DART-ID increases single-cell proteome coverage

Albert T. Chen[1], Alexander Franks[2], Nikolai Slavov[1,3,✉]

[1] Department of Bioengineering, Northeastern University, Boston, MA 02115, USA
[2] Department of Statistics and Applied Probability, UC Santa Barbara, CA 93106, USA
[3] Department of Biology, Northeastern University, Boston, MA 02115, USA

✉ – Corresponding author: nslavov@northeastern.edu

Abstract

Analysis by liquid chromatography and tandem mass spectrometry (LC-MS/MS) can identify and quantify thousands of proteins in microgram-level samples, such as those comprised of thousands of cells. Identifying proteins by LC-MS/MS proteomics, however, remains challenging for lowly abundant samples, such as the proteomes of single mammalian cells. To increase the identification rate of peptides in such small samples, we developed DART-ID. This method implements a data-driven, global retention time (RT) alignment process to infer peptide RTs across experiments. DART-ID then incorporates the global RT-estimates within a principled Bayesian framework to increase the confidence in correct peptide-spectrum-matches. Applying DART-ID to hundreds of samples prepared by the Single Cell Proteomics by Mass Spectrometry (SCoPE-MS) design increased the peptide and proteome coverage by 30 – 50% at 1% FDR. The newly identified peptides and proteins were further validated by demonstrating that their quantification is consistent with the quantification of peptides identified from high-quality spectra. DART-ID can be applied to various sets of experimental designs with similar sample complexities and chromatography conditions, and is freely available online.

Author Summary

Identifying and quantifying proteins in single cells gives researchers the ability to tackle complex biological problems that involve single cell heterogeneity, such as the treatment of solid tumors. However, the mass spectra from analysis of single cells do not support sequence identification for all analyzed peptides. To improve identification rates, we utilize the retention time of peptide sequences from liquid chromatography – a process used before to separate peptides before their analysis with mass spectrometry. We present both a novel method of aligning the retention times of peptides across experiments, as well as a rigorous framework for using the estimated retention times to enhance peptide sequence identification. Incorporating the retention time as additional evidence leads to a substantial increase in the number of proteins that can be quantified and biologically analyzed by single-cell mass spectrometry.

Introduction

Advancements in the sensitivity and discriminatory power of protein mass-spectrometry (MS) have enabled the quantitative analysis of increasingly limited amounts of samples. Recently, we have developed Single Cell Proteomics by Mass Spectrometry (SCoPE-MS) that allows quantifying proteins in single mammalian cells [1]. One of the challenges for SCoPE-MS is identifying the sequences of low abundance peptides, which generate limited numbers of the fragment ions that are necessary for confident identification by mass spectra [2, 3]. We sought to overcome this challenge by using the retention time (RT) of an ion as an additional feature to its spectra to achieve more confident peptide identifications. While this idea has been previously implemented, as discussed below, we enhanced its implementation by developing a novel data-driven method for aligning RTs and updating peptide confidence. DART-ID uses fewer assumptions than previous implementations, substantially reduces residual error in aligned RTs, resulting in median error below 3 seconds, and increases the peptide identification rate by 30 – 50% at 1% FDR.

The RT of a peptide is a specific and informative feature of its sequence, and this specificity has motivated approaches aiming to estimate peptide RTs. These approaches either (i) predict RTs from peptide sequences or (ii) align empirically measured RTs. Estimated peptide RTs have a wide

range of uses, such as scheduling targeted MS/MS experiments, building efficient inclusion and exclusion lists for LC-MS/MS, or augmenting MS2 mass spectra to increase identification rates.

Peptide RTs can be estimated from physical properties such as sequence length, constituent amino acids, and amino acid positions, as well as chromatography conditions, such as column length, pore size, and gradient shape. These features predict the relative hydrophobicity of peptide sequences and thus RTs for LC used with MS [4–10]. The predicted RTs can be improved by implementing machine learning algorithms that incorporate confident, observed peptides as training data [11–16]. Predicted peptide RTs are mostly used for scheduling targeted MS/MS analyses where acquisition time is limited, e.g., multiple reaction monitoring [17]. They can also be used to aid peptide sequencing, as exemplified by "peptide fingerprinting" that identifies peptides based on ion's RT and mass over charge (m/z) [7, 18–20]. While peptide fingerprinting has been successful for low complexity samples, where MS1 m/z and RT space is less dense, it requires carefully controlled conditions and rigorous validation with MS2 spectra [19–21]. Predicted peptide RTs have more limited use with data-dependent acquisition, i.e., shotgun proteomics. They have been used to generate data-dependent exclusion lists that spread MS2 scans over more diverse subset of the proteome [22, 23], as well as to aid peptide identification from MS2 spectra, either by incorporating the RT error (difference between predicted and observed RTs) into a discriminant score [24], or filtering out observations by RT error to minimize the number of false positives selected [13, 25, 26]. In addition, RT error has been directly combined with search engine scores [16, 27]. Besides automated methods of boosting identification confidence, proteomics software suites such as Skyline allow the manual comparison of measured and predicted RTs to validate peptide identifications [28].

The second group of approaches for estimating peptide RTs aligns empirically measured RTs across multiple experiments. Peptide RTs vary due to variation in sample complexity, matrix effects, column age, room temperature and humidity. Thus, estimating peptide RTs from empirical measurements requires normalization, or alignment, that compensates for RT variation across experiments. Usually, RT alignment methods align the RTs of two experiments at a time, and typically utilize either a shared, confidently-identified set of endogenous peptides, or a set of spiked-in calibration peptides [29, 30]. Pairwise alignment approaches must choose a particular set of RTs to

which to align all other experiment, and the choice of such reference RT set is not obvious. Alignment methods are limited by the availability of RTs measured in relevant experimental conditions, but can result in more accurate RT estimates when such empirical measurements are available [30–32]. Generally, RT alignment methods can provide more accurate estimations than RT prediction methods, discussed earlier, but also generally require more extensive data and cannot estimate RTs of peptides without empirical observations.

Methods for RT alignment are various, and range from linear shifts to non-linear distortions and time warping [33]. Some have argued for the necessity of non-linear warping functions to correct for RT deviations [34], while others have posited that most of the variation can be explained by simple linear shifts [35]. More complex methods include multiple generalized additive models [36], or machine-learning based semi-supervised alignments [37]. Once experiments are aligned, peptide RTs can be predicted by applying experiment-specific alignment functions to the RT of a peptide observed in a reference run.

Peptide RTs estimated by alignment can be used to schedule targeted MS/MS experiments – similar to the use of predicted RTs estimated from the physical properties of a peptide [30]. The most common usage of RT alignments has been to align or compare features in the MS1 m/z and RT space, i.e., the ion map. These alignments are crucial when performing calibrations, difference detection in features, or label-free quantitation [31, 38–41]. The match-between-runs function of MaxQuant [31, 32] can match observed MS1 m/z-RT pairs across experiments, without considering the MS2 spectra of the matched ion. Percolator, a support vector machine classifier, uses aligned peptide RTs as a predictive feature, along with many other features, to estimate the confidence of peptide identifications [42].

With DART-ID, we compare estimated RTs and observed RTs as additional pieces of evidence, independent from the MS2 mass spectra, to update the confidence in peptide sequence identification. Unlike previous methods which incorporate the RT error into discriminants [24], filters [13, 25, 26], or scores [16, 27], we update the ID confidence directly with a Bayesian model [43, 44]. Crucial to this method is the accuracy of the alignment method, the higher the accuracy of RT estimates, the more informative they are for the peptide sequence.

With DART-ID, we implement a novel global RT alignment method that takes full advan-

tage of SCoPE-MS data, which feature many experiments run with analogous samples and on the same nano-LC (nLC) system [1, 45]. These experimental conditions yield many RT estimates per peptide with relatively small variability across experiments. In this context, we used empirical distribution densities and obviated the functional dependence between peptide properties and RTs to maximize the statistical power of highly reproducible RTs.

Results

Model for global RT alignment and PSM confidence update

Using RT for identifying peptide sequences starts with estimating the RT for each peptide, and we aimed to maximize the accuracy of RT estimation by optimizing RT alignment. Many existing methods can only align the RTs of two experiments at a time, i.e., pairwise alignment, based on partial least squares minimization, which does not account for the measurement errors in RTs [46]. Furthermore, the selection of a reference experiment is non-trivial, and different choices can give quantitatively different alignment results. In order to address these challenges, we developed a global alignment method, sketched in Fig. 1a,b. The global alignment infers a *reference RT* for each peptide as a latent variable. The reference RT is initialized simply as the mean RT of all observed instances of the peptide from the unaligned experiments, Fig. 1a. This approach allows the alignment to take advantage of any peptide observed in at least two experiments, regardless of the number of missing observations. Thus it maximizes the data used for alignment and obviates the need for spiked-in standards. Furthermore, the reference RT provides a principled choice for a reference (rather than choosing a particular experiment) that is free of measurement noise. The alignment process accounts for the error in individual observations by inferring a per peptide RT distribution, as opposed to aligning to a point estimate, as well as for variable RT deviations in experiments by using experiment-specific weights.

For experiments with similar sample complexities and chromatography conditions, it is reasonable to model the relationship between experiments with a monotonic function. In DART-ID, we use a two-piece linear model for the alignment function, and we found that this model captured

the majority of the inter-experiment variation in RTs. We also observed that our alignment errors follow a Laplace distribution, and that RT residuals generally increase along the LC gradient. Incorporating these observations, we modelled the RT density for the alignment with a Laplace distribution, and modelled the variance of that distribution as an experiment-specific term.

The conceptual idea based on which we incorporate RT information in sequence identification is illustrated in Fig. 1c and formalized with Bayes' theorem in Fig. 1d. We start with a peptide-spectrum-match (PSM) from a search engine and its associated probability to be incorrect (PEP; posterior error probability) and correct, 1-PEP. If the RT of a PSM is far from the RT of its corresponding peptide, as PSM1 in Fig. 1c, then the spectrum is more likely to be observed if the PSM is incorrect, and thus we can decrease its confidence. Conversely, if the RT of a PSM is very close to the RT of its corresponding peptide, as PSM2 in Fig. 1c, then the spectrum is more likely to be observed if the PSM is correct, and thus we can increase its confidence. To estimate whether the RT of a PSM is more likely to be observed if the PSM is correct or incorrect, we use the conditional likelihood probability densities inferred from the alignment procedure in Fig. 1b. Combining these likelihood functions with the Bayes' theorem in Fig. 1d allows to formalize this logic and update the confidence of analyzed PSMs, which we quantify with DART-ID PEPs.

Global alignment process reduces RT deviations

To evaluate the global alignment by DART-ID, we used a set of 46 LC-MS/MS 60 min runs. Each run was a 1% injection from a $100 \times M$ SCoPE-MS sample with two carrier channels (126C - Jurkat cells; 127N - U-937 cells) and 6 interleaved single-cell channels (3 Jurkat and 3 U-937 cells), so that the injections represented the protein abundances expected for single-cell SCoPE-MS set [1, 45]; see Methods for more details. The experiments were run over a span of two months so that the measured RTs captured expected variance in the chromatography. The measured RTs were compared to RTs predicted from peptide sequences [9, 10, 14], and to the reference RTs from DART-ID and top-performing aligning methods [30–32]. All methods estimated RTs that explained the majority of the variance of the measured RTs, Fig. 2a. The alignment methods provided closer estimates, explaining over 99% of the variance.

To evaluate RT similarity more rigorously, we compared the distributions of differences be-

tween the reference RTs and measured RTs, shown in Fig. 2b. This comparison again underscores that the differences are significantly smaller for alignment methods, and smallest for DART-ID. We further quantified these differences by computing the mean and the median RT deviation, i.e., $|\Delta\text{RT}|$, which is defined as the absolute value of the difference between the observed RT and the reference RT. For the prediction methods – SSRCalc, BioLCCC, and ELUDE – the average deviations exceed 2 min, and ELUDE has the smallest deviation of 2.5 min. The alignment methods result in smaller deviations, all below < 1 min, and DART-ID shows the smallest deviation of 0.044 min (2.6 seconds).

Incorporating RTs increases proteome coverage

Search engines such as MaxQuant [31] use the similarity between theoretically predicted and experimentally measured MS1 and MS2 spectra of ions to match them to peptide sequences, i.e., peptide-spectral-matches (PSM). The confidence of a PSM is commonly quantified by the probability of an incorrect match: the posterior error probability (PEP) [44]. Since the estimation of PEP does not include RT information, we sought to update the PEP for each PSM by incorporating RT information within the Bayesian framework displayed in Fig. 1c,d. This approach allowed us to use the estimated RT distributions for each peptide with minimal assumptions.

