Journal of proteome research

pubs.acs.org/jpr

Article

Comparison of Database Searching Programs for the Analysis of Single-Cell Proteomics Data

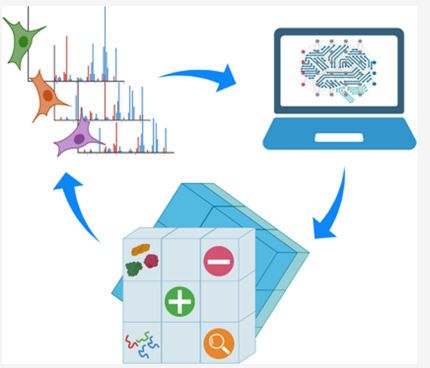
Jiaxi Peng, Calvin Chan, Fei Meng, Yechen Hu, Lingfan Chen, Ge Lin, Shen Zhang,* and Aaron R. Wheeler*

Cite This: *J. Proteome Res.* 2023, 22, 1298−1308

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Single-cell proteomics is emerging as an important subfield in the proteomics and mass spectrometry communities, with potential to reshape our understanding of cell development, cell differentiation, disease diagnosis, and the development of new therapies. Compared with significant advancements in the "hardware" that is used in single-cell proteomics, there has been little work comparing the effects of using different "software" packages to analyze single-cell proteomics datasets. To this end, seven popular proteomics programs were compared here, applying them to search three single-cell proteomics datasets generated by three different platforms. The results suggest that MSGF+, MSFragger, and Proteome Discoverer are generally more efficient in maximizing protein identifications, that MaxQuant is better suited for the identification of low-abundance proteins, that MSFragger is superior in elucidating peptide modifications, and that Mascot and X!Tandem are better for analyzing long peptides. Furthermore, an experiment with different loading amounts was carried out to investigate changes in identification results and to explore areas in which single-cell proteomics data analysis may be improved in the future. We propose that this comparative study may provide insight for experts and beginners alike operating in the emerging subfield of single-cell proteomics.



KEYWORDS: *single-cell proteomics, database searching comparison, data analysis, protein identification, peptide modification, protein abundance distribution*

INTRODUCTION

The cell is the smallest structural unit of living organisms. Although cells work together, intercellular heterogeneity (even among cells of the same type) exists because of complex genetic and environmental factors.[1−5] Therefore, the capacity to evaluate single cells has long been thought to have the potential to revolutionize our understanding of cell development, cell differentiation, disease diagnosis, and therapy.[6−9]

Unlike single-cell genomics and transcriptomics, single-cell proteomics has proven to be a substantial technical challenge because of the intrinsic lack of amplification tools, the wide variety of analyte physicochemistries, the low abundance of particularly important analytes, and the exceptionally wide dynamic range of the system.[10−12] With recent innovations in mass spectrometry (MS) and ion mobility, especially in combination with ultrasensitive sample preparation technologies, it is now possible to identify hundreds to over one thousand proteins from a single cell by MS analysis.[13−17] Additionally, signal matching or boosting strategies such as match between run (MBR) or isobaric tandem-mass-tag (TMT) carrier channels show promise for future gains in the number of proteins that can be identified in single cells.[1,18−25]

Despite the excitement for single-cell proteomics and the advances in data generation techniques, data analysis for this application remains relatively unexplored. Considering that current proteomics database searching programs were developed based on spectra from bulk samples, it is important to investigate their performance when applied to single-cell data. For example, a recent study[26] reported that single-cell peptide fragmentation spectra have fewer annotated fragment ions, reduced signal-to-background ratios, and reduced spectral consistency compared to spectra from bulk samples. Furthermore, feature detection in spectra generated from single cells is particularly hindered by background interference.[14,27] In sum, optimizing data analysis of the single-cell proteome is particularly important,[28] which motivated the work here.

In this work, seven proteome database searching programs were evaluated: Comet,[29] Mascot,[30] X!Tandem,[31,32] MSGF+,[33,34] MSFragger,[35] Proteome Discoverer,[36] and MaxQuant.[37] Of this group, Comet, Mascot, X!Tandem, Proteome

Received: December 14, 2022
Published: March 9, 2023

Discoverer, and MaxQuant are well established, having been used (and improved) by the proteomics community for many years. Notably, recent Proteome Discoverer updates have enabled the analysis of high-field asymmetric waveform ion mobility spectrometry (FAIMS)[38−40] data, a technique that is emerging as being particularly powerful for mass-limited samples like single cells. On the other hand, MSGF+ and MSFragger are relative newcomers but are growing in popularity as a result of their unique features. For example, MSGF+ is uniquely designed to deliver universal protein identification from diverse types of spectra from different MS instruments, and MSFragger focuses on identifying peptide modifications. A particularly useful tool for the work reported here is ProHits,[41] an open-source service that supports analyses by Comet, Mascot, X!Tandem, MSFG+, and MSFragger. In particular, ProHits allows standardized data cutoffs via the Trans-Proteomics Pipeline[42] (TPP), which allowed us to set the data analysis parameters and statistics cutoffs for these five algorithms to be as similar as possible.

Each of the algorithms evaluated here was used to evaluate three sets of published single-cell proteomics data, DISCO,[43] autoPOTS,[13] and nanoPOTS,[14] which were generated from two different cell lines using three different sample processing techniques and three different Thermo Fisher Scientific Orbitrap mass spectrometers (Q Exactive HFX, Orbitrap Exploris 480, and Orbitrap Eclipse). Among other differences, the DISCO[43] and autoPOTS[13] datasets were collected using standard HPLC-MS/MS, while the nanoPOTS[14] dataset was collected using a FAIMS Pro Interface (San Jose, CA) between the chromatography system and the mass spectrometer. The results of each program's analysis of each dataset, including identifications, peptide modifications, protein abundance distributions, and distributions of peptide properties, were compared as a step toward identifying strengths and weaknesses among the algorithms. Additionally, in new work, cell lysate samples with total protein amounts varying from 250 pg (the amount expected in a single cell) to 10 ng were evaluated by MaxQuant to investigate changes in identification results and to explore areas in which single-cell proteomics data analysis may be improved. We propose that this report will be useful for researchers who are considering which database searching program to apply to their single-cell proteomics data, as well as providing valuable insights for those who develop the next generation of protein identification algorithms for single-cell proteomics analysis.

EXPERIMENTAL SECTION

Datasets

Proteomics data generated from nine mammalian cells were collected from the DISCO[43] (3 U87 cells), autoPOTS[13] (3 HeLa cells), and nanoPOTS[14] (3 HeLa cells) studies. These data are freely available from the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with identifiers PXD019958, PXD021882, and PXD019515, respectively.

Database Searching on the ProHits Platform

RAW data files were converted to mzML using ProteoWizard (3.0.4468) and analyzed through the TPP[42] via the iProphet[44] pipeline implemented within ProHits[41] as follows. The database consisted of the HEK293 sequences in the RefSeq protein database (version 57) supplemented with "common contaminants" from the Max Planck Institute (http://lotus1.gwdg.de/mpg/mmbc/MaxQuant_input.nsf/7994124a4298328fc125748d0048fee2/$FILE/contaminants.fasta) and the Global Proteome Machine (GPM; https://www.thegpm.org/crap/). The search database consisted of forward and reverse sequences (labeled "gil9999" or "DECOY"); in total, 72,226 entries (including decoys) were searched. Spectra were analyzed separately using Mascot (2.3.02; Matrix Science), Comet (2018.01 rev.4), X!Tandem (2013.06.15), MSGF+ (v2019.02.28), and MSFragger (3.2) for trypsin specificity with up to two missed cleavages; deamidation (NQ), oxidation (M), and protein N-terminal acetylation as variable modifications; single-, double-, and triple-charged ions allowed mass tolerance of the parent ion to 12 ppm; and the fragment bin tolerance at 0.6 amu. The search results from the different programs were analyzed and combined through the TPP (v.4.7 POLAR VORTEX rev 1) via the iProphet pipeline.[42,44] All proteins with a minimal iProphet protein probability of 0.05 were parsed to the relational module of ProHits. This process was applied for each of the algorithms individually, and also for all five together in a combinatorial approach. For a fair and consistent comparison, only proteins with iProphet protein probability $\geq$0.95 were reported, corresponding to an estimated protein level false-discovery rate (FDR) of approximately 1%.

