DIAlignR Provides Precise Retention Time Alignment Across Distant Runs in DIA and Targeted Proteomics

Authors
Shubham Gupta, Sara Ahadi, Wenyu Zhou, and Hannes Röst

Correspondence
hannes.rost@utoronto.ca

In Brief
Retention time alignment across large number of runs is important for reproducibility of SWATH-MS runs. We developed a hybrid RT alignment approach to establish correspondence among MS2-peaks. After testing on a gold-standard validation dataset, we were also able to align peaks from large-scale human plasma runs, which have elution-order-swapped peaks, with 98% accuracy compared to 76% by the existing method.

Highlights
- Retention time shift can lead to inversion of elution order of peptides.
- Global alignment methods are suboptimal for alignment of distant runs.
- DIAlignR employs hybrid (global + local) RT alignment approach.
- DIAlignR can align swapped peaks accurately across distant runs.
DIAlignR Provides Precise Retention Time Alignment Across Distant Runs in DIA and Targeted Proteomics*

Shubham Gupta‡§, Sara Ahadi¶, Wenyu Zhou¶, and Hannes Röst‡§

Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra (SWATH-MS) is widely used for proteomics analysis given its high throughput and reproducibility, but ensuring consistent quantification of analytes across large-scale studies of heterogeneous samples such as human plasma remains challenging. Heterogeneity in large-scale studies can be caused by large time intervals between data acquisition, acquisition by different operators or instruments, and intermittent repair or replacement of parts, such as the liquid chromatography column, all of which affect retention time (RT) reproducibility and, successively, performance of SWATH-MS data analysis. Here, we present a novel algorithm for RT alignment of SWATH-MS data based on direct alignment of raw MS2 chromatograms using a hybrid dynamic programming approach. The algorithm does not impose a chronological order of elution and allows for alignment of elution-order-swapped peaks. Furthermore, allowing RT mapping in a certain window around a coarse global fit makes it robust against noise. On a manually validated dataset, this strategy outperformed the current state-of-the-art approaches. In addition, on real-world clinical data, our approach outperformed global alignment methods by mapping 98% of peaks compared with 67% cumulatively. DIAlignR reduced alignment error up to 30-fold for extremely distant runs. The robustness of technical parameters used in this pairwise alignment strategy is also demonstrated. The source code is released under the BSD license at https://github.com/Roestlab/DIAlignR.

In translational research, protein biomarkers and therapeutic targets are usually discovered by data-driven methods such as by linking protein abundance patterns with disease conditions. A large sample cohort is essential in these studies as substantial biological variability exists in the population and enough statistical power is required to identify disease-specific events (1, 2). Blood plasma is a good source of clinical information of a patient as it can be obtained noninvasively and proteins from affected tissue can potentially leak into the blood. Plasma samples, unfortunately, are highly challenging for proteomic analysis due to the diversity of peptides within the samples and high dynamic range of plasma proteins (3). Therefore, quantification of plasma proteins requires a highly reproducible reduction of complexity and measurement within a wide dynamic range. The situation is exacerbated across large-scale studies, which makes development of plasma biomarker challenging (2, 3).

In the past two decades, mass spectrometry (MS)-based proteomics has made rapid advances with a high degree of innovation in obtaining identification and quantification of proteins in various biological samples (2, 4). Targeted proteomics methods, specifically selected reaction monitoring (SRM), can provide high reproducibility across multiple runs. However, it is limited by low throughput and can measure abundance of only a few tens to low hundreds of proteins per study (1, 5).

Recently, we developed SWATH-MS, an approach for targeted analysis of data-independent acquisition (DIA) data, which can reproducibly quantify larger sets of peptides in large-scale clinical studies (5, 6). Implementing this method in the clinical field could provide comprehensive characterization of samples across various conditions. It has allowed reproducible quantification of about 2,000 proteins in a biomarker study on tumorous kidney and healthy tissues (1, 7) and has the potential to record a molecular inventory of samples comprising a large number of proteotypes, thus making longitudinal monitoring of a patient possible (1).