We refer to the PEP assigned by the search engine as "Spectral PEP", and after it is updated by the Bayesian model from Fig. 1d as "DART-ID PEP". Comparing the Spectral and DART-ID PEPs indicates that the confidence for some PSMs increases while for others decreases; see density plot in Fig. 3a. Reassuringly, all PSMs with low Spectral PEPs have even lower]dart PEPs, meaning that all confident PSMs become even more confident. On the other extreme, many PSMs with high Spectral PEPs have even higher DART-ID PEP, meaning that some low-confidence PSMs are further downgraded. Confidence upgrades, where DART-ID PEP $<$ Spectral PEP, range within 1–3 orders of magnitude.

The density plot in Fig. 3b displays a subset of peptides with Spectral PEP > 0.01 and DART-ID PEP < 0.01. These peptides have low confidence of identification based in their MS/MS spectra alone, but high confidence when RT evidence is added to the spectral evidence. To visualize how these peptides are distributed across experiments, we marked them with red dashes in Fig. 3b. The

results indicate that sparsity decreases; thus DART-ID helps mitigate the missing data problem of shotgun proteomics. In addition to PSMs of peptides with at least one confident spectral PSM, DART-ID increases the confidence for PSMs of peptides without confident PSMs, see as marked in Fig. 3b. While these PSMs very likely represent the same peptide sequence – since by definition they share the same RT, MS1 m/z and MS2 fragments consistent with the sequence – we cannot be confident at this point in the sequence assignment. Thus, they are labeled separately and their sequence assignment further tested in the next section. The majority of PSMs whose confidence is increased by DART-ID have multiple confident Spectral PSMs, and thus reliable sequence assignment. Analysis of newly identified peptides in Fig. 3c shows that DART-ID helps identify about 50% more PSMs compared to spectra alone at an FDR threshold of 1%. This corresponds to an increase of $\sim 40 - 60\%$ in the fraction of PSMs passing an FDR threshold of 1%, as shown in the bottom panel of Fig. 3c. Furthermore, the number of distinct peptides identified per experiment increases from an average of $\sim$1000 to an average of $\sim$1600, Fig. 3d. Percolator also increases identification rates, albeit to a lesser degree than DART-ID, Fig. 3c,d.

Importantly, the total number of false positives in the entire data set changed by $< 1\%$ when deriving the number of false positives from either the Spectral PEP or the DART-ID PEP. This demonstrates that DART-ID is merely separating the true and false positives further apart without changing the underlying assumptions of the data as made by the search engine. Previous methods have criticized the use of Bayes' theorem as a means of incorporating RT evidence because of its tendency to land in "grey areas" between known true positives and known false positives [24]. However, we observe that DART-ID PEPs are bimodaly distributed (Fig. 3e), suggesting that DART-ID performs as an efficient binary classifier. This difference is likely due to the high accuracy of reference RTs estimated by DART-ID. This accuracy results in likelihood functions with small dispersion, and thus substantial changes in the confidence of Spectral PEPs, as evident from Fig. 3a.

Validation of new identifications

We sought to evaluate whether the confident DART-ID PSMs without confident Spectral PSMs are matched to the correct peptide sequences. To this end, we sought to evaluate whether the RTs of

such PSMs match the RTs for the corresponding peptides identified from high-quality, confident spectra. For this analysis, we split a set of experiments into two subsets, A and B, Fig. 4a. The application of DART-ID to A resulted in two disjoint subsets of PSMs: A_1, corresponding to PSMs with confident spectra (Spectral PEP < 0.01), and A_2, corresponding to "upgraded" PSMs (Spectral PEP > 0.01 and DART-ID PEP < 0.01). We overlapped these subsets with PSMs from B having Spectral PEP < 0.01, so that the RTs of PSMs from B can be compared to the RTs of PSMs from subsets A_1 and A_2, Fig. 4a. This comparison, shows excellent agreement of the RTs for both subsets A_1 and A_2 with the RTs for high quality spectral PSMs from B, Fig. 4b,c. This result suggests that even peptides upgraded without confident spectral PSMs are matched to the correct peptide sequences.

Validation by internal consistency

We ran DART-ID on SCoPE-MS experiments [1, 45], all of which contain quantification data in the form of 10-plex tandem-mass-tag (TMT) reporter ion (RI) intensities. Out of the 10 TMT "channels", six represent the relative levels of a peptide in simulated single cells, i.e., small bulk cell lysate diluted to a single cell-level level. These six channels are made from blood cancer cell lines, of which three are from the Jurkat cell line, and the other three from the U-937 cell line. We used the normalized TMT RI intensities to validate upgraded PSMs by analyzing the consistency of protein quantification from distinct peptides.

Internal consistency is defined by the expectation that the relative intensities of PSMs reflect the abundances of their corresponding proteins. If upgraded PSMs are consistent, then their relative RI intensities will have lower coefficients of variation (CV) within a protein than across different proteins [47]. CV is defined as σ/μ, where σ is the standard deviation and μ is the mean of the normalized RI intensities of PSMs belonging to the same protein. A negative control is constructed by creating a decoy dataset where PSM protein assignments are randomized.

For this and later analyses, we filter PSMs from a collection of $\sim$200 experiments into the following disjoint sets:

- *Spectra* – Spectral PEP < 0.01
- *DART-ID* – (Spectral PEP > 0.01) $\cap$ (DART-ID PEP < 0.01)

- *Percolator* – $(\text{Spectral PEP} > 0.01) \cap (\text{Percolator PEP} < 0.01)$

where *Spectra* is disjoint from the other two sets, i.e., Spectra $\cap$ DART-ID $= \emptyset$ and Spectra $\cap$ Percolator $= \emptyset$. These sets of PSMs, as depicted in Fig. 5a, are intersected with each other through a set of shared proteins between the three sets of PSMs.

The protein CVs of the *Spectra*, *DART-ID*, and *Percolator* PSM sets, depicted in Fig. 5b, show similar distributions and smaller CVs than those from the decoy set. In addition, Fig. 5 shows agreement between the protein CVs of the *Spectra* and *DART-ID* PSM sets, as opposed to the CVs of the *Spectra* set and *Decoy* set. This demonstrates that the protein-specific variance in the relative quantification, due to either technical or biological noise, is preserved in these upgraded PSMs.

Proteins identified by DART-ID separate cell types

The upgraded PSMs from the *DART-ID* set are not just representative of proteins already quantified from confident spectral PSMs, but when filtering at a given confidence threshold (e.g., 1% FDR), they allow for the inclusion of new proteins for analysis. As the quantification of these new proteins from the *DART-ID* PSMs cannot be directly compared to that of the proteins from the *Spectra* PSMs, we instead compare how the new proteins from *DART-ID* can explain the biological differences between the two cell types – Jurkat and U-937 – present in each sample and experiment. The data was split into sets in the same manner as the previous section, as shown in Fig. 5a, where the *Spectra* and *DART-ID* sets of PSMs are disjoint. We then filtered out all PSMs from *DART-ID* that belonged to any protein represented in *Spectra*, so that the sets of proteins between the two sets of PSMs were disjoint as well.

We normalized relative RI intensities between the Jurkat and U-937 cell-equivalents in their respective TMT channels. For both the *Spectra* and *DART-ID* PSM sets, we correlated the protein-level RI intensities of the 3 Jurkat and 3 U-937 channels against each other. The resulting correlations in Fig. 6a show similarity within their respective cell type and expected dissimilarity when compared to their corresponding cell type.

The cell-type correlation distributions shown in Fig. 6b from the *DART-ID* set of PSMs are slightly closer to 0 than those from the *Spectra* set. This is likely due, at least in part, to the fact that features that reduce the quality of spectral identifications, such as low peptide abundance and

high coisolation, are likely to undermine quantification. Indeed, we found that *DART-ID* PSMs tend to be less abundant, as shown by the MS1 precursor ion areas, and more likely to contain coisolation, as shown by the precursor ion fraction (PIF), where a lower PIF indicates a higher level of contamination in the MS2 spectra, Fig. 6c. Additionally, the *DART-ID* PSMs are more likely to have a missed cleavage, i.e. they are more susceptible to sample loss and TMT labelling inefficiencies. These factors culminate in a poorer quantification signal and greater fraction of missing data in the *DART-ID* PSMs.

Discussion

Here we present DART-ID as a novel method that improves on existing alignment techniques and utilizes the added alignment accuracy to make peptide RT inferences which are then used to improve peptide sequence identification in the context of ulta-low abundant samples, diluted SCoPE-MS sets. We demonstrate that DART-ID has both the capacity to predict RTs to within seconds for a 60 min LC-MS run, and the ability to increase the number of confidently identified PSMs by up to 50% at 1% FDR.

The motivation to develop DART-ID was the large number of LC-MS/MS analyses on-hand and the high consistency of the LC setup [1, 3]. DART-ID derives more statistical power from more consistent LC setups. Furthermore, since DART-ID alignment takes into account the confidence of PSMs and RT is inferred as a distribution and not a point estimate, the mixture model used for the alignment can derive useful information even from low-confidence PSMs.

There is also motivation to increase the complexity of the alignment model to capture more of the variation within the data. The two-piece linear regression currently used by DART-ID to align experiments captures more variation than a simple linear regression, but DART-ID is not constrained to these two and can implement any monotone function. Non-linear functions that are monotonically constrained, such as the *logit* function, have been implemented in our model in the past. More complex models, such as monotonically-constrained general additive models, could increase alignment accuracy further given that the input data motivates added complexity.

Another possible improvement to alignment accuracy is the removal of peptides with shifting

elution rank. While some physical factors lead to systematic changes in peptide retention across experiments, other factors such as column age and gradient length have been observed to change the elution rank of peptides in a non-systematic fashion. These rank order shifts will align poorly to the monotonically-constrained functions currently used in DART-ID. Future alignments could filter out peptides that have high variance in their elution rank orders, and this could also allow the alignment method to analyze LC experiments with different run times.

While DART-ID is focused on aligning and utilizing RTs from LC-MS experiments, the alignment method could potentially be applied to other separation methods, including but not limited to, gas and supercritical fluid chromatography, 2D liquid chromatography, and capillary electrophoresis. Additionally, ion drift time obtained from instruments with an ion mobility cell could be aligned by DART-ID and applied in the same fashion as it is currently with RT.

Comparing aligned RTs with observed RTs has been used to update spectral confidence previously, but with DART-ID we incorporate the added RT evidence with a straightforwards application of Bayes' theorem. Our method of building conditional RT densities combines information from the spectral error, the RT variance of the experiment, and the uncertainty of the reference RTs (i.e., uncertainty originating from the number of points used to estimate the reference RT). The current version of DART-ID, however, uses all information available in the input set when building the posterior predictive distribution for a peptide in an experiment. For larger sets of experiments, observations from more similar experiments should be preferentially chosen or weighted higher than an observation from a distant, dissimilar experiment. This approach, as well as potentially aligning experiments in clusters of similarity, could allow for the alignment of different chromatographies, and could account for the shifting peptide elution rank order.

While DART-ID is currently limited to updating the best match for an MS2 spectrum as determined by the search engine, a future application is to use RT evidence from the DART-ID alignment in the spectral matching process. Search engines can provide multiple peptide sequence hits for an individual spectrum, and applying the RT evidence to individual hits could save a spectrum from being filtered out due to ambiguity, or change the best hit for a spectrum by boosting a second or third-rank hit. This concept is not new, and has previously been shown to increase identification rates even with RT prediction methods that have significantly higher variance [16, 24, 27].

We present in DART-ID a novel RT alignment that derives statistical power from reproducibility in liquid chromatography runs, and an implementation into an established Bayesian framework from which the peptide identification confidence can be updated using RT evidence. The alignment method provided is a simple implementation of the concept, and has potential to become more complex and capture more variance. The overall concept is also applicable to other features of ions, such as MS1 intensity, that can be integrated into the identification confidence of peptides. DART-ID significantly boosts the proteome coverage of single cell proteomic sets, without having to make any experimental changes or considerations.