Proteome Discoverer Database Searching

Proteome Discoverer 2.4 software (Thermo Fisher Scientific) was used for analyses of RAW files. Searching was performed using the SEQUEST HT database search engine and the same database indicated (above) for ProHits. Trypsin was selected as the digestion enzyme, and a maximum number of two missed cleavages was allowed. The mass tolerance of precursor ions and fragment ions was set as 10 ppm and 0.08 Da, respectively. Carbamidomethylation (cysteine) was set as a static modification, and acetylation (protein N-terminus), oxidation (methionine), and deamination (asparagine or glutamine) were selected as dynamic modifications. The Percolator tool integrated in the Proteome Discoverer 2.4 software was used to validate the database search results based on the q-value. The identifications were filtered with a peptide confidence value as high as possible to obtain FDR less than 1% on the peptide level. Protein grouping was enabled, and the strict parsimony principle was applied.

MaxQuant Database Searching

MaxQuant (Version 1.6.4.0) software was used to search RAW files, using the same protein database indicated above. Trypsin was selected as the digestion enzyme, and a maximum number of two missed cleavages was allowed. The minimum peptide length was set as six amino acids, and the maximum peptide mass was 4600 Da. Methionine oxidation, N-terminal protein acetylation, and deamination of asparagine or glutamine were set as variable modifications, while cysteine carbamidomethylation was set as a fixed modification. Both peptides and proteins were filtered with a maximum FDR of 0.01. The default settings of MaxQuant were used for all other parameters not mentioned.

Analysis of HeLa Protein Digests

A HeLa protein digest standard (Pierce, Thermo Fisher Scientific) was diluted in 0.1% formic acid in water (v/v) to contain 250, 500, 1000, or 10,000 pg in 1 μL aliquots. Each aliquot was loaded onto an EASY-nLC 1200 system coupled with a QE HFX mass spectrometer (Thermo Fisher Scientific).

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308

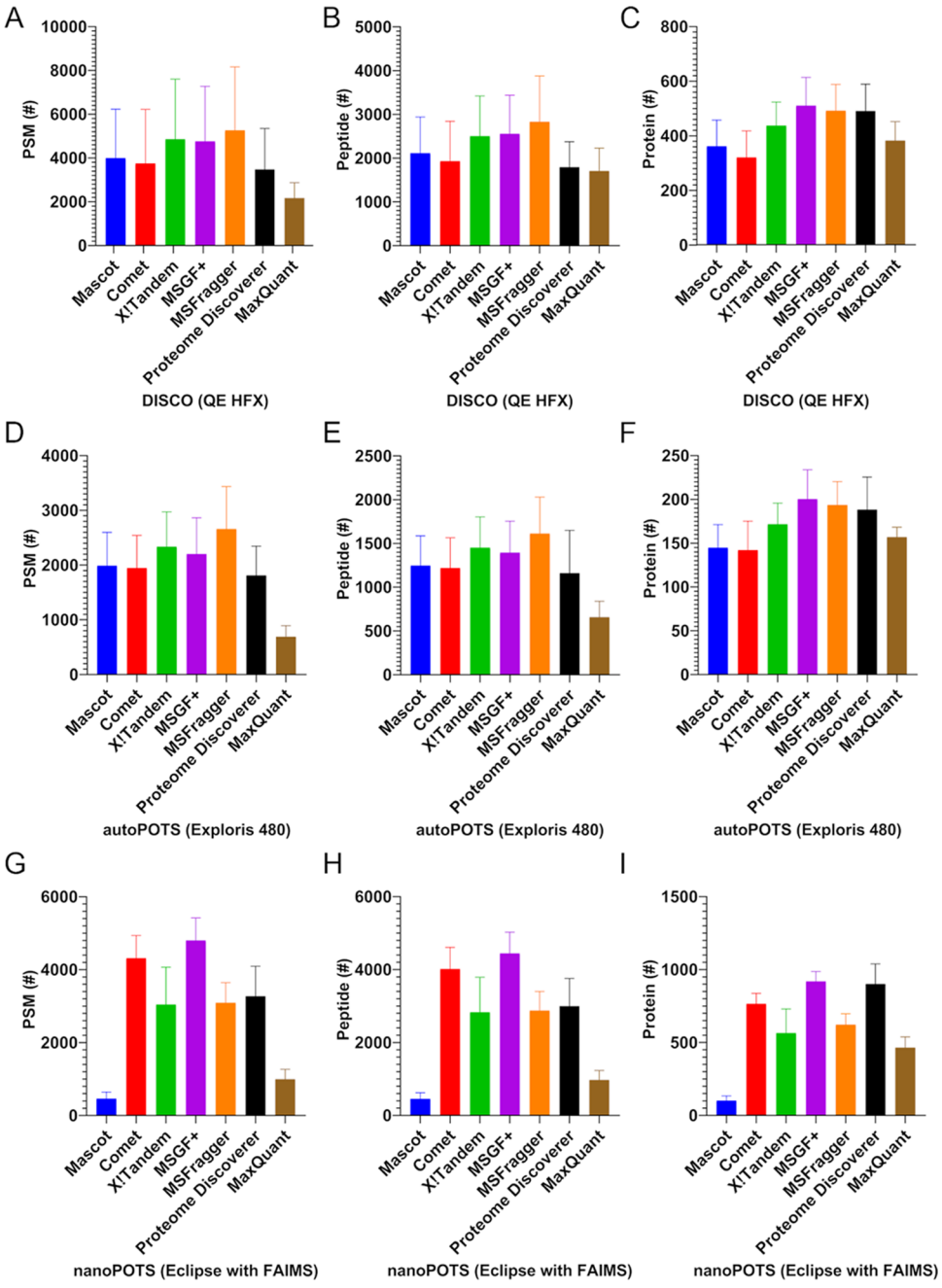


Figure 1. Plots of PSMs, peptide, and protein identifications from datasets reported for U87 cells evaluated by DISCO[43] (A–C), HeLa cells evaluated by autoPOTS[13] (D–F), and HeLa cells evaluated by nanoPOTS[14] (G–I) by Mascot (blue), Comet (red), X!Tandem (green), MSGF+ (purple), MSFragger (orange), Proteome Discoverer (black), and MaxQuant (brown). The error bars represent standard deviations from the analysis of three different cells in each study (n = 3).

The sample was separated in a fused silica microcapillary column (12 cm, 100 μm i.d., Polymicro Technologies) packed with 1.9 μm diameter reversed phase C18 beads (ReproSil-Pur 120 Å, Dr. Maisch GmbH). Mobile phase A was 0.1% (v/v) formic acid in water, mobile phase B was 80/19.9/0.1% (v/v/v) ACN/water/formic acid, and the two were used to generate a linear gradient of 3–30% B for 90 min, 30–45% B for 20 min, 45–95% B for 1 min, and 95% B for 14 min. The flow rate was set as 300 nL/min. The full mass scan range was set from m/z 375 to 1575 at a resolution of 120,000, while the automatic gain control (AGC) target was 5×10^5 and maximum injection time was set as 50 ms. Precursor ions with charges of +2 to +6 were fragmented using high energy collision with 27% normalized energy at a resolution of 60,000, AGC of 5×10^4, and maximum injection time of 250 ms. Previously selected precursor ions were excluded from further identification for 20 s. RAW data were searched by MaxQuant (version 1.6.4.0) as described above.

