Bioinformatics, 31(22), 2015, 3645–3652
doi: 10.1093/bioinformatics/btv434
Advance Access Publication Date: 27 July 2015
Original Paper

Systems biology

MetTailor: dynamic block summary and intensity normalization for robust analysis of mass spectrometry data in metabolomics

Gengbo Chen[1], Liang Cui[2], Guo Shou Teo[1], Choon Nam Ong[1,3], Chuen Seng Tan[1] and Hyungwon Choi[1]*

[1]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore, [2]Interdisciplinary Research Group in Infectious Diseases, Singapore-MIT Alliance for Research & Technology, Singapore, Singapore and [3]National University of Singapore Environment Research Institute, Singapore, Singapore

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 30, 2015; revised on June 13, 2015; accepted on July 21, 2015

Abstract

Motivation: Accurate cross-sample peak alignment and reliable intensity normalization is a critical step for robust quantitative analysis in untargetted metabolomics since tandem mass spectrometry (MS/MS) is rarely used for compound identification. Therefore shortcomings in the data processing steps can easily introduce false positives due to misalignments and erroneous normalization adjustments in large sample studies.

Results: In this work, we developed a software package MetTailor featuring two novel data preprocessing steps to remedy drawbacks in the existing processing tools. First, we propose a novel dynamic block summarization (DBS) method for correcting misalignments from peak alignment algorithms, which alleviates missing data problem due to misalignments. For the purpose of verifying correct realignments, we propose to use the cross-sample consistency in isotopic intensity ratios as a quality metric. Second, we developed a flexible intensity normalization procedure that adjusts normalizing factors against the temporal variations in total ion chromatogram (TIC) along the chromatographic retention time (RT). We first evaluated the DBS algorithm using a curated metabolomics dataset, illustrating that the algorithm identifies misaligned peaks and correctly realigns them with good sensitivity. We next demonstrated the DBS algorithm and the RT-based normalization procedure in a large-scale dataset featuring >100 sera samples in primary Dengue infection study. Although the initial alignment was successful for the majority of peaks, the DBS algorithm still corrected ∼7000 misaligned peaks in this data and many recovered peaks showed consistent isotopic patterns with the peaks they were realigned to. In addition, the RT-based normalization algorithm efficiently removed visible local variations in TIC along the RT, without sacrificing the sensitivity of detecting differentially expressed metabolites.

Availability and implementation: The R package MetTailor is freely available at the SourceForge website http://mettailor.sourceforge.net/.

Contact: hyung_won_choi@nuhs.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Mass spectrometry (MS) coupled with gas or liquid chromatography (GC-MS or LC-MS) is already the technology of choice in the metabolomics literature (Dettmer *et al.*, 2007; Griffiths and Wang, 2009). In the experiment, hundreds to thousands of compounds elute through the chromatography column at varying rates, resulting in separation of compounds across the retention time (RT). The compounds are then ionized and analysed by the MS, in which the mass to charge ratio (*m*/*z*) and the intensity of ions are determined. In the untargeted setting, compound identification is typically achieved by searching the mono isotopic mass of each peak against a large-scale database of compounds by exact mass values, such as the HMDB (Wishart *et al.*, 2007, 2009), MassBank (Horai *et al.*, 2010) and Metlin (Smith *et al.*, 2005), to name a few. Quantitative data analysis is then performed using the corresponding peak intensity or integrated peak area in each sample.

Despite conceptual similarity, extraction of peak features with precise compound identification has proved to be a challenging task in untargeted metabolomics. In particular, the lack of the identification step via global-scale MS/MS fragmentation makes cross-sample matching of peak feature data, so called alignment, as the most crucial data extraction step. This is because multiple compounds of a similar or identical mass with slightly different chemical structure (isomers) and different elemental composition, can co-elute and they are simultaneously analysed, creating mixed signals. Without the MS/MS evidence, the intensity signals for such molecules will be aligned together, being treated as the same compound. On the other hand, when thousands of features are processed across a large number of samples, it is also possible that the same compound can be misaligned across the samples and subsequently treated as different compounds in the statistical analysis stage. Hence accurate data extraction and robust preprocessing steps are prerequisite for successful untargeted metabolomics analysis.

The general workflow for MS data preprocessing consists of several steps. First, peak picking algorithms are applied to identify the ion chromatograms with robust isotopic patterns and each of them is reported as a peak feature with three-dimensional coordinates (*m*/*z* value, retention time and peak areas/intensities, often aggregated over isotopes and adducts into a major peak feature). Next, the peak alignment step removes the variations in both RT and *m*/*z* axes for the same compounds across the samples, aligning the extracted peaks to the same *m*/*z* and RT grid to a single identifier. The aligned peak intensity data are then further processed by a normalization procedure prior to statistical analysis, to remove the systematic bias introduced during sample preparation and the variation in the ionization efficiency and instrumental analysis. Common choices for normalizing metabolomics MS data include total intensity sum (TIS) (Manna *et al.*, 2013), internal standard calibration (ISC) (Cui *et al.*, 2013), or a statistical model that combines standards and TIS (Sysi-Aho *et al.*, 2007).

