Journal of proteome research

pubs.acs.org/jpr

Technical Note

Comparative Evaluation of Proteome Discoverer and FragPipe for the TMT-Based Proteome Quantification

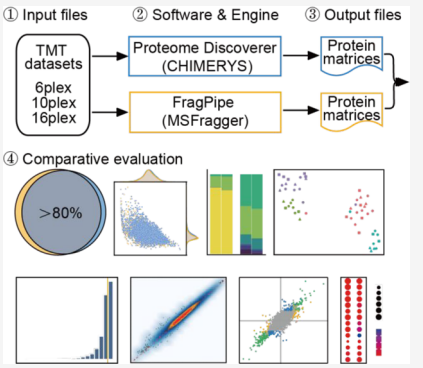
Tianen He,[#] Youqi Liu,[#] Yan Zhou, Lu Li, He Wang, Shanjun Chen, Jinlong Gao, Wenhao Jiang, Yi Yu, Weigang Ge, Hui-Yin Chang, Ziquan Fan, Alexey I. Nesvizhskii,* Tiannan Guo,* and Yaoting Sun*

Cite This: *J. Proteome Res.* 2022, 21, 3007−3015

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Isobaric labeling-based proteomics is widely applied in deep proteome quantification. Among the platforms for isobaric labeled proteomic data analysis, the commercial software Proteome Discoverer (PD) is widely used, incorporating the search engine CHIMERYS, while FragPipe (FP) is relatively new, free for noncommercial purposes, and integrates the engine MSFragger. Here, we compared PD and FP over three public proteomic data sets labeled using 6plex, 10plex, and 16plex tandem mass tags. Our results showed the protein abundances generated by the two software are highly correlated. PD quantified more proteins (10.02%, 15.44%, 8.19%) than FP with comparable NA ratios (0.00% vs. 0.00%, 0.85% vs. 0.38%, and 11.74% vs. 10.52%) in the three data sets. Using the 16plex data set, PD and FP outputs showed high consistency in quantifying technical replicates, batch effects, and functional enrichment in differentially expressed proteins. However, FP saved 93.93%, 96.65%, and 96.41% of processing time compared to PD for analyzing the three data sets, respectively. In conclusion, while PD is a well-maintained commercial software integrating various additional functions and can quantify more proteins, FP is freely available and achieves similar output with a shorter computational time. Our results will guide users in choosing the most suitable quantification software for their needs.



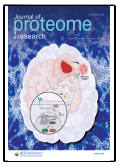
KEYWORDS: *FragPipe, Proteome Discoverer, tandem mass tag, labeled quantitative proteomics, mass spectrometry*

INTRODUCTION

Tandem mass tag (TMT) labeling is a tandem mass spectrometry (MS/MS)-based protein quantification method developed in 2003.[1] This technology can be used to analyze multiple labeled samples in one experiment, and over the years, the multiplexity of TMT has expanded from 2plex[1] to 18plex.[2] TMT enables the accurate measurement of the relative abundances of proteins from different samples pooled together while decreasing random variations between runs.[3] The increase in sample throughput demands efficient analysis pipelines that support TMT-based spectral data. Therefore, numerous computational tools have been developed for MS/MS-based protein quantification. As a commercial software first released in 2007, Proteome Discoverer (PD, Thermo Fisher Scientific) is well established and widely used to process proteome data and supports both isobaric labeling and label-free quantification.[4,5] PD is flexible, stable, and allows access to the original MS/MS spectral data from various protein qualitative and quantitative methods.[6] The release of PD version 3.0 in 2021 has further combined the artificial intelligence-based search algorithm CHIMERYS (MSAID GmbH) to increase the identification numbers.[7] However, the license of PD is costly and its execution is time-consuming, making it suboptimal for large-scale studies.[8,9] Therefore, new software tools have been developed due to the unmet need for a free and more efficient proteome quantification. FragPipe (FP) is one of the few software (PD, Mascot Server, OpenMS, etc.) compatible with the TMTpro 16plex labeling method launched in 2020.[10] This tool provides a complete pipeline for identifying, validating, and quantifying proteins from spectral files. FP integrates many tools, including the novel ultrafast proteome database search engine MSFragger,[11] whose search speed and peptide-spectrum match identifications have improved further.[12] Different built-in workflows are supplied in FP, supporting a wide range of experimental designs, such as isobaric labeling-based quantification and data-independent acquisition. Compared with PD, FP is available freely for noncommercial use (download from GitHub at https://github.com/Nesvilab/FragPipe) in addition to the advantage of providing advanced analysis workflows, e.g., for glycoproteomics and for comprehensive analysis of post-translational modification via open database searching.

In this study, we compared the performances of PD and FP in quantifying TMT-based proteomic data. In particular, we