RESULTS AND DISCUSSION

Single-Cell Proteomics Data Analysis Programs

The central aim of this work was to evaluate the differences between single-cell protein identification programs. This is a challenging task, given that the types of cells, the nature of the sample pretreatment, and the model of mass spectrometer used

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298–1308

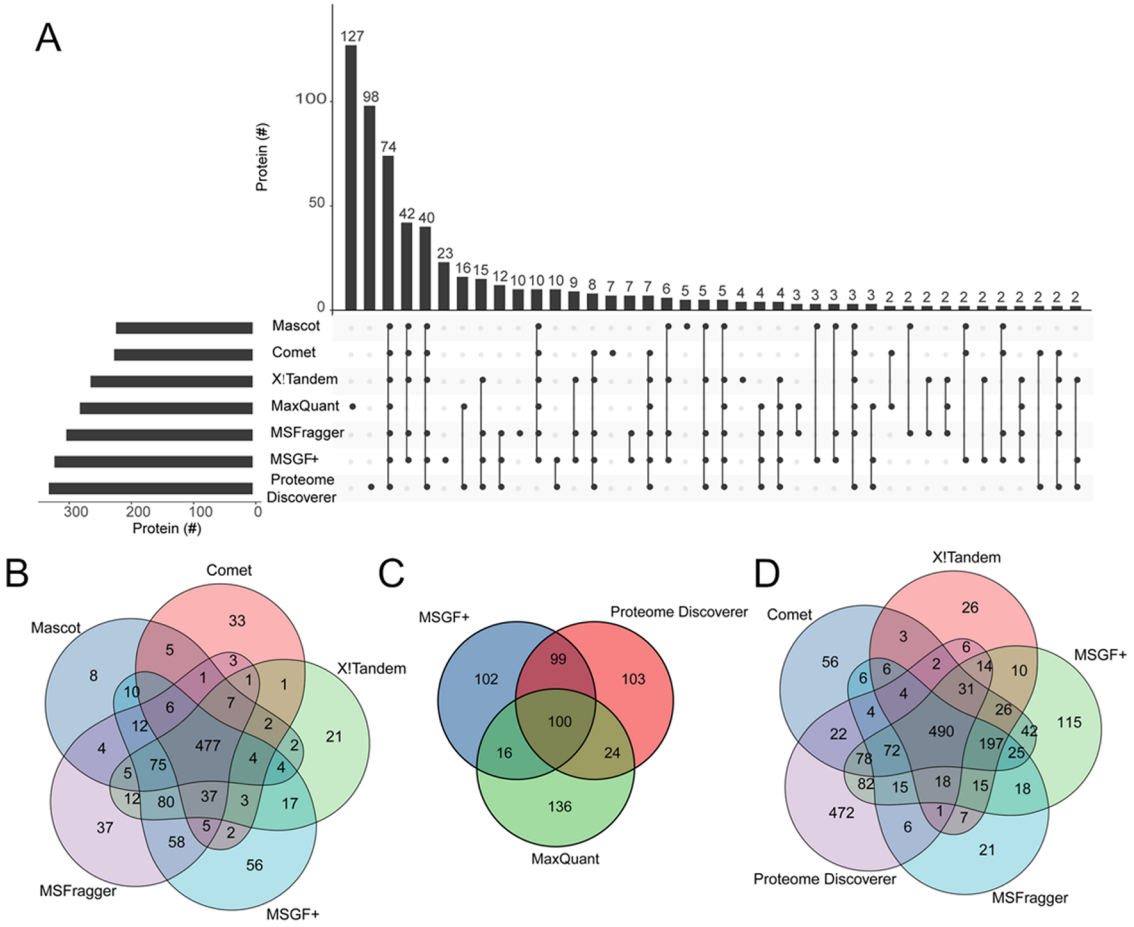


Figure 2. UpSet chart illustrating the overlap of proteins identified in HeLa cells in the autoPOTS dataset by Mascot, Comet, X!Tandem, MaxQuant, MSFragger, MSGF+, and Proteome Discoverer (A). The total numbers of proteins identified by each program are plotted in the bar chart to the left. The numbers of proteins identified in one or more programs are plotted in the bar chart above (each representing a subset of the total list of proteins identified). The markers and lines in the table indicate the source(s) of the proteins in each subset. Venn diagrams of the proteins identified in the autoPOTS dataset by Mascot (dark blue), Comet (red), X!Tandem (green), MSGF+ (light blue), and MSFragger (pink) (B) and by MSGF+ (blue), Proteome Discoverer (red), and MaxQuant (green) (C). Venn diagram of the proteins identified in HeLa cells in the nanoPOTS (with FAIMS) dataset by Comet (dark blue), X!Tandem (red), MSGF+ (green), MSFragger (light blue), and Proteome Discoverer (pink) (D).

have great influence on peptide and protein identification in proteomic studies, especially when the study is scaled down to tiny samples or even single cells. To try to accommodate these differences and generate generalizable results, we used data generated from three different single-cell sample processing platforms, DISCO,[43] autoPOTS,[13] and nanoPOTS,[14] applied to two different cell types evaluated on three different mass spectrometers to compare the performance of seven popular database searching programs for single-cell proteomics analysis.

As the most direct assessment criteria of each program, we first tested the numbers of peptide spectral matches (PSMs), peptides identified, and proteins identified for the three datasets (Figure 1). Although there are some variations, each program's performance is consistent for DISCO[43] (Figure 1A–C) and autoPOTS[13] (Figure 1D–F), with MSGF+ identifying the most proteins and MSFragger identifying the greatest number of peptides and PSMs. Using the DISCO[43] results as an example, MSGF+ identified 4% more proteins than both MSFragger and Proteome Discoverer and 17, 33, 41, and 51% more proteins than X!Tandem, MaxQuant, Mascot, and Comet, respectively. At the peptide level, MSFragger identified 11, 13, 34, 46, 58, and 66% more peptides than MSGF+, X!Tandem, Mascot, Comet, Proteome Discoverer, and MaxQuant, respectively.

In evaluating the data in Figure 1, we speculate that the reduced number of peptide and PSM identifications for Proteome Discoverer and MaxQuant may be attributed to the bespoke peptide and PSM statistical cutoffs that are employed by these programs. In contrast, identical/standardized cutoffs can be applied to the five algorithms that can be managed in ProHits[41] (with TPP[42]). Thus, for peptide and PSM level comparisons, we focused on the five programs (Comet, Mascot, X!Tandem, MSGF+, and MSFragger) that are integrated in the ProHits platform. Although the spectral features found in bulk samples and single-cell samples are different, the performance in peptide and PSM identifications of these programs has similar trends to their reported performances[29,34,35,45] in analyzing bulk samples, indicating that the effects of the changed spectral features on the identification may be similar across the different programs.

The nanoPOTS[14] dataset is quite different than the others, as the results were generated by mass spectrometry combined with FAIMS. In these data (Figure 1G–I), MSGF+ identified a similar number of proteins as Proteome Discoverer and 20, 47,

