bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fast quantitative analysis of timsTOF PASEF data with MSFragger and IonQuant

Fengchao Yu[1#], Sarah E. Haynes[1#], Guo Ci Teo[1], Dmitry M. Avtonomov[1], Daniel A. Polasky[1], Alexey I. Nesvizhskii[1,2*]

[1]Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA

[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

[#] These authors contributed equally

* Correspondence to A.I.N. (nesvi@med.umich.edu)

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Running title:

Analysis of timsTOF PASEF data with MSFragger and IonQuant

Abbreviations:

LC-MS: liquid chromatography-mass spectrometry

IM: ion mobility

TIMS: trapped ion mobility spectrometry

TOF: time-of-flight

PASEF: parallel accumulation-serial fragmentation

DDA: data-dependent acquisition

DIA: data-independent acquisition

MS/MS: tandem mass spectrometry

PSM: peptide-spectrum match

LC-IMS-MS: liquid chromatography-ion mobility-mass spectrometry

XIC: extracted ion chromatogram

CV: coefficient of variation

LFQ: label free quantification

FDR: False Discovery Rate

CID: collision-induced dissociation

ISF: In-source fragmentation

PDV: proteomics data viewer

CPU: central processing unit

PTM: post-translational modification

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Ion mobility brings an additional dimension of separation to liquid chromatography-mass spectrometry, improving identification of peptides and proteins in complex mixtures. A recently introduced timsTOF mass spectrometer (Bruker) couples trapped ion mobility separation to time-of-flight mass analysis. With the parallel accumulation serial fragmentation (PASEF) method, the timsTOF platform achieves promising results, yet analysis of the data generated on this platform represents a major bottleneck. Currently, MaxQuant and PEAKS are most commonly used to analyze these data. However, due to the high complexity of timsTOF PASEF data, both require substantial time to perform even standard tryptic searches. Advanced searches (e.g. with many variable modifications, semi- or non-enzymatic searches, or open searches for post-translational modification discovery) are practically impossible. We have extended our fast peptide identification tool MSFragger to support timsTOF PASEF data, and developed a label-free quantification tool, IonQuant, for fast and accurate 4D feature extraction and quantification. Using HeLa dataset published by Meier et al. (2018), we demonstrate that MSFragger identifies significantly (~30%) more unique peptides than MaxQuant (1.6.10.43), and performs comparably or better than PEAKS X+ (~10% more peptides). IonQuant outperforms both in terms of number of quantified proteins while maintaining good quantification accuracy. Runtime tests show that MSFragger and IonQuant can fully process a typical two hour PASEF run in under 50 minutes on a modern desktop (6 CPU cores, 32 GB RAM), significantly faster than other tools. Finally, through semi-tryptic searching, we annotate significantly (63%) more peptides. Within these semi-tryptic identifications, we report evidence of gas-phase fragmentation prior to MS/MS analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

A major challenge to identification and quantification of proteins from tissue or cultured cells is the immense complexity of the peptide mixtures that result from enzymatic preparation of these samples for liquid chromatography-mass spectrometry (LC-MS) analysis. Ion mobility (IM) spectrometry brings an additional dimension of separation to LC-MS proteomics, significantly improving peptide identification. Following electrospray ionization, IM differentiates gas-phase peptide ions by their size and charge prior to mass analysis. IM separation occurs on the millisecond timescale, improving selectivity without adding to analysis times. Recently, a commercially available instrument that couples trapped ion mobility spectrometry (TIMS) to time-of-flight (TOF) mass analysis (1) has achieved promising depth of coverage, routinely identifying over 6000 proteins from individual 120-minute LC gradients (2, 3).

Owing to the dual TIMS design of this instrument, where the first region is used for storing ions and the second for IM separation, peptides can be continually selected for sequencing with minimal reduction in duty cycle. This data acquisition method has been termed parallel accumulation-serial fragmentation (PASEF) (2, 3). For typical data-dependent acquisition (DDA) measurements, a survey scan is performed, and the N-highest abundance precursor ions are targeted for tandem mass spectrometry (MS/MS) analysis based on their mass-to-charge ratio (m/z) and mobility. Fast quadrupole switching times allow multiple peptide ions to be targeted for fragmentation during a single ion mobility scan. As a target precursor exits the TIMS region, the quadrupole switches to transmit the corresponding m/z determined by the survey scan. Synchronization of the TIMS device and quadrupole mass filter reduces chimeric spectra and enables removal of singly-charged contaminant ions. Additionally, because of the fast acquisition speed (50-200 ms for a full scan), low-abundance precursors can be repeatedly re-targeted to improve MS/MS spectrum quality (2, 3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A current major limitation of the PASEF proteomics method is long post-acquisition analysis time due to the high dimensionality of the data and large number of acquired MS/MS scans. MaxQuant (4, 5) and PEAKS (6) are both capable of processing PASEF data but require a substantial amount of time to perform standard tryptic searches. Neither MaxQuant nor PEAKS are practical for nonspecific digest searches or open searches (7), which are helpful in discovering post-translational modifications, in these data. We have recently introduced a fragment ion indexing method and its implementation in an ultrafast database search tool MSFragger (8). The speed of MSFragger makes it well suited for the analysis of large and complex data sets such as those from timsTOF PASEF. As conversion from Bruker's raw liquid chromatography-ion mobility-mass spectrometry (LC-IMS-MS) format (.d) to an open, searchable format (.mzML) represents another significant computational challenge (up to 90 minutes per single two-hour LC-MS gradient raw file), we also extended MSFragger to read the raw format directly. Here we demonstrate that MSFragger can now perform peptide identification from raw timsTOF PASEF data in a fraction of time required by other tools.

