Triqler for Protein Summarization of Data from Data Independent Acquisition Mass Spectrometry

Patrick Truong[1], Matthew The[2], and Lukas Käll[1,*]

[1]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Royal Institute of Technology – KTH, Solna, Sweden
[2]Chair of Proteomics and Bioanalytics, Technical University of Munich (TUM), Freising, Germany
[*]Corresponding author, lukas.kall@scilifelab.se

February 23, 2023

Abstract

A frequent goal, or subgoal, when processing data from a quantitative shotgun proteomics experiment is a list of proteins that are differentially abundant under the examined experimental conditions. Unfortunately, obtaining such a list is a challenging process, as the mass spectrometer analyses the proteolytic peptides of a protein rather than the proteins themselves. We have previously designed a Bayesian hierarchical probabilistic model, Triqler, for combining peptide identification and quantification errors into probabilities of proteins being differentially abundant.

However, the model was developed for data from data-dependent acquisition. Here we show that Triqler is also compatible with data-independent acquisition data after applying minor alterations for the missing value distribution. Furthermore, we find that it has better performance than ~~other protein summarization tools when compared to~~ a set of compared state-of-the-art ~~DIA processing methods~~protein summarization tools when evaluated on data-independent acquisition data.

Introduction

Mass spectrometry (MS)-based proteomics enables efficient detection of proteins in complex mixtures. There are several different techniques to make the technology quantitative.[1,2] Out of these, label-free quantification (LFQ)[3] has the advantage that it can smoothly handle large sample sizes~~in a straightforward manner~~. For data from LFQ experiments, just as in other quantification schemes, there exists a plethora of processing options, all containing several processing steps, each subject to their different error sources, all affecting the result of the processing.

We have previously designed a hierarchical Bayesian model, Triqler, able to control for errors from both the identification and quantification process in LFQ experiments[4]. By integrating the error probabilities from identification and quantification, one can obtain better accuracy in calling differentially abundant proteins. Triqler ~~was~~ does so by assuming that peptide abundances follow a probability distribution. This abundance distribution is initiated according to a prior distribution, but updated based on each sample's registered peptide abundance values. This results in peptide abundance distributions that are integrated into protein abundance and subsequently fold change distributions. In the processing Triqler weights in information on differences in peptide abundances within sample groups and search engine identification error probabilities. Data that indicate uncertainty results in wider abundance distributions, while certainty results in tighter abundance distributions. Triqler also integrates the resulting fold change distributions to derive posterior probabilities for each protein having an actual fold change larger than a preset threshold value. These posterior probabilities are also averaged into q values [4].

Triqler was originally designed for handling LFQ data from Data-dependent acquisition (DDA). However, ~~many labs prefer~~ in principle, once peptide abundances and identities are estimated, there are only small differences between data from DDA and Data-independent acquisition (DIA) mass spectrometry[5] ~~as they find that it gives more reproducible peptide detection, and allows for a broader dynamical range in quantification , compared to DDA.~~

. Here, we set out to investigate Triqler's ability to summarize protein concentrations from peptide abundances derived from DIA data. We used the LFQBench from Navarro et al.[6] for the evaluation. In the original benchmark, Navarro et al. included a comparison of different protein summarization strategies and found that the so-called Top3 method generally resulted in lower variance and better quantification accuracy than the built-in methods from OpenSwath, SWATH2.0, Skyline, Spectronaut, and DIA-Umpire.[6] However, there are reasons to believe that more sophisticated methods would yield better protein quantification than the ~~Top3 method~~Top3 method. Simple summarization methods based on mean and median peptide intensity have been shown to produce unreliable protein abundance estimates,[7] and more advanced summarization strategies for LFQ data have been proposed in the literature.[8,9] Summarization techniques such as PQPQ,[10] MSstats,[11] Diffacto,[12] MSqRob2[13] and Triqler[4] have all been shown to outperform Top3 and there are no theoretical reasons why the stated methods would not perform well for DIA data. Hence, we found it apt to ~~benchmark Triqler~~ modify Triqler to handle DIA data, and to benchmark it against a set of state-of-the-art protein summarization methods~~using peptide quantities from the LFQBench DIA data set~~.

Materials and methods

~~Data description~~

LFQBench mass spectrometry data

We downloaded the LFQBench dataset[6] from ~~PRIDE~~ ProteomeXchange (PXD002952). Here we used the TripleTOF6600 section of the study, which was harvested with a setup of 32 fixed windows MS2-windows. We also restricted ourselves to the low ratio difference samples, referred to as the HYE124 hybrid proteome samples in the original study. These consist of triplicates of Sample A composed of tryptically digested proteins from 65% w/w HeLa, 30% w/w yeast, and 5% w/w *E. coli* cells, and triplicates of Sample B, composed of 65% w/w, 15% w/w yeast, and 20% w/w *E. coli* proteins. Samples from HYE110 and the TripleTOF5600 section of PXD002952 were omitted in this study. Further details about mass spectrometric instrumentation and data acquisition are available in Navarro et al.[6] The `.wiff` files were converted to `.mzML` files in a centroided format using msconvert (using Windows OS msconvert version 3.0) with peakPicking filter msLevel=1-).