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308

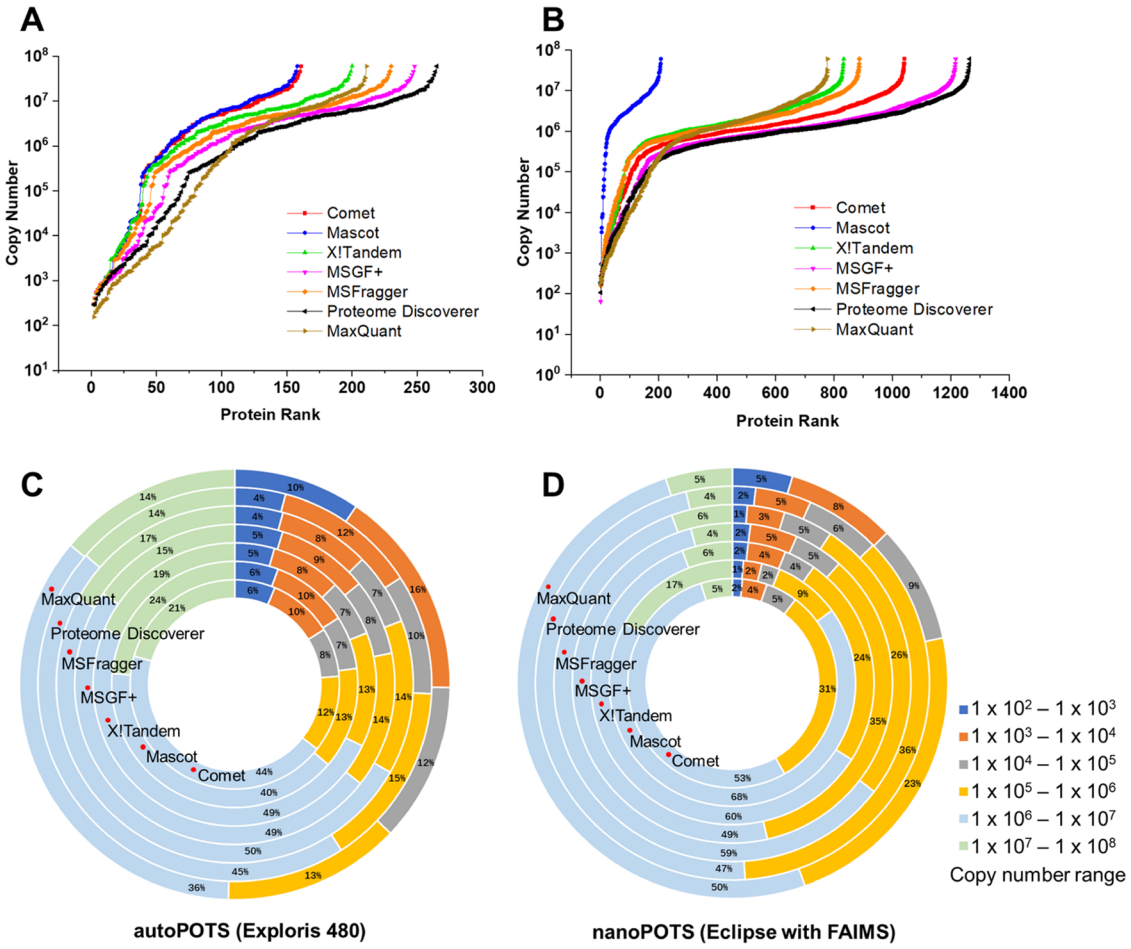


Figure 3. Abundance rank plots of copy number (from a comprehensive HeLa cell proteome dataset[52] as a function of the algorithm rank in the current analysis) for proteins in single HeLa cells in the AutoPOTS (A) and nanoPOTS (B) datasets by Comet (red), Mascot (blue), X!Tandem (green), MSGF+ (purple), MSFragger (orange), Proteome Discoverer (black), and MaxQuant (brown). Circular abundance plots of the percentage distributions for proteins with copy numbers in the ranges of 1×10^2–1×10^3 (dark blue), 1×10^3–1×10^4 (dark orange), 1×10^4–1×10^5 (gray), 1×10^5–1×10^6 (light orange), 1×10^6–1×10^7 (light blue), and 1×10^7–1×10^8 (light green) in the autoPOTS (C) and nanoPOTS (D) datasets.

and 63% more proteins than Comet, MSFragger, and X!Tandem, respectively. Likewise, MSGF+ identified 11% more peptides than Comet and 48−57% more peptides than Proteome Discoverer, MSFragger, and X!Tandem. The trend of nanoPOTS identifications by MSGF+, Comet, and Proteome Discoverer is similar to the trend observed in data collected without FAIMS (e.g., DISCO and autoPOTS). Since Proteome Discoverer is specifically designed to support data generated with FAIMS, it is likely that MSGF+ and Comet are also compatible with FAIMS data based on this result. On the other hand, MSFragger and X!Tandem identified fewer PSMs, peptides, and proteins than Proteome Discoverer, which is not consistent with the trend observed in data collected without FAIMS. Note that the version of MSFragger used here supports the analysis of FAIMS data and X!Tandem was used in a previous study to search FAIMS data,[46] suggesting that Proteome Discoverer may provide deeper analysis of FAIMS data compared to MSFragger and X!Tandem. Many fewer identifications were observed for MaxQuant and Mascot results compared to the other programs; this result is expected as neither the MaxQuant nor the Mascot version used here support data collected with FAIMS.

Finally, we note that the different sets of proteins and peptides that are identified by the different algorithms suggests important questions about the reliability of the results. On one hand, if all of the algorithms are equally reliable, we might consider combining the results to maximize the numbers. For example, we can use iProphet[44] to combine the identifications from the five algorithms supported by TPP; as shown in Figure S1, this results in substantial increases for each of the single-cell proteome datasets evaluated here. But on the other hand, we propose that skepticism is warranted for this type of combinatorial approach, given that the different algorithms operate from different assumptions and boundary conditions. In this study, we chose to focus on the unique behavior of each algorithm when applied to single-cell proteome data rather than to rigorously assess algorithm reliability. In the future, we propose that it will be useful to validate these algorithms by comparing to actual[47] and predicted[48,49] retention times, which has been shown to improve peptide/protein identification reliability and confidence.[17,50,51]

Distribution of Protein PSMs across Sample Processing Platforms

We plotted PSM distributions for the proteins identified by MSGF+ and Proteome Discoverer (as they both support

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308

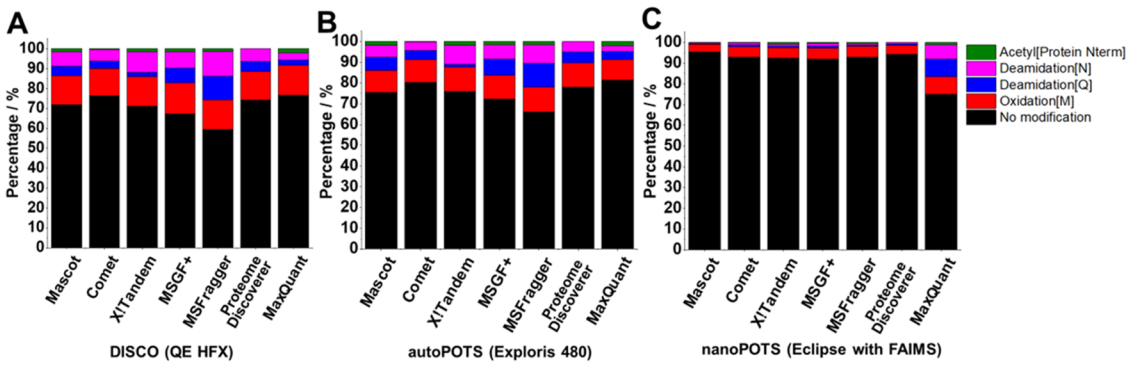


Figure 4. Proportion of four common peptide modifications (acetyl [protein N-term] in green, deamidation [N] in purple, deamidation [Q] in blue, and oxidation [M] in red and peptides with no modification in black) in single U87 cells in the DISCO dataset (A) and in single HeLa cells in the autoPOTS (B) and nanoPOTS (C) datasets by Mascot, Comet, X!Tandem, MaxQuant, MSFragger, MSGF+, and Proteome Discoverer. The measure in each bar is normalized by the total number of identified peptides by each program.

FAIMS data) for all three datasets (Figure S2). Strikingly, compared with DISCO and autoPOTS results, almost no proteins with high PSMs (>100) were observed in the nanoPOTS (with FAIMS) dataset. The average PSM for proteins identified from both the DISCO and autoPOTS datasets were much higher than that for the nanoPOTS (with FAIMS) dataset; however, the median PSM for proteins identified by each platform was similar, suggesting that FAIMS may reduce the intensity compression from high-abundance proteins, facilitating the identification of middle- to low-abundance proteins.

Comparison of Identifications across Seven Database Searching Programs

In addition to the numbers of identifications, the overlap of the identities of the proteins and peptides between different programs is another important factor to consider for single-cell proteomics data. When identification results from three different single cells are combined, Proteome Discoverer identified the greatest number of proteins from the autoPOTS dataset (326), which is even more than the number identified by MSGF+ (317), suggesting that Proteome Discoverer identified more unique proteins. MSGF+ identified 92.2% of the proteins (201/218) identified by Mascot and 89.6% of proteins (198/221) identified by Comet. Strikingly, although Proteome Discoverer identified the greatest number of proteins in triplicate analyses, it only identified 62.8% of proteins (137/218) identified by Mascot and 65.2% of proteins (144/221) identified by Comet (Figure 2A). A Venn diagram of the proteins identified in the five programs (Comet, Mascot, X!Tandem, MSGF+, and MSFragger) on the ProHits platform (Figure 2B) indicates much greater overlap than was the case for the others (MSGF+, Proteome Discoverer, and MaxQuant) (Figure 2C). This leads us to hypothesize that the differences reported by the latter set are not only a result of the different search algorithms but also may be related to the (non-standardized) statistical cutoff methods that are used in these programs. Finally, a Venn diagram of the proteins identified in Comet, X!Tandem, MSGF+, MSFragger, and Proteome Discoverer analyses of the nanoPOTS (with FAIMS) dataset (Figure 2D) indicated a similar level of overlap observed for the non-FAIMS data.

