Title of Paper

DIAlign provides precise retention time alignment across distant runs in DIA and targeted proteomics

Authors

Shubham Gupta[1,2], Sara Ahadi[3], Wenyu Zhou[3], Hannes Röst[1,2]

Affiliations:

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON M5G 1A8, Canada
[2]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada
[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA.

Abstract

SWATH-MS has been widely used for proteomics analysis given its high-throughput and quantitative 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 due to large time intervals between acquisition, acquisition by different operators or instruments or intermittent repair or replacement of parts, such as the liquid chromatography column, all of which affect retention time (RT) reproducibility and successively quantitative performance of SWATH-MS. Here, we present a novel algorithm 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 aligning of elution-order swapped peaks. Furthermore, allowing RT mapping in a certain window around coarse global fit makes it robust against noise. On a manually validated dataset, this strategy outperforms the current state-of-the-art approaches. In addition, on a real clinical data, our approach outperforms global alignment methods by mapping 98% of peaks compared to 67% cumulatively, is capable of reducing alignment error up to 30-fold for extremely distant runs. The robustness of technical parameters used in this strategy has also been demonstrated. The source code is released under the BSD license at https://github.com/Roestlab/DIAlign.

Introduction:

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 huge biological variability exists in the population and enough statistical power is required to identify disease specific events (Uzozie and Aebersold 2018; Surinova et al. 2011). Plasma is a good source of clinical information of a patient as it is noninvasive 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(Nigjeh et al. 2017). Therefore, quantification of plasma proteins requires a highly reproducible reduction of complexity and measurement within a wide dynamic range. The situation exacerbates across large-scale studies and make development of plasma biomarker challenging(Nigjeh et al. 2017; Surinova et al. 2011).

In the past two decades, mass spectrometry (MS) based proteomics has made rapid advances to obtain near-exhaustive identification and quantification of proteins in various biological samples(Surinova et al. 2011; Schubert et al. 2017). Targeted proteomics methods, specifically selected reaction monitoring (SRM), can provide reproducible protein quantification across multiple runs. However, they are limited by low throughput and can measure abundance of only a few tens of proteins per study(Röst et al. 2016; Uzozie and Aebersold 2018).

Recently, we developed SWATH-MS, an approach for targeted analysis of data-independent acquisition (DIA) data, which has potential to reproducibly quantify large sets of peptides in large-scale clinical study(Röst et al. 2016; Gillet et al. 2012). Implementing this method in the clinical field could provide comprehensive characterization of sample across various clinical conditions. It also facilitates creating a digital inventory of a tissue proteome(Uzozie and Aebersold 2018; Guo et al. 2015) and opens up the potential for clinics to record a molecular inventory of samples through which longitudinal monitoring of patient is possible(Uzozie and Aebersold 2018).

In DIA mode, after MS1 precursors are selected for a predetermined m/z range and fragmented non-specifically. This produces a multiplexed MS2 spectra of fragment-ions of all selected precursors. The DIA data can then be analyzed by either using a library-based approach (Röst et al. 2014; Röst et al. 2016) or a library-free approach (Tsou et al. 2015). Library-based approaches have shown to be capable of accurate peptide and protein quantification in complex samples (Navarro et al. 2016; Röst et al. 2016; Röst et al. 2014; Liu et al. 2015). Nonetheless, obtaining reproducible protein quantification from clinical plasma sample has been challenging even with SWATH-MS, as large variations in number of proteins in individual runs were observed (Nigjeh et al. 2017, (Röst et al. 2016; Liu et al. 2015; Navarro et al. 2016). One of the major factors driving variability is retention time deviation between the assay library and plasma peptides' DIA elution profiles. In experiments carried by Nigjeh and coworkers, most of the peptides had RT variation of about 10 minutes between technical replicates, affecting the robustness of peptide quantification (Nigjeh et al. 2017) . RT variations, if left uncorrected, may also produce incorrect and inconsistent identification of the peptides(Nigjeh et al. 2017).

Current DIA data analysis softwares are capable of finding multiple peak-groups in MS2 extracted ion chromatograms (XICs), however it is often challenging to efficiently integrate this information across multiple runs. By establishing peak correspondence among runs, correct peptide elution time could be determined in each MS run (Smith et al. 2015; Röst et al. 2016). A shift in retention time (RT) is often considered as system-level variation which is then modelled using monotonic functions between two runs (Smith et al. 2015). However, this assumption may not always be accurate, and specifically among distant runs singularities specific to a single peptide are also common that produces relative peptide switching (peak switching) where the elution order of two peptides is swapped across two runs (Smith et al. 2015; Spicer et al. 2010; Wu et al. 2016)]. 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.

Current RT alignment algorithms were mostly developed before the arrival of SWATH-MS technique (Smith et al. 2015) and therefore mostly rely on either MS1 chromatograms or picked features or combination of both (Listgarten et al. 2005; Prince and Marcotte 2006; Sandin et al. 2013). In SWATH runs, MS2 data has high signal-to-noise ratio and reproducible across multiple runs. Previous research on RT alignment in DIA data has also relied on feature finding, using either a bipartite matching approach (Wu et al. 2016) to match MS2 features or using a global function calculated either by a global local weighted regression (LOESS) (Chambers et al. 1992) or a kernel density approach to match features between two runs (Röst et al. 2016; Searle et al. 2018). 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 retention time reversal between any two peptides(Wu et al. 2016).

Here, we present an algorithm which does not require features and is capable of directly aligning the raw multiplexed chromatographic traces from targeted proteomic data. Our approach uses dynamic programming to obtain an optimal mapping between two chromatograms which contain local information in the form of multiple, close-by peaks around the elution peak-group. Independent RT alignment of each precursor facilitates the alignment for elution-order swapped peaks. Our method is also capable of using a global whole-run alignment for guidance, making it robust against noise. With this flexibility, our DIAlign tool provides a knob to select between extremes of global and local alignment fit as user decides.

We provide free-access to our source-code and our R-package can be downloaded from CRAN. We have tested our algorithm on manually validated dataset which shows improved performance over existing methods. We also have tested our algorithm on 24 randomly selected distant plasma runs. We observed that our algorithm outperforms global alignment methods and is capable of correcting mis-annotations introduced by feature detection algorithms. For very distant runs, it could also align switched peaks precisely which is not possible using global alignment methods (Escher et al. 2012; Röst et al. 2014).

Material and Methods:

Validation dataset

For benchmarking of the developed algorithm, previously published and manually validated dataset of *Streptococcus Pyogenes* bacterial strain is used (Röst et al. 2016). Out of 452 transition IDs from (Röst et al. 2016), eight IDs has annotated peak for less than two runs out of 16 runs, making them unsuitable for benchmarking. Seven transition IDs from the remaining set had extracted fragment-ion chromatograms (XICs) outside of annotated peak, making them inapplicable and hence were removed from benchmarking (see the supplemental Section S1). Total 437 transition IDs annotated in 16 runs were considered for testing the performance of the developed DIAlign tool against global alignment approaches. Retention time of these transition IDs in all 16 runs and run names are available in the supplemental Table 1 and supplemental Section S2, respectively. For global alignment, LOESS fit with optimum span value is obtained between two runs (Chambers et al. 1992; Röst et al. 2016). ⅓ cross-validation is performed to obtain the optimum span value. Steps to obtain a global fit (monotone mapping function) are detailed in the supplemental Section S3.