Methods

Data sources and experimental design

The data used for the development and validation of the DART-ID method were mostly method-development experiments for SCoPE-MS and its related projects. All samples were digested with trypsin. All experiments used either 10 or 11-plex TMT for quantification. The experiments used to generate figures for this paper were run on a Thermo Fisher Easy-nLC system with a Waters 25cm x 75µm, 1.7µ BEH column with 130Å pore diameter. Gradients were run from 5-35%B in 48 minutes with a 12 minute wash step to 100%B. Solvent composition was 0% acetonitrile for A and 80% acetonitrile for B, with 0.1% formic acid in both. A subset of later experiments included the use of a trapping column, which extended the total run-time to 70 minutes.

Searching raw MS data

Searching was done with MaxQuant v1.6.1.0 [31] against a UniProt protein sequence database with 443722 entries. The database contained only SwissProt entries and was downloaded on 5/1/2018. Searching was also done on a contaminant database provided by MaxQuant, which contained contaminants such as common laboratory contaminants and keratins. MaxQuant was run with Trypsin specificity which allowed for two missed cleavages, and methionine oxidation and protein N-terminus acetylation as variable modifications and no fixed modifications apart from TMT tags.

Observations were selected at a false discovery rate (FDR) of 100% at both the protein and PSM level to obtain as many spectrum matches as possible, regardless of their spectral confidence. All raw files, MaxQuant output files, and search parameters will be made available upon request.

Data filtering

Only a subset of the input data is used for the alignment of experiments and the inference of RT distributions for peptides. First, decoys and contaminants are filtered out of the set. Contaminants are especially problematic for RT alignment as their retention is not strongly defined. Then, observations are selected at a threshold of PEP < 0.5.

Observations are additionally filtered through a threshold of retention length, which is defined by MaxQuant as the range of time between the first matched scan of the peptide and the last matched scan. Any peptide with retention length > 1 min for a 60 min run is deemed to have too wide of an elution peak, or chromatography behavior more consistent with contaminants than retention on column.

Only peptide sequences present in at least two experiments are passed through to the alignment process. The model can allow peptides only present in one experiment to be included in the alignment, but the inclusion of this data adds no additional information to the alignment and only serves to slow it down. For our set of 226 experiments, only peptide sequences present in 10 or more experiments were allowed to participate in the alignment process. The definition of a peptide sequence in these cases is dynamic, and can include modifications, charge states, or any other feature that would affect the retention of an isoform of that peptide.

Preliminary alignments revealed certain experiments where chromatography was extremely abnormal, or where peptide identifications were sparse enough as to prevent an effective alignment. These experiments were manually removed from the alignment procedure in future runs of DART-ID.

Global alignment model

Let ρ_{ik} be the RT assigned to peptide i in experiment k. In order to infer peptide and experiment-specific RT distributions, we assume that there exists a set of reference retention times, μ_i, for all peptides. Each peptide has a unique reference RT, independent of experiment. We posit that for each experiment, there is a simple monotone increasing function, g_k, that maps the reference RT to the predicted RT for peptide i in experiment k. An observed RT can then be expressed as

$$\rho_{ik} = \mu_{ik} + \epsilon_{ik}$$

where $\mu_{ik} \triangleq g_k(\mu_i)$ and ϵ_{ik} is an independent mean-zero error term expressing residual (unmodeled) RT variation. As a first approximation, we assume that the observed RTs for any experiment can be well approximated using a two-segment linear regression model:

$$g_k(\mu_i) = \begin{cases} \beta_{0k} + \beta_{1k}\mu_i & \text{if } \mu_i < s_k \\ \beta_{0k} + \beta_{1k}s_k + \beta_{2k}(\mu_i - s_k) & \text{if } \mu_i \geq s_k \end{cases}$$

where s_k is the split point for the two segment regression in each experiment, and the parameters are constrained to not produce a negative RT.

To factor in the spectral PEP given by the search engine, and to allow for the inclusion of low probability PSMs, the marginal likelihood of an RT in the alignment process can be described using a mixture model. For a PSM assigned to peptide i in experiment k the RT density is

$$P(\rho_{ik}) \propto \mathbf{1}\{\rho_{ik} > 0\} \left((1 - \lambda_{ik}) \times f_{ik}(\rho_{ik} \mid \mu_{ik}, \sigma_{ik}) + (\lambda_{ik}) \times f_k^0(\rho_{ik})\right) \tag{1}$$

where λ_{ik} is the error probability (PEP) for the PSM returned by MaxQuant, f_{ik} is the inferred RT density for peptide i in experiment k and f_k^0 is the null RT density. In our implementation, we let $f_{ik} \sim \text{Laplace}(\mu_{ik}, \sigma_{ik}^2)$ and $f_k^0 \sim \text{Normal}(\mu_k, \sigma_k^2)$, which we found worked well in practice. However, our framework is modular and it is straightforward to utilize different residual RT and null distributions if appropriate. For example, with non-linear gradients that generate more a uniform distribution of peptides across the LC run [17], it may be sensible for the null distribution to be

uniformly distributed, i.e. $f_k^0 \sim \text{Uniform}(\text{ RT}_{min},\ \text{RT}_{max})$.

Finally, to reflect the fact that residual RT variation increases with mean RT, we model the standard deviation of a peptide RT distribution, σ_{ik}, as a linear function of the reference RT:

$$\sigma_{ik} = a_k + b_k \mu_i$$

where μ_i is the reference RT of the peptide sequence, and a_k and b_k are the intercept and slope which we infer for each experiment. a_k, b_k and μ_i are constrained to be positive, and hence $\sigma_{ik} > 0$ as well.

Alignment Comparison

We compared the DART-ID alignment accuracy against five other RT prediction or alignment algorithms. As some methods returned absolute predicted RTs (such as BioLCCC) and others returned relative hydrophobicity indices (such as SSRCalc), a linear regression was built for each prediction method. Alignment accuracy was evaluated using three metrics: R^2, the Pearson correlation squared, and the mean and median of $|\Delta\text{RT}|$, the absolute value of the residual RT, which is defined as $(\text{Observed RT} - \text{Predicted RT})$. We only selected confident PSMs (PEP < 0.01) for this analysis, and used data that consisted of 33383 PSMs from 46 LC-MS/MS experiments run over the course of 90 days.

SSRCalc was run from the SSRCalc Online (http://hs2.proteome.ca/SSRCalc/SSRCalcQ.html), with the "100Å C18 column, 0.1% Formic Acid 2015" model, "TMT" modification, and "Free Cysteine" selected. No observed RTs were inputted along with the sequences.

BioLCCC was run online from http://www.theorchromo.ru/ with the parameters of 250mm column length, 0.075mm column inner diameter, 130Å packing material pore size, 5% initial concentration of component B, 35% final concentration of component B, 48 min gradient time, 0 min delay time, 0.0001 ml/min flow rate, 0% acetonitrile concentration in component A, 80% acetontrile concentration in component B, "RP/ACN+FA" solid/mobile phase combination, and no cysteine carboxyaminomethylation. As BioLCCC could only take in one gradient slope as the input, all peptides with observed RT $>$ 48 min were not inputted into the prediction program.

ELUDE was downloaded from the percolator releases page https://github.com/percolator/percolator/releases. The 33383 PSMs were split into two, equal sets with distinct peptide sequences to form the training and test sets. The `elude` program was run with the `--no-in-source` and `--test-rt` flags. Predicted RTs from ELUDE were obtained from the testing set only, and training set RTs were not used in further analysis.

For iRT, the same raw files used for the previous sets were searched with the Pulsar search engine, with iRT alignment turned on and filtering at 1% FDR. From the Pulsar search results, only peptide sequences in common with the previous set searched in MaxQuant were selected. Predicted RT was taken from the "PP.RTPredicted" column and plotted against the empirical RT column "PP.EmpiricalRT". Empirical RTs were not compared between those derived from MaxQuant and those derived from Pulsar.

MaxQuant match-between-runs was run by turning the respective option on when searching over the set of 46 experiments, and given the options of 0.7 min match time tolerance and a 20 min match time window. The "Calibrated retention time" column was used as the predicted RT, and these predicted RTs were related to observed RTs with a linear model for each experiment run.

For DART-ID, predicted RTs are the same as the mean of the inferred RT distribution, and no linear model was constructed to relate the predicted RTs to the observed RTs.

Confidence update

We update the confidence for PSM i in experiment k according to Bayes' theorem. Let $\delta_{ik} = 1$ denote that PSM i in experiment k is assigned to the correct sequence (true positive), $\delta_{ik} = 0$ denotes that the PSM is assigned to the incorrect sequence (a false positive), and as above, ρ_{ik} is an observed RT assigned to peptide i. At a high level, the probability that the peptide assignment is as true positive is

$$P(\delta_{ik} = 1 \mid \rho_{ik}) = \frac{P(\rho_{ik} \mid \delta_{ik} = 1) \times P(\delta_{ik} = 1)}{P(\rho_{ik})} \tag{2}$$

Each term is described in more detail below:

δ_{ik} An indicator for whether or not the peptide sequence assignment, i in

experiment k is correct (i.e. a true or false positive).

$P(\delta_{ik} = 1|\rho_{ik})$ The posterior probability that the PSM is assigned to the right sequence, given the observed RT, ρ_{ik}.

$P(\rho_{ik} \mid \delta_{ik} = 1)$ The RT density for peptide i in experiment k given the assignment is correct (true positive). Conditional on the alignment parameters, the true positive RT density $f_{ik}(\rho_{ik} \mid \mu_{ik}, \sigma_{ik})$ is Laplace(mu_{ik}, σ^2_{ik}).

$P(\rho_{ik} \mid \delta_{ik} = 0)$ The RT density given the assignment is incorrect (false positive). We assume that a false positive match is assigned to a peptide at random and thus take $f^0_k(\rho_{ik})$ to be a broad distribution reflecting variation in *all* RTs in experiment k. We model this distribution as Normal(μ_k, σ^2_k), where μ_k is approximately the average of all RTs in the experiment and σ^2_k is the variance in RTs.

$P(\delta_{ik} = 1)$ The prior probability that the PSM's assigned sequence is correct, i.e. one minus the posterior error probability (PEP) provided by MaxQuant, $1 - \lambda_{ik}$.

$P(\rho_{ik})$ The marginal likelihood for observing the RT assigned to peptide i in experiment k. By the law of total probability, this is simply the mixture density from Equation 1.

The confidence update depends on the global alignment parameters. Let θ consist of the global alignment parameters and reference RTs, i.e. $\beta_{0k}, \beta_{1k}, \sigma_{ik}$ and μ_i. If θ were known, then the Bayesian update could be computed in a straightforward manner as described above. In practice the alignment parameters are not known and thus must be estimated using the full set of observed RTs across all experiments, $\boldsymbol{\rho}$. The PSM confidence update can be expressed unconditional on θ, by integrating over the uncertainty in the estimates of the alignment parameters:

$$(\delta_{ik} = 1 \mid \boldsymbol{\rho}) = \int p(\delta_{ik} = 1 \mid \rho_{ik}, \theta) p(\theta \mid \boldsymbol{\rho}) d\theta_{ik} \tag{3}$$

Although we can estimate this posterior distribution using Markov Chain Monte Carlo (MCMC), it is prohibitively slow given the large number of peptides and experiments that we analyze. As

such, we estimate maximum a posteriori (MAP) estimates for the reference RTs μ_i, alignment parameters β_{0k}, β_{1k}, and RT standard deviation σ_{ik} using an optimization routine implemented in `STAN` [48][1]. This approach is computationally efficient but is limited in that parameter uncertainty quantification is not automatic.

To address this challenge, we incorporate estimation uncertainty using a computationally efficient procedure based on the parametric bootstrap. Note that uncertainty about the alignment parameters β_{0k} and β_{1k} is small since they are inferred using thousands of RT observations per experiment. By contrast, the reference RTs, μ_i, have much higher uncertainty since we observe *at most* one RT associated with peptide i in each experiment (usually far fewer). As such, we choose to ignore uncertainty in the alignment parameters and focus on incorporating uncertainty in estimates of μ_i.

