Label-free quantification with FDR-controlled match-between-runs

Fengchao Yu[1], Sarah E. Haynes[1], Alexey I. Nesvizhskii[1,2]*

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

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

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

Running title:

match-between-runs with false discovery rate control

Abbreviations:

LC-MS: liquid chromatography-mass spectrometry

DDA: data-dependent acquisition

DIA: data-independent acquisition

MBR: match-between-runs

FDR: false discovery rate

LDA: linear discriminant analysis

EM: expectation-maximization

LFQ: of label-free quantification

CV: coefficient of variation

Abstract

Missing values weaken the power of label-free quantitative proteomic experiments to uncover true quantitative differences between biological samples or experimental conditions. Match-between-runs (MBR) has become a common approach to mitigate the missing value problem, where peptides identified by tandem mass spectra in one run are transferred to another by inference based on m/z, charge state, retention time, and ion mobility when applicable. Though tolerances are used to ensure such transferred identifications are reasonably located and meet certain quality thresholds, little work has been done to evaluate the statistical confidence of MBR. Here, we present a mixture model-based approach to estimate the false discovery rate (FDR) of peptide and protein identification transfer, which we implement in the label-free quantification tool IonQuant. Using several benchmarking datasets generated on both Orbitrap and timsTOF mass spectrometers, we demonstrate that IonQuant with FDR-controlled MBR results in superior performance compared to MaxQuant. We further illustrate the need for FDR-controlled MBR in sparse datasets such as those from single-cell proteomics experiments.

Introduction

Due to its sensitive and high-throughput nature, liquid chromatography-mass spectrometry (LC-MS) is a commonly used technology to identify and quantify peptides and proteins from complex samples. Various approaches to LC-MS data acquisition (1-3) have been developed, among which data-dependent acquisition (DDA) remains the most commonly used strategy (4). In the course of a DDA run, eluted peptides are introduced into a mass spectrometer, where peptide ions are sampled for fragmentation and identified from the resulting tandem mass (MS/MS) spectra. Precursor peptide ion intensities are assumed to be correlated with the actual peptide amount, yielding relative peptide and, after an additional peptide to protein roll-up step, protein quantification. Peptide ions successfully targeted and identified by MS/MS are used to calculate peptide and then protein abundances. However, due to the stochastic nature of intensity-based sampling of peptide ions for MS/MS analysis, not all peptides are consistently identified in all runs. This in turn gives rise to missing quantification values, weakening essential comparisons between different biological samples or experimental conditions. The prevalence of missing values in DDA proteomics is generally higher than that in genomics or transcriptomics. The issue of missing data can be alleviated to some degree using the data-independent acquisition (DIA) strategy (5-8). However, as label-free quantification using DDA data remains popular, there is a critical need to improve computational solutions for this method.

To address the missing value problem in DDA-based proteomics, a number of “identification transfer” approaches have been devised (9-12), exemplified by the match-between-runs (MBR) option of MaxQuant (13, 14) that allows “transfer” of identified precursor peptide peaks from one run (referred to below as donor run) to another (acceptor). Given a peak identified by MS/MS in the donor run, attributes such as m/z, charge state, and retention time, are used to locate a corresponding peak in the acceptor run that is most likely the same peptide. The intensity of the

donor peak is then assigned to the acceptor peak, thus filling in the missing value. With more quantified features in common between runs, a greater number of peptides and proteins can be compared among different runs and experiments, increasing the depth of experimental findings (15-17).

While the goal of MBR is to mitigate the missing value problem, it has the potential to introduce false positives, as transferred peaks have not been rigorously identified using MS/MS spectra in the acceptor run. Lim et al. (18) evaluated the false transfer rate of MBR using a two-organism dataset. They concluded that there was a considerable proportion of false positives from MBR when using MaxQuant, yet most were removed with additional filtering as part of the LFQ calculations. However, in practical settings, even with the additional filtering, FDR of MBR may still be unacceptably high. Thus, this subject deserves a more rigorous treatment that can be generalized across different samples and experimental designs. Here, we propose a supervised semi-parametric approach to control the FDR of MBR, extending our earlier work on FDR for protein identification (3, 19) and DIA quantification (20). We implement FDR-controlled MBR in IonQuant (21), which has been extended to support LC-MS data both with and without ion mobility. We also implement a new protein abundance calculation module in IonQuant, based on the MaxLFQ strategy (13), improving upon our previously described top-N approach. Using the dataset from Lim et al. (18), we reproduce the authors findings and demonstrate that IonQuant with FDR-controlled MBR has a lower false positive rate and higher sensitivity compared to MaxQuant. With two additional datasets from timsTOF Pro mass spectrometers, we demonstrate that FDR-controlled MBR results in higher quantification precision (lower CV), accuracy, and sensitivity. At last, we demonstrate that IonQuant displays high sensitivity and precision in single cell data, but that FDR control for MBR is crucial in such datasets. Overall, we propose an efficient approach to perform MBR with FDR control while maintaining high

accuracy and precision. We implement the new methods as a default option in IonQuant, readily available as a standalone tool or within our integrated computational platform FragPipe (https://fragpipe.nesvilab.org/).

Experimental Procedures

Experimental Design and Statistical Rationale

We used four datasets in this work. In all datasets, we estimated the identification false-discovery rate using the target-decoy approach (3). For MSFragger, PSMs, peptides, and proteins were filtered at 1% PSM and 1% protein identification FDR. For MaxQuant, PSMs and peptides were filtered at 1% PSM FDR, and proteins were filtered at 1% protein FDR, which is MaxQuant's default setting. A two-organism dataset (*H. sapiens* and *S. cerevisiae*) with 40 LC-MS runs from Lim et al. (18) was generated on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). In this dataset, 20 runs include only *H. sapiens* proteins, while the remaining 20 runs contain a mixture of *H. sapiens* and *S. cerevisiae* proteomes. *S. cerevisiae* peptides transferred to the 20 *H. sapiens*-only runs by MBR are false positives and were used to evaluate the false positive rate. We also employed two datasets from timsTOF Pro (Bruker), as in our previous work (21). A HeLa dataset with 4 replicate injections from Meier et al. (22) was used to evaluate the sensitivity (i.e., quantified protein count) and precision (i.e., coefficient of variation (CV)) of quantification across replicate runs. A three-organism timsTOF dataset (*H. sapiens*, *S. cerevisiae*, and *E. coli*) with 6 runs from Prianichnikov et al. (23) was used to evaluate quantification accuracy, and contains two experimental conditions with ground truth protein ratios: 1:1 (*H. sapiens*), 2:1 (*S. cerevisiae*), and 1:4 (*E. coli*). The final dataset used in this study contains 26 runs from the single-cell proteomics experiment published by Williams et al. (24) generated on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific).

This dataset contains 3 replicate runs with 0 cell (blank runs), 11 replicates with 1 cell, 4 replicates with 3 cells, 4 replicates with 10 cells, and 4 replicates with 50 cells. Numbers of quantified peptides and proteins were used to evaluate sensitivity, and quantification CV was used to evaluate precision.

Indexing-based MBR

We developed a fast MBR algorithm based on indexing. In IonQuant (21), an index of each run is built and written to the disk for fast feature extraction, which supports data with and without ion mobility information. Given a run with possible missing values that will accept ions (acceptor run) and a separate run that will be used to fill these missing values (donor run), correlations between the two runs are calculated using overlapped ions' retention times, intensities, and ion mobilities if applicable: $(o \times r_1 + o \times r_2)/2$ or $(o \times r_1 + o \times r_2 + o \times r_3)/3$, where o is the overlapping ratio (25); r_1, r_2, and r_3 are Spearman's rank correlation coefficients of retention time, intensity, and ion mobility, respectively. Up to *n* (user-specified 'MBR top runs' parameter, 3 by default) donor runs with the highest correlations (which must be greater than user-specified 'MBR min correlation' parameter, 0 by default) are selected.