Although the existing open source software packages such as MZmine (Pluskal *et al.*, 2010) and XCMS (Smith *et al.*, 2006) perform peak picking and alignment, there is room for further improvement (Smith *et al.*, 2013). In our observations, for example, even the most sophisticated multi-sample alignment algorithms have been prone to misalignment error at the individual compound level, especially for less abundant compounds or multiply charged compounds. Misalignment tends to happen when (i) signal detection algorithms fail to separate co-eluting compounds, (ii) peak picking algorithms identify incorrect major isotopic peaks or (iii) minor temporal variation outlasts the alignment step by a few seconds of RT or a few decimals in the *m*/*z* beyond the tolerance level specified in those algorithms.

Besides the misalignment issue, the existing methods for data normalization also have limitations. For example, the internal standard measurements can be inaccurate in some samples, or mixed with co-eluting compounds. The TIS method may fluctuate due to poor chromatographic separation or the influence of a few dominantly abundant compounds, and it can also be inapplicable when the detected compounds are genuinely heterogeneous between comparison groups. Most importantly, the procedures mentioned above are corrections by a single constant, which adjusts all intensity values for global bias only and does not account for temporal or local variations along the RT axis across different samples (Rudnick *et al.*, 2014).

To address these two key limitations in the current data processing pipeline, we developed a software package MetTailor, which implements two post-extraction processing steps including a method for block-wise quantitative summary and a novel RT-based local normalization procedure.

2 Approach

2.1 Dynamic block summary algorithm

To address misalignment events in large sample datasets, we developed an algorithm called dynamic block summary (DBS) that takes de-isotoped and pre-aligned peak intensity data reported from data extraction software packages and dynamically reduces local misalignment errors in the chromatographic space (*m*/*z* and RT grids). The DBS algorithm creates rectangular *blocks* with *m*/*z* and RT grid coordinates that are small enough to capture a single compound in the majority or all of the samples, and then acquires quantitative summary for each block by the peak apex or the largest integrated peak area for downstream analysis in each sample. The resulting quantitative data corrects a certain degree of alignment errors when merging a large number of samples, yielding fewer missing data in the data table reported by blocks.

Figure 1 illustrates how the DBS algorithm works. The data shown are the raw peak intensity data from Agilent 6520 Q-TOF instrument, prior to any signal extraction in the Dengue dataset (see Section 4). In each example, XCMS reported at least two consecutive rows across the samples for the same compound. The point where the horizontal line and the first vertical line crosses indicates the alignment coordinate for a particular row in the XCMS table (the two other vertical lines are expected *m*/*z* coordinates for their isotopes). In all panels, alignment was successful in the first three samples at the fixed *m*/*z*-RT coordinate (i.e. thus reported in the same row in the original table), while it failed in the fourth sample despite having a very similar *m*/*z*-RT coordinate (horizontal or vertical shift in the alignment point). Since the DBS algorithm places rectangular blocks, it is able to bring the fourth sample back into the correct alignment and to summarize the maximal intensity data inside the box across all the samples. As a result, this process merges the two rows representing the same compound in different samples into a single row.

2.2 Isotopic pattern validation

The local realignment achieved by the DBS algorithm does not always merge the intensity data of the same compound. We thus propose to use the isotopic patterns, represented by the intensity ratios between the mono isotopic peak to the next two isotopic peaks, and the charge state as the quality metric of alignments. In other words,

