Near-noiseless ion networks from multiple data independent acquisition mass spectrometry samples allow a fragment-centric perspective and single window ion mobility DIA

Sander Willems[1], Simon Daled[1], Bart van Puyvelde[1], Laura de Clerck[1], Sofie vande Casteele[1], Filip van Nieuwerburgh[1], Dieter Deforce[1], and Maarten Dhaenens[1]

[1]*Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium*

Poor reproducibility of data-dependent acquisition (DDA) in mass spectrometry based proteomics can be complemented with data-independent acquisition (DIA). To date, DIA needs to balance comprehensive data acquisition and comprehensive data analysis. Frequently, DIA is analysed peptide-centric instead of spectrum-centric such as in DDA. While peptide-centric approaches can handle chimericy and incorporate fragment retention times, they unfortunately show limited false discovery rate (FDR) control.

Here we present single window ion mobility (SWIM)-DIA, wherein all precursors are continuously fragmented in a single 2000 m/z window to acquire all fragments without precursor selection or survey scans. While this provides a fully comprehensiveness data acquisition, the data complexity requires deconvolution. Therefore, we leverage the improved reproducibility of SWIM-DIA to convert data of multiple samples into a single near-noiseless ion network based on consistent co-elution of individual ions.

This near-noiseless ion network offers a novel fragment-centric perspective. As a proof-of-concept, we show that a naive database search algorithm can be applied to a near-noiseless ion network of a mixture of commercial human, yeast and ecoli tryptic peptides. Hereby, we annotated more than a 120 thousand reproducible ions as individual b- and y-ions at an experimentally verified 1% FDR, accounting for 10 thousand peptides of 2000 proteins.

1 Introduction

Mass spectrometry (MS)-based proteomics is traditionally done with data-dependent acquisition (DDA). Herein MS is preceded by liquid chromatography (LC) to acquire multiple low energy (LE) scans for analytes that continuously elute over time. For each of these LE scans, a few precursors are selected for high energy (HE) scans, in which fragmentation occurs. The selection of precursors for fragmentation generally relies on the intensity and charge state of precursors and hence is data dependent. This acquisition methodology has several inherent limitations, such as 1) poor reproducibility due to stochastic precursor selection, 2) no temporal information on fragments obtained in HE scans, and 3) a limited duty cycle, defined as the ratio of ions formed in electrospray ionization (ESI) that enter the mass spectrometer and finally reach the detector, as limited time is spent on LE scans.

In recent years, several data-independent acquisition (DIA) techniques have been developed that replace the data-dependent precursor selection of DDA with a partitioning of predefined mass over charge ratio (m/z) windows for fragmentation. The main differences between most of these techniques are the number and size of the predefined m/z windows. Currently, the most popular technique is probably sequential window acquisition of all theoretical mass spectra (SWATH), introduced by AB Sciex in 2012, in which each LE scan is typically followed by 32 or 64 HE scans of 20 or 10 m/z wide. Another technique is Waters' MS^e that was introduced in 2004, in which each LE scan is followed by a single HE scan in which all detected precursors, typically between 0 and 2000 m/z, are fragmented. They have since improved upon this technique with the introduction of an ion mobility separation (IMS) cell between their ion guide and collision cell, defining this technique as high definition MS^e (HDMS^e). The IMS cell separates precursors based on their collisional cross section (CCS). This separation, with drift time (DT) as metric, is achieved in milliseconds so that it fits exactly between the LC in which retention time (RT) is measured in seconds and the time of flight (TOF) detector in which m/z is measured in microseconds. Both SWATH and HDMS^e, among several others, have experimentally proven to provide more reproducible data than DDA and temporal information on fragments, while only HDMS^e has improved upon the duty cycle.

Even though these seem like clear advantages of DIA compared to DDA, there is also reason for caution. Since the precursor selection is replaced with predefined m/z windows in DIA, the subsequent HE spectra are more chimeric, meaning they contain fragments from multiple precursors. This chimericy can be reduced by taking smaller windows, but this comes at a cost of needing more windows to cover fragmentation of all precursors detectable in LE. This subsequently means either shorter scan times are used for HE scans, or that the cycle time, defined as the time needed to return

to the same window, increases while the duty cycle decreases. The former has the disadvantage that lower intensities with higher coefficient of variation (CV) are measured [1], possibly even below linear detector range. The latter means fewer points can be used to define an extracted ion chromatogram (XIC), typically below the recommended standard of nine points. In either case, there is a trade-off between data chimericy and comprehensiveness of temporal and intensity information.

Once DIA data is acquired, there are several approaches to process it. Most of these approaches can be divided by two characteristics: library-based versus library-free and spectrum-centric versus peptide-centric. In a library-based approach a pre-annotated library with known fragmentation patterns and intensities is used, as opposed to library-free approaches relying only on in-silico information. Libraries introduce the potential to gain specificity from prior information, but are often build with DDA data, either directly or indirectly, and thereby partially negate the advantages of DIA. In spectrum-centric approaches each query spectrum is annotated with target peptides that best explain these query spectra, whereas in peptide-centric approaches evidence for query peptides is obtained from the acquired data. Many spectrum-centric approaches have origins in DDA and are not tailor-made for DIA data. As such they can be susceptible to chimericy and therefore require deconvolution to partition the spectra by precursor origin. Peptide-centric approaches on the other hand tend to exhibit limited false discovery rate (FDR) control as they are often based on multiple reaction monitoring (MRM)-like approaches that are not always scalable to DIA data. With the exception of GROUP-DIA, all approaches are performed on a per-run basis, even though the reproducibility of DIA is generally considered as its strongest trait.

Here, we present HistoPyA, a tool that demultiplexes DIA data into a near noiseless ion-network based on replicate samples. As this ion-network has minimal noise, it eliminates the need for specifity obtained through DDA-based spectral libraries. Furthermore the ion-network is neither scan- nor peptide-centric, allowing a fragment-centric annotation approach developed especially for DIA data with an intuitive FDR control. The creation of this ion-network builds on the following hypothesis: *Poor separation of analytes is the primary cause of chimeric HE spectra. However, there are stochastic differences between runs, even between replicate injections from a single sample vial. As fragmentation of precursors occurs after separation (in LC and IMS or precursor window partitioning), HE data can be demultiplexed based on (in)consistency of co-separation.* As such, HistoPyA uses the greatest advantage of DIA over DDA, namely its reproducibility and temporal information of fragments, as an additional dimension in the data.

In brief, HistoPyA consists of the following steps (Figure 1). First, raw data of each individual sample is peak-picked in all dimensions to obtain a list of ions and their intensities. Each ion of each sample is now defined by three coordinates: 1) an m/z apex, 2) an RT apex, and 3) a DT

apex or precursor window. After a quick calibration and estimation of inter-run differences, all ions of all samples are simultaneously aligned to obtain a list of aggregate ions, where each aggregate ion is composed of ions from different samples with equal m/z, RT, and DT or precursor window. Hereafter, an ion-network is created with aggregate ions as vertices and edges between aggregate ions if and only if they are consistently co-separated in all overlapping samples of their individual ions. The definition of co-separation is estimated from pairs of likely isotopes in LE scans. Finally, an X!Tandem-like hyperscore is calculated for each HE aggregate ion [2], based on the aggregate m/z of all its neighbors and itself, as well as the existence of a potential LE precursor. Using Percolator, an FDR can then be calculated for each aggregate ion, which can be extrapolated to the precursor, peptide, and protein level [3]. Optionally, a relative quantification can be performed for each aggregate ion, again allowing an extrapolation to precursor, peptide, and protein level.

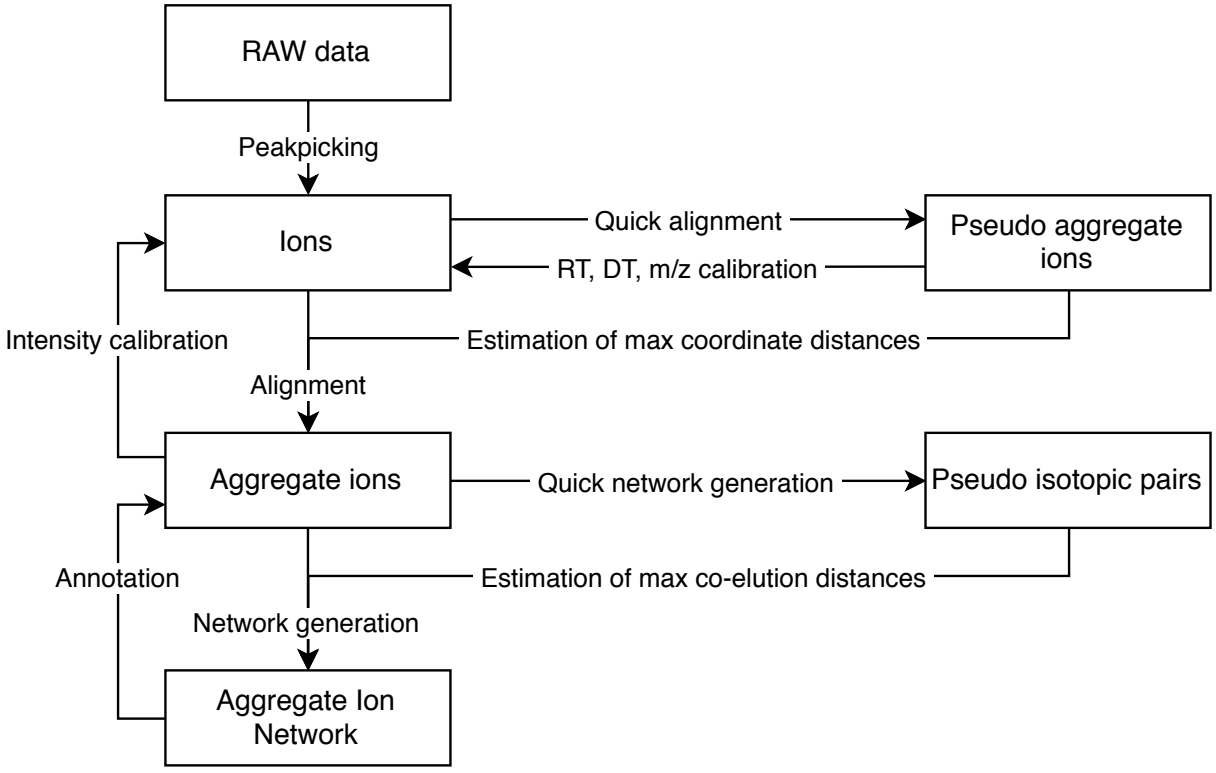

Figure 1: Schematic overview of HistoPya's workflow.

To test the performance of HistoPyA, two different hybrid proteomes of human, yeast and ecoli peptides were prepared with a logarithmic fold change (logFC) between condition A and B of respectively 0, 1 and -2 [4]. For each condition 3 samples were prepared, as well as a quality control (QC) sample in which all samples were pooled. For each normal sample 3 replicates were acquired and for the QC sample 9 replicates were acquired, resulting in a total of 27 samples. Based on these samples, 5 million different reproducible HE aggregate ions could be detected within a network containing 100 million edges. Depending on how well an aggregate ion was reproducible, more than

98% of these edges existed between ions with a similar logFC, indicating the same LE origin and a very high specificity with low noise. The annotation of all these aggregate ions equally showed more than 99% with a correct logFC for more than 200 thousand aggregate ions, resulting in 25 thousand and 4 thousand proteins, each at their respective 1% FDR.

These results indicate there is much to gain from such ion networks. The low noise especially would allow for many new DIA annotation algorithms, including *de novo* approaches or partitioning algorithms to trim the network by precursor origin to obtain traditional spectra. Moreover, the low noise in the network was obtained without precursors and LE precursors are only used in the last annotation step as specificity boost. This effectively means simpler hardware can be developed, in which an LC is coupled directly to an IMS cell followed by a collision cell and detector, skipping any precursor selection mechanisms such as a quadrupole. From a software perspective the statistics could be developed further, as all aggregate ions have multiple measurement over different samples and the quantification can be done on a HE level instead of a LE level. Finally HistoPyA is very modular, meaning each part of the algorithm (Figure 1) can be easily substituted by other approaches, allowing many improvements.

In DDA the measured analytes are peptides and thus a spectrum-centric or peptide-centric approach is justifiable. However, in DIA the direct connection between precursor- and fragment-ions is not present anymore. As such, we argue that a peptide-centric approach is not fully applicable. At the same time, DIA spectra are of high complexity and not independent from one another, thereby also invalidating a spectrum-centric approach. What is truly measured in DIA, are independent ions in multiple dimensions. As such, a more intuitive perspective than spectrum- or peptide-centric would be fragment/ion-centric.

2 Results

2.1 Peak picking and inter-run ion alignment

The core premise of HistoPyA is that HE ions which are consistently co-eluting in all their representative samples are likely to be derived from the same LE precursor. Therefore, a primary component of HistoPyA is its ability to determine which ions are *the same* in different samples.

Using Water's Apex 3D peak picking algorithm the raw data from 27 samples (nine per condition A, B and QC) were peak picked in all of its dimensions: m/z, RT, DT and intensity. Even though these samples had different ratios for human, yeast and ecoli proteins, the resulting number of

detectable ions was similar for all of them, with each sample resulting in roughly 6 million ions in both LE and HE ions (Supplementary Table 1) at the lowest thresholds of 1 for both LE and HE.

2.1.1 Calibration and estimation of mass over charge ratio, retention time and drift time with pseudo aggregate ions

To compensate for small differences in acquisition, all 27 samples were calibrated with a quick alignment. To perform this quick alignment the peak picked ions of all samples were merged into a single list and only those ions with an intensity larger than 2^{14} were used. Of all these ions more than 10 thousand pseudo aggregate ion could be formed as they had a unique m/z which was found exactly once in each sample (Supplementary Figure 7). A thousand of these pseudo aggregate ions had either RT or DT outliers and were removed, and the remaining pseudo aggregate ions were equally partitioned in a group for calibration and a group for estimation. With the pseudo aggregate ions for calibration both the m/z and DT of all peak-picked ions were corrected by at most 10 parts per million (ppm) and 10000 ppm respectively, depending on their sample origins. For the RT calibration the pseudo aggregate ions were grouped in more than 500 groups so that each group had a distinct RT and the RT of each ion of each pseudo aggregate ion in the group was strictly smaller than the RT of each ion in the next pseudo aggregate ion group. With these pseudo aggregate ions groups a piece-wise linear transformation was performed to calibrate the ions of each sample, resulting in a more consistent retention time (Supplementary Figure 8).