For each ion in every selected donor run, we locate the target region within the acceptor run using an approach similar to FlashLFQ (26). First, pairs of retention times from the overlapping ions are collected and sorted according to the value from the donor run. Using d_i and a_i to denote the retention times of i-th pair of ions from the donor and acceptor runs, respectively, we have pairs from (d_1, a_1) to (d_N, a_N) sorted by d_i, where N is the number of overlapped ions. Given a donor ion with retention time t, we find its position in the sorted pairs satisfying $d_i \leq t < d_{i+1}$. Then, we collect all pairs satisfying $d_i - \tau \leq d_j \leq d_i + \tau$, where τ is a predefined

tolerance ('MBR RT window' parameter, 1 minute by default) . With those pairs, we generate a list whose elements are $a_j - d_j$, and calculate the median (m) and median absolute deviation (σ) of that list. The possible target range in the retention time dimension is then:

$$[d_i + m - 2\sigma, d_i + m + 2\sigma] \#(1)$$

If ion mobility data are used, we take the same approach to locate the target range in the ion mobility dimension (controlled by the 'MBR IM window' parameter, 0.05 by default). The transferred ion's m/z equals the donor ion's m/z adjusted by mass calibration error (mass calibration is performed by MSFragger (27)). After locating the target region in m/z, retention time, and ion mobility if applicable, we trace all peaks within the region using our recently described algorithm (21). Four isotope peaks (-2, -1, +1, and +2) are also traced to check the charge state and the isotope distribution. Peak boundaries are allowed to extend beyond the target region's retention time and ion mobility bounds. Peak tracing is performed rapidly using the index, after which the donor ion's peptide information is assigned to the traced monoisotopic peak.

MBR false discovery rate estimation

To estimate the rate at which false transfers occur, we adopted a supervised semi-parametric mixture model that we previously applied in a number of related applications (19, 20). For each successfully transferred donor ion (i.e., target ion), we try to transfer a decoy ion, created to have the same retention time and ion mobility (if applicable) but with a large m/z shift (28, 29). To generate a decoy, we first shift the m/z by +11×1.0005 Th. If there is no traceable peak in that region, we keep decreasing the m/z shift by 1.0005 Th until we successfully trace a peak or until the m/z shift reaches +4 Th.

For all transferred target and decoy ions, we calculate nine (without ion mobility) or ten (with ion mobility) scores (**Table 1**). For three of these scores (using the 0/+1/+2, -1/0/+1, and -2/-1/0 peaks), Kullback-Leibler divergence is used to compare the quality of the traced isotopic distribution to a theoretical one given m/z and charge state, where the Poisson distribution is used as theoretical (30).

We classify all transferred ions (identified with sequence, charge, and modification information) into four types: a target ion that has not been identified by MS/MS in the acceptor run (type 1); a decoy ion that is from a m/z-shifted type 1 ion (type -1); a target ion that has already been identified by MS/MS (type 2); or a decoy ion that is from a m/z-shifted type 2 ion (type -2). Following the strategy we previously used for DIA data (20), we train a linear discriminant analysis (LDA) model using scores from type 2 and -2 ions. From the trained LDA, we calculate a final score for each type 1 and -1 ion:

$$s = \sum_i w_i \, b_i \#(2)$$

where s is the final score, w_i are the weights from LDA, and b_i are the scores detailed in **Table 1**. If multiple ions were transferred to one location, the top scoring one is kept.

Using the final scores from type 1 and -1 ions, we estimate a posterior probability of correct identification transfer by fitting a mixture model:

$$f(s) = \pi_0 f_0(s) + \pi_1 f_1(s) \#(3)$$

where f_0 is the distribution of correctly transferred ions, f_1 is the distribution of incorrectly transferred ions, π_0 and π_1 are the respective priors of false and true transferred ions. We use the expectation-maximization (EM) algorithm (20) to estimate the coefficients and distributions in Equation (3).

After fitting the mixture model, we calculate a posterior probability for each transferred ion using

$$p(s_i) = \frac{\pi_1 f_1(s_i)}{\pi_0 f_0(s_i) + \pi_1 f_1(s_i)} \#(4)$$

where s_i is the score of the transferred ion. Then, we calculate an ion-level MBR FDR using the posterior probability (31) of type 1 ions:

$$\widehat{FDR}(t) = \frac{\sum_{s_i \geq t}\left(1 - p(s_i)\right)}{\sum_i \mathbf{1}_{s_{i \geq t}}} \#(5)$$

where t is a score threshold and $\sum_i \mathbf{1}_{s_{i \geq t}}$ is the number of type 1 ions whose score is larger than t. We can also calculate peptide- and protein-level FDR for MBR by collapsing ions with the same sequence or protein and using the highest probability entry in the FDR calculation.

Calculating protein intensity using MaxLFQ algorithm

Cox et al. proposed MaxLFQ (13) algorithm to calculate protein intensity with peptide intensities. It has a high precision (low CV) according to our previous study (21). We implemented it in IonQuant to provide a new (default) option in addition to the top-N approach.

Given a study with *N* experiments (samples), and a protein with *M* quantified peptide ions, for each peptide ion $p \in [1, M]$ we calculate a log-ratio of its intensities between experiments i and j:

$$r_{i,j}(p) = \log \frac{I_i(p)}{I_j(p)} = \log I_i(p) - \log I_j(p) \#(6)$$

where $I_i(p)$ is the intensity of peptide ion p from i-th experiment. If the ion is not quantified in experiment i or j, we do not calculate the corresponding log ratio. Then, we have a linear relationship among the log-transformed protein intensities and their peptide ion log-ratios:

$$x_i - x_j = m_{i,j} \#(7)$$

where x_i is the (unknown) log-transformed protein intensity in i-th experiment and $m_{i,j}$ is the median of the log-ratios $r_{i,j}(p)$ among all peptide ions p from 1 to M. Given the set of 1 to N experiments, Equation (7) can be expressed in a matrix form

$$\boldsymbol{Ax} = \boldsymbol{b} \#(8)$$

where

$$\boldsymbol{A}_{i,j} = \begin{cases} -1 & (i \neq j) \\ \sum_{i=1}^{N-1} \mathbf{1}(i,j) & (i = j) \end{cases} \#(9)$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

$$\boldsymbol{b_i} = \begin{cases} \sum_{j=i+1}^{N} m_{i,j} & (i = 1) \\ \sum_{j=i+1}^{N} m_{i,j} - \sum_{j=1}^{i} m_{j,i} & (i > 1) \end{cases}$$

In Equation (9), $\mathbf{1}(i,j)$ equals 1 if there is a peptide ion quantified in both experiment i and j, and 0 otherwise. Equation (8) can be efficiently solved with Cholesky decomposition to get the log-transformed protein intensity x_i. Then, the protein intensity in experiment i equals e^{x_i}.

Validation of the FDR for MBR approach using two-organism dataset

We used 40 runs from Lim et al. (18) (ProteomeXchange (32) identifier PXD014415) to evaluate the sensitivity and precision of FDR-controlled MBR. This dataset contains 20 runs with only *H. sapiens* proteins and 20 with a mixture of *H. sapiens* (90%) and *S. cerevisiae* (10%) proteins, all acquired on an Orbitrap Fusion Lumos mass spectrometer. Further sample preparation and data acquisition details can be found in the original publication (18). We used FragPipe (version 13.0) with MSFragger (33) (version 3.0), Philosopher (34) (version 3.2.7), and IonQuant (21) (version 1.4.4) to analyze this dataset. For this analysis pipeline, raw spectral files were first converted to mzML using ProteoWizard (version 3.0.20066) with vendor's peak picking. We used MaxQuant (35) (version 1.6.14.0) for comparison, using raw spectral files as MaxQuant is optimized for vendor formats. A protein sequence database of reviewed *H. sapiens* (UP000005640) and *S. cerevisiae* (UP000002311) from UniProt (36) (reviewed sequences only; downloaded on Jan. 15, 2020) and common contaminant proteins (26448 proteins total) was used. For the MSFragger analysis, precursor and (initial) fragment mass tolerance were set to 50 ppm and 20 ppm, respectively. Reversed protein sequences were appended to the original database as decoys. Mass calibration and parameter optimization were enabled. Isotope error was set to 0/1/2, and one missed trypsin cleavage was allowed. Peptide length was set from 7 to 50, and peptide mass was set to 500 to 5000 Da. Oxidation of methionine and acetylation of protein-N termini were set as variable modifications. Carbamidomethylation of cysteine was set as a fixed modification. Maximum allowed variable modifications per peptide was set to 3. Philosopher (34) with PeptideProphet (37) and ProteinProphet (38) was used to estimate identification FDR. The PSMs were filtered at 1% PSM and 1% protein identification FDR. Quantification and MBR was performed with IonQuant. Minimum number of ions parameter required for quantifying a protein was set to 2 (default). To test the performance of FDR control for MBR, the maximum number of runs used for transfer was set to 40, and the minimum required correlation between the donor and acceptor run was set to 0. Ion-, peptide-, and protein-level MBR FDR thresholds were all set to 1% unless otherwise noted. Protein intensities

were computed using the re-implementation of MaxLFQ protein intensity calculation algorithm described above. Default values were used for all remaining parameters. For MaxQuant comparisons, the parameters were set as close to those described above as possible, with maximum modifications per peptide set to 3, maximum missed cleavages set to 1, LFQ enabled with default settings, maximum peptide mass set to 5000, built-in contaminant proteins were not used, and the second peptide option was not used. Default values were used for all remaining MaxQuant parameters.