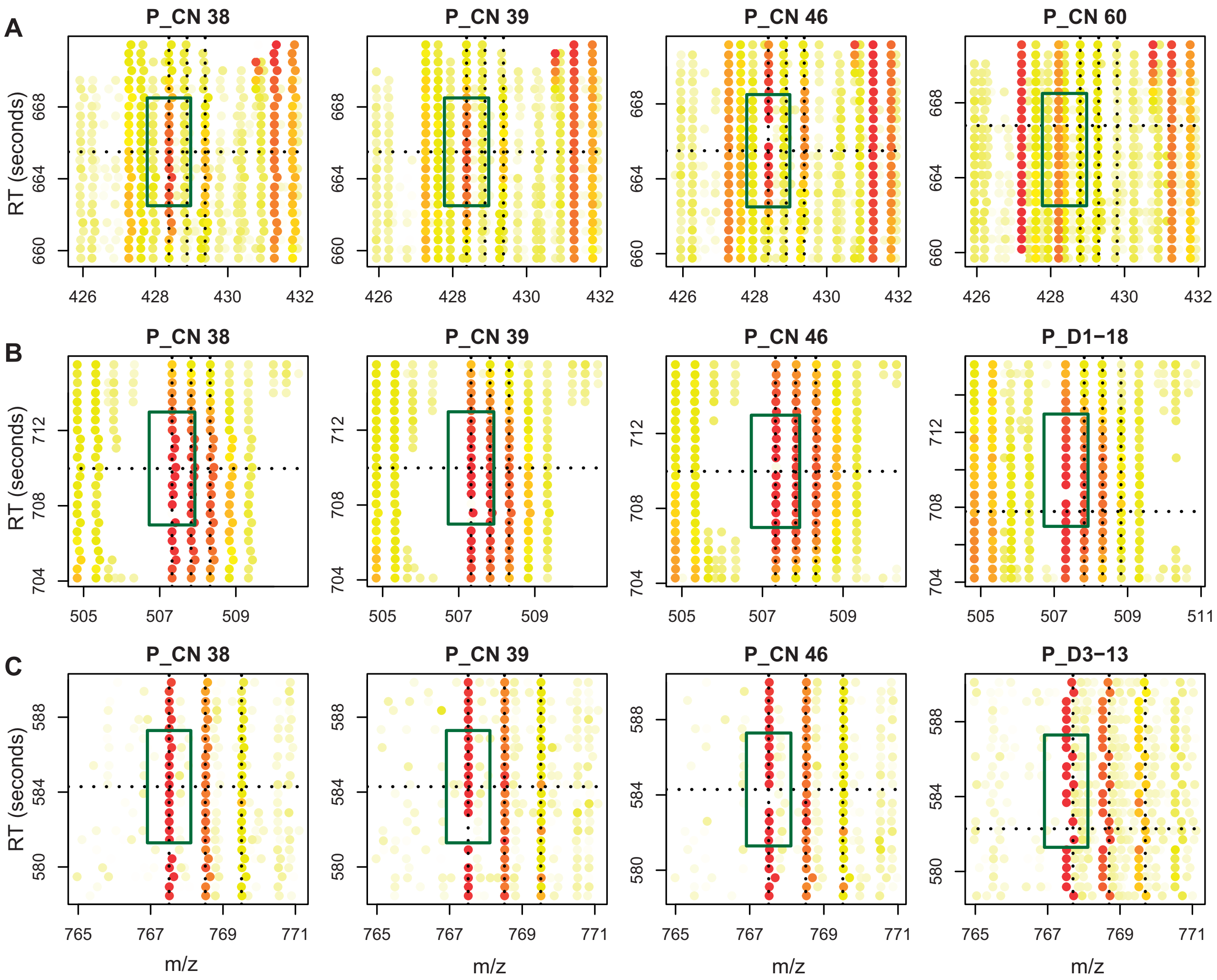

Fig. 1. Raw Q-TOF data prior to preprocessing by XCMS. Each circle is a unique peak extracted from the corresponding mzML file, whose intensity is indicated by the heat color (white: low intensity; red: high intensity). The green boxes are the blocks placed by the DBS algorithm. The case of failed alignment is presented in the fourth sample in each row, where the corresponding intensity data in the misaligned sample appeared in a separate row for that sample in the table reported by XCMS. The three panels show the different misalignment scenarios, but the DBS corrects the misalignment error in each case

if the DBS algorithm correctly realigns a peak in a sample to a different *m*/*z*-RT coordinate, then its isotopic pattern and charge state should be consistent with the peaks from other samples that had been well aligned to the same coordinate. To do this, we provide a program as a part of MetTailor distribution to identify peak clusters around fixed *m*/*z*-RT coordinates from the raw MS data and extract their isotopic patterns and the charge states.

2.3 Retention time-based local normalization

We also propose a novel normalization procedure that does not adjust all intensity values by a constant normalization factor. When a large number of samples are processed through a GC-MS/LC-MS platform, there are numerous *temporal* factors that can affect intensity measurements over the chromatographic time and subsequent steps such as the ionization efficiency in electro-spray ionization (ESI)-MS platforms. To address this, we developed a RT(δ)-based normalization method, in which the normalizing factor is acquired by the weighted sum of peak area values in the RT neighborhood (of size δ) of each compound at a specific RT.

The algorithm is capable of automatically determining the optimal window size δ (see Section 3 for details). In the RT regions where few compounds are found, the algorithm recognizes the scarcity of data and borrows the information from other RT regions with a sufficient amount of data. By design this procedure can adaptively remove temporal variations in the intensity data, and it becomes equivalent to the TIS normalization in the absence of such local variations. It is also advantageous in that it prevents a few intense compounds from dominating the normalizing factor across all ions within each sample.