After this calibration, the other half of the pseudo aggregate ions were used to estimate the maximum inter-run distance for each of RT, m/z, DT, resulting in respectively 5 ppm, 6000 ppm and 0.2 minutes (Supplementary Figures 9, 10 and 11). These estimates are relatively small, indicating that the calibration performed well or that the acquisition was very robust. While the latter is subjective to evaluate without comparison, the estimates with uncalibrated coordinates are more than twice as large (Supplementary Figures 12, 13 and 14), indicating that at least the former statement is true.

2.1.2 Inter-run ion alignment

With the estimation parameters obtained from the pseudo aggregate ions, more than 100 million pairs of ion neighbors could be defined. However, the definition of neighboring implies a transitive relation over a path of multiple neighboring pairs. With this transitivity, there are clusters with over a 100 ions in them, sometimes with a sample being represented 10 times. To avoid such ambiguous

situations, the pairs of ion neighbors were trimmed so that each remaining cluster, hereafter defined as an aggregate ion, has at most one ion per sample. With this trimming 10 million pairs of ion neighbors were removed. This relatively low percentage of trimming implies the estimation parameters are stringent enough to obtain a good specificity. Most of the resulting aggregate ions were either fully reproducible or random noise (Supplementary Figures 15 and 16). As expected, the fully reproducible aggregate ions have the highest average intensities, while the noisy aggregate ions are less intense (Supplementary Figures 17 and 17).

2.1.3 Intensity calibration and validation

Intensity was not used to create the aggregate ions and was therefore not calibrated with the pseudo aggregate ions. This calibration is done after the full alignment to include as much information as possible. On average, this calibration reduced the intensity CV of anchors by a factor 3 (Supplementary Figures 19, 20, 21, 22, 23 and 24), resulting in e.g. more than 80% of all fully reproducible anchors with a CV below 20 for each of condition A, B and QC.

With this calibrated intensity, the logFC values between condition A and B could be determined. As expected, these logFC values indeed show three groups of -2, 0 and 1, corresponding to the original mix ratios of the samples (Supplementary Figure 25).

2.1.4 Robustness to noise

To assess HistoPyA's ability to deal with noise, different peak picking thresholds in Waters' Apex3D software were tested. For nearly all partially reproducible aggregate ions, additional samples could be found at lower ion count thresholds. Fully reproducible aggregate ions were not hindered by the potential extra interference (Supplementary figure 26). However, halving the ion count threshold in Apex3D results in doubling the amount of noisy ions in HistoPyA. Nearly all ions are reproducible signal at an ion count threshold of 100.

2.2 Aggregate ion network

With the 5 million aggregate ions obtained at a peak picking ion count threshold of 1, a quick assessment of potential isotopic pairs was made. From the 1 million fully reproducible aggregate HE ions, 100 thousand pairs of pseudo isotopic pairs were obtained. Based on these isotopes, the maximum difference in RT and DT peaks was estimated to be smaller than 0.05 minutes and 7000

ppm respectively for all of the samples (Supplementary Figures 27 and 28).

With these maximum difference per sample the potential *consistent* co-elution between all HE aggregate ions was determined. Hereby, a total 400 million pairs of aggregate ions could be defined as *consistently* co-eluting, implying a probable same precursor in LE. While there are some aggregate ions with more than a 100 neighbors, most aggregate ions have between 1 and 13 neighbors showing a high specificity (Supplementary Figure 29). Of note, 15% of all anchor ions do not have a single neighbor and could be considered noise, even though they are reproducible over multiple runs.

Based on the logFC values between condition A and B many aggregate ions can be classified by most likely organism origin as human, yeast or ecoli. Since neighboring aggregate ions are expected to come from the same precursor, they are also expected to have the same organism origin. Moreover, neighboring aggregate ions with a different logFC are most likely false positives. While neighboring aggregate ions with a small sample overlap have many false positives, this false positive rate decreases when aggregate ions overlap in more samples. Especially fully reproducible aggregate ions that are neighbors show a low false positive rate of less than 1% (Figure 2).

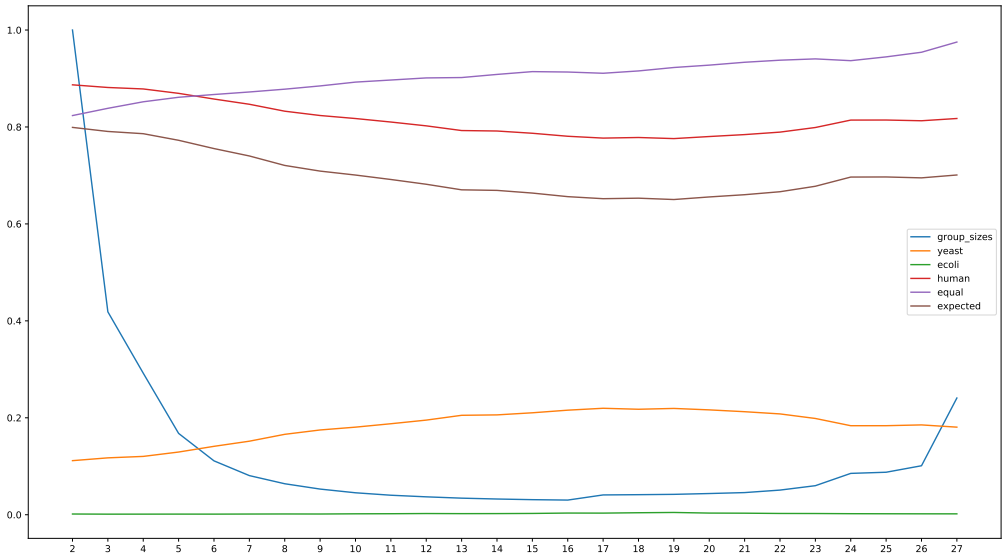

Figure 2: Neighbor organism origins.

Add a chimeric network (@m/z) that declusters by reprodicibility. When HE aggregate ions are deisotoped based on their neighbors, the delta mass frequencies are most prevalent for amino acid delta masses (Supplementary methods TODO). When only LE aggregate ions are considered and deisotoped, the singly charged aggregate ions have drift time that confirm the charge state as being 1.

Delta masses are amino acids, ergo the neighbors belong together.

To determine consistent co-elution between LE and HE, only consistency in RT was considered. This was essential because there is a DT shift between LE and HE scans of roughly 3 (Supplementary Figure 31)

TODO

2.3 Fragment-centric Annotation

Based on the ion network, a ion count (noiseless) was calculated for each individual ion. For each possible b- or y-ion explanation of a single fragment, the count of neighboring ions from the same peptide was determined. The frequency of all unique counts was determined and log transformed, and the e-value for the most extreme count was determined similar as to XTandem. No intensity was used (ion network). As opposed to spectrum-centric, i.e. precursor-centric. We do not use presurors, but fragments create similar specificity... (PNM as opposed to PIMs) This leaves us with PTMs and SNPs, i.e. open mass search. Which we will address through m/z estimation with dt.

2.3.1 Model validation

Count distributions (NOVEL) follow good linear distribution with outlier (XTandem).

Fragment-centric annotation of DDA data gives similar results as Mascot annotation. Precusorless?

2.3.2 Results

96000 PNMs on 95000 fragments were significant out of 10e06. 1+ is annotated. But reproducibility filtering: 100000 on 1e6 (plot). More DIA peptides found than DDA. All highest scoring PNMs are retained per fragment ion (identical number of ion counts matched peptide)

Dechimerization (4).

We annotate 10% (40.000 are annotated)

Peptide annotation goes through percolator, i.e. best scoring PNM = peptide score Overview of fragment features (used for percolator). Percolator increased significance by many percent.

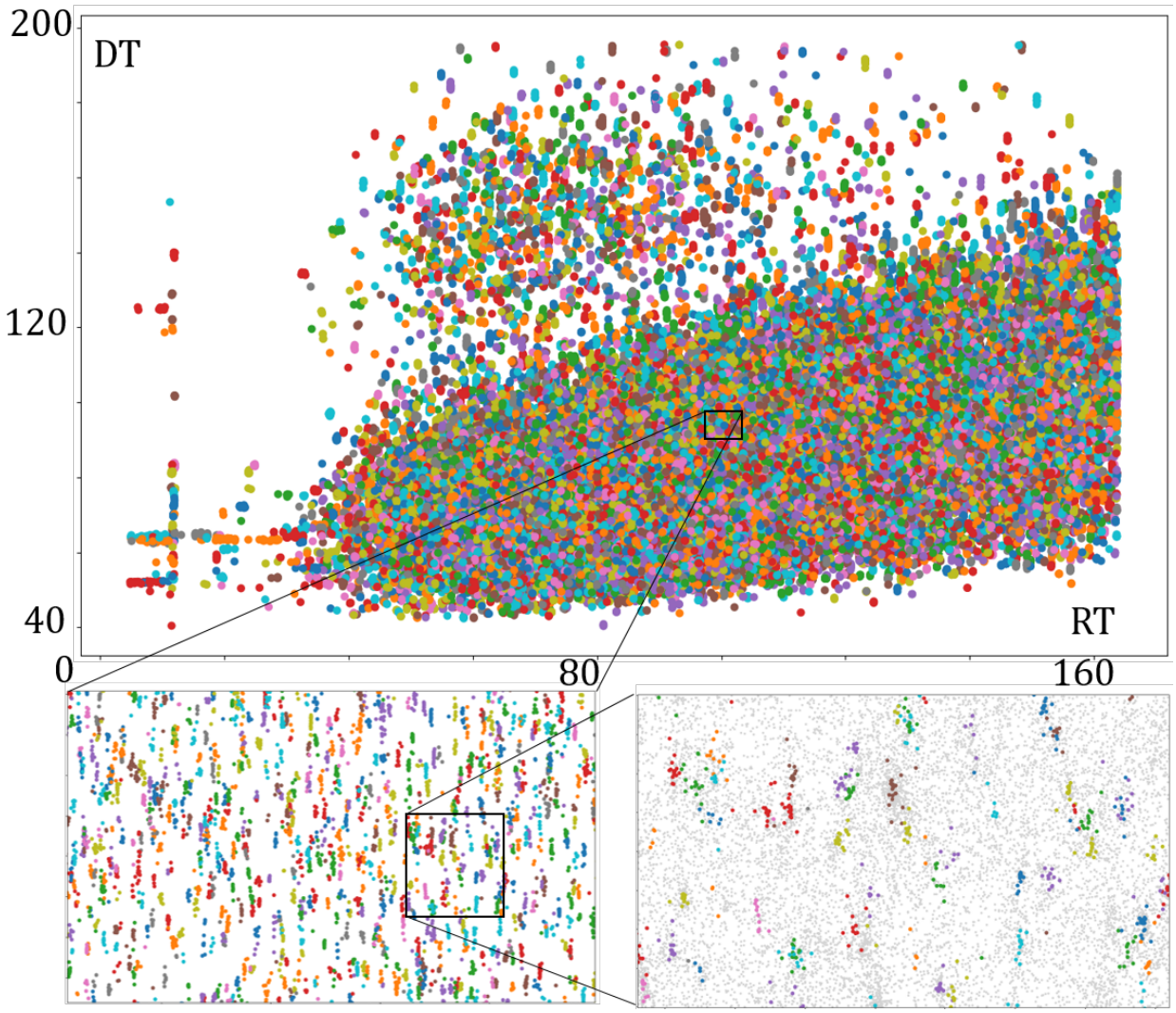

Figure 3: Annoteted fragments.

Precursor existence filter very useful, but not essential. Alternatively You percolate your precursor mass out of drift (allowing for SWIM)

Unannotated stuff can be annotated thanks to noiselessness, i.e. fully reproducibles (400.000)

250.000 have an annotated neighbor = unannotated ions: isotopes: 20.000 => 15% Non-b-y-ions neutral losses non-significant PNMs

MEANING 150.000 belong to unannotated clusters = unannotated peptides (ISD, PTMs, missed cleavage,...):

AND THERE ARE 41 loners

Vendor specs of >12000 peptides for human reached.

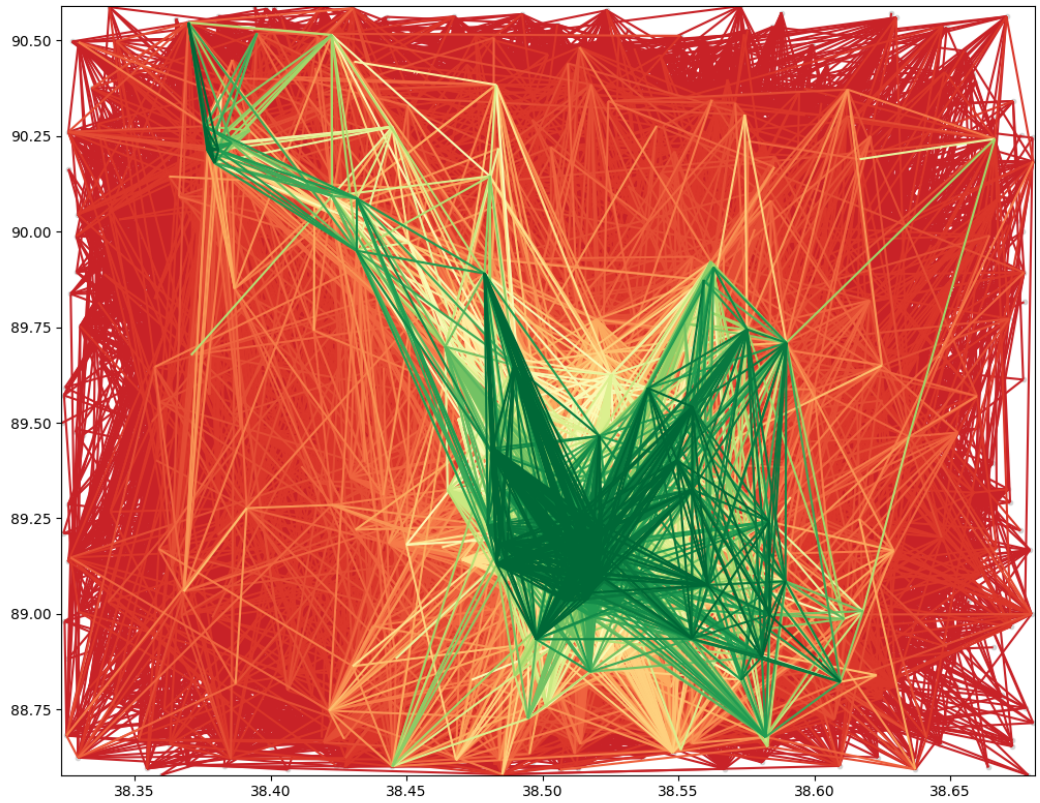

Figure 4: Dechemerization.

