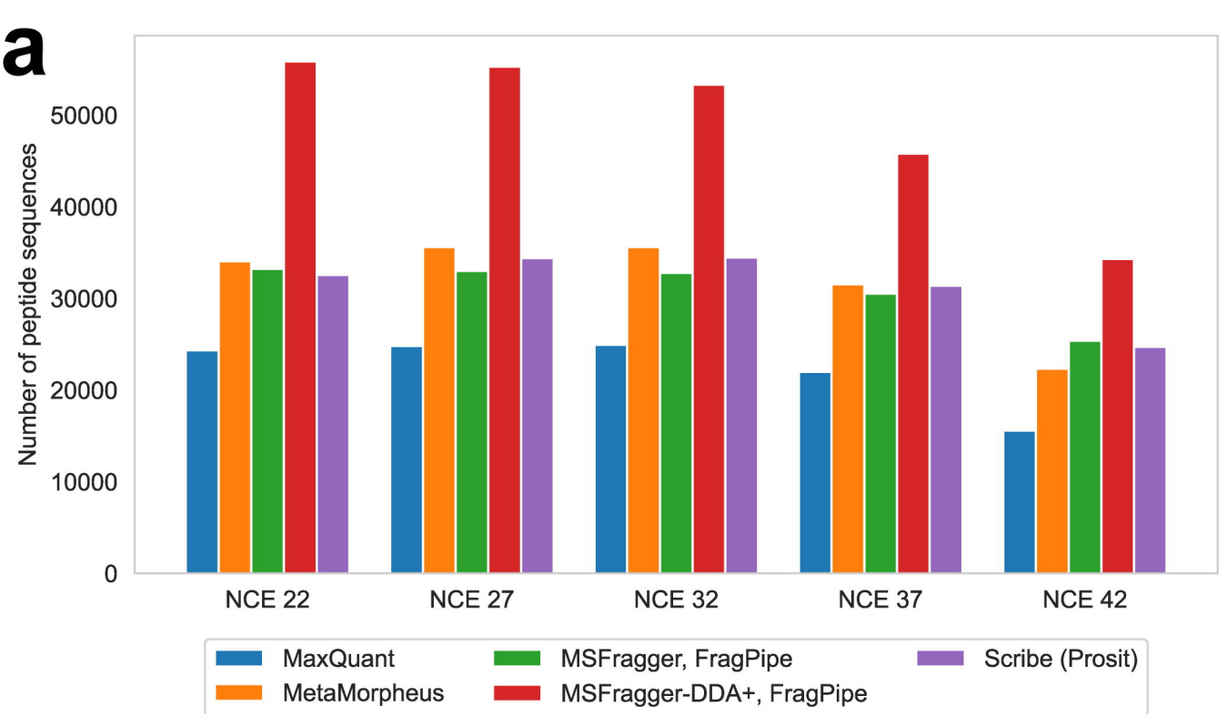
Figure 3. Performance benchmarking using timsTOF ddaPASEF data. (a) Number of quantified proteins from the DDA and DDA+ workflows in the A549 cell line dataset (with three biological replicates, and four technical replicates for each). “MBR+” and “MBR-” refer to IonQuant run with and without MBR, respectively. **(b)** Numbers and CVs of overlapped and non-overlapped proteins quantified from the DDA and DDA+ workflows using the A549 cell line. The blue box plots are from the unique proteins of the DDA mode, the green box plots are from the common proteins of the DDA mode, the red box plots are from the common proteins of the DDA+ mode, and the yellow box plots are from the unique proteins of the DDA+ mode. The common proteins are the overlapping proteins quantified in both DDA and DDA+ modes. The numbers on the right are the quantified proteins. The box in each plot captures the interquartile range (IQR) with the bottom and top edges representing the first (Q1) and third quartiles (Q3), respectively. The median (Q2) is indicated by a horizontal line within the box. The whiskers extend to the minima and maxima within 1.5 times the IQR below Q1 or above Q3. **(c)** Two plots showing the protein non-missing values and missing values from the DDA and DDA+ workflows.