Received: July 1, 2022
Published: October 31, 2022



https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

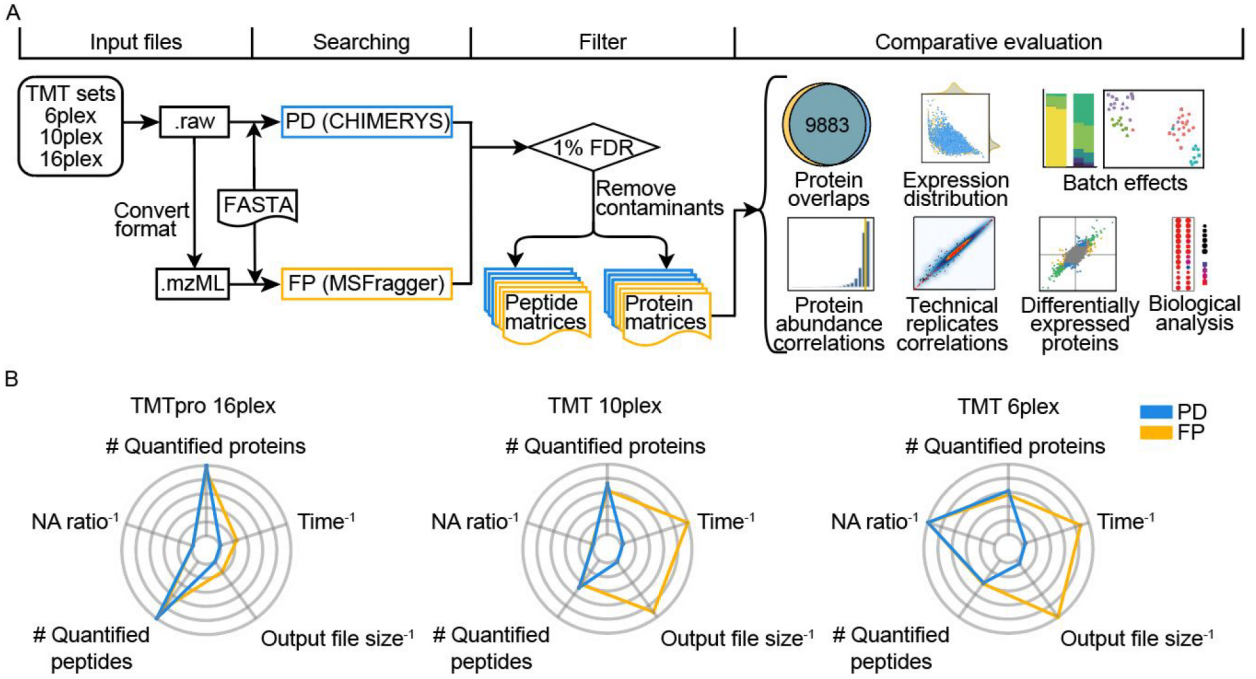


Figure 1. Comparison of Proteome Discoverer (PD) and FragPipe (FP) in the protein quantification of tandem mass tag (TMT)-labeled data sets. (A) Study design. (B) Comparison of PD and FP based on their execution time^{-1}, output file size^{-1}, # quantified proteins, # quantified peptides, and nonavailable (NA) values ratio^{-1} using three data sets. As the reciprocals of time, NA ratios, and output file sizes are taken, larger values of all five parameters indicate better performances. NA ratios were added with 0.01 to avoid taking the reciprocal of zero. For each variable, the values are normalized to range between 0.00 and 1.00 by dividing them by the maximum value. The six circles indicate, from the inside to the outside, the normalized values of 0.00, 0.20, 0.40, 0.60, 0.80, and 1.00.

Table 1. Description of the Three Public TMT-Based Proteomics Datasets Used in Our Work

TMT label	Species	Sample type	Disease	Instrument	# Batches	# Raw files/batch	Publication
6plex	Human	Cell culture	Acute myeloid leukemia	Orbitrap Fusion	1	10	Wang et al.[13]
10plex	Human	Liver	Hepatocellular carcinoma	Q Exactive HF	1	40	Gao et al.[14]
16plex	Human	Kidney	COVID-19	LTQ Orbitrap	4	30	Nie et al.[15]

looked at their software execution statistics and output results. For this purpose, we downloaded and analyzed three published MS/MS data sets, whose samples were labeled by TMT 6plex,[13] TMT 10plex,[14] and TMTpro 16plex,[15] using PD and FP with the same main parameters (see the Supporting Information ZIP for files). We first evaluated the correlation and differences between the two resulting protein expression matrices. Next, the batch effects and functional analyses of the output results for the two software were compared (Figure 1A). Our findings provide a reference for selecting and further improving protein quantification software.

MATERIALS AND METHODS

Data Sets

Three publicly available TMT-based proteomic data sets were used for the comparative evaluation of PD (version 3.0.0.757) and FP (version 18.0). The TMT 6plex, TMT 10plex, and TMTpro 16plex data sets were thus downloaded from the iProx database (https://www.iprox.cn/) with the identifiers IPX0002486000,[13] IPX0002067000,[14] and IPX0002393000,[15] respectively. Details are summarized in Table 1.

Protein Identification and Quantification

Databases. The human proteome database used by FP comprised 20,413 sequences and their corresponding decoys, including Swiss-Prot entries and common contaminant proteins listed in cRAP (https://www.thegpm.org/crap/). For PD, the database mentioned above was used after removing all the decoys because PD will generate the decoy version of the database. In addition, to identify the contaminants in the outputs, we used a FASTA contaminant database downloaded from cRAP.

PD Software Settings. Our PD protein identification used a processing workflow modified from the default version "PWF_Hybrid_TMT_MS2_CHIMERYS" or "PWF_Hybrid_TMTpro_MS2_CHIMERYS" and a consensus workflow modified from "CWF_Comprehensive_Enhanced Annotation_Reporter_Quan". Specifically, the provided quantification methods "TMTpro 16plex", "TMT 10plex", and "TMTe 6plex" were selected to quantify their corresponding data sets. Among the Spectrum Selector node parameters, mass analyzer, MS order, activation type, and polarity mode were set to FTMS, MS2, HCD, and positive, respectively. For the CHIMERYS Identification node, the parameter fragment mass tolerance was set to 0.02 Da, and the N-terminal loss of methionine and acetylation are inherent in the modifications. Regarding quantification, for TMT-labeled spectral data acquired from Orbitrap mass spectrometers, PD defaults to use the signal-to-noise values to calculate the abundances. In order to make the quantification more comparable between PD and FP, which calculates abundances from peak intensities, reporter abundance is set to be based on intensity in the Reporter Ions Quantifier