2.3.3 Annotation validation

Annotated (significant and by proxy) fragment/precursor/peptide organisms are correct with LFQ organism classifaction.

RTs coincide with mascot DDA rts.

Decoy-decoy (search against pyrococcus) gives no hits.

2.4 Fragment-centric Quantification

You have the experiment in a single network: performance Specificity with reduced interference? Removing stochisticity give ultra clean data Intensity is for Quant, not for any other purpose

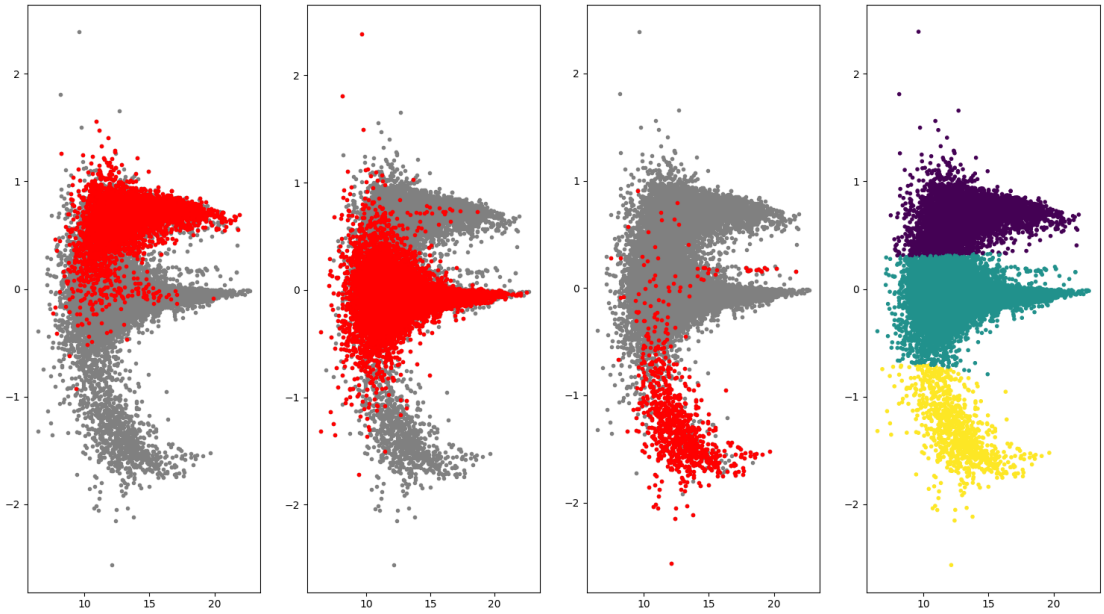

Figure 5: Experimental annotation and theoretical organism classification

2.5 Fragment-centric Application

Public data? LOPIT Kathryn Lilley

2.6 SWIM

Precursor existence filter very useful, but not essential. Alternatively You percolate your precursor mass out of drift (allowing for SWIM)

3 Discussion

Creating pseudo MSMS spectra for traditional annotation tools (without precursor)?

HistoPyA is modular, meaning peak picking, calibration, ion alignment, aggregate ion network creation, annotation van all be easily replaced.

Cosmology vs quantum mechanics

Noiseless data simplifies many annotation algorithms, including de novo for DIA.

In theory, many replicates (with high overlap requirements) would partition the aggregate ion network in fragment groups all belonging to the same precursor (noise-free pseudo MSMS spectra).

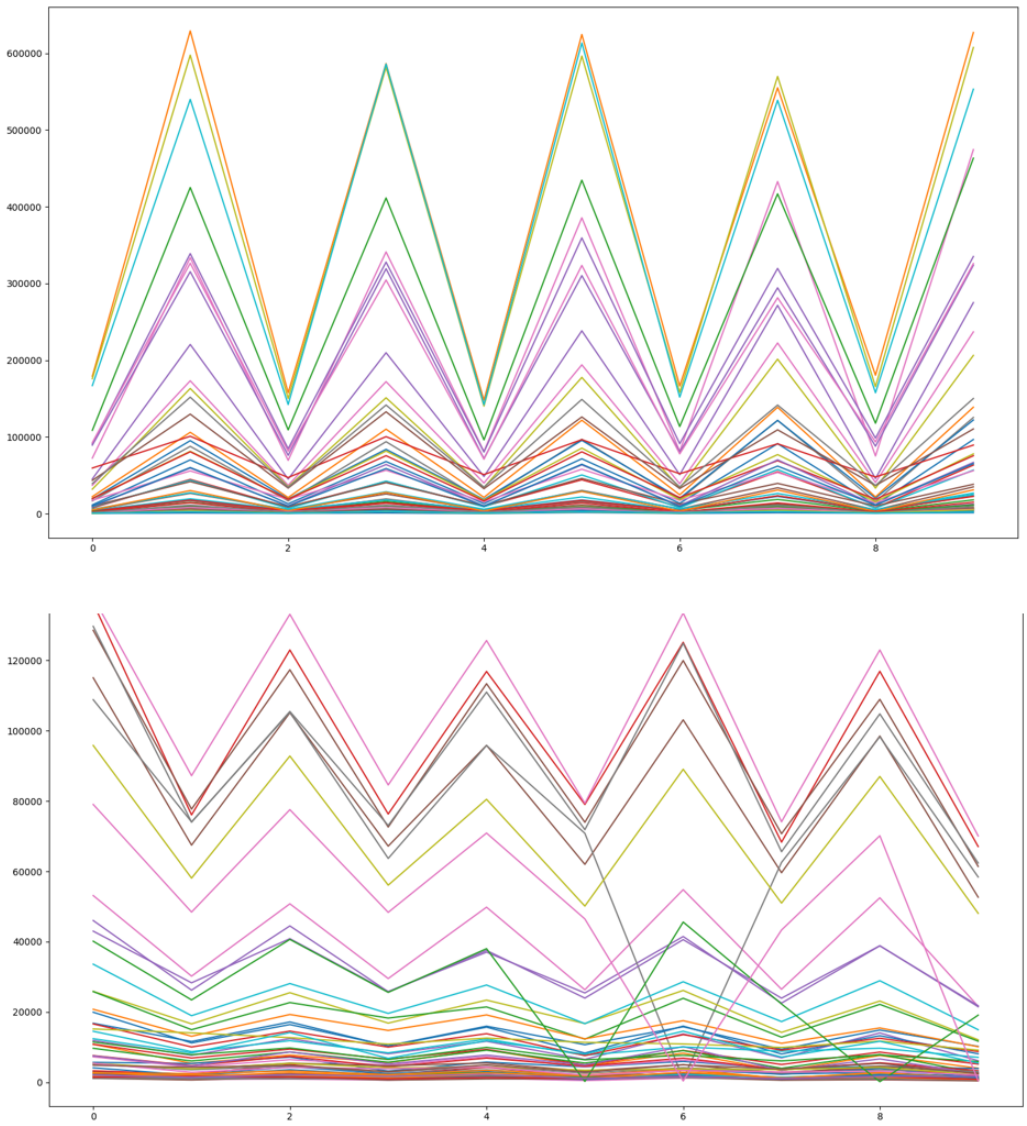

Figure 6: fragment quantifications per peptide

Results are better than DDA and more transparent (especially FDR) than other DIA approaches.

HE only could allow better/simpler/cheaper instrumentation.

Annotations match MS2PIP theoreticals? We start from aggregate spectra, as is done in other tools as well

Data and software availability

All data is available at the ProteomeXchange consortium with identifier TODO. This includes raw data and data peakpicked with Waters' commercial Apex3D software. Complete algorithmic results as presented in this manuscript, including parameters, logs, figures, and other in/output files are deposited alongside this data.

The complete source code for version TODO of single window ion mobility (SWIM)-DIA is available at GitHub TODO. In-house scripts performing label-free quantification (LFQ) validation and recreating figures are included in an additional sub folder in the GitHub repository, but require original result files to be downloaded from ProteomeXchange. A minor test case illustrating how to use the software on novel samples provided by the user is included in the GitHub repository.

QC files monitoring general MS performance are available at the Panorama website with identifier TODO.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This research was funded by Research Foundation Flanders (FWO) research project grant G013916N, FWO mandate 12E9716N (MD), FWO mandate 3F016517 (BVP), and Flanders Innovation & Entrepreneurship (VLAIO) mandate SB-141209 (LDC).

The authors would like to express their gratitude to the Waters informatics team (including, but not limited to; Hans Vissers, Scott Geromanos and Steve Cievarini) and Lennart Martens (Ghent University (UGhent)) for their critical feedback throughout the project. Samples were acquired at the ProGenTomics facility. Computational assistance was provided by Yannick Gansemans (UGhent) and Laurentijn Tilleman (UGhent).

Author contributions

SW and MD conceived the ideas of creating noiseless ion-networks with replicates and annotating this fragment-centric. SD and BVP performed all sample preparation and data acquisition. SW performed all computational analysis. SW and MD wrote the draft manuscript. MD and DD supervised the project. All authors provided critical feedback on the manuscript and approved the final version.

Acronyms

m/z mass over charge ratio

CCS collisional cross section

CPU central processing unit

csv comma separated values

CV coefficient of variation

DDA data-dependent acquisition

DIA data-independent acquisition

DT drift time

DTT dithiothreitol

ESI electrospray ionization

FDR false discovery rate

FWO Research Foundation Flanders

GB gigabytes TODO bits?

HDMS^e high definition MS^e

HE high energy

IMS ion mobility separation

LC liquid chromatography

LE low energy

LFQ label-free quantification

logFC logarithmic fold change

MRM multiple reaction monitoring

MS mass spectrometry

PIM peptide-ion match

ppm parts per million

PSM peptide-spectrum match

ptp point-to-point

QC quality control

RAM random-access memory

RANSAC random sample consensus

RT retention time

SWATH sequential window acquisition of all theoretical mass spectra

SWIM single window ion mobility

TOF time of flight

UGhent Ghent University

VLAIO Flanders Innovation & Entrepreneurship

w/w weight for weight

XIC extracted ion chromatogram

References

[1] Limonier F, Willems S, Waeterloos G, Sneyers M, Dhaenens M, Deforce D. Estimating the reliability of low-abundant signals and limited replicate measurements through MS2 peak area in SWATH. PROTEOMICS;0.ja:1800186.

[2] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20.9:1466–1467. eprint: `/oup/backfile/content_public/journal/bioinformatics/20/9/10.1093/bioinformatics/bth092/2/bth092.pdf`.

[3] The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. Journal of The American Society for Mass Spectrometry 2016;27.11:1719–1727.

[4] Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Röst HL, Tate SA, Tsou C-C, Reiter L, Distler U, Rosenberger G, Perez-Riverol Y, Nesvizhskii AI, Aebersold R, Tenzer S. A multicenter study benchmarks software tools for label-free proteome quantification. Nature Biotechnology 2016;34:1130.

1 Material and methods

1.1 Sample preparation

Lyophilized whole cell protein extracts of yeast and human were acquired from Promega and lyophilized whole cell protein extract of ecoli was acquired from Waters. All extracts were already reduced with dithiothreitol (DTT), alkylated with iodoacetamide and digested with Trypsin/Lys-C Mix by their respective manufacturers. These extracts were reconstituted in 0.1% formic acid and two master samples were created as in Navarro et al. [4], each in triplicate: A) a mixture of 65% weight for weight (w/w) human, 15% w/w yeast and 20% w/w ecoli and B) a mixture of 65% w/w human, 30% w/w yeast and 5% w/w ecoli. The resulting samples have logarithmic fold changes (logFCs) of 0, 1 and -2 for respectively human, yeast and ecoli. One third of each of the six master batches was mixed as a quality control (QC), resulting in ratios of 65% w/w human, 22.5% w/w yeast and 12.5% w/w ecoli.

1.2 Data acquisition

For each of the six master samples three technical replicates injections were acquired to obtain nine samples in total for both condition A and B. Nine technical replicate injections of the QC were also acquired. All 27 samples were acquired in a randomized design in three different acquisition mass spectrometry (MS) modes on three different mass spectrometers: 1) high definition MS^e ($HDMS^e$) mode on a Synapt G2-Si (Waters), 2) data-dependent acquisition (DDA) mode on a Q-Exactive (Thermo), and 3) in SWATH mode on a TripleTOF 5600 (AB Sciex) (Supplementary Figure 30). For each of the $3 \cdot 9 \cdot 3 = 81$ samples, five μg was injected. All data was acquired in res mode. The acquisition on the Synapt G2-Si was preceded by a nano-acquity (Waters) set up in microflow liquid chromatography (LC), the acquistion on the Q-Exactive was preceded by a TODO micro LC, an the acquistion on TripleTOF was preceded by an Eksigent micro LC. All samples were acquired on a 150 minute gradient. After each three samples, an acoli autoQC sample was run to assess the performance of the mass spectrometers. For the sequential window acquisition of all theoretical mass spectra (SWATH) acquisition, TODO windows of TODO mass over charge ratio (m/z) were used.