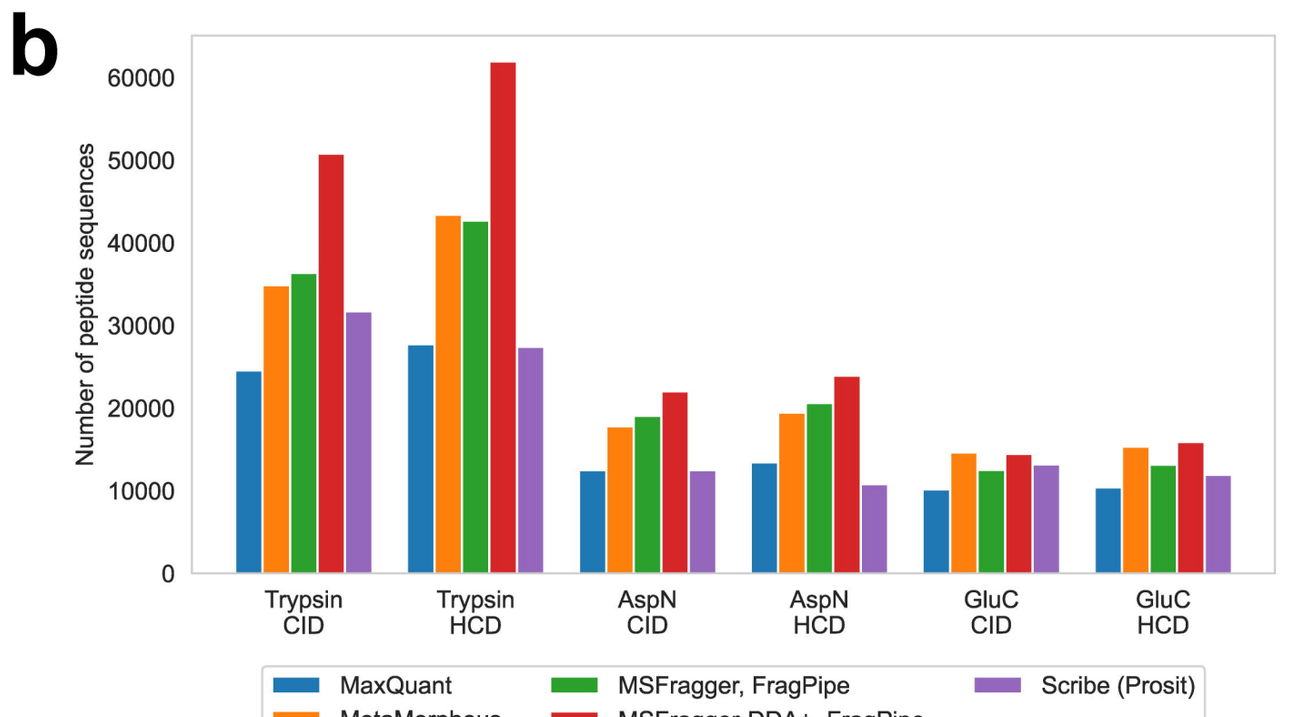
A549 cell line data. MBR is disabled. The number of columns equals to the number of proteins quantified in the specific setting and listed at the top of each plot. The proteins with non-zero intensities are in orange, and the proteins with zero intensities are in black. **(d)** Similar to **(c)** but the MBR is enabled.

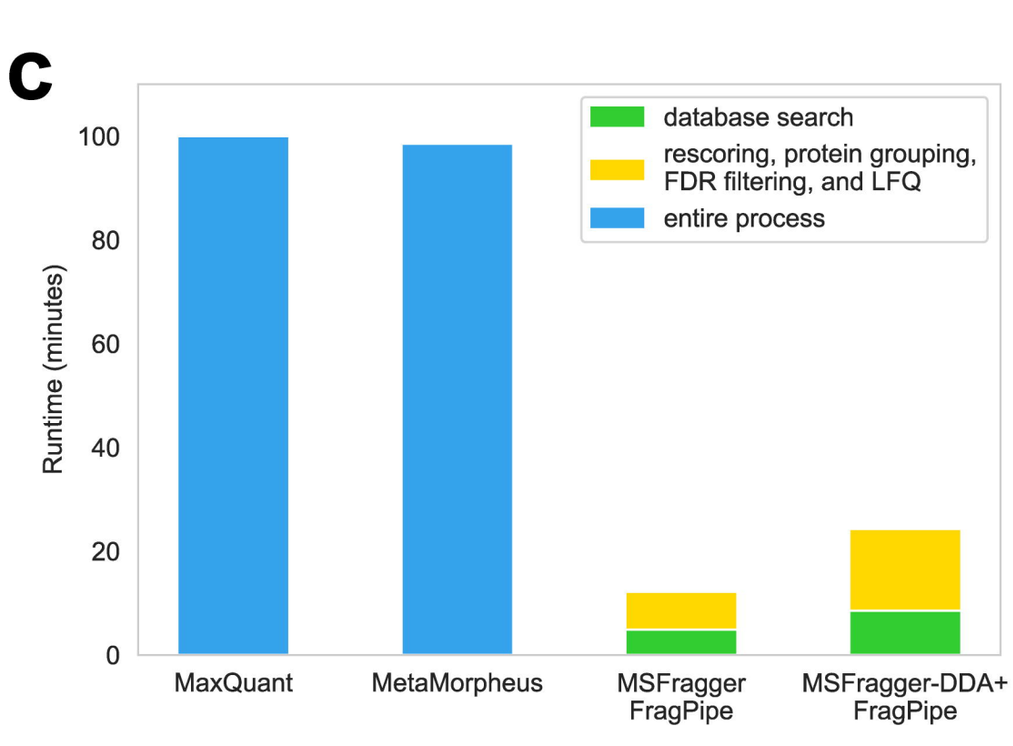
Figure 4. Sensitivity assessment using three WWA datasets. (a) and **(b)** Numbers of peptides from the first WWA dataset of Truong et al. There are 14 samples with different isolation windows, maximum injection time, and MS2 resolutions. Each sample contain two technical replicates. MBR is disabled. **(c)** and **(d)** Numbers and CVs of quantified proteins from the second dataset of Truong et al. The dark color is with CV < 20% and the light color is with CV ≥ 20%. The samples are from K562 and HeLa cell lines, respectively. There are four samples with different combinations of cell line and gradient length. Each sample has eight technical replicates. MBR is disabled. **(e)** and **(f)** Numbers of identified peptides from the WWA dataset published by Matzinger et al. There are 17 samples with different isolation windows and sample amounts. Each sample has three technical replicates.