A second challenge is related to quantification in timsTOF PASEF data. Due to the added IM dimension, previously developed quantification tools need to be extended to LC-IMS-MS data. In MaxQuant it is done by slicing a 4-D space (ion mobility, m/z, retention time, and intensity) into multiple 3-D sub-spaces (m/z, retention time, and intensity) and tracing peaks within each sub-space (5). Though MaxQuant only uses every third TOF scan in feature detection, it represents a significant fraction of the overall analysis time. Similarly, PEAKS (6) has extended its functionality to support quantification of timsTOF PASEF data, with the analysis times similar to that in MaxQuant. To address this challenge, we introduce IonQuant, a quantification tool that takes Bruker's raw files and database search results as input to perform fast extracted ion chromatogram (XIC)-based quantification. Using spectral data indexing, for XIC tracing in retention and IM dimension, IonQuant requires between 10-20 minutes per file on a desktop computer. IonQuant is integrated seamlessly with MSFragger (8) and the Philosopher data validation toolkit.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Using timsTOF PASEF HeLa data recently published by Meier et al. (3), we show the application of MSFragger and IonQuant to measure the analysis speed and quantitative reproducibility across replicate injections, and to compare these results to PEAKS and MaxQuant. We demonstrate how more comprehensive (including semi-tryptic and open) searches enabled by MSFragger enable deep dives in these data, revealing interesting trends and recovering large numbers of peptides missed in the original analysis. Additionally, our pipeline is fully compatible with the Skyline environment for subsequent visualization and targeted exploration of the data, and also has its own spectral library building capabilities. Overall, we showcase a fast, flexible, and accurate computational platform for analyzing timsTOF PASEF proteomics data.

Experimental Procedures

Experimental Design and Statistical Rationale

We used data from five experimental conditions (25, 50, 100, 150, and 200 ms TIMS accumulation time) published by Meier et al.(3) in the experiments. Each experimental condition has four technical replicates. Meier et al.(3) concluded that the 100 ms accumulation time gave the best results. We used these four replicates with 100 ms accumulation time extensively (performing closed tryptic search, closed semi-tryptic search, open search, and label free quantification comparisons). For identification, we estimated the false-discovery rate (FDR) using the target-decoy based approach (9, 10). For quantification, we evaluated the quality with coefficient of variation (CV) and Pearson correlation coefficient.

Data Analysis

Raw data files from four replicate injections each of HeLa lysate acquired at five different TIMS ramp (accumulation) times on a Bruker timsTOF Pro (3) were downloaded from ProteomeXchange (11)

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(PXD010012). For all searches, a protein sequence database of reviewed Human proteins (accessed 09/30/2019 from UniProt; 40926 entries including decoys and 115 common contaminant sequences) was used unless otherwise noted. Decoy sequences were generated and appended to the original database for MSFragger. PEAKS and MaxQuant only need target sequences. Tryptic cleavage specificity was applied, along with variable methionine oxidation, variable protein N-terminal acetylation, and fixed carbamidomethyl cysteine modifications. The allowed peptide length and mass ranges were 7-50 residues and 500-5000 Da, respectively. PEAKS and MaxQuant search parameters were set as close as possible to those used by MSFragger. For MSFragger searches, peptide sequence identification was performed with version 2.2 and FragPipe version 12.1 with mass calibration and parameter optimization enabled. PeptideProphet and ProteinProphet in Philosopher (version 2.0.0; https://philosopher.nesvilab.org/) were used to filter all of peptide-spectrum matches (PSMs), peptides, and proteins to 1% PSM and 1% FDR. For PEAKS X+ searches, version 10.5 was used, and PSMs and peptides were filtered to 1% peptide FDR by clicking the FDR button on the "Summary" page. Since there is no option in PEAKS to automatically filter the proteins, we tried different protein "-10logP" scores from the smallest to the largest until the reported protein FDR was equal to 1%. MaxQuant v1.6.10.43 was used. The PSMs and peptides were filtered to 1% PSM FDR, and the protein groups were filtered to 1% protein FDR, which are the default settings.

Closed searches

Within MSFragger, precursor tolerance was set to 50 ppm and fragment tolerance was set to 20 ppm, with mass calibration and parameter optimization enabled. Two missed cleavages were allowed, and two enzymatic termini were specified. Isotope error was set to 0/1/2. The minimum number of fragment peaks required to include a PSM in modelling was set to two, and the minimum number required to report the match was four. The top 150 most intense peaks and a minimum of 15 fragment peaks required to search a spectrum were used as initial settings.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Semi-specific searches

The parameters used by MSFragger for semi-tryptic searches were equivalent to those used in the closed searches (detailed above) but with only one enzymatic peptide terminus required. MaxQuant only supports zero missed cleavage with semi-tryptic digestion. For further investigation of the identified semi-tryptic peptides, variable pyro-glutamic acid and pyro-carbamidomethyl cysteine (-17.03 Da from glutamine and cysteine), and variable water loss (-18.01) allowed on any peptide N-terminus were also included in the semi-enzymatic MSFragger search parameters.

Open searches

Precursor mass tolerance was set from -150 to +500 Da, and precursor true tolerance and fragment mass tolerance were set to 20 ppm. Mass calibration and parameter optimization were enabled. Two missed cleavages were allowed, and the number of enzymatic termini was set to two. Isotope error was set to 0. The minimum number of fragment peaks required to include a PSM in modelling was set to two, and the minimum number required to report the match was four. A minimum of 15 fragment peaks and the top 100 most intense peaks were used as initial settings.

Label-free quantification

In IonQuant, mass tolerance was set to 10 ppm, retention time tolerance was set to 0.05 minutes, and IM 1/k0 tolerance was set to 0.05. In PEAKS, identification directed quantification was performed with retention time alignment, with no CV filter nor outlier removal. Mass error, retention time shift, and ion mobility tolerances were set to 20 ppm, 20 minutes, and 0.05 1/k0, respectively. In MaxQuant, Fast LFQ (label free quantification) was performed with large ratio stabilization, min ratio count set to one (except where noted), three minimum neighbors, and six average number of neighbors. The remaining parameters were set to default values.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Protein quantification with MSstats

MSstats was used to calculate protein abundances based on the ion abundances reported by each tool. For MSFragger and PEAKS, ions (filtered at 1% PSM and 1% protein FDR for MSFragger; 1% peptide FDR for PEAKS) in the MSstats compatible format were provided to MSstats. For MaxQuant, evidence.txt (filtered at 1% PSM FDR) and proteinGroup.txt (filtered at 1% protein FDR) were provided to MSstats. The dataProcess function with log10 intensity transformation was used to calculate protein abundances.