1.3 Peak picking of raw data

Raw data from all samples were peak-picked to obtain one comma separated values (csv) file per sample in which all its ions, noth low energy (LE) and high energy (HE), and intensities were defined by their m/z apex, retention time (RT) apex, and drift time (DT) apex. In case of DDA or SWATH, the DT apex is replaced by the m/z the precursor selection.

Waters' HDMS^e data was peak-picked with their Apex3D software, version 3.1.0.9.5 on a Windows 10 Workstation with 160 gigabytes TODO bits? (GB) random-access memory (RAM) and 16 central processing units (CPUs). Selected parameters were a lockMass of 785.8426 for charge 2 with m/z tolerance of 0.25, apexTrackSNRThreshold of 1, and write to Apex3D csv file instead of default Apex2D csv file. Different counts thresholds of 1, 5, 10, 20, 50, and 100 were used for both LE and HE to test the influence of noise on HistoPyA.

All resulting csv files were imported simultaneously in a Python environment to obtain a single list containing all ions from all samples.

1.4 Sample calibration and estimation of ion inter-run differences

To calibrate the m/z, RT and optionally DT of each sample, all LE ions with an intensity larger than 2^{14} were selected and ordered by their m/z, regardless of sample origin. Between each consecutive pair of ions, their m/z parts per million (ppm) error was calculated. Whenever a set of consecutive ions, in which each sample was represented by exactly one ion, had smaller m/z ppm errors than the left and right flanking m/z ppm errors, it was defined as a pseudo aggregate ion.

For each pseudo aggregate ion the point-to-point (ptp) distance in RT and optionally DT dimension of their representative ions was calculated. Based on the distribution of the median absolute deviation of all RT or DT ptp errors, individual z-scores were calculated per pseudo aggregate ion. Each pseudo aggregate ion with a z-score exceeding 5 was considered an outlier and removed. This process of outlier removal was repeated until only pseudo aggregate ions with z-scores below 5 for both their RT and DT remained.

50% of the pseudo aggregate ions were selected for calibration of the m/z and DT between each sample. For each pseudo aggregate ion, the average RT, m/z, and DT was calculated. Per pseudo aggregate ion the median ppm error of m/z and DT of all representative ions compared to the pseudo aggregate ions average was calculated. These median sample ppm errors were subtracted from the original m/z and DT of each ion present in the complete ion list. As a result, the median

error between all pseudo aggregate ions and the representative ions of each sample was zero.

The same 50% of pseudo aggregate ions were partitioned in groups to calibrate the RT between samples. Two pseudo aggregate ions a and b belong to the same group if there exists a sample α in which $RT_{a,\alpha} < RT_{b,\alpha}$ and a sample β in which $RT_{a,\beta} > RT_{b,\beta}$. Thus, two pseudo aggregate ions c and d from two different groups always have representative ions so that for each sample γ the statement $RT_{c,\gamma} < RT_{d,\gamma}$ is true. Per sample the average RT of each pseudo aggregate ion group are taken as y-values, while the average RT of all representative ions of each pseudo aggregate ion group are taken as x-values. Per sample, these x and y-values are then used to perform a piece-wise linear transformation on the RT of all the ions in the complete ion list.

The remaining 50% of the pseudo aggregate ions were used to obtain an unbiased estimate of the inter-run errors of the calibrated m/z, RT and DT errors. Per pseudo aggregate ion the ptp distance (largest minus smallest of the representative ions) of the calibrated m/z, RT and DT were calculated. The 99^{th} percentile of each characteristic is now defined as the maximum allowed inter-run error between two ions from different samples.

1.5 Ion inter-run alignment and noise definition

A network was created wherein each ion was a vertex. Between two ions an edge was set if and only if the ions originated from different samples, were both acquired in either LE or HE, and had calibrated m/z, RT and DT errors smaller than the maximum estimated inter-run errors.

Subsequently this network was trimmed, so that no path existed between two ions from the same sample. This trimming was done iteratively on paths of increasing length. Whenever a path of the specified length existed between two vertices from the same sample, all edges of the path were removed. For each remaining connected components it was checked whether all ions originated from different samples. If this was true, no further trimming happened on this connected component, otherwise all edges which are not part of an edge-triangle are removed and the specified path length was increased by one for the next trimming iteration.

The resulting network now consists of multiple connected components, in which each ion originates from a different sample. Note that there may be connected components in which not all vertices are connected, meaning that either some calibrated m/z, RT or DT exceed their respective maximum allowed errors, or their connection got trimmed. The maximum allowed errors were determined on the 99^{th} percentile of pseudo aggregate ions, which in turn were defined with ions with intensity above 2^{14}, meaning their apices were likely to be peak-picked more accurately than ions with lower

intensity. As such, these maximum allowed errors can be considered quite stringent and some missing edges should be expected. Finally, each connected component was defined as an aggregate ion. For all of these aggregate ions, their average calibrated m/z, calibrated RT and calibrated DT was calculated. Each aggregate ion also has a weight that is defined by the number of samples where it was detected. This property is proportional to the probability that this ion is a true signal. Finally, all aggregate ions with only a single ions are considered noise and removed for subsequent analyses.

To normalize intensity difference between samples, the average intensity of all aggregate ions expressed in all samples was calculated, as well as the logFC distance of each individual sample to this average. For each sample, the median of these logFC distances was determined and subsequently subtracted from all ions in the complete ion list. Finally, the logFC of the average calibrated intensity from ions in condition A compared to the average calibrated intensity from ions in condition B was calculated per aggregate ion, or set to $-\infty, \text{null}, +\infty$ when no average could be calculated for condition A and/or B.

1.6 Estimation of intra-run differences between high energy aggregate ions of the same precursor

To estimate maximum RT and DT intra-run differences between aggregate ions derived from the same precursor (e.g. fragments), HE isotopic aggregate ion pairs with ion respresentatives in all samples are used. Two aggregate ions are defined as an isotopic pair if and only if their difference in aggregate calibrated m/z is $1.002861 \pm x$ ppm (averagine isotope) with x the maximum inter-run m/z error. Furthermore, the difference in original RT and DT per sample should be smaller than the inter-run maximum error for each sample, assuming intra-run errors are smaller than inter-run errors. Finally, this pair should be unique, meaning no other potential isotopic pair can be formed with either of the aggregate ions. For this estimation, this generally implies only the mono-isotopic and first isotope can be detected and that the second isotope is not present as an aggregate ion expressed in all samples, or that a charge other than 1 was accidentally used.

Two ions from the same sample are now defined as co-eluting if and only if their distance in RT and DT is smaller than the 99^{th} percentile of the isotopic aggregate ion pair distribution per sample.

A special situation arises when determining co-elution between LE and HE scans for e.g. fragments and precursors, as there is a drift shift between those channels. To correct this drift shift, unfragmented pairs of fully reproducible LE and HE aggregate ions with equal m/z, within intra-

run ppm error, are determined in a similar way as isotopic pairs where original RT per sample should be smaller than the inter-run maximum error for each sample. As with isotopic pairs, each unfragmented pair should be unique. Hereafter, the relative drift shift, i.e. difference in drift time divided by LE drift in ppm, per sample between LE and HE ions is determined and only those within the 10^{th} and 90^{th} percentile are retained. Furthermore ions with DT below 50 or greater than 190 are removed to avoid boundary issues. Optimal parameters a, b, c and d are then determined such that for all the retained ions the error between the relative drift shift y and the function $y = a \cdot \|dt, mz\| + b \cdot \arctan(dt/mz) + c \cdot dt/mz + d$ has an optimal least squares fit. Finally, this function is applied to all ions of all aggregate ions per sample.

TODO PPM difference calibration between HE-LE

1.7 Aggregate ion network generation

A network was created in which all aggregate ions were vertices. An edge is set between two aggregate ions if and only if they consistently co-elute. Two aggregate ions are defined as *consistently co-eluting* if and only if they co-elute for each overlapping sample. However, as the intra-run differences within each sample are independent, a large sample count can introduce a dimensionality curse, meaning it is unlikely that representative ions co-elute in each sample even if they originate from the same precursor. Therefore the definition of *consistently co-eluting* is weakened to mean that they should have a probability of at least 0.999 to overlap in at least x out of y samples. Herein the probability is calculated by binomials, i.e. $\sum_{x \geq i}^{y} \binom{y}{i} \cdot 0.99^{y-i} \cdot 0.01^{i} > 0.999$. As a final constraint, two aggregate ions should co-elute in at least two samples to be considered *consistently co-eluting*.

1.8 Fragment-centric annotation of an aggregate ion network

At this point, the complete experiment has been collapsed into a single (noiseless) aggregate ion network. Here, we used the aggregate ions for the final analysis, but this can easily be split into a separate ion network for each sample. A fasta file containing all SwissProt entries from human, yeast and ecoli was downloaded. The crap database was appended to this fasta, as well as a decoy with all reversed protein sequences. A standard in silico tryptic digest with one miscleavage and default amino acids masses, with the exception of a cysteine to which a carbamido mass of 57.021464 was added, was made to obtain a list of peptides and their masses. Duplicate peptides from different proteins were merged to obtain a list of unique peptide sequences. Peptides originating solely from

decoy proteins were classified as decoy peptides, while all others were classified as targets. For each peptide, the masses of all b- and y-ions was calculated.

For each HE aggregate ion, all potential singly, doubly and triply charged b- or y-ion explanations were determined within the inter-run ppm error. Moreover, each of these explanations belong to a peptide, so every aggregate ion has a list of peptides from where it could have originated.

For each aggregate ion with at least three peptide explanations and at least two edges in the aggregate ion network a hyperscore was determined in an X-Tandem! like fashion for all its potential peptide explanations. For each of the peptide explanations it was counted how often it occurred in the peptide explanations of the neighboring aggregate ions. Hereafter, the cumulative log frequency of all but the highest of these counts was determined and used for a robust random sample consensus (RANSAC) regression. A hyperscore equal to minus the regressed prediction of the highest count was then determined for all peptides with this highest corresponding count. Note that some aggregate ions have no peptides with a hyperscore, for instance when no regression can be made.

For each aggregate ion with at least one peptide with a hyperscore, it is checked whether there is an LE aggregate ion that consistently co-elutes.

All aggregate ions and their remaining peptide explanations, meaning with a hyperscore and co-eluting precursor, are now considered as a peptide-ion match (PIM) and given to percolator where they are treated as if they were peptide-spectrum matchs (PSMs). Percolator features are set to RT, fragment delta mass (ppm), precursor delta mass (ppm), neighbor count, peptide count, hyperscore, precursor charge and fragment ion type with e.g. b7 as 7 and y4 as -4. Percolator was run with default parameters with the addition of post processing tdc (Y-flag) to correct for an imbalance in targets and decoys, and all predicted features (D-flag set to 15). Finally, all PIMs with a q-value below 0.01 are retained.

For each of the aggregate ions belonging to a PIM with q-value below 0.01, an exhaustive annotation is done for all its neighbors, which can thus be annotated as singly, doubly or triply charged precursor, b-NH3, b-H2O, c, a, a-NH3, a-H2O, y, y-NH3, y-H2O or x fragment of a specific peptide.

1.9 AutoQC, data deposition and source code availability

TODO

References

[1] Limonier F, Willems S, Waeterloos G, Sneyers M, Dhaenens M, Deforce D. Estimating the reliability of low-abundant signals and limited replicate measurements through MS2 peak area in SWATH. PROTEOMICS;0.ja:1800186.

[2] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20.9:1466–1467. eprint: `/oup/backfile/content_public/journal/bioinformatics/20/9/10.1093/bioinformatics/bth092/2/bth092.pdf`.

[3] The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. Journal of The American Society for Mass Spectrometry 2016;27.11:1719–1727.

[4] Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Röst HL, Tate SA, Tsou C-C, Reiter L, Distler U, Rosenberger G, Perez-Riverol Y, Nesvizhskii AI, Aebersold R, Tenzer S. A multicenter study benchmarks software tools for label-free proteome quantification. Nature Biotechnology 2016;34:1130.

Supplementary

Peak picking

Table 1: Apex 3D peakpicking results on the 27 LFQ samples.

552 **Schematic overview of quick alignment to detect pseudo aggregate ions**

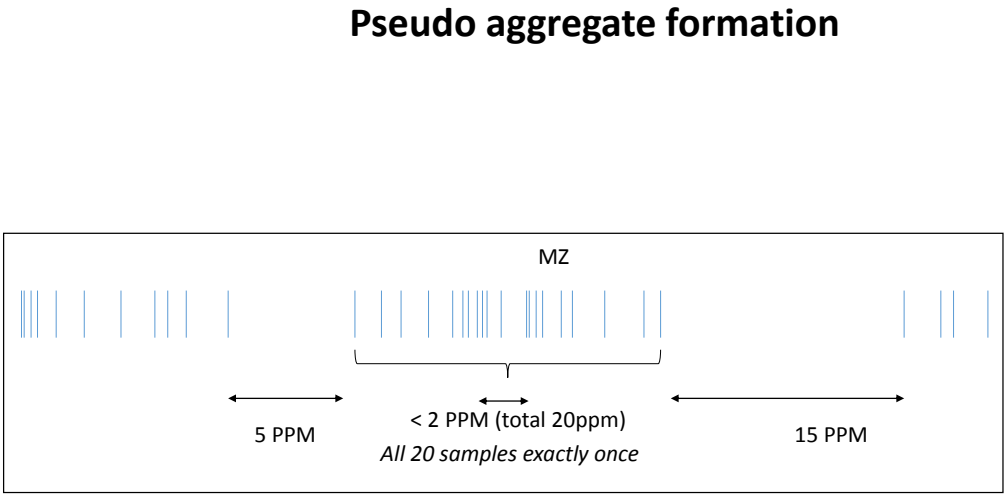

Figure 7: Schematic overview of quick alignment to detect pseudo aggregate ions.