Figure 5. Performance demonstration using the glioma dataset. (a) Heatmap of the gene intensities from the DDA+ workflow. **(b)** PCA plot of the quantitative results from the DDA+ workflow. **(c)** Volcano plot using the gene intensities from the DDA+ workflow. **(d)** Box plots showing the percentage of protein-level missing values from the DDA and DDA+ workflows. The box in each plot captures the IQR with the bottom and top edges representing the Q1 and Q3, respectively. The median (Q2) is indicated by a horizontal line within the box. The whiskers extend to the minima and maxima within 1.5 times the IQR below Q1 or above Q3. **(e)** Venn diagrams showing the number of upregulated and downregulated genes in the DDA and DDA+ workflows.


a
Fragmentation
isolation window
peptide 1
peptide 2
peptide 3
...
peptide N
Database search
Detect precursors
Intensity
RT
M/Z
Rescore using greedy algorithm
peptide 1
peptide 2
peptide 3
peptide 3
peptide 2
peptide 2
b
FRAGPIPE
DDA, WWA
MSFRAGGER
(DDA+ mode)
MSBooster
Percolator
ProteinProphet
FDR filtering
IonQuant
EasyPQP
spectral library

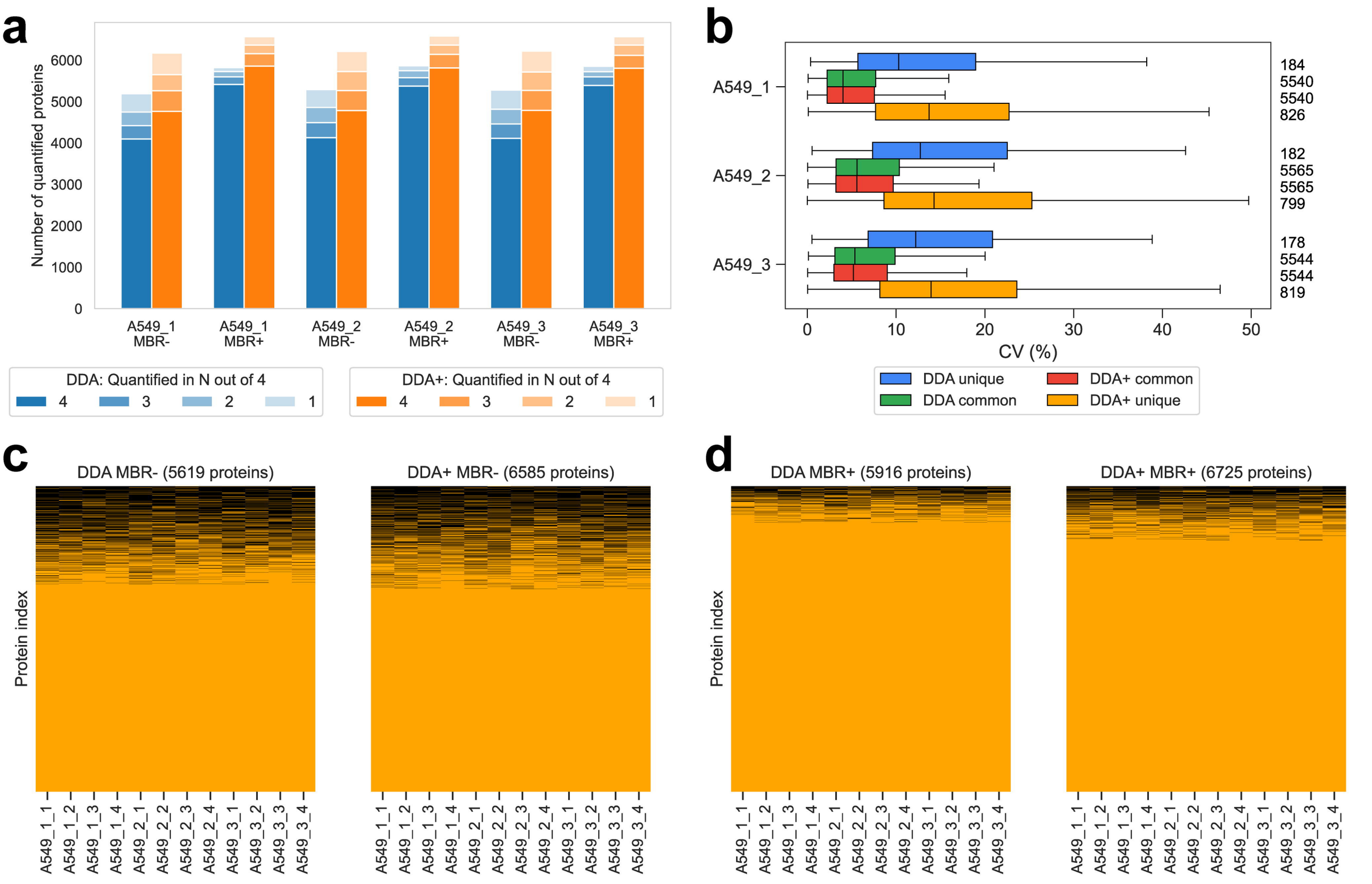


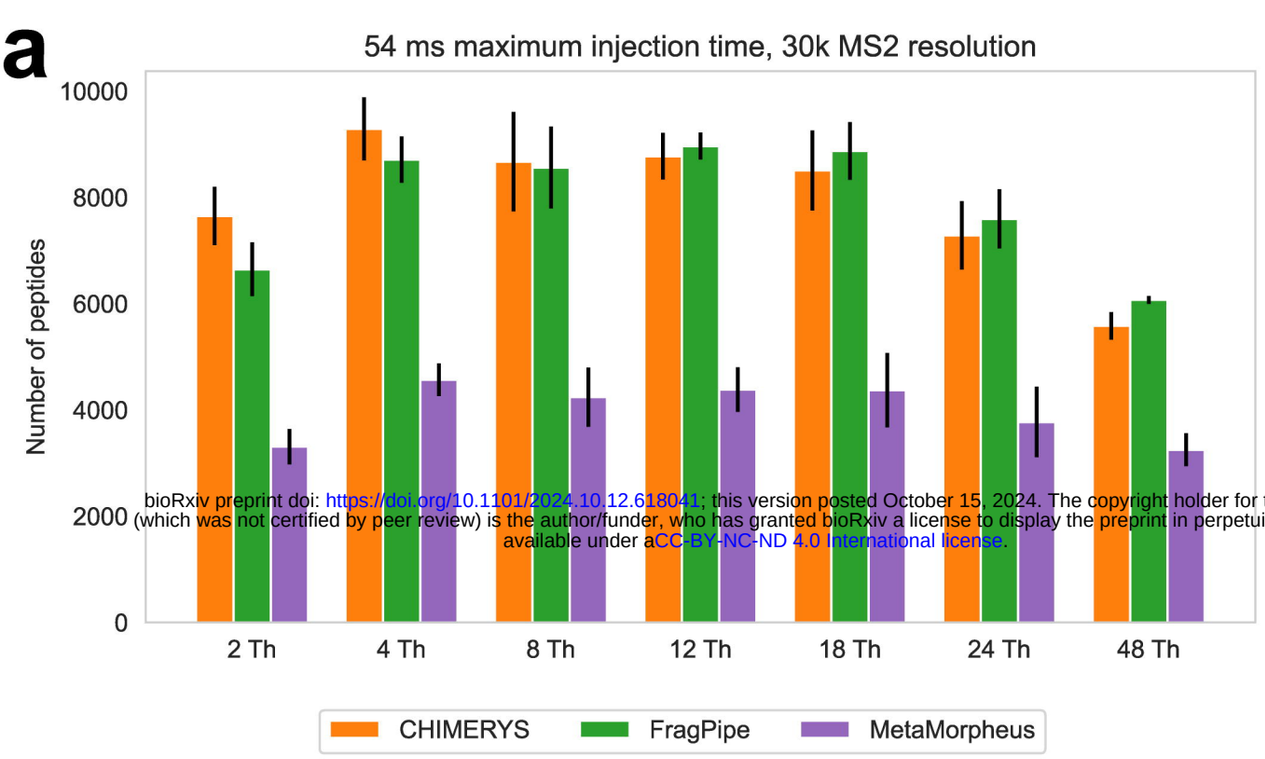


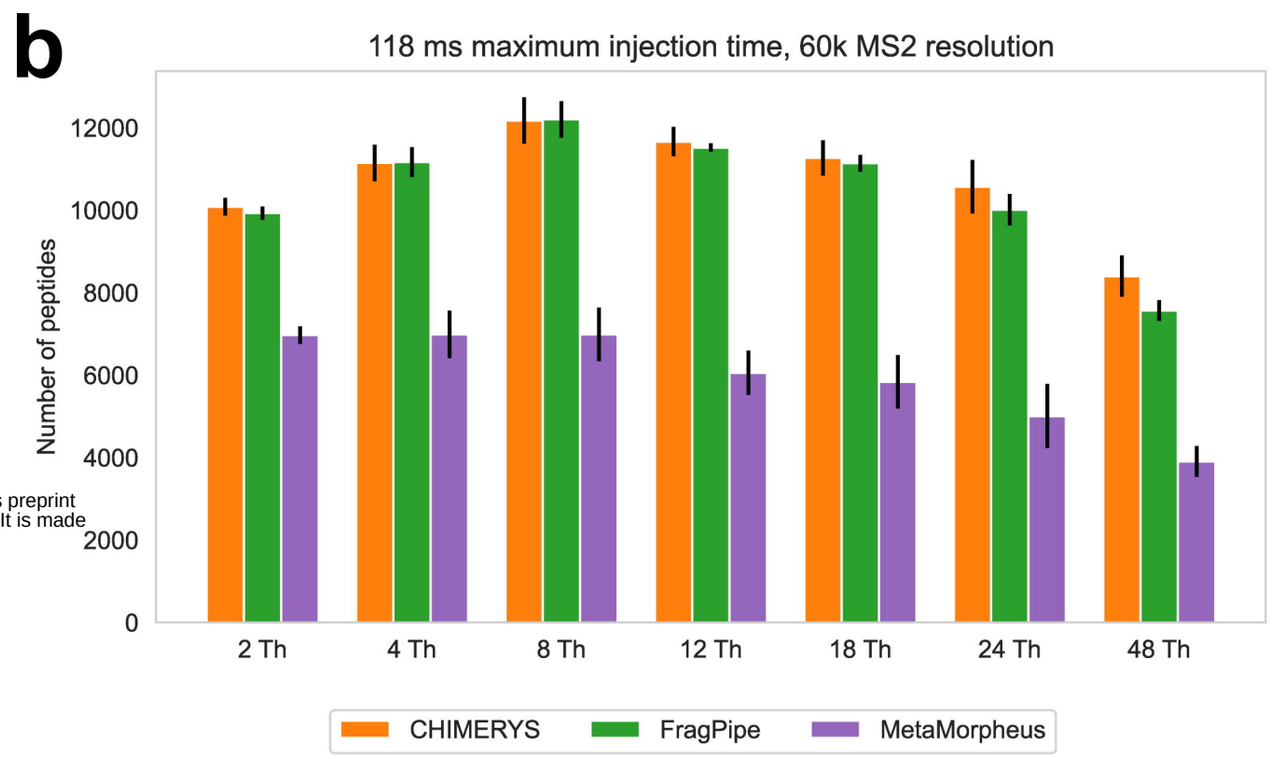


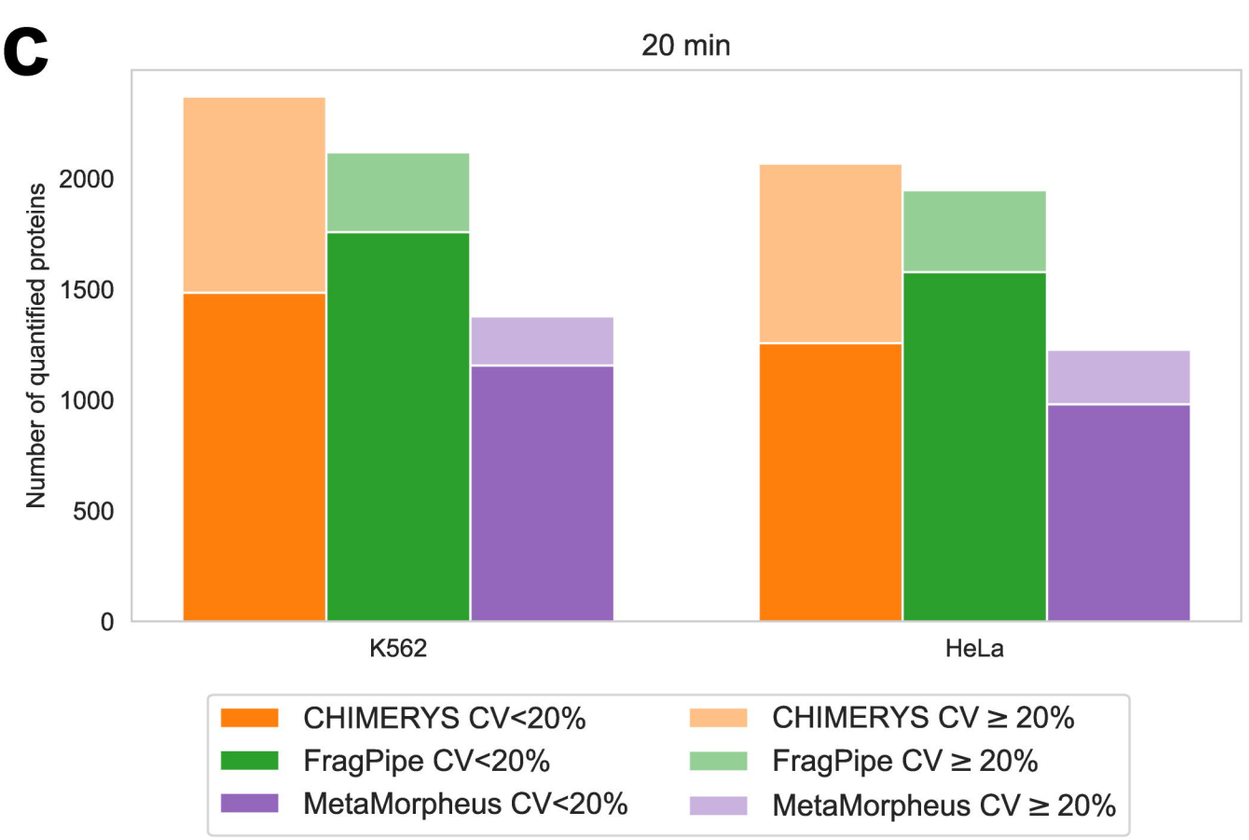
