diaTracer enables spectrum-centric analysis of diaPASEF proteomics data

Kai Li[1], Guo Ci Teo[2], Kevin L. Yang[1], Fengchao Yu[2*], Alexey I. Nesvizhskii[1,2*]

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

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

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

Abstract

Data-independent acquisition (DIA) has become a widely used strategy for peptide and protein quantification in mass spectrometry-based proteomics studies. The integration of ion mobility separation into DIA analysis, such as the diaPASEF technology available on Bruker's timsTOF platform, further improves the quantification accuracy and protein depth achievable using DIA. We introduce diaTracer, a new spectrum-centric computational tool optimized for diaPASEF data. diaTracer performs three-dimensional (m/z, retention time, ion mobility) peak tracing and feature detection to generate precursor-resolved "pseudo-MS/MS" spectra, facilitating direct ("spectral-library free") peptide identification and quantification from diaPASEF data. diaTracer is available as a stand-alone tool and is fully integrated into the widely used FragPipe computational platform. We demonstrate the performance of diaTracer and FragPipe using diaPASEF data from triple-negative breast cancer (TNBC), cerebrospinal fluid (CSF), and plasma samples, data from phosphoproteomics and HLA immunopeptidomics experiments, and low-input data from a spatial proteomics study. We also show that diaTracer enables unrestricted identification of post-translational modifications from diaPASEF data using open/mass-offset searches.

Introduction

Liquid chromatography-mass spectrometry (LC-MS) is a commonly used method to analyze proteomics data[1]. There are two primary approaches for acquiring high-throughput data: data-dependent acquisition (DDA) and data-independent acquisition (DIA). DDA isolates and fragments each peptide ion individually to produce a tandem mass spectrum (MS/MS) that contains information on the peptide sequence. In contrast, DIA isolates and fragments multiple peptide ions within specific mass and retention time ranges, resulting in more complex spectra[2-4]. As mass spectrometers have become faster and more sensitive over the years, the popularity of DIA has increased. Furthermore, the use of ion mobility (IM) separation, in addition to liquid chromatography, to further separate peptides and reduce co-elution[5] has made DIA an even more attractive strategy. The Bruker timsTOF mass spectrometer platform, which couples trapped ion mobility separation with the parallel accumulation serial fragmentation (PASEF)[6] technique and supports both DDA (ddaPASEF) and DIA (diaPASEF) modes, has proven to offer excellent performance for a wide range of quantitative proteomics applications.

Although ion mobility separation helps resolve co-eluted peptides, it presents new challenges for computational analyses. Most peptide identification and quantification tools were not initially designed to support additional ion mobility dimensions. We have proposed an efficient solution[7, 8] for analyzing timsTOF data in DDA mode, but there are still open questions regarding the analysis of diaPASEF data generated by timsTOF. The first critical step in any DIA workflow is the creation of a so-called “spectral library” containing a list of peptide ions to be quantified in the acquired DIA data. For each target peptide ion, the library contains its separation coordinates (LC retention time and ion mobility value) and an empirical or predicted fragmentation spectrum (fragment m/z values and intensities) that are used to locate and quantify the peptide ion signal in each of the acquired DIA runs. A spectral library can be built using DDA data acquired from the same or similar samples, often using additional fractionation steps to increase the proteome depth. This remains an effective strategy for building spectral libraries, including for diaPASEF data analysis[9], however, additional efforts are required to acquire DDA data for building the library.

An alternative computational strategy, first proposed by us and implemented in the DIA-Umpire algorithm[10], is to identify peptides from DIA data directly and then use these identifications to build a spectral library for the subsequent targeted protein quantification step. Known as direct

DIA or “library-free” strategy (contrasting with the requirement of having a DDA-based library in some DIA analysis tools[11, 12]), it has now become the main mode of analysis of DIA data in a typical shotgun proteomics experiment. Direct identification of peptides from DIA data can be performed using either the spectrum-centric approach of DIA-Umpire[10, 13] (also adopted by Spectronaut[14]), which starts with the detection of extracted ion chromatogram (XIC) features to generate DDA-like pseudo-MS/MS spectra[10] that are compatible with conventional search engines such as MSFragger. Alternatively, direct identification can be performed using a peptide-centric data analysis strategy[15-17], including the library-free mode of DIA-NN[9, 18]. Furthermore, MSFragger-DIA[19] implements a hybrid strategy in which full DIA MS/MS scans are searched against a protein sequence database[19, 20], followed by targeted precursor and fragment peak tracing and rescoring of the top peptide candidates for each MS/MS scan[19]. Both spectrum-centric and peptide-centric strategies take advantage of deep-leaning-based predictions[21, 22] of peptide spectra and separation coordinates (LC and IM), but with significant differences in the use of prediction models. In the fully spectrum-centric approach and in the hybrid MSFragger-DIA strategy, predictions of peptide properties are performed only for a restricted set of top scoring peptides identified from the data[19, 21], whereas peptide-centric methods such as DIA-NN require *in-silico* predictions for all possible candidate peptides given the user-specific protein sequence database and search parameters. One important advantage of the spectrum-centric DIA-Umpire strategy of deconvoluting full DIA MS/MS spectra into pseudo-MS/MS spectra is that it facilitates analyses requiring a large search space. These include studies focusing on post-translational modifications (PTMs), studies requiring nonspecific (e.g., HLA or endogenous peptidomics) or semi-enzymatic (e.g., N-terminomic) searches, or when spectra and/or retention times of chemically labelled peptides cannot be accurately predicted (e.g., in chemoproteomics). However, although successful and frequently used for the analysis of DIA data from the Thermo Fisher Scientific Orbitrap and Sciex tripleTOF mass spectrometry platforms, neither DIA-Umpire nor MSFragger-DIA are capable of processing diaPASEF data because of the challenges of detecting XIC features with the additional ion mobility dimension. Although Spectronaut supports direct identification from diaPASEF data the details of the algorithm, to the best of our knowledge, have not been publicly disclosed.

To address the need for an efficient and versatile method for analyzing diaPASEF data, we developed a new spectrum-centric tool called diaTracer. diaTracer has been designed to perform highly efficient processing of three-dimensional (m/z, retention time, ion mobility)

information encoded in diaPASEF data to detect precursor and fragment ion features, followed by the generation of pseudo-MS/MS spectra that can be easily analyzed with peptide identification tools such as MSFragger[23]. diaTracer was fully integrated into the widely used FragPipe computational platform. We demonstrate the performance of diaTracer using diaPASEF data from triple-negative breast cancer (TNBC), cerebrospinal fluid (CSF), and plasma proteome datasets, using data from phosphoproteomics and immunopeptidomics experiments, and low-input data from a spatial proteomics study. We also show that our strategy enables unrestricted PTM searches via the application of the open/mass-offset mode of MSFragger to diaTracer-extracted files. Overall, our study demonstrates that diaTracer and FragPipe computational platform enable fast and sensitive analysis of diaPASEF data from a wide range of biological applications.

Results

diaTracer workflow in FragPipe

diaTracer takes diaPASEF data (as input .d files) and generates pseudo-MS/MS spectra[24] (in mzML format) for DDA-like database searching. It employs a three-dimensional feature detection method that leverages retention time and ion mobility dimensions to enhance signals and reduce noise interference. In our previous study[25], we demonstrated that summing RT frames with closed ion mobilities contributed to the accurate identification of the correct XIC features. Therefore, diaTracer starts with the aggregation of neighboring RT frames, followed by a two-dimensional (m/z-1/K0) adaptive Gaussian feature detection algorithm for the merged frames. The same approach is used to detect all precursor and fragment XIC features. Ion mobility and m/z values of the signals are determined using the apex. An isotope grouping method filters precursor signals and identifies potential charge states in the MS1 data. diaTracer then calculates the Pearson correlation coefficient of the XIC features between all detected precursors and fragments within user-defined ion mobility and retention time tolerance to assemble pseudo-MS/MS spectra (see **Methods**).

diaTracer was fully integrated into FragPipe for streamlined, direct analysis of diaPASEF data (**Figure 1**). FragPipe (https://fragpipe.nesvilab.org) is a comprehensive computational platform that automates all the data analysis steps. The pseudo-MS/MS spectra extracted by diaTracer

are searched using MSFragger[26] in the DDA mode, enabling tryptic, semi-tryptic, nonspecific, mass-offset, or even open database searches for comprehensive PTM characterization[27-29]. Peptide-spectrum matching (PSM) with MSFragger[23, 27] is followed by deep learning-based rescoring of PSMs using MSBooster[21] and Percolator[30] (default option; alternatively, using PeptideProphet[31]), protein inference with ProteinProphet[32], and FDR filtering (by default 1% FDR at the PSM, ion, peptide, and protein levels) using Philosopher[33]. PTMProphet[34] is enabled for PTM workflows that require site localization (e.g., phosphoproteomics). The spectral library is generated using EasyPQP, followed by the extraction of quantification from the DIA data using DIA-NN[9, 18] (alternatively, using Skyline[35]). Furthermore, a hybrid spectral library[10, 19] can also be generated from both diaPASEF data and additional ddaPASEF data when available. Users have a choice to run FragPipe using an easy-to-use graphical user interface (GUI) or a command line version of the pipeline that is fully compatible with high-performance computing or cloud-based environments. A data visualization module, FragPipe-PDV (an extension of the PDV viewer[36]), is integrated into FragPipe and provides a convenient way to view pseudo-MS/MS spectra and peptide-to-spectrum matches. Quantitative matrices (with quantification extracted by DIA-NN) can be conveniently analyzed using FragPipe-Analyst[37] or MSStats[38].

Performance evaluation: Deep proteome profiling

We first sought to assess the performance of FragPipe with diaTracer on a dataset representative of typical, deep proteome profiling experiments. We downloaded a triple-negative breast cancer (TNBC) dataset[39] generated on a timsTOF Pro platform and containing 1) ddaPASEF data from 12 HILIC peptide fractions of a pooled TMBC sample and 2) 16 diaPASEF runs on individual tissue lysate samples. In the original manuscript, these data were analyzed using Spectronaut version 18.5 using direct DIA approach, using DDA-based library, and using a hybrid (DDA plus direct DIA) library. In addition, the diaPASEF data was processed using DIA-NN 1.8.1 in the “library-free” (direct DIA) mode. We reanalyzed these data using FragPipe using just the 16 diaPASEF files processed with diaTracer (abbreviated “FragPipe” in **Figure 2**) and using the hybrid library build from these 16 diaPASEF and the additional 12 fractionated ddaPASEF runs (“FragPipe hybrid”). The Spectronaut 18.5 results (.sne files) and the DIA-NN 1.8.1 library-free results (report.tsv files) were download from the original study.