In DIA mode, precursors in MS1 are selected for a pre-determined m/z range and fragmented nonspecifically. This produces multiplexed MS2 spectra of fragment ions of all selected precursors. The DIA data can be analyzed by using either a library-based approach (5, 8) or a library-free ap-
Library-based approaches have shown to be capable of accurate peptide and protein quantification in complex samples (5, 10, 11). Nonetheless, obtaining reproducible and robust analysis of clinical plasma samples is challenging even with SWATH-MS, as large variations in the number of proteins in individual runs are observed (5, 10, 11). One of the major factors driving variability is the retention time (RT) deviation between assay library and plasma peptides’ elution profiles. In experiments carried out by Nigjeh and coworkers, most of the peptides had RT variation of about 10 min between technical replicates, affecting the robustness of peptide quantification (3). This variation, if left uncorrected, may also result in incorrect and inconsistent identification of the peptides (3).

Current DIA data analysis software use iRT peptides to calculate a monotonic RT function (linear regression (12) or segmented regression (13)) with respect to a library. Using this mapping, extracted ion chromatograms (XICs) from MS2 spectra are obtained for peak picking. Software usually finds multiple potential peak groups in XICs, which makes downstream analysis challenging. By establishing peak correspondence among runs, correct peptide elution time can be determined for each MS run (5, 14). A shift in RT is often considered as a system-level variation, which is modeled using a monotonic function between two runs (14). However, this assumption may not always be accurate, and specifically among distant runs, singularities specific to a single peptide are common that produce relative peak switching where the elution order of two peptides is swapped across two runs (14–16). This phenomenon is increasingly likely in larger studies and very probable in large-scale clinical studies in which data acquisition happens over a span of years.

There are many methods in the literature for establishing correspondence in RTs. Current RT alignment algorithms in metabolomics and proteomics were mostly developed before the development of SWATH-MS (14) and, therefore, rely on either MS1 chromatograms (17–23), picked features (24–27), or a combination of both (28, 29). These algorithms usually find a global pairwise alignment function using dynamic programming on raw MS1 chromatograms (20–23, 28) or on feature lists (24, 25), which include methods using band constraints. For complex samples, so-called “landmark peaks” (28, 29) have been used to improve RT alignment accuracy. However, most of these approaches rely on MS1 data, and the resulting alignment functions are influenced by all constituent peptides. In SWATH-MS runs, MS2 data have high signal-to-noise ratio and are reproducible across multiple runs.

Previous research on RT alignment of SWATH runs has relied on MS2 feature finding software. These either align MS2 features using bipartite matching (16) or use the features to calculate a global function by local weighted regression (LOESS) (30) or by a kernel density approach (5, 31) (see supplemental Section S1). These approaches, however, provide suboptimal results in case of high noise, missing features, or when feature detection algorithms malfunction. Furthermore, the global monotone functions do not account for peptide switching as a monotone function disallows RT reversal between any two peptides (16).

Here we present DIAlignR, an RT alignment algorithm that addressed these shortcomings of previous methods. Our algorithm does not require features and is capable of directly aligning the raw multiplexed MS2 chromatographic traces from targeted proteomics data. Our approach uses dynamic programming to obtain an optimal mapping between chromatograms that contain local information as multiple, close-by peaks around the eluted peak group. Independent RT alignment of each precursor facilitates the alignment of elution-order-swapped peaks. Our method is also capable of using a global whole-run alignment for guidance, making it robust against noise. Thus, DIAlignR can flexibly handle user preference of selecting between extremes of global and local alignment.

We provide free access to our source code and our R-package at https://github.com/Roestlab/DIAlignR. We tested our tool on a manually validated dataset of over 7,000 chromatograms and demonstrated improved performance over existing methods. We also tested it on 24 randomly selected blood plasma runs, selected from a heterogeneous cohort measured across many months. For both datasets, our algorithm outperformed global alignment methods and was capable of correcting misannotations introduced by feature detection algorithms. For very distant runs, it could also precisely align switched peaks, which is not possible using global alignment methods (8, 12).

MATERIALS AND METHODS

Alignment algorithms are useful for mapping signals, which are close to noise, over several runs. Few datasets exist that can be used for benchmarking as often the ground truth is unknown.