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

Table 2. Software Parameters for Database Searching

	PD	FP
Basis	Processing step: PWF_Hybrid_TMTpro_MS2_CHIMERYS for TMTpro 16plex data set, PWF_Hybrid_TMT_MS2_CHIMERYS for TMT 10plex and TMT 6plex data sets; Consensus step: CWF_Comprehensive_Enhanced Annotation_Reporter_Quan	TMT16 workflow for TMTpro 16plex data set, TMT10 workflow for TMT 10plex and TMT 6plex data sets
Fragment mass tolerance (Da)	0.02	
Peptide length	7−30	
Peptide mass range	350−5000	
Fixed modifications	TMT on lysine and peptide N-terminal (16plex: + 304.207146 Da; 10plex, 6plex: + 229.162932 Da), cysteine carbamidomethylation (+57.021464 Da)	
Variable modifications	methionine oxidation (+15.994915 Da)	methionine oxidation (+15.994915 Da), protein N-terminal acetylation (+42.010565 Da)
Database	20,413 human proteins containing 116 common contaminants	
Parallelism	Max. number of processing workflows in parallel execution = 8, max. number of consensus workflows in parallel execution = 8	Parallelism = 16
Input	.raw files	.mzML files
Other	Spectrum Selector: Mass Analyzer = Is FTMS, MS Order = Is MS2, Activation Type = Is HCD, Polarity Mode = Is + ; Reporter Ions Quantifier: Reporter Abundance Based On = Intensity	Database Splitting available, Command line option -minprob removed for ProteinProphet

node. No normalization or scaling was applied in this node. The maximum numbers of processing or consensus workflows in parallel execution in PD were both set to eight, which is half of the number of logical cores on our computer, the largest possible value for the two parameters. The final output was exported from the PD results after filtering the output to achieve a <1% false discovery rate (FDR).

FP Software Settings. The built-in workflows "TMT16" and "TMT10" were used with the adjustments next described. The mzML files were first generated from the original raw data using the ProteoWizard[16] MSConvert tool (version 3.0.21193-ccb3e0136) and then searched using MSFragger (version 3.5) with mass calibration and parameter optimization. To maximize the comparability between the two software, the search engine parameters, including fragment mass tolerance, peptide length, peptide mass range, and modifications, were given the same value as their equivalent parameters from the PD workflow. Notably, MSFragger defaults to the trimming of protein N-terminal methionine as a variable modification. MSFragger search results were processed using Percolator[17] (version 3.5) for peptide-spectrum match validation, followed by Philosopher[18] (version 4.4.0) for protein inference (using ProteinProphet[19]) and FDR filtering. For the TMTpro 16plex data set, TMTpro-126-labeled pooled samples were used as a reference for each batch within TMT-Integrator[20] (version 3.3.3), while virtual references were used for the TMT 10plex and 6plex data sets due to the lack of pooled sample. The quantification reports of the quantified proteins with FDR < 1% were generated without normalization. Parallelism was set to 16. The parameters used for the PD and FP workflows are summarized in Table 2.

Data Analysis of the Protein Quantification Results. Statistical analyses and data visualization were performed using R (version 4.2.1) and Python (3.8.8). For the FP outputs, the columns of protein abundances and abundance ratios were separately extracted from the report files abundance_protein_None.tsv and ratio_protein_None.tsv to generate abundance matrices and abundance ratio matrices in a log2 scale. For the PD outputs, only protein abundances were provided in the output files. To calculate the abundance ratios, we first replaced the nonavailable (NA) values (i.e., the missing values) in the reference column of the protein abundances with zeros and then divided the remaining 15 columns of the abundances by the reference column. In the abundance ratio matrices, infinite values were replaced with 100. The "Ratio + Median + Limma" workflow was applied to differential expression analysis as Huang et al.[21] described. First, the log2 abundance ratios from FP were transformed to the nonlogarithmic scale for consistency with the abundance ratios from PD. Then, the zeros and NAs in the abundance ratio matrices were imputed with 0.80 times the minimum nonzero ratio. Next, we performed a log2 transformation of the abundance ratios and a zero median normalization of the log2 ratios in each column. A principal variance component analysis (PVCA)[22] in BatchServer[23] (https://www.guomics.com/BatchServer/) and a uniform manifold approximation and projection (UMAP)[24] were then used for evaluating the batch effects, which were corrected using Combat[25] in the R package sva. The batch-corrected data underwent another zero median normalization, and their batch effects were re-evaluated with PVCA and UMAP. Subsequently, the proteins differentially expressed in COVID-19 and non-COVID-19 samples were identified using the R package limma.[26,27] ClusterProfiler[28] was used for the GO (Gene Ontology) and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analyses of the differentially expressed proteins (DEPs). The R packages VennDiagram[29] and ggplot2[30] were used for data visualization.

RESULTS AND DISCUSSION

Protein Identifications Obtained with PD or FP

Using three different mass spectrometry data sets with TMTpro 16plex, TMT 10plex, and TMT 6plex labeling, we assessed the protein quantification performances of PD and FP. For this purpose, we measured the following parameters: the total processing time of each software, the number of proteins and peptides quantified, the ratio of NAs in the protein quantification results, and the output file sizes (Table S1). FP requires the format conversion of the Thermo raw files to perform TMT-based proteomic quantifications, while PD can process the raw files. However, the FP processing time was shorter than that of PD, even including the time for file format conversion. Specifically, the lengths of time taken by PD were