LFQBench sequence database

Uniprot FASTA files with one protein sequence per gene were downloaded for each species (UP000005640, UP000000625, and UP000002311, acquired on 2021-06-16). The unfiltered FASTA files contained 20 590 human proteins, 6 046 yeast proteins, and 4 373 *E. coli* proteins. To reduce the effect of the different protein inference strategies for the tested protein summarization tools, a modified FASTA file, without shared peptides, was used for database search. The filter removed protein sequences with shared peptides so that the final database did not contain any two proteins sharing tryptic peptides longer than 7 amino acids. After filtering the FASTA file contained 20 302 proteins (288 human proteins fewer proteins than in an unfiltered database), 5 848 yeast proteins (198 yeast proteins fewer proteins than in an unfiltered database), and 4 306 *E. Coli* proteins (67 *E. Coli* proteins than in an unfiltered database). Replacing the I/L amino acids to handle the mass equivalence did not result in any considerable differences (See Supplementary Table S1). We also added pseudo-reverse sequences to the database as decoys for target-decoy analysis using OpenSwathDecoyGenerator.

Cancer proteomics data

We also downloaded two DIA data cancer proteomics-related datasets from ProteomeXchange, with accession numbers PXD004684 and PXD022992. The first set consists of data from four frozen lung squamous cell carcinomas and four adjacent tissues harvested with a Thermo QExactive Plus [14]. The second set consists of data from a comparison of three primary cell lines (SK-MEL-1, A375, and G-361) to three metastatic (RPMI-7951, SH-4, and SK-MEL-3) cell lines analyzed (two samples per cell line) by a Thermo Orbitrap Fusion Lumos Tribrid [15].

General workflow

We used ~~two~~ three separate strategies to generate peptide abundances from the DIA runs. First, we used a spectral library consisting of selected spectra from separate DDA runs, we refer to this workflow as ID~~hereon, and~~, secondly, we searched pseudo-spectra generated directly from the DIA data, we will refer to this workflow as PS~~from hereon. The workflows are shown in Figure ??~~, and thirdly we applied a DIA-NN centered pipeline that also operated directly on DIA data. We will describe the parameter choices of ~~both these two~~ these methods below.

~~The DDA identification-driven matching (ID) and pseudo-spectrum workflow pipelines (PS).In the pipelines DIA-Umpire(SE), MSFragger and EasyPQP are run from Fragpipe (). A 1% PSM-level identification threshold was applied to Top3 and MSstats, a 1% peptide-level threshold to MSqRob2, while all the identifications were provided unfiltered to Triqler.~~

DDA identification-driven (ID) spectral library

For the ID workflow, we searched the ~~LFQBench provided~~ LFQBench-provided DDA runs with MSFragger[16] with default settings~~(Precursor mass tolerance of -20, 20ppm, fragment tolerance of 20 ppm, Calibration and Optimization: "Max calibration, parameter optimization", isotope error: 0/1, Data type: DDA, Load rules: strict trypsin, Cut after: KR, Cleavages: ENZYMATIC, Missed cleavages: 2, Clip N-term M: True, Peptide length of 7, 50, Peptide mass range of 500, 5000, Split database: 1, and allowing for oxidation on methionine and protein N-terminus modifier as variable modifications)~~, and a spectral library was constructed with EasyPQP[17] with the default setting~~(RT Calibration: "Automatic selection of a run as reference RT", RT Lowess Fraction: 0.1, UniMod annotation tol(Da), Fragment annotation tol(ppm): 15, and the default PSM-level threshold of 0.01, peptide-level false discovery rate (FDR) of 0.01, and protein-level FDR of 0.01)~~. No transition refinement was used. OpenSwatchDecoyGenerator was used to generate decoys for the spectral library with a pseudo-reverse method. The DIA data ~~was~~ were searched with the spectral library through OpenSwath Workflow, and the `m_score` was computed using PyProphet.[18] The `m_score` cut-off for a user-specified peptide identification ~~false-discovery~~ false discovery rate was computed with SWATH2STATS.[19] This process resulted in a set of detected peptides together with their assessed peptide identification accuracies and abundance estimates (Supplementary Table S2 shows the number of identified peptides and proteins).

Pseudo-spectra generation (PS) spectral library

For the PS workflow, we also used the FragPipe software, employing DIA-Umpire to extract pseudo-spectra from the DIA data. The DIA-Umpire parameters ~~was~~ were set to default values~~(MS1 PPM: 10, MS2 PPM: 20, Max Missed Scans: 1, Mass Defect Filter: True, RP max: 25, RF max: 500, Corr Threshold: 0, Delta Apex: 0.2, RT Overlap 0.3, Mass Defect Offset 0.1, Isotope Pattern: 0.3, MS1 SN: 1.1, MS2 SN: 1.1, Adjust Fragment Intensity: True)~~. The pseudo-spectra were subsequently searched using MSFragger with default settings~~(Precursor mass tolerance of -20, 20ppms, fragment tolerance of 20 ppms, Calibration and Optimization: "Max calibration, parameter optimization", isotope error: 0/1, Data type: DDA, Load rules: strict trypsin, Cut after: KR, Cleavages: ENZYMATIC, Missed cleavages: 2, Clip N-term M: True, Peptide length of 7, 50, Peptide mass range of 500, 5000, Split database: 1, and allowing for oxidation on methionine and protein N-terminus modifier as variable modifications)~~. A spectral library was built from the resulting PSMs using easyPQP with default ~~setting (RT Calibration: "Automatic selection of a run as reference RT", RT Lowess Fraction: 0.1, UniMod annotation tol(Da), Fragment annotation tol(ppm): 15, and the default PSM-level threshold of 0.01, peptide-level FDR of 0.01, and protein-level FDR of~~

0.01)settings. DIA-NN [20] (version 1.7.12) was used for peptide quantifications with settings specified in (Protein inference: "Off", Quantification strategy: "Robust LC (High accuracy)", Precursor FDR (%): 1.0, Protease: "Trypsin/P", Missed cleavages: 1, Maximum number of variable modifications: 0, N-term M excision: True, C carbamidomethylation: True, Peptide length: 7, 30, Precursor m/z range: 300, 1800, Fragment ion m/z range: 200, 1800, Mass accuracy: 0.0default settings.