The number of quantified proteins using the 16 diaPASEF runs alone (direct DIA), was similar between all the tools. Note that the results from all tools were filtered to global protein, global

precursor, and run-specific precursor Q-value of 0.01. Under these settings, FragPipe with diaTracer, direct DIA analysis, quantified an average of 9,296 proteins per diaPASEF run (**Figure 2a**), with 10,341 proteins in total across all runs (**Figure 2b**). In comparison, Spectronaut directDIA analysis quantified fewer proteins, 8,997 on average per run (10,032 proteins in total). DIA-NN library-free analysis quantified slightly more proteins on average (9,520) and in total (10,628) in these data. The number of proteins quantified increased with the use of the ddaPASEF data. Using the hybrid DDA/DIA library, FragPipe again outperforming Spectronaut both in the total number of proteins quantified (11,029 vs.10,552) and per run (9,653 vs 9,021). Furthermore, FragPipe with the hybrid DDA/DIA analysis strategy quantified the most proteins with no missing values or with less than 50% missing values across the individual runs (**Figure 2b**).

Cerebrospinal fluid data

We then assessed the performance of diaTracer and FragPipe using the cerebrospinal fluid (CSF) dataset from Mun et al.[40]. This dataset contained 24 ddaPASEF and 34 diaPASEF runs. We performed quantification of 34 diaPASEF samples using the following three FragPipe workflows: 1) Using 34 diaPASEF runs, direct DIA analysis (abbreviated “FragPipe” in **Figure 3**). 2) Using the library built from 24 ddaPASEF fractionated runs (abbreviated “FragPipe DDALib”). 3) Using the hybrid library obtained by combining 24 ddaPASEF fractionated runs and 34 diaPASEF runs (“FragPipe HybridLib”). We also used the standalone DIA-NN 1.8.1 in the library-free mode (“DIA-NN lib-free”). Additionally, we used the diaTracer-extracted files to conduct semi-tryptic and comprehensive PTM searches (using the mass-offset and open search modes of MSFragger). We did not attempt to perform semi-tryptic searches using the DIA-NN library-free mode.

Using the FragPipe workflow (tryptic search), using the 34 diaPASEF files only, we quantified an average of 955 proteins per file, with 1,023 proteins in total (**Figure 3a**; **Supplementary Figure 1a**). In comparison, DIA-NN library-free analysis quantified fewer proteins per file, with 846 proteins on average, despite reporting a higher total number of proteins (**Supplementary Figure 1a**). The number of proteins quantified per file using the DDA-based library was 23% (221 proteins) higher than that obtained from direct analysis of the diaPASEF data. Using both ddaPASEF and diaPASEF data (i.e., building the hybrid library) yielded the best results, with an

18% increase (207 proteins) in the average number of quantified proteins per file compared with using the DDA-based library. This also resulted in the highest total number of proteins (**Supplementary Figure 1a)** and precursors (**Supplementary Figure 1b)** across the four workflows**.** Interestingly, the average number of quantified precursors per file in the direct DIA analysis with diaTracer was 14% (1,217 precursors) higher than that using the DDA-based library (**Figure 3b**).

The semi-tryptic search of these data did not result in a significant change in the average number of quantified proteins per file (**Figure 3a**) and a minor increase (e.g., 83 in the case of direct DIA analysis) in the total number of proteins (**Supplementary Figure 1a)**. However, this resulted in a noticeable increase in the number of quantified precursors (39% on average per file) compared to the tryptic search (e.g., 9,771 and 13,997 precursors per file, on average, in tryptic and semi-tryptic searches for the direct diaPASEF analysis workflow, respectively). A deeper investigation into the nature of the additional peptides identified in the semi-tryptic search revealed that many of them were derived from immunoglobulin proteins and secreted proteins (**Supplementary Table 1**). For 158 proteins, we found semi-tryptic peptides mapping to the region immediately following the signal peptides in the N-terminal portion of the corresponding protein (**Supplementary Table 2**). **Figure 3c** shows an example of protein hemopexin (HPX), in which the removal of the signal peptide in the mature form of the protein creates a peptide that does not conform to the trypsin cleavage rules. Multiple additional semi-tryptic peptides were identified for this and other proteins (**Figure 3c**). Overall, this confirms the presence of proteolytic cleavage events expected in CSF (and plasma, see below) samples and highlights the importance of monitoring such events using our diaTracer-enabled direct diaPASEF analysis workflow.

We also benchmarked the running time of diaTracer and the rest of the FragPipe workflow **(Figure 3d)**. diaTracer required 661 min to generate pseudo-MS/MS spectra (19.4 min per diaPASEF file, on average), and the rest of FragPipe (tryptic search) required 87 min to complete the analysis (including running DIA-NN for quantification). In contrast, the DIA-NN library-free analysis took 10.5 min to generate the in-silico predicted spectral library and 1,486 min to perform the rest of the analysis, more than twice the time of the diaTracer workflow. Importantly, repeating FragPipe analysis (starting with the existing pseudo-MS/MS mzML files) using the semi-tryptic search took only 110 min. Because diaTracer extracts all possible signals, the generated pseudo-MS/MS mzML file can be reused in any search setting. In contrast, the *in-*

silico spectral library generated by DIA-NN can only be reused if the search settings are the same.

We also tested the possibility of performing mass-offset searches using the FragPipe's "Mass-Offset-CommonPTMs" workflow and the same pseudo-MS/MS spectra from diaTracer. **Figure 3e** displays the most abundant PTMs identified in these data, as summarized by PTM-Shepherd[28]. The list of PTMs included common chemical artifacts, but also biological modifications such as phosphorylation. The analysis took 341 min (or 10 min per file), from MSFragger search to PTM-Shepherd reports. A similar mass shift histogram (**Supplementary Figure 2**) was observed from the open search results using the FragPipe's "Open" workflow, and the computational analysis took similar time (317 min). These mass-offset and open search analyses demonstrate the feasibility of performing comprehensive PTM searches using diaPASEF data processed using diaTracer.

Plasma proteomics

Another evaluation was conducted using the plasma dataset from Vitko et al[41]. In the original study, the authors performed a plasma profiling workflow using Seer's Proteograph Assay Kit, which includes five distinct nanoparticles (NP1-5). We selected 40 diaPASEF runs of NP2 enriched plasma samples analyzed with a 60SPD gradient on a timsTOF HT mass spectrometer. Pseudo-MS/MS spectra were generated from these 40 diaPASEF runs using diaTracer, and the rest of the analysis was completed in FragPipe using MSFragger tryptic and semi-tryptic search settings. The data were also analyzed using the DIA-NN library-free mode. With at least 50% non-missing value filtering (i.e., proteins must be quantified in at least 50% of the runs) and performing tryptic searches, the DIA-NN library-free analysis quantified 2,347 proteins, whereas diaTracer coupled with FragPipe quantified 2,922 proteins (**Figure 4a**). Similarly, although the DIA-NN library-free mode quantified most precursors across all runs in total, the diaTracer workflow in FragPipe resulted in 77% more precursors when requiring a precursor to be quantified in at least 50% of the runs. This demonstrates the improved ability of the diaTracer workflow in FragPipe to consistently quantify peptides across a large sample cohort.

Furthermore, as with the CSF dataset described above, performing a semi-tryptic search significantly increased the number of quantified precursors (but not proteins). Using the semi-

tryptic search setting, we identified 30,158 peptides, compared to 26,469 peptides identified using the tryptic search. Of these, 7,642 peptides were uniquely identified in the semi-tryptic search (**Supplementary Figure 3**). Among these 7,642 peptides, 6,800 (89%) were semi-tryptic. Additionally, 95% (6,458) of the semi-tryptic peptides, corresponding to 1,104 proteins, were quantified. Among the 40 samples in this plasma dataset, 20 were from patients with stage IV non-small cell lung cancer (NSCLC) and 20 were from patients in the non-cancer control group. We applied FragPipe-Analyst[42] to conduct differential expression analysis between the NSCLC and control groups using both tryptic and semi-tryptic search results. We detected 69 more significantly differentially expressed proteins from the semi-tryptic results under the same filtering criteria (**Supplementary Table 3**). In the semi-tryptic results, 35 proteins were upregulated, and 140 proteins were downregulated (**Figure 4c**). **Figure 4d** shows two examples of proteins showing statistically significant upregulation in the NSCLC vs. control group with the semi-tryptic search, for which a non-significant difference was observed with the tryptic search. These two proteins, AKT2 and H2AX, have been previously reported to be associated with NSCLC[43, 44]. Upon examining peptide-level quantification, we found the difference in H2AX quantification was due to two semi-tryptic peptides showing significant upregulation in the NSCLC group (**Supplementary Figure 4**). For AKT2, however, there were no semi-tryptic peptides quantified. One tryptic peptide identified in the tryptic search was missing from the semi-tryptic search as it did not pass the FDR filters. In the absence of this (low confidence) peptide, AKT2 expression was significantly upregulated in the NSCLC group. Overall, while in-depth interpretation of these examples goes beyond the scope of this work, our results highlight the importance of accounting for proteolytic cleavage events in the analysis of plasma data.

HLA immunopeptidomics

Applying peptide-centric strategies to the analysis of HLA or endogenous peptidome data is not practically feasible, given the time required to generate an *in-silico* predicted spectral library for all candidate database peptides under nonspecific protein cleavage rules. The diaTracer approach of extracting pseudo-MS/MS spectra offers a clear advantage for analyzing this type of data. We used an HLA dataset[45] to evaluate the performance of our method and compare it with the results reported in the original study. We selected three replicates of an HLA peptidome sample purified from a healthy donor's plasma sample and analyzed using a Whisper40 gradient (31 min length) on a timsTOF Ultra instrument. As shown in **Figure 5a,** from the three replicates**,** we identified and quantified 2,651 immunopeptides with length from 7-14, which was

18% more than the 2,288 peptides reported in the original publication using Spectronaut version 17 directDIA.

To further evaluate our results, we used NetMHCPan-4.1[46] to predict the binding affinity of the identified immunopeptides with the donor's HLA allele information provided in the original study. We identified 2,327 (90%) strong and 130 (5%) weak binders from 2,572 immunopeptides of lengths 8–12 (**Figure 5b**). Most of these binders were derived from the HLA-A*24:02 allele (**Supplementary Figure 5**), as reported in the original publication. We further summarized the length and charge state distributions of the quantified immunopeptides (**Figure 5c** and **5d**). Most immunopeptides had a length of nine and were observed as singly or doubly charged ions. This finding aligns well with the general expectations for HLA data and the specific findings reported for this dataset in the original study. We also compared how the identified immunopeptides overlapped with those reported in the original study using the experimental DDA library and using the panlibrary (**Supplementary Figure 6**). While it is expected that the analysis of diaPASEF data using a more comprehensive spectral library would result in more peptide identifications, our FragPipe results showed good overlap with the published data. We also selected one peptide to demonstrate the utility of our integrated FragPipe-PDV[36] viewer for inspecting PSMs and MS/MS spectra. The peptide VYQHLFTRI is predicted to be a strong binder. We visualized one of its PSMs along with the predicted spectrum generated by MSBooster[21] (**Figure 5e**). The pseudo-MS/MS spectrum extracted by diaTracer exhibited a high degree of similarity (spectrum entropy score of 0.9863) with the predicted spectrum, demonstrating the high quality of the pseudo-MS/MS spectra.