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

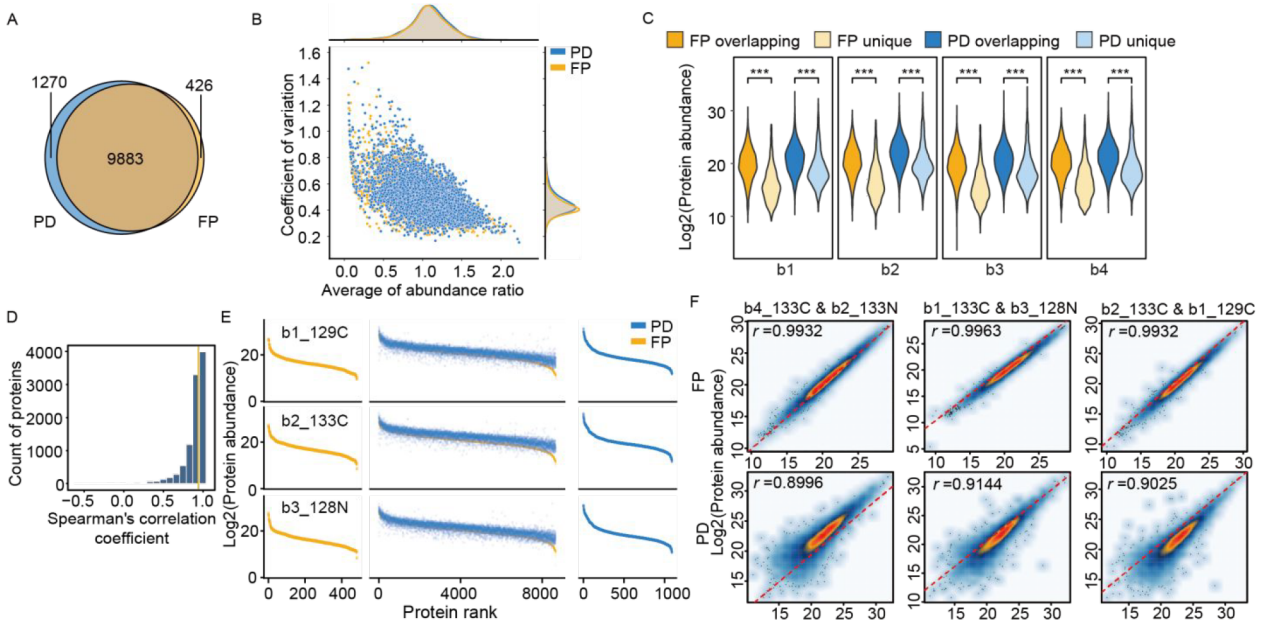


Figure 2. Comparison of the quantification results from the PD and FP analyses of the TMTpro 16plex data set. (A) Overlap between the proteins quantified using PD and FP. (B) Average values and coefficients of variation (CVs) of the abundance ratios of each protein across all samples. Each dot represents a protein. The abundance ratio outliers that are higher than 1.5 times the difference between the third quartile (Q3) and the first quartile (Q1) of the ratio matrix (i.e., the interquartile range, IQR) plus Q3 were given the value Q3 + 1.5 × IQR; the outliers lower than Q1−1.5 × IQR were given the value Q1−1.5 × IQR. The curves on the x- and *y*-axes are the distributions of the average values and the CVs, respectively. (C) The distribution of the log2 abundances of the overlapping and the uniquely quantified proteins show highly significant differences (Wilcoxon rank-sum test, *** $P < 0.001$). The labels b1, b2, b3, and b4 indicate the batches. (D) Histogram of the Spearman's correlation coefficients (SCCs) between the abundance ratios of the same protein quantified by PD and FP. The yellow vertical line represents the median of the SCCs. (E) Comparison of the log2 abundances of the same proteins quantified by PD (middle and right) and FP (left and middle). The proteins identified by FP were ranked by their log2 abundances in FP results (left and middle), while the proteins uniquely identified by PD were ranked by their log2 abundances in PD results (right). (F) The correlations of the log2 abundances of three pairs of technical replicates. The SCCs between the quantifications of each pair of technical replicates are indicated on the plots.

27.89, 29.83, and 16.46 times that of FP for the TMTpro 16plex, TMT 10plex, and TMT 6plex data sets, respectively. Furthermore, PD quantified higher protein numbers with similar NA ratios to FP. In particular, for the TMTpro 16plex, TMT 10plex, and TMT 6plex data set, PD vs. FP quantified the following protein numbers: 11,153 vs. 10,309, 8088 vs. 7006, and 6938 vs. 6306, respectively; the NA ratios were: 11.74% vs. 10.52%, 0.85% vs. 0.38%, and 0.00% vs. 0.00%, respectively. The numbers of quantified peptides were 19.69% and 13.70% larger using PD with the 16plex and 10plex data sets, while 3.50% larger using FP with the 6plex data set. In addition, the output file sizes of FP were only ∼5% of those of PD (4.81, 3.86, and 6.62%, for the 16plex, 10plex, and 6plex data sets, respectively). These results show that PD quantified more proteins, while FP provided time- and space-saving features when analyzing large data sets (Figure 1B).

Correlation and Differences between the Protein Quantification Results of PD and FP

Using the TMTpro 16plex data set, we evaluated the proteins quantified by both PD and FP and the consistency of their output expression profiles. Of all the quantified proteins, ∼ 90% were shared by PD and FP (88.61% and 95.87% of the proteins quantified using PD and FP, respectively), whereas 1270 and 426 proteins were uniquely quantified by PD and FP, respectively (Figure 2A). Moreover, for the abundance ratios of each protein quantified in different samples, the distribution of the average values and the coefficients of variation (CVs) were similar in the PD and the FP outputs (Figure 2B). To further investigate the differences between the overlapping and the uniquely quantified proteins, the abundance distributions were compared within each batch (Figure 2C) and each sample (Figure S1), evading the potential batch effects. The proteins uniquely quantified by either software were generally less abundant than those quantified by both.

Next, we studied the consistencies of the quantifications performed by PD and FP. To this aim, we calculated the Spearman's correlation coefficients (SCCs) of the abundance ratio vectors of each protein, the abundance ratio vectors of each sample, and the overall abundance ratio matrices measured with PD and FP (the uniquely quantified proteins were removed from the abundance ratio matrices before computing the SCCs). For the vectors of each protein, the median of the SCCs on the diagonal of the output correlation matrix was 0.9411 (Figure 2D). For the vectors of each sample, the median of the SCCs on the diagonal of the output correlation matrix was 0.8327 (Figure S2F). Moreover, the SCC between the two overall abundance ratio matrices was 0.9149. These results showed that the quantifications of the same proteins using PD and FP were highly correlated. The evaluation of the quantification output of the TMT 6plex and 10plex data sets generated similar results to the TMTpro 16plex data set ones (Figure S2A−E).