Directly extracted peptide abundances

For this workflow, we used the quantms nf-core pipeline (`https://nf-co.re/quantms`), MS1 accuracy: 0.0, Scan window: 0, Use isotopologues: True, Remove likely interferences: True, Neural network classifier: "Single-pass mode", Protein inference: Genes, Cross-run normalization: "RT-dependent", Library generation: "Smart profiling"). To compute false discovery rates, which converts raw files to mzML with thermorawfileparser [21] and subsequently extracts peptide abundances using DIA-NNuses a built-in custom implementation of the mProphet algorithm(Supplementary Table S3 shows the number of identified peptides and proteins) [21]. The pipeline generated decoys for FDR calculations which were discarded after DIA-NN processing. To circumvent the lack of decoys in output for Triqler we concatenated shuffled entrapment sequences in the FASTA database [22]. This workflow was used with the two cancer proteomics-related datasets.

Protein summarization

The peptide quantities were summarized to proteins using the average of the three most intense peptides (We call it Top3), MSstats, MSqRob2, and Triqler. This is done for both the ID and PS all three pipelines.

Top3

We implemented a short script that takes the average of the 3 most abundant PSMs for each protein and sample. In samples only having two PSMs these were also included, still represented by their average. Proteins with one or zero PSMs per sample were excluded. PSMs were filtered at a 1% PSM-level FDR before performing the Top3 protein summarization.

MSstats

We installed MSstats version 3.18.5 using R/Bioconductor (available at `https://www.bioconductor.org/packages/release/bioc/html/MSstats.html`). MSstats use feature-level data, allowing for multiple PSM hits per peptide identification. We use the disaggregate function to disaggregate the PSMs to fragment-level data. We filtered so that every protein had at least 2 peptides and a maximum of 10 peptides, and we thresholded with a `m_score` which corresponds to a peptide identification FDR lower than 0.01 for both ID and PS pipelines. For the cancer data set we allowed up a maximum of 100 peptides. In the PS pipeline and nf-core pipeline, the data was filtered on the `Q.Value` columns from the DIA-NN output file. In the ID pipeline, we computed an `m_score` corresponding to a 1% FDR and used this `m_score` to filter the data. MSstats was run using the MSstats command `dataProcess`. The significance testing between conditions was performed using the MSstats function `groupComparison`.

MSqRob2

We installed MSqRob2 version 0.9 using R/Bioconductor (available at `https://github.com/statOmics/msqrob2`). MSqRob2 takes peptide-level input. The output from OpenSwath and DIA-NN is at PSM-level. We select the top PSM hit as our peptide and filter the data on 1% peptide level FDR to get peptide-level data. The highest scoring highest-scoring PSMs are selected by the highest `m_score` for OpenSwath and highest `CScore` for DIA-NN. MsqRob2 was run using the MSqRob2 command `msqrob`, where the contrast was set to `condition`. MSqRob2 uses `lme4` to construct a linear-mixed model with random effect, but without fixed effect. The models specified by the variable formulas construct the models $y = Z_1\mu + \epsilon$ and $y = Z_2\mu + \epsilon$, where Z_1 contains the random effects condition, sample and feature (peptides for protein) and Z_2 contains only a random effect for the condition. Some proteins have only intensities from one

~~peptide. This can cause the first model $y = Z_1\mu + \epsilon$ parameters to fail to converge. For these proteins, we use the reduced model $y = Z_2\mu + \epsilon$.~~

Limma analysis on DIA-NN protein groups

We downloaded limma[23] version 3.50.3 from Bioconductor. We applied limma with default parameters to the protein groups inferred from DIA-NN in the PS pipeline.

Triqler

We downloaded Triqler from `https://github.com/statisticalbiotechnology/triqler`. We used Triqler v0.6.1 for the tests described in this paper. We selected a lower bound estimate for the `fold_change_eval` parameter as described in The&Käll,[24] which ended up as 0.76 for the spectral library data, and 0.51 for the pseudo-spectra enabled data.

As Triqler's model accounts for the assessed uncertainty in the prior identification steps of the data, it does not threshold the PSMs but instead takes all of the PSMs as input. The `searchScore` column should reflect increasing certainty in PSMs. Therefore we apply the negative log-transform to the `m_score` or `Q.Value` to indicate `searchScore` for the two different pipelines.

When evaluating Triqler we estimated a fold-change valuation thresholds using a previously described heuristic that ensures that 99% of the probability distribution of the normal distribution is contained [24], as,

$$L = \frac{\log_2\left(10^{\sigma_y}\right) \cdot \sqrt{2}}{\sqrt{N}} \cdot 2.5, \tag{1}$$

where σ_y is the standard deviation of the protein prior in $\log_{10}$ abundance units, N is the number of samples in the group. Triqler estimates and prints this minimum advisable fold change threshold by default, and it can be retrieved from Triqlers standard output. Also, Triqler removes any peptides that are attributed to more than one protein.