d

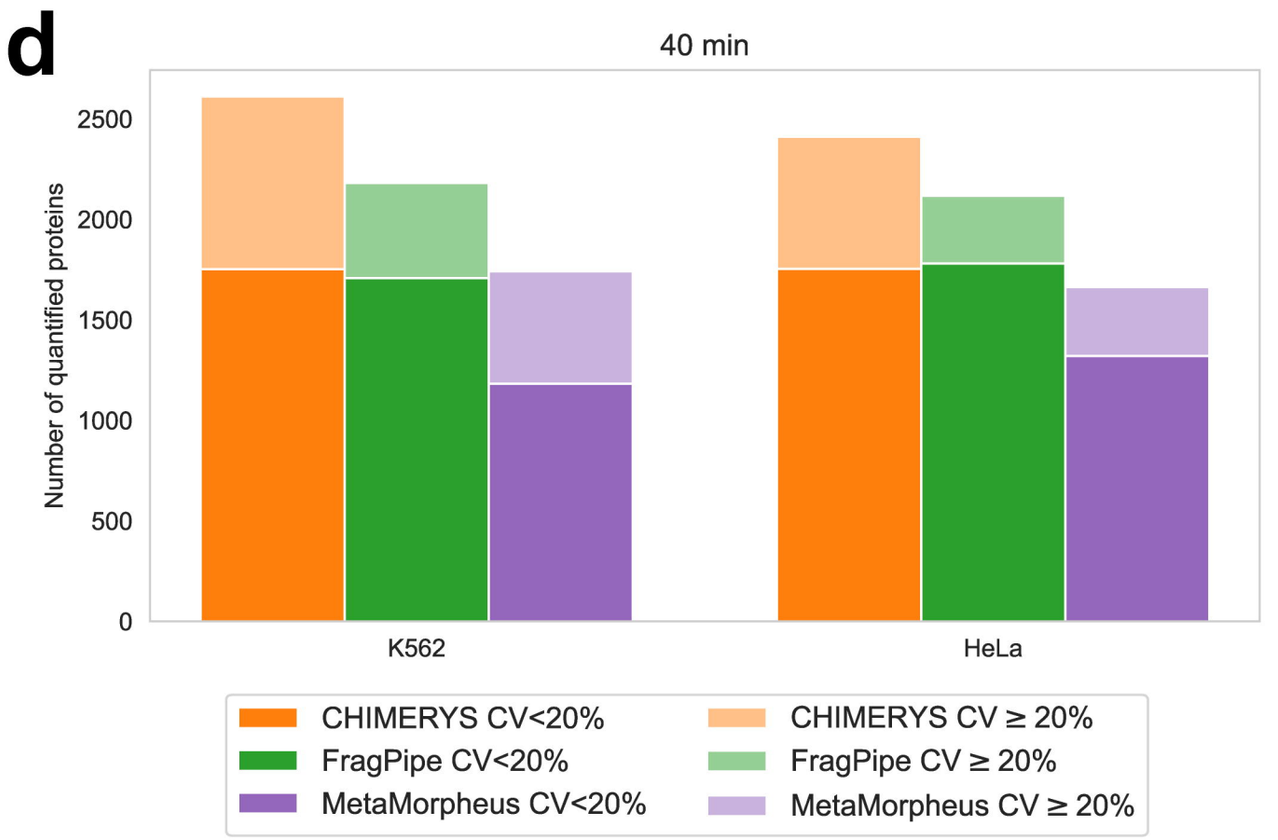
Tool	Target proteins	Entrapment proteins	FDP estimation	
			upper bound	lower bound
MaxQuant	3509	27	1.53%	0.76%
MetaMorpheus	4345	68	3.08%	1.54%
MSFragger FragPipe	4522	16	0.71%	0.35%
MSFragger-DDA+ FragPipe	5656	29	1.02%	0.51%

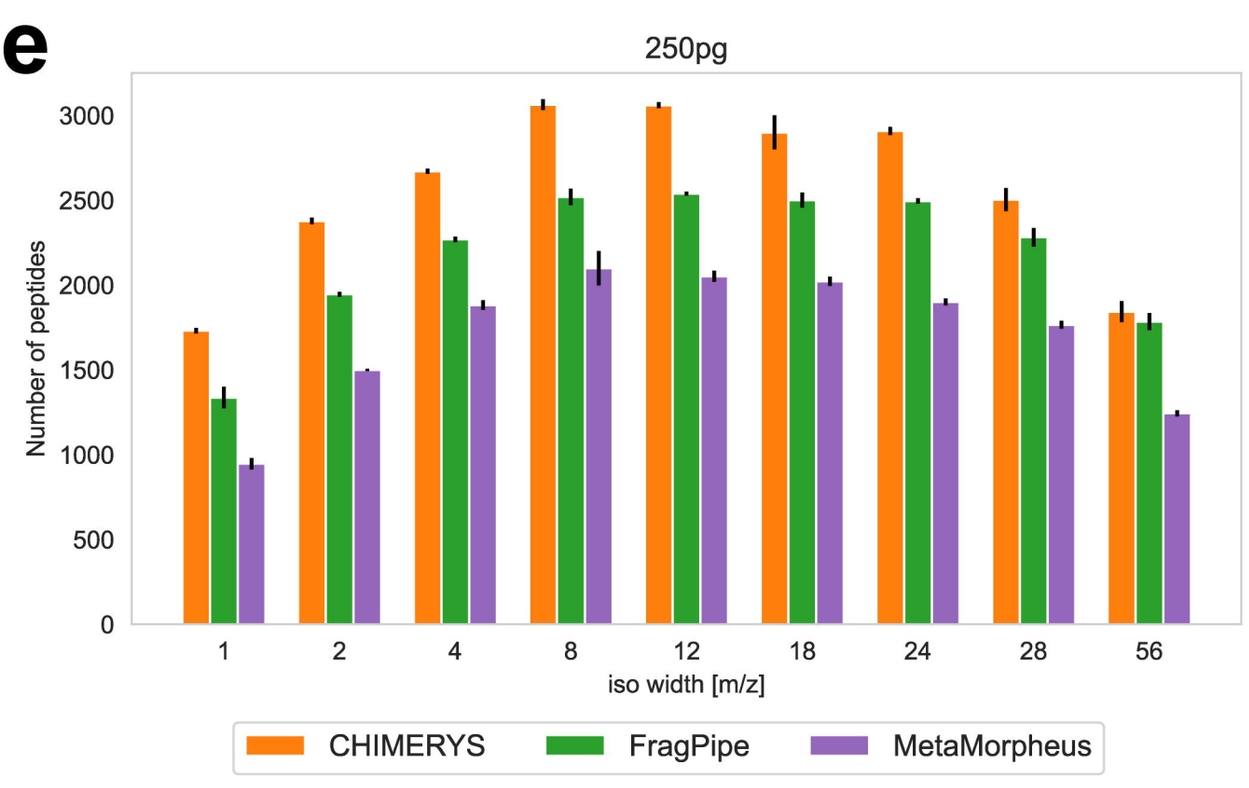

a
Number of quantified proteins
0
1000
2000
3000
4000
5000
6000
A549_1 MBR-
A549_1 MBR+
A549_2 MBR-
A549_2 MBR+
A549_3 MBR-
A549_3 MBR+
DDA: Quantified in N out of 4
4
3
2
1
DDA+: Quantified in N out of 4
4
3
2
1
b
A549_1
A549_2
A549_3
184
5540
5540
826
182
5565
5565
799
178
5544
5544
819
0
10
20
30
40
50