Validation Dataset—For benchmarking, we used a previously published dataset (8) of 16 SWATH runs from Streptococcus pyogenes bacterial strains. In these runs, 452 randomly selected precursors were manually annotated (5). Out of these, eight precursors had annotation for less than two runs and were thus removed. Seven other precursors from the remaining set had annotated peaks outside of the XICs, making them inapplicable and, hence, were removed from benchmarking (see supplemental Section S2). Therefore, 437 annotated precursors were considered for performance testing of the developed DIAlignR tool against global alignment approaches. Annotated RTs of the precursors are available in the supplemental Table 1. Since annotated peaks were selected randomly from the validation dataset, this dataset had 4.9% peaks with signal-to-noise ratio less than 1 (supplemental Fig. 1A).

Large-Scale Human Plasma Dataset—We performed SWATH-MS on 975 human plasma samples in 12 batches from February 17, 2017...
The RT of the peptides in all 24 runs is provided in the supplemental OpenSWATH (8, 32) and "mzR" package (33) with default parameters. Of the selected 406 precursors were extracted and parsed using peak-group rank score were considered for performance evaluation. Therefore, peaks from each batch were selected; their metadata and OpenSWATH output files are described in the supplemental Table 2. Since peaks were not visually validated, only peaks with a low false discovery rate were selected per run for alignment.

To reduce the number of pairwise alignments, randomly, two runs from each batch were selected; their metadata and OpenSWATH output files are described in the supplemental Table 2. Since peaks were not visually validated, only peaks with a low false discovery rate score were considered for performance evaluation. Therefore, peaks of target precursors with a q-value less than $10^{-3}$ (m-score < 1e-03, peak-group rank = 1) were selected, and precursors were required to be present in all 24 runs. Successively, fragment-ion chromatograms of the selected 406 precursors were extracted and parsed using OpenSWATH (8, 32) and "mzR" package (33) with default parameters. The RT of the peptides in all 24 runs is provided in the supplemental Table 3. The chromatograms are available on PeptideAtlas (ftp.peptidatlas.org PASS01280:KQ2592b).

A tabular description of both datasets is provided as Table I.

Table I
Summary description of validation and human plasma datasets

| Validation dataset | Human plasma dataset |
|--------------------|----------------------|
| Biological sample  | S. pyogenes strain SF370 | Plasma from blood samples |
| Mass-spectrometer  | SciEX 5600 TripleTOF | SciEX 6600 TripleTOF |
| LC-gradient        | Linear               | Linear |
| Total run time     | 135 minutes          | 55 minutes |
| LC-column replaced | No                   | Yes |
| Mass spectrometer repaired | No             | Yes (replaced quadrupole after 7 batches) |
| Data acquisition date | August 8, 2012-August 9, 2012 | February 17, 2017-July 20, 2017 |
| Number of runs acquired | 16              | 975 |
| Number of batches  | 1                    | 12 |
| Runs selected for alignment | 16             | 24 |
| Total number of run pairs | 120            | 276 |
| Software used for feature detection and XIC extraction | OpenSWATH | OpenSWATH |
| Number of common precursors selected per run for alignment | 437 | 406 |
| Total number of alignments | 49,505          | 112,056 |
| Manual Annotation  | Yes (Skyline)        | No |

1A, the similarity between all time points can be represented as a similarity matrix s. Thus,

$$s = f(ChromA, ChromB)$$

The function $f$ is termed as a similarity measure and can be selected by the user (see below).

**Similarity Measure**—In our R package DIAAlignR, we implemented several similarity measures that have been suggested in previous literature for chromatograms such as covariance, dot product, Pearson's correlation, spectral angle, and Euclidean distance (8, 28). We observed that the dot product between all $I$ and $J$ data points provided information about both magnitude and angle between two time points, hence segregating elution signal from the background. If each data point of the chromatogram is represented by a vector in $n$ dimensional space ($n = 3$ in Fig. 1A), the resulting dot product of the two vectors will be as shown in Fig. 1B. Thus, with dot-product similarity, the matrix $s$ from all vectors of both chromatogram groups is defined as

$$s_{ij} = \sum_{k=1}^{n} a_k b_k$$

where $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J\}$ represent the index of vectors in ChromA and ChromB, respectively. A color-coded similarity matrix of size $I \times J$ is shown in Fig. 1C. However, to reduce the impact of noise peaks, a modified dot product, termed as "masked dot-product," was used where higher similarity scores were checked again for spectral angle similarity (see supplemental Section S5). A path in the resulting similarity matrix was calculated using dynamic programming, which directly translates to an RT alignment that mapped indices/time from ChromA to ChromB and vice versa.