Protein Abundance Distribution in Single-Cell Data

A key challenge for single-cell proteomics is improving the ability to detect low-abundance proteins. While there are important innovations related to sample processing techniques, liquid phase separation parameters, and mass spectrometry settings for low-abundance protein detection, the data analysis methods used can also have great effect on the number of identifications and the detection accuracy of low-abundance proteins.

We examined the distribution of protein abundance in both autoPOTS and nanoPOTS datasets by matching the identification results with protein copy numbers from a comprehensive HeLa proteome dataset.[52] Summaries of the results of this study are found in Tables S1 and S2. In both the autoPOTS and nanoPOTS datasets (Figure 3A,B), MaxQuant identified the most low-abundance proteins (note that the total number of identified proteins for MaxQuant was lower than some of the other programs). Circular abundance plots of the same data (Figure 3C,D) make this more clear—as shown, MaxQuant identified the highest percentage of proteins with the global HeLa copy number between 1×10^2 and 1×10^5 in both the autoPOTS (38%) and nanoPOTS (22%) datasets, compared to the other programs which identified 16−26 and 5−13% in this range, respectively. Finally, although not every program supports FAIMS data, the abundance distributions from autoPOTS (Figure 3C) and nanoPOTS (Figure 3D) suggest that the use of FAIMS may improve the identification of midabundance proteins (1×10^5−1×10^6) in single-cell samples.

Peptide Modifications in Single-Cell Data

In addition to the identification of peptides at low abundance, identification of peptides with particular modifications is another important metric in proteomics research. Here, to investigate the ability of different programs to identify basic peptide modifications, we compare the proportion of the four most common modifications in proteomics database searching in the search results for the seven programs. In the DISCO dataset (Figure 4A), the proportion of peptide oxidation across the seven programs is very similar, ranging from 13.6 to 15.6%. In contrast, the proportion of peptide deamidation (N and Q) differs greatly across the different programs. For example, in the MaxQuant results for this dataset, only 6.2% of identified peptides were deamidated, while in the MSFragger results, 24.3% of peptides showed deamidation, and in the remaining programs, the percentage ranged from 9.5 to 15.6%. Finally, Mascot, X!Tandem, MSGF+, MSFragger, and MaxQuant identified a similar proportion of peptide acetylation in this

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308

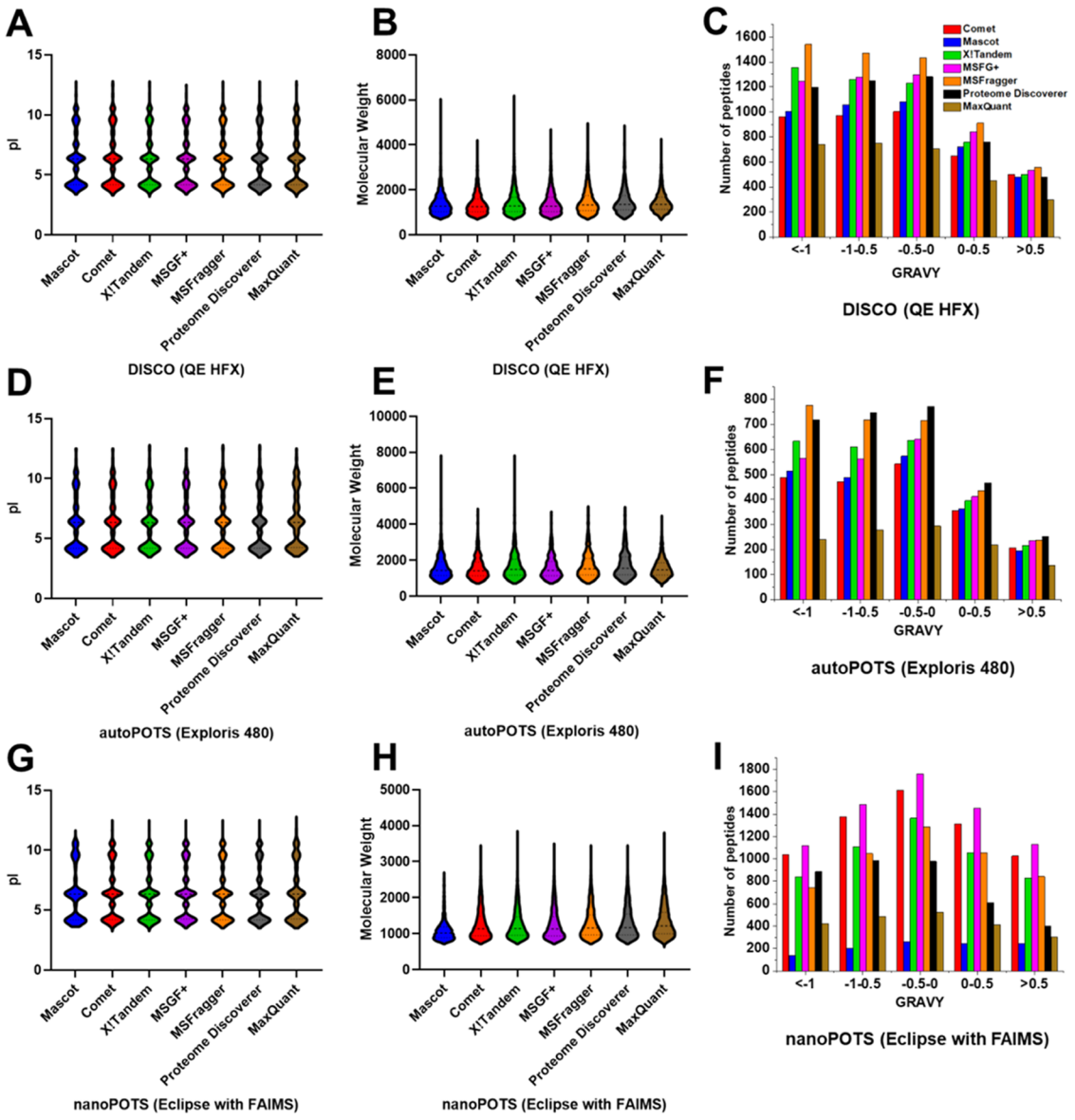


Figure 5. Properties of peptides identified in single U87 cells generated by DISCO (A–C) and HeLa cells generated by autoPOTS (D–F) and nanoPOTS (G–I) after evaluation by Mascot (blue), Comet (red), X!Tandem (green), MSGF+ (purple), MSFragger (orange), Proteome Discoverer (black), and MaxQuant (brown) from DISCO, autoPOTS, and nanoPOTS datasets. Distributions of peptide pI (A, D, G) and molecular weight (B, E, H) are shown in violin plots. Numbers of peptides in the GRAVY ranges of <−1, −1 to −0.5, −0.5 to 0, 0–0.5, and >0.5 (C, F, I) are shown as bar plots.

dataset (1.4–2.2%), while Comet identified only 0.5%, and Proteome Discoverer 0%.

In the autoPOTS dataset, the proportion of modifications identified by each program was similar to that observed in the DISCO dataset (Figure 4B). These results suggest that MSFragger identifies more peptides with basic modifications, especially deamidations (which only produce a mass shift of 0.98 Da). In the nanoPOTS dataset, the proportions of peptide modifications were also similar, ranging from 5.8 to 8.1% across the five FAIMS-supporting programs (Figure 4C). However, the proportion of total modifications observed in the nanoPOTS dataset is much smaller than that from the two datasets without FAIMS. As fewer modifications were also found in the published[14] Proteome Discoverer search results for the nanoPOTS dataset, we speculate that the specific compensation voltages used in FAIMS may filter out modified peptides. Finally, although the proportion of these modifications in MaxQuant was 25% (much greater than that from other software), the absolute number of peptides with modifications was only 538, which is less than the 562 modified peptides identified by MSGF+. This is likely because MaxQuant does not support FAIMS data, and the total number of proteins identified in this dataset by MaxQuant is much smaller than most of the other algorithms, as shown in Figure 1H.