Large-scale human plasma dataset

We have performed SWATH-MS on 975 human plasma samples in 12 batches from 17 February 2017 to 20 July 2017. Tryptic peptides of plasma samples were separated on a NanoLC™ 425 System (SCIEX). 5ul/min flow was used with trap-elute setting using a 0.5 x 10 mm ChromXP™ (SCIEX). LC gradient was set to a 43 minute gradient from 4-32% B with 1 hour total run. Mobile phase A was 100% water with 0.1% formic acid. Mobile phase B was 100% acetonitrile with 0.1% formic acid. 8ug load of undepleted plasma on 15cm ChromXP column. MS analysis were performed using SWATH®Acquisition on a TripleTOF®6600 System equipped with a DuoSpray™Source and 25μm I.D. electrode (SCIEX). Variable Q1 window SWATH Acquisition methods (100 windows) were built in high sensitivity MS/MS mode with Analyst®TF Software 1.7.

To reduce the number of pairwise alignment, randomly two runs from each batch are selected; their metadata and OpenSWATH output files are described in the supplemental Table 3 and SelectedOSW.tar.gz, respectively. In the absence of manually annotated peaks, peaks with low FDR score and highest peak-group rank are considered for performance evaluation. Therefore, the best peak of target precursors with a q-value less than 10^{-3} (m-score < 1e-03, peak-group rank = 1) and common in all 24 runs are selected, and successively, their fragment-ion chromatograms are extracted using OpenSWATH with default parameters (Rosenberger et al. 2017; Röst et al. 2014). Fragment ion chromatograms were parsed from OpenSWATH output using "mzR" package (Chambers et al. 2012). The retention time of these 406 peptides in all 24 runs is provided in the supplemental Table 4. The chromatograms are available in the file SelectedChroms.tar.gz. The global alignment function was fit as described above (see the Supplemental Section S4).

Chromatogram Alignment Algorithm

In targeted proteomics or SWATH experiments, each precursor is measured using one or more fragment ions (transitions) which are measured using an extracted fragment-ion chromatogram (XIC or chromatogram). A collection of one or more chromatogram is called a "chromatogram group" which all map to the same precursor ion. If the same precursor is measured across multiple runs, each run produces a "chromatogram group" for that precursor and this constitutes the raw data for our alignment procedure.

A chromatogram group can be considered a collection of time-series signals. The similarity of the time-series signals between chromatogram groups from runA (ChromA) and runB (ChromB) can be calculated. If a precursor has *n* fragment-ions, therefore, *n* XICs; where each XIC has *I* and *J* time-points in *ChromA* and *ChromB*, respectively as shown in Fig. 1*a*, the similarity between all time-points is represented as a similarity matrix *s,* where

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

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

a. Similarity measure:

In our R package DIAlign, we have implemented several similarity measures which have been suggested in previous literature for chromatograms such as covariance, dot-product, Pearson's correlation, spectral angle and euclidean distance (Röst et al. 2014; Prince and Marcotte 2006). We observed that the dot-product between all *I* and *J* data points provides information about both magnitude and angle between two data-vector, hence segregating elution signal from the background. If each data point of chromatogram is represented by a vector in *n* dimensional space (*n* = 3 in Fig. 1*a*), the resulting dot-product of the two vectors is shown in Fig. 1*b*. Thus, in the case of the dot-product, a similarity matrix s from all vectors of both chromatogram-groups is defined as,

$$s_{ij} = \sum_{k=1}^{n} a_{ik} b_{jk}$$

Where $i \in \{1, ..., I\}$ and $j \in \{1, ..., J\}$ represents index of vectors in *ChromA* and *ChromB*, respectively. A color-coded similarity matrix of size *I* x *J* is shown in Fig. 1*c*. However, to reduce the impact of noise peaks, a modified dot-product is used where higher similarity scores are checked again for spectral angle similarity (see the Supplemental Section S5). A path in the resulting similarity matrix is calculated using dynamic programming which directly translates to a retention time alignment that maps indices/time from *ChromA* to *ChromB* and vice-versa.

b. Penalizing similarity matrix with global alignment:

While dynamic programming will find a path which results into 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 non-linear alignment. To make alignment robust against noise and in order to incorporate information from a global context, we have added an option in our algorithm to modify the the similarity matrix *s* using feature-based global alignment (such as LOESS). Residual Sum of Error (RSE) of fit is utilized to define a region of non-interference in the similarity matrix and values outside of it punished with negative score (see the Supplemental Section S5). This allows us to find an alignment path within a reasonable time window relative to global prediction and avoid large deviations as shown in Fig. 1*d*.

c. Overlap Alignment with affine gap penalty:

The optimal alignment path is 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 (Durbin et al. 1998). Chromatogram-groups *ChromA* and *ChromB* may not have end-to-end mapping as these may only be partial chromatograms which were extracted around the expected peptide elution (as determined by iRT peptides for example). Therefore, overlap alignment instead of a global alignment of MS2 chromatogram groups is employed. This approach allows free end-gaps and thus allows to slide 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 is utilized with higher gap penalty for gap length of more than one. In this approach, three matrices (Matrix M, A, B) are defined which recursively calculates score for gaps of more than unit length (Durbin et al. 1998). The overlap alignment path using affine gap-penalty is presented in Fig. 1*e*. 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 the Supplemental Section S5). Mapping the alignment path to the initial time values provides aligned chromatograms as depicted in Fig. 1*f*.

d. Running time for alignment:

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

e. Optimization of algorithm parameters:

There are various parameters used in DIAlign. A description of these parameters is available in the Supplemental Section S5. We have used a manually validated dataset of 437 *S. pyogenes* peptides acquired with SWATH-MS across 16 LC-MS/MS runs for parameter optimization, using the number of peaks aligned within half chromatographic peak-width and cumulative RT alignment error as our optimization target.

Performance metrics for comparison with current algorithms:

We used the manually validated dataset (Röst et al. 2016) to compare DIAlign to the current state-of-the-art method which utilizes a set of high confidence peaks (“anchor peptides”) to compute a linear or non-linear alignment function that transforms RT values of run1 to RT values of run2. We chose LOESS (local regression) as well as linear regression for our evaluation of a nonlinear alignment function. For LOESS, both optimized spanvalue from cross-validation (as used in TRIC) and default spanvalue (= 0.75) of the R software environment are tested (Chambers et al. 1992).

Retention time error is calculated by comparing against the manual annotation of the *S. Pyogenes* dataset (Röst et al. 2016) and the resulting distribution of the number of peptides aligned within a certain RT tolerance is used as a measure of overall accuracy of the alignment algorithm. Manual annotations are not available for the iPOP dataset, therefore, the high quality results (peaks with low FDR cutoff) of the automated OpenSWATH tool is used for benchmarking.

Results:

Parameter optimization.

Here, we present an algorithm for multi-trace chromatographic alignment that can directly use raw data from targeted proteomics or DIA experiments for retention time alignment. To optimize the performance of our algorithm, we used a manually validated dataset of 7,232 peakgroups (Röst et al. 2016) to investigate the effect of algorithmic parameters on the accuracy of the results. First, we evaluated the performance for different similarity measures of chromatogram groups. The dot product masked with spectral angle as a similarity measure provides the highest fraction of peptides aligned for 120 possible run pairs on the validation dataset (Fig. 2*a*) . Within RT error tolerance of half peak-width (15.3 sec), this similarity measure aligns 94.33% of annotated peaks with the highest area under the curve (see the supplemental Table 6 and 7).

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

Our algorithm is capable of constraining the similarity matrix using a global alignment function. Constraining the alignment in a certain window (given by RSEdistFactor) about the global fit improves the alignment accuracy. We observed that with a constrained similarity matrix 95.4% peaks get aligned compared to 94.3% with non-constrained one (see the supplemental Figure 10). An example of such alignment is shown in Fig. 2*c* and 2*d*, in which the similarity matrix has two high similarity hot-spots. Constraining similarity outside of dashed region in Fig. 2*d*, the alignment path goes through correct hot spot. With the unconstrained similarity matrix, an incorrect alignment resulted as shown in the supplemental Figure 11.