**Penalizing Similarity Matrix with Global Alignment**—While dynamic programming will find a path of the highest cumulative score, in some instances, the score is driven by alignment to noise and can lead to a solution where the alignment is highly divergent from a global linear or nonlinear alignment. To make the alignment robust against noise and in order to incorporate information from the global context, we added an option in our algorithm to modify the similarity matrix $s$ (Fig. 1D) using feature-based global alignment such as LOESS. Residual standard error of the fit was utilized to define a region of noninterfer-
ence in the similarity matrix and values outside of it were punished with a negative score (see supplemental Section S5). This allowed us to find an alignment path within a reasonable time window relative to global prediction and avoid large deviations.

**Overlap Alignment with Affine Gap Penalty**—The optimal alignment path was found by recursively calculating all possible optimal paths from the start of the similarity matrix \((1, 1)\) to the end of it \((I, J)\) using dynamic programming (34). Chromatogram groups ChromA and ChromB may not have end-to-end mapping as these could be partial chromatograms that were extracted around the expected peptide elution (as determined by indexed retention time \([iRT]\) peptides, for example). Therefore, overlap alignment instead of a global alignment of MS2 chromatogram groups was employed. This approach allowed free end gaps and thus, allowed sliding chromatograms freely without incurring any gap penalty for it.

To widen or shrink chromatogram peaks, a gap of unit length is a reasonable choice as it will distribute gaps along the complete peak. Therefore, an affine gap penalty scheme was utilized with higher gap penalty for a gap length of more than one. In this approach, three matrices \((\text{Matrix } M, \text{ Matrix } A, \text{ Matrix } B)\) were defined, which recursively calculated a score for gaps of more than unit length (34). The overlap alignment path using affine gap penalty is presented in Fig. 1E. The running time of such alignment is \(O(max(I, J)^3)\). A heuristic data-driven approach is employed to obtain suitable affine gap penalties from the similarity matrix (see supplemental Section S5). Mapping the alignment path to the initial time values provides aligned chromatograms as depicted in Fig. 1F.

**Running Time for Alignment**—Alignment of MS2 chromatograms of each peptide/precursor has a running time of order \(O(max(I, J)^3)\); however, chromatograms of different precursors can be aligned independently. Therefore, we employed parallelization for different peptides to obtain a much faster speed for complete run-time mapping.

**Optimization of Algorithm Parameters**—There are various parameters used in DIAAlignR. A description of these parameters is available in supplemental Section S5. We employed a validation dataset for parameter optimization and used the number of peaks aligned within half chromatographic peak-width and cumulative RT alignment error as an optimization target.

**Performance Metrics for Comparison with Current Algorithms**—We used the manually validated dataset (5) to compare DIAAlignR to the current state-of-the-art method (e.g. TRansfer of Identification Confidence [TRIC] (5)), which utilizes a set of high confidence peaks ("anchor peptides") to compute a linear or nonlinear alignment function that transforms RT values from run1 to run2. We chose LOESS (local regression) as well as linear regression for evaluation. For LOESS, both optimized spanvalue from cross-validation (as used in
TRIC) and default spanvalue (= 0.75) of the R software environment were tested (30). For LOESS fit, cross-validation was performed to obtain the optimum span value between two runs (5, 30). Steps to obtain a global fit (monotone mapping function) are detailed in the supplemental Section S3 and S4.

RT error was calculated by comparing against the manual annotation of the S. pyogenes dataset (5), and the resulting distribution of the number of peptides aligned within a certain RT tolerance was used as a measure of overall accuracy of the alignment algorithm. Manual annotations are not available for the human plasma dataset; therefore, the high-quality results (peaks with low false discovery rate cutoff) of OpenSWATH were used for benchmarking.

RESULTS

Parameter Optimization—Here, we present an algorithm for multitrace chromatographic alignment that only uses raw MS2 data from targeted proteomics or DIA experiments for RT alignment. To optimize the performance of our algorithm, we investigated the effect of algorithmic parameters on the accuracy of the alignment of runs from the validation dataset (5). First, we evaluated the performance for different similarity measures of chromatogram groups. The dot product masked with spectral angle (see supplemental Section S5) as a similarity measure provided the highest fraction of peptides aligned for all 120 possible run pairs (Fig. 2A). Within an RT error tolerance of half-peak width (here: 15.3 s), this similarity measure aligned 94.33% of annotated peaks with the highest area under the curve (AUC) of all approaches (see supplemental Tables 7 and 8).