3 Methods

The starting material for MetTailor is the aligned data reported by the software packages such as MZmine and XCMS. We denote the intensity data as $\mathbf{Y} = \{\mathbf{y}_j\} = \{y_{jtm}\}$, where the subscripts $t = (1, \ldots, T)$, $m = (1, \ldots, M)$ and $j = (1, \ldots, n)$ are the indexes for aligned retention times, mass-to-charge ratio (*m*/*z*) and samples, respectively. Hence y_{jtm} is interpreted as the intensity of a peak at the aligned grid of RT r_t and *m*/*z* x_m in sample *j*. *T* and *M* are the number of unique RT and *m*/*z* values appearing in the aligned data respectively. We also define the block center as the middle of the *m*/*z*-RT grid coordinates.

3.1 Dynamic block summary

Step A *Block Initialization*: Given the aligned data, the DBS algorithm starts by initializing blocks, where each block is uniformly sized covering m_0 in *m/z* and r_0 seconds in the chromatographic space (default values $m_0 = 1.2$ and $r_0 = 6$ seconds). For each block, the algorithm goes through a series of decision criteria to determine whether the current block definition is optimal or not. To explain the criteria, we have to define the optimality of peak alignment first: we consider the original processing software has aligned a peak well if, under the reported alignment, the peak intensity is above a minimal intensity threshold I_* (default 2000) in at least p% of the samples (default $p = 80$). We say that a peak intensity is missing in a sample if its intensity is below I_* or unreported. After initialization of blocks, each initial block may contain no well-aligned peak, one such peak, or multiple peaks, and the iterative block updates have customized steps for block position adjustment or segmentation in each of the three cases.

Step B1 *Iterative Updates: One well-aligned peak*. If the initial block already contains a single well-aligned peak and the block center is located on the *m/z*-RT grid of the peak, then the DBS terminates the block update for the current block and moves onto the block summary step (Step C). If not, it repositions the block so that the block center is at the grid of the well-aligned peak and the algorithm returns to Step A to restart the checklist (whether the modified block still contains only one well-aligned peak).

Step B2 *Iterative Updates: No well-aligned peak*. If the initial block contained no well-aligned peak meeting the criteria, then the DBS identifies a ‘lead’ peak, defined as the peak with the smallest number of missing data across the samples. If there are more than one such lead peaks, then the mean *m/z*-RT coordinate is defined as the lead peak. If the current block center coincides with this lead peak, the DBS terminates the block update and moves on to the block summary step. If not, it updates the grids of the block so that its center is placed on the lead peak, and returns to Step A to restart the checklist.

Step B3 *Iterative Updates: Multiple well-aligned peaks*. If the initial block contains two or more well-aligned peaks of different *m/z* and RT grids, then the DBS removes the current block definition, creates one block for each of the peaks, keeping the block size the same in both *m/z* and RT axes, and then returns to Step A (recording that this duplication move was made). If the DBS returns to this same point of multiple peaks situation (D peaks), with a record of duplication move, then it breaks the current block into D segments. Each segment retains the same *m/z* range as the parent block, but the RT range will now be defined by setting break points at the mid points in the RT axis between the D well-aligned peaks. After block segmentation, the algorithm returns to Step A. These updates are iterated until there are no more blocks containing multiple well-aligned peaks.

Step C *Summary of intensities and removal of replicated peaks*. Once the DBS moves onto this stage for each block, then it summarizes the intensity value for the block by the largest intensity value, which corresponds to either the apex of an elution profile or the integrated peak area, depending on the initial data processing scheme. Finally, the DBS goes through the list of summarized blocks and removes duplicated blocks.

3.2 Isotopic pattern validation

As a means to extract and compare charge states and isotopic patterns of realignments, we also implemented a program to first extract the MS1 raw peaks from the *m/z*-RT coordinates in the aligned data. The peak extraction is performed for *m/z* neighborhood with default parameters of -2 amu to $+4$ amu on the *m/z* axis and -20 to $+40$ s in the RT axis. It then routinely determines the charge states of the candidate compound and extracts the isotopic peak intensity pattern in the form of ratio between the mono isotopic peak and the subsequent two isotopic peaks. For each aligned peak, the program reports the mean and standard deviation of the isotopic ratios across the samples. For a newly recovered peak, we consider the peak is well recovered if its isotopic ratio falls within a reasonable range of the distribution of the ratios in the samples with good initial alignments (e.g. 95% percentile).

3.3 RT(δ) normalization

Following the DBS algorithm, MetTailor offers a data normalization step, where the user can choose either the TIS normalization or a novel method called RT(δ), or opts to skip the normalization step.

TIS normalization: For the TIS normalization, we first compute the TIS for sample j

$$T_j = \sum_{s,\ell} y_{js\ell}$$

and transform the data by

$$y_{jtm} \rightarrow y_{jtm}/T_j.$$

Following this step, we rescale the entire dataset so that the total normalized intensity across all compounds and samples is equal to the total sum of pre-normalization intensities. This step ensures that the normalized intensity values are on a comparable scale with the raw intensity values.