Distribution of Peptide Properties in Single-Cell Data

Peptide properties such as isoelectric point (pI), molecular weight (MW), and grand average of hydropathicity (GRAVY) are of great interest as they can predict peptide retention in chromatography and peptide ionization in mass spectrometry. To examine the differences between the properties of peptides identified by different programs on different platforms, we plotted their respective pI, MW, and GRAVY value

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298–1308

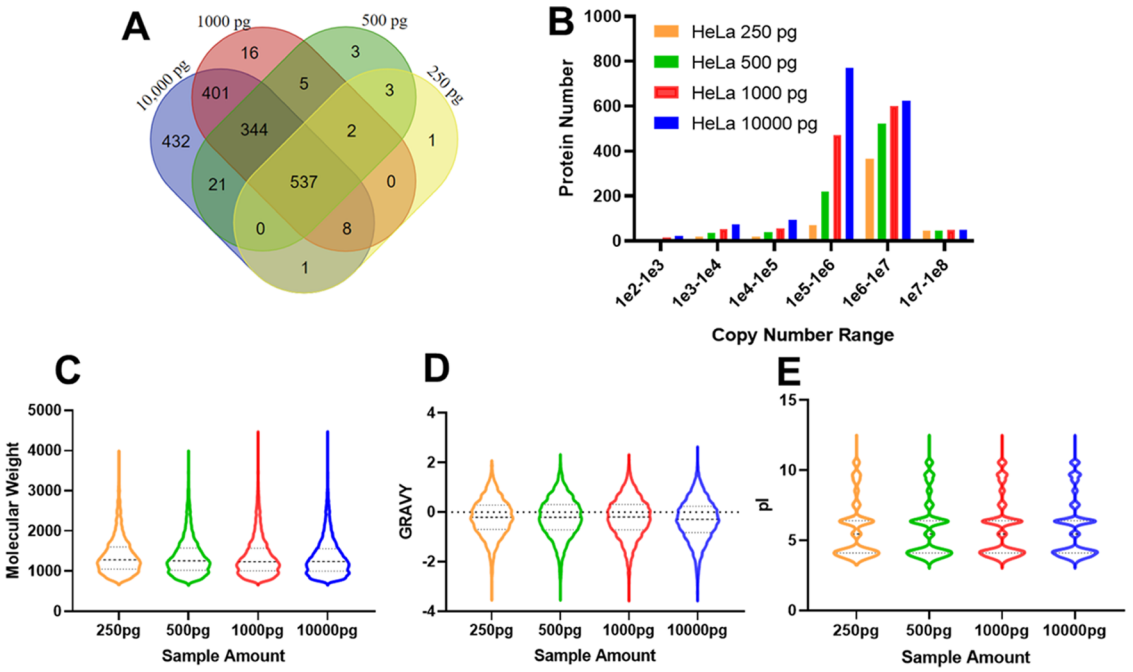


Figure 6. Mass-loading experiments of HeLa digest dilutions containing 250 pg (yellow), 500 pg (green), 1000 pg (red), and 10,000 pg (blue) total protein acquired by HPLC-MS/MS (QE HFX) and searched by MaxQuant. Proteins represented in each group are the combined set identified in any of three replicates per condition. Venn diagrams of the proteins identified at each sample loading (A). Bar chart of numbers of proteins identified at different loading levels as a function of protein copy numbers from a comprehensive HeLa cell proteome dataset (B).[52] Violin plots of molecular weight (C), GRAVY (D), and pI (E) distributions of identified peptides for each sample loading amount.

distributions (Figure 5). Overall, the pI, MW, and GRAVY value distributions of the peptides identified by each of the seven programs are consistent. The distributions of these properties for peptides identified from DISCO and autoPOTS datasets are also similar, which suggests that the DISCO and autoPOTS sample preparation platforms have minimal bias for peptides with specific properties. Interestingly, the datasets collected without FAIMS (Figure 5B,E) identify many peptides with MW larger than 4000 Da; in contrast, data from the nanoPOTS platform with FAIMS (Figure 5H) did not identify any peptides with MW larger than 4000 Da, suggesting a bias for smaller peptides or potentially the elimination of peptides with MW larger than 4000 Da because of the FAIMS compensation voltage settings. On the other hand, data collected using the conventional (non-FAIMS) workflows (Figure 5C,F) suffered from the well-known problem[16,53] (attributed to loss of hydrophobic peptides during processing) of identifying few peptides with GRAVY scores greater than 0. In contrast, the nanoPOTS workflow with FAIMS provides much better identification of peptides with GRAVY scores greater than 0 (Figure 5I), indicating that FAIMS may help improve the identification of hydrophobic peptides. This effect has been reported previously,[54] with the explanation that FAIMS allows for improved resolution between coeluting hydrophobic peptides (and background constituents), allowing the peptides to be more readily identified.

Experimental Results for Varied Loading of HeLa Protein Digest

Finally, to investigate the changes associated with loading different amounts of samples and to explore directions that might improve single-cell proteomics analyses in the future, we generated HeLa digest dilutions with total protein amounts ranging from the average expected in a single-cell level (250 pg) to samples containing many-fold higher than that amount (10,000 pg). In particular, we were interested in the reproducibility in detecting the same proteins in replicates at these different mass-loading levels. In this case, a single database searching program, MaxQuant, was used to evaluate the protein identifications from samples processed using standard HPLC-MS/MS with a Thermo Fisher Scientific Q Exactive HFX spectrometer.

Figure S3 illustrates the differences in reproducibility of protein identification for different loading amounts. As the loading amount increases from 250 to 10,000 pg, the percentage of proteins identified by all three replicates increases from 63.2 to 76.6%, and the percentage of proteins identified in only one replicate decreases from 22.8 to 1.2%. This change in reproducibility of identification within replicates is especially obvious when comparing the difference between loading 1000 and 10,000 pg, indicating that, even for replicates analyzed on the same HPLC-MS/MS system running the same algorithm, the variation in HPLC separation and MS acquisition may significantly reduce the reproducibility when the loading amount is less than 1000 pg. This is unfortunate, given that the single mammalian cells evaluated here (and most often) contain less protein than this cutoff, and provides justification for the use of signal matching or boosting strategies MBR or TMT techniques.[1,18−25]

A closer look at the mass-loading data reveals a number of interesting results. First, as shown in Figure 6A,B, for proteins with abundance greater than 1×10^7 copies, there is almost no effect on identification when the loading amount increases. This makes intuitive sense, as the signal of high-abundance proteins is likely high enough such that increasing sample loading does not provide benefit for identification. Second, for proteins with copy numbers from 1×10^6 to 1×10^7, the increase in the number of identifications is relatively small

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308

when the loading amount becomes larger than 1000 pg; however, for proteins with copy numbers less than 1×10^6, the increase in the number of identifications is substantially impacted by increased sample loading. This suggests that researchers should develop improved sample processing techniques to minimize loss of low-copy number proteins in future single-cell proteomics studies. Third, although the MW (Figure 6C) and GRAVY (Figure 6D) value distributions were wider with increased loading, the number of newly identified peptides with large MW or GRAVY values at high sample loading values is small, suggesting that it might be more efficient to improve the identification of these peptides using other database searching programs (as shown in Figure 5B,E) or using methods relying on FAIMS (as shown in Figure 5I). Finally, the distribution of pI across different loading amounts is consistent (Figure 6E), which suggests that more peptides in each pI region could be identified uniformly with the increase of loading amount.