Validation using "gold standard" reference dataset

We then used the manually validated dataset to compare DIAlign to the current state-of-the-art method which utilizes a set of high confidence peaks (“anchor peptides”) to compute a linear or non-linear alignment function that transforms RT values of run1 to RT values of run2. In terms of number of peptides aligned and alignment precision, chromatogram alignment outperforms LOESS and linear regression methods (see the Fig. 3*a* and Table I). On the benchmark dataset, DIAlign improves error rates by 1.8-fold compared to the state-of-the-art. Cumulatively, chromatographic alignment only mis-aligns 4.3% of all peaks within 15.3 seconds (half peak width) of the true RT compared to 7.9% for LOESS (while LOESS with default parameters mis-aligns 22.8% of all peaks and linear regression mis-aligns 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*). When human plasma (10% volume) is added to the sample, the performance of both alignment methods degrades compared to samples without plasma (Fig. 3*b*). Additional drift in LC retention time is expected for a sample of increased complexity (Nigjeh et al. 2017). However, the LOESS performance drop (4.93%) is substantially larger than the corresponding performance drop of DIAlign (2.7%) (Fig 3*b*).

To evaluate the consistency of alignment approaches across multiple run-pairs, we computed the number of aligned-peaks (time mapping falls within half peak-width from annotated RT) for each run-pair. This distribution is shifted towards the right with low standard deviation for chromatogram alignment method compared to the same for LOESS, indicating that the former is consistent in its performance (Fig. 3*c*). In terms of the precision of the alignment, chromatogram alignment consistently performs better than global alignment methods such as LOESS as the former has higher area under the cumulative peptide frequency curve for each run-pair (see the supplemental Figure 12*c*). Similarly, we observed a larger RT variation (standard deviation = 18.45 sec) with the LOESS approach which chromatogram alignment is able to correct satisfactorily with standard deviation being 11.68 sec (Fig 3*d* and supplemental Fig. 12*a*,*b*). We conclude that on the validation dataset, DIAlign performs consistently better in terms of accuracy of alignment and number of aligned-peaks across a range of different RT cutoffs.

Next, we were interested 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 (49,505 alignments) and found that chromatographic alignment outperforms LOESS in 4.7% of all cases (see the supplemental Fig. 13*b*). On average, DIAlign reduces the RT error by 2.3 seconds with a median of 1.7 seconds (see the supplemental Fig. 13*c*). Overall, our method aligns 47.3k peaks compared to 45.6k by an optimized global lowess method within 15.3 seconds (half peak-width). However, in general we observed that on the validation dataset both methods perform with similar consistency which may be due to the low complexity of a bacterial sample and the high homogeneity of the data which was acquired within a single week 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 a large-scale SWATH-MS experiments on human plasma. This experiment provided a more challenging dataset as the data was acquired over the period of six months with an intermittent repair of the instrument and change of LC column. We selected 2 LC-MS/MS runs from each of the 12 batches at random and used 406 peptides for testing our algorithms. Since we did not have manual validations, we selected high confidence peak groups (q-value < 10^{-3}) as our validation peptide set.

Comparing our chromatogram alignment algorithm (DIAlign) with the LOESS method on a highly heterogeneous human plasma dataset, we found that our approach aligns 97.92% of peaks compared to 76.03% using the LOESS method with a maximal error of 20 seconds (half chromatographic peak-width) as depicted in Fig. 4*a*. All tested 276 pairwise alignments shown 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 same and different columns. We found that for runs acquired on different columns, chromatogram alignment method aligns 97.7% of peaks compared to 63.38% by LOESS method (Fig. 4*b*), suggesting that DIAlign retains performance even for highly heterogeneous datasets. New column and instrument repair adds more features in the LC/MS-MS output (see supplemental Figure 2 and supplemental Table 5), therefore, we observed an improvement in LOESS' performance for "column2 pair". However, despite such changes our approach had steady response validating its robustness to such events.

After validating the performance of chromatogram alignment cumulatively, we decided to investigate its consistency across individual run-pair alignments. Fig. 4*c* presents the distribution of the number of peaks aligned in all 276 pairs. DIAlign is capable of aligning 400 peaks on average within half-peak width (while LOESS aligned 309 peaks on average), a 29% improvement. This indicates the inconsistency of LOESS approach which was not observed with DIAlign.

To validate the performance on individual alignment, we further computed the alignment error for each pairwise alignment of each peptide. The standard deviation of alignment error for LOESS was 22.91 sec compared to 13.7 sec for DIAlign (Fig. 4*d*). This indicates the higher precision of RT alignment with our approach. Out of 112,056 alignments, we found that DIAlign outperforms LOESS in 23% of all cases (see the supplemental Figure 14). Thus, testing of chromatogram alignment approach on the heterogeneous human plasma dataset again validates its consistent and improved RT alignment performance.

Switching of peptide elution order

In liquid chromatography, retention time drift is often observed from one run to another run. However, the drift can be variable for different peptides and thus will result in reversal of retention order (Spicer et al. 2010). In such a scenario, two peptides which are eluting in order in one run may reverse their elution order in other run. Since our approach does not make an assumption of order preservation of peptide elution, we hypothesized that DIAlign would be capable of uncovering instances of non-order preserving chromatographic alignment. Specifically, we analyzed the heterogeneous and distant blood plasma runs for peptide pairs that switch elution order.

To confirm the alignment for such peak-switching cases by chromatogram alignment algorithm, we have specifically looked at the alignment of the pair “run4_run23” as it had highest number of peak switching pairs. run4 was part of batch V4 and was acquired on February 28th, 2017 whereas run23 was from batch M3 and was acquired on July 20th, 2017. The LOESS fitting from common high scoring training peptides for this pair is presented in Fig 5*a*. Most of the test peptides are scattered around the global fit line, instead of being directly on the line. This graph quickly suggests 407 peptide pairs (one from either side of the line) compromising of 237 out of 406 peptides which have switched their elution order (see the supplemental Section S6). We thus found that overall, 58.4% of peptides were involved in at least one event of non-order preserving elution.

One of the peak switching cases is presented in Fig 5*b*. In run4 peptide AQLVDMK/2 elutes after HYDGSYSTFGER/2, whereas in run23 the elution order has been reversed. Both peptides have seen positive RT drift in run23 from run4, however, HYDGSYSTFGER/2 had shift of 1070-850 = 270 seconds whereas AQLVDMK/2 had shift of only 1050-900 = 150 seconds. This varying RT drift between two runs has caused the peptides to elute in different order. This peptide pair cannot be aligned with a global alignment approach, which in the best-case scenario will be off by 120 seconds -- however, our chromatogram alignment method has mapped the peaks correctly from run4 to run23 (see the supplemental Figure 17).

To compare DIAlign against other state-of-the-art approaches, we calculated the cumulative fraction of peptides aligned for pair “run4_run23” (Fig 5*c*). Chromatogram alignment correctly aligned 98% peaks compared to LOESS which was able to align only 37.93%, thus decreasing error by up to 30-fold. Eight peaks which were not aligned were further inspected visually by the authors and found to be cases of incorrect annotation of OpenSWATH, mainly due to the mis-annotations of peptides carrying of post-translational modifications (see the supplemental Section S7 and supplemental Figure 18).

Discussion:

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

In this paper, we have presented a novel algorithm that uses the raw fragment ion chromatogram data directly to perform retention time alignment for targeted proteomics and DIA data. Our algorithm uses extracted-ion chromatograms to map peaks across multiple runs and improves accuracy compared to current state-of-the-art methods. We have furthermore extended the algorithm and implemented a hybrid approach that also uses a feature-based global alignment to condition the similarity matrix *s* which led to further gains in accuracy (see the Supplemental Fig. 8). This hybrid approach provides the best of both worlds with a flexible "knob" which 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 non-linear (or gapped) alignment as distant runs also have varying drift even for local peaks. Using a feature-based LOESS method to partially constrain the alignment makes our algorithm more stable and provide the robustness of global alignment methods.