Using three randomly selected technical replicate samples from the TMTpro 16plex data set, the protein abundances generated by PD and FP showed similar distributions (Figure 2E). Examining the quantification results from the three pairs of technical replicates, the respective SCCs of the log2 abundance vectors were 0.9932, 0.9963, and 0.9932 with FP, which were significantly higher (Welch two-sample *t*-test, *P*-value = 0.0009)

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

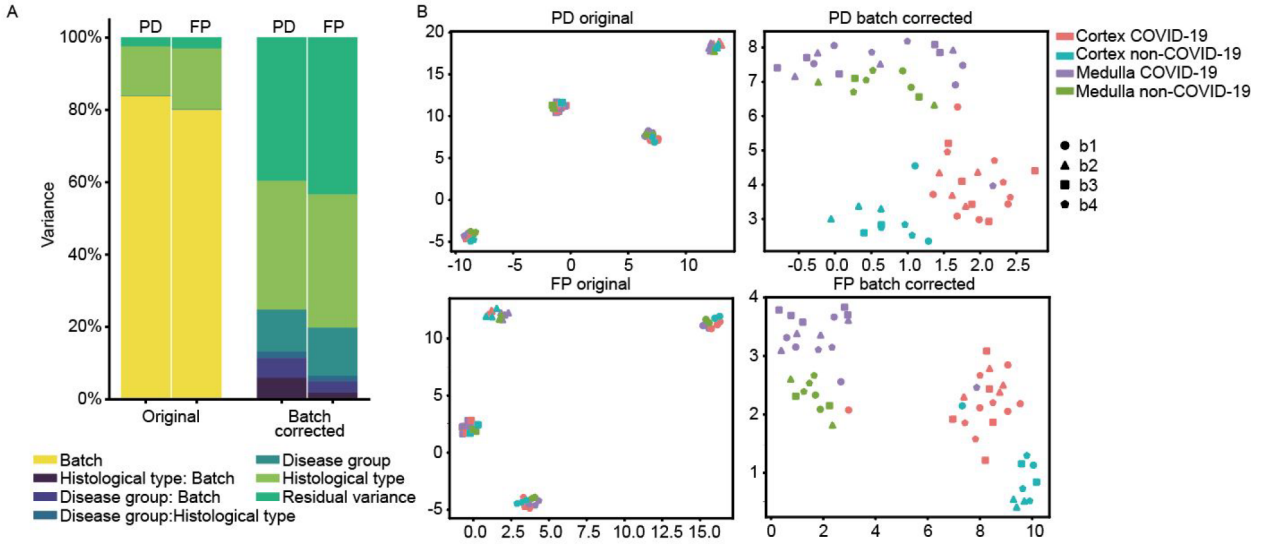


Figure 3. Batch effects in the data generated by PD and FP. (A) Contribution of different grouping variables and their interaction terms on the variance of the expression data. (B) The expression profiles of the samples are visualized using uniform manifold approximation and projection (UMAP) and labeled by disease group, histological type, and batch. The shape indicates the batch, while the color indicates the histological type and the disease group.

than those obtained using PD (0.8996, 0.9144, 0.9025) (Figure 2F). FP thus showed stronger consistency between technical replicates. These results showed that the proteomes quantified using PD and FP were highly overlapping and that their expression profiles are highly correlated and show good data quality.

Batch Effects in Protein Quantification by PD and FP

To further assess PD and FP, we analyzed the batch effects in their protein expression matrix outputs. As variations caused by batching may interfere with the data analysis, assessment and removal of batch effects are fundamental. The batch effects were thus evaluated, visualized, and corrected using PVCA, UMAP, and ComBat after preprocessing the abundance ratio matrices, as described in the Materials and Methods section. Our samples had three grouping variables: disease group (COVID-19 and non-COVID-19 (control) patients), histological type (renal cortex and renal medulla), and batch (b1, b2, b3, and b4). The PVCA analysis revealed the contribution of the three grouping variables and the pairwise interaction between the variables to the variance of the matrices. In addition, UMAP showed the impact of each grouping variable on the expression data. Batching predominantly resulted in the variance of the expression matrices, although its effects were lighter in the output of FP (Figure 3A). After we removed the batch effects, the samples from the different batches were evenly distributed for both software (Figure 3B). Indeed, the corrected expression data from PD and FP were minimally affected by batch (0.00%, 0.00%) and the interaction of batch with the disease group (5.39%, 3.09%) or the histological type (6.00%, 1.87%) according to PVCA analysis. By eliminating this confounding effect, sample discrimination became reasonable, and thus, the data is suitable for downstream analyses. Notably, the cortex and medulla samples formed two relatively isolated clusters, highlighting the differences in protein expression between the two kidney histological types. PD and FP quantification results thus have comparable batch effects before and after correction.

Differentially Expressed Proteins and Enriched Functional Terms in the PD and FP Outputs

To further elucidate the similarities and differences between the two software, a functional analysis was performed on the protein expression matrices resulting from PD and FP, using the same TMTpro 16plex data set based on the COVID-19 and the non-COVID-19 samples. Specifically, a differential expression analysis was performed using the batch-corrected data. The cortex and medulla samples were analyzed separately since the histological type was the main contributing factor to the data variance. Similar numbers of DEPs, using a fold-change threshold of 1.50 (adjusted *P*-value <0.05), were identified by both software in the cortex (411 by PD and 372 by FP) and the medulla (474 by PD and 402 by FP) (Figure 4A). Specifically, FP identified 149 up-regulated and 223 down-regulated DEPs in the cortex, and 222 up-regulated and 180 down-regulated DEPs in the medulla; PD identified 137 up-regulated and 274 down-regulated DEPs in the cortex, and 249 up-regulated and 225 down-regulated DEPs in the medulla. Among them, 534 DEPs were identified by both PD and FP (275 in the cortex and 259 in the medulla), and their regulatory directions were consistent in both PD and FP outputs (Figure 4B).