Let $\hat{\mu}_{ik}$ and $\hat{\sigma}_{ik}$ denote the MAP estimates of the location and scale parameters for the RT densities. To approximate the posterior uncertainty in the estimates of μ_i, we use the parametric bootstrap. First, we sample $\rho_{ik}^{(b)}$ from $f_{ik}(\rho_{ik} \mid \hat{\mu}_{ik}, \hat{\sigma}_{ik})$ with probability $1 - \lambda_{ik}$ and $f_k^0(\rho_{ik})$ with probability λ_{ik}. We then map $\rho_{ik}^{(b)}$ back to the reference space using the inferred alignment parameters as $\hat{g}^{-1}(\rho_{ik})$ and compute a bootstrap replicate of the reference RT associated with peptide i as the median (across experiments) of the resampled RTs: $\mu_i^{(b)} = \underset{k}{\text{median}}\ \hat{g}^{-1}(\rho_{ik}^{(b)})$[2]. For each peptide we repeat this process B times to get several bootstrap replicates of the reference RT for each peptide. We use the bootstrap replicates to incorporate the uncertainty of the reference RTs into the Bayesian update of the PSM confidence. Specifically, the we approximate the confidence update in Equation 3 as

$$
\begin{aligned}
p(\delta_{ik} = 1 \mid \rho_{ik}) &\approx \frac{1}{B} \sum_{b=1}^{B} p(\delta_{ik} = 1 \mid \rho_{ik}, \mu_{ik}^{(b)}, \hat{\sigma}_{ik}) \\
&= \frac{(1 - \lambda_{ik}) \left(\frac{1}{B} \sum_{b=1}^{B} f_{ik}(\rho_{ik} \mid \mu_{ik}^{(b)}, \hat{\sigma}_{ik} \right)}{(1 - \lambda_{ik}) \left(\frac{1}{B} \sum_{b=1}^{B} f_{ik}(\rho_{ik} \mid \mu_{ik}^{(b)}, \hat{\sigma}_{ik} \right) + \lambda_{ik} f_k^0(\rho_{ik})}
\end{aligned} \tag{4}
$$

[1] If computation time is not a concern, is straightforward to generate posterior samples in our model by running MCMC sampling in STAN, instead of MAP optimization.

[2] The maximum likelihood estimate of the location parameter of a Laplace distribution is the median of independent observations.

This process is depicted in Fig. S2.

TMT reporter ion intensity normalization

Reporter ion (RI) intensities were obtained by selecting the tandem-mass-tag (TMT) 10-plex labels in MaxQuant, for both attachment possibilities of lysine and the peptide N-terminus, and with a mass tolerance of 0.01 Da. Data from different experiments and searches are all combined into one matrix, where the rows are observations (PSMs) and the 10 columns are the 10 TMT channels. Observations are filtered at a confidence threshold, normally 1% FDR, and observations with missing data are thrown out.

Before normalization, two channels, 127N and 128C, are removed from the matrix. Each column of the matrix is divided by the median of that column, to correct for the total amount of protein in each channel and any biases between the respective TMT tags. Then, each row of the matrix is divided by the median of that row, to obtain the relative enrichment between the samples in the different TMT channels.

Assuming that the relative RI intensities of PSMs are descriptive of their parent peptide, the peptide intensity can be estimated as the median of the RI intensities of its constituent PSMs. Similarly, if protein levels are assumed to correspond to the levels of its constituent peptides, then protein intensity can be estimated as the median of the intensities of its constituent peptides. The previous steps of RI normalization makes all peptide and protein-level quantitation relative between the conditions in each channel.

Protein inference

Any interpretation of the DART-ID data on the protein-level was first run through the Fido protein inference algorithm [49], which gives the probability of the presence of a protein in a sample given the pool of observed peptides and the probabilities of their constituent PSMs. The Python port of Fido was downloaded from `https://noble.gs.washington.edu/proj/fido` and modified to be compatible with Python 3. The code was directly interfaced into DART-ID and is available to run as a user option.

For the data in this paper, protein-level analyses first had their proteins filtered at 1% FDR, where the FDR was derived from the probabilities given to each protein by the Fido algorithm.

Implementation

The DART-ID pipeline is roughly divided into three parts. First, input data from search engine output files are converted to a common format, and PSMs unsuitable for alignment are marked for removal. Second, initial values are generated by running a simple estimation of reference RTs and linear regression parameters for f_{ik} for each experiment. The initial values and the data are fed into the alignment model. Third, alignment parameters from the alignment model are used to update the PEP of a PSM.

The model was implemented using the STAN modeling language [50]. All densities were represented on the log scale. STAN was interfaced into an R script with `rstan`. STAN was used with its `optimizing` function, which gave maximum a posteriori (MAP) estimates of the parameters, as opposed to sampling from the full posterior. R was further used for data filtering, PEP updating, model adjustment, and figure creation. The code is also ported to Python3 and `pystan`, and is available as a `pip` package `DART-ID` that can be run from the command-line. `DART-ID` is run with a configuration file that specifies inputs and options. All model definitions and related parameters such as distributions are defined in a modular fashion, which supports the addition of other models or fits.

All code, instructions, and figure scripts are made available on GitHub: https://github.com/blahoink/DART-ID. The data are available from a supplemental data website https://web.northeastern.edu/slavovlab/2018_DART-ID/

Acknowledgments

We thank H. Specht, T. Chen, and members of the Slavov laboratory for discussions and constructive comments. This work was funded by the National Institute of Health to N.S. under project number 1DP2GM123497-01.

Competing interests: The authors declare that they have no competing financial interests.

Corresponding author: Correspondence and requests for materials should be addressed to N.S. (nslavov@northeastern.edu)

Author Contributions

Conceptualization A.F. and N.S.

Data Curation A.C.

Formal Analysis A.F. and A.C.

Funding Acquisition N.S.

Investigation A.C. and A.F.

Methodology A.F. and N.S.

Project Administration N.S.

Resources N.S.

Software A.C. and A.F.

Supervision N.S.

Validation A.C.

Visualization A.C.

Writing - Original Draft Preparation A.C.

Writing - Review & Editing A.C., A.F., and N.S.

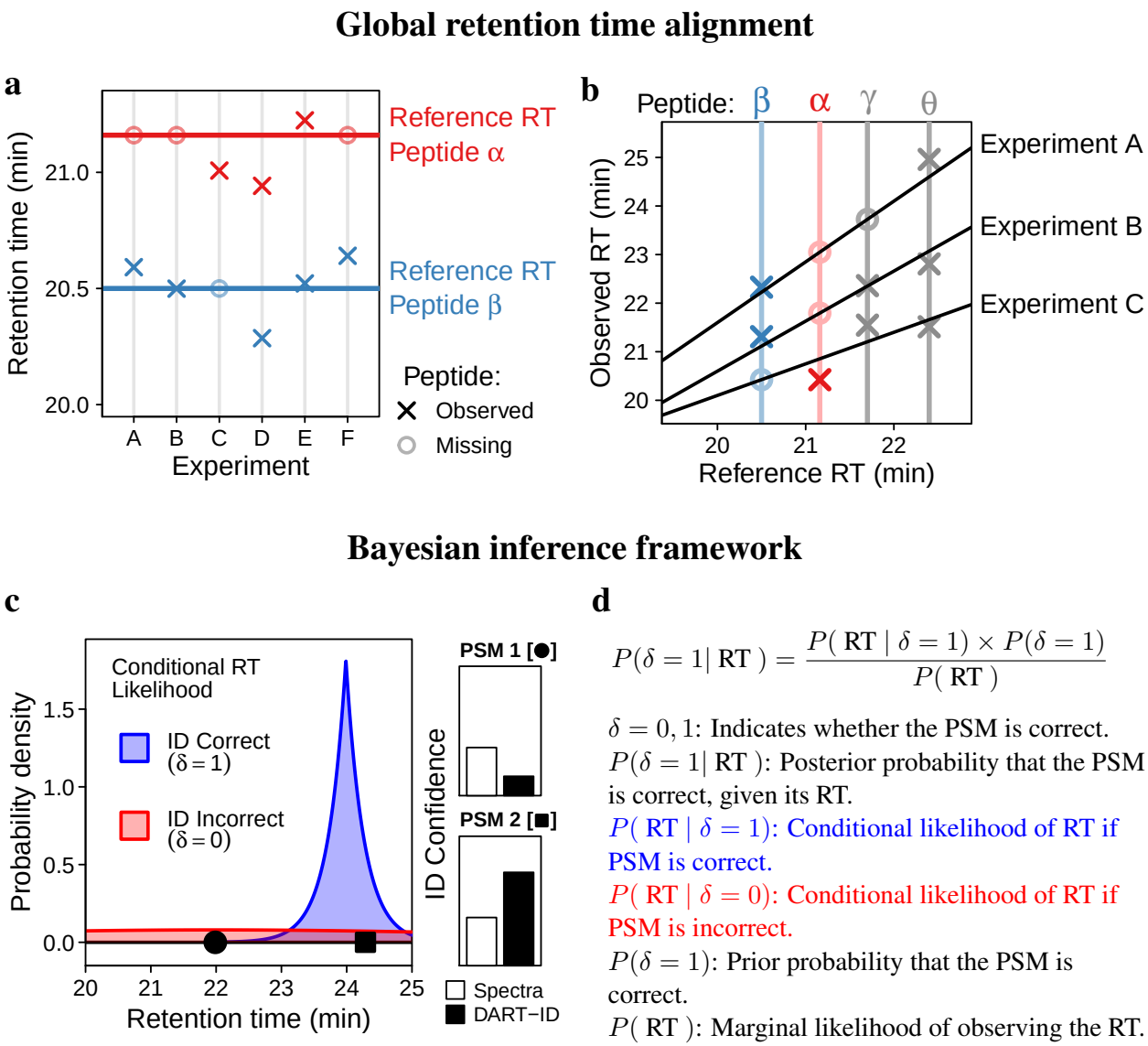
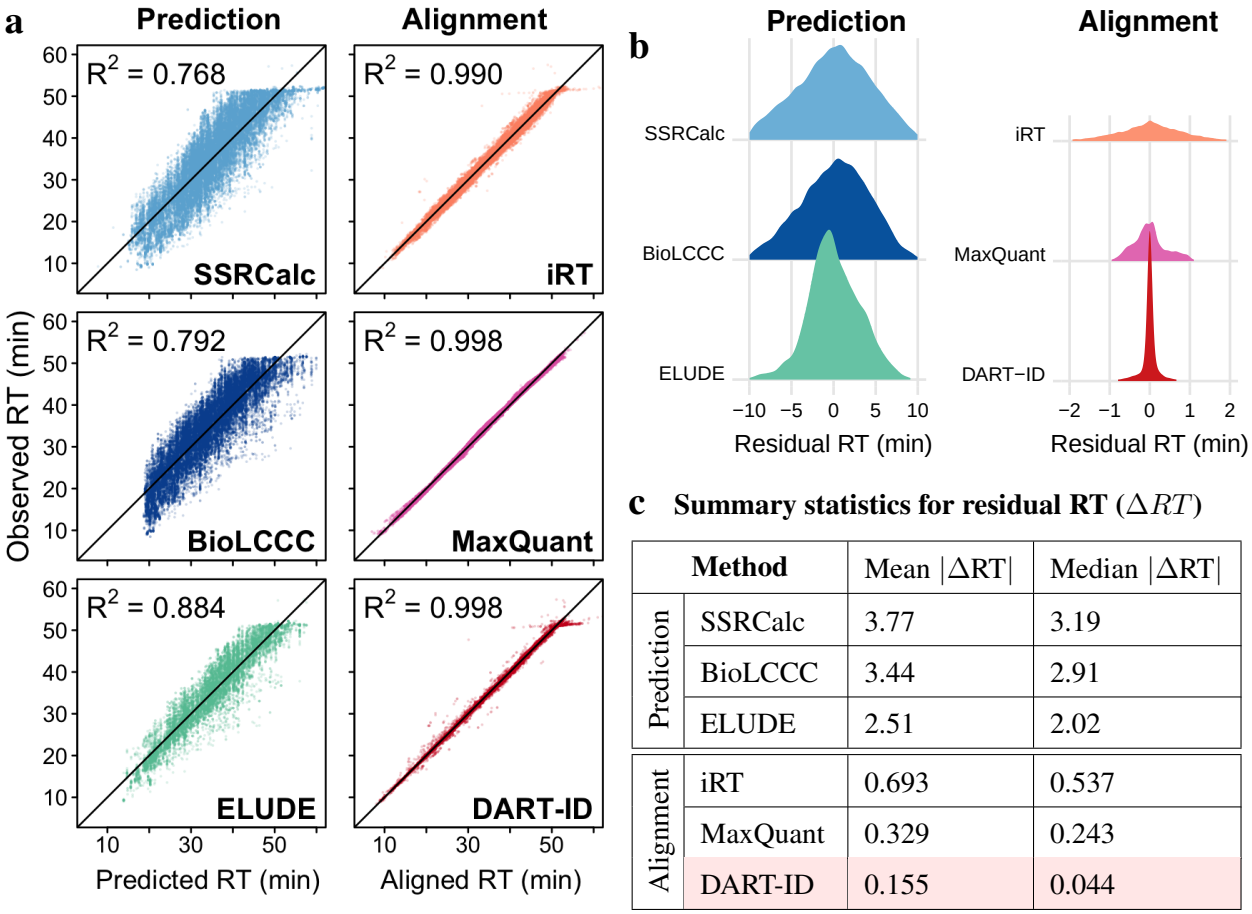