We have shown that on a "gold-standard" validation dataset, our method consistently outperforms a global alignment method (using either linear or non-linear approaches), the current state-of-the-art (Supplemental Figure 12*c*). The DIAlign tool is able to decrease error rates from 7.9% to 4.3% overall. Interestingly, we find that our method is less sensitive to changes in chromatographic condition or sample matrix than global alignment approaches (Fig. 3*b*).

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 blood plasma acquired over several months. On this heterogeneous dataset, DIAlign reduces 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. 4*b*). DIAlign picks correct peak-group and reduces identification errors; an example of wrong peak-group picking by global alignment method is presented in the Supplemental Figure 19. We have also shown that in the case of a peak being outside of chromatograms, our method is able to map retention time outside of it as it also uses global alignment (the Supplemental Figure 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.

We believe that our approach is most useful for large-scale heterogeneous targeted proteomics studies where runs are acquired by different personnel and data is 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 in nature assume chronological order of peptide elution and, therefore, cannot align switched peptides. However, we have shown that our hybrid approach aligns these peptides accurately as it mostly relies on additional dimensions of fragment-ion m/z to align peaks. It is possible that switching peptides may share fragment-ion m/z, however, this is very rare scenario and in that case our method will perform no worse than global alignment methods.

Applying mass spectrometry-based proteomics in large-scale systems biology studies, high reproducibility and quantitation of large number of analytes is imperative. We present a tool that can be used to establish correspondence between analytes across large number of samples, making DIA amenable for multi-center and longitudinal studies. We also expect that this tool can be utilized by existing proteomics softwares to streamline analyte identification and improve the quantification.

Acknowledgements:

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

Author contributions:

S.G. designed and wrote code, performed data analysis and produced the figures. S.A. and W.Z. performed human plasma related experiments, acquired MS data. H.R. designed and supervised the study. S.G. and H.R. contributed to writing the manuscript. All authors have contributed to the final manuscript.

Competing Financial Interest:

The authors declare no competing financial interests.

Supplementary Material:

References:

Chambers, J. M., Hastie, T. J. & Others. *Statistical models in S*. **251,** (Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA, 1992).

Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30,** 918–920 (2012).

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. (Cambridge university press, 1998).

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

Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11,** O111.016717 (2012).

Guo, T. *et al.* Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Med.* **21,** 407–413 (2015).

Listgarten, J., Neal, R. M., Roweis, S. T. & Emili, A. Multiple Alignment of Continuous Time Series. in *Advances in Neural Information Processing Systems 17* (eds. Saul, L. K., Weiss, Y. & Bottou, L.) 817–824 (MIT Press, 2005).

Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11,** 786 (2015).

Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34,** 1130–1136 (2016).

Nigjeh, E. N. *et al.* Quantitative Proteomics Based on Optimized Data-Independent Acquisition in Plasma Analysis. *J. Proteome Res.* **16,** 665–676 (2017).

Omenn, G. S. *et al.* Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5,** 3226–3245 (2005).

Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.* **14,** 3452–3460 (2015).

Prince, J. T. & Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* **78,** 6140–6152 (2006).

Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14,** 921–927 (2017).

Rost, H. L., Malmstrom, L. & Aebersold, R. A computational tool to detect and avoid redundancy in selected reaction monitoring. *Mol. Cell. Proteomics* (2012). doi:10.1074/mcp.M111.013045

Röst, H. L., Schmitt, U., Aebersold, R. & Malmström, L. pyOpenMS: a python-based interface to the OpenMS mass-spectrometry algorithm library. *Proteomics* **14,** 74–77 (2014).

Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32,** 219–223 (2014).

Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13,** 741–748 (2016).

Röst, H. L. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13,** 777–783 (2016).

Sandin, M. *et al.* An adaptive alignment algorithm for quality-controlled label-free LC-MS. *Mol. Cell. Proteomics* **12,** 1407–1420 (2013).

Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat. Protoc.* **12,** 1289–1294 (2017).

Searle, B. C. *et al.* Comprehensive peptide quantification for data independent acquisition mass spectrometry using chromatogram libraries. *bioRxiv* 277822 (2018). doi:10.1101/277822

Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief. Bioinform.* **16,** 104–117 (2015).

Spicer, V., Grigoryan, M., Gotfrid, A., Standing, K. G. & Krokhin, O. V. Predicting retention time shifts associated with variation of the gradient slope in peptide RP-HPLC. *Anal. Chem.* **82,** 9678–9685 (2010).

Surinova, S. *et al.* On the development of plasma protein biomarkers. *J. Proteome Res.* **10,** 5–16 (2011).

Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12,** 258 (2015).

Uzozie, A. C. & Aebersold, R. Advancing translational research and precision medicine with targeted proteomics. *J. Proteomics* (2018). doi:10.1016/j.jprot.2018.02.021

Voss, B. *et al.* SIMA: simultaneous multiple alignment of LC/MS peak lists. *Bioinformatics* **27,** 987–993 (2011).

Wu, L., Amon, S. & Lam, H. A hybrid retention time alignment algorithm for SWATH-MS data. *Proteomics* **16,** 2272–2283 (2016).

FIG. 1. **Alignment algorithm for targeted proteomics MS2 chromatograms.** *a*, Fragment-ions chromatograms of a peptide for two runs; *run A* at top and *run B* at bottom. Correct peak, typically, has all library fragment-ions (n = 3) coeluting. *b*, similarity between chromatograms of both runs is calculated by dot-product of intensity vector; defined in n dimensional space. *c*, outer dot-product of chromatograms provides an I x J similarity score matrix (S). *d*, feature-based complete run alignment is used as an approximate path for alignment. Time points farther from an allowed window in similarity score matrix are penalized by adding negative score. *e*, Affine gap penalty based overlap alignment strategy is employed for calculating best scoring path through the similarity matrix. This dynamic programming based strategy utilizes three matrices for recursively calculating multiple gap length scores. Calculated alignment path is indicated using black arrow. *f*, Chromatograms recreated by mapping intensity back to aligned time path.

FIG. 2. **Comparison of different similarity measurements, technical parameters and effect of penalizing similarity using global prior on the accuracy of alignment in *S. Pyogenes* dataset.** *a*, Performance of various similarity measures as the cumulative fraction of peptides having error less than RT difference is plotted. *b*, the effect of gap penalty selection using gapQuantile on the percentage of peaks aligned within certain RT difference tolerance is depicted. *c*, The penalized similarity matrix for peptide DGSVSVADSGR/2 between run11 and run12 is presented. From available two high similarity vectors, alignment path passes through high similarity vectors B. *d*, The end-points of extracted ion chromatograms (XICs) for the peptide are shown as green dots. Penalizing similarity gives preference to alignment within a certain window around LOESS fit, depicted as dashed green lines. Here, alignment of high similarity vectors B (solid red circle ●) is preferred over high similarity vectors A (red cross □).

FIG. 3. **Alignment accuracy of MS2 chromatogram alignment on a validation dataset of 16 runs with manually annotated 437 peak groups in each run.** *a*, cumulative fraction of peptides having error less than RT difference is plotted for all possible C(16,2) = 120 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 biological conditions. Strep0 pair constitutes both 0% plasma runs, Strep10 pair is composed of both 10% plasma runs and Strep0_Strep10 pair have one run with 0% plasma and other with 10% plasma. There are 28 Strep0 pairs, 28 Strep10 pairs and 64 pairs for Strep0_Strep10 case in the validation dataset. *c*, histogram of number of peptides matched within half peak-width for LOESS and chromatogram alignment. *d*, Histogram of retention time (RT) prediction error is plotted for chromatogram alignment and LOESS. RT difference standard deviation for both approaches is 9.56 sec and 10.98 sec, respectively.