We classified a peptide as *S. cerevisiae* peptide if it only maps to *S. cerevisiae* proteins. We classified a peptide as *H. sapiens* if it maps to at least one *H. sapiens* protein. The classification was done based on the protein name in the searched protein sequence database: those ending with "_HUMAN" were classified as *H. sapiens* protein and those ending with "_YEAST" were classified as *S. cerevisiae* protein.

Quantification precision

We used four replicate HeLa cell lysate runs acquired on a timsTOF Pro mass spectrometer (22) with 100 ms TIMS accumulation time to evaluate quantification precision when MBR is used. As in the previous section, we used FragPipe (version 13.0) with MSFragger (version 3.0), Philosopher (version 3.2.7), and IonQuant (version 1.4.4) to analyze this dataset. MaxQuant (version 1.6.14.0) was used to perform a benchmark comparison. Raw spectral files (.d extension) were used. The sequence database contained reviewed *H. sapiens* (UP000005640) proteins and common contaminants from UniProt (downloaded on Sep. 30, 2019; 20463 sequences). Minimum number of ions parameter required for quantifying a protein was set to 2 unless otherwise noted. For MBR in IonQuant, MBR top runs parameter was set to 3 and MBR

min correlation was set to 0. Ion-, peptide-, and protein-level MBR FDR threshold were set to 1%. Remaining parameters were identical to those in the previous section. We used the number of quantified proteins (considering proteins quantified in at least two runs) and quantification CV across replicates to evaluate the performance.

Quantification accuracy

We used the three-organisms dataset by Prianichnikov et al. (23) to demonstrate the accuracy of IonQuant with MBR. There are six runs from two experimental conditions in which *H. sapiens*, *S. cerevisiae*, and *E. coli* proteins are mixed at known ratios. The ratios between conditions are 1:1 (*H. sapiens*), 2:1 (*S. cerevisiae*), and 1:4 (*E. coli*). These data were acquired on a timsTOF Pro mass spectrometer, and details of the sample preparation and data generation can be found in the original publication (23). We used FragPipe (version 13.0) with MSFragger (version 3.0), Philosopher (version 3.2.7), and IonQuant (version 1.4.4) to analyze the data. MaxQuant results published by Prianichnikov et al. (23) were used as a benchmark comparison. Using the latest MaxQuant (version 1.6.14.0), a reviewed UniProt protein sequence database, and parameters closest to those of MSFragger and IonQuant yielded results similar to those in the original publication (**Supporting Figure S1**). A combined database of reviewed *H. sapiens* (UP000005640), *S. cerevisiae* (UP000002311), and *E. coli* (UP000000625) sequences from UniProt (30788 sequences downloaded Apr. 18, 2020) was used. Ion-, peptide-, and protein-level MBR FDR thresholds were set to 1%. Minimum number of ions parameter required for quantifying a protein was set to 2. Allowed missed cleavages was set to 2, and all other parameters were the same as those in the previous section. We used LFQbench (39) to plot the protein quantification results.

Single-cell dataset

We used 26 runs published by Williams et al. (24) to demonstrate IonQuant's performance with single-cell data. There are 3 replicates from 0 cell sample served as negative control, 11 replicates from 1 cell sample, 4 replicates from 3 cells sample, 4 replicates from 10 cells sample, and 4 replicates from 50 cells sample. The data was generated on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific), with MS/MS spectra acquired in the ion trap, 30 minute LC gradient. Details of the sample preparation and data acquisition can be found in Williams et al. (24). The raw data was converted to mzML format using ProteoWizard (version 3.0.19302) with vendor's peak picking. We used FragPipe (version 13.0) with MSFragger (version 3.0), Philosopher (version 3.2.7), and IonQuant (version 1.4.4) to analyze the data. We also used MaxQuant (version 1.6.14.0) as a benchmark. The database was downloaded along with the data (20129 proteins, ProteomeXchange (32) identifier MSV000085230). In MSFragger analysis, common contaminants and reversed protein sequences were appended by Philosopher. In MaxQuant analysis, the built-in contaminants were used. The precursor mass tolerance was set to 20 ppm, and the initial fragment mass tolerance was set to 0.6 Da. Two missed cleavages were allowed. IonQuant (version 1.4.4) with and without MBR was used. The MBR top runs parameter for MBR transfer was set to 26 and the minimum required correlation was kept at 0. MaxLFQ protein intensity calculation algorithm was used. Minimum number of ions parameter required for quantifying a protein was set to 1. Various ion-level MBR FDR thresholds were applied. The rest of the parameters are the same as those used in the previous section. MaxQuant's parameters were set as close as possible to those used in MSFragger and IonQuant. We used the numbers of peptides and proteins to evaluate the sensitivity, and we used CV to evaluate the precision of label free quantification with MBR.

Run time comparison

We used the two-organism dataset with 40 Orbitrap Fusion Lumos runs and the HeLa dataset with 4 timsTOF Pro runs to demonstrate the speed of label-free quantification coupled with FDR-controlled MBR in IonQuant (version 1.4.4). MaxQuant (version 1.6.14.0) was used for comparison. For the two-organism dataset, we used a combined database of reviewed *H. sapiens* (UP000005640) and *S. cerevisiae* (UP000002311) sequences from UniProt (36) plus common contaminants (26448 proteins downloaded Jan. 15, 2020). For the HeLa dataset, a database of reviewed *H. sapiens* (UP000005640) proteins from UniProt (20463 proteins downloaded on Sep. 30, 2019) and common contaminants was used. Reversed proteins sequences were appended to both databases as decoys for MSFragger analysis. All other parameters are identical to those used in the previous section. All analyses were run on a desktop with 4 CPU cores (Intel Xeon E5-1620 v3, 3.5 GHz, 8 logical cores) and 128 GB memory. We isolated quantification-specific run times from MaxQuant log files.

Results and Discussion

FDR-controlled MBR

We developed an MBR module in IonQuant enabling accurate and fast label-free quantification with match-between-runs peptide ion transfer with the help of the indexing functionality in IonQuant (see **Figure 1** for an overview). For each experiment (acceptor run) in the analysis, ion-level Spearman's rank correlation coefficients with all other experiments are calculated, where ion is defined as the combination of peptide sequence, modification pattern, and charge state. The percentage of ions overlapping between two runs is used as a weight in the calculation (25). For each acceptor run, IonQuant picks the top N runs with correlation larger than a certain threshold as donor runs. Both parameters ('MBR top runs' and "MBR min correlation' can be adjusted by the user). Given an ion from a donor run, IonQuant locates a

region in the acceptor run where the transferred ion is likely to be using m/z, retention time, and ion mobility (if applicable) distributions from both runs (see **Figure 1** and **Experimental Procedures**). For simplicity, we use retention time to describe the region-finding process. Given an ion from a donor run, all ions within a predefined retention time tolerance are collected. Retention time differences from pairs of ions overlapping between the runs are calculated, and the median and median absolute deviation of these differences are found. Then the region for transfer is determined using Equation (1). We use the same approach to locate the ion mobility region. After getting a 1-D (without ion mobility) or 2-D (with ion mobility) region, IonQuant traces peaks using the donor ion's m/z, taking any mass calibration correction into account. In addition to the monoisotopic peak, four additional isotope peaks (-2, -1, +1, and +2) are also included in peak tracing so that the isotopic distribution and charge state can be used in evaluation. Finally, IonQuant assigns the donor ion's peptide to each traced peak and calculates nine (without ion mobility) or ten (with ion mobility) scores (**Table 1**) measuring the quality of the peptide ion transfer.

In conventional MBR, most notably in MaxQuant, ions matching tolerance criteria are transferred without statistically assessing the confidence in the transfer. Here, we propose a supervised semi-parametric approach to estimate the FDR of transferred ions (see **Experimental Procedures**). Briefly, decoy ion transfers are generated by transferring ions with a significant m/z shift. All transferred ions are classified into four types: the ion has not been identified by MS/MS and is not a decoy (type 1); the ion is a decoy type 1 (type -1); the ion has been identified by MS/MS and is not a decoy (type 2); and the ion is a decoy type 2 ion (type -2). IonQuant trains a linear discriminant analysis (LDA) model with type 2 and -2 ions to separate the target and decoy ions. Using the trained model, a final score is calculated for each of the type 1 and -1 ions (Equation (2)). A mixture model (Equation (3)) is built using type 1 and -1 ions,

and the expectation-maximization (EM) algorithm is used to fit the model and subsequently calculate the posterior probability. Finally, global ion-level FDR (Equation (5)) is calculated using the local FDR, equal to one minus the posterior probability (Equation (4)). IonQuant also calculates peptide and protein level FDR by collapsing ions with the same peptide and protein, respectively.