Multiple Hypothesis Correction

There was a slight difference in some of the benchmarking metrics, as the multiple test correction is performed with q value for Triqler and Top3, while MsStats and MSqRob2 use Benjamini-Hochberger[25] corrections. The q value approach aims to give an unbiased estimator of FDR, while the Benjamini-Hochberger approach estimates the upper bound of the FDR (and will therefore result in an FDR equal to or higher than the q value). As a consequence, the statistics from MSstats and MSqRob2 should be more conservative than the estimates from Triqler and Top3.[26]

Results

To establish that Triqler is useful when evaluating DIA data, we first examined the DIA data to establish that the abundance values derived from such processes are in line with the assumptions that Triqler makes about input data. Second, we benchmarked Triqler against a set of state-of-the-art methods for protein summarization. For both tasks, we used selected parts of the LFQBench dataset, which we processed in two different ways. We used a pipeline with a DDA identification-driven (ID) spectral library, as well as one using a Pseudo-spectra generation (PS) spectral library (See Methods).

Validations of properties of DIA peptide abundance

DIA data is assumed to encompass a larger dynamic range than DDA data . This could affect one of Triqler's assumptions, that the noise structure is mainly multiplicative, i.e. that the standard deviation within a sample group is proportional to its mean. When investigating all the peptide abundance measurements at a 1% PSM-level identification false discovery rate (FDR) from the TripleTOF6600 section

of the LFQBench dataset, we found a relatively linear relationship between standard deviation and mean (Supplementary Figure S1A-D).

However, when ~~Triqler tried to estimate~~ estimating the distribution of missing values with Triqler, we noted that the lower number of missing values in DIA compared to DDA (~~¡2%~~ < 2% for DIA), ~~lead~~ led to parameter estimates out of the expected range. We hence introduced an alternative estimation procedure for DIA data. Instead of fitting a censored normal distribution function over all XIC values as for DDA data, we use the mean value of the missing values to fit a censored normal distribution function (See Supplementary Figure S2)~~that can be~~. The alternative curve fitting procedure is enabled by a command line option in Triqler ~~.~~ named `--missingValuePrior=DIA`

Harmonization of protein inference procedures.

Encouraged by the finding that peptide abundances from DIA data have similar properties to those of DDA data, we wanted to see how well Triqler works in practice. Particularly, we wanted to compare the performance of Triqler to that of other protein summarization methods. However, to do so we first needed to establish some principles for how to benchmark summarization methods on data from whole-cell mixtures, when comparing methods that use different protein inference procedures.[27] When reporting the number of differentially abundant proteins in data sets from mixtures from whole-cell extracts, a protein inference scheme that infers any protein containing a detected peptide will report more differentially abundant proteins than a more restrictive scheme that just reports a parsimonious set of proteins. There are no restrictive mechanisms detecting situations where non-present proteoforms are reported as long as they are reported with protein abundance rates compatible with the right proteome. To alleviate, or at least minimize, this problem from our comparison, we restricted the searched sequence database by removing proteins with shared peptides in such a manner that no two proteins shared a tryptic peptide longer than 7 amino acids. This operation aims at a fairer comparison of protein summarization regardless of the protein inference method.

Distributions of the estimated fold changes

We applied Triqler, MSstats, MSqRob2, and Top3 (see Methods) to the peptide abundances derived from the ID and PS workflows from the varying concentrations of *E. Coli* and yeast concentrations in a background of HeLa-cells of the LFQBench set. Although Triqler is not meant as a tool to give point estimates, we here used Triqler's maximum a ~~posterioris~~ posteriori probability (MAPs) as point estimates of protein abundance to enable comparisons to other methods. Also, for the comparisons, we removed the methods' fold change selection mechanisms. For an overview of the results, we made histograms of the methods' estimated protein-level fold changes (Figure 1 and Supplementary Figure S7),

Triqler and Top3 ~~appears~~ appear to have less fold change bias than MSstats and MSqRob2, that is, the apex of the distributions is centered more closely to the lysate mixture rates. MSstats and MSqRob2 have distribution apexes that, for each species, are centered at lower fold changes than the true lysate mixture rates. Reprocessing the data with replaced labels led to the reversed result, i.e. the apexes centered around values higher than the pipetted mixture rates for both MSstats and MSqRob2 (data not shown). We also observe that the apex of the distributions has higher values for Triqler than the other methods in Figure 1. We provide comparisons of the fold change estimates in Figure ~~??~~ 2 and Supplementary Figure S3.

~~Triqler has~~ As visible in Figure 1, Triqler's fold change distributions have longer tails for the estimated yeast and *E. Coli* ~~fold change distributions~~than the compared methods. These values are explained by proteins quantified with ~~fewer peptides, i.e. less certain protein quantification~~less certainty. Proteins with correctly estimated fold changes are identified by more peptides than those differing significantly from their expected fold change (data not shown). The technical aspect of this is explained by how Triqler computes the posteriors for differentially abundant proteins; with better data, more confidence is put on data and less confidence is put on the prior distributions, resulting in narrower fold change distributions with means more centered at the observed mean values.