RT(δ) normalization: In this new normalization procedure, we transform the data as follows. At the block indexed by (t, m), we first compute the sample-specific local weighted intensity sum

$$W_{jt} = \sum_{s,\ell} y_{js\ell} g_{\delta_j}(r_t - r_s)$$

for all j, where $g_{\delta_j}(\cdot)$ is the Gaussian kernel function with standard deviation δ_j. We transform the data by

$$y_{jtm} \rightarrow y_{jtm}/W_{jt}$$

for all peaks (with *m/z* value x_m). Here the ‘size’ of the weighting function δ_j can be provided by the user, or determined by our automated algorithm (see below). Following this correction, we rescale the data in a similar way as the TIS normalization, but at a specific RT grid. In other words, we rescale the data at each RT grid so that the total intensity sum is the same between the raw data and the adjusted data at the specific RT. This step ensures that the order of absolute abundance levels is roughly retained across the compounds after normalization. Note that the window size δ should not be too small since the scaling factor will be dominated by the intensity of the compound itself, especially in highly abundant compounds. On the other hand, larger δ will draw this algorithm closer to the TIS normalization.

Automated selection of $\{\delta_j\}$. We search for the optimal δ_j for each sample as follows. We first transform the data by

$$y_{jtm} \rightarrow y_{jtm}/T_j * (n^{-1}\sum_{\ell=1}^{n} T_\ell),$$

so that the data is scaled to have the same TIS across all the samples. We then place a sliding window with size 1 min and slide the window by 30 s each time, recording the intensity sum in each sample at each position of the sliding window. For sample j, we look for the sample with the most distinct total ion chromatogram (TIC) profile, say k, for which we compute the intensity sum differences across all the windows and find the longest streak of consecutive windows with an identical sign (positive or negative). We iterate the same search for each sample to determine the optimal window size, and then set δ_j as a quarter of the window size, based on the fact that four times the standard deviation δ of a Gaussian kernel will cover 95% of such a window size.

3.4 Model-based significance analysis

For the differential expression analysis of block-summarized and normalized data, we used the hierarchical Bayesian model proposed by Wei and Li (2007), implemented in the mapDIA software (see Supplementary Information). mapDIA was initially developed for quantitative proteomics data analysis obtained from data independent acquisition MS (DIA-MS), but the differential expression analysis component is directly applicable to any intensity data. In the model, two possible probability models of intensity data are proposed for each compound, namely differential expression (DE) model and non-DE model, respectively, and the posterior probability of DE is calculated. The posterior probability scores are used to select the differentially expressed compounds and is used to compute the false discovery rates (FDR) (Newton *et al.*, 2004).

4 Results

4.1 Datasets

We used two published datasets to evaluate the performance of MetTailor in this work. The first is the metabolomics dataset of 24 samples with extracted feature map (featureXML files) and curated alignment from Lange *et al.* (2008). We will call this dataset M2 data following the reference in the original paper. To replicate the steps in the paper, we also made the XCMS read the featureXML files skipping its peak detection step, and perform alignment and retention time correction with the same parameters as reported in the paper. For MetTailor, we used all default parameters other than setting the minimal intensity 0 to avoid removal of any data points. Then we benchmarked all DBS-recovered peaks to the curated peak alignment for verification.

The second dataset is from a recently published metabolomics study for Dengue virus infection (Cui *et al.*, 2013) (Dengue data hereafter). The goal of this study was to identify differentially expressed metabolites in the serum samples of Dengue patients at different phase of infection and recovery. In the study, blood samples were taken from 49 control subjects and 27 subjects diagnosed with acute Dengue fever (DF) but not life threatening Dengue haemorrhagic fever/Dengue shock syndrome, where samples were collected from the patients multiple times at early febrile (rising fever), defervescence (abating fever) and convulscant (recovery) stages. This comprises 115 MS runs in total. The samples were analysed using Agilent 6520 Q-TOF mass spectrometer coupled with ultrahigh pressure LC.

We processed the data using XCMS (Smith *et al.*, 2006) to perform peak detection and alignment, and applied MetTailor to the output data. In the XCMS processing, the default options were used (see Supplementary Information). For MetTailor processing of this data, we set default parameters. For quality assessment of realigned peaks, charge states and isotopic peaks intensity patterns were extracted from the raw data using the *m*/*z*-RT coordinates reported by MetTailor. RT(δ) normalization was performed with the automatically optimized window size δ. For the statistical significance analysis using mapDIA, we set the score thresholds associated with 5% FDR.