Phosphoproteomics

To evaluate the performance of our method in quantifying phosphorylated peptides, we applied our diaTracer-based direct diaPASEF workflow in FragPipe to a phosphorylation-enriched dataset acquired using six gradient lengths[47]. It took diaTracer, on average, 13.5 min to extract signals and produce pseudo-MS/MS spectra for a single 60-minute gradient file. For site localization, PTMProphet was enabled, and the site localization information was propagated from the PSM files to the final quantification matrices generated by DIA-NN. We identified 6,844 and 11,281 phosphopeptide sequences in the 7-minute and 60-minute gradient data, respectively (**Figure 6a**). In contrast to the original study, where the number of phosphorylated peptide sequences plateaued at a 21-minute gradient (**Figure 6a**), we observed an increase in

the number of quantified phosphorylated peptides with longer gradients. This trend also holds for class I phosphosites (sites with site localization probability > 0.75) (**Supplementary Figure 7**). We also found that the quantification results reported by diaTracer-based direct diaPASEF workflow in FragPipe showed high correlations across four technical replicates (**Figure 6b**). We also compared the quantification performance with Spectronaut. **Supplementary Figure 8** shows the CV distribution, with box plots illustrating the quartiles and median of the CVs. The result reveal that our tool has comparable quantification precision with Spectronaut in these data.

Although our pipeline computes site localization scores at the spectral library building stage, manual inspection of peptide elution profiles and site localizing fragments observed in a specific sample may be desired. In addition to the FragPipe-PDV viewer mentioned above, the output from FragPipe is fully compatible with the Skyline computational environment for advanced visualization of peptide precursors and fragment ion signals in diaPASEF data. We visualized the PSMs and XICs of two co-eluted isobaric positional isomers, S(pho)PSPPDGSPAATPEIR and SPSPPDGS(pho)PAATPEIR, quantified in a 60-minute gradient run. diaTracer successfully separated these two precursor signals along with their corresponding fragment signals into two separate pseudo-MS/MS scans. Both isomers were identified with a site probability > 0.75. They eluted approximately 35 s apart in retention time and were 0.02 1/K0 apart in the ion mobility dimension (**Figures 6c-f**). Several fragment peaks that contributed to the localization of the phosphorylation site were clearly visible in the spectra.

Low-input, spatial proteomics data

We then investigated the performance of our method using low-input data from a recent spatial proteomics study[48]. The study profiled 148 microregions from four areas (Epithelium, Germinal center, Mantel zone, and T-cell zone) of a human tonsil obtained using laser capture microdissection (LCM). In the original study, the authors first used the DIA-NN library-free mode to generate a comprehensive human tonsil spectral library using 46 runs acquired in the diaPASEF mode from samples with a high protein amount (referred to as high-input samples below). The authors then applied this high-input spectral library to quantify proteins in the 148 low-input samples. The authors did not attempt to perform a direct analysis of the low-input samples, that is, without using the high-input spectral library. For comparison with the original study, we performed two analyses using the diaTracer workflow in FragPipe. In the first analysis,

we used diaTracer and FragPipe to build the library from the 46 high-input samples ("FragPipe high-input Lib"), similar to the original study that used DIA-NN. Second, we performed a direct analysis of the 148 low-input samples without using high-input data ("FragPipe").

The results of our diaTracer-FragPipe workflow, with high-input data used for building the library, produced identification numbers very similar to those based on DIA-NN library-free analysis, using the reports taken from the original publication (**Figure 7a**; **Methods**). As expected, the high-input library resulted in more proteins, 3,349 in total (1,992 proteins per sample), compared to 1,602 proteins (1,303 proteins per sample) identified using direct analysis of low-input samples. However, the total number of identified proteins is a misleading metric; proteins detected in one or a few samples have little value for downstream analysis. Indeed, all analyses reported by the authors in the original paper were performed after the application of a missing value filter. More specifically, they required at least 70% non-missing values across all samples in at least one group (the four sample groups mentioned above). With this filter in place, the use of high-input data resulted in 2,175 proteins (1,692 per sample) compared with 1,475 proteins (1,265 per sample) quantified directly from low-input samples. Increasing the stringency of the missing value filter further diminishes the impact of using high-input data. Requiring 90% non-missing values, we identified 1,652 (1,398 per sample) and 1,336 proteins (1,182 per sample) with and without using high-input samples alone, respectively.

Figure 7b displays the overlap between proteins quantified with at least 70% non-missing values in at least one class in the original study and using FragPipe with diaTracer with and without high-input sample data. All these approaches exhibited a high degree of overlap. Overall, the biological observations were consistent across all analyses. We specifically investigated whether similar results could be obtained with direct identification from low-input samples, despite quantifying fewer proteins compared to using a larger high-input library. Following the same data filtering procedure as in the original study, the quantified proteins clearly separated into four distinct cell types (**Figure 7c**). We also examined cell type markers reported in the original study, including CD19 (B-cell marker), CD3D (T-cell marker), and CDH1 (epithelial marker), which were all highly expressed in the anticipated sample groups (**Figure 7d**). Furthermore, a global comparison between Mantel zone and T cells revealed significant proteomic differences (**Figure 7e**). The list of differentially expressed proteins derived from the results with and without using high-input data, when submitted to FragPipe-Analyst, produced a similar list of enriched pathways (**Supplementary Figure 9**). Overall, our analysis demonstrates

that FragPipe with diaTracer can be effectively used to identify and quantify proteins from a large cohort of low-input diaPASEF samples, even in the absence of high-input data (although having such data is beneficial), especially if focusing on a subset of proteins consistently identified across a large fraction of the samples in the cohort.

Discussion

We presented a new computational method and software, diaTracer, which can deconvolute diaPASEF data to produce DDA-like pseudo-MS/MS spectra. These pseudo-MS/MS spectra can be searched using conventional DDA peptide identification tools, such as MSFragger. diaTracer is available as a stand-alone tool and is integrated into FragPipe, enabling a complete diaPASEF data analysis workflow, including peptide identification, deep-learning-based rescoring, protein inference, FDR filtering, PTM site localization, spectral library generation, quantification, and data visualization. We used several datasets from different types of biological studies to demonstrate the speed and high sensitivity of our new computational strategy for analyzing diaPASEF data.

diaTracer enables direct, spectrum-centric analysis of diaPASEF data, which is more versatile than the peptide-centric approach. Advances in deep learning methods have enabled the creation of tools that can predict fragment intensities, ion mobilities, and retention times of peptides with high accuracy, thereby facilitating the generation of *in-silico* spectral libraries. However, current deep learning prediction tools have limitations. Peptide-centric strategies coupled with *in-silico* library prediction work well for conventional experiments involving quantitative protein profiling of cell lines and tissues. In contrast, their application to endogenous peptidome or immunopeptide data is generally not feasible because of the large number of peptide candidates that require predictions. Even semi-tryptic searches, which are often beneficial for certain samples, such as plasma samples and samples from N-terminomics[49, 50] studies, remain challenging. Furthermore, peptide-centric analyses struggle with searches involving more than a few common PTMs. Even for phosphorylation, a PTM that receives a great deal of attention, the peptide-centric approach can be computationally prohibitive. Furthermore, for peptides harboring less common PTMs or chemical labels that are not included in the training set used to build prediction models there is an additional risk of poor prediction accuracy.

In contrast, the spectrum-centric approach of DIA-Umpire, and now diaTracer for diaPASEF data, does not have these limitations. After diaTracer generates pseudo-MS/MS spectra, well-established DDA peptide identification tools can be used to perform semi-enzymatic or nonspecific searches. We demonstrated that the semi-enzymatic search provides a significant boost in the number of quantified precursors. It also increases the accuracy and precision of quantification, thereby identifying more differentially expressed proteins in the analysis of CSF and plasma datasets. We also illustrated the high sensitivity and accuracy of our strategy in the analysis of HLA immunopeptidome data. We note that spectrum-centric strategies still benefit from deep-learning predictions and rescoring; however, predictions need to be made only after the search of pseudo-MS/MS spectra and for a relatively small number of peptides. The accuracy of the predictions is also less critical (e.g., a very inaccurate retention time prediction would not automatically exclude the peptide from being identified), making the pipeline more robust for peptides harboring less common PTMs or chemical labels. Thus, we hope that our strategy will be especially useful for chemoproteomics applications involving different types of chemical labels, such as affinity-based proteome profiling (ABPP) workflows[51, 52]. Finally, the spectrum-centric approach enables unrestricted PTM analysis via open or mass-offset searches of pseudo-MS/MS spectra, which is currently not possible using the peptide-centric approach.

The spectrum-centric approach of diaTracer (and DIA-Umpire) also has limitations. First, diaTracer starts with MS1 feature detection, which means it relies on the detection of precursor peptide's signal in MS1 data to group fragments and build a pseudo-MS/MS spectrum. To minimize this limitation, diaTracer aggregates the neighboring RT frames to amplify the MS1 signals. This helps to extract weak signals and filter out noise. Second, diaTracer extracts multiple features for each signal, including ion mobility range, m/z, precursor XIC, XICs of the corresponding fragments and their correlation with the precursor and with each other. However, this information is currently used only at the pseudo-MS/MS generation step. Additionally using it in the subsequent database search and scoring, as it is done in e.g. MSFragger-DIA[19], would likely improve the results. We plan to address these limitations in future work.

The computational time required for processing diaPASEF data is an important consideration. In our tests of diaTracer and FragPipe, the analysis time was less than the time it took to acquire the mass spectrometry data. Furthermore, the conversion of diaPASEF raw files to pseudo-MS/MS spectra can start with the already acquired files while the data acquisition is still ongoing. Finally, the conversion needs to be performed only once. The resulting deconvoluted

files can be used in multiple ways, for example, searched using different search parameters, or using different search engines or their combination, all without the need to reprocess the files.

In this study, we focused on the conventional diaPASEF data. At the same time, the rapid technological development of the timsTOF platform led to the introduction of several related data acquisition methods, including Synchro-PASEF[53], Slice-PASEF[54], and midiaPASEF[55]. These techniques aim to increase the efficiency of signal detection by redesigning the schema of the quadrupole selection windows. These new methods better preserve the relationships between the precursor peptide and its fragments, facilitating downstream data analysis. The spectrum-centric approach of diaTracer is well-suited for analyzing data from these new acquisition methods. We plan to extend diaTracer to support these new data types in the future.

Methods

diaTracer feature extraction algorithm

The data from one isolation window ($1/k0 \times m/z$) in each retention time (RT) frame is stored as a sparse matrix sized at $N \times W$, where W is the number of ion mobility (IM) bins determined by the IM window width and resolution. It equals the isolation window frame scan range divided by the smallest IM difference between two signals. N is the number of m/z bins, which is determined by the isolation window width and mass resolution. It equals the isolation window m/z range divided by the smallest m/z difference between two signals. To enhance the signal and reduce noise interference[25], data in the same isolation window of adjacent neighbor RT frames are aggregated to form a composite 3D matrix by summing the intensity of signals and subsequently removing the data points lacking proximate neighbors within a certain range. A 2D Gaussian filter is then employed to smooth the $N \times W$ matrix data. Local maxima are identified within the matrix and used as seeds for Gaussian fitting to determine 2D peaks. The center of the 2D Gaussian peak determines the m/z and IM values of the corresponding feature.