CONCLUSIONS

In this work, seven database searching programs for proteomic data were applied to analyzing three single-cell proteomics datasets that were generated using different workflows (DISCO, autoPOTS, and nanoPOTS). The seven programs were compared on the basis of their identifications for PSMs, peptides, proteins, and peptide modifications, as well as their ability to analyze low-abundance proteins and peptides with different properties. In addition, a new experiment comparing different loading amounts, from single-cell level up to 10 ng of total protein, was conducted to investigate the effects of loading on the results. The comparison suggests that MSGF+, MSFragger, and Proteome Discoverer are generally more efficient in identifying proteins, that MaxQuant is better suited for analyzing low-abundance proteins, that MSFragger is superior in elucidating PTM, and that Mascot and X!Tandem are more appropriate for identifying longer peptides. In the future, tools such as iProphet may be useful for generating combinatorial results from the different algorithms, but we note that additional validation studies should be carried out, both to evaluate combinatorial approaches and to confirm which algorithms have the highest identification accuracies. In addition, new experiments indicated that improving protein recovery during sample preparation may significantly improve the identification of middle- to low-abundance proteins relative to high-abundance proteins from small initial sample amounts. Likewise, the use of FAIMS seems to substantially improve the identification of hydrophobic peptides, which are usually difficult to analyze by traditional HPLC-MS/MS strategies. We propose that the results described here may serve as a reference for researchers who are choosing the database searching programs to apply to their single-cell proteomics data, as well as to guide future programmers who develop the next generation of single-cell proteome profiling algorithms.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00821.

Peptide and protein identifications integrated by iProphet (Figure S1); PSM numbers of proteins identified by Proteome Discoverer and MSGF+ from three datasets (Figure S2); plot of percentages of proteins identified in replicates of HeLa digest dilutions (Figure S3); percentage distributions of protein numbers in each abundance rank in the AutoPOTS dataset (Table S1); and percentage distributions of protein numbers in each abundance rank in the nanoPOTS dataset (Table S2) (PDF)

AUTHOR INFORMATION

Corresponding Authors

Shen Zhang − *Clinical Research Center for Reproduction and Genetics in Hunan Province, Reproductive and Genetic Hospital of CITIC-XIANGYA, Changsha, Hunan 410000, China*; Email: szhang231@126.com

Aaron R. Wheeler − *Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada; Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada; Institute of Biomedical Engineering, University of Toronto, Toronto, Ontario M5S 3G9, Canada*; orcid.org/0000-0001-5230-7475; Email: aaron.wheeler@utoronto.ca

Authors

Jiaxi Peng − *Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada; Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada; Institute of Biomedical Engineering, University of Toronto, Toronto, Ontario M5S 3G9, Canada*

Calvin Chan − *Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada*; orcid.org/0000-0003-0640-0921

Fei Meng − *Clinical Research Center for Reproduction and Genetics in Hunan Province, Reproductive and Genetic Hospital of CITIC-XIANGYA, Changsha, Hunan 410000, China*

Yechen Hu − *Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada; Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada; Institute of Biomedical Engineering, University of Toronto, Toronto, Ontario M5S 3G9, Canada*

Lingfan Chen − *Fujian Province New Drug Safety Evaluation Centre, Fujian Medical University, Fuzhou, Fujian 350108, China*

Ge Lin − *Clinical Research Center for Reproduction and Genetics in Hunan Province, Reproductive and Genetic Hospital of CITIC-XIANGYA, Changsha, Hunan 410000, China; Laboratory of Reproductive and Stem Cell Engineering, NHC Key Laboratory of Human Stem Cell and Reproductive Engineering, Central South University, Changsha, Hunan 410075, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00821

Author Contributions

J.P., S.Z., G.L, and A.R.W. conceived the concept of program comparison for single-cell data analysis. S.Z., J.P., and C.C. performed the database searching experiments. S.Z., J.P., L.C., Y.H., and F.M. carried out the data analysis. J.P., C.C., S.Z., and A.R.W. wrote and edited the manuscript. All authors discussed the results and commented on the manuscript.

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from the Reproductive and Genetic Hospital of CITIC-XIANGYA (YNXM-202108 and YNXM-202211 to S.Z.), the Hundred Youth Talents Program of Hunan Province (to S.Z.), the National Science and Engineering Research Council of Canada (RGPIN 2019-04867 to A.R.W.), and the Canadian Foundation for Innovation/Province of Ontario (to A.R.W.). J.P. and Y.H. acknowledge the Precision Medicine initiative (PRiME) for postdoctoral fellowships, and J.P. acknowledges MITACS for a postdoctoral fellowship. S.Z. was supported by an ASMS Postdoctoral Career Development Award. A.R.W. acknowledges the Canada Research Chair (CRC) program. The authors thank Prof. Anne-Claude Gingras and Brett Larsen for fruitful discussions and assistance with database searching on ProHits.

■ REFERENCES

(1) Schoof, E. M.; Furtwangler, B.; Uresin, N.; Rapin, N.; Savickas, S.; Gentil, C.; Lechman, E.; Keller, U. A. D.; Dick, J. E.; Porse, B. T. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nat. Commun.* **2021**, *12*, No. 3341.

(2) Buettner, F.; Natarajan, K. N.; Casale, F. P.; Proserpio, V.; Scialdone, A.; Theis, F. J.; Teichmann, S. A.; Marioni, J. C.; Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **2015**, *33*, 155−160.

(3) Colhoun, H. M.; McKeigue, P. M.; Smith, G. D. Problems of reporting genetic associations with complex outcomes. *The Lancet* **2003**, *361*, 865−872.

(4) Sun, L.; Dubiak, K. M.; Peuchen, E. H.; Zhang, Z.; Zhu, G.; Huber, P. W.; Dovichi, N. J. Single Cell Proteomics Using Frog (Xenopus laevis) Blastomeres Isolated from Early Stage Embryos, Which Form a Geometric Progression in Protein Content. *Anal. Chem.* **2016**, *88*, 6653−6657.

(5) Zhang, X.; Deeke, S. A.; Ning, Z.; Starr, A. E.; Butcher, J.; Li, J.; Mayne, J.; Cheng, K.; Liao, B.; Li, L.; Singleton, R.; Mack, D.; Stintzi, A.; Figeys, D. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* **2018**, *9*, No. 2873.

(6) Shapiro, E.; Biezuner, T.; Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **2013**, *14*, 618−630.

(7) Paik, D. T.; Cho, S.; Tian, L.; Chang, H. Y.; Wu, J. C. Single-cell RNA sequencing in cardiovascular development, disease and medicine. *Nat. Rev. Cardiol.* **2020**, *17*, 457−473.

(8) Vistain, L. F.; Tay, S. Single-Cell Proteomics. *Trends Biochem. Sci.* **2021**, *46*, 661−672.

(9) Hunter, K. Host genetics influence tumour metastasis. *Nat. Rev. Cancer* **2006**, *6*, 141−146.

(10) Kelly, R. T. Single-cell Proteomics: Progress and Prospects. *Mol. Cell Proteomics* **2020**, *19*, 1739−1748.

(11) Marx, V. A dream of single-cell proteomics. *Nat Methods* **2019**, *16*, 809−812.

(12) Doerr, A. Single-cell proteomics. *Nat. Methods* **2019**, *16*, 20.

(13) Liang, Y.; Acor, H.; McCown, M. A.; Nwosu, A. J.; Boekweg, H.; Axtell, N. B.; Truong, T.; Cong, Y.; Payne, S. H.; Kelly, R. T. Fully Automated Sample Processing and Analysis Workflow for Low-Input Proteome Profiling. *Anal. Chem.* **2021**, *93*, 1658−1666.

(14) Cong, Y.; Motamedchaboki, K.; Misal, S. A.; Liang, Y.; Guise, A. J.; Truong, T.; Huguet, R.; Plowey, E. D.; Zhu, Y.; Lopez-Ferrer, D.; Kelly, R. T. Ultrasensitive single-cell proteomics workflow identifies >1000 protein groups per mammalian cell. *Chem. Sci.* **2021**, *12*, 1001−1006.

(15) Specht, H.; Slavov, N. Transformative Opportunities for Single-Cell Proteomics. *J Proteome Res* **2018**, *17*, 2565−2571.

(16) Li, Z. Y.; Huang, M.; Wang, X. K.; Zhu, Y.; Li, J. S.; Wong, C. C. L.; Fang, Q. Nanoliter-Scale Oil-Air-Droplet Chip-Based Single Cell Proteomic Analysis. *Anal. Chem.* **2018**, *90*, 5430−5438.