We then investigated the effect of the gap penalty used in dynamic programming. In DIAlignR, the gap penalty is calculated heuristically as a fixed quantile value of the distribution of similarity scores. We found that the selection of quantile value did not have a considerable impact on the percentage of peaks aligned within a certain RT tolerance (Fig. 2B). From the figure, 20th to 90th quantile values yielded approximate 95.6% of aligned peaks within half-peak width. The effect of gapQuantile was less pronounced for wider RT tolerance. For further analysis, the 65th quantile was selected as the base gap penalty for chromatogram alignment. For the affine gap penalty, a gap opening factor of 0.125 was used, while a gap extension factor of 40 was considered (see supplemental Section S5).

We next investigated the impact of the number of transitions on the alignment (see supplemental Section S5 and supplemental Fig. 10A). We found that, as the number of fragment ions increased, alignment accuracy improved, with the best alignment 94.3% from using all library fragment ions compared with 88.7% for only one transition (supplemental Table 15). However, the largest improvement was going from one transition (88.7%) to two transitions (92.8%) with smaller improvements for using all transitions. This indicates that a single transition was not sufficient for this algorithm, but the relative gain was more modest after two transitions.

Our algorithm constrains the similarity matrix using a global alignment function. Constraining the alignment in a certain window (given by RSEdistFactor) around the global fit derived from “anchor peptides” improves the alignment accuracy. We observed that, with a constrained similarity matrix 95.4% peaks were aligned compared with 94.3% with the nonconstrained one (see supplemental Fig. 10B). An example of such alignment is shown in Figs. 2C and 2D in which the similarity matrix had two high similarity hot spots. By constraining the alignment inside the dashed region in Fig. 2D, the alignment path went through the correct hot spot. With the unconstrained similarity matrix, an incorrect alignment resulted, as shown in the supplemental Fig. 11.

Validation Using “Gold Standard” Reference Dataset—Using a validation dataset, we compared DIAlignR to current alignment methods. In terms of number of peptides aligned and alignment precision, chromatogram alignment by DIAlignR outperformed LOESS and linear regression methods (see Fig. 3A and Table II). On the S. pyogenes benchmark dataset, DIAlignR decreased error rates by 1.8-fold compared with the state-of-the-art methods. Cumulatively, chromatographic alignment only misaligned 4.3% of all peaks within 15.3 s (half-peak width) of the true RT compared with 7.9% for LOESS (while LOESS with default parameters misaligned 22.8% of all peaks, and linear regression misaligned 44.8%; see Fig. 3A).

We next investigated the effect of experimental perturbation on the performance of the alignment method. We compared within-condition alignments with between-condition alignments (in the validation dataset, the conditions were 0% and 10% human plasma added to S. pyogenes during growth). For both alignment methods, we observed decreased performance for between-condition alignments compared with within-condition alignments (Fig. 3B). However, the performance drop of the LOESS method (4.93%) was substantially larger than the corresponding performance drop of DIAlignR (2.7%), indicating increased robustness to sample heterogeneity for DIAlignR.

To evaluate the consistency of alignment approaches across multiple run pairs, we computed the number of correctly aligned peaks (defined as instances where the alignment was correct within half-peak width) for each run pair. This distribution was shifted toward the right with the linear regression method compared with LOESS, indicating that the former was consistent in its performance (Fig. 3C). In terms of the precision of the alignment, chromatogram alignment consistently performed better than global alignment methods as the former had a higher area under the cumulative peptide frequency curve for each run pair (higher AUC for 120 out of 120 pairs; see supplemental Fig. 12A). Similarly, we observed a larger RT variation (standard deviation = 18.45 s) with the LOESS approach, which the chromatogram alignment corrected satisfactorily with the standard deviation being 11.68 s (Fig. 3D and supplemental Figs. 12A and 12B). We concluded that, on the validation dataset, DIAlignR performed consistently better in terms of accuracy of alignment and number of correctly
aligned peaks across a range of different RT cutoffs and LC-MS/MS runs.