Runtime comparisons

MSFragger (v2.2, via FragPipe v12.1) and MaxQuant (v1.6.10.43) were compared on a desktop with Intel Optane SSD 900P series hard disk, Intel Core i7-8700 3.2 GHz 6 CPU cores (12 logical cores), and 32 GB memory. Due to installation and licensing constraints, PEAKS Studio X+ was used on an Intel Xeon Gold 2.4 GHz 20 CPU cores (40 logical cores) workstation with 96 GB RAM.

Results and Discussions

Workflow Overview

The overview of the computational workflow in shown in **Figure 1**. MS/MS spectral files acquired in PASEF mode can be read directly by MSFragger. MSFragger loads the raw format (.d) using our original spectral reading library MSFTBX (12), extended here to interact with the Bruker's native library. During loading, Bruker's native library functions are called to perform scan combining, peak picking, and de-noising. After loading, MSFragger writes all extracted scans in to a binary format, mzBIN, for fast data access in the future re-analyses of the same data. After database searching with MSFragger (see Experimental Procedures), PSMs are saved in the pepXML file format. PSMs are processed using PeptideProphet (13) and ProteinProphet (14) as part of the Philosopher toolkit. Philosopher is also used for FDR filtering, and for generating summary reports at the PSM, peptide

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ion, peptide, and protein levels (**Figure 1a**). Finally, IonQuant (see below) is used to extract peptide ion intensities for all PSMs passing the FDR filter, and adds quantification information to the PSM, peptide, and protein-level tables.

IonQuant Algorithm

Spectral files generated by timsTOF PASEF are larger and more structurally complex than traditional LC-MS data due to the fast TOF scan rate and additional IM dimension. IonQuant, written in Java, traces and quantifies features from the four-dimensional space (ion mobility, m/z, retention time, and intensity) quickly and accurately using indexing technology (**Figure 1b**). IonQuant first digitizes the ion mobility dimension with a predefined bin width (0.002 $1/K_0$; Vs/cm^2). Then, IonQuant indexes all peaks within this 4D space according to their ion mobility, m/z, and retention time, which reduces memory usage and accelerates subsequent peak tracing. Given precursor m/z and retention time from an identified MS/MS spectrum, IonQuant uses the index to collect all related peaks. Then, it generates a curve with respect to retention time by tracing and performing Gaussian smoothing. After tracing all peaks in the retention time dimension, IonQuant traces the ion mobility dimensions by clustering adjacent peaks to form 4-D features. Finally, IonQuant reports the boundaries, apex location, and apex intensity of each detected ion feature.

IonQuant takes spectral files (.d, Bruker's raw format, using MSFTBX as in MSFragger) and peptide identifications (pepXML) as input and outputs a csv file containing quantified results for each spectral file. When used with Philosopher summary tables as input, IonQuant adds quantification information directly to the tables containing validated PSM, peptide, and protein results. In combining protein intensities across multiple experiments, IonQuant uses an approach similar to that of DIA-Umpire (15). Each protein's intensity is the summed intensity of top *n* ions identified in *t* percentage of all experiments, where *n* and *t* are parameters with default values of infinity (i.e. using all) and 50%, respectively. In addition, IonQuant also uses the quantified features and the PSM table from Philosopher to generate an MSstats (16) compatible file for downstream analysis using that tool.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Peptide and Protein identification

We monitored runtime and sensitivity of database searching and quantification using four replicate injections of HeLa cell digest (see **Experimental Procedures**). The data was analyzed using MSFragger with IonQuant and compared to the results from MaxQuant and PEAKS. MSFragger identified 58954 peptides and 6525 proteins from a standard tryptic search, more than the other tools (**Figure 2a, Supporting Table S1-S4**). Uniqueness of the peptide identifications obtained by PEAKS, MaxQuant, and MSFragger from four replicate injections of HeLa cell digest is shown in **Figure 2b**. MSFragger with IonQuant also required significantly less total analysis time than PEAKS or MaxQuant (**Figure 2c**). Furthermore, when MSFragger was used to perform subsequent searches on the same raw files (i.e. starting with mzBIN files), the total processing times were below 20 minutes per file, more than nine times faster than PEAKS or MaxQuant (**Figure 2c**). We also note that a similarly fast speed can be achieved when using MGF files as input to MSFragger (generation of MGF files can be scheduled as an additional post-processing step in the instrument's Data Analysis software immediately following data acquisition). In such a workflow, protein quantification would be limited to MS/MS-based spectral counts only, which is nevertheless sufficient for certain applications such as sample quality control or interactome analysis using affinity-purification mass spectrometry (17).

Protein Quantification (Tryptic Search)

We evaluated the quantitative performance of MSFragger with IonQuant, and compared with MaxQuant and PEAKS, using the tryptic search results (see **Experimental Procedures**) from the same four HeLa replicates (**Table 1**). Because each tool groups peptides and performs protein quantification differently, we used MSstats (16) to independently calculate protein abundances from ions quantified by these tools. Across the four replicate injections, IonQuant with MSstats demonstrated excellent reproducibility, with Pearson correlation between replicates of 0.979 or above (**Figure 3a**), higher than that from PEAKS and MaxQuant (**Supporting Figure S1**). The distribution of CVs for each protein among the tools is shown in **Figure 3b**. Considering proteins quantified in at

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

least two replicates, IonQuant with MSstats quantified the most proteins (5961) while exhibiting the smallest median CV across replicates of 0.059, compared to PEAKS-MSstats (0.070) and MaxQuant-MSstats (0.072). Protein abundances reported by IonQuant correlated with those reported by PEAKS and MaxQuant with Pearson correlations of 0.873 and 0.736, respectively (**Figure 3c, Supporting Figure S2**). We noticed that both MaxQuant and PEAKS X+ report the volume/area of the traced peaks while IonQuant reports the apex intensity(18) of the traced peaks. We demonstrate that using apex resulted in higher accuracy and lower noise. Each tool, including IonQuant, can also perform peptide to protein roll-up and report protein-level quantification ('native' quantification in **Table 1**). However, our analysis shows that post-processing using MSstats performed as well or better than native protein-level quantification methods for all three tools. For MaxQuant, applying an addition filter of min 2 peptides per protein for quantification (which is a default option in MaxQuant) reduced the mean protein CV to 0.057. However, this was associated with a very significantly drop in the total number of proteins quantified in at least two replicates (from 5335 to 4040, **Table 1**).