Figure 1 | Global retention time alignment method and Bayesian inference framework
(**a**) DART-ID defines the global reference RT as a latent variable. It is initialized as the average RTs of the observed RTs from the unaligned experiments. (**b**) The global alignment model aligns all experiments simultaneously to the reference RT and improves its estimate. The model incorporates experiment specific weights amd takes into account uncertainty in measured RTs and peptide identification. See Methods for details. (**c**) A conceptual diagram for updating the confidence in a peptide-spectrum-match (PSM). The probability to observe each PSM is estimated from the conditional likelihoods for observing the RT if the PSM is assigned correctly (blue density) or incorrectly (red density). For PSM 1, $P(\delta = 1 \mid \text{RT}) < P(\delta = 0 \mid \text{RT})$, and thus the confidence decreases. Conversely, for PSM 2, $P(\delta = 1 \mid \text{RT}) > P(\delta = 0 \mid \text{RT})$, and thus the confidence increases. (**d**) The Bayes' formula used to formalize the model from panel c and to update the error probability of PSMs.



c **Summary statistics for residual RT** (ΔRT)

	Method	Mean $\lvert\Delta\text{RT}\rvert$	Median $\lvert\Delta\text{RT}\rvert$
Prediction	SSRCalc	3.77	3.19
	BioLCCC	3.44	2.91
	ELUDE	2.51	2.02
Alignment	iRT	0.693	0.537
	MaxQuant	0.329	0.243
	DART-ID	0.155	0.044

Figure 2 | Comparison of inferred reference RTs to empirical RTs
(**a**) Scatter plots of observed RTs versus inferred RTs. The comparisons include 33,383 PSMs with PEP < 0.01 from 46 LC-MS/MS runs over the span of 90 days. The left column displays comparisons for RT prediction methods – SSRCalc, BioLCCC, and ELUDE. The right column displays comparisons for alignment methods – iRT, MaxQuant match-between-runs, and DART-ID. Note the different scales of the x-axes. (**b**) Distributions of residual RTs (Observed RT − Reference RT). (**c**) Mean and median of the absolute values from panel (b).

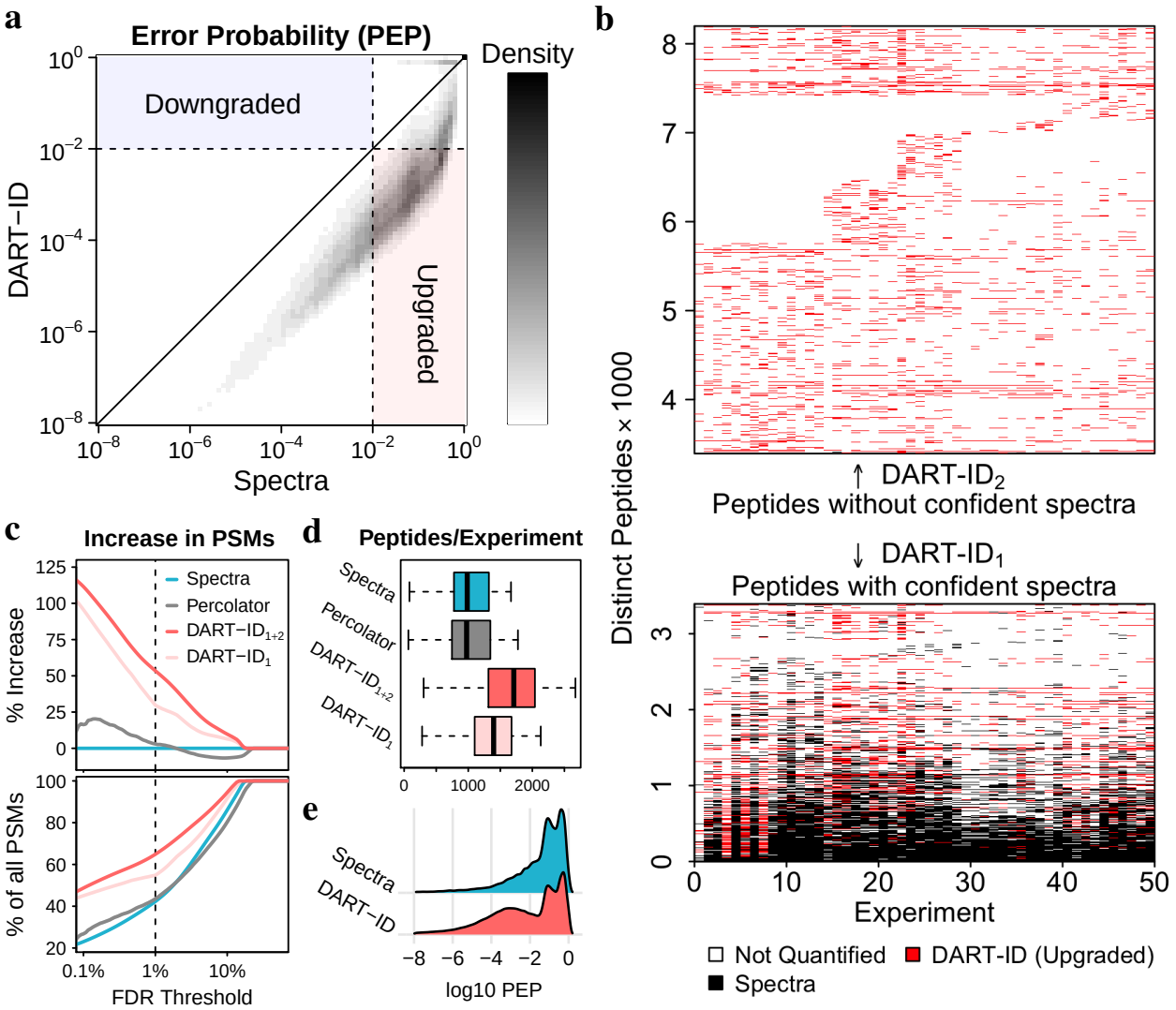


Figure 3 | Incorporating RTs increases confident peptide identifications
(**a**) A 2D density distribution of error probabilities derived from spectra alone (Spectral PEP), compared to that after incorporating RT evidence (DART-ID PEP). (**b**) Map of all peptides observed across all experiments. Black marks indicate peptides with Spectral FDR $< 1\%$, and red marks peptides with DART-ID FDR $< 1\%$. (**c**) Increase in confident PSMs (top), and in the fraction of all PSMs (bottom) across the confidence range of the x-axis. The curves correspond to PEPs estimated from spectra alone, from spectra and RTs using percolator and from spectra and RTs using DART-ID. DART-ID identifications are split into DART-ID$_1$ and DART-ID$_2$ depending on whether the peptides have confident spectral PSMs as marked in panel (b). (**d**) Distributions of number of unique peptides identified per experiment. (**e**) Distributions of Spectral PEPs and DART-ID PEPs. The bimodality of the DART-ID distribution suggests that DART-ID's use of RTs helps cleanly separate incorrect from correct PSMs.

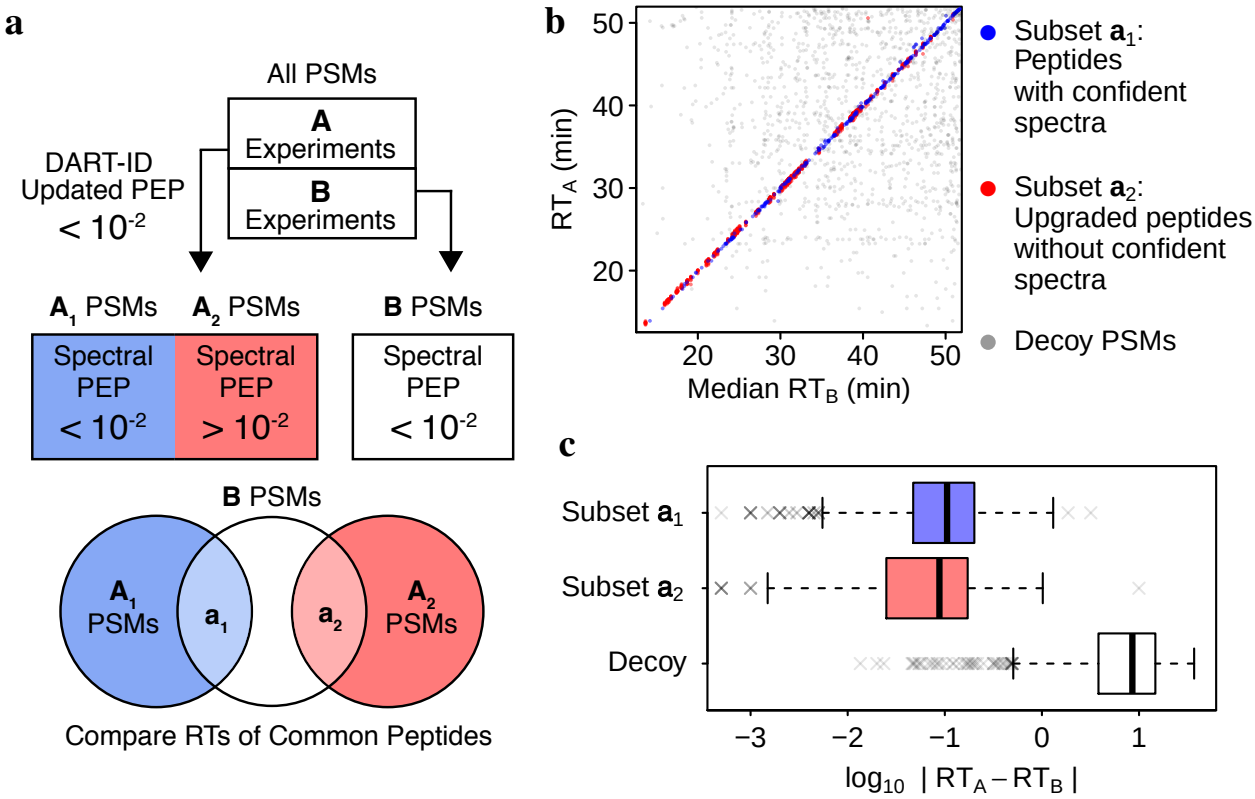


Figure 4 | Validation of newly identified peptides with retention time of technical replicates (**a**) Schematic design of this validation experiment. It used 11 technical replicate LC-MS/MS experiments that were run on the same day. (**b**) Comparison of the RTs of subsets a_1 and a_2 to the RTs of corresponding peptides from B. Decoy PSMs have randomly sampled RTs and are included here as a null model. (**c**) Residual RT distributions for the two subsets of data a_1 and a_2 as defined in panel a and for a decoy subset.