To investigate whether the identified DEPs of PD and FP had similar functional profiles, enrichment analyses of GO biological processes (BPs), molecular functions (MFs), cellular components (CCs), and KEGG pathways were conducted. For the cortex samples, the DEPs identified from the PD output enriched more GO terms and KEGG pathways than the DEPs from the FP output. The enriched terms from the DEPs of PD and FP had a partial overlap of 99 (46.05%) GO BPs, 46 (58.97%) GO MFs, 23 (88.46%) GO CCs, and 17 (50.00%) KEGG pathways (Figure 4C, S3A). The ten most significant BPs in either PD or FP were contained in the 99 shared BPs, and nine of the ten most significant KEGG pathways in PD and all the ten most significant KEGG pathways in FP were contained in the 17 shared pathways (Figure 4D). Similarly, the 46 shared MFs included all the ten most significant MFs in PD and in FP, and the 23 shared CCs covered the ten most significant CCs in PD

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

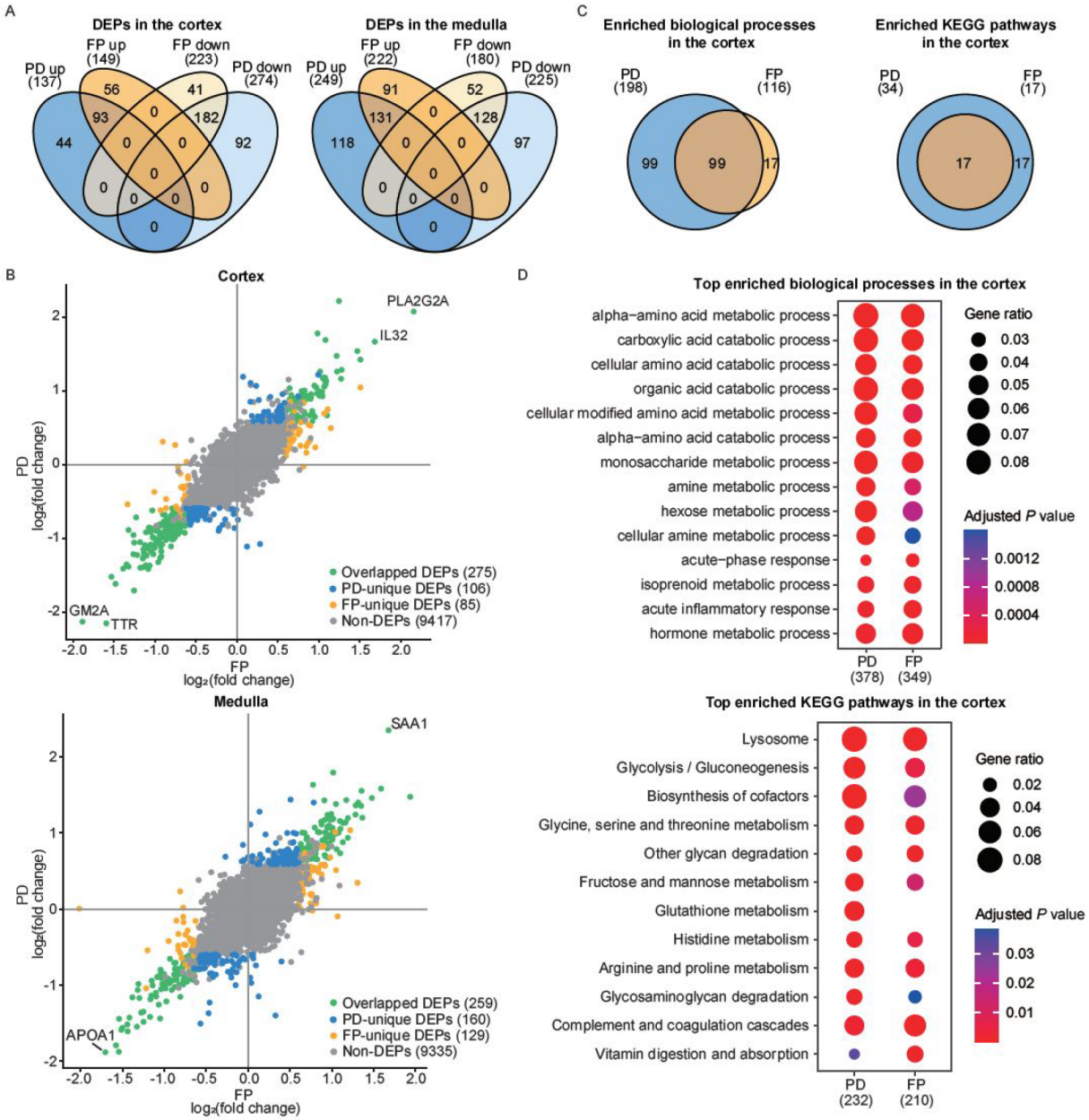


Figure 4. Differentially expressed proteins (DEPs) and functional enrichment in the PD and FP quantification results. (A) Overlap between the DEPs identified by PD and FP. Proteins with fold change (FC) > 1.50 and Benjamini-Hochberg adjusted *P*-value <0.05 were classified as differentially expressed. (B) Comparison between the DEPs calculated using PD and FP. Only proteins identified by both software were shown. The gene names of the proteins with FC > 3.00 in both the PD and the FP quantification results are labeled in the plots. (C) Gene Ontology (GO) biological processes (BPs) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched by PD and FP in the cortex DEPs. (D) The ten most significantly enriched GO BPs and KEGG pathways in the cortex DEPs identified by PD and FP. If a GO BP or KEGG pathway was enriched by both software but was among the ten most significantly enriched categories of one software only, it is here marked with a dot for the other software as well. The numbers within parentheses indicate the total numbers of DEPs provided for the enrichment analyses that have GO BP or KEGG pathway annotations.

and in FP (Figure S3B). On the other hand, fewer GO terms and KEGG pathways were enriched in the medulla samples. However, the agreement between the PD and FP enrichment results was similar to that for the cortex (Figure S3C,D). Differential expression and functional enrichment analyses of the PD and the FP quantification outputs produced different results, which imply the non-negligible impact of software on data-driven biological discoveries. Nonetheless, the most significantly enriched terms were relatively robust in both the PD and the FP outputs.

CONCLUSIONS

Our analysis of three TMT-labeled proteomic data sets revealed that PD has an advantage in the number of protein identification, FP stands out for generating similar quantification results while requiring shorter computational time, which is particularly important for large-scale studies. Also, unlike PD, FP is freely available for academic, educational, or other noncommercial uses. However, as a well-maintained commercial software suite, PD is robust and integrates various tools providing powerful functions. PD and FP also support the analysis of data sets from many other proteomic workflows that were not evaluated in this

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

study, such as label-free quantification. Additionally, our study may not be comprehensive due to the limited number of analyzed data sets. The present conclusion is based on the current version of the software. Conclusions may change if there are major updates of the software to the algorithm. Despite these limitations, our comparative analysis provides clear information for choosing the TMT-proteomic data analysis software that best suits the needs of a specific group or study.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00390.