Open Search Analysis

Using MSFragger and IonQuant, we performed a quantitative open search on the four HeLa replicates acquired with 100 ms accumulation time. After statistical evaluation and filtering by Philosopher, mass shifts corresponding to water and ammonia losses (-17 and -18 Da, respectively) were the most prominent, followed by a +52.91 Da mass shift that corresponds to substitution of three protons with Fe(III), possibly an artefact from sample handling. Open search also revealed the presence of many semi-tryptic (neutral loss) peptides. Plots displaying the number of PSMs for each of these mass shifts are shown in **Supporting Figure S3 (Supporting Table S5-S6)**. Note that MSFragger and IonQuant analysis times were not significantly longer for open search.

Semi-tryptic Peptide Monitoring

From the open search, we observed a significant number of semi-tryptic PSMs, and PSMs with water and ammonia loss. Intrigued by these observations, we investigated whether these observations were

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

indicative of ion activation prior to MS/MS analysis. To this end, we performed semi-tryptic searches (also allowing -17 and -18 Da losses, see **Experimental Procedures**) on the HeLa data acquired with different TIMS accumulation times (3), during which trapping in the first TIMS region and mobility separation in the second occur. Across the five different accumulation times tested in the publication (25, 50, 100, 150, and 200 ms), we observed that the number of PSMs with only one enzymatic terminus increases with accumulation time (**Figure 4a**). The relationship between accumulation time and semi-tryptic peptides is likely due in part to increased sensitivity. The number of peptide ions that can be targeted for fragmentation increases with accumulation time (3), so low-intensity ions are more likely to be detected when longer accumulation times are used. This can be seen in **Figure 4b,** where the share of total ion intensity from semi-tryptic peptides increases as the instrument has more time to interrogate these lower-abundance ions.

At 100 ms accumulation time, which was selected as optimal by the original manuscript authors, a semi-tryptic MSFragger search resulted in an astonishing 63% increase in the number of identified peptides (from 58954 to 95967) across four replicates (**Figure 2** and **Figure 4c**). The number of identified proteins in MSFragger search increased as well (from 6525 to 6749). Both PEAKS and MSFragger identified more unique peptides with a semi-tryptic search (**Figure 2a,b**), PEAKS identified ~63% more and MSFragger identified ~58% more, while MaxQuant results did not reflect a noticeable increase. This may be partially due to the fact that MaxQuant does not allow missed cleavages in semi-tryptic searches. Of those peptides with a single enzymatic terminus identified by the semi-tryptic search, the majority (67%) were found alongside their full-length tryptic form. We also demonstrate that MSFragger with IonQuant quantifies more proteins in semi-tryptic vs. tryptic search without compromising accuracy (**Table 1**). It is also worth noting that, due to fast fragment ion indexing, MSFragger's runtime advantage over MaxQuant and PEAKS is even greater when performing semi-enzymatic searches (**Figure 2c**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We further investigated the source of these semi-enzymatic peptides by comparing the observed cleavage sites to established gas-phase peptide fragmentation behavior. In all cases where a peptide was found to be semi-enzymatic, proline was found C-terminal to the cleavage site, a well-known product of fragmenting positively-charged peptides (19, 20). In the semi-tryptic searches, we allowed a neutral loss of H_2O from any N-terminal residue. We observed an increase in the percentage of PSMs containing a neutral water loss with longer accumulation times (**Figure 4d**), as would be expected for a gas-phase fragmentation event. As described previously (21-23), water loss from N-terminal glutamine and glutamate is frequently observed following collision-induced dissociation (CID) of peptides. Of the peptides identified with N-terminal semi-tryptic cleavages across the entire dataset (four replicates each of five accumulation times), we observed that water loss occurred preferentially when glutamine or glutamate were present C-terminal to the cleavage site (**Figure 4e**). As the semi-enzymatic peptides identified in this data set display neutral losses characteristic of CID, it appears peptide ion activation occurred in the dual TIMS device, resulting in the majority of the semi-enzymatic peptides we observe.

The high rates of semi-enzymatic PSMs may be specific to the timsTOF datasets used in this work, and these analyses should be repeated when more datasets become publicly available. We expect improvements in instrument tuning to provide gentler peptide ion handling and therefore less fragmentation within the instrument. Despite the clear reduction in semi-enzymatic PSMs with altered tuning settings, reducing the energy imparted by the source and initial ion optics can reduce ion transmission, in some cases dramatically. In many analyses, it may thus be preferable to use higher energies in the instrument source (or later ion optics such as the TIMS device itself) to improve transmission efficiency despite increased fragmentation of some peptides, making a semi-enzymatic search necessary to recover the identities of all peptides analyzed (24) and maximize the sensitivity of the instrument. Furthermore, certain analyses, such as those of glycopeptides (25) may also benefit from in-source pseudo-MS^3 capabilities to enable advanced analysis methods. As the in-TIMS

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

fragmentation level appears to be tunable, the instrument appears to have the capability to perform these pseudo-MS^3 methods as well.