FIG. 4. **Alignment accuracy of MS2 chromatogram alignment on 24 runs of clinical plasma measurement dataset annotated with OpenSWATH. 406 peak groups are selected in each run with m-score < 0.001.** *a*, cumulative fraction of peptides having error less than 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 quadrupole replacement. 14 runs were acquired on column1 which makes 91 pairs, labeled as "column1". 10 runs were acquired after quadrupole replacement on column2 which results into 45 pairs, labelled as "column2". There are 140 pairs composed of "column1" and "column2" labelled runs; these pairs are labelled as "column1-column2". *c*, histogram of number of peptides matched within half peak-width for LOESS and chromatogram alignment. *d*, Histogram of retention time (RT) prediction error is plotted for chromatogram alignment and LOESS. RT difference standard deviation for both approaches is 22.91 sec and 13.7 sec, respectively.

FIG. 5. **Alignment of 406 peptides in pair *run4 and run23* from clinical plasma measurement dataset.** run4 "022817_V4_Plasma_8ug_C11_010−05−02−2−V3−Plasma083" was acquired on February 28th, 2017 whereas run23 "072017_M3_Plasma_8ug_C4_69−090−1031−M3−Plasma027" was acquired on July 20th, 2017**.** *a*, LOESS fit between two runs is obtained using confident peaks. Test peptides are shown in red color around the fit line. Span value = 0.27 for fit is obtained by ⅓ cross-validation. Precursors AQLVDMK/2 and HYDGSYSTFGER/2 are shown in magenta and orange circle-cross symbols, respectively. *b*, Two peptides AQLVDMK/2 and HYDGSYSTFGER/2 have their elution order reversed in these runs. This phenomenon makes alignment of peaks theoretically impossible for global monotonic methods. Chromatogram alignment uses fragment-ions as additional dimensions and hence can align them precisely. *c*, fraction of peptides having error less than RT difference is plotted for pair *run4 and run23* for chromatogram alignment, linear fit, k-nearest neighbor smoothing (LOESS) with and without optimum span and without any alignment.

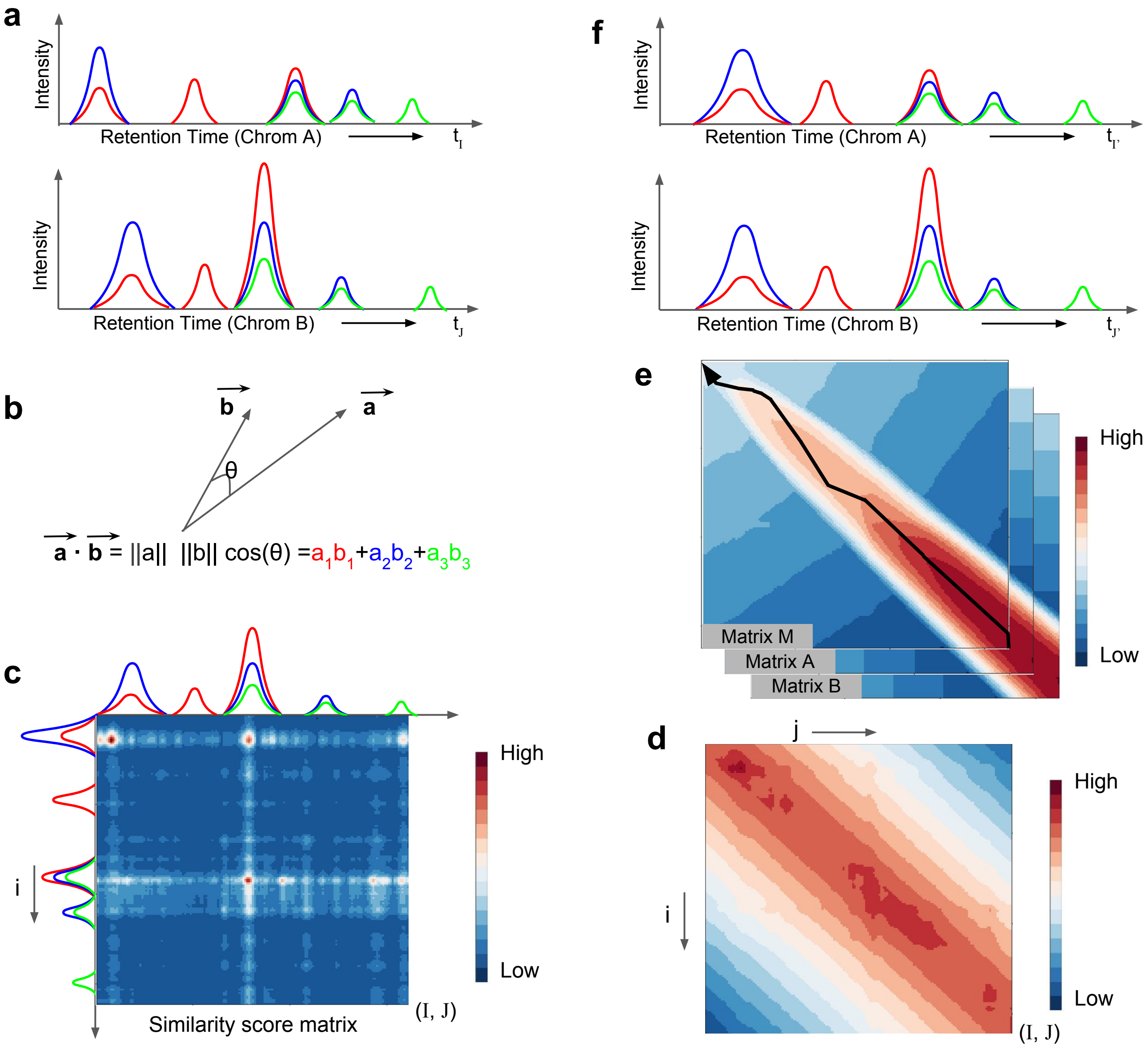
a
Intensity
Retention Time (Chrom A)
t_I
Intensity
Retention Time (Chrom B)
t_J
b
$\vec{b}$
$\vec{a}$
θ
$\vec{a} \cdot \vec{b}$ = ||a|| ||b|| cos(θ) = a_1b_1+a_2b_2+a_3b_3
c
i
Similarity score matrix
(I, J)
High
Low
f
Intensity
Retention Time (Chrom A)
$t_{I'}$
Intensity
Retention Time (Chrom B)
$t_{J'}$
e
Matrix M
Matrix A
Matrix B
High
Low
d
j
i
(I, J)
High
Low

FIG. 1. **Alignment algorithm for targeted proteomics MS2 chromatograms.** *a*, Fragment-ions chromatograms of a peptide for two runs; *run A* at top and *run B* at bottom. Correct peak, typically, has all library fragment-ions (n = 3) coeluting. *b*, similarity between chromatograms of both runs is calculated by dot-product of intensity vector; defined in n dimensional space. *c*, outer dot-product of chromatograms provides an I x J similarity score matrix (S). *d*, feature-based complete run alignment is used as an approximate path for alignment. Time points farther from an allowed window in similarity score matrix are penalized by adding negative score. *e*, Affine gap penalty based overlap alignment strategy is employed for calculating best scoring path through the similarity matrix. This dynamic programming based strategy utilizes three matrices for recursively calculating multiple gap length scores. Calculated alignment path is indicated using black arrow. *f*, Chromatograms recreated by mapping intensity back to aligned time path.