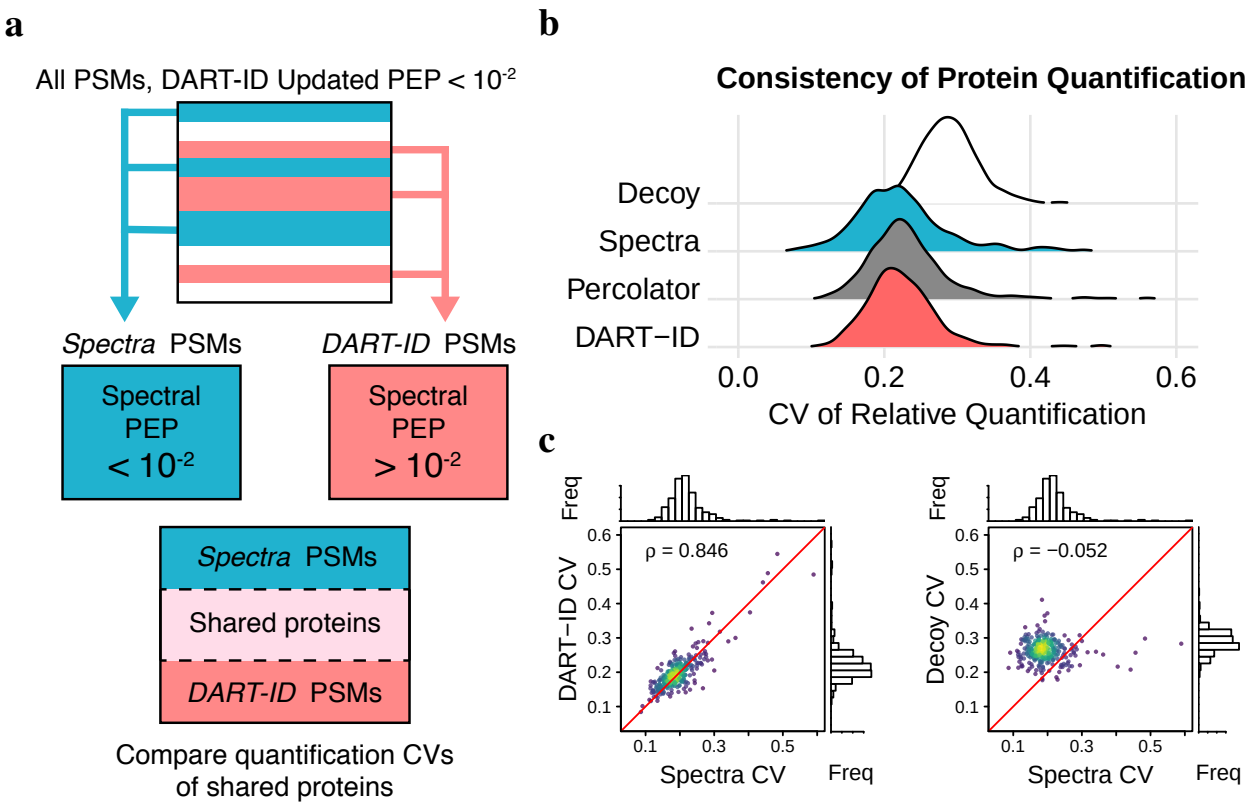


Figure 5 | Validation of boosted PSMs by internal consistency
(**a**) Schematic for separating PSM subsets, where *Spectra* and *DART-ID* subsets of PSMs are disjoint. (**b**) Distributions of coefficient of variation (CVs) for each protein in each subset. The *Decoy* subset is a subset of PSMs with their protein assignments randomized. (**c**) Comparing protein CVs of $n = 275$ proteins between the *Spectra* and *DART-ID* PSM subsets, and from the *Spectra* and *Decoy* subsets.

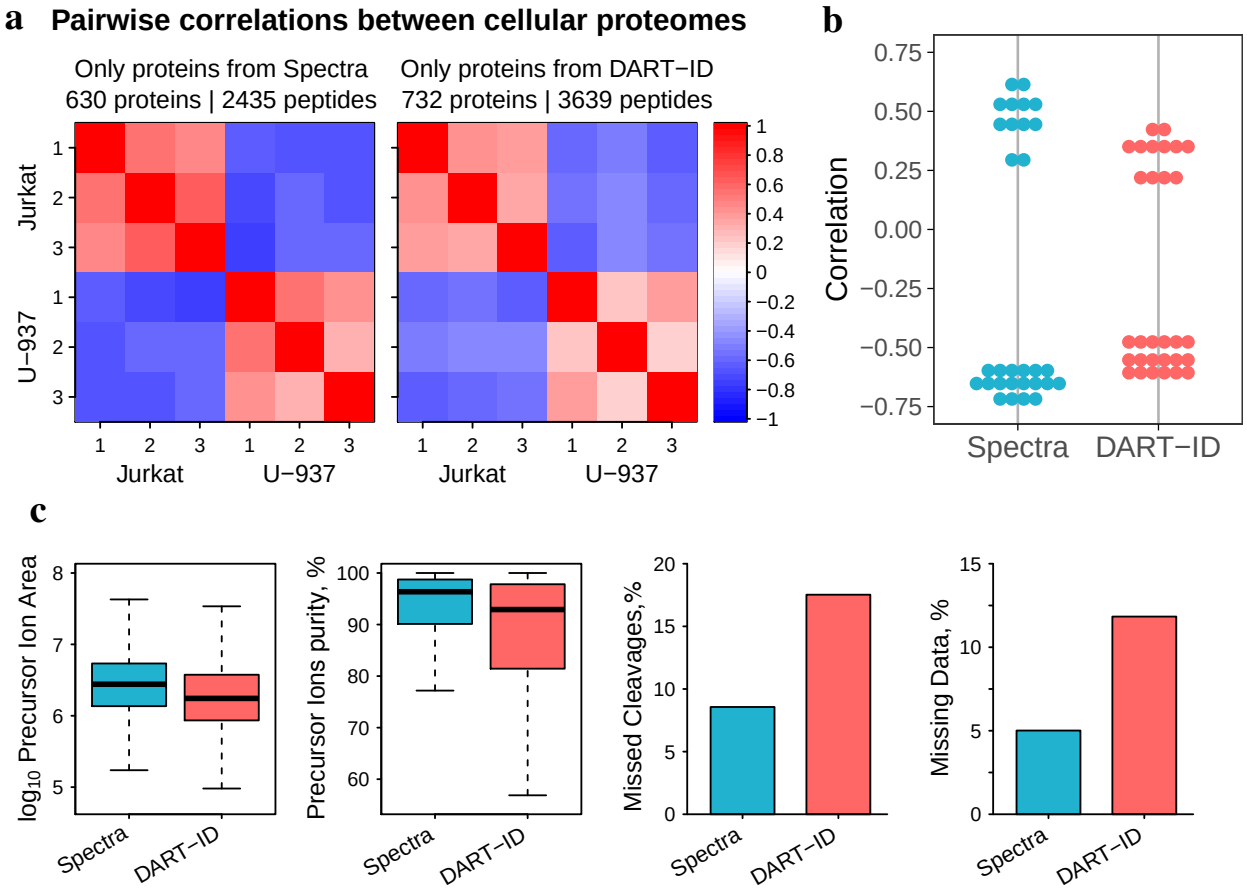


Figure 6 | Quantification of proteins identified by spectra alone and by DART-ID
(**a**) Pairwise correlations between relative protein levels of 2 blood cancer cell lines, Jurkat and U-937. The *Spectra* set contains proteins with Spectral PSMs filtered at 1% FDR, and the *DART-ID* set contains a disjoint set of proteins quantified from PSMs with high Spectral PEP but low DART-ID PEP. (**b**) Distributions of the correlations from panel (a). (**c**) Features of the *Spectra* and *DART-ID* sets of PSMs, where precursor ion area is the area under the MS1 elution peak, precursor ion fraction is a measure of MS2 spectral purity, missed cleavages is the average number of internal lysine and arginine residues, and % missing data is the average fraction of missing TMT reporter ion quantitation per PSM.

Supporting Information

Mixture Model for RT Likelihood

$$P(\rho_{ik} | \, \delta_{ik} = 1) = f_{ik}(\rho_{ik})$$

$$P(\rho_{ik} | \, \delta_{ik} = 0) = f_k^0(\rho_{ik})$$

$$f_{ik} \sim \text{Laplace}(\rho_{ik} \mid \mu_{ik}, \sigma_{ik}^2)$$
$$f_k^0 \sim \text{Normal}(\rho_{ik} \mid \mu_k, \sigma_k^2)$$

δ_{ik} – Indicator of true positive (1) or false positive (0) for peptide ik
ρ_{ik} – RT for peptide ik
λ_{ik} – Error probability (PEP) for peptide ik
μ_{ik} – Aligned canonical RT for peptide i in experiment k.
σ_{ik} – Standard deviation of the positive portion of the inferred RT distribution for peptide i in experiment k
μ_k, σ_k – Mean and standard deviation of all retention times in experiment k.

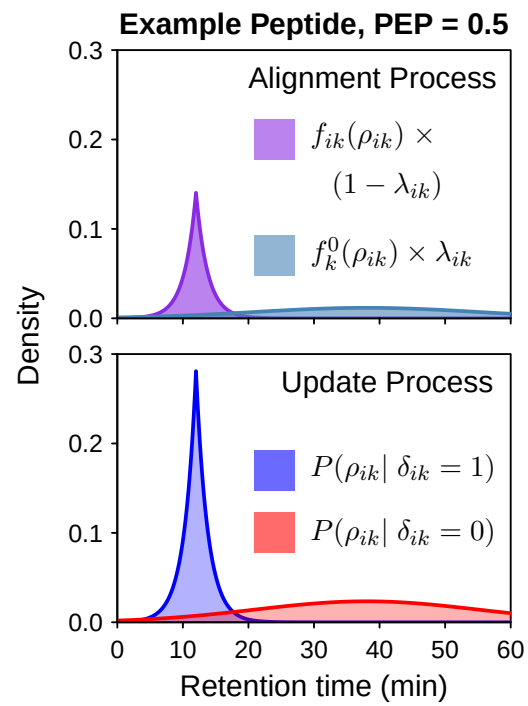


Figure S1 | Mixture model incorporates spectral confidence to estimate likelihood of observing retention times.
In the global alignment process, the likelihood of the alignment function and the reference RT is estimated from a mixture model, which combines the two possibilities of whether the peptide is assigned the correct or incorrect peptide sequence. These two distributions are then weighted by the error probability (PEP). This is similar to the update process, which updates the error probability and incorporates the previous error probability, as well as the two conditional probability distributions.

a

Experiment	Peptide α	RT (min)	Aligned RT (min)	Std (min)	PEP
A	EQSAAER	35.12	30.11	0.10	0.05
C	EQSAAER	26.25	30.02	0.02	0.48
F	EQSAAER	27.82	30.21	0.22	0.42
M	EQSAAER	32.99	29.94	0.15	0.07
Y	EQSAAER	33.14	29.97	0.05	0.13

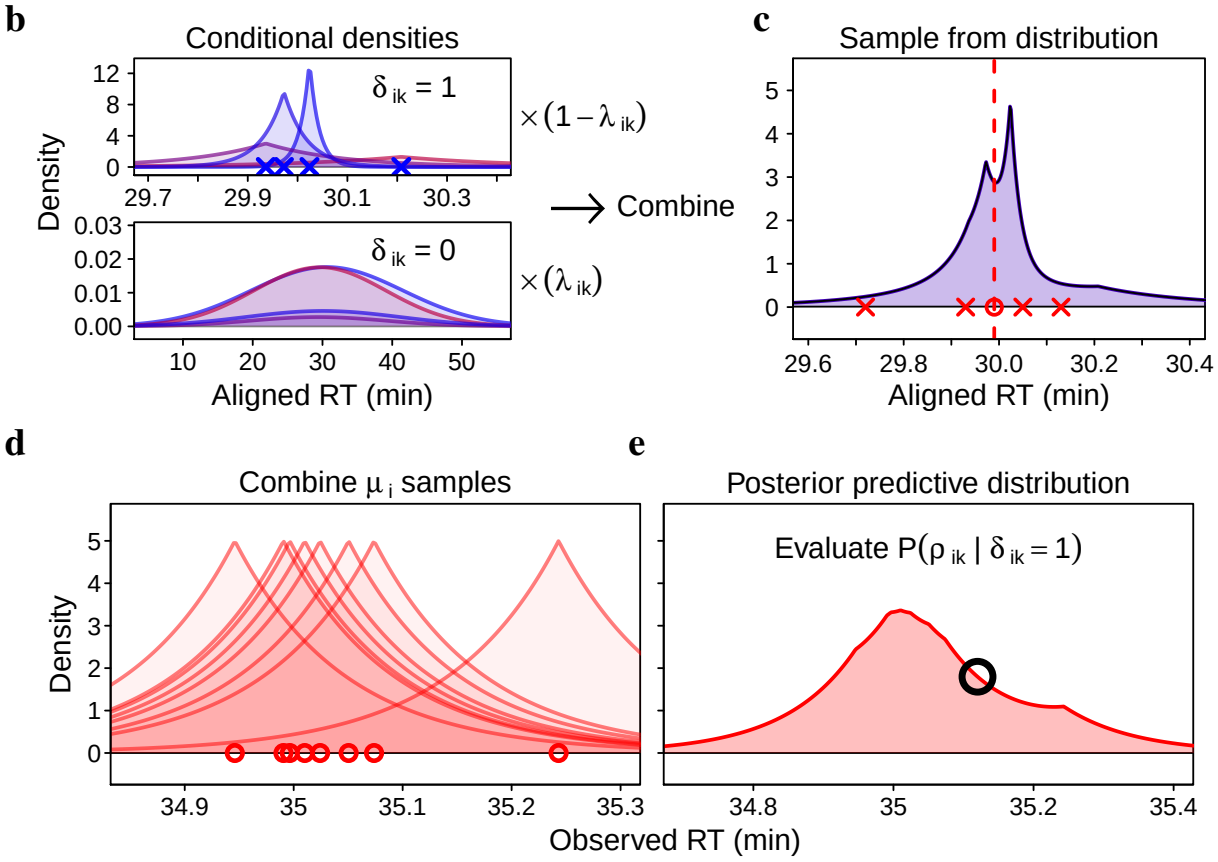


Figure S2 | Deriving conditional probability of RT given a correct match
(**a**) The conditional probability distribution of RT given a correct peptide sequence assignment incorporates evidence about that peptide sequence across many different experiments. "Aligned RT" is the RT after applying the alignment function, and "Std" is inferred RT standard deviation for the peptide in the given experiment. (**b**) For each RT observation for a sequence in an experiment, we infer two distributions: one corresponding to RT density given a correct PSM and the other to an incorrect PSM match. These densities are weighted by the 1-PEP and the PEP respectively and summed to produce the marginal RT distribution. (**c**) The marginal RT distribution is then used to sample B bootstrap replicates of of the observed RTs. Each bootstrapped RT is then used to construct a bootstrapped reference RT for a given sequence. The reference RT is the median of the resampled RTs (inn the aligned space). (**d**) The B bootstrap samples of μ_i are used to build distributions where the variance is determined by the model-derived variance of the peptide in an experiment. (**e**) The combination of the distributions in panel (d) forms a posterior predictive distribution for the observed RT, given that the peptide sequence assignment is correct.