Spectral Library Generation

The search results from MSFragger (after processing with Philosopher/PeptideProphet) can also be fed into Skyline (26) to generate spectral libraries and inspect peptide features in three dimensions (**Supporting Figure S4**). By providing Skyline with 1% FDR filtered protein list (generated by Philosopher, in fasta format), Skyline libraries can be effectively created with a desired protein level and peptide ion FDR filters (e.g. 1% protein FDR and 1% peptide ion FDR). A detailed tutorial for importing and visualizing the results from MSFragger search in Skyline can be found on the MSFragger webpage (https://msfragger.nesvilab.org/tutorial_pasef_skyline.html). Furthermore, the spectral library building tool EasyPQP (https://github.com/grosenberger/easypqp) has been adapted to be used with ion mobility data, and we incorporated this capability into the MSFragger user interface FragPipe. This feature allows building spectral libraries from DDA data as part of the data-independent acquisition (DIA) data analysis workflows, e.g. for subsequent quantification from diaPASEF data using OpenSWATH or Spectronaut tools (27). Running EasyPQP on MSFragger tryptic search results of the four HeLa replicates (100 ms accumulation time) resulted in a spectral library containing 58931 peptides.

Conclusions

Due to the efficient parallel accumulation strategy and the added selectivity of trapped ion mobility, the timsTOF PASEF method has achieved highly sensitive proteomics measurements. We have extended MSFragger to directly read raw PASEF data for rapid database searching, and developed IonQuant to accurately quantify peptides and proteins from these data. For standard tryptic searches, MSFragger requires less than half the analysis time needed by other tools that currently support PASEF data, and is three to five times faster for semi-enzymatic searching. MSFragger is the only PASEF-compatible

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

search engine with the ability to conduct open searches in reasonable time. The flexibility afforded by MSFragger's modest analysis times can be applied for post-translational modification (PTM) discovery or screening for artefacts of sample preparation or data acquisition. Overall, we report data analysis times that remove a primary bottleneck in the usability of timsTOF PASEF data. MSFragger and IonQuant enable fast, sensitive, and precise quantitative proteomic analyses, including semi-specific and open searches, as well as spectral library generation for diaPASEF analysis workflows. A match-between-runs capability for IonQuant is also currently under development. This entire pipeline can be accessed through a graphical user interface FragPipe (http://fragpipe.nesvilab.org/) or with the command line for high-throughput applications. Outputs are also compatible with tools such as Skyline, MSstats, and with proteomics data viewer PDV (28) for visualization of peptide assignments to MS/MS spectra, enabling a variety of complete workflows.

Acknowledgements

The authors would like to thank Markus Lubeck and Florian Meier for helpful discussions. We would like to thank George Rosenberger for assistance with EasyPQP. We also thank the users of our tools for their feedback. This work was funded in part by NIH grants R01-GM-094231 and U24-CA210967.

Data and Software Availability

The data used in the manuscript are published by Meier et al. (3) and can be found from the ProteomeXchange Consortium via the PRIDE partner repository (29) with the dataset identifier PXD010012 (https://www.ebi.ac.uk/pride/archive/). MSFragger and IonQuant programs were developed in the cross-platform Java language and can be accessed at http://msfragger.nesvilab.org/ and https://github.com/Nesvilab/IonQuant.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Author Contributions

F.Y. adopted MSFragger to timsTOF data and developed the algorithm of IonQuant, with contribution form G.C.T. and D.A.; S.E.H., F.Y., and A.I.N. analyzed the data; A.I.N. supervised the entire project; S.E.H., F.Y., D.A.P., and A.I.N. wrote the manuscript with input from all authors.

Competing Interests Statement

The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Silveira, J. A., Ridgeway, M. E., Laukien, F. H., Mann, M., and Park, M. A. (2017) Parallel accumulation for 100% duty cycle trapped ion mobility-mass spectrometry. *International Journal of Mass Spectrometry* 413, 168-175

2. Meier, F., Beck, S., Grassl, N., Lubeck, M., Park, M. A., Raether, O., and Mann, M. (2015) Parallel accumulation–serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *Journal of proteome research* 14, 5378-5387

3. Meier, F., Brunner, A. D., Koch, S., Koch, H., Lubeck, M., Krause, M., Goedecke, N., Decker, J., Kosinski, T., Park, M. A., Bache, N., Hoerning, O., Cox, J., Rather, O., and Mann, M. (2018) Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* 17, 2534-2545

4. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* 26, 1367–1372

5. Prianichnikov, N., Koch, H., Koch, S., Lubeck, M., Heilig, R., Brehmer, S., Fischer, R., and Cox, J. (2019) MaxQuant software for ion mobility enhanced shotgun proteomics. *bioRxiv*, 651760

6. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. (2012) PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics* 11, M111. 010587

7. Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology* 33, 743–749

8. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature methods* 14, 513

9. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4, 207-214

10. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics* 73, 2092-2123

11. Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., and Bandeira, N. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* 32, 223

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

12. Avtonomov, D. M., Raskind, A., and Nesvizhskii, A. I. (2016) BatMass: a Java Software Platform for LC-MS Data Visualization in Proteomics and Metabolomics. *J Proteome Res* 15, 2500-2509

13. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* 74, 5383-5392

14. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry* 75, 4646-4658

15. Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., and Nesvizhskii, A. I. (2015) DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* 12, 258-264, 257 p following 264

16. Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B., and Vitek, O. (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 30, 2524-2526

17. Choi, H., Larsen, B., Lin, Z. Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z. S., Tyers, M., Gingras, A. C., and Nesvizhskii, A. I. (2011) SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature methods* 8, 70-73

18. Argentini, A., Goeminne, L. J., Verheggen, K., Hulstaert, N., Staes, A., Clement, L., and Martens, L. (2016) moFF: a robust and automated approach to extract peptide ion intensities. *Nat Methods* 13, 964-966

19. Breci, L. A., Tabb, D. L., Yates, J. R., and Wysocki, V. H. (2003) Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Analytical chemistry* 75, 1963-1971

20. Huang, Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2005) Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Analytical chemistry* 77, 5800-5813

21. Neta, P., Pu, Q.-L., Kilpatrick, L., Yang, X., and Stein, S. E. (2007) Dehydration versus deamination of N-terminal glutamine in collision-induced dissociation of protonated peptides. *Journal of the American Society for Mass Spectrometry* 18, 27-36