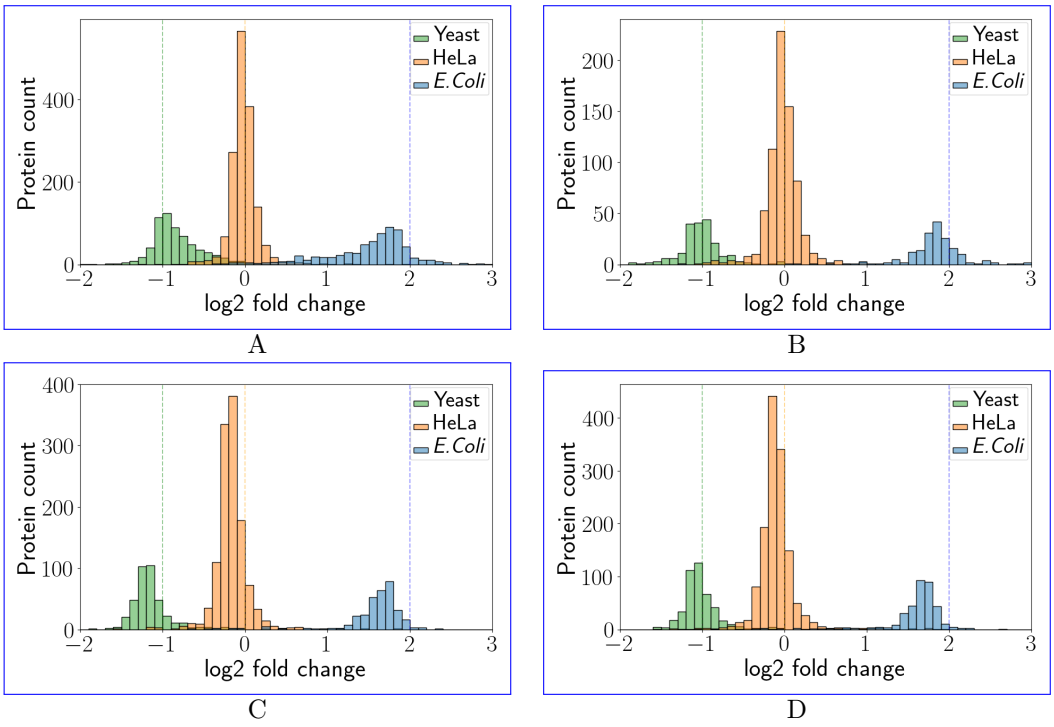


Figure 1: **Comparison of reported fold change distributions.** The summarized protein abundances from the ID spectral library pipeline for (A) Triqler, (B) Top3, (C) MSstats, and (D) MSqRob2 were binned and plotted as histograms. The dashed lines indicate the pipetted log2-fold change ratio between a specie and HeLa samples. We noted that the empirical densities of the protein counts are less biased for Triqler and Top3 than MSstats and MSqRob2, as the apex of their distribution was found closer to the true fold change difference (see Supplementary Figure S7 for reported fold change distributions for both ID and PS workflows). For these histograms, we included all the summarized proteins and did not remove any proteins based on the significance or fold change thresholds.

Comparison of ability to discriminate differentially from equivalently abundant proteins

As the first quantitative test of performance, we compared the methods' reported number of differentially abundant *E. Coli* and yeast proteins as a function of the number of HeLa proteins (Figure 3). As the former two lysates were injected in different concentrations and the HeLa was at constant concentration over the sample groups, a higher number of non-HeLa proteins for a similar number of HeLa-proteins is seen as better performance. Overall, Triqler reports more ~~differentially abundant protein for every non-differentially abundant protein~~ such expected differentially abundant non-HeLa proteins per HeLa protein than the compared methods for both the peptide abundances generated by the ID spectral library and PS spectral library pipelines. The results can be found in Supplementary Figure S4. We also ran a second experiment where we removed shared peptides after matching, to ensure that the order of removal of the shared peptides does not affect performance. However, we found no large differences between removing peptides before or after database searching (Supplementary Figure S6).

Comparison of statistical calibration

Further, we tested the statistical calibration of the summarization methods. We hence investigated the relationship between the fraction of wrongly reported ~~differential~~ differentially abundant proteins (i.e. the fraction of HeLa proteins among all reported ~~differential~~ differentially abundant proteins), and each inference method's estimated false discovery rate (See Figure 4 and Figure S8). We observed that Triqler was better calibrated (closer to the diagonal line) than MSstats and MSqRob2. Top3 showed an even

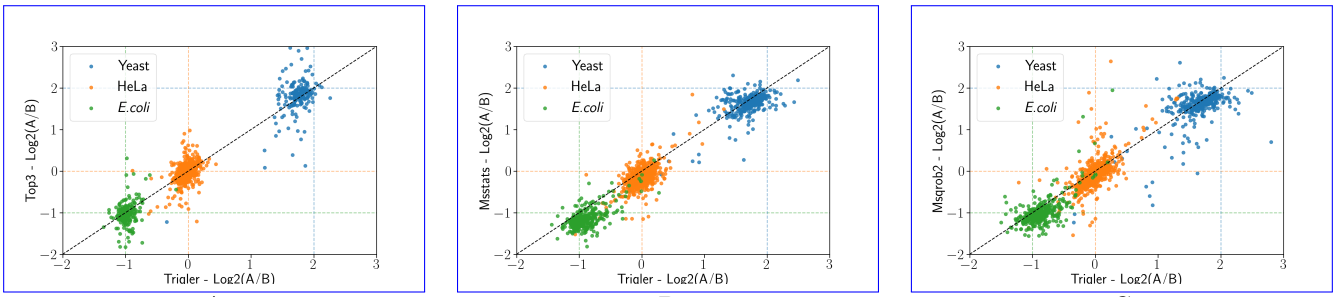


Figure 2: **Comparison of the summarized protein fold changes between the methods.** The comparison between the distributions of the MAP estimates from Triqler and (A) Top3, (B) MSstats, and (C) MSqRob2 for the data in the ID pipeline. The actual pipetted fold changes are indicated by dashed lines. Proteins not quantified by both compared methods were excluded from the plots. See Supplementary Figure ~~??~~ S3 for protein-level results for both ID and PS workflows.

better calibration than Triqler but, as previously demonstrated, has a much lower sensitivity.