In diaTracer, several attributes are defined for each extracted signal, including the m/z and IM centers, as well as the m/z and IM signal ranges. Subsequently, these attributes are partially initialized based on the 2D Gaussian distribution. The sum of the intensities within the signal range represents the peak intensity of the current RT frame. While traversing the frames along the retention time, the algorithm determines the apex for every signal. Upon locating the apex, its signal attributes are finalized, and the intensities from the other RT frames are adjusted

relative to it. Signal boundaries are determined based on predefined criteria to ensure a bell-shaped extracted ion chromatogram (XIC). However, if a defined endpoint of the peak is missed, an extended signal is recorded. diaTracer then applies Savitzky-Golay smoothing and the Z-Score[56] peak detection algorithm to segment an elongated peak into multiple separated peaks.

To separate authentic signals from background noise and assign charge states to precursors, diaTracer employs an isotope peak-grouping technique for the MS1 data. The traced MS1 peaks are organized into isotopic clusters based on RT and IM tolerances. The minimum number of the isotopes is 1. The isotope mass difference is based on the number of C13 atoms. Peaks that do not belong to any cluster are discarded. We consider the correlation of the theoretical and experimental isotope intensity distributions, and the correlation of the isotopes' XICs as metrics to control the quality of isotopic clusters. To minimize redundancy, highly confident isotopic peaks passing the threshold are removed from subsequent isotopic clustering. The charge state is determined by comparing against the theoretical isotopic distribution. When the charge state cannot be confidently determined, multiple values are retained. The extraction of MS2 peaks follows a similar process but with more relaxed constraints, including reduced neighbor requirements and no isotopic clustering.

After detecting precursor and fragment XIC features, they are grouped into clusters. Additionally, an optional fractional mass filter is applied to remove peptide features with a fractional mass outside of the expected mass range for tryptic peptides, as previously described[13] ("Mass Defect Filter" parameter). However, this filter is not recommended for certain PTM searches, such as phosphopeptide searches, or nonspecific searches. Clustering begins with the precursor peaks, considering all fragment peaks that met the IM and RT tolerances within the same isolation window ("Delta Apex IM" default 0.02 and "Delta Apex RT" parameters default: 3). Subsequently, the Pearson correlation coefficient is calculated between the smoothed XIC of the monoisotopic precursor and raw XICs of the fragments. Only fragment peaks that surpass the correlation threshold ("Corr threshold" parameter, default 0.3) are included in the construction of the pseudo-MS/MS spectrum. To control the number of fragment peaks in the pseudo-MS/MS spectrum, only the top N highest intensity peaks ("RF max" parameter, default: 500) are included. These default parameters were optimized based on datasets from various LC-MS/MS settings and perform well in most cases. The intensity values in the pseudo-MS/MS spectrum are derived from the peak intensities of the fragment XICs. All produced pseudo-MS/MS spectra

are used to generate an mzML file, which can be processed by MSFragger in FragPipe or any other search engine.

Deep proteome profiling triple-negative breast cancer data analysis

The triple-negative breast cancer (TNBC) dataset from Lapcik et al[39] contains 12 ddaPASEF runs from a fractionated pooled sample used to generate a DDA-based library and 16 diaPASEF runs from 16 individual TNBC peptide samples. The data was acquired using timsTOF Pro mass spectrometer. We ran FragPipe 22 (pre-release version) with the "DIA_SpecLib_Quant_diaPASEF" workflow to generate a spectral library and perform quantification for the raw d files. Within diaTracer, "Delta Apex IM" was set to 0.01, "Delta Apex RT" was set to 3, "RF max" was set to 500, and "Corr threshold" was set to 0.3. The mass defect filter was enabled. In the MSFragger database search, the initial precursor and fragment mass tolerances were set to 10 ppm and 20 ppm, respectively. Spectrum deisotoping[57], mass calibration, and parameter optimization[27] were enabled. The isotope error was set to “0/1/2”. The reviewed *Homo sapiens* protein sequence database obtained from UniProt (downloaded on November 15, 2023; 20,461 proteins), appended with common contaminants and decoys, was used in the search. Enzyme specificity was set to “stricttrypsin” and the maximum allowed missed cleavages were set to 2. Oxidation of methionine and N-terminal acetylation were set as variable modifications. The maximum number of variable modifications for each peptide was set to 3. The following change was made compared to the default settings of the DIA_SpecLib_Quant_diaPASEF workflow: pyro-Glu at peptide N-terminus was added as a variable modification and methylthiolation of cysteine was used as a fixed modification. MSBooster and Percolator were used to predict the RT and MS/MS spectra, and to rescore PSMs. The final FDR-filtered PSMs and the pseudo-MS/MS mzML spectral files (and DDA mzML files in the hybrid workflow) were used by EasyPQP to generate the spectral library. The spectral library was then used with the DIA-NN quantification module to quantify the library peptide ions in the individual diaPASEF data. The same settings were employed for all experiments, unless mentioned otherwise. Spectronaut 18.5 directDIA result and DIA-NN 1.8.1 library-free result were downloaded from the original study. The Spectronaut report file was exported using Spectronaut Viewer 18.5. DIA-NN generated report.tsv files were processed using the *iq* R package[58] to count the number of peptides and proteins identified passing the specified confidence thresholds (see **Result processing and statistical analysis).**

Cerebrospinal fluid data analysis

The cerebrospinal fluid (CSF) dataset from Mun et al.[40] contains 24 ddaPASEF runs used to generate a DDA-based library and 34 diaPASEF runs from 15 patients with Alzheimer's disease (AD) and 19 control subjects. FragPipe with diaTracer was run using the built-in "DIA_SpecLib_Quant_diaPASEF" workflow. Compared to the parameters for the TNBC dataset, carbamidomethylation of cystine set as a fixed modification. For semi-tryptic searches, the cleavage parameter was adjusted to "SEMI". The peptide length was set from 7 to 50 with a peptide mass range of 500 to 5,000 Da. For comparison, DIA-NN (version 1.8.1) as a standalone tool was run in the library-free mode using default settings. The precursor m/z range was set 300 to 1800. Methionine oxidation was set as a variable modification. Carbamidomethylation of cysteine was used as a fixed modification. A maximum of 1 missed cleavage was allowed. An *in-silico* predicted spectral library was generated from the same database file as described above except without adding decoys. Quantification results from FragPipe and DIA-NN 1.8.1 were processed using the *iq* R package[58] to count the number of peptides and proteins identified passing the specified confidence thresholds.

To conduct comprehensive PTM searches using the mass-offset and open-search mode of MSFragger, FragPipe's "Mass-Offset-CommonPTMs" and "Open" workflows were used under default settings. The details regarding the mass-offset and open search workflows in FragPipe can be found elsewhere[27-29, 59]. In brief, in the mass-offset search, MSFragger takes a list of modification masses as mass offsets and searches the spectrum against peptides within a narrow mass tolerance window placed around each specified mass offset. This approach can be used to sensitively and rapidly detect hundreds of known modifications. In the open search mode, MSFragger searches the spectrum against the peptides with a wide mass tolerance window, by default from -150 to 500 Da. It is an efficient approach for detecting peptides with any mass shift, including unknown modifications.

Plasma data analysis

The subset of the plasma dataset from Vitko et al.[41] used in this study contained 40 files (20 patients with stage IV non-small cell lung cancer, NSCLC, and 20 control samples) acquired using the Seer Proteograph nanoparticle conjugation kit NP2 and a timsTOF HT mass spectrometer. Data was processed as described for the CSF dataset. Differential expression analysis was performed using FragPipe-Analyst[42]. When loading data in FragPipe-Analyst, the "data type" parameter was set to DIA, and both "report.pg_matrix.tsv" and "experiment_annotation.tsv" files were uploaded to the FragPipe-Analyst website. The "min

percentage of non-missing values globally" and "Min percentage of non-missing values in at least one condition" parameters were set to 25%, with all other parameters at default values. "DE Adjusted p-value cutoff" was set to 0.05. "DE Log2 fold change cutoff" was set to 1. The imputation type was set to "Perseus-type". The FDR correction type was set to "Benjamini Hochberg". The DE_result.tsv file was downloaded and sued for downstream analysis.

HLA data analysis

We used a subset of the HLA dataset from Wahle et al.[45], using a sample from one donor profiled in triplicate. FragPipe with diaTracer was run using the built-in "Nonspecific-HLA-diaPASEF" workflow. Compared to the "DIA_SpecLib_Quant_diaPASEF" workflow described above, the enzyme cleavage was set to "nonspecific" and the peptide length was restricted to 7 to 25. Carbamidomethylation of cysteine was not specified. Cysteinylation (+119) was added as a variable modification. The isotope error was set to 0/1. NetMHCpan 4.1[46] was used to predict the binding affinities of the identified peptides (considering peptides of length 8-12 only). Six MHC class I alleles of the donor were provided as additional input information. The immunopeptides were ranked according to their MHC-binding affinities expressed as a percentile rank. Immunopeptides with a percentage rank of less than 0.5% were classified as strong binders, whereas those with a percentage rank of less than 2% were considered weak binders. PDV[36], integrated into FragPipe, was used to visualize the PSMs. Both the predicted spectrum and spectral entropy scores were generated using MSBooster[21].

Phosphoproteome data analysis

The phosphopeptide enriched dataset from Oliinyk et al.[47] contained runs from six gradient lengths ranging from 7 to 60 min. Each gradient condition was replicated four times. To analyze the dataset using FragPipe with diaTracer, we employed the built-in "DIA_SpecLib_Quant_Phospho_diaPASEF" workflow. Compared to the "DIA_SpecLib_Quant_diaPASEF" workflow described above, Phosphorylation on STY was added as a variable modification in MSFragger, and site localization with PTMProphet was enabled. FragPipe propagated the site localization information from the PSM.tsv files to the quantification files produced by DIA-NN (report.tsv and report_pr_matrix.tsv files). Phosphopeptide counts, representing the number of non-redundant sequences of all identified phosphopeptides, were obtained from FDR-filtered results. When counting the number of phosphorylation sites, sites with the localization probability greater than 0.75 were considered

as localized. The quantification intensity correlation shown was generated by R package protti using phosphorylated peptides with the localization probability greater than 0.75 and measured in all 4 technical replicates of the 7 min gradient time experiment. The PSMs for selected peptides were visualized using PDV, whereas XICs were visualized using Skyline[60].

Low-input, special proteomics data

We used the dataset from Anuar et al.[48], 148 microregions from a patient' tonsil sections. In addition, there were 42 diaPASEF runs (from seven different sizes of human tonsil tissue, six replicates each) that were used to build a comprehensive tonsil library. FragPipe with diaTracer results were compared to the DIA-NN results from the original study. Differential expression analysis was performed using FragPipe-Analyst based on "report.pg_matrix.tsv" and "experiment_annotation.tsv" files generated by FragPipe.