TMT-based protein quantification statistics using PD and FP (Table S1); abundances of the overlapping or uniquely quantified proteins in each sample (Figure S1); comparison of the quantification results from the PD and FP analyses using the TMT 6plex, 10plex, and 16plex data sets (Figure S2); functional enrichments using the output DEPs from PD and FP (Figure S3) (PDF)

Analysis code, PD and FP output matrices, list of contaminants, sample information for TMTpro 16plex data set (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Alexey I. Nesvizhskii − *Department of Pathology; Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, United States*; Email: nesvi@med.umich.edu

Tiannan Guo − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China;* orcid.org/0000-0003-3869-7651; Email: guotiannan@westlake.edu.cn

Yaoting Sun − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China;* orcid.org/0000-0001-7613-648X; Email: sunyaoting@westlake.edu.cn

Authors

Tianen He − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China; School of Life Sciences, Peking University, Beijing 100871, China;* orcid.org/0000-0001-6864-0723

Youqi Liu − *Westlake Omics (Hangzhou) Biotechnology Co., Ltd., Hangzhou 310024, China*

Yan Zhou − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China*

Lu Li − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China*

He Wang − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China*

Shanjun Chen − *Westlake Omics (Hangzhou) Biotechnology Co., Ltd., Hangzhou 310024, China*

Jinlong Gao − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China*

Wenhao Jiang − *Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310024, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China; Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China*

Yi Yu − *Westlake Omics (Hangzhou) Biotechnology Co., Ltd., Hangzhou 310024, China*

Weigang Ge − *Westlake Omics (Hangzhou) Biotechnology Co., Ltd., Hangzhou 310024, China*

Hui-Yin Chang − *Department of Pathology; Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, United States; Department of Biomedical Sciences and Engineering, National Central University, Taoyuan City 320317, Taiwan;* orcid.org/0000-0003-1767-1874

Ziquan Fan − *Thermo Fisher Scientific, Shanghai 201203, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00390

Author Contributions

#T.H. and Y.L. are co-first authors. Y.S., T.G., A.I.N., and T.H. designed the study. T.H. and Y.L. conducted software execution and data analysis. T.H. and Y.S. wrote the paper with inputs from all other coauthors. Y.L., Y.Z., L.L., H.W., S.C., J.G., W.J., Y.Y., W.G., and H.C. helped to improve the paper. Z.F. helped with PD execution. T.G., A.i.N. and Y.S. supervised the project.

Notes

The authors declare the following competing financial interest(s): T.G. is a shareholder of Westlake Omics, Inc. Y.L., S.C., Y.Y., and W.G. are employees of Westlake Omics, Inc. Z.F. is an employee of Thermo Fisher Scientific.

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

■ ACKNOWLEDGMENTS

We thank for the assistance from the Thermo Fisher Scientific team (Ziquan Fan and Jie Cui) and CHIMERYS platform who provided us with free access to PD 3.0 and CHIMERYS. Proteome Discoverer is a trade-marked property of Thermo Fisher Scientific. CHIMERYS is a trademark of MSAID GmbH. This work is supported by grants to T.G. from the National Key R&D Program of China (No. 2021YFA1301603).

■ ABBREVIATIONS

TMT, tandem mass tag; FDR, false discovery rate; FP, FragPipe; PD, Proteome Discoverer; cRAP, common Repository of Adventitious Proteins; DEP, differentially expressed protein; GO, Gene Ontology; BP, biological process; MF, molecular function; CC, cellular component; KEGG, Kyoto Encyclopedia of Genes and Genomes; PVCA, principal variance component analysis; UMAP, uniform manifold approximation and projection; FC, fold change; SCC, Spearman's correlation coefficient; CV, coefficient of variation; Q1, the first quartile; Q3, the third quartile; IQR, interquartile range

■ REFERENCES

(1) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895−1904.

(2) Li, J. M.; Cai, Z. Y.; Bomgarden, R. D.; Pike, I.; Kuhn, K.; Rogers, J. C.; Roberts, T. M.; Gygi, S. P.; Paulo, J. A. TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J. Proteome Res.* **2021**, *20* (5), 2964−2972.

(3) Rauniyar, N.; Yates, J. R. Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *J. Proteome Res.* **2014**, *13* (12), 5293−5309.

(4) Casado-Vela, J.; Martinez-Esteso, M. J.; Rodriguez, E.; Borras, E.; Elortza, F.; Bru-Martinez, R. iTRAQ-based quantitative analysis of protein mixtures with large fold change and dynamic range. *Proteomics* **2010**, *10* (2), 343−347.

(5) Goel, R.; Murthy, K. R.; Srikanth, S. M.; Pinto, S. M.; Bhattacharjee, M.; Kelkar, D. S.; Madugundu, A. K.; Dey, G.; Mohan, S. S.; Krishna, V.; Prasad, T. S. K.; Chakravarti, S.; Harsha, H. C.; Pandey, A. Characterizing the normal proteome of human ciliary body. *Clin. Proteomics* **2013**, *10* (1), 9.

(6) Orsburn, B. C. Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **2021**, *9* (1), 15.

(7) Frejno, M.; Zolg, D. P.; Schmidt, T.; Gessulat, S.; Graber, M.; Seefried, F.; Rathke-Kuhnert, M.; Fredj, S. B.; Premnadh, S.; Samaras, P.; Fritzemeier, K.; Berg, F.; Nasir, W.; Horn, D.; Delanghe, B.; Henrich, C.; Kuster, B.; Wilhelm, M. CHIMERYS: An AI-Driven Leap Forward in Peptide Identification. In *the 69th ASMS Conference on Mass Spectrometry and Allied Topics*, Philadelphia, USA, October 31−November 4, 2021.