Benchmark on cancer proteomics-related data

To further compare the different protein inference methods, they were finally tested on two sets of less engineered data — sets without known relative abundance differences. Data was downloaded from Stewart *et al.*, a set consisting of DIA data contrasting lung squamous cell carcinomas to adjacent tissues [14], and from Gao *et al.* a set comparing primary to metastatic cell lines[15]. Peptide abundances were derived with the nf-core quantms pipeline followed by our protein summarization strategies and plotted the number of differentially abundant proteins as a function of each method's estimated quantitative protein-level FDR of q value. We used fold change evaluation of 0.92 resp. 0.63 for Triqler, based on its lower bound estimate. We used the the same lower bound estimate as a fold change threshold when reporting performance of the compared methods. As can be seen in Figure 5, Triqler reports more differential proteins than the other methods for similair estimated q values.

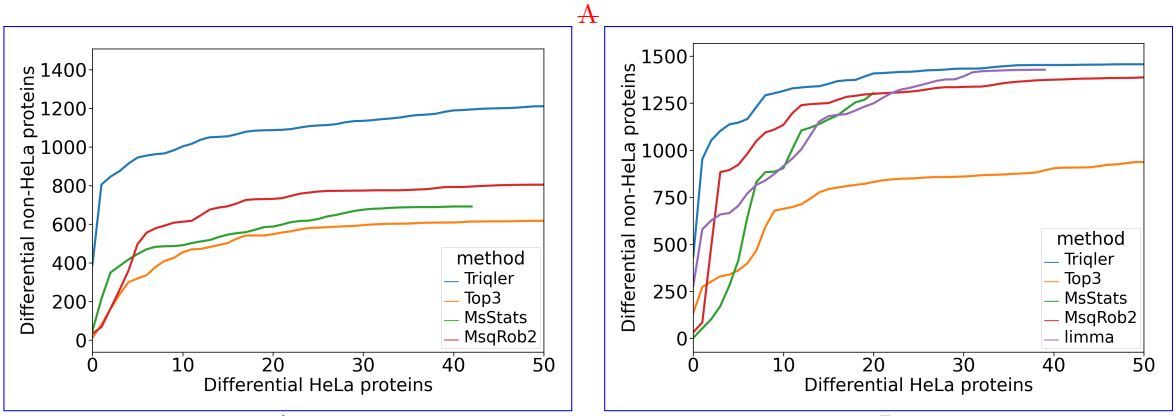
Discussion

Here we have shown that Triqler operates well for DIA data, despite being originally intended for DDA data. We also find that Triqler outperforms other protein summarization methods on an engineered benchmark set, both in terms of sensitivity and accuracy in its error estimates. Triqler was also able to detect a higher number of differentially abundant proteins at a more accurately reported false discovery rate than the compared methods. The absence of filtering and imputation steps before Triqler benefits the analysis by making it both more user-friendly by reducing parameter choices and inducing less bias into the result.

The analytes in shotgun proteomics are peptides and not proteins or proteoforms. Nevertheless, most users of mass spectrometry use and will continue to find reasons to report findings on a protein level. It makes sense to put efforts into a better understanding of which protein inference tools to use at what occasion and how to summarize peptide abundances into protein relative concentration values. Also, protein summarization gives lower variance than peptide-level analysis, and it reduces the number of hypotheses tested and reduced the number of missing values, which can have a major impact on the quality of the analysis.?

One important remark is that the sequence database that we used for matching the spectra was filtered so that only one protein per peptide was kept, to control for any difference in protein inference strategies used by our compared protein summarization methods. There is currently no consensus on how to handle multiple proteoforms in bottom-up proteomics. Hence, we believe that protein inference strategies that can

~~A~~



A B

Figure 3: **Comparison of ability to differentiate proteins with ~~differential~~ differentially abundance between conditions.** We plotted the number of reported differentially abundant *E. Coli* and yeast proteins as a function of the number of proteins from the HeLa background when sorting according to significance for (A) ID pipeline and (B) PS pipeline. For the test, we selected a fold change evaluation of 0.51 for Triqler and a fold change threshold of 0.51 for Top3, MSstats and MSqRob2. ~~All methods had a protein-level FDR threshold of 0.01. (~~See the Supplement ~~Figure~~ Figures S4~~shows differential abundance -S6 for the number of differentially abundant proteins~~ for each ~~specie)~~proteome.

account for multiple proteoforms would greatly benefit the field by improving the quality of the quantitative analysis.

We also see some differences in how DIA and DDA peptide-level abundance data appear. For instance, there are more missing values in the DDA than in the DIA data. We addressed this issue by providing an alternative method for estimating the censoring function for missing values in Triqler.

It is quite hard to evaluate how the performance of a data processing pipeline is influenced by its components. This should not stop the field from trying to establish the features of the different processing steps.[6,28,29] Unbiased comparisons of software tools are challenging for several reasons.[28] Methods can be assessed by scientists lacking relevant expertise, the tested methods may be lacking sufficient documentation and the interpretation of test results may be subjective.[30–33] By using the same data set we can assure that the data set is processed consistently and further the analysis by extending it to protein summarization procedures.

Lastly, we want to highlight the benefit and importance of datasets such as the one provided by Navarro et al.[6] These benchmarking datasets make it easy for the scientific community to investigate computational tools by providing a golden standard and significantly facilitating benchmark studies.