CV (%)
DDA unique
DDA common
DDA+ common
DDA+ unique
c
DDA MBR- (5619 proteins)
DDA+ MBR- (6585 proteins)
Protein index
A549_1_1
A549_1_2
A549_1_3
A549_1_4
A549_2_1
A549_2_2
A549_2_3
A549_2_4
A549_3_1
A549_3_2
A549_3_3
A549_3_4
d
DDA MBR+ (5916 proteins)
DDA+ MBR+ (6725 proteins)
Protein index
A549_1_1
A549_1_2
A549_1_3
A549_1_4
A549_2_1
A549_2_2
A549_2_3
A549_2_4
A549_3_1
A549_3_2
A549_3_3
A549_3_4

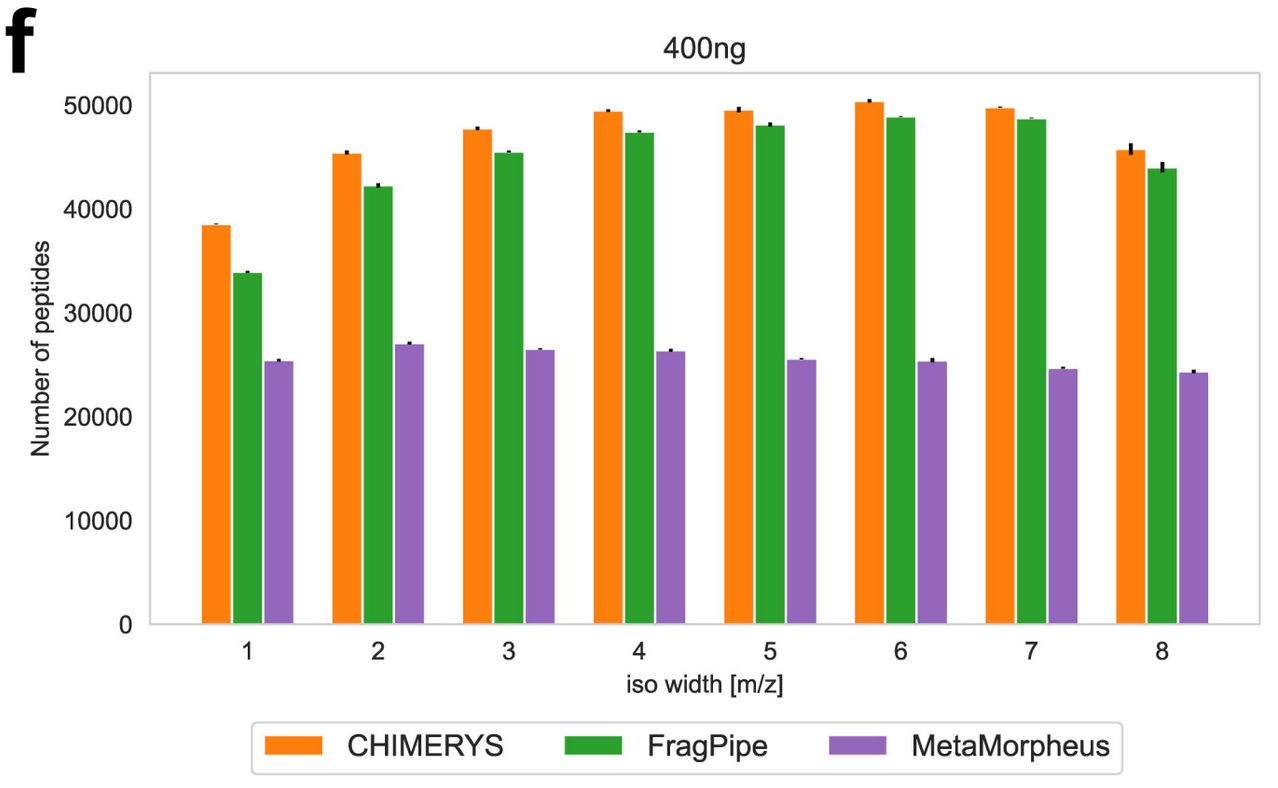


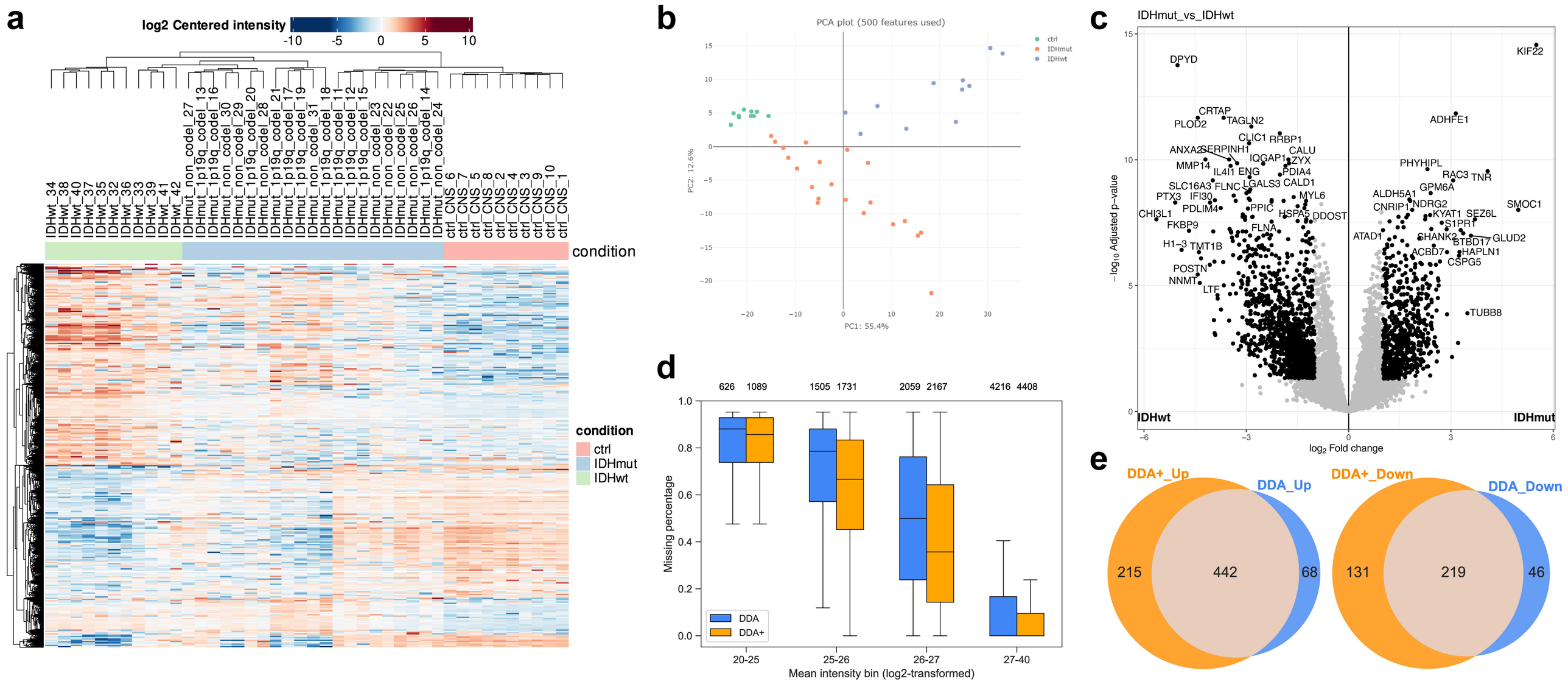











a
log2 Centered intensity
-10
-5
0
5
10
IDHwt_34
IDHwt_38
IDHwt_40
IDHwt_37
IDHwt_35
IDHwt_32
IDHwt_36
IDHwt_33
IDHwt_39
IDHwt_41