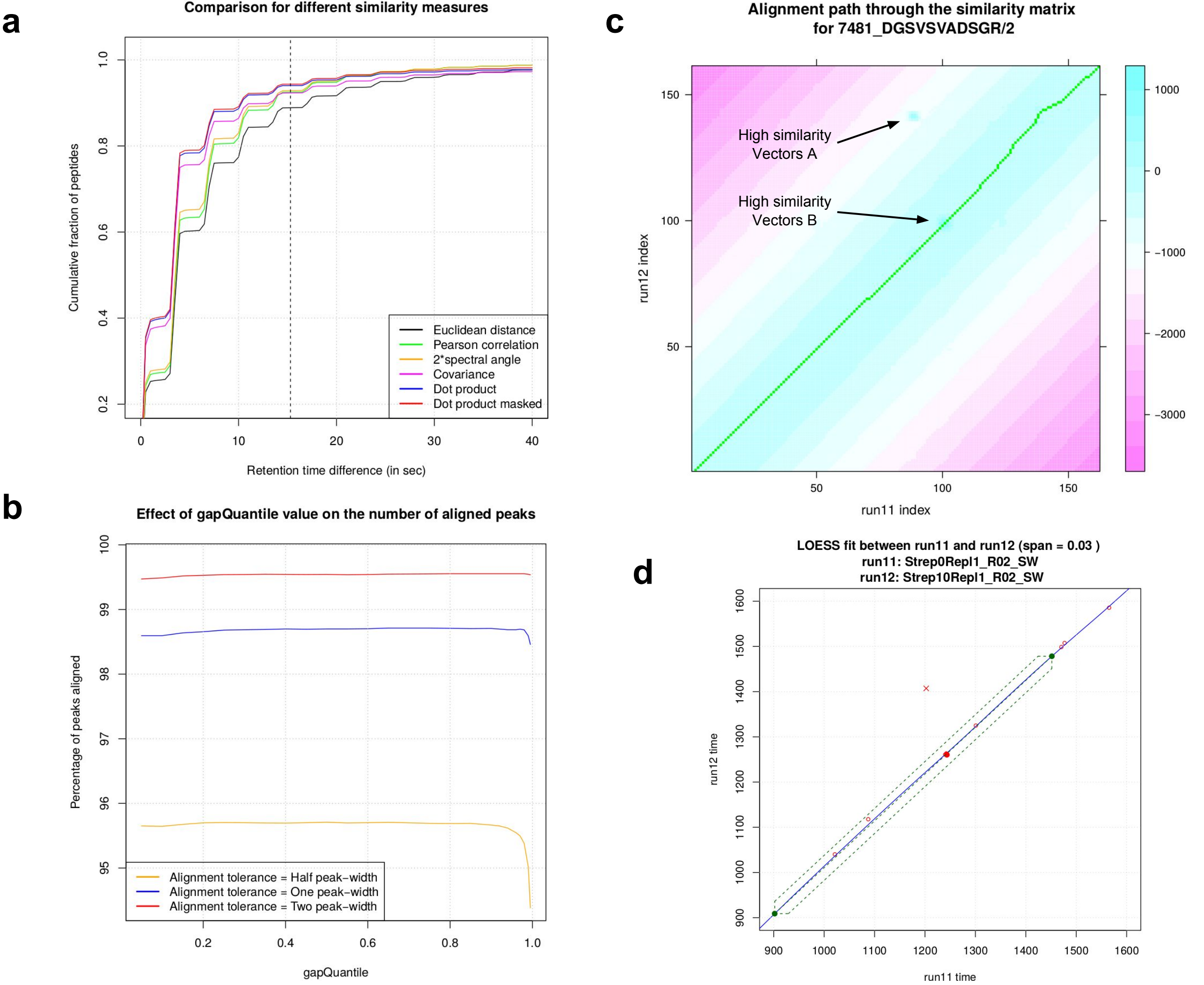
a
Comparison for different similarity measures
Cumulative fraction of peptides
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Retention time difference (in sec)
Euclidean distance
Pearson correlation
2*spectral angle
Covariance
Dot product
Dot product masked
b
Effect of gapQuantile value on the number of aligned peaks
Percentage of peaks aligned
95
96
97
98
99
100
0.2
0.4
0.6
0.8
1.0
gapQuantile
Alignment tolerance = Half peak−width
Alignment tolerance = One peak−width
Alignment tolerance = Two peak−width
c
Alignment path through the similarity matrix
for 7481_DGSVSVADSGR/2
High similarity
Vectors A
High similarity
Vectors B
run12 index
50
100
150
run11 index
50
100
150
1000
0
−1000
−2000
−3000
d
LOESS fit between run11 and run12 (span = 0.03)
run11: Strep0Repl1_R02_SW
run12: Strep10Repl1_R02_SW
run12 time
900
1000
1100
1200
1300
1400
1500
1600
run11 time
900
1000
1100
1200
1300
1400
1500
1600

FIG. 2. **Comparison of different similarity measurements, technical parameters and effect of penalizing similarity using global prior on the accuracy of alignment in *S. Pyogenes* dataset.** *a*, Performance of various similarity measures as the cumulative fraction of peptides having error less than RT difference is plotted. *b*, the effect of gap penalty selection using gapQuantile on the percentage of peaks aligned within certain RT difference tolerance is depicted. *c*, The penalized similarity matrix for peptide DGSVSVADSGR/2 between run11 and run12 is presented. From available two high similarity vectors, alignment path passes through high similarity vectors B. *d*, The end-points of extracted ion chromatograms (XICs) for the peptide are shown as green dots. Penalizing similarity gives preference to alignment within a certain window around LOESS fit, depicted as dashed green lines. Here, alignment of high similarity vectors B (solid red circle ●) is preferred over high similarity vectors A (red cross □).