In the remainder of the manuscript, we demonstrate the accuracy of FDR-controlled MBR using a two-organism dataset, and the precision and accuracy of subsequent label-free quantification by comparing HeLa replicate runs and using a three-organism dataset, respectively.

Evaluation of FDR-controlled MBR method

We used the dataset published by Lim et al. (18) to evaluate the false positive rate of FDR-controlled MBR (see **Experimental Procedures**). The dataset is comprised of 20 LC-MS files from *H. sapiens*-only proteins ("H") and 20 from a mixture of *H. sapiens* (90%) and *S. cerevisiae* (10%) proteins ("HY"). With MBR, *S. cerevisiae* peptides transferred from HY to H runs are known to be false positives, and can be used to evaluate the false positive rate, equal to false positives (*S. cerevisiae* peptides in H runs) divided by negatives (*S. cerevisiae* peptides in total). To ensure all *S. cerevisiae* peptides in the HY runs have the chance to be transferred, the number of top runs used in transferring was set to 40 and minimum required correlation was set to 0. In evaluation, a peptide was assigned to *S. cerevisiae* if all proteins it maps to are from *S. cerevisiae*, or to *H. sapiens* if at least one of its proteins is from *H. sapiens*.

Overall, IonQuant coupled with MSFragger identified 45875 unique *H. sapiens* peptides and 4610 unique *S. cerevisiae* peptides, ~19% and ~31% more *H. sapiens* and *S. cerevisiae* peptides compared to MaxQuant, respectively (**Table 2, Supporting Table S1, Supporting Data S1**). More peptides were also identified or transferred in individual runs with MSFragger and IonQuant. In transferring ions between the runs, IonQuant had a lower false positive rate than MaxQuant, 1.76% compared to 2.78%. The numbers listed in **Table 2** differ slightly from Figure S1 in Lim et al. (18) because of small differences in data analysis settings and version of the tools used. **Figure 2(a)** shows average peptide coverage, average peptide false positive rate, average protein coverage, and average protein false positive rate with respect to different MBR FDR thresholds. We used the same threshold for ion-, peptide-, and protein-level MBR FDR. The actual MBR FDR corresponding to no FDR filtering are 8% (ion-level), 7% (peptide-level), and 4% (protein-level). The peptide/protein coverage values shown are *H. sapiens* peptides/proteins in each H run divided by total *H. sapiens* peptides/proteins identified in the dataset. Peptide coverage increases from ~57% to ~81% with the inclusion of MBR, and protein coverage increases from ~87% to ~96%. As the MBR FDR threshold is increased, neither peptide nor protein coverage increase significantly, indicating most *H. sapiens* peptides have been successfully transferred by IonQuant already at 1% MBR FDR. On the other hand, the false positive rate continues to rise when the MBR FDR threshold is increased, demonstrating the need for FDR-controlled MBR.

Improved protein quantification with FDR-controlled MBR

We used four HeLa cell lysate replicates acquired on a timsTOF Pro published by Meier et al. (22) to demonstrate the sensitivity and precision of label-free quantification coupled to FDR-controlled MBR (see **Experimental Procedures**). We previously (21) performed a similar analysis of the same dataset, but without MBR and with protein abundances calculated from

peptide ion intensities using top-N peptide approach. In this work we use a new protein abundance calculation module in IonQuant implemented according to the MaxLFQ (13) algorithm (see **Experimental Procedures**).

Table 3 lists the numbers of quantified proteins and the CV from each method. A protein is used for CV analysis and counted as quantified if it has non-zero LFQ intensity in at least two runs. The results from IonQuant and MaxQuant (both with MaxLFQ method, indicated as 'native' quantification method in **Table 3**) are shown, which were run under similar settings of requiring either a minimum of 1 or 2 peptide ions (IonQuant) or peptides (MaxQuant) in pair-wise ratio calculation in MaxLFQ method. Enabling MBR (MBR+) improved the number of quantified proteins without a significant increase in protein quantification CV. For example, with min 2 ion setting, IonQuant MBR+ quantified ~12% more proteins (5674 vs 5044) while maintained a similar CV (3.8% vs 3.6%) than IonQuant MBR-. Compared to MaxQuant, IonQuant quantified more proteins and with greater precision (lower CVs) in all pair-wise comparisons between the tools under matched settings. For example, IonQuant with MBR+ and min 1 ion quantified 6449 proteins with CV of 4.2%, compared to 5950 proteins with CV of 5.3% for MaxQuant with MBR+ and min 1 peptide. We also compared IonQuant's native (MaxLFQ-based) protein abundance with that from IonQuant with MSstats (40) for peptide to protein intensity roll-up. Note that MSstats does not have a simple way to set a minimum peptide ion count requirement. **Table 3** shows that IonQuant's maxLFQ-based protein abundance calculation method results in slightly lower CVs compared to using MSstats, whereas our initial (top-N peptide based) strategy for protein abundance calculation in IonQuant was inferior to that of MSstats (21).

We also used the three-organism mixture dataset published by Prianichnikov et al. (23) to demonstrate the accuracy of label-free quantification when FDR-controlled MBR is employed (see **Experimental Procedures**). There are three replicates each of two experimental conditions, where the ratios between the two conditions are 1:1 (*H. sapiens*), 2:1 (*S. cerevisiae*), and 1:4 (*E. coli*). Since these proteomes were mixed at known ratios, we can evaluate the accuracy of the label-free quantification algorithm by comparing the estimated ratio against the ground truth. MaxQuant results published by Prianichnikov et al. (23) were used as a benchmark. We also performed the analysis with the latest MaxQuant (version 1.6.14.0), a newer reviewed protein database, and parameters as close as possible to those used in MSFragger and IonQuant, and got similar results (**Supporting Figure S1**). We used LFQbench (39) to summarize the analyses and visualize the results (**Figure 3**). As expected, both MaxQuant and IonQuant quantified more proteins with MBR than without MBR. IonQuant quantified ~21% and ~9% more proteins compared to MaxQuant with and without MBR, respectively (**Figure 3, Supporting Table S3, Supporting Data S3**). IonQuant also had fewer outliers than MaxQuant.

FDR-controlled MBR in single-cell data

We used 26 runs from a single-cell proteomics dataset (24) to demonstrate the performance of FDR-controlled MBR in sparse datasets (i.e. datasets with low overlap between different samples, resulting in a large number of missing values). There are 5 biological replicates with 0, 1, 3, 10, and 50 cells. The 0-cell (blank runs) sample is used as negative control in MBR. MaxQuant with and without MBR were used as benchmark.

We first evaluated the number of proteins quantified in at least two runs (**Figure 4(a)**). Without MBR (MBR-), MSFragger with IonQuant identified and quantified a higher number of proteins per sample on average than MaxQuant across all groups of samples. MaxQuant with MBR (MBR+) got on average 68 proteins from a replicate of the 0-cell sample, which is much more than MaxQuant MBR- (14 proteins) and IonQuant MBR- (19 proteins) (**Supporting Table S4**, **Supporting Data S4**). This by itself indicates a noticeable false transfer rate of MaxQuant's MBR in these data. IonQuant MBR+ did not transfer any proteins to the 0-cell sample due to its stringent quality control (which effectively disabled MBR transfer to 0-cell sample, even with MBR option turned on). As expected, as the number of cells per sample increases, the average number of proteins quantified per sample, with or without MBR, increases for both MaxQuant and IonQuant. Comparing the numbers from MaxQuant MBR+ and IonQuant MBR+ with FDR set to 100% (i.e. no FDR control) shows that IonQuant has a higher number of transferred proteins than MaxQuant under equivalent setting of no FDR control applied in both tools. However, the actual FDR, as estimated using IonQuant, is high, up to 26% for MBR transfers to the single cell samples. Applying increasingly stringent MBR FDR filtering in IonQuant gradually reduces the number of quantified proteins per sample.