4.2 DBS algorithm

4.2.1 The DBS algorithm recovers misalignments with high sensitivity and specificity

The M2 data contains on average 16 122 features in the 24 samples and the authors have curated 2630 gold standard peaks by consensus peak groups with annotations from the CAMERA package (Kuhl *et al.*, 2012). As reported in the reference paper, the initial alignment by XCMS had recall rate 98% and precision 78%. However, MetTailor discovered a small proportion of misalignment events (21 misalignments within 8 unique *m*/*z*-RT coordinate, Supplementary Table S1) and 19 recoveries of these 21 were consistent with the curated alignment. Hence, through MetTailor, we verified that XCMS achieved the initial alignment really well for the majority of the gold standard peaks and a small number of misaligned cases were re-aligned. We note that, however, the curated peaks comprise merely ~16% of the entire peak list and these high quality peaks with consistent annotation are expected to be aligned well, and therefore our evaluation based on this dataset is quite limited.

In the Dengue data, the table initially reported by XCMS contained 7920 aligned peaks appearing in at least one of the 115 MS runs. Similar to the M2 data, many peaks were localized to highly consistent *m*/*z* and RT coordinates across the samples, if not perfectly aligned across the samples. In the reported table, misaligned peaks appeared as different peaks, and if the intensity levels are compared between groups of samples (DF stages), then the samples with misalignments will be considered as missing data, affecting the statistical analysis. The DBS algorithm alleviated this burden considerably by merging these misaligned peaks into single blocks, while decreasing the frequency of missing data at the same time. Specifically, the DBS algorithm reduced the number of aligned peaks from 7920 to 6915. In this process, the algorithm corrected total of 6989 misalignments, accounting for 2.2% of 315 454 non-missing data points in the final data. To verify that the realignments by the DBS algorithm are likely recoveries into the correct line of the table, we compared the isotopic patterns of the recovered peaks to those of samples with good initial alignment. We could extract clear isotopic patterns from 2380 recovered peaks at 285 unique *m*/*z*-RT coordinate, and the isotopic ratios between the first and second isotopic peaks fell in 95% of the ratio distributions from the previously well aligned peaks in 1938/2380 of the cases (81.4%) (see Supplementary Table S2 for summary at 285 *m*/*z*-RT coordinates). Combined with the similarity of *m*/*z*-RT cooridnates, this reflects high quality realignments by the DBS algorithm across the dataset.

4.2.2 Misalignments arise from a few major sources

The fact that merely 2.2% of the peaks were corrected confirms that the initial data extraction and alignment was of a high quality in this data. To gain insights for the source of misalignments, we plotted the distances to the misaligned peaks on both *m*/*z* and RT axis from the *m*/*z*-RT coordinate to which the majority of the samples are aligned (Fig. 2). 4601/6989 misaligned peaks showed *m*/*z* shift of 0.4–0.6 amu and 681/6989 misaligned peaks showed a *m*/*z* shift by 0.27–0.4. This suggests that these compounds are likely doubly or triply charged and the misalignment was caused by incorrect grouping of the isotopic peaks. This occurred frequently in the data

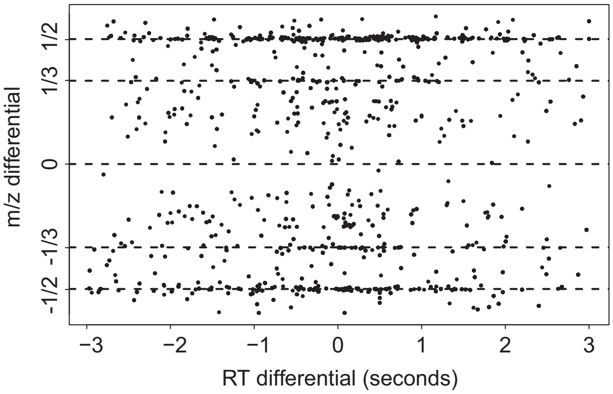

Fig. 2. The RT and *m/z* differences of misaligned peaks to the *m/z*-RT coordinate to which the majority of the samples aligned in Dengue data. The dashed lines were drawn at ± 1/2 and ± 1/3, which indicates that misaligned peaks are likely from doubly or triply charged compounds

when multiple co-eluting compounds in the vicinity of *m/z*-RT regions, since the elution profiles of the isotopic peaks overlap. Misalignment also happened when minor temporal variations outlast the alignment step by a few seconds of RT or a few decimals in the *m/z* beyond the tolerance level specified in those algorithms.

4.2.3 Block-wise recovery reduces missing data in RT regions with active chromatographic elution

To better understand the benefit of the DBS algorithm, we further investigated where in the chromatographic space (RT and *m/z*) and in what abundance range the DBS algorithm recovered misalignments. Figure 3 shows the number of recovered misalignment events across 115 samples in the Dengue data, in terms of the chromatographic time (RT) and the abundance range. Figure 3A shows that the frequency of re-alignments by the DBS algorithm is correlated with the cross-sample TICs (dashed line), which was computed as the sum of TICs across the RTs. At the same time, Figure 3B shows that misalignments occurred mostly for low to medium intensity peaks. Taken together, this suggests that the misalignment occurs more frequently for low abundance compounds in active RT regions with elution of multiple compounds. This also indicates that the alignment procedure built in XCMS is primarily driven by the most intense peaks in local RT regions.