22. Savitski, M. M., Kjeldsen, F., Nielsen, M. L., and Zubarev, R. A. (2007) Relative specificities of water and ammonia losses from backbone fragments in collision-activated dissociation. *Journal of proteome research* 6, 2669-2673

23. Harrison, A. G. (2003) Fragmentation reactions of protonated peptides containing glutamine or glutamic acid. *Journal of mass spectrometry* 38, 174-187

24. Kim, J.-S., Monroe, M. E., Camp, D. G., Smith, R. D., and Qian, W.-J. (2013) In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *Journal of proteome research* 12, 910-916

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

25. Zhao, J., Song, E., Zhu, R., and Mechref, Y. (2016) Parallel data acquisition of in-source fragmented glycopeptides to sequence the glycosylation sites of proteins. *Electrophoresis* 37, 1420-1430

26. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26, 966-968

27. Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., MacLean, B., Rost, H. L., Tate, S. A., Tsou, C. C., Reiter, L., Distler, U., Rosenberger, G., Perez-Riverol, Y., Nesvizhskii, A. I., Aebersold, R., and Tenzer, S. (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nature biotechnology* 34, 1130-1136

28. Li, K., Vaudel, M., Zhang, B., Ren, Y., and Wen, B. (2019) PDV: an integrative proteomics data viewer. *Bioinformatics* 35, 1249-1251

29. Vizcaino, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* 44, D447-456

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1. Comparison of protein quantification within MSstats to protein quantification reported by each tool (native). Median protein coefficient of variation (CV) across replicates is shown. The number of quantified proteins refers to those quantified in at least two replicates. For all searches, two missed cleavages are allowed except for MaxQuant's semi-tryptic search that only support zero missed cleavage.

Tool	**MSstats proteins quantified**	**MSstats CV**	**Native proteins quantified**	**Native CV**
PEAKS (tryptic)	5227	0.070	5359	0.203
MaxQuant (tryptic)	5261	0.072	5335	0.072
MaxQuant (tryptic, min 2 pep)	5261	0.072	4040	0.057
MSFragger-IonQuant (tryptic)	5961	0.059	5940	0.091
PEAKS (semi-tryptic)	5406	0.066	5527	0.194
MaxQuant (semi-tryptic)	4740	0.072	4839	0.071
MaxQuant (semi-tryptic, min 2 pep)	4740	0.072	3526	0.054
MSFragger-IonQuant (semi-tryptic)	6118	0.055	6088	0.090

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

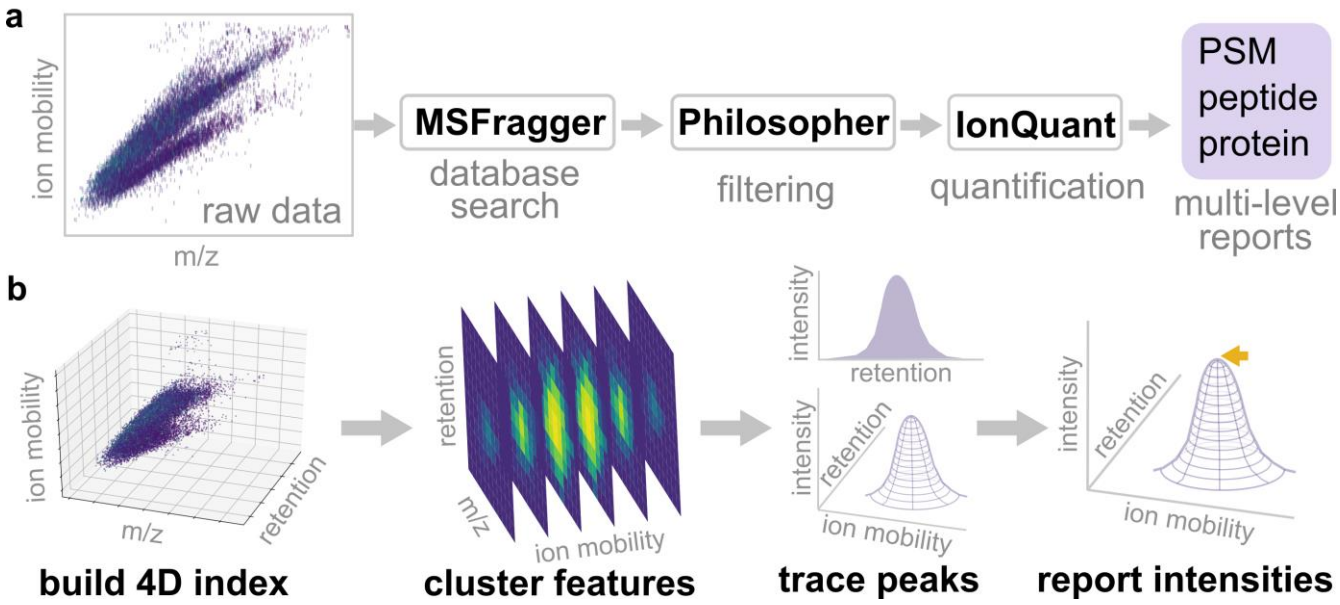

Figure 1. (a) Overview of the analysis workflow. Raw Bruker timsTOF data are converted (to mzBIN format) and searched with MSFragger to identify peptides from MS/MS spectra. Identifications are processed with Philosopher (PeptideProphet, ProteinProphet, FDR estimation) and FDR-filtered reports are generated at the PSM, peptide ion, peptide, and protein levels. IonQuant performs quantification and generates final reports. (b) Schematic of the IonQuant algorithm. Raw Bruker timsTOF data are loaded and indexed. Then, IonQuant clusters features and traces peaks (for all identified peptide ions) in IM and retention time dimensions. Finally, IonQuant locates the apex of each peak (peptide ion) and reports its apex intensity.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