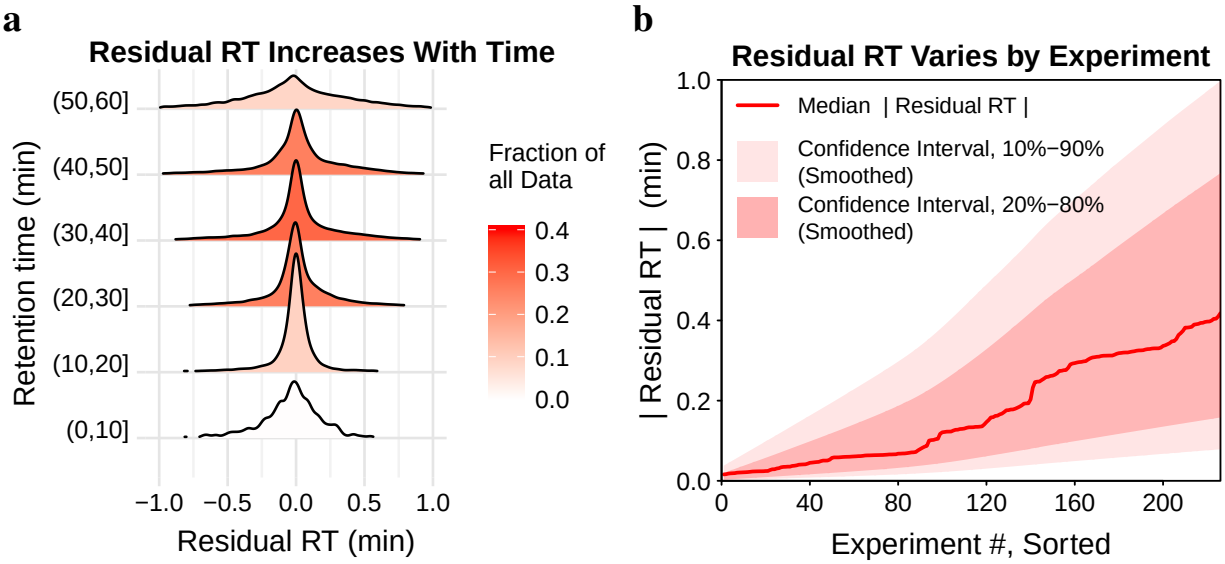


Figure S3 | Accuracy of RT inferences varies with time and between experiments. (**a**) Residual RT (observed RT - aligned RT) binned by RT for 60 min LC-MS runs. The gradient run is 5 – 35%B from 0 – 48 min, with a wash step of 35 – 100%B from 48 – 60 min. (**b**) Residual RT varying between different experiments. Experiments are 226 LC-MS experiments, with run times of 60 min each. The experiments were run over the course of 3 months.

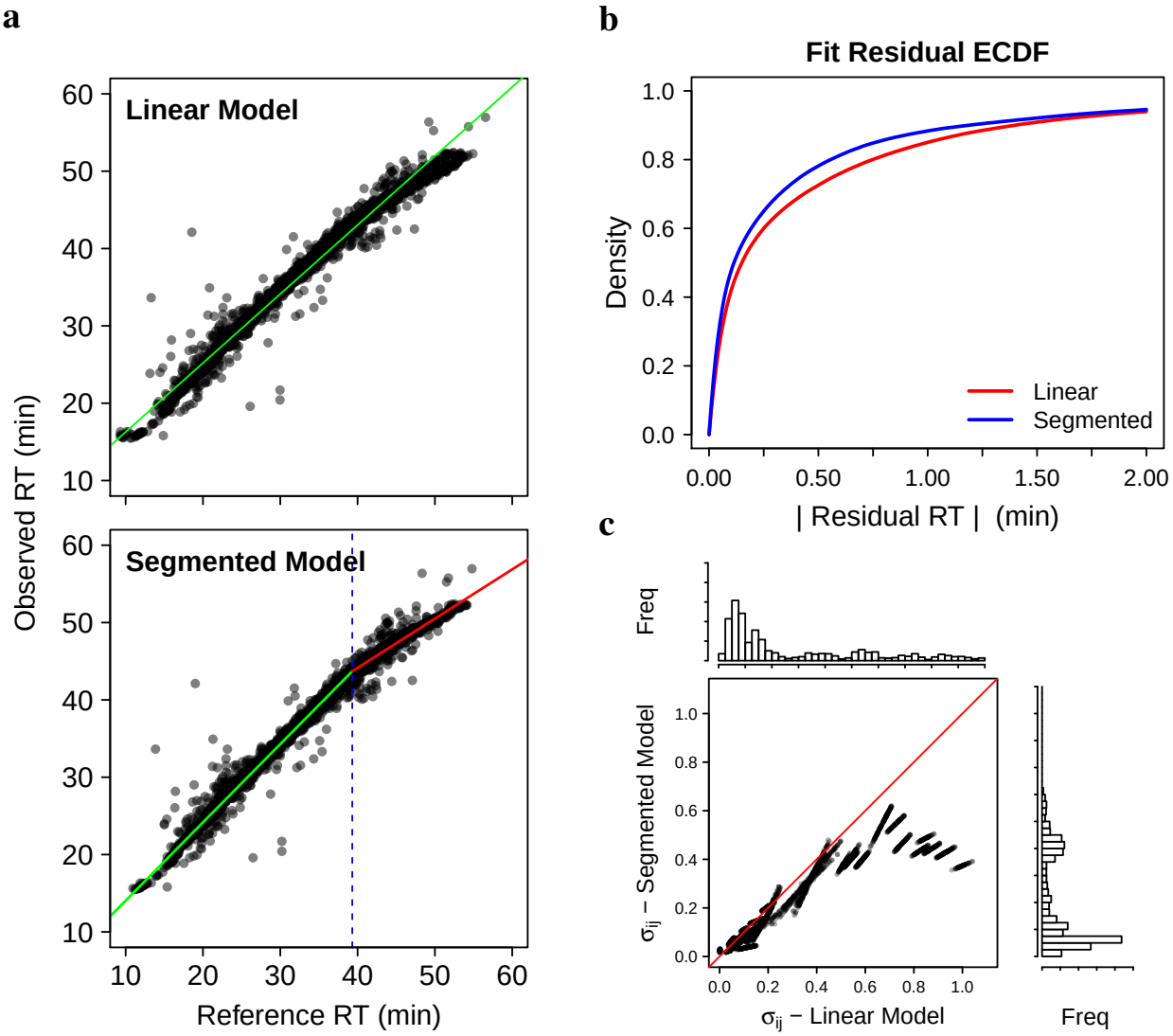


Figure S4 | Comparison of linear and segmented fits for reference retention times in experiments

(**a**) The reference RT of PSMs compared with their observed RTs, and the model plotted in green line (linear fit), or green and red lines (segmented fit, representing the two segments). For the segmented fit, the inflection point is marked with the dotted blue line. Both fits were specified separately and run separately with the same input data. (**b**) Empirical cumulative density function (ECDF) of the residual RTs for both fits. The residual RT is defined as the Observed RT − Inferred RT, where the inferred RT is the reference RT aligned to that particular experiment via. the model function – linear or segmented. (**c**) Model-fitted standard deviations, σ_{ik}, for each PSM as estimated by both linear and segmented fits. Points below the $45°$ line indicate a lower modeled RT standard deviation for the segmented fit, and vice versa. Clusters of points correspond to PSMs belonging to a particular experiment, as the PSM-specific variance of σ_{ik} is mostly reliant on the experiment in which the PSM is observed.

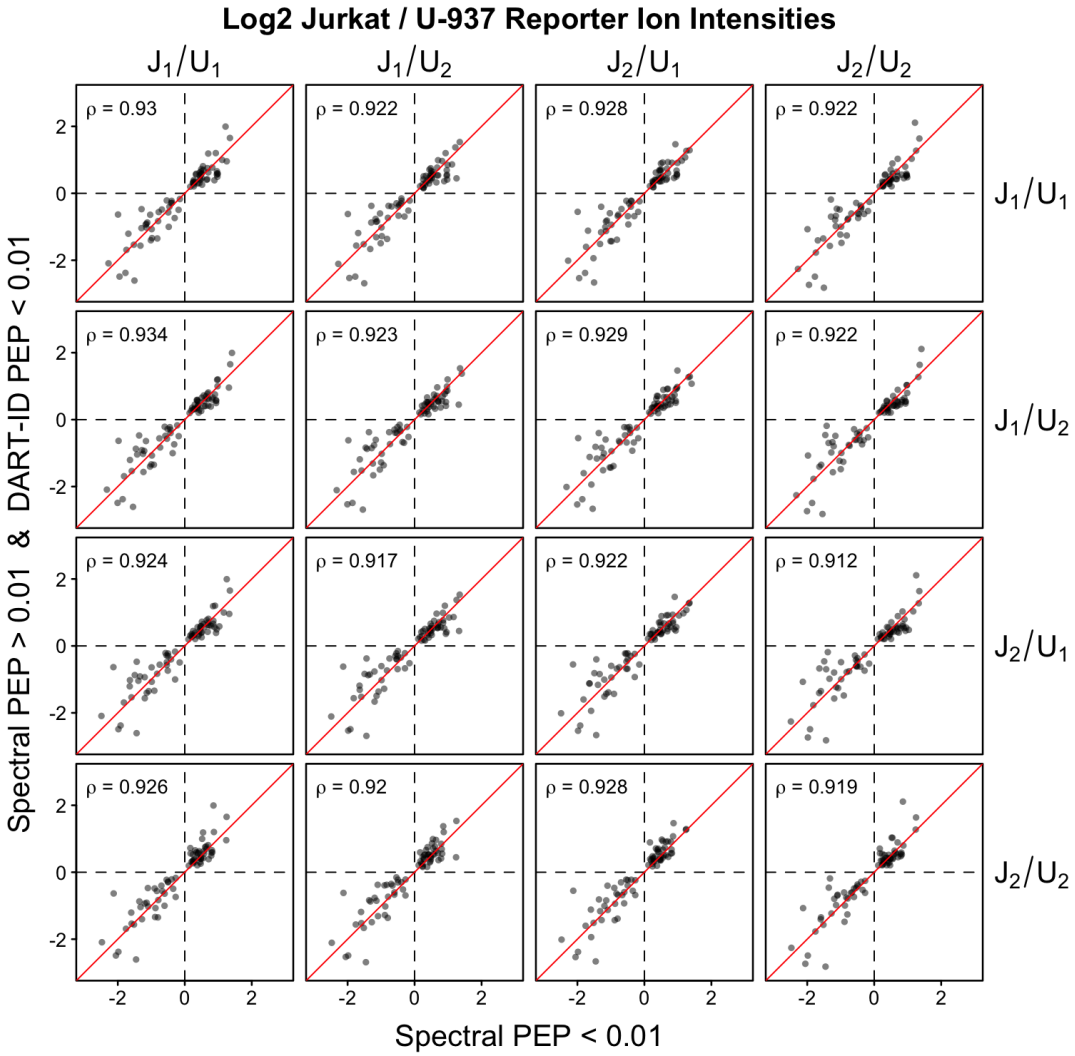


Figure S5 | Ratios of relative enrichment between two Jurkat and two U-937 cells from SCoPE-MS experiments

Spectra and *DART-ID* sets of PSMs were selected using the same method in Fig. 5. PSMs were further filtered to only represent proteins that showed sufficient fold-change, and this was measured by applying a two-sided t-test between the relative intensity of the protein in Jurkat cells versus U-937 cells, and then selecting at a significance of $p = 0.05$ after applying a Bonferroni (FWER) correction. Shown above are the comparisons of four ratios between the *Spectra* and *DART-ID* sets of distinct peptides.