4.3 Normalization

Next we tested three different normalization methods, including the internal standard calibration, TIS normalization and RT(δ) normalization. Note that the first two normalization methods uses a single constant as the normalizing factor for each sample, while the RT(δ) adjusts the normalizing factor locally at each RT grid in each sample. The first two methods are also different from each other in the sense that the first uses a known compound injected before sample preparation for LC-MS whereas the TIS is calculated from the extracted intensity data.

Before we compared the normalization methods, we first examined the raw data and discovered that there was no visible variation in terms of the TIC profiles across the samples, indicating that there was no constant systematic variation in the chromatography across a large number of MS runs [mean correlation 0.971 and 95% confidence interval (0.925,0.994)]. However, a closer examination revealed temporal variations in TIC across many samples. Figure 4A shows the TIC plots in the raw intensity data, with a pair of samples highlighted in purple and orange colors. It is easy to see that the sample in pink had much higher intensities between 600 and 800 s, while the pattern dissipated after 800 s. Even though these two samples were from different patients, the overall total intensity differences were very specific to certain RT periods and this pattern was consistently observed across many pairs of samples.

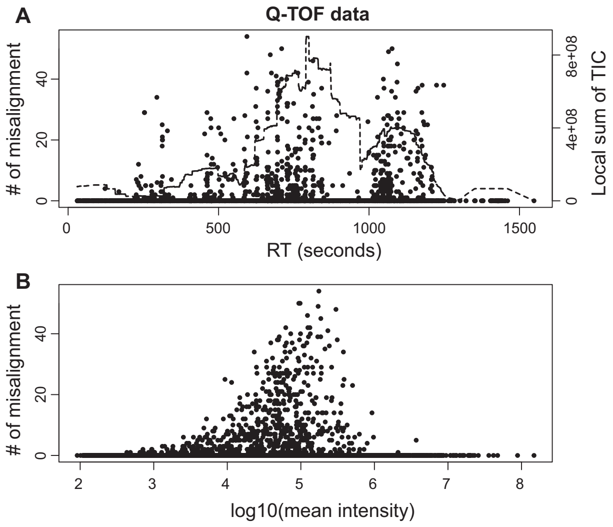

Fig. 3. The frequency of missing data recovery by the DBS algorithm in the Dengue data (115 runs) across the RT (**A**) and the abundance range (**B**). The dashed line is the cross-sample local sums of TICs in sliding windows of RT (after scaling to the misalignment frequency plot)

4.3.1 Internal standard calibration and TIS normalization

The acetic acid standard (9-fluorenylmethoxycarbonyl-glycine) showed little variation across the samples in the Dengue data (Supplementary Fig. S1). The internal standard appeared in the middle of the RT range (408 s) and their intensities were not dominant (the standards accounted for 0.3% of the TIS). Meanwhile, the TIS values were also calculated for normalization. Interestingly, the TIS values were not correlated with the intensity values of the internal standard (Pearson correlation 0.11; Supplementary Fig. S1). As a result, the normalized data were substantially different between the two methods (Fig. 4B, C) and this also led to unique sets of compound selected in the DE analysis (see next section).

4.3.2 RT(δ) normalization removes temporal variation along the RT axis

For the RT normalization, δ values chosen by the automated search varied in the range of 2–4 min, which is around 10% of the total RT. Figure 4A suggests that there is a strong evidence of temporal variation in the raw data between 400 and 800 s. The two single constant-based normalization (TIS and ISC) reduced this temporal variation at the expense of amplifying the variation in the 800–1000 second period across the samples. By contrast, the proposed RT(δ) normalization have substantially reduced such variations localized to specific RT periods (Fig. 4D). In the RT regions with few detected peaks, stable normalizing factors could not be calculated and were possibly dominated by the compound itself due to the lack of a sufficient number of background. In these regions, our procedure borrowed information from the median pattern of local intensity sums from other RT regions and normalized the data against those values.