Supporting Information

The following supporting information is available free of charge at ACS website `http://pubs.acs.org`

- Table S1: Protein count in the Uniprot FASTA protein database.
- Table S2: Number of identified peptides and proteins for the ID workflow.
- Table S3: Number of identified peptides and proteins for the PS workflow
- Figure S1: Standard deviation of peptide abundance as a function of mean abundance in DIA experiments.
- Figure S2: Comparison of actual missing values against fit to the censoring distribution used in Triqler.

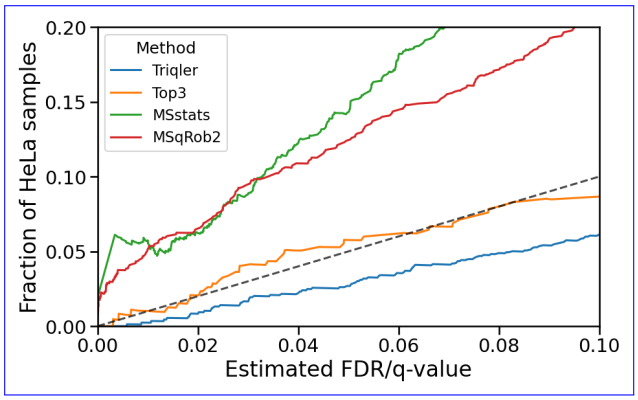


Figure 4: **Comparison of calibration of the compared summarization methods.** We plotted the fraction of reported differentially abundant HeLa proteins as a function of the q value threshold for protein abundances without any restrictions on the fold change. We used q value for Triqler and Top3, and Benjamini-Hochberg corrections for MSstats and MSqRob2.

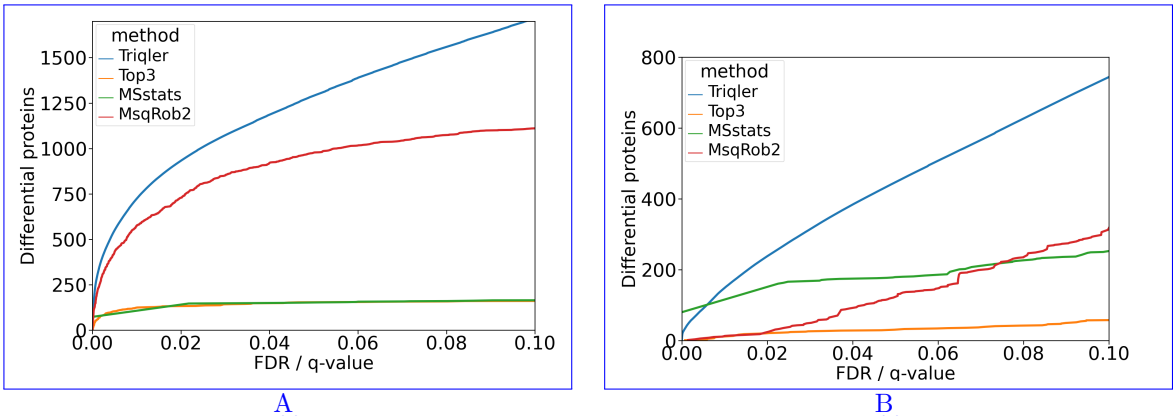


Figure 5: **Reported performance of the compared methods on two cancer proteomics-related datasets.** We plotted the reported differentially abundant proteins as a function of their reported FDR/q value for the (A) Stewart *et al.* set (lung squamous cell carcinomas vs. adjacent tissues) and (B) Gao *et al.* (primary vs. metastatic cell lines).

- Figure S3: Comparison of reported log2-fold change between Triqler and the compared methods.
- Figure S4: The compared methods' ability to distinguish differentially abundant proteins when applying protein-level fold-change tresholds.
- Figure S5: The compared methods' ability to distinguish differentially abundant proteins when not applying fold change treshholds.
- Figure S6: The compared methods' ability to distinguish differentially abundant proteins when removing shared peptides after database matching.
- Figure S7: Comparison of reported fold change distributions.
- Figure S8: Comparison of calibration of the compared summarization methods.

Funding

This work was supported by grants from the Swedish Research Council (grant 2017-04030).

~~Supporting information~~Acknowledgements

We would like to thank Yasset Perez-Riverol, EMBL-EBI for helpful discussions.

References

[1] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical Chemistry*, 389(4):1017–1031, 2007.

[2] Lukas Käll and Olga Vitek. Computational mass spectrometry–based proteomics. *PLoS Computational Biology*, 7(12):e1002277, 2011.

[3] Pavel V Bondarenko, Dirk Chelius, and Thomas A Shaler. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Analytical Chemistry*, 74(18):4741–4749, 2002.

[4] Matthew The and Lukas Käll. Integrated identification and quantification error probabilities for shotgun proteomics. *Molecular & Cellular Proteomics*, 18(3):561–570, 2019.

[5] John D Venable, Meng-Qiu Dong, James Wohlschlegel, Andrew Dillin, and John R Yates. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, 1(1):39–45, 2004.

[6] Pedro Navarro, Jörg Kuharev, Ludovic C Gillet, Oliver M Bernhardt, Brendan MacLean, Hannes L Röst, Stephen A Tate, Chih-Chiang Tsou, Lukas Reiter, Ute Distler, et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nature biotechnology*, 34(11):1130–1136, 2016.