Next, we were interested in how the global differences between the two methods translate to individual alignments. We, therefore, computed the alignment error for each pairwise alignment of each peptide (total 49,505 alignments) and found that chromatographic alignment outperformed LOESS in 4.7% of all cases, whereas, LOESS achieved better results in 1.1% of cases, while comparable performance was achieved in the remaining 94.2% cases (see supplemental Fig. 13B). On average, DIAlignR reduced the RT error by 2.3 s with a median of 1.7 s (see supplemental Fig. 13C). Overall, our method aligned 47.3k peaks correctly compared with 45.6k by an optimized LOESS within half-peak width (15.3 s). However, in
In general, we observed that, on the validation dataset, both methods performed with similar consistency, which may be due to the low complexity of a bacterial sample and the high homogeneity of the data as they were acquired within two consecutive days on the same LC column.

Application to Large-Scale Heterogeneous Human Plasma Measurements—After demonstrating consistently improved performance on the *S. pyogenes* validation dataset, we investigated the performance of our algorithm on the data from large-scale SWATH-MS experiments on human plasma.
These experiments provided a more challenging dataset as the data were acquired over a period of six months with an intermittent repair of the instrument and replacement of the old column (column1) with a new column (column2). Two LC-MS/MS runs from each of the 12 batches were selected at random, and 406 peptides were used for testing our algorithm. Since we did not have manually validate peaks, high-confidence peak groups (q-value < $10^{-5}$) were used as the validation set.

Comparing our chromatogram alignment algorithm (DIAAlignR) with the LOESS method on a highly heterogeneous human plasma dataset, we found that our approach aligned 97.92% of peaks compared with 76.03% by LOESS with a maximal error of 20 s (half chromatographic peak width) as depicted in Fig. 4A. All tested 276 pairwise alignments showed improved performance using chromatographic alignment (see supplemental Fig. 15). Next, we were interested in the performance of our method on the alignment of runs acquired on the two different columns. We found that, for runs acquired on different columns, chromatogram alignment method aligned 97.7% (compared with 97.84% within-column alignment) of peaks compared with 63.38% (89.06% for within-column alignment) by the LOESS method (Fig. 4B), suggesting that DIAAlignR retained performance even for highly heterogeneous datasets while the LOESS approach did not. Specifically, we found not only that DIAAlignR outperformed LOESS on between-column alignments but also that the performance loss for DIAAlignR was much less pronounced than for LOESS compared with within-column alignments. After validating the performance of chromatogram alignment cumulatively, we decided to investigate its consistency across individual run-pair alignments. Fig. 4C presents the distribution of the number of peaks aligned in all 276 pairs. DIAAlignR was capable of aligning 400 peaks on average within half-peak width (while LOESS aligned 309 peaks on average), a 29% improvement. This indicated substantial alignment error when using LOESS, which could be reduced drastically by DIAAlignR.

To validate the performance on individual alignment, we further computed the alignment error for each pairwise alignment for every peptide. The standard deviation of alignment error for LOESS was 22.91 s compared with 13.7 s for DIAAlignR (Fig. 4D). This indicated the higher precision of RT alignment with our approach. Out of 112,056 alignments, we found that DIAAlignR outperformed LOESS in 23% of all cases and performed similarly in 76% cases (see supplemental Fig. 14—while performing worse in 1.13% cases). Upon manual validation, several of these worse-performing peaks were found to be due to wrong annotation by OpenSWATH (supplemental Fig. 18). Thus, testing of the chromatogram alignment approach on the heterogeneous human plasma dataset again validated its consistent and improved RT alignment performance.

**Switching of Peptide Elution Order**—In liquid chromatography, RT drift is often observed from one run to another run. However, we were interested whether this drift may be variable for different peptides and thus could result in reversal of retention order (15). In such a scenario, two peptides that are eluting in order in one run may reverse their elution order in another run. Since our approach did not make an assumption of order preservation of peptide elution and facilitated independent alignment, we hypothesized that DIAAlignR would be capable of uncovering instances of non-order-preserving chromatographic alignment. Specifically, we analyzed peptide pairs that switch elution order from the heterogeneous and distant plasma runs.