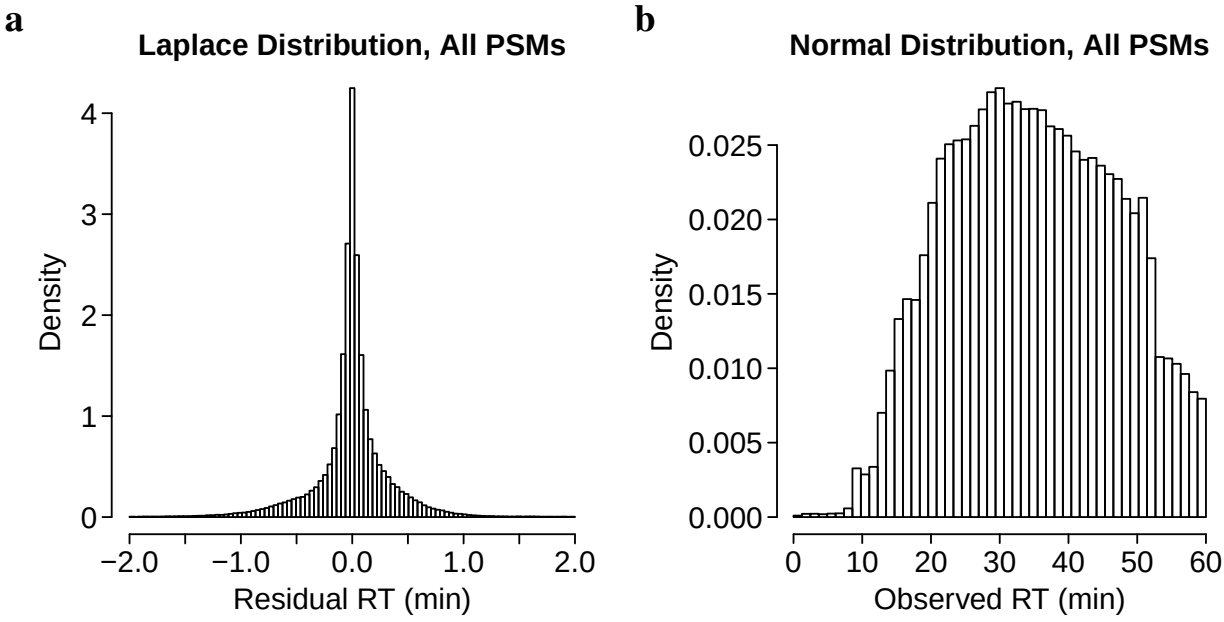


Figure S6 | Distribution choice for inferred RT distribution and null RT distribution
(**a**) Empirical distribution of all residual RTs, i.e., Observed RT − Predicted RT, and (**b**) all retention times, from a set of 226 SCoPE-MS experiments.

References

1. Budnik, B., Levy, E., Harmange, G. & Slavov, N. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *bioRxiv* **1,** DOI: 10.1101/102681 (2017).

2. Levy, E. & Slavov, N. Single cell protein analysis for systems biology. *Essays In Biochemistry* **62.** doi:10.1042/EBC20180014 (2018).

3. Specht, H. & Slavov, N. Transformative opportunities for single-cell proteomics. *Journal of Proteome Research* **17,** 2563–2916 (8 June 2018).

4. Meek, J. L. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. en. *Proceedings of the National Academy of Sciences* **77,** 1632–1636. ISSN: 0027-8424, 1091-6490 (Mar. 1980).

5. Guo, D., Mant, C. T., Taneja, A. K., Parker, J. & Rodges, R. S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides. *Journal of Chromatography A* **359,** 499–518. ISSN: 0021-9673 (1986).

6. Sakamoto, Y., Kawakami, N. & Sasagawa, T. Prediction of peptide retention times. en. *Journal of Chromatography A* **442,** 69–79. ISSN: 00219673 (Jan. 1988).

7. Krokhin, O. *et al.* An Improved Model for Prediction of Retention Times of Tryptic Peptides in Ion Pair Reversed-phase HPLC: Its Application to Protein Peptide Mapping by Off-Line HPLC-MALDI MS. en. *Molecular & Cellular Proteomics* **3,** 908–919. ISSN: 1535-9476, 1535-9484 (Sept. 2004).

8. Baczek, T., Wiczling, P., Marszall, M., Heyden, Y. V. & Kaliszan, R. Prediction of Peptide Retention at Different HPLC Conditions from Multiple Linear Regression Models. en. *Journal of Proteome Research* **4,** 555–563. ISSN: 1535-3893, 1535-3907 (Apr. 2005).

9. Krokhin, O. V. Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-Å Pore Size C18 Sorbents. *Analytical Chemistry* **78.** PMID: 17105172, 7785–7795 (2006).

10. Gorshkov, A. V. *et al.* Liquid Chromatography at Critical Conditions: Comprehensive Approach to Sequence-Dependent Retention Time Prediction. en. *Analytical Chemistry* **78,** 7770–7777. ISSN: 0003-2700, 1520-6882 (Nov. 2006).

11. Petritis, K. *et al.* Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses. en. *Analytical Chemistry* **75,** 1039–1048. ISSN: 0003-2700, 1520-6882 (Mar. 2003).

12. Petritis, K. *et al.* Improved Peptide Elution Time Prediction for Reversed-Phase Liquid Chromatography-MS by Incorporating Peptide Sequence Information. *Analytical Chemistry* **78,** 5026–5039 (2006).

13. Klammer, A. A., Yi, X., MacCoss, M. J. & Noble, W. S. Improving Tandem Mass Spectrum Identification Using Peptide Retention Time Prediction across Diverse Chromatography Conditions. *Analytical Chemistry* **79.** PMID: 17622186, 6111–6118 (2007).

14. Moruz, L., Tomazela, D. & Käll, L. Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. *Journal of Proteome Research* **9.** PMID: 20735070, 5209–5216 (2010).

15. Lu, W. *et al.* Locus-specific Retention Predictor (LsRP): A Peptide Retention Time Predictor Developed for Precision Proteomics. en. *Scientific Reports* **7,** 43959. ISSN: 2045-2322 (Mar. 2017).

16. Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: Boosting peptide identifications by chimeric spectra identification and retention time prediction. *Journal of Proteome Research* **0.** PMID: 29863353, null (2018).

17. Moruz, L. & Käll, L. Peptide retention time prediction. en. *Mass Spectrometry Reviews* **36,** 615–623. ISSN: 02777037 (Sept. 2017).

18. Palmblad, M., Ramström, M., Markides, K. E., Håkansson, P. & Bergquist, J. Prediction of Chromatographic Retention and Protein Identification in Liquid Chromatography/Mass Spectrometry. en. *Analytical Chemistry* **74,** 5826–5830. ISSN: 0003-2700, 1520-6882 (Nov. 2002).

19. Palmblad, M. *et al.* Protein identification by liquid chromatography-mass spectrometry using retention time prediction. en. *Journal of Chromatography B* **803,** 131–135. ISSN: 15700232 (Apr. 2004).

20. Silva, J. C. *et al.* Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs. *Analytical Chemistry* **77.** PMID: 15801753, 2187–2200 (2005).

21. Conrads, T. P., Anderson, G. A., Veenstra, T. D., Paša-Tolić, L. & Smith, R. D. Utility of Accurate Mass Tags for Proteome-Wide Protein Identification. en. *Analytical Chemistry* **72,** 3349–3354. ISSN: 0003-2700, 1520-6882 (July 2000).

22. Krokhin, O. V. *et al.* Use of Peptide Retention Time Prediction for Protein Identification by off-line Reversed-Phase HPLC-MALDI MS/MS. en. *Analytical Chemistry* **78,** 6265–6269. ISSN: 0003-2700, 1520-6882 (Sept. 2006).

23. McQueen, P. *et al.* Information-dependent LC-MS/MS acquisition with exclusion lists potentially generated on-the-fly: Case study using a whole cell digest of *Clostridium thermocellum*. en. *PROTEOMICS* **12,** 1160–1169. ISSN: 16159853 (Apr. 2012).

24. Strittmatter, E. F. *et al.* Application of Peptide LC Retention Time Information in a Discriminant Function for Peptide Identification by Tandem Mass Spectrometry. *Journal of Proteome Research* **3,** 760–769 (4 2004).

25. Pfeifer, N., Leinenbach, A., Huber, C. G. & Kohlbacher, O. Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. en. *BMC Bioinformatics* **8,** 468. ISSN: 1471-2105 (2007).

26. Pfeifer, N., Leinenbach, A., Huber, C. G. & Kohlbacher, O. Improving Peptide Identification in Proteome Analysis by a Two-Dimensional Retention Time Filtering Approach. en. *Journal of Proteome Research* **8,** 4109–4115. ISSN: 1535-3893, 1535-3907 (Aug. 2009).

27. Li, G.-Z. *et al.* Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. en. *Proteomics* **9,** 1696–1719. ISSN: 16159853, 16159861 (Mar. 2009).

28. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968 (2010).

29. Krokhin, O. V. & Spicer, V. Peptide Retention Standards and Hydrophobicity Indexes in Reversed-Phase High-Performance Liquid Chromatography of Peptides. en. *Analytical Chemistry* **81,** 9522–9530. ISSN: 0003-2700, 1520-6882 (Nov. 2009).

30. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12,** 1111–1121. ISSN: 1615-9861 (2012).

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

32. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11,** 2301–2319 (2016).

33. Van Nederkassel, A., Daszykowski, M., Eilers, P. & Heyden, Y. V. A comparison of three algorithms for chromatograms alignment. en. *Journal of Chromatography A* **1118,** 199–210. ISSN: 00219673 (June 2006).

34. Podwojski, K. *et al.* Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **25,** 758–764 (2009).

35. Lange, E., Tautenhahn, R., Neumann, S. & Gröpl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. en. *BMC Bioinformatics* **9,** 375. ISSN: 1471-2105 (2008).

36. Stanstrup, J., Neumann, S. & Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. en. *Analytical Chemistry* **87,** 9421–9428. ISSN: 0003-2700, 1520-6882 (Sept. 2015).

37. Fischer, B. *et al.* Semi-supervised LC/MS alignment for differential proteomics. en. *Bioinformatics* **22,** e132–e140. ISSN: 1367-4803, 1460-2059 (July 2006).

38. Bellew, M. *et al.* A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. en. *Bioinformatics* **22,** 1902–1909. ISSN: 1367-4803, 1460-2059 (Aug. 2006).

39. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. en. *Analytical Chemistry* **78,** 779–787. ISSN: 0003-2700, 1520-6882 (Feb. 2006).

40. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. en. *BMC Bioinformatics* **11,** 395. ISSN: 1471-2105 (2010).

41. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. en. *Nature Biotechnology* **32,** 219–223. ISSN: 1087-0156, 1546-1696 (Mar. 2014).

42. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4,** 923–925 (2007).

43. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry* **74,** 5383–5392 (2002).

44. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. en. *Journal of Proteome Research* **10,** 1794–1805. ISSN: 1535-3893, 1535-3907 (Apr. 2011).

45. Specht, H. *et al.* Automated sample preparation for high-throughput single-cell proteomics. *bioRxiv* (2018).

46. Malioutov, D. & Slavov, N. Convex Total Least Squares. *Journal of Machine Learning Research* **32,** 109–117 (1 2014).

47. Franks, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. *PLoS computational biology* **13,** e1005535 (2017).

48. Carpenter, B. *et al. Stan: A Probabilistic Programming Language*

49. Serang, O., MacCoss, M. J. & Noble, W. S. Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data. en. *Journal of Proteome Research* **9,** 5346–5357. ISSN: 1535-3893, 1535-3907 (Oct. 2010).

50. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76.** doi:10.18637/jss.v076.i01. http://dx.doi.org/10.18637/jss.v076.i01 (2017).