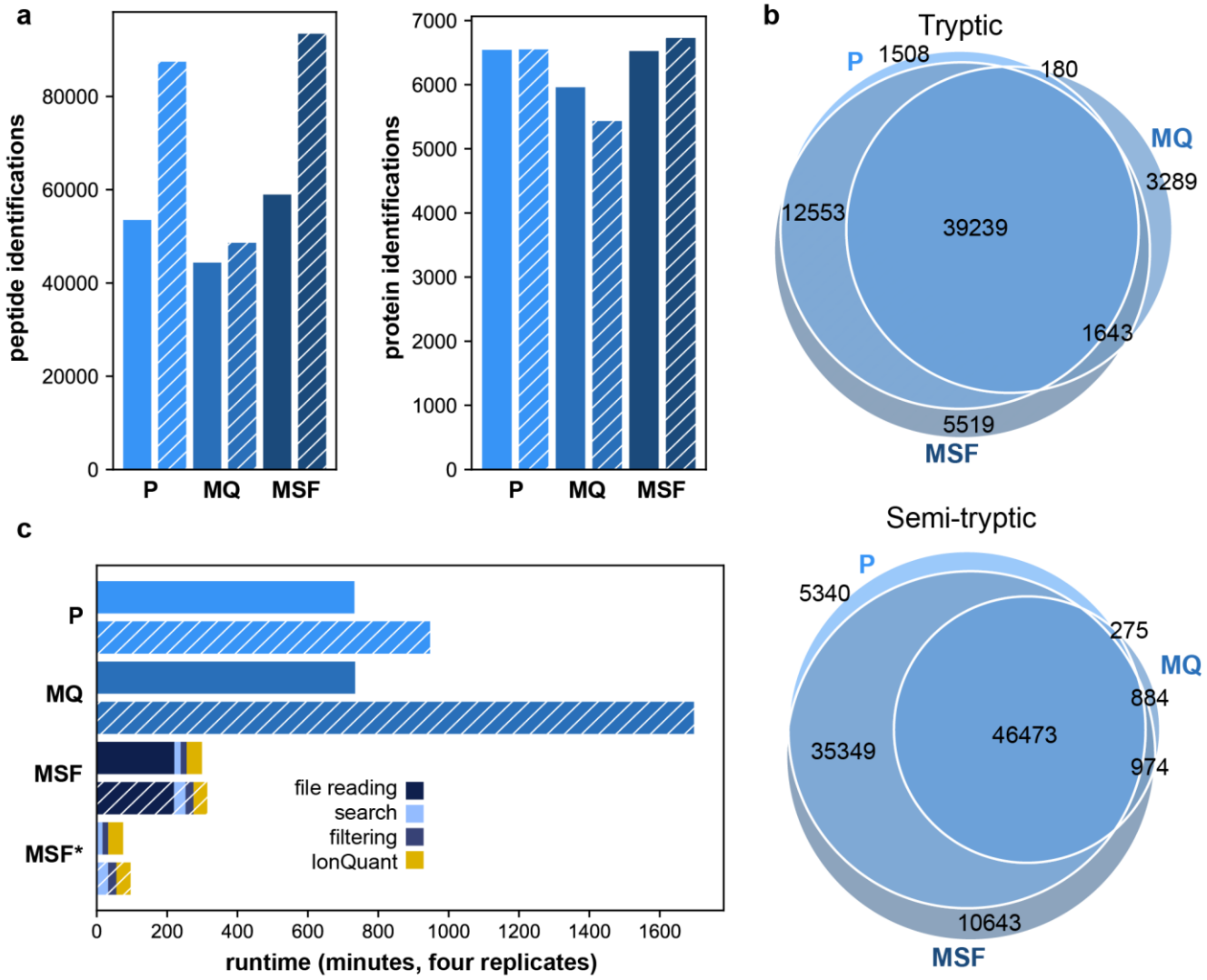

Figure 2. Feature identification and run time comparison. PEAKS Studio X+ ("P"), MaxQuant v1.6.10.43 ("MQ"), and MSFragger 2.2 ("MSF") results for four HeLa replicates acquired with 100 ms accumulation time. Hatching indicates results from semi-enzymatic search. (a) Peptide (left) and protein (right) identifications. (b) Comparison of non-redundant peptide sequences identified by each tool. (c) Total analysis times for each tool. MSF* denotes MSFragger search when mzBIN files are available. MSFragger analysis times are broken down into raw file reading (i.e. conversion to mzBIN), database searching, filtering, and quantification with IonQuant.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