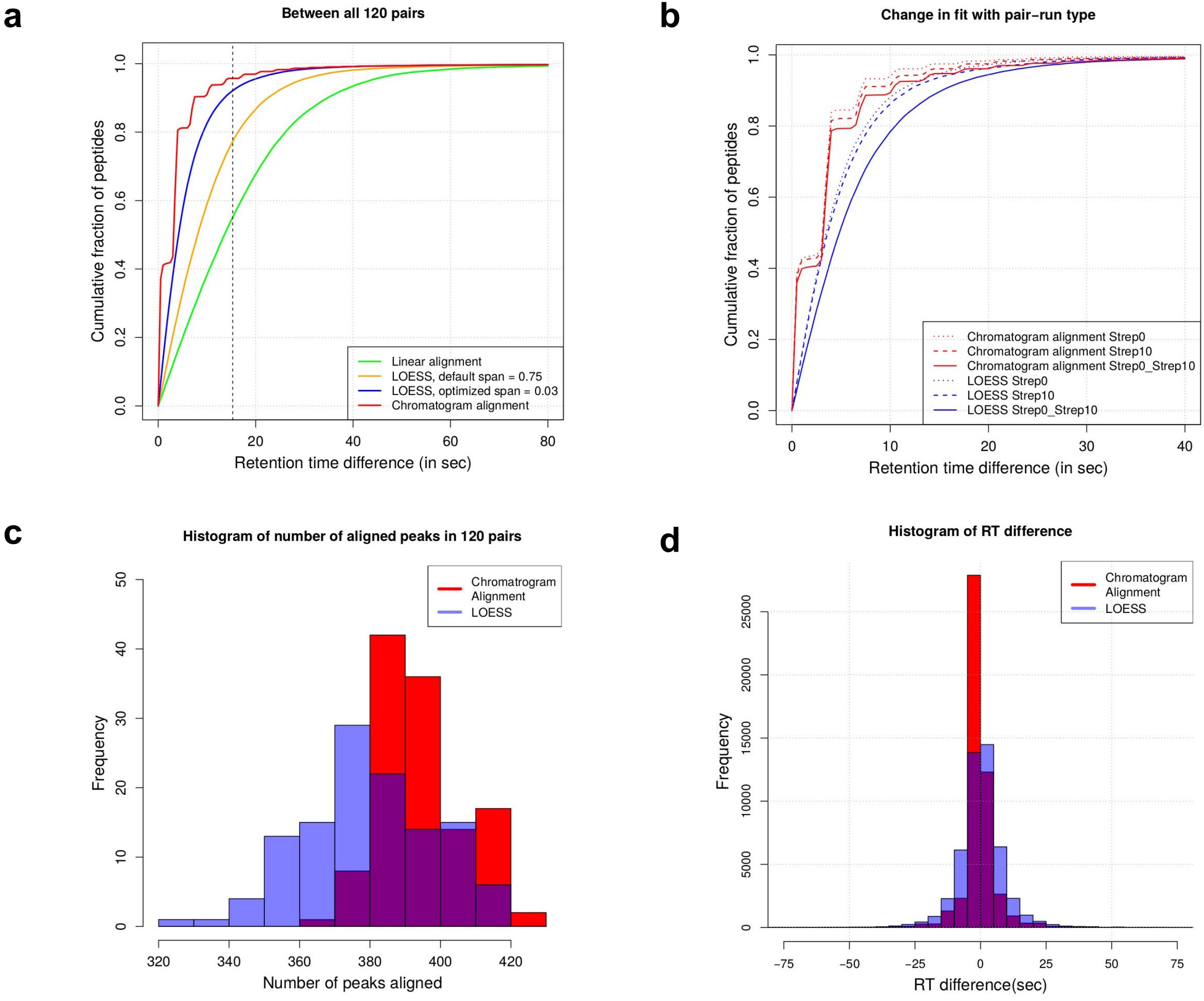
a
Between all 120 pairs
Cumulative fraction of peptides
0.0
0.2
0.4
0.6
0.8
1.0
Linear alignment
LOESS, default span = 0.75
LOESS, optimized span = 0.03
Chromatogram alignment
0
20
40
60
80
Retention time difference (in sec)
b
Change in fit with pair−run type
Cumulative fraction of peptides
0.0
0.2
0.4
0.6
0.8
1.0
Chromatogram alignment Strep0
Chromatogram alignment Strep10
Chromatogram alignment Strep0_Strep10
LOESS Strep0
LOESS Strep10
LOESS Strep0_Strep10
0
10
20
30
40
Retention time difference (in sec)
c
Histogram of number of aligned peaks in 120 pairs
Frequency
0
10
20
30
40
50
Chromatrogram Alignment
LOESS
320
340
360
380
400
420
Number of peaks aligned
d
Histogram of RT difference
Frequency
0
5000
10000
15000
20000
25000
Chromatogram Alignment
LOESS
−75
−50
−25
0
25
50
75
RT difference(sec)

FIG. 3. **Alignment accuracy of MS2 chromatogram alignment on a validation dataset of 16 runs with manually annotated 437 peak groups in each run.** *a*, cumulative fraction of peptides having error less than RT difference is plotted for all possible C(16,2) = 120 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 biological conditions. Strep0 pair constitutes both 0% plasma runs, Strep10 pair is composed of both 10% plasma runs and Strep0_Strep10 pair have one run with 0% plasma and other with 10% plasma. There are 28 Strep0 pairs, 28 Strep10 pairs and 64 pairs for Strep0_Strep10 case in the validation dataset. *c*, histogram of number of peptides matched within half peak-width for LOESS and chromatogram alignment. *d*, Histogram of retention time (RT) prediction error is plotted for chromatogram alignment and LOESS. RT difference standard deviation for both approaches is 9.56 sec and 10.98 sec, respectively.

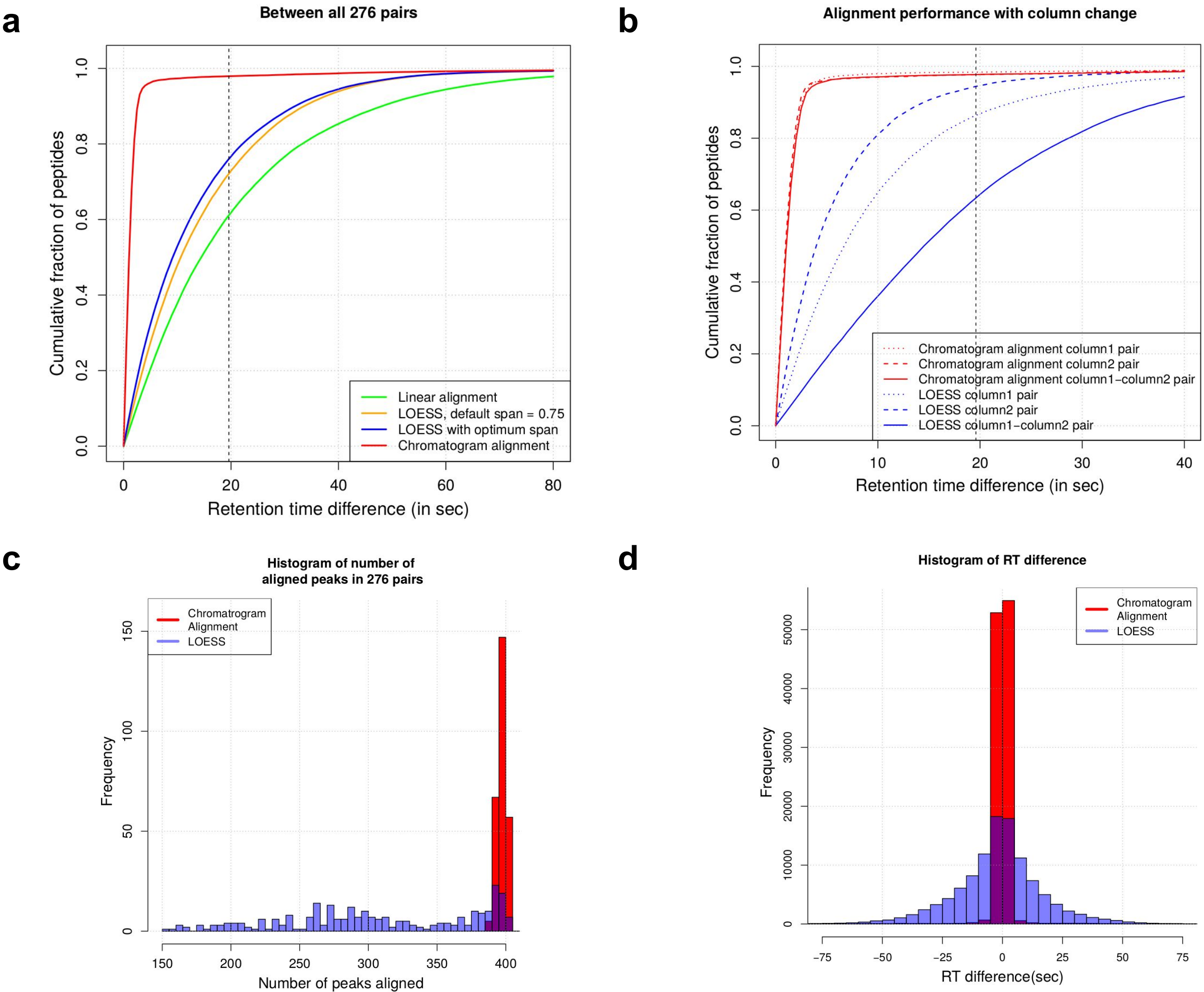
a
Between all 276 pairs
Cumulative fraction of peptides
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Retention time difference (in sec)
Linear alignment
LOESS, default span = 0.75
LOESS with optimum span
Chromatogram alignment
b
Alignment performance with column change
Cumulative fraction of peptides
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Retention time difference (in sec)
Chromatogram alignment column1 pair
Chromatogram alignment column2 pair
Chromatogram alignment column1−column2 pair
LOESS column1 pair
LOESS column2 pair
LOESS column1−column2 pair
c
Histogram of number of
aligned peaks in 276 pairs
Chromatrogram
Alignment
LOESS
Frequency
0
50
100
150
150
200
250
300
350
400
Number of peaks aligned
d
Histogram of RT difference
Chromatogram
Alignment
LOESS
Frequency
0
10000
20000
30000
40000
50000
−75
−50
−25
0
25
50
75
RT difference(sec)