IDHwt_42
IDHmut_non_codel_27
IDHmut_1p19q_codel_13
IDHmut_1p19q_codel_16
IDHmut_non_codel_30
IDHmut_non_codel_29
IDHmut_1p19q_codel_20
IDHmut_non_codel_28
IDHmut_1p19q_codel_21
IDHmut_1p19q_codel_17
IDHmut_1p19q_codel_19
IDHmut_non_codel_31
IDHmut_1p19q_codel_18
IDHmut_1p19q_codel_11
IDHmut_1p19q_codel_12
IDHmut_1p19q_codel_15
IDHmut_1p19q_codel_23
IDHmut_non_codel_22
IDHmut_non_codel_25
IDHmut_non_codel_26
IDHmut_1p19q_codel_14
IDHmut_non_codel_24
ctrl_CNS_6
ctrl_CNS_7
ctrl_CNS_5
ctrl_CNS_8
ctrl_CNS_2
ctrl_CNS_4
ctrl_CNS_3
ctrl_CNS_9
ctrl_CNS_10
ctrl_CNS_1
condition
condition
ctrl
IDHmut
IDHwt
b
PCA plot (500 features used)
ctrl
IDHmut
IDHwt
PC2: 12.6%
PC1: 55.4%
c
IDHmut_vs_IDHwt
DPYD
KIF22
CRTAP
PLOD2
TAGLN2
ADHFE1
CLIC1
RRBP1
ANXA2
SERPINH1
CALU
IQGAP1
ZYX
MMP14
IL4I1
ENG
PDIA4
PHYHIPL
RAC3
TNR
SLC16A3
FLNC
LGALS3
CALD1
GPM6A
PTX3
IFI30
MYL6
ALDH5A1
CNRIP1
NDRG2
SMOC1
PDLIM4
PPIC
HSPA5
DDOST
KYAT1
SEZ6L
CHI3L1
S1PR1
FKBP9
FLNA
ATAD1
SHANK2
GLUD2
H1–3
TMT1B
BTBD17
ACBD7
HAPLN1
POSTN
CSPG5
NNMT
LTF
TUBB8
−log10 Adjusted p-value
log2 Fold change
IDHwt
IDHmut
d
626 1089
1505 1731
2059 2167
4216 4408
Missing percentage
DDA
DDA+
20-25
25-26
26-27
27-40
Mean intensity bin (log2-transformed)
e
DDA+_Up
DDA_Up
215
442
68
DDA+_Down
DDA_Down
131
219
46