(17) Zhu, Y.; Clair, G.; Chrisler, W. B.; Shen, Y.; Zhao, R.; Shukla, A. K.; Moore, R. J.; Misra, R. S.; Pryhuber, G. S.; Smith, R. D.; Ansong, C.; Kelly, R. T. Proteomic Analysis of Single Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS. *Angew. Chem.* **2018**, *130*, 12550−12554.

(18) Cheung, T. K.; Lee, C. Y.; Bayer, F. P.; McCoy, A.; Kuster, B.; Rose, C. M. Defining the carrier proteome limit for single-cell proteomics. *Nat Methods* **2021**, *18*, 76−83.

(19) Budnik, B.; Levy, E.; Harmange, G.; Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **2018**, *19*, 161.

(20) Specht, H.; Emmott, E.; Petelski, A. A.; Huffman, R. G.; Perlman, D. H.; Serra, M.; Kharchenko, P.; Koller, A.; Slavov, N. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* **2021**, *22*, No. 50.

(21) Dou, M.; Clair, G.; Tsai, C. F.; Xu, K.; Chrisler, W. B.; Sontag, R. L.; Zhao, R.; Moore, R. J.; Liu, T.; Pasa-Tolic, L.; Smith, R. D.; Shi, T.; Adkins, J. N.; Qian, W. J.; Kelly, R. T.; Ansong, C.; Zhu, Y. High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform. *Anal. Chem.* **2019**, *91*, 13119−13127.

(22) Petelski, A. A.; Emmott, E.; Leduc, A.; Huffman, R. G.; Specht, H.; Perlman, D. H.; Slavov, N. Multiplexed single-cell proteomics using SCoPE2. *Nat. Protoc.* **2021**, *16*, 5398−5425.

(23) Zhu, Y.; Piehowski, P. D.; Zhao, R.; Chen, J.; Shen, Y.; Moore, R. J.; Shukla, A. K.; Petyuk, V. A.; Campbell-Thompson, M.; Mathews, C. E.; Smith, R. D.; Qian, W. J.; Kelly, R. T. Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat. Commun.* **2018**, *9*, No. 882.

(24) Cong, Y.; Liang, Y.; Motamedchaboki, K.; Huguet, R.; Truong, T.; Zhao, R.; Shen, Y.; Lopez-Ferrer, D.; Zhu, Y.; Kelly, R. T. Improved Single-Cell Proteome Coverage Using Narrow-Bore Packed NanoLC Columns and Ultrasensitive Mass Spectrometry. *Anal. Chem.* **2020**, *92*, 2665−2671.

(25) Gebreyesus, S. T.; Siyal, A. A.; Kitata, R. B.; Chen, E. S.; Enkhbayar, B.; Angata, T.; Lin, K. I.; Chen, Y. J.; Tu, H. L. Streamlined single-cell proteomics by an integrated microfluidic chip and data-independent acquisition mass spectrometry. *Nat. Commun.* **2022**, *13*, No. 37.

(26) Boekweg, H.; Van Der Watt, D.; Truong, T.; Johnston, S. M.; Guise, A. J.; Plowey, E. D.; Kelly, R. T.; Payne, S. H. Features of Peptide Fragmentation Spectra in Single-Cell Proteomics. *J. Proteome Res.* **2022**, *21*, 182−188.

(27) Orsburn, B. C. Evaluation of the Sensitivity of Proteomics Methods Using the Absolute Copy Number of Proteins in a Single Cell as a Metric. *Proteomes* **2021**, *9*, 34.

(28) Huffman, R. G.; Chen, A.; Specht, H.; Slavov, N. DO-MS: Data-Driven Optimization of Mass Spectrometry Methods. *J. Proteome Res.* **2019**, *18*, 2493−2500.

(29) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22−24.

(30) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.

(31) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−1467.

(32) Fenyö, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768−774.

(33) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7*, 3354−3363.

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308

(34) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, No. 5277.
(35) 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*, 513−520.
(36) Orsburn, B. C. Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **2021**, *9*, 15.
(37) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.
(38) Barnett, D. A.; Ells, B.; Guevremont, R.; Purves, R. W. Application of ESI-FAIMS-MS to the analysis of tryptic peptides. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 1282−1291.
(39) Saba, J.; Bonneil, E.; Pomies, C.; Eng, K.; Thibault, P. Enhanced sensitivity in proteomics experiments using FAIMS coupled with a hybrid linear ion trap/Orbitrap mass spectrometer. *J. Proteome Res.* **2009**, *8*, 3355−3366.
(40) Hebert, A. S.; Prasad, S.; Belford, M. W.; Bailey, D. J.; McAlister, G. C.; Abbatiello, S. E.; Huguet, R.; Wouters, E. R.; Dunyach, J.-J.; Brademan, D. R.; et al. Comprehensive single-shot proteomics with FAIMS on a hybrid orbitrap mass spectrometer. *Anal. Chem.* **2018**, *90*, 9529−9537.
(41) Liu, G.; Zhang, J.; Larsen, B.; Stark, C.; Breitkreutz, A.; Lin, Z.-Y.; Breitkreutz, B.-J.; Ding, Y.; Colwill, K.; Pasculescu, A.; et al. ProHits: integrated software for mass spectrometry−based interaction proteomics. *Nat. Biotechnol.* **2010**, *28*, 1015−1017.
(42) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, *1*, No. 2005.0017.
(43) Lamanna, J.; Scott, E. Y.; Edwards, H. S.; Chamberlain, M. D.; Dryden, M. D. M.; Peng, J. X.; Mair, B.; Lee, A.; Chan, C.; Sklavounos, A. A.; Heffernan, A.; Abbas, F.; Lam, C.; Olson, M. E.; Moffat, J.; Wheeler, A. R. Digital microfluidic isolation of single cells for -Omics. *Nat. Commun.* **2020**, *11*, No. 5632.
(44) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell Proteomics* **2011**, *10*, No. M111.007690.
(45) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Mol. Cell Proteomics* **2008**, *7*, 962−970.
(46) Swearingen, K. E.; Hoopmann, M. R.; Johnson, R. S.; Saleem, R. A.; Aitchison, J. D.; Moritz, R. L. Nanospray FAIMS fractionation provides significant increases in proteome coverage of unfractionated complex protein digests. *Mol. Cell Proteomics* **2012**, *11*, No. M111.014985.
(47) Palmblad, M.; Ramstrom, M.; Bailey, C. G.; McCutchen-Maloney, S. L.; Bergquist, J.; Zeller, L. C. Protein identification by liquid chromatography-mass spectrometry using retention time prediction. *J Chromatogr B* **2004**, *803*, 131−135.
(48) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **2021**, *18*, 1363−1369.
(49) Zeng, W. F.; Zhou, X. X.; Willems, S.; Ammar, C.; Wahle, M.; Bludau, I.; Voytik, E.; Strauss, M. T.; Mann, M. AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **2022**, *13*, No. 7238.
(50) Yang, Z.; Sun, L. Recent technical progress in sample preparation and liquid-phase separation-mass spectrometry for proteomic analysis of mass-limited samples. *Anal. Methods* **2021**, *13*, 1214−1225.
(51) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteomics* **2014**, *13*, 2513−2526.
(52) Bekker-Jensen, D. B.; Kelstrup, C. D.; Batth, T. S.; Larsen, S. C.; Haldrup, C.; Bramsen, J. B.; Sorensen, K. D.; Hoyer, S.; Orntoft, T. F.; Andersen, C. L.; Nielsen, M. L.; Olsen, J. V. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **2017**, *4*, 587−599 e4.
(53) Deng, J.; Julian, M. H.; Lazar, I. M. Partial enzymatic reactions: A missed opportunity in proteomics research. *Rapid Commun. Mass Spectrom.* **2018**, *32*, 2065−2073.
(54) Bagag, A.; Giuliani, A.; Canon, F.; Refregiers, M.; Le Naour, F. Separation of peptides from detergents using ion mobility spectrometry. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 3436−3440.

https://doi.org/10.1021/acs.jproteome.2c00821
J. Proteome Res. 2023, 22, 1298−1308