Our results above suggest that application of the conventional MBR strategy (i.e. with no FDR control) to sparse datasets such as single-cell data may result in a high rate of false transfers. IonQuant, with its ability to estimate and control FDR, provides the users a way to minimize false transfers by applying an FDR threshold of their choice. These results also warrant discussion regarding a reasonable FDR threshold to apply in such scenarios. **Figure 4(b)** shows the number of quantified peptides and proteins, and protein quantification CV from analyzing 11 replicates of 1-cell sample with MaxQuant and IonQuant, respectively. Without MBR, IonQuant measured more peptides (1407 vs 1208) and more proteins (409 vs 371), while

maintaining a similar CV (29.2% vs 27.0%) compared MaxQuant. With MBR but without FDR control, IonQuant also measured more peptides (4686 vs 3937) and more proteins (1087 vs 918) with a similar CV (27.8% vs 26.0%) compared to MaxQuant. As previously discussed, FDR control reduces the IonQuant numbers. Raising the FDR threshold increases the number of proteins until roughly 10% FDR, when saturation becomes apparent. Notably, saturation is reached at a much smaller FDR threshold in the whole cell lysate data (around 1% FDR, **Figure 2(a)).** This is reflective of the fact that single-cell data is naturally more sparse, with more peptides and proteins that can be transferred from other runs (especially when transferring identifications from “boosting” samples, i.e. samples with much higher number of cells). This also suggests that setting a higher FDR threshold (e.g. 5% FDR) in single-cell or other kinds of similarly sparse data may be justifiable. At the same time, it is likely sufficient to restrict FDR MBR to 1% in more typical datasets, where loosening the FDR threshold serves to reduce overall quantification accuracy with no noticeable improvement in the number of quantified proteins.

Speed of indexing-based MBR in IonQuant

Finally, we compared the computational time required by IonQuant (version 1.4.4) and MaxQuant (version 1.6.14.0), both with MBR enabled. The HeLa dataset (timsTOF Pro) and the two-organism dataset from (Orbitrap Fusion Lumos) were used, comprised of 4 and 40 LC-MS files, respectively. (**Experimental Procedures**). For MaxQuant, only jobs related to quantification and MBR were counted (**Supporting Table S5**). **Figure 2(b)** displays the run time of these tools in minutes. IonQuant is ~40x faster and ~17x faster than MaxQuant in analyzing the data without and with ion mobility, respectively. The reason for IonQuant exhibits a smaller gain in speed compared to MaxQuant in analyzing the timsTOF Pro data is that most of the IonQuant runtime is spent loading the raw data via the vendor-provided library (21).

Conclusions

Match-between-runs (MBR) is a commonly used approach to quantify additional features by transferring information across different samples. It largely mitigates the missing value problem of DDA-based label-free quantification, increasing data completeness for improved differential analyses. To our knowledge, there was previously no method to control the rate of false transfers in DDA-based MBR in practical settings. To address this issue, we described a method to estimate and control the FDR for MBR with the help of mixture modeling and the target-decoy concept. We implemented MBR with FDR control in our quantification tool, IonQuant. Our experiments and comparisons with a frequently used tool MaxQuant showed that IonQuant allowed fewer false positive transfers while maintaining high sensitivity. We also highlight the importance of FDR control when MBR is applied to sparse datasets such as those from single-cell proteomics experiments. Furthermore, by way of advanced indexing technology, IonQuant performs MBR with unmatched speed, making it well-suited even for analysis of large-scale datasets composed of thousands of LC-MS runs.

Acknowledgements

This work was funded in part by NIH grants R01-GM-094231 and U24-CA210967. We thank Brett Phinney, Roman Fischer, and Tobias Kockmann for useful discussions.

Data and Software Availability

The two-organism data was published by Lim et al. (18) and can be found from the ProteomeXchange Consortium (32) with identifier PXD014415. The HeLa cell lysate data was

published by Meier et al. (22) and can be found from the ProteomeXchange Consortium with the identifier PXD010012. The three-organism data was published by Prianichnikov et al. (23) and can be found from the ProteomeXchange Consortium with identifier PXD014777. The single-cell data was published by Williams et al. (24) and can be found from the ProteomeXchange Consortium with identifier MSV000085230. MSFragger and IonQuant programs were developed in the cross-platform Java language and can be accessed at http://msfragger.nesvilab.org/ and https://ionquant.nesvilab.org/.

Author Contributions

F.Y. developed IonQuant and its match-between-runs module; F.Y. and A.I.N. analyzed the data; F.Y., S.E.H., and A.I.N. wrote the manuscript with input from all authors; A.I.N. supervised the entire project.

Competing Interests Statement

The authors declare no competing financial interests.

References

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198-207
2. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 4, 787-797
3. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 73, 2092-2123
4. Leitner, A., and Aebersold, R. (2013) SnapShot: mass spectrometry for protein and proteome analyses. *Cell* 154, 252-252.e251
5. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., and Aebersold, R. (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular systems biology* 14, e8126
6. Rosenberger, G., Bludau, I., Schmitt, U., Heusel, M., Hunter, C. L., Liu, Y., MacCoss, M. J., MacLean, B. X., Nesvizhskii, A. I., Pedrioli, P. G. A., Reiter, L., Röst, H. L., Tate, S., Ting, Y. S., Collins, B. C., and Aebersold, R. (2017) Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Methods* 14, 921-927
7. Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., and Nesvizhskii, A. I. (2015) DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* 12, 258-264, 257 p following 264
8. Searle, B. C., Pino, L. K., Egertson, J. D., Ting, Y. S., Lawrence, R. T., MacLean, B. X., Villén, J., and MacCoss, M. J. (2018) Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat Commun* 9, 5128
9. Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M. Y., Vitek, O., Aebersold, R., and Müller, M. (2007) SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 7, 3470-3480
10. Tsou, C. C., Tsai, C. F., Tsui, Y. H., Sudhir, P. R., Wang, Y. T., Chen, Y. J., Chen, J. Y., Sung, T. Y., and Hsu, W. L. (2010) IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Mol Cell Proteomics* 9, 131-144
11. Zimmer, J. S., Monroe, M. E., Qian, W. J., and Smith, R. D. (2006) Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass spectrometry reviews* 25, 450-482
12. Andreev, V. P., Li, L., Cao, L., Gu, Y., Rejtar, T., Wu, S. L., and Karger, B. L. (2007) A new algorithm using cross-assignment for label-free quantitation with LC-LTQ-FT MS. *J Proteome Res* 6, 2186-2194
13. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 13, 2513-2526
14. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 11, 2301-2319
15. Deshmukh, A. S., Murgia, M., Nagaraj, N., Treebak, J. T., Cox, J., and Mann, M. (2015) Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. *Mol Cell Proteomics* 14, 841-853
16. Rieckmann, J. C., Geiger, R., Hornburg, D., Wolf, T., Kveler, K., Jarrossay, D., Sallusto, F., Shen-Orr, S. S., Lanzavecchia, A., Mann, M., and Meissner, F. (2017) Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat Immunol* 18, 583-593
17. Anders, S., and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol* 11, R106
18. Lim, M. Y., Paulo, J. A., and Gygi, S. P. (2019) Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J Proteome Res* 18, 4020-4026

19. Choi, H., and Nesvizhskii, A. I. (2008) Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J Proteome Res* 7, 254-265
20. Tsou, C. C., Tsai, C. F., Teo, G. C., Chen, Y. J., and Nesvizhskii, A. I. (2016) Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics* 16, 2257-2271
21. Yu, F., Haynes, S. E., Teo, G. C., Avtonomov, D. M., Polasky, D. A., and Nesvizhskii, A. I. (2020) Fast quantitative analysis of timsTOF PASEF data with MSFragger and IonQuant. *Molecular & Cellular Proteomics*, mcp.TIR120.002048
22. Meier, F., Brunner, A. D., Koch, S., Koch, H., Lubeck, M., Krause, M., Goedecke, N., Decker, J., Kosinski, T., Park, M. A., Bache, N., Hoerning, O., Cox, J., Rather, O., and Mann, M. (2018) Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* 17, 2534-2545
23. Prianichnikov, N., Koch, H., Koch, S., Lubeck, M., Heilig, R., Brehmer, S., Fischer, R., and Cox, J. (2020) MaxQuant software for ion mobility enhanced shotgun proteomics. *Mol Cell Proteomics*
24. Williams, S. M., Liyu, A. V., Tsai, C. F., Moore, R. J., Orton, D. J., Chrisler, W. B., Gaffrey, M. J., Liu, T., Smith, R. D., Kelly, R. T., Pasa-Tolic, L., and Zhu, Y. (2020) Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics. *Anal Chem* 92, 10588-10596
25. Freksa, C., Newcombe, N. S., Gärdenfors, P., and Wölfl, S. (2008) *Spatial Cognition VI. Learning, Reasoning, and Talking about Space: International Conference Spatial Cognition 2008, Freiburg, Germany, September 15-19, 2008. Proceedings*, Springer
26. Millikin, R. J., Solntsev, S. K., Shortreed, M. R., and Smith, L. M. (2018) Ultrafast Peptide Label-Free Quantification with FlashLFQ. *J Proteome Res* 17, 386-391
27. Yu, F., Teo, G. C., Kong, A. T., Haynes, S. E., Avtonomov, D. M., Geiszler, D. J., and Nesvizhskii, A. I. (2020) Identification of modified peptides using localization-aware open search. *Nat Commun* 11, 4065
28. Stanley, J. R., Adkins, J. N., Slysz, G. W., Monroe, M. E., Purvine, S. O., Karpievitch, Y. V., Anderson, G. A., Smith, R. D., and Dabney, A. R. (2011) A Statistical Method for Assessing Peptide Identification Confidence in Accurate Mass and Time Tag Proteomics. *Analytical Chemistry* 83, 6135-6140
29. The, M., and Käll, L. (2020) Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. *Nat Commun* 11, 3234
30. Breen, E. J., Hopwood, F. G., Williams, K. L., and Wilkins, M. R. (2000) Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis* 21, 2243-2251
31. Ma, K., Vitek, O., and Nesvizhskii, A. I. (2012) A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* 13 Suppl 16, S1
32. Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P. A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H. J., Albar, J. P., Martinez-Bartolome, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 32, 223-226
33. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* 14, 513-520
34. Leprevost, F. V., Haynes, S. E., Avtonomov, D. M., Chang, H.-Y., Shanmugam, A. K., Mellacheruvu, D., Kong, A. T., and Nesvizhskii, A. I. (2020) Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nature Methods*

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

36. Consortium, U. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 47, D506–D515

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

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

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

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

Table 1. List of scores used in linear discriminant analysis (LDA).