Result processing and statistical analysis

To ensure a fair comparison of peptide and protein identification numbers, DIA-NN report.tsv files from the diaTracer workflows in FragPipe and from DIA-NN stand-alone analyses were filtered using the *iq* R package[58] to achieve a 1% FDR at run-specific precursor, global precursor, and global protein levels. Differential expression analyses conducted using FragPipe-Analyst were based on DIA-NN's report_pg_matrix.tsv files. Subsequent result processing and plot generation were performed within the RStudio build 402 environment using the R version 4.3.3 statistical software. The R packages ggplot2, tidyverse, ggrepel, plotly, eulerr, and protti were used in our analysis.

Runtime comparison

The runtimes for FragPipe with diaTracer and for the DIA-NN library-free mode analyses were measured on a Linux desktop with an Intel Core i9-13900K CPU (32 logical cores) and 128GB memory. All analyses were configured to utilize all 32 CPUs.

Acknowledgements

This work was supported in part by National Institutes of Health grants R01-GM-094231 and U24-CA271037. We thank Maria Wahle for providing additional results files for the HLA dataset.

Abbreviations

TNBC: triple-negative breast cancer; CSF: cerebrospinal fluid; DIA: data-independent acquisition; DDA: data-dependent acquisition; FDR: False discovery rate; GUI: graphical user interface; HLA: Human leukocyte antigens; IM: ion mobility; LC-MS: Liquid chromatography-mass spectrometry; MS/MS: tandem mass spectrum; NP: nanoparticles; NSCLC: non-small cell lung cancer; PASEF: parallel accumulation serial fragmentation; PTM: post-translational modifications; PSM: Peptide-spectrum matching; RT: retention time; XIC: extracted ion chromatogram.

Author Contributions

K.L. and A.I.N. developed the diaTracer algorithm. K.L. wrote the software and analyzed the results. F.Y. and G.T. assisted with the algorithm and software development. K.L.Y helped modify MSBooster. F.Y. assisted with the integration of diaTracer into FragPipe. K.L., F.Y., and A.I.N. wrote the manuscript. A.I.N. conceived the study. A.I.N. and F.Y. supervised the study.

Data availability

The raw MS/MS files used in this study can be found at the ProteomeXchange Consortium and PRIDE partner repository[61] or at the MassIVE repository with the following accession codes:

- TNBC data: PXD047793
- CSF data: PXD035249
- Plasma data: PXD047839
- HLA data: MSV000092557
- Phosphoproteomics data: PXD033904
- Low-input, spatial proteomics data: PXD042367

diaTracer converted mzML files and FragPipe results have been uploaded to MassIVE repository with the identifier MSV000094803 (reviewer username: MSV000094803_reviewer, password: diatracer_fp).

Code availability

The standalone version of diaTracer can be downloaded as a single JAR file at https://diatracer.nesvilab.org. FragPipe is available at GitHub https://github.com/Nesvilab/FragPipe. Python and R scripts for summarizing the results and generating the figures are available at https://github.com/nesvilab/diaTracer-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. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355 (2016).
2. 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 (2012).
3. Panchaud, A. et al. Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean. *Analytical Chemistry* **81**, 6481-6488 (2009).
4. 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. *Nature Methods* **1**, 39-45 (2004).
5. May, J.C. & McLean, J.A. Ion mobility-mass spectrometry: time-dispersive instrumentation. *Anal Chem* **87**, 1422-1436 (2015).
6. Meier, F. et al. Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* **17**, 2534-2545 (2018).
7. Yu, F. et al. Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Molecular & cellular proteomics : MCP* **19**, 1575-1585 (2020).
8. 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* **20**, 100077 (2021).
9. Demichev, V. et al. dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat Commun* **13**, 3944 (2022).
10. Tsou, C.C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **12**, 258-264, 257 p following 264 (2015).
11. Rost, H.L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature biotechnology* **32**, 219-223 (2014).
12. Wang, J. et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nature Methods* **12**, 1106-1108 (2015).
13. Tsou, C.C., Tsai, C.F., Teo, G.C., Chen, Y.J. & Nesvizhskii, A.I. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics* **16**, 2257-2271 (2016).
14. Bruderer, R. et al. New targeted approaches for the quantification of data-independent acquisition mass spectrometry. *Proteomics* **17** (2017).
15. Ting, Y.S. et al. Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics* **14**, 2301-2307 (2015).
16. Searle, B.C. et al. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications* **9**, 5128 (2018).
17. Ting, Y.S. et al. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature Methods* **14**, 903-908 (2017).
18. 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* **17**, 41-44 (2020).
19. Yu, F. et al. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat Commun* **14**, 4154 (2023).
20. Lu, Y.Y., Bilmes, J., Rodriguez-Mias, R.A., Villén, J. & Noble, W.S. DIAmeter: matching peptides to data-independent acquisition mass spectrometry data. *Bioinformatics* **37**, i434-i442 (2021).

21. Yang, K.L. et al. MSBooster: improving peptide identification rates using deep learning-based features. *Nature Communications* **14** (2023).
22. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **16**, 509-518 (2019).
23. 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* **14**, 513-520 (2017).
24. Tsou, C.C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* **12**, 258-+ (2015).
25. Avtonomov, D.M., Polasky, D.A., Ruotolo, B.T. & Nesvizhskii, A.I. IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry. *Analytical Chemistry* **90**, 2369-2375 (2018).
26. 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. *Nature Methods* **14**, 513-+ (2017).
27. Yu, F.C. et al. Identification of modified peptides using localization-aware open search. *Nature Communications* **11** (2020).
28. Geiszler, D.J. et al. PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications From Open Search Results. *Molecular & Cellular Proteomics* **20**, 100018 (2021).
29. Polasky, D.A. et al. MSFragger-Labile: A Flexible Method to Improve Labile PTM Analysis in Proteomics. *Molecular & cellular proteomics : MCP* **22**, 100538 (2023).
30. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S. & Maccoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923-925 (2007).
31. 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* **74**, 5383-5392 (2002).
32. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658 (2003).
33. da Veiga Leprevost, F. et al. Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nature Methods 2020 17:9* **17**, 869-870 (2020).
34. Shteynberg, D.D. et al. PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. *J Proteome Res* **18**, 4262-4272 (2019).
35. Pino, L.K. et al. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom Rev* **39**, 229-244 (2020).
36. Li, K., Vaudel, M., Zhang, B., Ren, Y. & Wen, B. PDV: an integrative proteomics data viewer. *Bioinformatics* **35**, 1249-1251 (2019).
37. Hsiao, Y. et al. Analysis and visualization of quantitative proteomics data using FragPipe-Analyst. *bioRxiv*, 2024.2003.2005.583643 (2024).
38. Kohler, D. et al. MSstats Version 4.0: Statistical Analyses of Quantitative Mass Spectrometry-Based Proteomic Experiments with Chromatography-Based Quantification at Scale. *J Proteome Res* **22**, 1466-1482 (2023).
39. Lapcik, P. et al. A hybrid DDA/DIA-PASEF based assay library for a deep proteotyping of triple-negative breast cancer. *Sci Data* **11**, 794 (2024).
40. Mun, D.G. et al. Four-dimensional proteomics analysis of human cerebrospinal fluid with trapped ion mobility spectrometry using PASEF. *Proteomics* **23** (2023).

41. Vitko, D. et al. timsTOF HT Improves Protein Identification and Quantitative Reproducibility for Deep Unbiased Plasma Protein Biomarker Discovery. *Journal of Proteome Research* **23**, 929-938 (2024).
42. Hsiao, Y. et al. Analysis and visualization of quantitative proteomics data using FragPipe-Analyst. *bioRxiv* (2024).
43. Liu, T. et al. AKT2 drives cancer progression and is negatively modulated by miR-124 in human lung adenocarcinoma. *Respiratory Research* **21** (2020).
44. Matthaios, D., Hountis, P., Karakitsos, P., Bouros, D. & Kakolyris, S. H2AX a promising biomarker for lung cancer: a review. *Cancer Invest* **31**, 582-599 (2013).
45. Wahle, M. et al. IMBAS-MS Discovers Organ-Specific HLA Peptide Patterns in Plasma. *Mol Cell Proteomics* **23**, 100689 (2024).
46. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **48**, W449-W454 (2020).
47. Oliinyk, D. & Meier, F. Ion mobility-resolved phosphoproteomics with dia-PASEF and short gradients. *Proteomics* **23**, e2200032 (2023).
48. Makhmut, A. et al. A framework for ultra-low-input spatial tissue proteomics. *Cell Systems* **14**, 1002-+ (2023).
49. Chang, C.-H., Chang, H.-Y., Rappsilber, J. & Ishihama, Y. Isolation of Acetylated and Unmodified Protein N-Terminal Peptides by Strong Cation Exchange Chromatographic Separation of TrypN-Digested Peptides. *Molecular & Cellular Proteomics* **20**, 100003 (2021).
50. Klein, T., Eckhard, U., Dufour, A., Solis, N. & Overall, C.M. Proteolytic Cleavage—Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification. *Chemical Reviews* **118**, 1137-1168 (2018).
51. Kuljanin, M. et al. Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nature biotechnology* **39**, 630-641 (2021).
52. Desai, H.S. et al. SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State–Dependent Redox-Sensitive Cysteines. *Molecular & Cellular Proteomics* **21** (2022).
53. Skowronek, P. et al. Synchro-PASEF Allows Precursor-Specific Fragment Ion Extraction and Interference Removal in Data-Independent Acquisition. *Mol Cell Proteomics* **22**, 100489 (2023).
54. Lukasz Szyrwiel, L.S., Markus Ralser, Vadim Demichev Slice-PASEF: fragmenting all ions for maximum sensitivity in proteomics. *bioRxiv* (2022).
55. Ute Distler, M.K.Ł., Michał Piotr Startek, David Teschner, Sven Brehmer, Jens Decker, Thilo Schild, Jonathan Krieger, Florian Krohs, Oliver Raether, Andreas Hildebrandt, Stefan Tenzer midiaPASEF maximizes information content in data-independent acquisition proteomics. *bioRxiv* (2023).
56. Brakel, J.P.G.v. (2014).
57. Teo, G.C., Polasky, D.A., Yu, F.C. & Nesvizhskii, A.I. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *Journal of Proteome Research* **20**, 498-505 (2021).
58. Pham, T.V., Henneman, A.A. & Jimenez, C.R.: an R package to estimate relative protein abundances from ion quantification in DIA-MS-based proteomics. *Bioinformatics* **36**, 2611-2613 (2020).
59. Ramírez, C.R., Espino, J.A., Jones, L.M., Polasky, D.A. & Nesvizhskii, A.I. Efficient Analysis of Proteome-wide FPOP Data by FragPipe. *bioRxiv* (2023).
60. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968 (2010).
61. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* **47**, D442-D442 (2019).