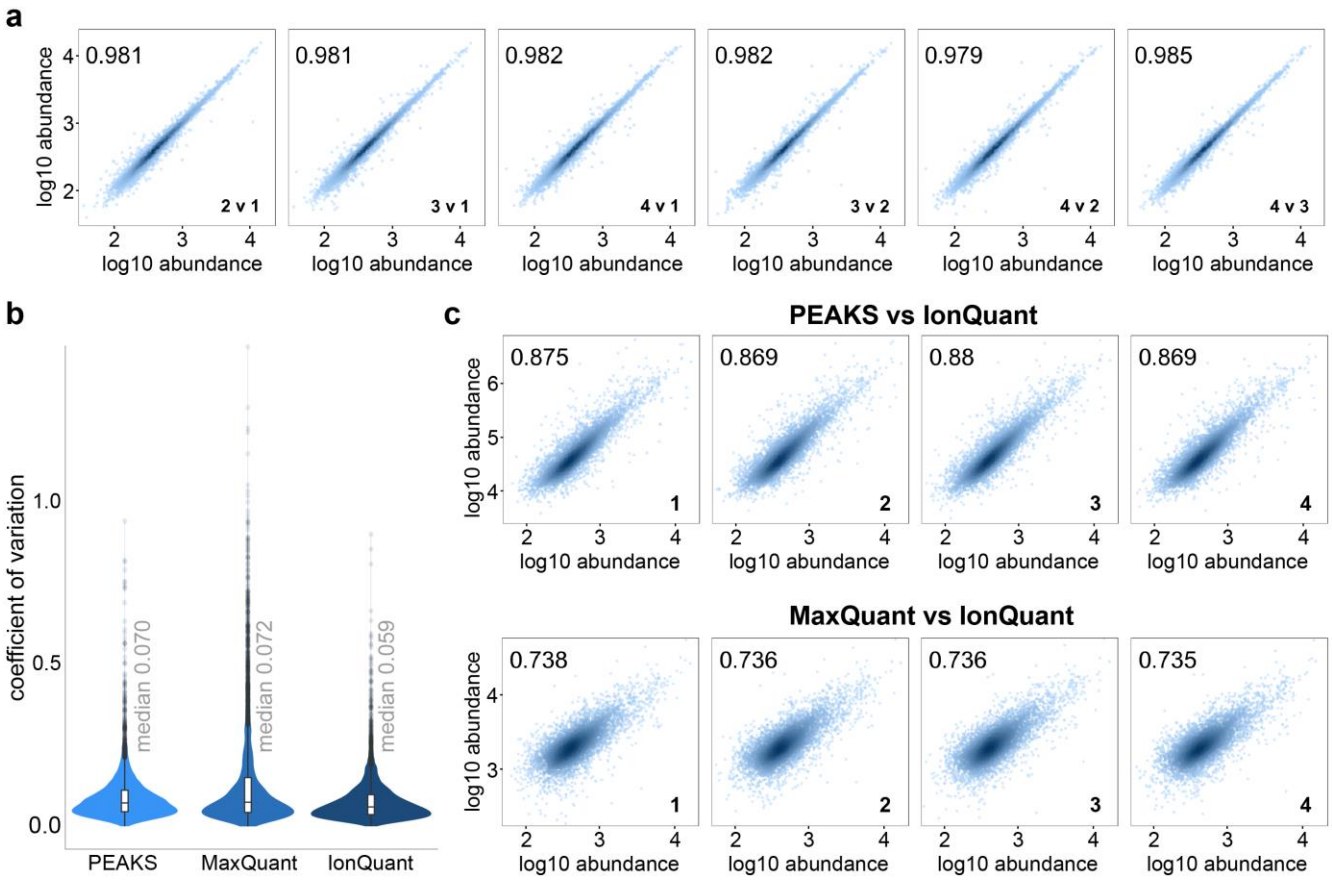

Figure 3. Protein quantification (with MSstats). (a) Correlation of quantified proteins between four technical replicates, MSFragger-IonQuant results. Each paired comparison is labeled in the bottom right-hand corner of the plot. (b) Protein coefficient of variation across the four replicates, comparing PEAKS, MaxQuant, and MSFragger-IonQuant. Replicates are labeled in the bottom right-hand corner of each plot. (c) Comparison of MSFragger-IonQuant protein abundances to PEAKS and MaxQuant for each replicate.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.19.999334; this version posted March 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

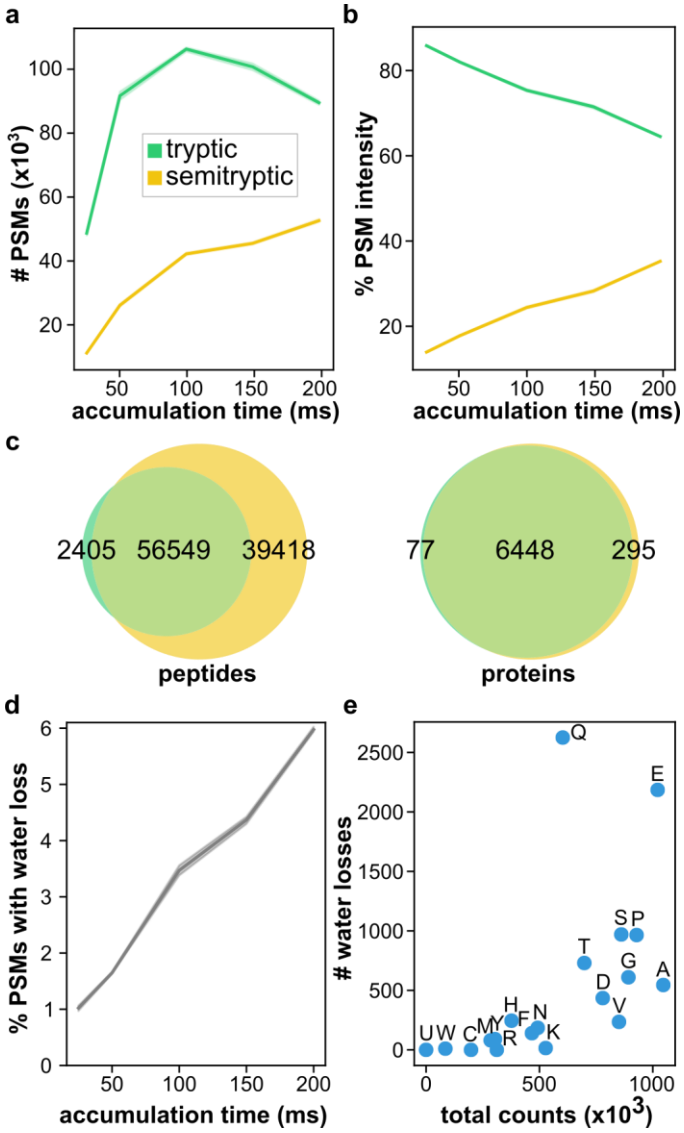

Figure 4. Semi-tryptic searching with MSFragger monitors fragmentation within dual TIMS device. The total number of semi-tryptic PSMs (a) and the percentage of total precursor intensity from semi-tryptic PSMs (b) increase with accumulation time. (c) More peptides and proteins are identified using semi-tryptic search with MSFragger (four pooled HeLa replicates, 100 ms accumulation time). For semi-tryptic search, variable pyro-glutamic acid and pyro-carbamidomethyl cysteine (-17.03 Da from glutamine and cysteine), and variable water loss (-18.01) allowed on any peptide N-terminus were added. (d) The percentage of PSMs displaying neutral water loss increases with accumulation time. (e) Water losses for each amino acid following the cleavage site are plotted against the total occurrences of the amino acid in the data set. For each line plot, shaded areas represent the 95% confidence interval from four replicates.