Score	**Explanation**
Log10(intensity)	Log-transformed intensity of a traced peak. The intensity can be from an area (without ion mobility) or a volume (with ion mobility).
Log10(KL)	Log-transformed Kullback-Leibler divergence of an experimental isotope distribution and the theoretical isotope distribution. 0, +1, and +2 isotope peaks are used.
Log10(KL_negative_1)	Same as Log10(KL) but -1, 0, and +1 isotope peaks are used.
Log10(KL_negative_2)	Same as Log10(KL) but -2, -1, and 0 isotope peaks are used.
Abs(ppm)	Absolute value of the mass error (in ppm) from a traced peak.
IM diff	Ion mobility difference between an acceptor ion and its donor ion.
RT diff	Retention time difference between an acceptor ion and its donor ion.
Log10(intensity) diff	Log-transferred intensity difference between an acceptor ion and its donor ion. The intensities are normalized by their runs' median intensity.
Correlation between runs	Weighted Spearman's rank correlation coefficient between an acceptor ion and the donor run where the ion contributes.
Matched run percentage	The number of donor runs contributing the ion divided by the number of qualified donor runs.

Table 2. Peptides identified by MaxQuant and IonQuant in analyzing the two-organism dataset with MBR. "Sample H" indicates *H. sapiens*-only samples and "Sample HY" indicates samples with a mixture of *H. sapiens* and *S. cerevisiae* proteins. There are 20 runs in each sample type. "MBR-" and "MBR+" indicate that the analysis was performed without and with match-between-runs (MBR), respectively. For each analysis, unique peptide counts are listed along with per run identification rates (% of all observed peptides found in each run).

MaxQuant		
Total unique *H. sapiens* peptides	38405	
Sample H, MBR -	19360±648	50.41%
Sample HY, MBR -	18945±522	49.33%
Sample H, MBR +	31129±637	81.05%
Sample HY, MBR +	29747±730	77.46%
Total unique *S. cerevisiae* peptides	3527	
Sample H, MBR -	20±5	0.57%
Sample HY, MBR -	1848±93	52.40%
Sample H, MBR +	98±10	2.78%
Sample HY, MBR +	2858±63	81.03%

IonQuant		
Total unique *H. sapiens* peptides	45875	
Sample H, MBR -	26032±499	56.75%
Sample HY, MBR -	25683±716	55.98%
Sample H, MBR +	37075±314	80.82%
Sample HY, MBR +	36134±666	78.77%
Total unique *S. cerevisiae* peptides	4610	
Sample H, MBR -	26±6	0.56%
Sample HY, MBR -	2597±82	56.33%
Sample H, MBR +	81±13	1.76%
Sample HY, MBR +	3756±75	81.48%

Table 3. Quantified protein counts and coefficients of variation (CV) from four HeLa cell lysate replicates. “MBR-” and “MBR+” indicate that the analysis was performed without and with match-between-runs (MBR), respectively. Both MSstats and the tools’ native modules were used to calculate protein intensities from ion intensities. MSstats had errors in processing the file from MaxQuant with MBR.

Tool			Native protein quantified	Native CV (%)	Msstats proteins quantified	MSStats CV (%)
MaxQuant	MBR-	min 1 peptide	5406	5.3	5334	5.6
		min 2 peptides	4186	4.3		
	MBR+	min 1 peptide	5950	5.3	(no result)	
		min 2 peptides	5073	4.7		
IonQuant	MBR-	min 1 ion	5952	4.1	5952	4.8
		min 2 ions	5044	3.6		
	MBR+	min 1 ion	6449	4.2	6451	4.7
		min 2 ions	5674	3.8		

Figure 1. (a) (a) Overview of match-between-runs in IonQuant. For each acceptor run (central point in teal) ion-level correlations with all other runs (blue and orange points) are calculated, where distance from the central point represents correlation. The top N runs (blue points) within the correlation threshold (gray area) are selected as eligible donor runs. For every ion in each eligible donor run, target and decoy (m/z-shifted) transfer regions are located using retention time (and ion mobility if applicable). Peak tracing in the acceptor run is used to determine the isotopic distribution and the charge state. All matches are evaluated, and the top scoring donor for each acceptor peak is selected for transfer. (b) All matches/transferred ions are classified into one of the four categories shown. Type 2 and -2 matches are used to train a linear discriminant analysis (LDA) model. The trained LDA is used to calculate the final score for type 1 and -1 matches. A posterior probability of correct transfer is estimated by fitting a mixture model, allowing ion, peptide, and protein-level false discovery rate (FDR) for match-between-runs estimation.

Figure 2. (a) Per-run proteome coverage and observed false positive rate as a function of the model-estimated false discovery rate (FDR) threshold. Coverage is equal to the number of *H. sapiens* peptides/proteins from one run divided by the total number of *H. sapiens* peptide/protein identifications in the entire experiment. The false positive rate is equal to the number of *S. cerevisiae* peptides/proteins from one run divided by the total number of *S. cerevisiae* peptides/proteins. (b) Run time comparison of quantification-related tasks using the HeLa dataset (4 timsTOF Pro runs) and the two-organism dataset (40 Orbitrap Fusion Lumos runs).

Figure 3. Ground-truth protein quantification results from MaxQuant and IonQuant from a mixture of three different proteomes. “MBR-” and “MBR+” indicate that the analysis was performed without and with match-between-runs (MBR), respectively. *S. cerevisiae* proteins are shown in orange, *H. sapiens* in green, and *E. coli* in purple. The known ratios of three organisms are 2:1 (*S. cerevisiae*), 1:1 (*H. sapiens*), and (1:4) *E. coli*. Box plots of the intensities are shown to the right of each scatter plot panel.

Figure 4. Quantified peptides and proteins from MaxQuant and IonQuant analysis of the single cell dataset. “MBR-” and “MBR+” indicate that the analysis was performed without and with match-between-runs, respectively. (a) Numbers of quantified proteins from samples with 0 cells (blank runs), 1 cell, 3 cells, and 10 cells, respectively. Two ion-level MBR false discovery rate (FDR) thresholds (1% and 5%) plus “no FDR” setting (corresponding to IonQuant-estimated FDR of 26%) were applied. (b) Peptide/protein numbers and protein quantification coefficient of variation (CV) from 11 replicates of 1 cell samples, as a function of FDR threshold. “MQ” indicates MaxQuant and “IQ” indicates IonQuant.

Supporting Table S1. Proteins from two-organism dataset analyzed by MSFragger and IonQuant.

Supporting Table S2. Proteins from HeLa dataset analyzed by MSFragger and IonQuant without match-between-runs.

Supporting Table S3. Proteins from HeLa dataset analyzed by MSFragger and IonQuant with match-between-runs and 1% FDR.

Supporting Table S4. Proteins from three-organism dataset analyzed by MSFragger and IonQuant without match-between-runs.

Supporting Table S5. Proteins from three-organism dataset analyzed by MSFragger and IonQuant with match-between-runs and 1% FDR.

Supporting Table S6. Proteins from single-cell dataset analyzed by MSFragger and IonQuant without match-between-runs.

Supporting Table S7. Proteins from single-cell dataset analyzed by MSFragger and IonQuant with match-between-runs and 1% FDR.

Supporting Table S8. Run time from MaxQuant analyzing two-organism dataset.