| Dataset            | Method    | Average number of peaks aligned within half-peak width | Average number of peaks aligned within one peak width | Average number of peaks aligned within two peak width | Total number of peaks per run | Peaks aligned within half-peak width (%) |
|--------------------|-----------|--------------------------------------------------------|------------------------------------------------------|------------------------------------------------------|-------------------------------|-----------------------------------------|
| Validation dataset | DIAAlignR | 394.81667                                              | 407.3                                                | 410.69167                                             | 437                           | 95.68765                                |
|                    | LOESS     | 380.07500                                              | 406.18333                                            | 410.81667                                             | 437                           | 92.10249                                |
| Heterogenous human plasma dataset | DIAAlignR | 397.56159                                               | 400.50362                                            | 403.88043                                             | 406                           | 97.92157                                |
|                    | LOESS     | 308.67754                                               | 382.17754                                            | 403.15217                                             | 406                           | 76.02895                                |

TABLE II

The average number of peptides aligned from manually validated *S. pyogenes* dataset and from heterogeneous human plasma measurements. For the plasma data, peaks with OpenSWATH m-score < 0.001 are used for evaluation of the algorithm.
AQLVDMK/2 drifts only $1,050/1,100 = 150$ s. This varying RT drift between two runs caused the peptides to elute in a different order. The peptide-pair cannot be aligned with a global alignment approach, which in the best-case scenario would be off by 120 s; however, our chromatogram alignment method mapped the peaks correctly from run4 to run23 (see supplemental Fig. 17).

We, further, calculated the cumulative fraction of peptides aligned for pair “run4_run23” (Fig. 5C). Chromatogram alignment correctly aligned 98% peaks compared with LOESS,

Fig. 4. Alignment accuracy of MS2 chromatogram alignment on 24 runs of clinical plasma measurement dataset annotated with OpenSWATH. 406 peak groups were selected in each run with $m$-score < 0.001. (A) Cumulative fraction of peptides having error less than a specific RT difference is plotted for all possible $C(24,2) = 276$ pairs for chromatogram alignment, linear fit, and $k$-nearest neighbor smoothing (LOESS) with and without optimum span. (B) Cumulative fraction of peptides with alignment accuracy is plotted for chromatogram alignment and LOESS for pairs with different data acquisition conditions. LC column was changed together with an quadrupole replacement. Fourteen runs were acquired on column1, which makes 91 pairs, labeled as “column1”. Ten runs were acquired after quadrupole replacement on column2, which results into 45 pairs, labeled as “column2”. There were 140 pairs composed of “column1” and “column2” labeled runs; these pairs are labeled as “column1-column2”. (C) Histogram of number of peptides matched within half peak-width for LOESS and chromatogram alignment. (D) Histogram of RT prediction error is plotted for chromatogram alignment and LOESS. RT difference standard deviation for both approaches is 22.91 s and 13.7 s, respectively.
which was able to align only 37.93%, thus, DIALignR was capable of decreasing the error by up to 30-fold. Eight peaks, which were not correctly aligned by chromatogram alignment, were further inspected visually by the authors and were found to be the cases of misannotation by OpenSWATH, mainly due to the posttranslational modifications (see supplemental Section S7 and supplemental Fig. 18).

**DISCUSSION**

Correcting for RT drift and aligning RTs between LC-MS/MS runs has been a long-standing problem in proteomics, and it has become of particular importance as proteomics moves toward large-scale analysis of human cohorts. However, most efforts so far focus on MS1 data, and few algorithms are available that can exploit the full information present in MS2 spectra produced by targeted methods or DIA/SWATH-MS.

In this paper, we presented a novel algorithm that used the raw fragment-ion chromatograms directly to perform RT alignment for targeted proteomics and DIA data. Our algorithm uses XICs to map peaks across a pair of runs and improves alignment accuracy compared to current state-of-the-art methods. We, furthermore, extended the algorithm and implemented a hybrid approach, which used a feature-
based global alignment to condition the similarity matrix so that led to further gains in accuracy (see supplemental Fig. 10B). This hybrid approach provides the best of both worlds with a flexible “knob” that allows the user to either put more focus on global features or rely more on local information. To our knowledge, researchers have not yet explored dynamic-programming-based alignment on raw fragment-ion chromatograms. The dynamic programming approach is essential for obtaining a nonlinear (or gapped) alignment as distant runs also have varying drift even for local peaks. Using LOESS to partially constrain the alignment made our algorithm more stable and provided the robustness of global alignment methods.