Figure 1: Overview of diaTracer and FragPipe computational platform. diaTracer applies a 3D feature detection algorithm to detect signals from all possible precursors and fragments in MS1 and MS2 diaPASEF data. Pseudo-MS/MS spectra are generated through precursor-fragment clustering and can be processed as DDA data using MSFragger and FragPipe to build a spectral library directly from the data. A hybrid spectral library can also be generated if DDA data are available. This spectral library is then used to extract quantification using DIA-NN.

Figure 2: Deep proteome profiling. a) Box plot showing the numbers of quantified proteins using different methods in the TNBC dataset. The lower and upper edges of the box represent the first (Q1) and the third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. **b)** Histogram showing the number of quantified proteins in the TNBC dataset using Spectronaut 18.5 and FragPipe with diaTracer, direct DIA and hybrid DDA/DIA analysis, and using DIA-NN in library-free mode, after application of different non-missing value filters.

Figure 3: CSF data. a) Box plot showing the numbers of quantified proteins using different methods, with colors representing cleavage types (green: trypsin cleavage; orange: allowing semi-tryptic peptides). In the boxplot, the central line represents the median of the numbers. The lower and upper edges of the box represent the first (Q1) and third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. **b)** Box plot showing the numbers of quantified precursors using different methods. The box plots' median, edges, and whiskers are same as the ones in a). **c)** Distribution of identified peptides for protein HPX, with the red segment indicating the signal peptide region at the N-terminal. Orange segments represent tryptic peptides, while green segments represent semi-tryptic peptides. **d)** Running time comparison. FragPipe* indicated the FragPipe run time (including DIA-NN for quantification) starting from diaTracer-extracted files. **e)** Modifications identified using the common mass-offset workflow in FragPipe using pseudo-MS/MS spectra generated by diaTracer.

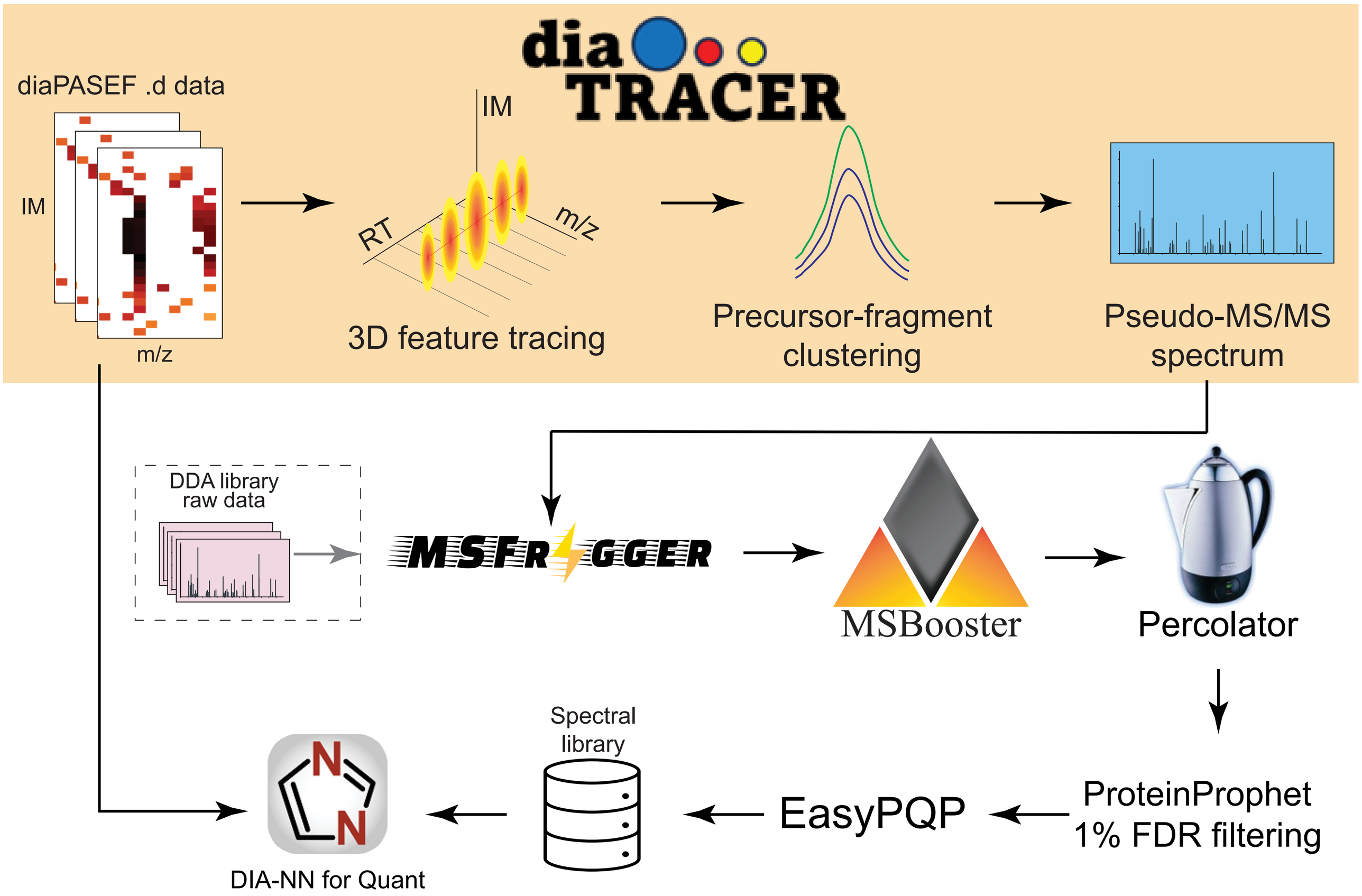
Figure 4: Plasma data. a) Histogram showing the number of quantified proteins using diaTracer-based FragPipe workflows (tryptic and semi-tryptic search), using DIA-NN library-free mode, colored from deep to light blue corresponding to different non-missing value filters. **b)** Same as a) for quantified precursors. **c)** Volcano plot comparing protein abundance between stage IV non-small cell lung cancer (NSCLC) and non-cancer control samples, highlighting NSCLC-overexpressed proteins (Log2 fold change >= 0.6; adjusted p-value <= 0.05). **d)** Boxplots of protein abundance distribution of AKT2 (top) and H2AX (bottom)proteins in semi-

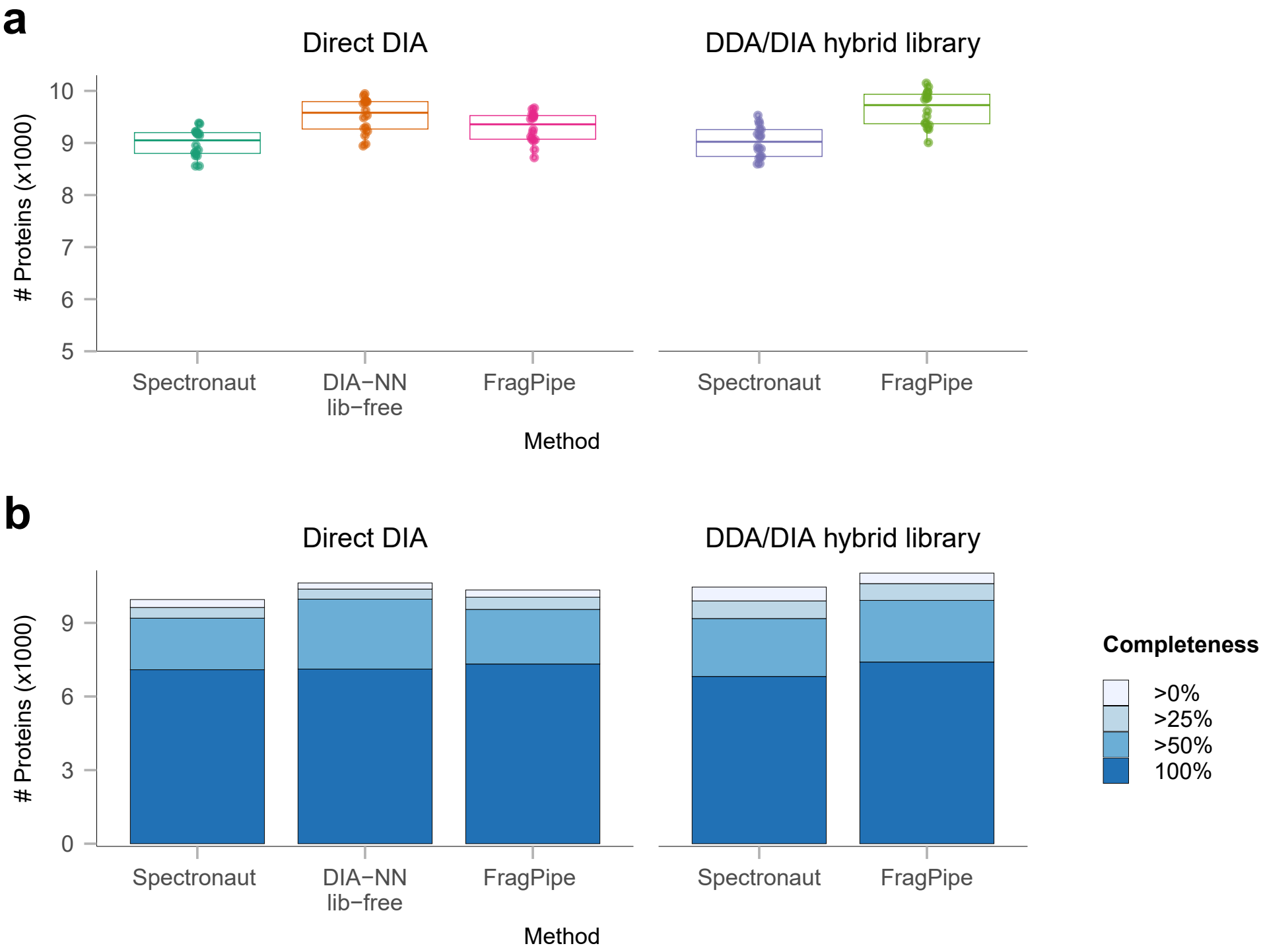
tryptic (left) and tryptic (right) searches between NSCLC (red) and control (blue) samples. In the boxplot, the central line represents the median of the numbers. The lower and upper edges of the box represent the first (Q1) and third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. Tr: tryptic search; Semi: semi-tryptic search.