Schematic overview of pseudo aggregate ion group piece-wise linear transformations for retention times

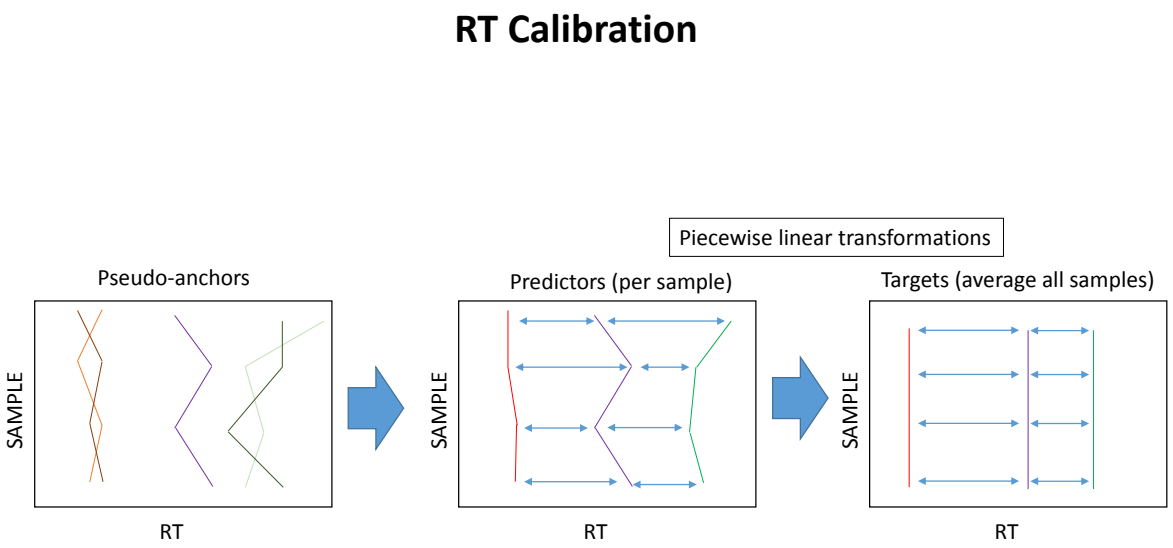

Figure 8: Schematic overview of pseudo aggregate ion group piece-wise linear transformations

Mass over charge ratio estimation

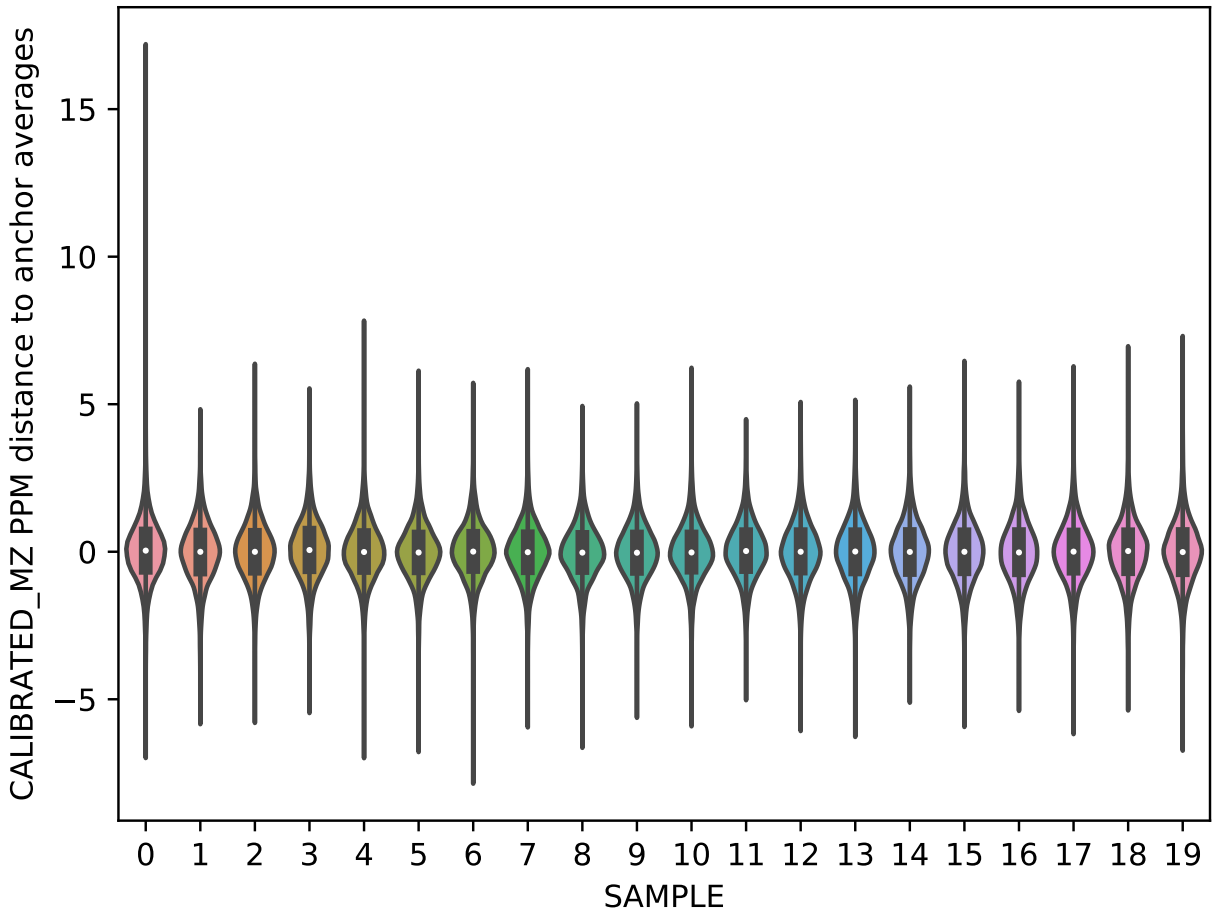

Figure 9: Estimation of m/z error per sample toward the pseudo aggregate ions.

Drift time estimation

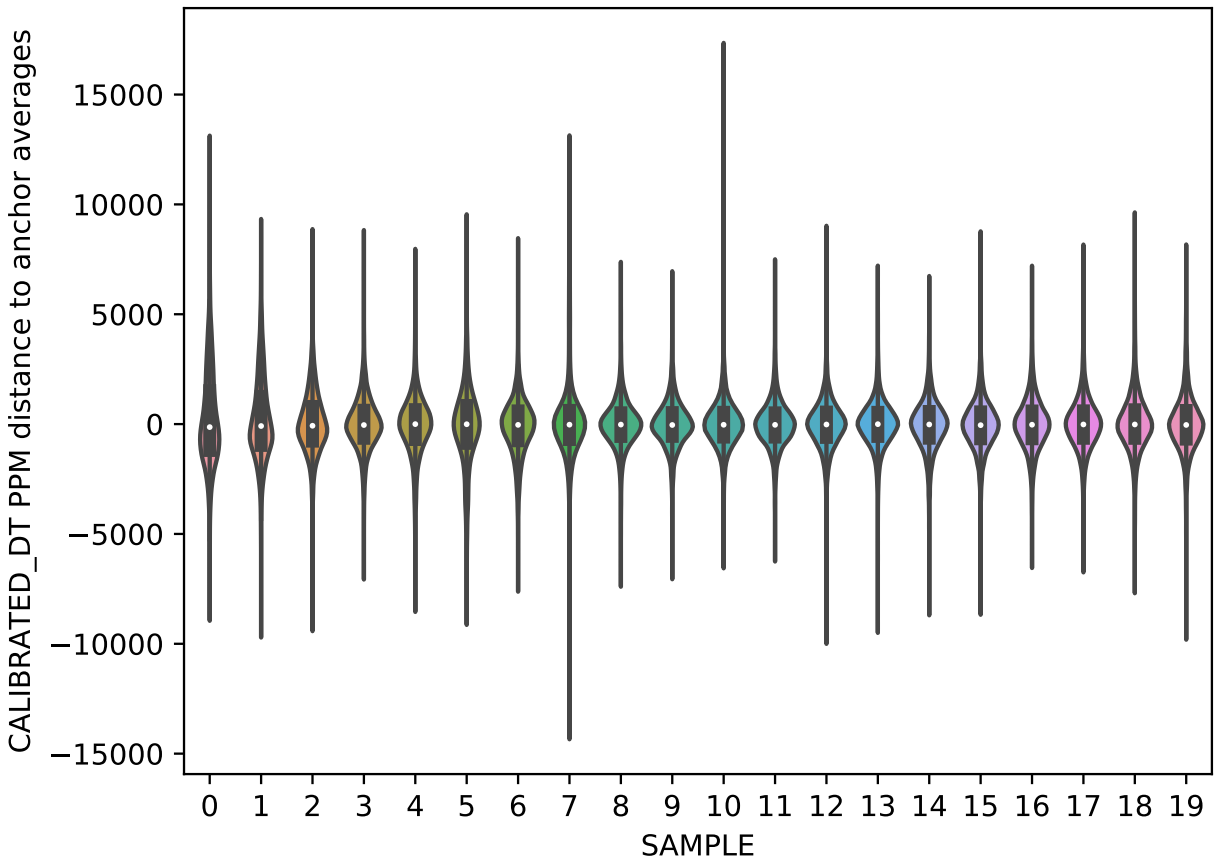

Figure 10: Estimation of DT error per sample toward the pseudo aggregate ions.

Retention time estimation

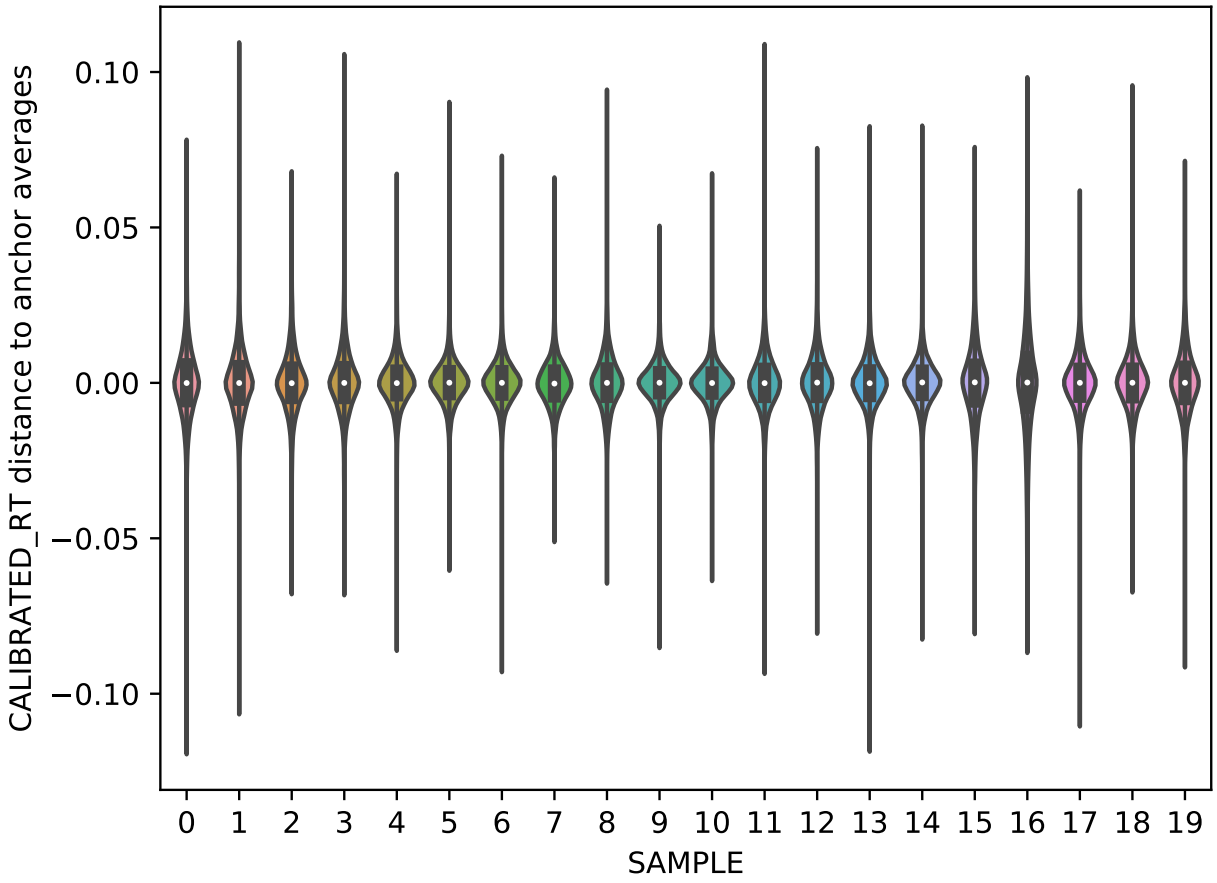

Figure 11: Estimation of RT error per sample toward the pseudo aggregate ions.