(8) Chang, C.; Zhang, J.; Han, M.; Ma, J.; Zhang, W.; Wu, S.; Liu, K.; Xie, H.; He, F.; Zhu, Y. SILVER: an efficient tool for stable isotope labeling LC-MS data quantitative analysis with quality control methods. *Bioinformatics* **2014**, *30* (4), 586−587.

(9) Li, Y.; Chi, H.; Wang, L. H.; Wang, H. P.; Fu, Y.; Yuan, Z. F.; Li, S. J.; Liu, Y. S.; Sun, R. X.; Zeng, R.; He, S. M. Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing. *Rapid Commun. Mass Spectrom.* **2010**, *24* (6), 807−814.

(10) Li, J.; Van Vranken, J. G.; Pontano Vaites, L.; Schweppe, D. K.; Huttlin, E. L.; Etienne, C.; Nandhikonda, P.; Viner, R.; Robitaille, A. M.; Thompson, A. H.; Kuhn, K.; Pike, I.; Bomgarden, R. D.; Rogers, J. C.; Gygi, S. P.; Paulo, J. A. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **2020**, *17* (4), 399−404.

(11) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14* (5), 513−520.

(12) Teo, G. C.; Polasky, D. A.; Yu, F.; Nesvizhskii, A. I. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *J. Proteome Res.* **2021**, *20* (1), 498−505.

(13) Wang, L.; Shao, X.; Zhong, T.; Wu, Y.; Xu, A.; Sun, X.; Gao, H.; Liu, Y.; Lan, T.; Tong, Y.; Tao, X.; Du, W.; Wang, W.; Chen, Y.; Li, T.; Meng, X.; Deng, H.; Yang, B.; He, Q.; Ying, M.; Rao, Y. Discovery of a first-in-class CDK2 selective degrader for AML differentiation therapy. *Nat. Chem. Biol.* **2021**, *17* (5), 567−575.

(14) Gao, H.; Zhang, F.; Liang, S.; Zhang, Q.; Lyu, M.; Qian, L.; Liu, W.; Ge, W.; Chen, C.; Yi, X.; Zhu, J.; Lu, C.; Sun, P.; Liu, K.; Zhu, Y.; Guo, T. Accelerated Lysis and Proteolytic Digestion of Biopsy-Level Fresh-Frozen and FFPE Tissue Samples Using Pressure Cycling Technology. *J. Proteome Res.* **2020**, *19* (5), 1982−1990.

(15) Nie, X.; Qian, L.; Sun, R.; Huang, B.; Dong, X.; Xiao, Q.; Zhang, Q.; Lu, T.; Yue, L.; Chen, S.; Li, X.; Sun, Y.; Li, L.; Xu, L.; Li, Y.; Yang, M.; Xue, Z.; Liang, S.; Ding, X.; Yuan, C.; Peng, L.; Liu, W.; Yi, X.; Lyu, M.; Xiao, G.; Xu, X.; Ge, W.; He, J.; Fan, J.; Wu, J.; Luo, M.; Chang, X.; Pan, H.; Cai, X.; Zhou, J.; Yu, J.; Gao, H.; Xie, M.; Wang, S.; Ruan, G.; Chen, H.; Su, H.; Mei, H.; Luo, D.; Zhao, D.; Xu, F.; Li, Y.; Zhu, Y.; Xia, J.; Hu, Y.; Guo, T. Multi-organ proteomic landscape of COVID-19 autopsies. *Cell* **2021**, *184* (3), 775−791.e14.

(16) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24* (21), 2534−2536.

(17) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923−925.

(18) da Veiga Leprevost, F.; Haynes, S. E.; Avtonomov, D. M.; Chang, H.-Y.; Shanmugam, A. K.; Mellacheruvu, D.; Kong, A. T.; Nesvizhskii, A. I. Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **2020**, *17* (9), 869−870.

(19) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646−4658.

(20) Djomehri, S. I.; Gonzalez, M. E.; da Veiga Leprevost, F.; Tekula, S. R.; Chang, H. Y.; White, M. J.; Cimino-Mathews, A.; Burman, B.; Basrur, V.; Argani, P.; Nesvizhskii, A. I.; Kleer, C. G. Quantitative proteomic landscape of metaplastic breast carcinoma pathological subtypes and their relationship to triple-negative tumors. *Nat. Commun.* **2020**, *11* (1), 1723.

(21) Huang, T.; Choi, M.; Tzouros, M.; Golling, S.; Pandya, N. J.; Banfai, B.; Dunkley, T.; Vitek, O. MSstatsTMT: Statistical Detection of Differentially Abundant Proteins in Experiments with Isobaric Labeling and Multiple Mixtures. *Mol. Cell. Proteomics* **2020**, *19* (10), 1706−1723.

(22) Boedigheimer, M. J.; Wolfinger, R. D.; Bass, M. B.; Bushel, P. R.; Chou, J. W.; Cooper, M.; Corton, J. C.; Fostel, J.; Hester, S.; Lee, J. S.; Liu, F. L.; Liu, J.; Qian, H. R.; Quackenbush, J.; Pettit, S.; Thompson, K. L. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics* **2008**, *9*, 285.

(23) Zhu, T.; Sun, R.; Zhang, F.; Chen, G. B.; Yi, X.; Ruan, G.; Yuan, C.; Zhou, S.; Guo, T. BatchServer: A Web Server for Batch Effect Evaluation, Visualization, and Correction. *J. Proteome Res.* **2021**, *20* (1), 1079−1086.

(24) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C. A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **2019**, *37* (1), 38−44.

(25) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8* (1), 118−127.

(26) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43* (7), e47.

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015

(27) Phipson, B.; Lee, S.; Majewski, I. J.; Alexander, W. S.; Smyth, G. K. Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression. *Ann. Appl. Stat.* **2016**, *10* (2), 946−963.
(28) Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; Fu, X.; Liu, S.; Bo, X.; Yu, G. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2021**, *2* (3), 100141.
(29) Chen, H.; Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **2011**, *12*, 35.
(30) Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed.; Springer-Verlag: New York, 2016; p 260.

https://doi.org/10.1021/acs.jproteome.2c00390
J. Proteome Res. 2022, 21, 3007−3015