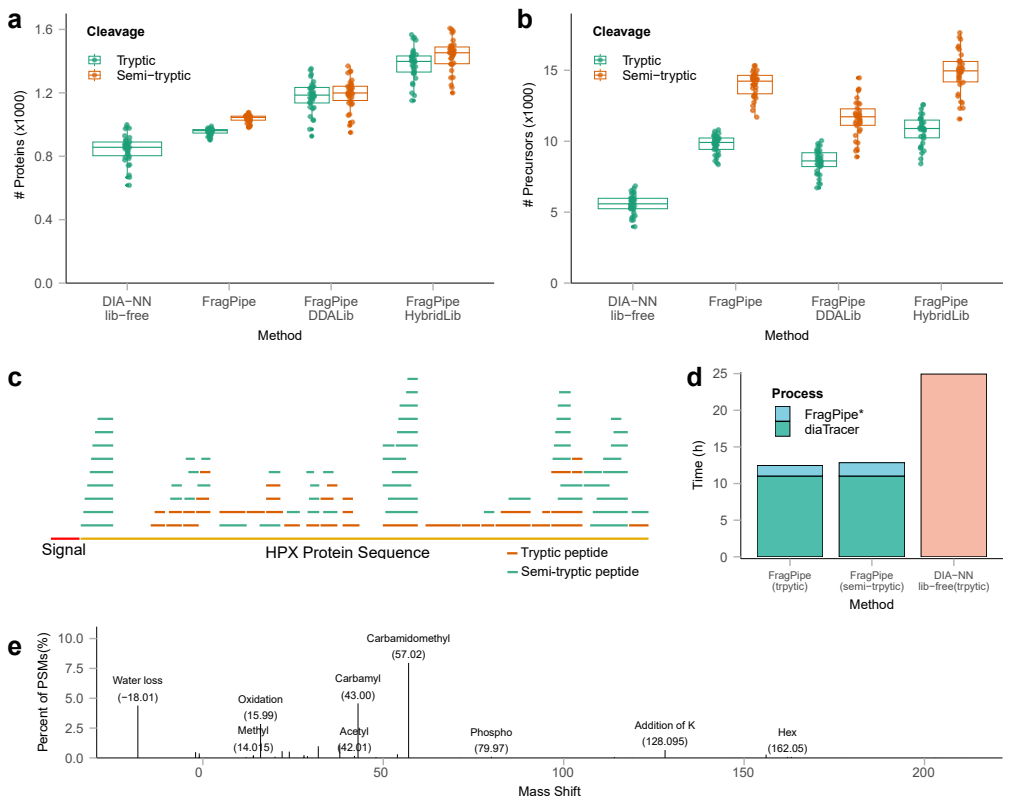
Figure 5: HLA immunopeptidomics data. a) Number of quantified immunopeptides obtained using FragPipe with diaTracer and those reported in the original study based on Spectronaut 17. **b)** Histogram of predicted binders for all HLA alleles of the corresponding sample donor, colored by binder type (light: weak binder; dark: strong binder). **c)** Length distribution of quantified immunopeptides using FragPipe with diaTracer and that reported in the original study based on Spectronaut 17 directDIA, DDA experimental library, and panlibrary. **d)** Charge state distribution of quantified immunopeptides using FragPipe with diaTracer and that reported in the original study based on Spectronaut 17 directDIA. **e)** Pseudo-MS/MS spectrum generated by diaTracer and the predicted spectrum for one of the identified immunopeptides, VYQHLFTRI. The entropy score between the two spectra is 0.9863. Visualization using FragPipe-PDV viewer.

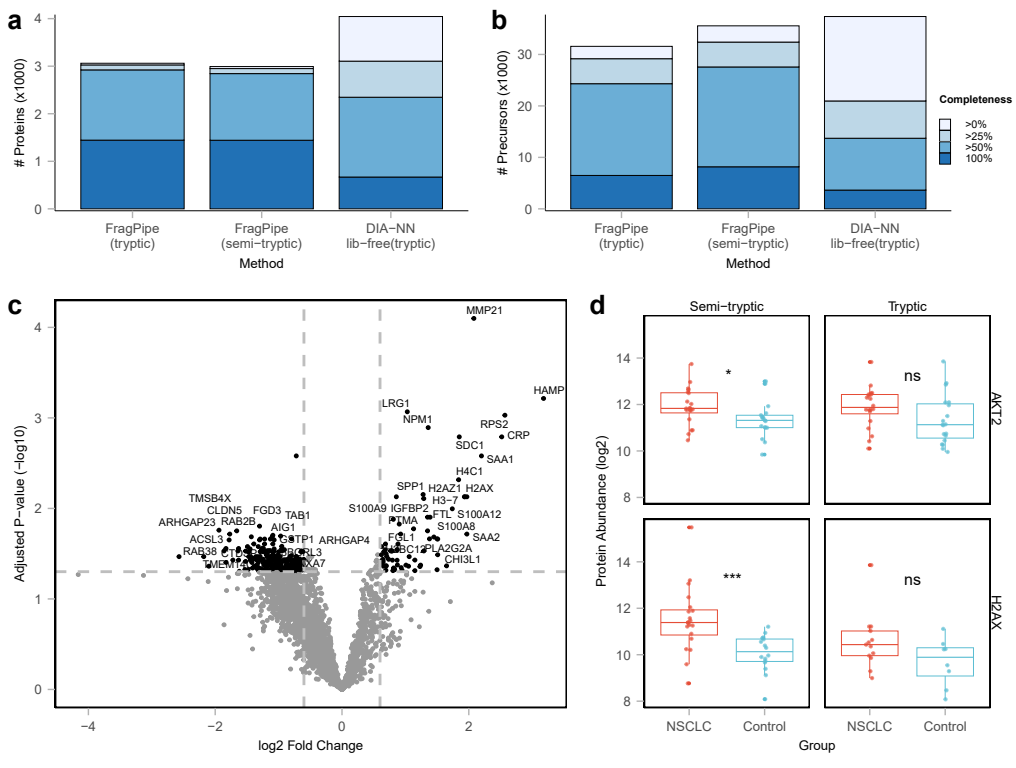
Figure 6: Phosphoproteomics data. **a)** Histogram of quantified phosphorylated peptide sequences across different gradients using FragPipe with diaTracer and those reported in the original study based on Spectronaut 16. **b)** Quantified phosphorylated peptides intensities and correlations in four replicates if the 7 min gradient time experiment. **c)** and **d)** PSM and fragment XICs of phosphorylated peptide S(Pho)PSPPDGSPAATPEIR. **e)** and **f)** PSM and fragment XICs of phosphorylated peptide SPSPPDGS(Pho)PAATPEIR. The XICs were generated by Skyline.

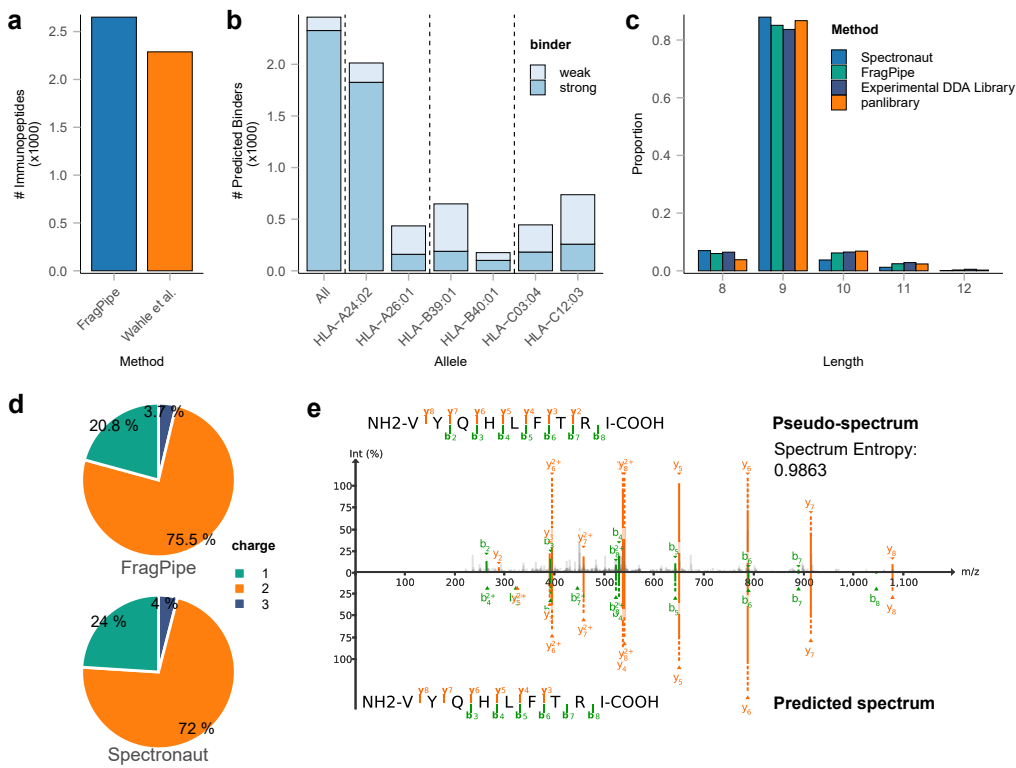
Figure 7: Low-input, spatial proteomics data. a) Number of quantified proteins after application of non-missing value (in at least one group) filter ranging from 0% to 100%, with line colors representing different methods. Red: results from the original study based on the library built using high-input samples; Green: results based on FragPipe with diaTracer, also using high-input data to build the library (“FragPipe high-input Lib”); Blue: result using diaTracer and FragPipe with low-input data only (“FragPipe”). **b)** Venn diagram of quantified proteins between the three methods, with data filtered to keep proteins with at least 70% non-missing values in at least one group. **c)** Principal-component analysis (PCA) plot of 148 samples based on 1471 proteins (after missing value filtering and data imputation) quantified using the FragPipe workflow (using low-input data only). **d)** Log2 transformed protein level abundance distribution of selected cell-type-specific proteins in different regions. In the boxplot, the central line represents the median of the numbers. The lower and upper edges of the box represent the first (Q1) and third quartiles (Q3). The interquartile range (IQR) is the box between Q1 and Q3. Whiskers extend from the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3, respectively. Data points outside this range are considered outliers and are shown as individual dots. **e)** Volcano plot showing protein abundance differences between Mantel zone and T cell zone, highlighting tissue-specific proteins (Log2 fold change >= 1; adjusted p-value <= 0.05).