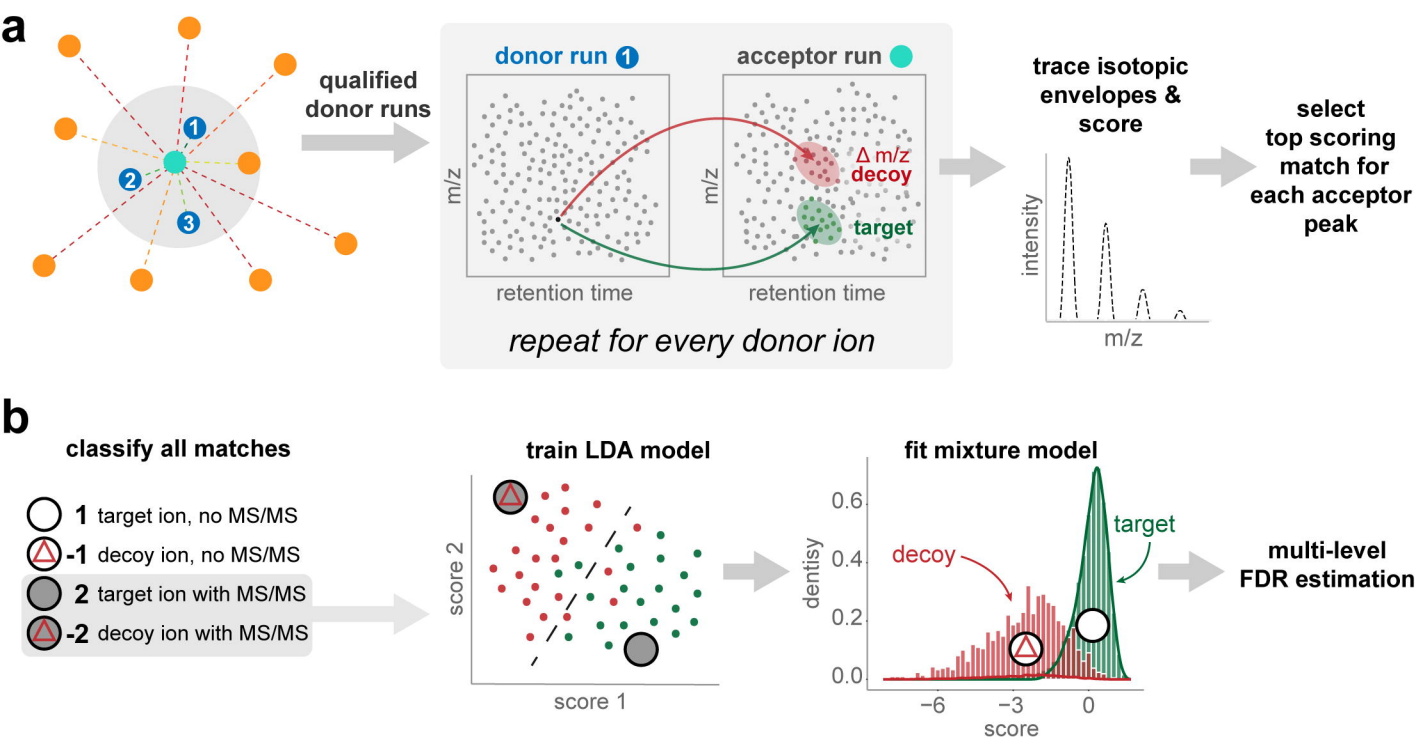
Supporting Table S9. Run time from MaxQuant analyzing HeLa dataset.

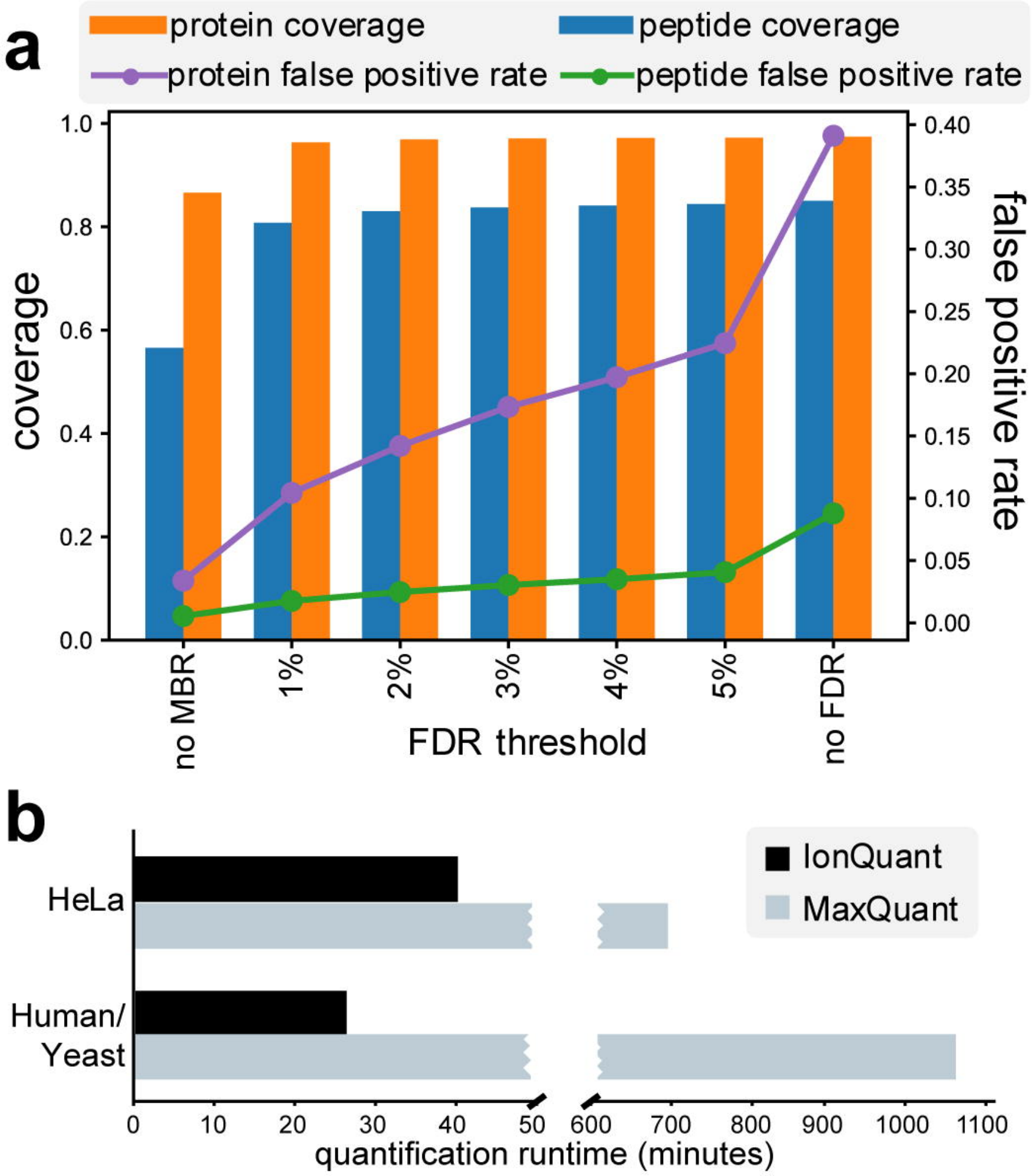
Supporting Data S1. Ions from two-organism dataset analyzed by MSFragger and IonQuant.