FIG. 4. **Alignment accuracy of MS2 chromatogram alignment on 24 runs of clinical plasma measurement dataset annotated with OpenSWATH. 406 peak groups are selected in each run with m-score < 0.001.** *a*, cumulative fraction of peptides having error less than 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 quadrupole replacement. 14 runs were acquired on column1 which makes 91 pairs, labeled as “column1”. 10 runs were acquired after quadrupole replacement on column2 which results into 45 pairs, labelled as “column2”. There are 140 pairs composed of “column1” and “column2” labelled runs; these pairs are labelled as “column1-column2”. *c*, histogram of number of peptides matched within half peak-width for LOESS and chromatogram alignment. *d*, Histogram of retention time (RT) prediction error is plotted for chromatogram alignment and LOESS. RT difference standard deviation for both approaches is 22.91 sec and 13.7 sec, respectively.

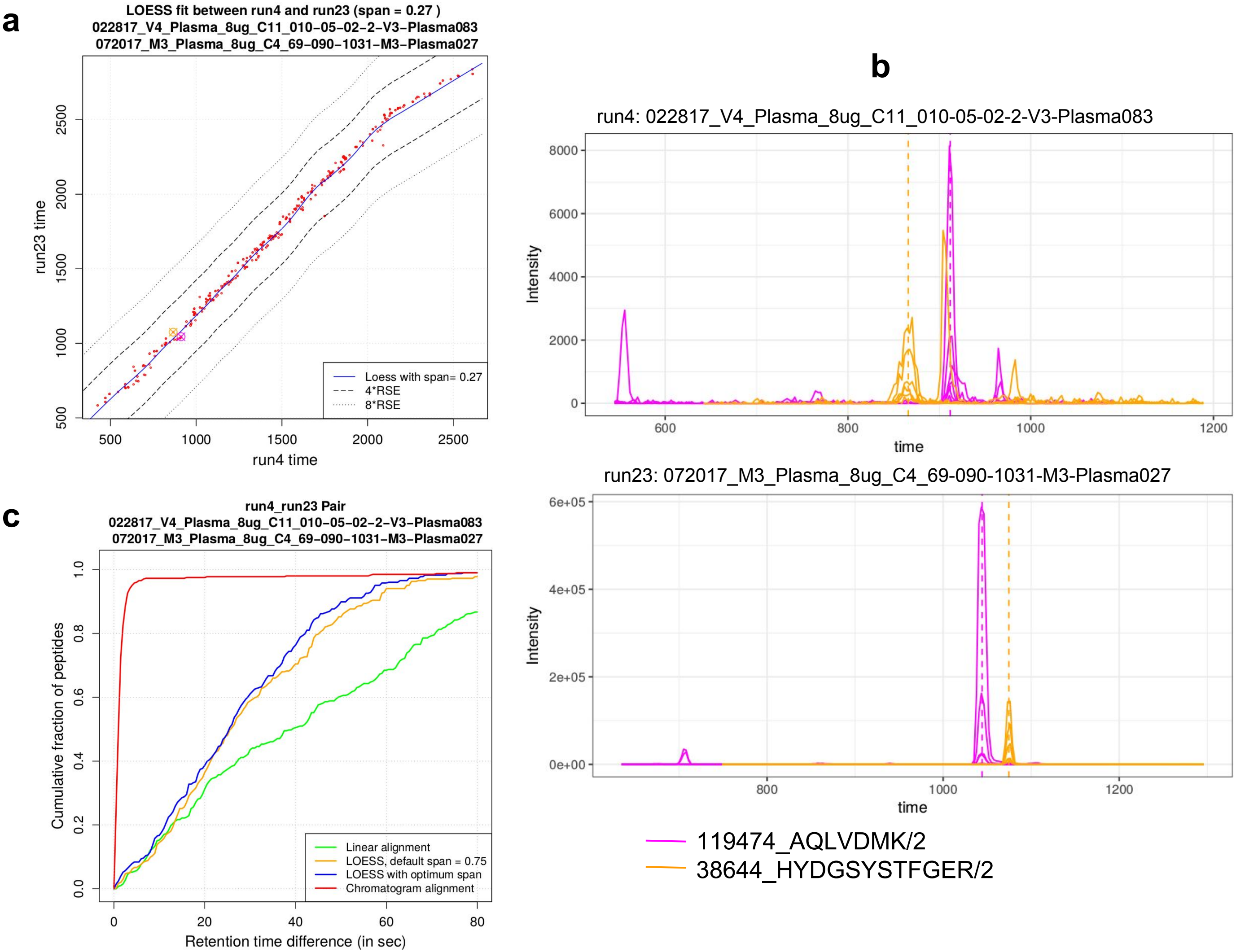
a
LOESS fit between run4 and run23 (span = 0.27)
022817_V4_Plasma_8ug_C11_010-05-02-2-V3-Plasma083
072017_M3_Plasma_8ug_C4_69-090-1031-M3-Plasma027
run23 time
500
1000
1500
2000
2500
run4 time
500
1000
1500
2000
2500
Loess with span= 0.27
4*RSE
8*RSE
b
run4: 022817_V4_Plasma_8ug_C11_010-05-02-2-V3-Plasma083
Intensity
8000
6000
4000
2000
0
600
800
1000
1200
time
run23: 072017_M3_Plasma_8ug_C4_69-090-1031-M3-Plasma027
Intensity
6e+05
4e+05
2e+05
0e+00
800
1000
1200
time
119474_AQLVDMK/2
38644_HYDGSYSTFGER/2
c
run4_run23 Pair
022817_V4_Plasma_8ug_C11_010-05-02-2-V3-Plasma083
072017_M3_Plasma_8ug_C4_69-090-1031-M3-Plasma027
Cumulative fraction of peptides
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Retention time difference (in sec)
Linear alignment
LOESS, default span = 0.75
LOESS with optimum span
Chromatogram alignment

FIG. 5. **Alignment of 406 peptides in pair *run4 and run23* from clinical plasma measurement dataset.** run4 “022817_V4_Plasma_8ug_C11_010−05−02−2−V3−Plasma083” was acquired on February 28th, 2017 whereas run23 “072017_M3_Plasma_8ug_C4_69−090−1031−M3−Plasma027” was acquired on July 20th, 2017**.** *a*, LOESS fit between two runs is obtained using confident peaks. Test peptides are shown in red color around the fit line. Span value = 0.27 for fit is obtained by ⅓ cross-validation. Precursors AQLVDMK/2 and HYDGSYSTFGER/2 are shown in magenta and orange circle-cross symbols, respectively. *b*, Two peptides AQLVDMK/2 and HYDGSYSTFGER/2 have their elution order reversed in these runs. This phenomenon makes alignment of peaks theoretically impossible for global monotonic methods. Chromatogram alignment uses fragment-ions as additional dimensions and hence can align them precisely. *c*, fraction of peptides having error less than RT difference is plotted for pair *run4 and run23* for chromatogram alignment, linear fit, k-nearest neighbor smoothing (LOESS) with and without optimum span and without any alignment.