Uncalibrated mass over charge ratio estimation

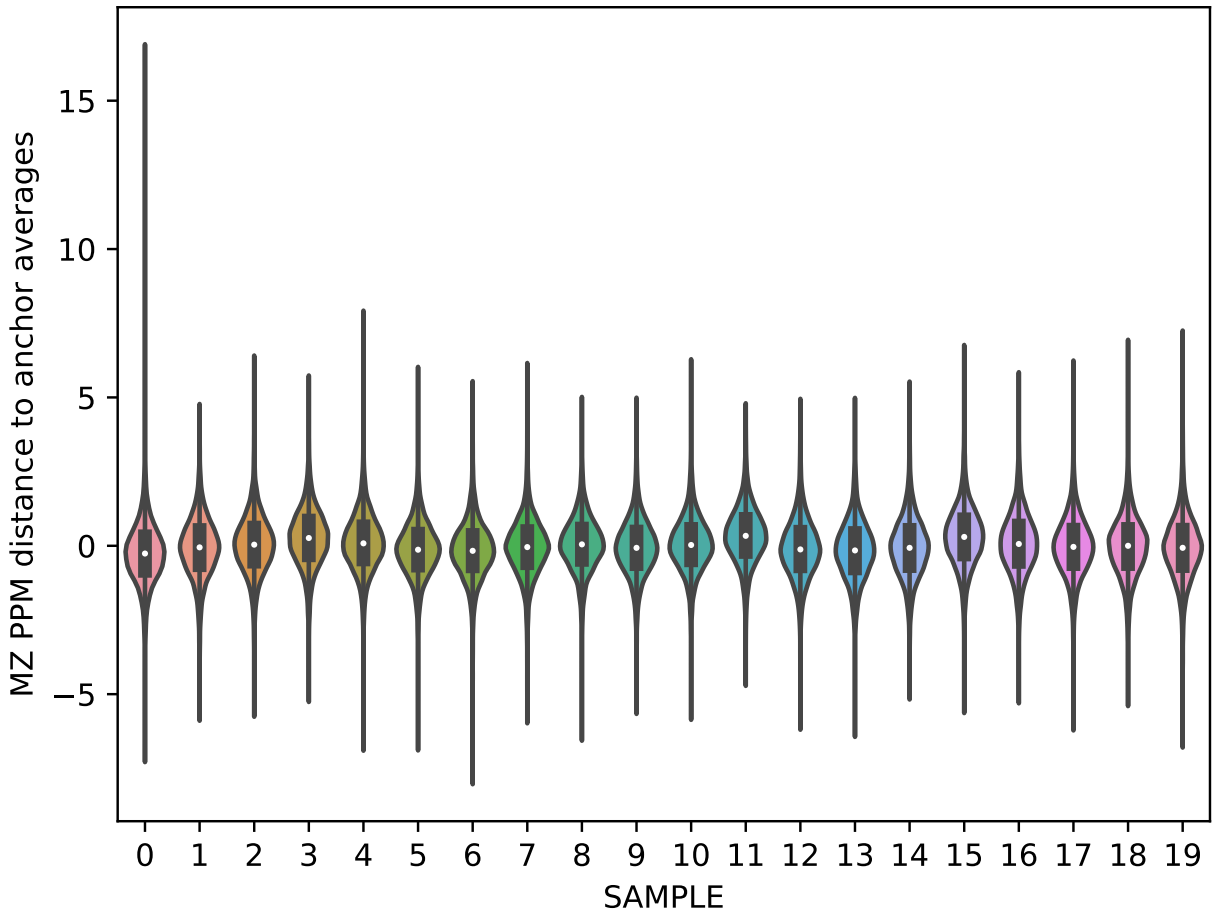

Figure 12: Uncalibrated m/z error per sample toward the pseudo aggregate ions.

Uncalibrated drift time estimation

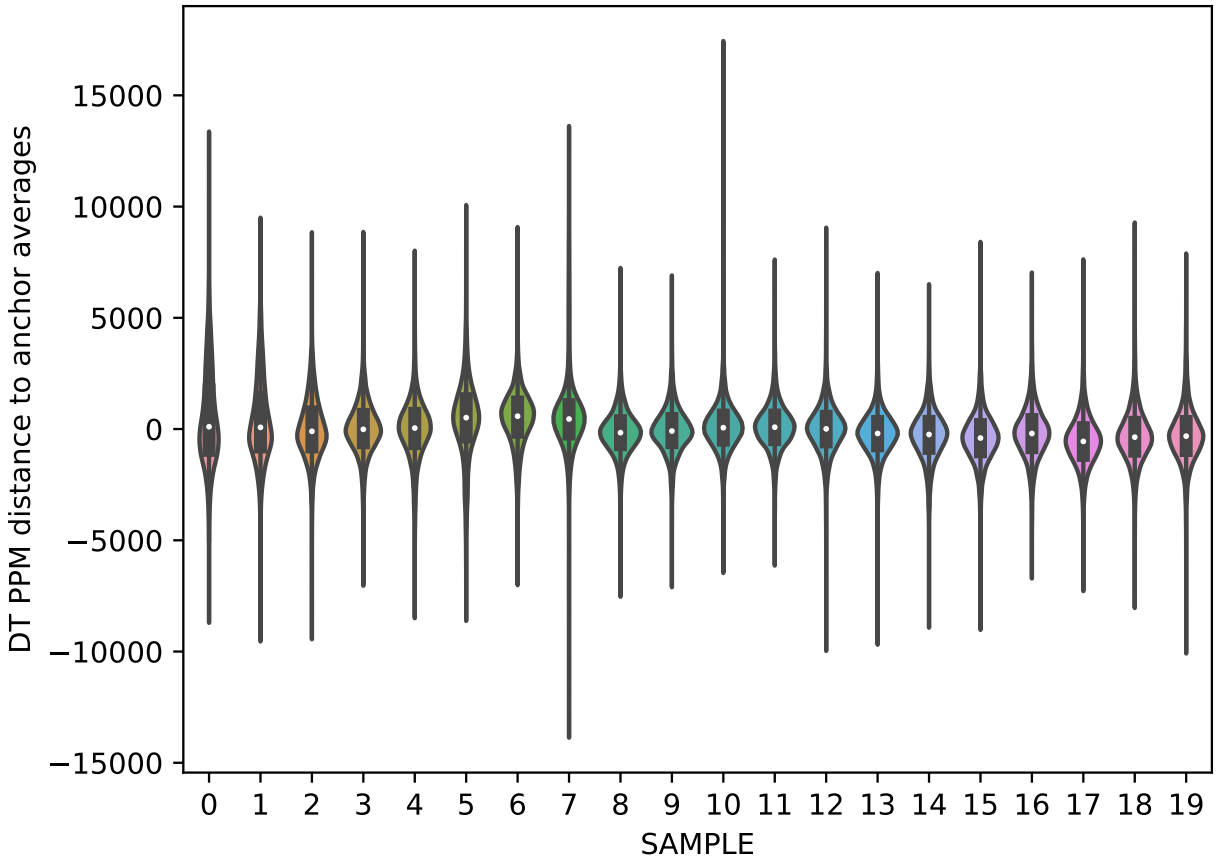

Figure 13: Uncalibrated DT error per sample toward the pseudo aggregate ions.

Uncalibrated retention time estimation

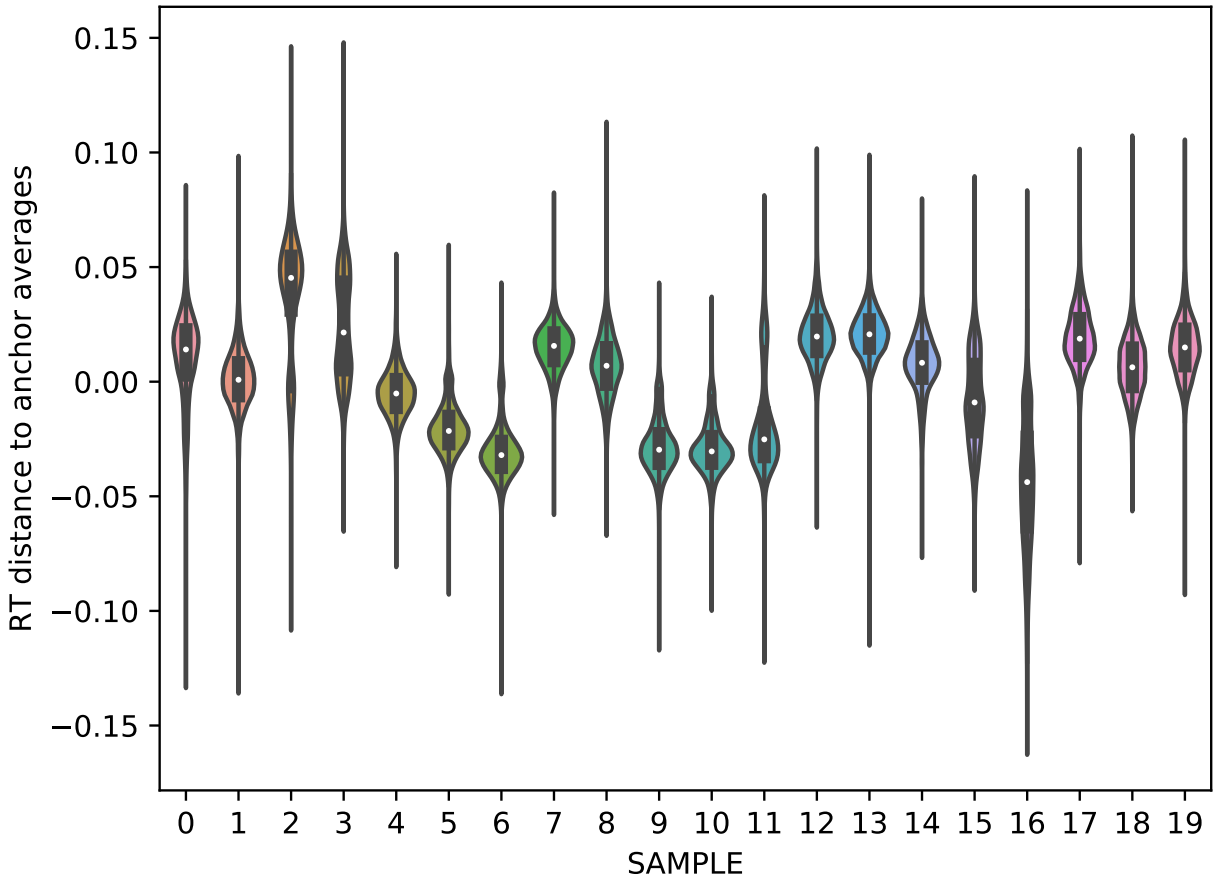

Figure 14: Uncalibrated RT error per sample toward the pseudo aggregate ions.

561 **Low energy aggregate ion sizes**

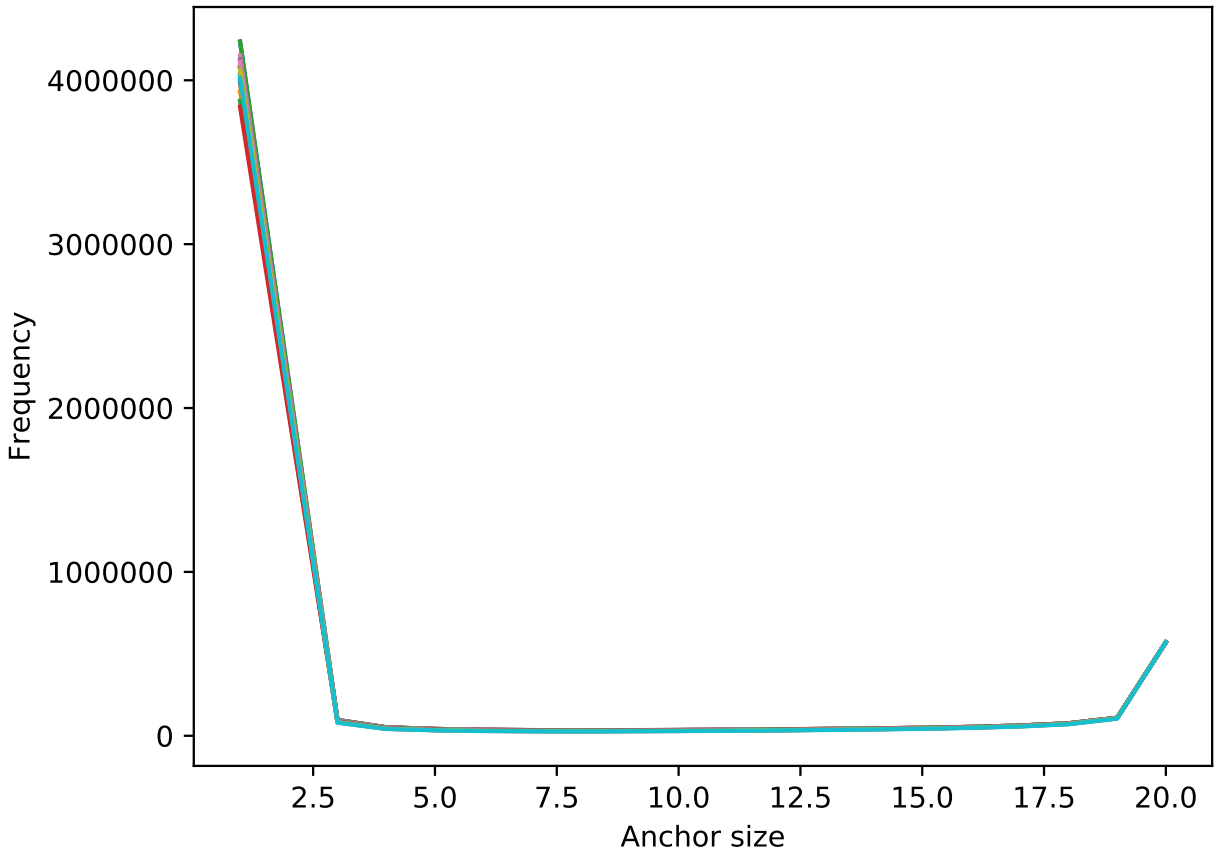

Figure 15: Low energy aggregate ion sizes.