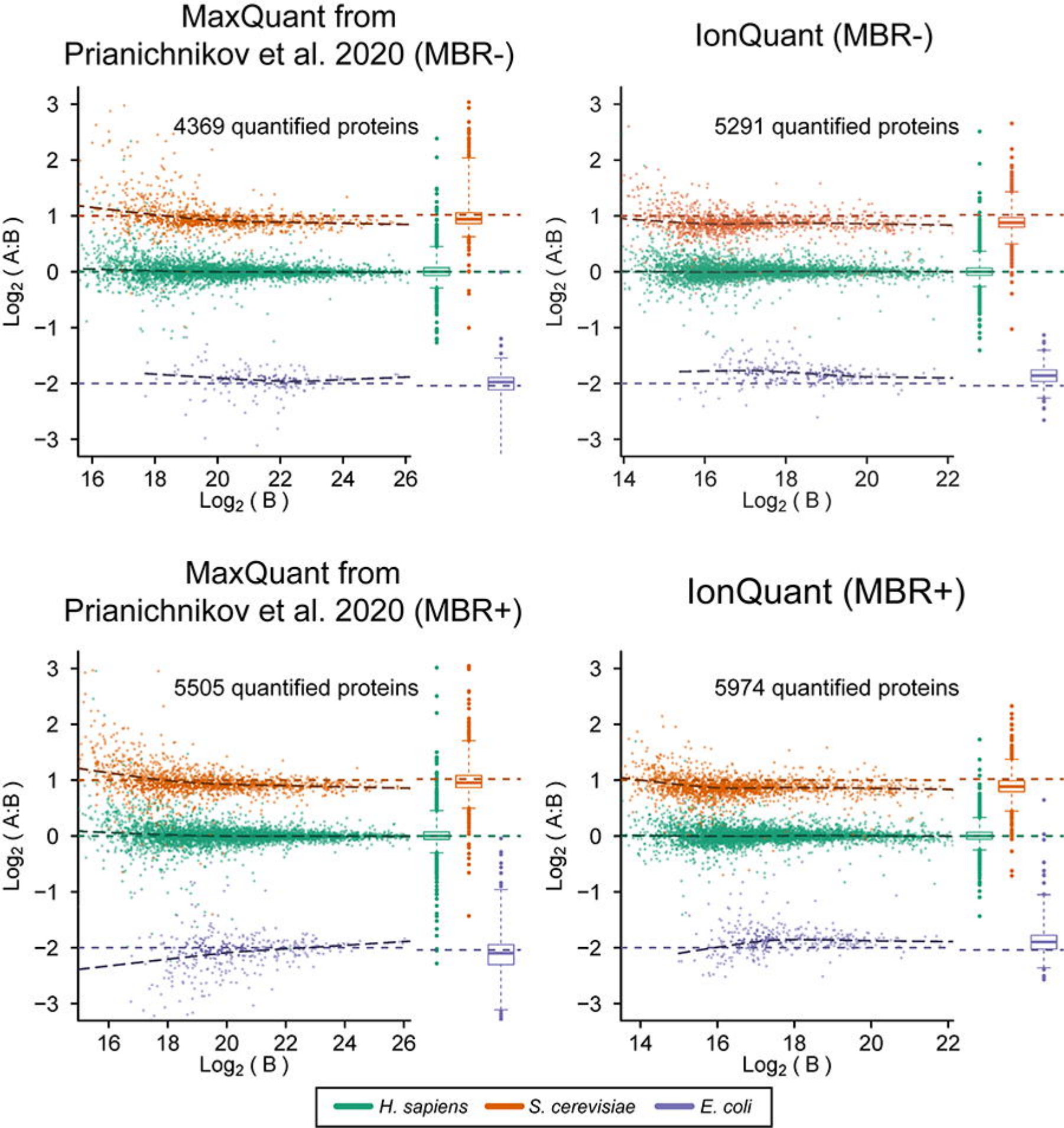
Supporting Data S2. Ions from HeLa dataset analyzed by MSFragger and IonQuant.

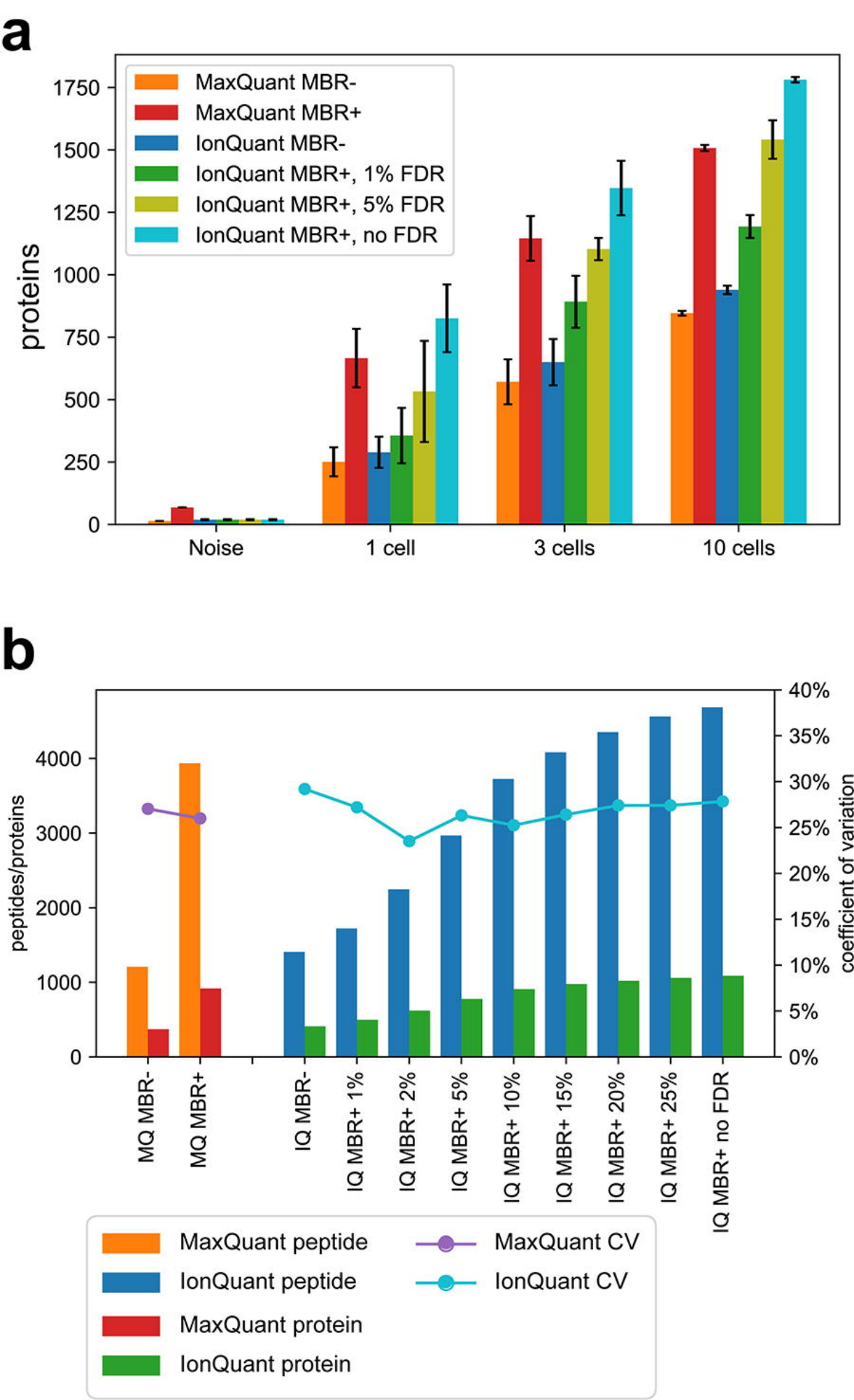
Supporting Data S3. Ions from three-organism dataset analyzed by MSFragger and IonQuant.

Supporting Data S4. Ions from single-cell dataset analyzed by MSFragger and IonQuant.


a
qualified donor runs
donor run 1
acceptor run
m/z
retention time
Δ m/z
decoy
target
repeat for every donor ion
trace isotopic envelopes & score
intensity
m/z
select top scoring match for each acceptor peak
b
classify all matches
1 target ion, no MS/MS
-1 decoy ion, no MS/MS
2 target ion with MS/MS
-2 decoy ion with MS/MS
train LDA model
score 2
score 1
fit mixture model
dentisy
decoy
target
score
multi-level FDR estimation


a
protein coverage
peptide coverage
protein false positive rate
peptide false positive rate
coverage
false positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
no MBR
1%
2%
3%
4%
5%
no FDR
FDR threshold
b
IonQuant
MaxQuant
HeLa
Human/
Yeast
0
10
20
30
40
50
600
700
800
900
1000
1100
quantification runtime (minutes)


MaxQuant from Prianichnikov et al. 2020 (MBR-)
4369 quantified proteins
IonQuant (MBR-)
5291 quantified proteins
MaxQuant from Prianichnikov et al. 2020 (MBR+)
5505 quantified proteins
IonQuant (MBR+)
5974 quantified proteins
Log2 (A:B)
Log2 (B)
H. sapiens
S. cerevisiae
E. coli


a
MaxQuant MBR-
MaxQuant MBR+
IonQuant MBR-
IonQuant MBR+, 1% FDR
IonQuant MBR+, 5% FDR
IonQuant MBR+, no FDR
proteins
0
250
500
750
1000
1250
1500
1750
Noise
1 cell
3 cells
10 cells
b
peptides/proteins
coefficient of variation
0
1000
2000
3000
4000
0%
5%
10%
15%
20%
25%
30%
35%
40%
MQ MBR-
MQ MBR+
IQ MBR-
IQ MBR+ 1%
IQ MBR+ 2%
IQ MBR+ 5%
IQ MBR+ 10%
IQ MBR+ 15%
IQ MBR+ 20%
IQ MBR+ 25%
IQ MBR+ no FDR
MaxQuant peptide
IonQuant peptide
MaxQuant protein
IonQuant protein
MaxQuant CV
IonQuant CV