We showed that on a “gold-standard” validation dataset, DIAlignR consistently outperformed a global alignment method (using either linear or nonlinear approaches), the current state-of-the-art (supplemental Fig. 12C). We observed that alignment accuracy increased almost 4% if a precursor had two transitions instead of one (supplemental Fig. 10A), and an additional 1.5% of aligned precursors by using all six fragment ions, which also corresponds to recommended guidelines (4). Since the DIAlignR approach puts more emphasis on local data, we also observed instances of overfitting where the global alignment function produced better alignment (supplemental Fig. 13B). Overall, however, we saw increased performance, and the DIAlignR algorithm decreased error rates from 7.9% to 4.3%. Interestingly, we found that our method was less sensitive to changes in chromatographic condition or sample matrix than global alignment approaches (Fig. 3B).

This finding led us to speculate that the novel chromatographic alignment would be less sensitive to heterogeneity in sample composition and chromatographic condition in large-scale studies. We tested our algorithm on a large-scale SWATH-MS experiment of human plasma acquired over several months. On this dataset, DIAlignR reduced RT alignment error from 24% to 2%, which is a significant improvement over current state-of-the-art methods. Our approach outperformed other methods and consistently mapped the highest number of peaks within half-peak width irrespective of acquisition time interval, column change, or instrument repair between two runs (Fig. 4B). DIAlignR improved RT alignment accuracy, which has the potential to improve peak-group identification and quantification through downstream tools. We manually identified an example of a wrongly aligned peak group as shown in supplemental Fig. 19. We also observed that, in the case of a peak being outside of an extracted chromatogram, our method was able to map RT outside of it as our hybrid approach also uses global alignment (see supplemental Fig. 18). Chromatograms can then be re-extracted and be used to correctly annotate peaks. Thus, this method can further be employed to extract chromatograms by OpenSWATH and other tools.

The improvement in RT alignment came with nonnegligible computation cost. For an alignment of 10-min-wide XICs with a 3.4-s cycle time, DIAlignR took, on average, 0.16 s. Thus, pairwise alignment of selected 437 peptides took approximately 1 min per run pair. The highest cost in the alignment of each peptide was the calculation of the alignment path using dynamic programming. However, this problem scales linearly with the number of peptides in the library and could be easily parallelized on a computing cluster.

We believe that our approach is most useful for large-scale heterogeneous targeted proteomics studies where runs are acquired by different personnel and data are collected over several months or even years. Applying a single mapping function in such experiments becomes a very challenging task, considering the switching of elution order of peptides. Global alignment functions, being monotonic, assume chronological order of peptide elution and, therefore, cannot align switched peptides. Hence, we observed substantial underfitting with the global function and an overall reduction of error using DIAlignR. Our hybrid approach aligned these peptides accurately as it mostly relied on additional dimensions of fragment-ion m/z to align peaks. It is possible that switching peptides may share fragment ions; however, this is very rare scenario if the library is designed carefully (4), and in such cases, our method will perform no worse than global alignment methods.

Accurate RT alignment has multiple uses in the application of mass-spectrometry-based proteomics for large-scale systems biology studies; correct identification and improved quantitation of a large number of analytes are two of them. This seems intuitive as most quantitative approaches currently available, at least to some degree, rely on accurate RT alignment. We present a tool that can align RTs of DIA data by establishing correspondence between analytes across large number of samples, making DIA amendable for multicenter and longitudinal studies. We also expect that this tool can be utilized by existing proteomics software to streamline analyte identification and to improve the quantification.

Acknowledgments—We are grateful to Michael Snyder for supervising data acquisition and providing access to the heterogeneous plasma dataset. We also thank Michael Brudno for valuable discussion on chromatogram alignment using dynamic programming.

DATA AVAILABILITY

Raw chromatograms and features extracted by OpenSWATH are available on PeptideAtlas under accession code PASS01280.

* The authors declare no competing financial interests.

To whom correspondence should be addressed: 160 College Street, Toronto, ON, M5S 3E1, Canada. Email: hannes.rost@utoronto.ca.

Author contributions: S.G. performed research; S.G. analyzed data; S.G. and H.R. wrote the paper; H.R. designed research; and S.A. and W.Z. acquired the mass-spectrometry dataset.