High energy aggregate ion sizes

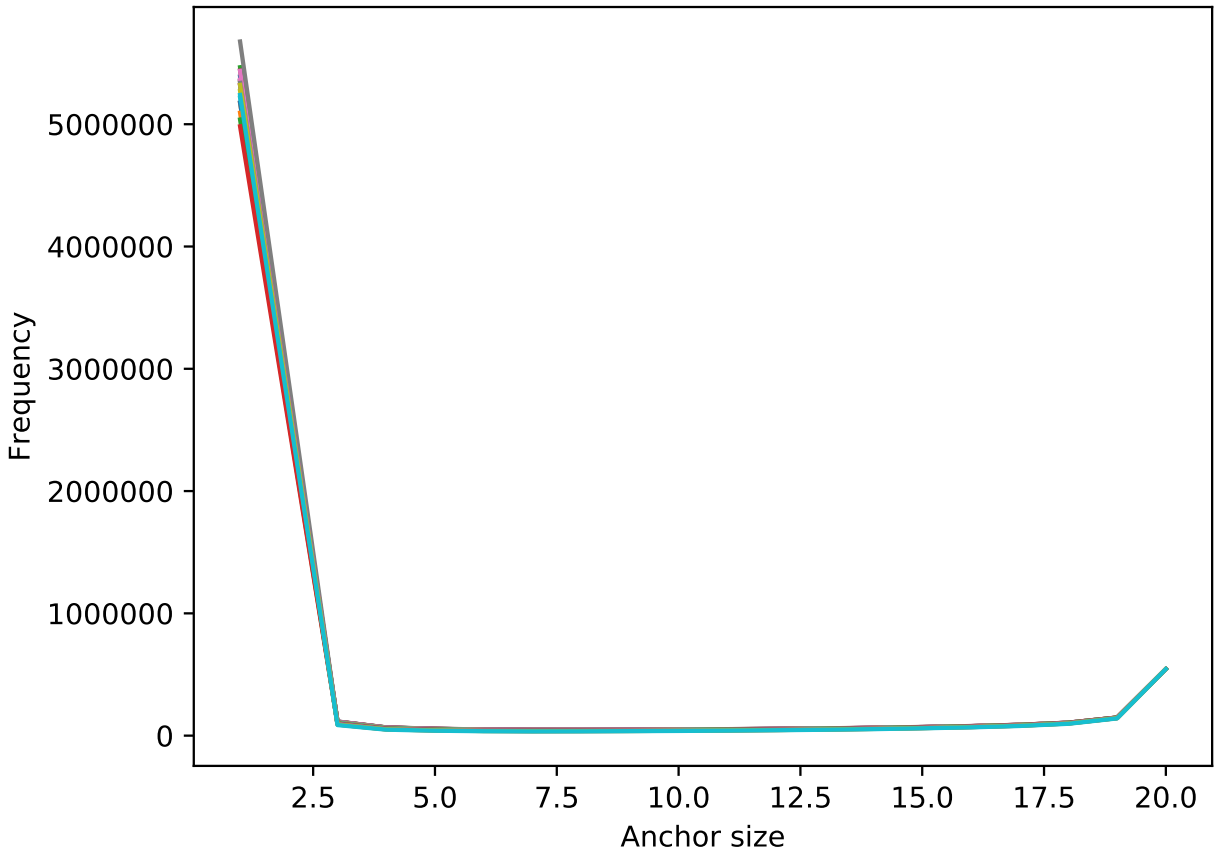

Figure 16: High energy aggregate ion sizes.

Low energy aggregate ion intensities

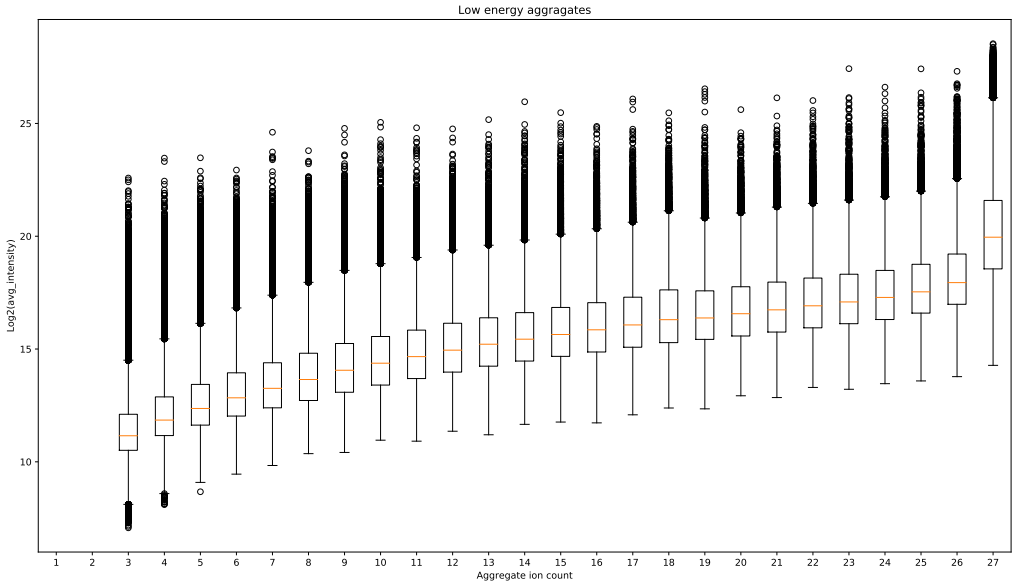

Figure 17: Low energy aggregate ion intensities.

High energy aggregate ion intensities

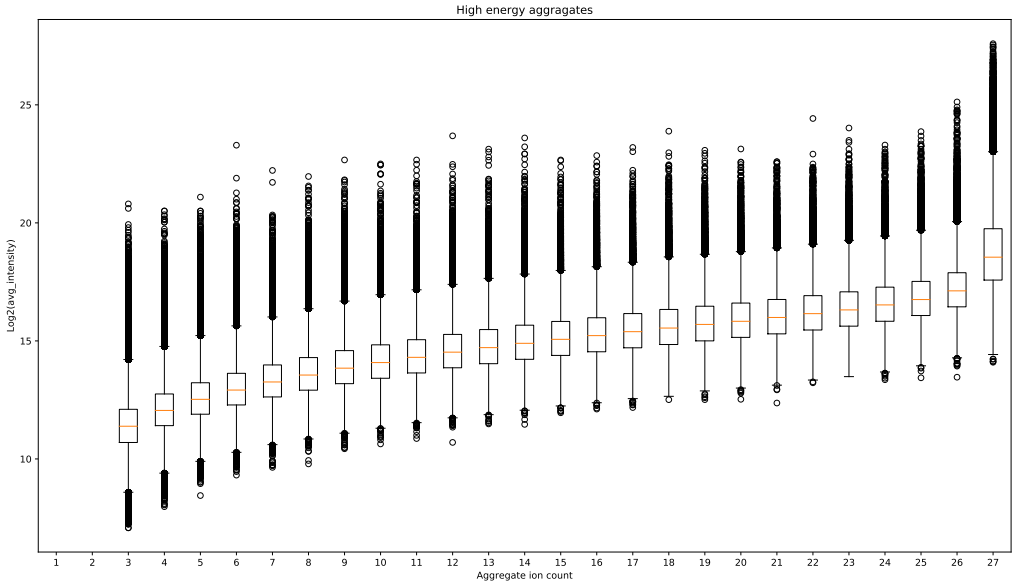

Figure 18: High energy aggregate ion intensities.

CV of uncalibrated intensities in condition A

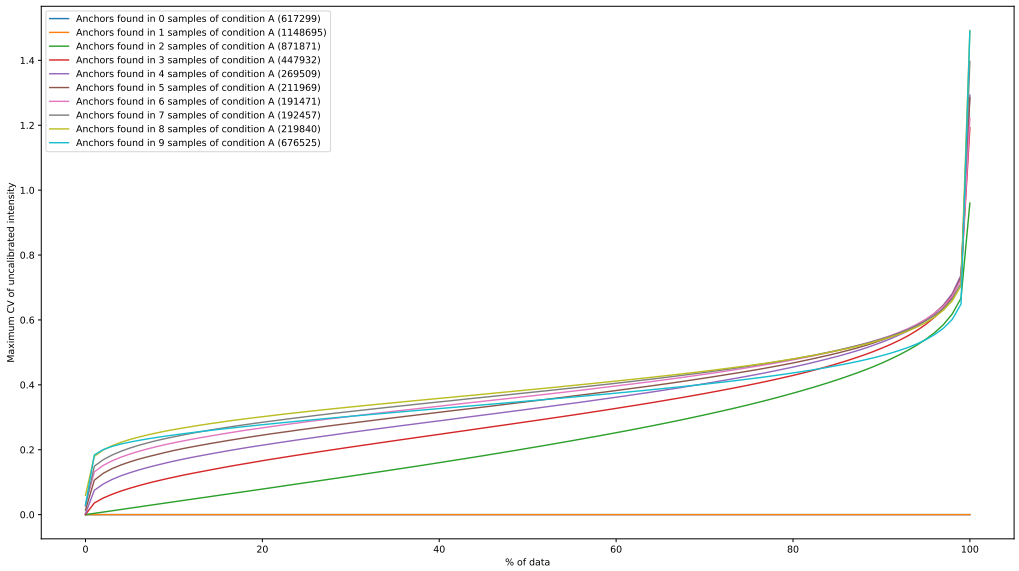

Figure 19: CV of uncalibrated intensities in condition A.

CV of uncalibrated intensities in condition B

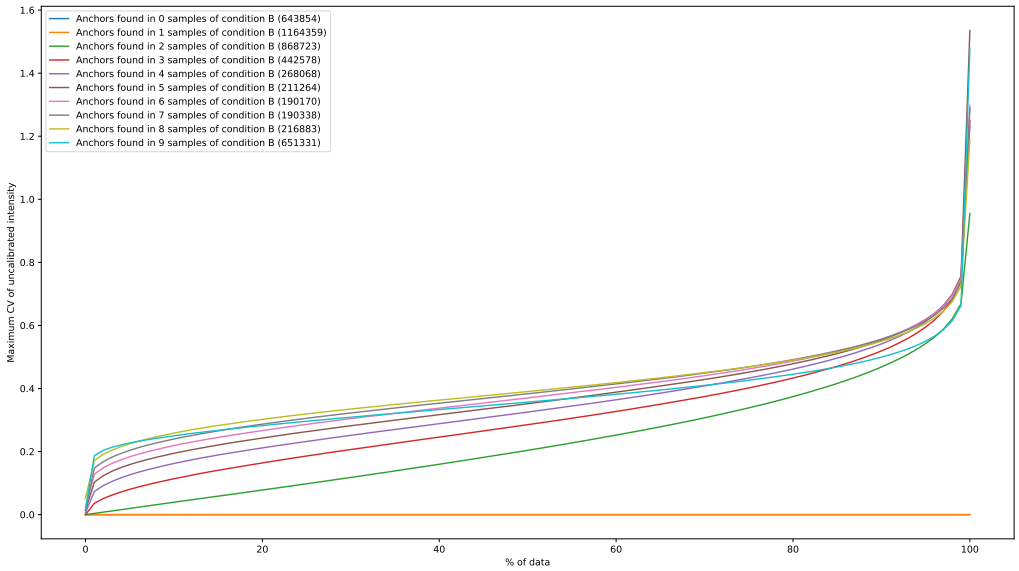

Figure 20: CV of uncalibrated intensities in condition B.

CV of uncalibrated intensities in condition QC

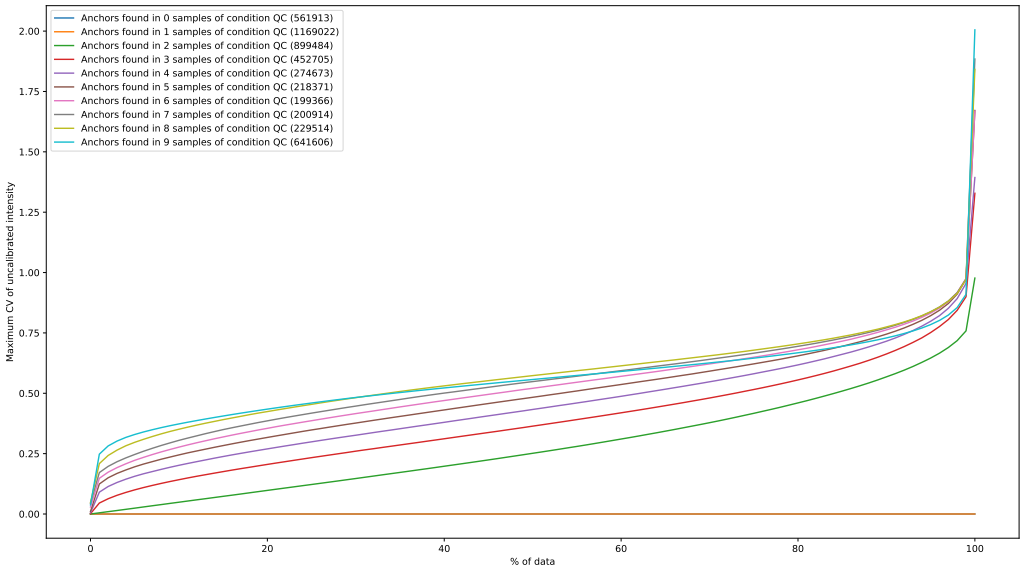

Figure 21: CV of uncalibrated intensities in condition QC.

CV of calibrated intensities in condition A

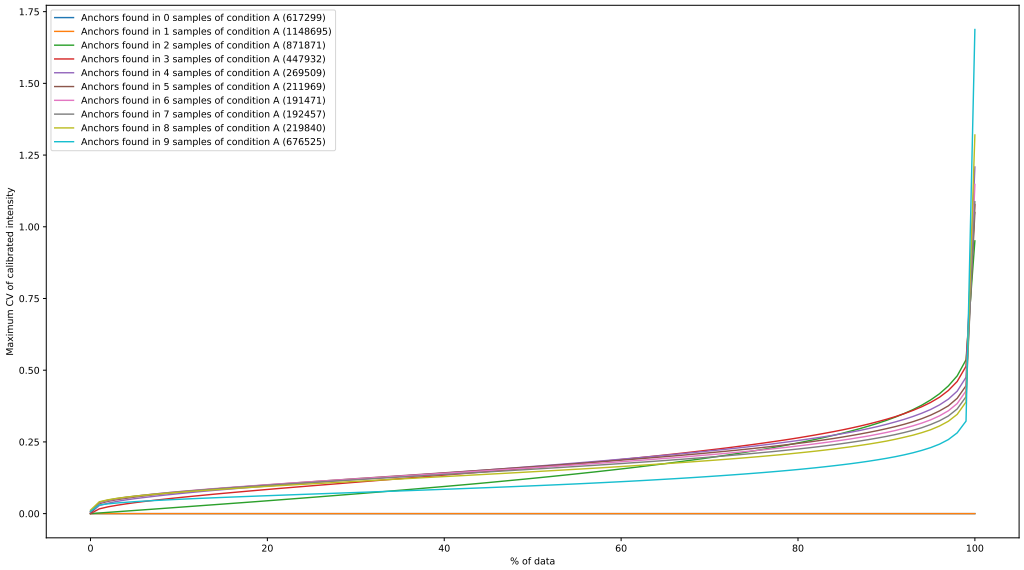

Figure 22: CV of calibrated intensities in condition A.