[7] Ludger JE Goeminne, Andrea Argentini, Lennart Martens, and Lieven Clement. Summarization vs peptide-based models in label-free quantitative proteomics: performance, pitfalls, and data analysis guidelines. *Journal of Proteome Research*, 14(6):2457–2465, 2015.

[8] Jeffrey C Silva, Marc V Gorenstein, Guo-Zhong Li, Johannes PC Vissers, and Scott J Geromanos. Absolute quantification of proteins by LCMSE: A virtue of parallel MS acquisitions. *Molecular & Cellular Proteomics*, 5(1):144–156, 2006.

[9] Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics*, 13(9):2513–2526, 2014.

[10] Jenny Forshed, Henrik J Johansson, Maria Pernemalm, Rui MM Branca, AnnSofi Sandberg, and Janne Lehtiö. Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (pqpq). *Molecular & Cellular Proteomics*, 10(10), 2011.

[11] Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, and Olga Vitek. Msstats: an r package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524–2526, 2014.

[12] Bo Zhang, Mohammad Pirmoradian, Roman Zubarev, and Lukas Käll. Covariation of peptide abundances accurately reflects protein concentration differences. *Molecular & Cellular Proteomics*, 16(5):936–948, 2017.

[13] Adriaan Sticker, Ludger Goeminne, Lennart Martens, and Lieven Clement. Robust summarization and inference in proteome-wide label-free quantification. *Molecular & Cellular Proteomics*, 19(7):1209–1219, 2020.

[14] Paul A Stewart, Bin Fang, Robbert JC Slebos, Guolin Zhang, Adam L Borne, Katherine Fellows, Jamie K Teer, Y Ann Chen, Eric Welsh, Steven A Eschrich, et al. Relative protein quantification and accessible biology in lung tumor proteomes from four lc-ms/ms discovery platforms. *Proteomics*, 17(6):1600300, 2017.

[15] Erli Gao, Wenxue Li, Chongde Wu, Wenguang Shao, Yi Di, and Yansheng Liu. Data-independent acquisition-based proteome and phosphoproteome profiling across six melanoma cell lines reveals determinants of proteotypes. *Molecular Omics*, 17(3):413–425, 2021.

[16] Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods*, 14(5):513–520, 2017.

[17] EasyPQP:simple library generation for openswath. `https://github.com/grosenberger/easypqp`. Accessed: 2021-09-03.

[18] Johan Teleman, Hannes L Röst, George Rosenberger, Uwe Schmitt, Lars Malmström, Johan Malmström, and Fredrik Levander. Diana—algorithmic improvements for analysis of data-independent acquisition ms data. *Bioinformatics*, 31(4):555–562, 2015.

[19] Peter Blattmann, Moritz Heusel, and Ruedi Aebersold. Swath2stats: an r/bioconductor package to process and convert quantitative swath-ms proteomics data for downstream analysis tools. *PloS one*, 11(4):e0153160, 2016.

[20] Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, and Markus Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1):41–44, 2020.

[21] Niels Hulstaert, Jim Shofstahl, Timo Sachsenberg, Mathias Walzer, Harald Barsnes, Lennart Martens, and Yasset Perez-Riverol. Thermorawfileparser: modular, scalable, and cross-platform raw file conversion. *Journal of Proteome Research*, 19(1):537–542, 2019.

[22] Viktor Granholm, José Fernández Navarro, William Stafford Noble, and Lukas Käll. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *Journal of Proteomics*, 80:123–131, 2013.

[23] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 01 2015.

[24] Matthew The and Lukas Käll. Triqler for MaxQuant: Enhancing results from MaxQuant by Bayesian error propagation and integration. *Journal of Proteome Research*, 20(4):2062–2068, 2021.

[25] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[26] Keegan Korthauer, Patrick K Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J Alm, and Stephanie C Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1):1–21, 2019.

[27] Oliver Serang, Luminita Moruz, Michael R Hoopmann, and Lukas Käll. Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. *Journal of Proteome Research*, 11(12):5586–5591, 2012.

[28] Craig Dufresne, David Hawke, Alexander R Ivanov, Antonius Koller, Brendan MacLean, Brett Phinney, Kristie Rose, Paul Rudnick, Brian Searle, Scott Shaffer, et al. Abrf research group development and characterization of a proteomics normalization standard consisting of 1,000 stable isotope labeled peptides. *Journal of Biomolecular Techniques: JBT*, 25(Suppl):S1, 2014.

[29] Laurent Gatto, Kasper D Hansen, Michael R Hoopmann, Henning Hermjakob, Oliver Kohlbacher, and Andreas Beyer. Testing and validation of computational methods for mass spectrometry. *Journal of Proteome Research*, 15(3):809–814, 2016.

[30] John R Yates, Sung Kyu Robin Park, Claire M Delahunty, Tao Xu, Jeffrey N Savas, Daniel Cociorva, and Paulo Costa Carvalho. Toward objective evaluation of proteomic algorithms. *Nature Methods*, 9(5):455–456, 2012.

[31] Felipe da Veiga Leprevost, Valmir C Barbosa, Eduardo L Francisco, Yasset Perez-Riverol, and Paulo C Carvalho. On best practices in the development of bioinformatics software. *Frontiers in genetics*, 5:199, 2014.

[32] Huisong Pak, Frederic Nikitin, Florent Gluck, Frederique Lisacek, Alexander Scherl, and Markus Muller. Clustering and filtering tandem mass spectra acquired in data-independent mode. *Journal of The American Society for Mass Spectrometry*, 24(12):1862–1871, 2013.

[33] The difficulty of a fair comparison. *Nature Methods*, 12(273), 2015.