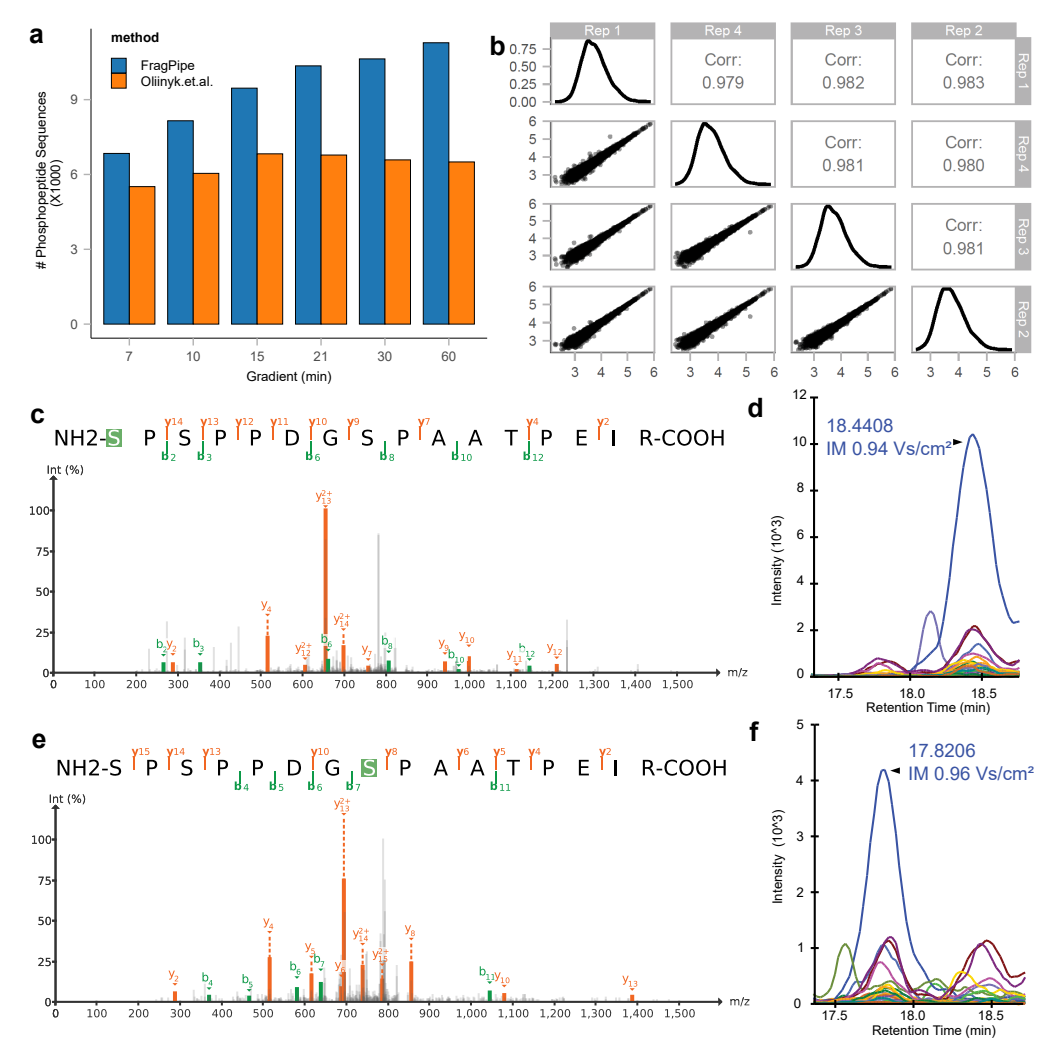

diaTRACER
diaPASEF .d data
IM
m/z
IM
RT
m/z
3D feature tracing
Precursor-fragment clustering
Pseudo-MS/MS spectrum
DDA library raw data
MSFRAGGER
MSBooster
Percolator
ProteinProphet 1% FDR filtering
EasyPQP
Spectral library
DIA-NN for Quant

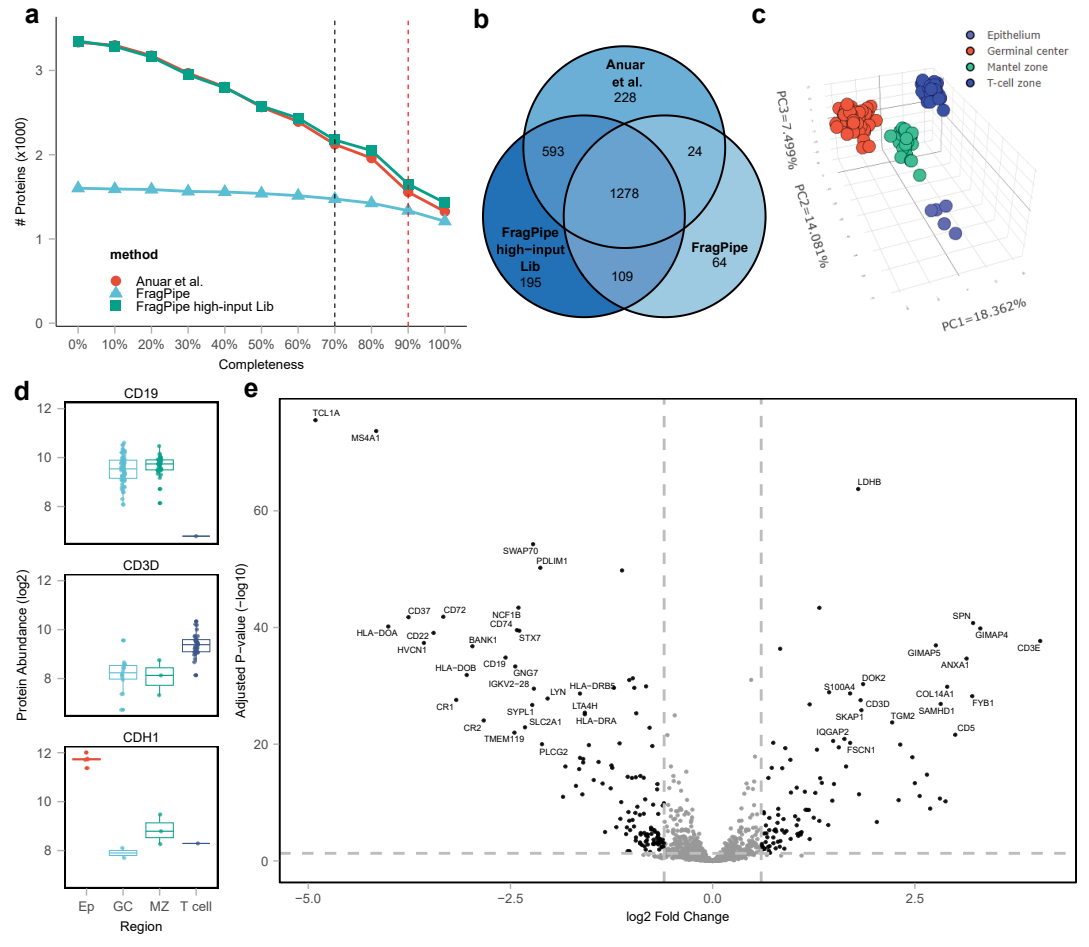

a
Direct DIA
DDA/DIA hybrid library
Proteins (x1000)
10
9
8
7
6
5
Spectronaut
DIA−NN lib−free
FragPipe
Spectronaut
FragPipe
Method
b
Direct DIA
DDA/DIA hybrid library
Proteins (x1000)
9
6
3
0
Spectronaut
DIA−NN lib−free
FragPipe
Spectronaut
FragPipe
Method
Completeness
>0%
>25%
>50%
100%


a
Cleavage
Tryptic
Semi−tryptic
Proteins (x1000)
1.6
1.2
0.8
0.4
0.0
DIA−NN lib−free
FragPipe
FragPipe DDALib
FragPipe HybridLib
Method
b
Cleavage
Tryptic
Semi−tryptic
Precursors (x1000)
15
10
5
0
DIA−NN lib−free
FragPipe
FragPipe DDALib
FragPipe HybridLib
Method
c
Signal
HPX Protein Sequence
Tryptic peptide
Semi-tryptic peptide
d
Process
FragPipe*
diaTracer
Time (h)
25
20
15
10
5
0
FragPipe (trpytic)
FragPipe (semi−trpytic)
DIA−NN lib−free(trpytic)
Method
e
Percent of PSMs(%)
10.0
7.5
5.0
2.5
0.0
Water loss (−18.01)
Oxidation (15.99)
Methyl (14.015)
Carbamyl (43.00)
Acetyl (42.01)
Carbamidomethyl (57.02)
Phospho (79.97)
Addition of K (128.095)
Hex (162.05)
0
50
100
150
200
Mass Shift


a
Proteins (x1000)
FragPipe (tryptic)
FragPipe (semi−tryptic)
DIA−NN lib−free(tryptic)
Method
b
Precursors (x1000)
Completeness
>0%
>25%
>50%
100%
c
Adjusted P−value (−log10)
log2 Fold Change
MMP21
HAMP
LRG1
NPM1
RPS2
CRP
SDC1
SAA1
H4C1
SPP1
H2AZ1
H2AX
H3−7
S100A9
IGFBP2
FTL
S100A12
PTMA
S100A8
SAA2
FGL1
PLA2G2A
CHI3L1
TMSB4X
CLDN5
FGD3
TAB1
ARHGAP23
RAB2B
AIG1
ACSL3
STP1
ARHGAP4
RAB38
d
Semi−tryptic
Tryptic
Protein Abundance (log2)
AKT2
H2AX
*
ns
***
ns
NSCLC
Control
Group


a
Immunopeptides (x1000)
2.5
2.0
1.5
1.0
0.5
0.0
FragPipe
Wahle et al.
Method
b
Predicted Binders (x1000)
2.0
1.5
1.0
0.5
0.0
binder
weak
strong
All
HLA-A24:02
HLA-A26:01
HLA-B39:01
HLA-B40:01
HLA-C03:04
HLA-C12:03
Allele
c
Proportion
0.8
0.6
0.4
0.2
0.0
Method
Spectronaut
FragPipe
Experimental DDA Library
panlibrary
8
9
10
11
12
Length
d
3.7 %
20.8 %
75.5 %
FragPipe
charge
1
2
3
4 %
24 %
72 %
Spectronaut
e
NH2-V Y Q H L F T R I-COOH
Pseudo-spectrum
Spectrum Entropy:
0.9863
Int (%)
m/z
NH2-V Y Q H L F T R I-COOH
Predicted spectrum


a
method
FragPipe
Oliinyk.et.al.
Phosphopeptide Sequences (X1000)
Gradient (min)
b
Rep 1
Rep 4
Rep 3
Rep 2
Corr: 0.979
Corr: 0.982
Corr: 0.983
Corr: 0.981
Corr: 0.980
Corr: 0.981
c
NH2-S P S P P D G S P A A T P E I R-COOH
Int (%)
m/z
d
18.4408
IM 0.94 Vs/cm²
Intensity (10^3)
Retention Time (min)
e
NH2-S P S P P D G S P A A T P E I R-COOH
Int (%)
m/z
f
17.8206
IM 0.96 Vs/cm²
Intensity (10^3)
Retention Time (min)


a
Proteins (x1000)
Completeness
method
Anuar et al.
FragPipe
FragPipe high-input Lib
b
Anuar et al. 228
FragPipe high-input Lib 195
FragPipe 64
593
24
1278
109
c
Epithelium
Germinal center
Mantel zone
T-cell zone
PC3=7.499%
PC2=14.081%
PC1=18.362%
d
CD19
CD3D
CDH1
Protein Abundance (log2)
Ep
GC
MZ
T cell
Region
e
Adjusted P-value (−log10)
log2 Fold Change
TCL1A
MS4A1
LDHB
SWAP70
PDLIM1
CD37
CD72
NCF18
CD74
HLA-DOA
CD22
HVCN1
BANK1
STX7
HLA-DOB
CD19
GNG7
IGKV2-28
LYN
HLA-DRB5
LTA4H
HLA-DRA
CR1
SYPL1
SLC2A1
CR2
TMEM119
PLCG2
SPN
GIMAP4
CD3E
GIMAP5
ANXA1
DOK2
S100A4
COL14A1
FYB1
CD3D
SAMHD1
SKAP1
TGM2
CD5
IQGAP2
FSCN1