CV of calibrated intensities in condition B

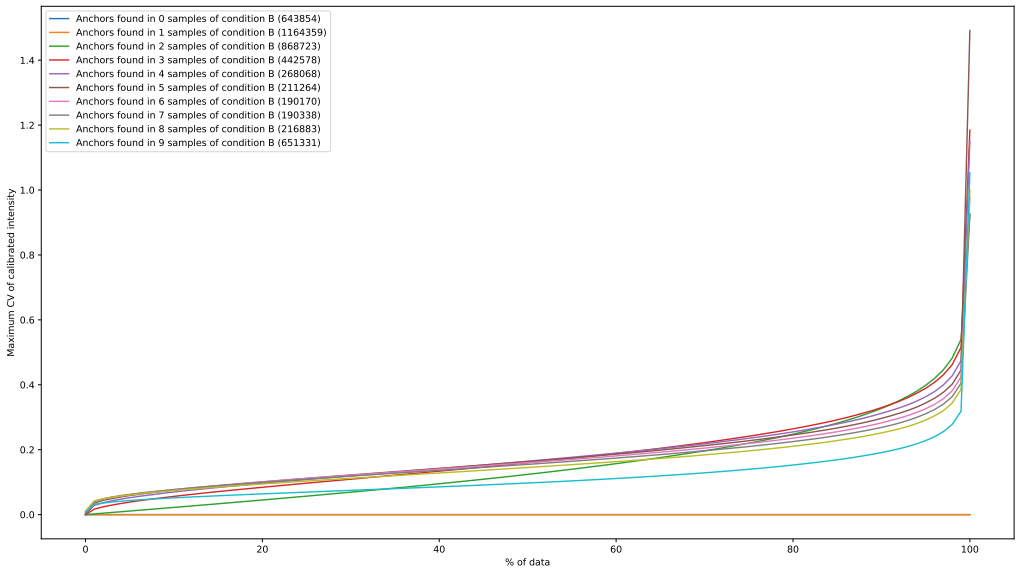

Figure 23: CV of calibrated intensities in condition B.

CV of calibrated intensities in condition QC

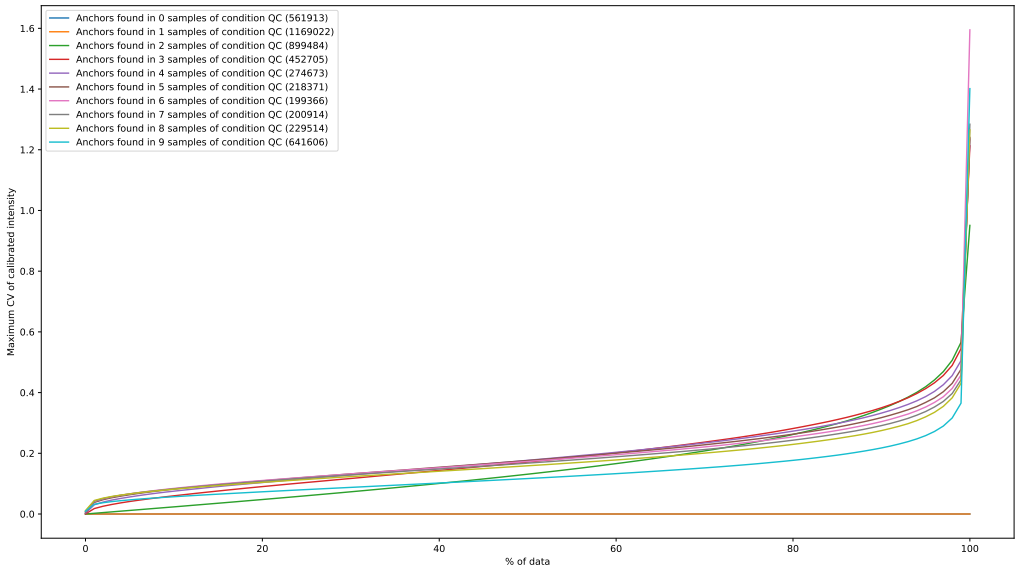

Figure 24: CV of calibrated intensities in condition QC.

Distribution of logarithmic fold changes

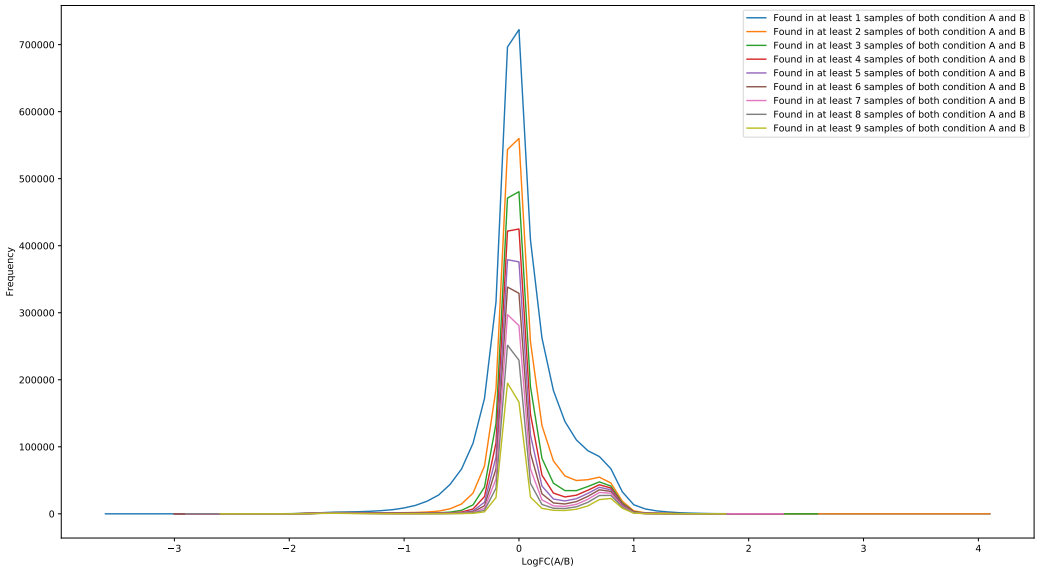

Figure 25: Distribution of logarithmic fold changes.

Noise robustness

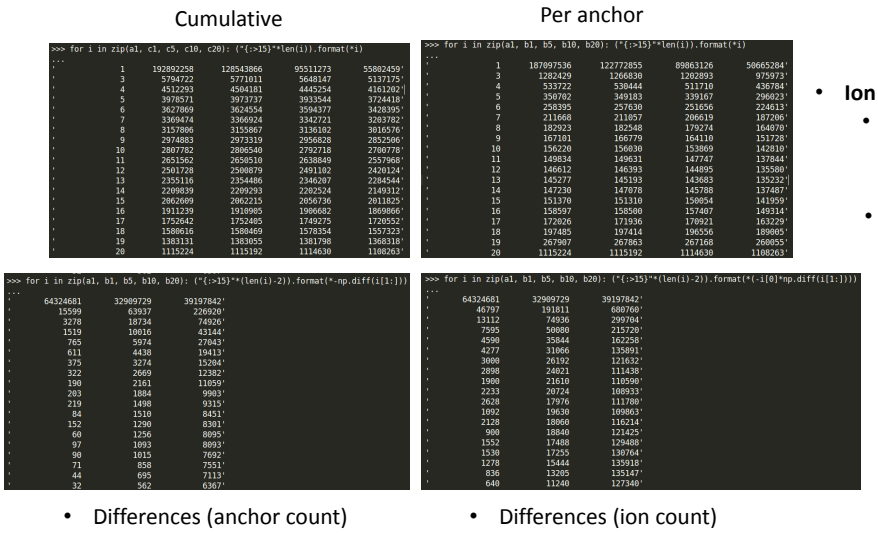

Figure 26: HistoPyA's robustness to peak picking noise.

Co-elution retention time distances

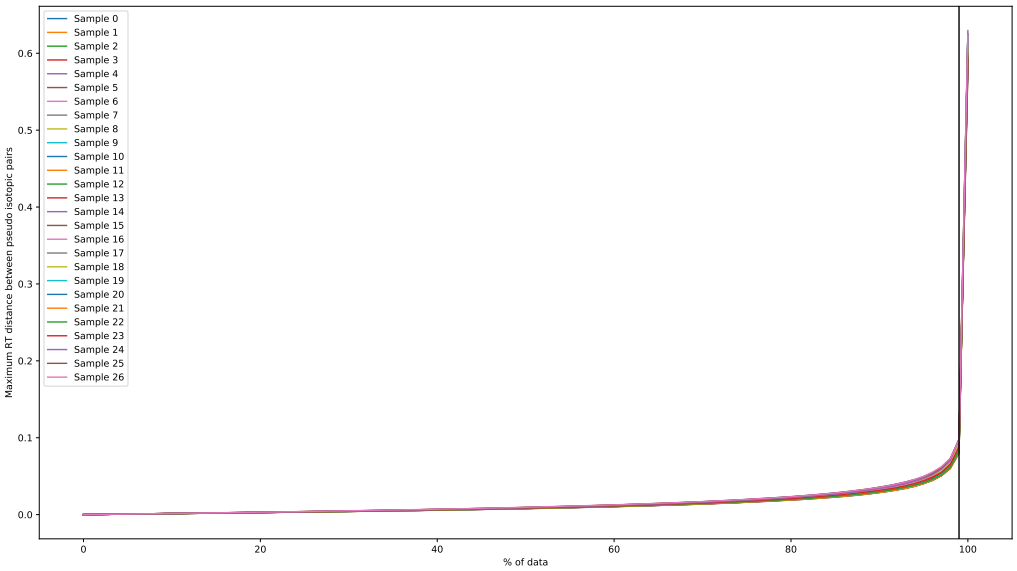

Figure 27: Distribution of intra-run retention time distances.

Co-elution drift time distances

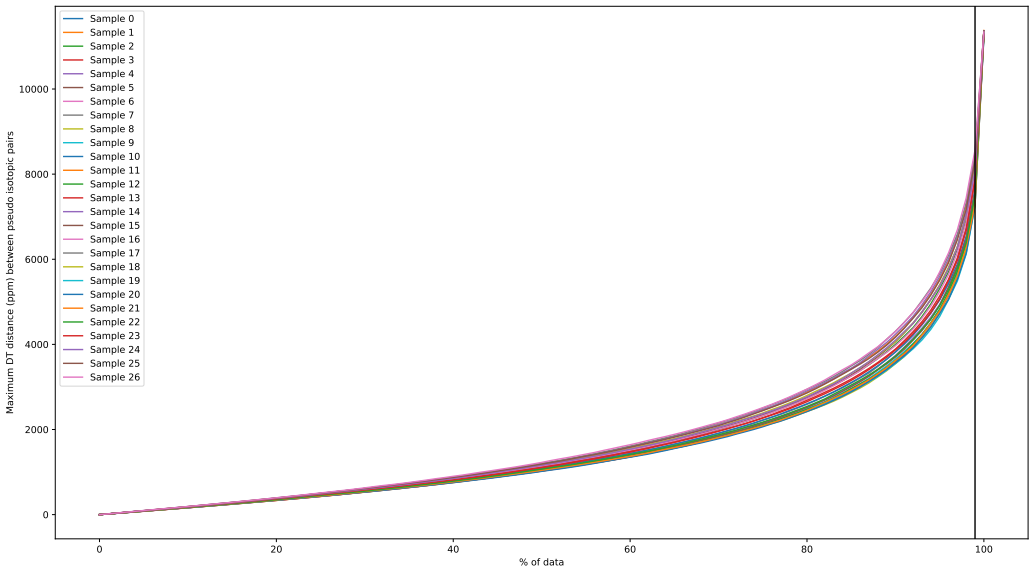

Figure 28: Distribution of intra-run drift time distances.

Neighbors distribution

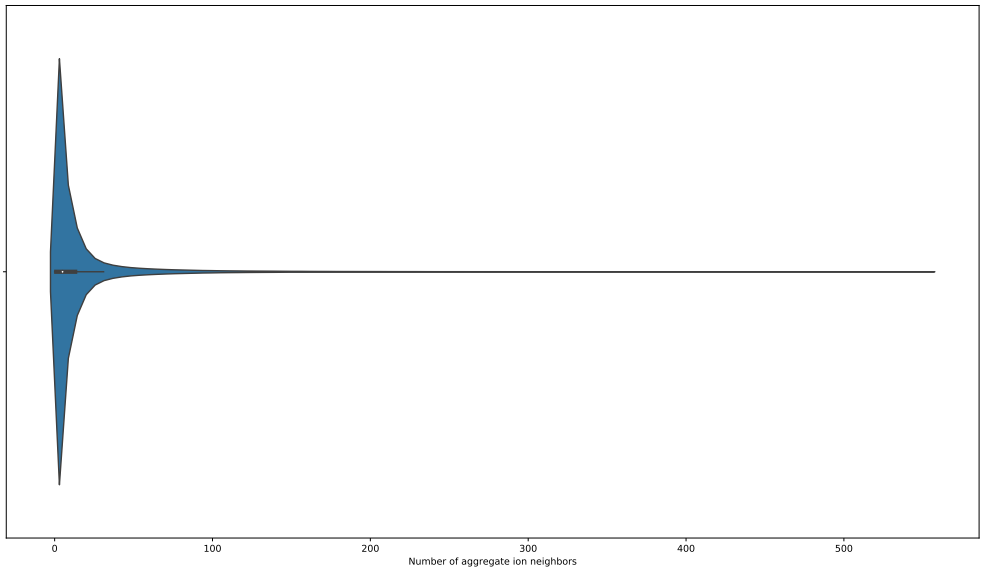

Figure 29: Distribution of aggregate ion neighbor counts.

Sample scheme

Figure 30: Sample scheme.

Drift time shift

Figure 31: Drift time shift between low energy and high energy channels.