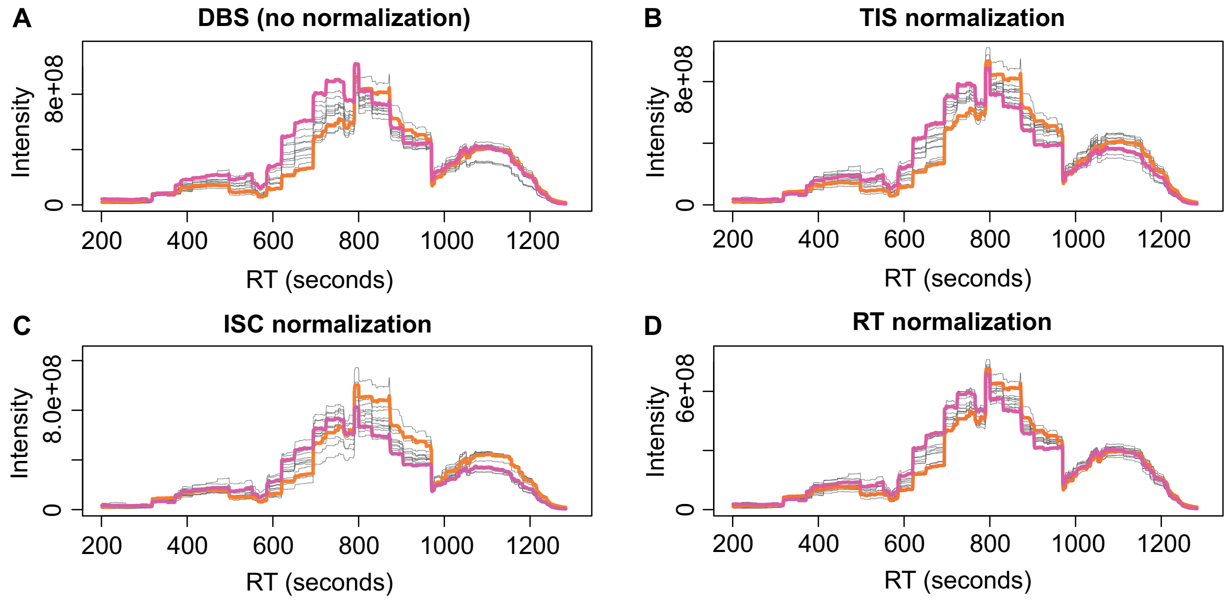

Fig. 4. TIC plots across random selected 14 samples in the Dengue data with (**A**) no normalization (DBS), (**B**) total intensity sum (TIS), (**C**) internal standard calibration (ISC) and (**D**) retention time (RT), respectively. Purple and orange lines are the pair of two samples with the lowest correlation

4.3.3 RT(δ) normalization does not lead to loss of statistical power

Although we demonstrated that the normalization procedure removed local chromatographic variations in the visual sense, it is important to verify that this correction neither removes real biological variations nor introduces false positives. To evaluate the possibility of such consequences, we first performed DE analysis between controls and patients in the febrile stage with normalization options. For the DE analysis, we set the significance score threshold to control the Bayesian FDR at 5%. We also considered the peaks or blocks that can be assigned at least one compound identifier assigned through the online HMDB search (tolerance ± 10 ppm, singly doubly or triply charged, [M+H], [M + 2 H] and [M + 3 H], no other adducts, peptides excluded).

The three normalization methods reported 476 DE compounds (significant in at least one method) and 62% were shared by all methods. Each method reported a handful of unique compounds nonetheless. In RT(δ) normalized data, which we deem to be the optimal normalization method in this data, led to a selection of unique DE compounds which were more spatially randomly distributed across the RT. Moreover, when compounds show DE in both TIS and ISC normalized data, a majority of them are also DE in RT(δ) normalized data. On the other hand, many of the compounds significant in the TIS and RT-normalized data were found to be non-significant in the ISC normalized data. Supplementary Figures S2 and S3 show the distribution of significant compounds in the data with each normalization method and the differences between the three outputs respectively. The latter figure verifies that the RT normalization did not lead to potential loss of statistical power compared to the other two normalization methods.

5 Discussion and conclusion

In this work, we developed a software package MetTailor, featuring a novel block summary method called DBS algorithm and an adaptive normalization method to remove temporal variations across a large number of untargeted MS experiments. These additional data processing steps were introduced as a complementary tool for quality assurance of the data reported by the extraction tools such as MZmine and XCMS. With inexpensive computational effort, these steps can address missing data problem and remove local variations that vary across retention time in a long series of MS experiments. We also provide implementations to extract the charge state and isotopic profiles from the raw data at specific *m*/*z*-RT coordinates so that the users can evaluate the quality of post-alignment peak summary tables as a validation step.

5.1 Utility of the DBS algorithm

Although the number of re-alignment events was relatively modest in our demonstration with both datasets, the DBS algorithm is computationally inexpensive and a few minutes of runtime can provide an important quality check for the initial alignment data, with additional recovery of thousands of misaligned peaks across the samples. We expect that the DBS algorithm will be equally useful for datasets generated with older generation mass spectrometers (e.g. linear quadrupole ion trap MS, old generation Q-TOF) and the next generation ones (e.g. Quadrupole-Orbitrap, QqTOF), since the alignment accuracy has more to do with temporal variations in the elution rate of individual compounds across the samples, rather than the mass accuracy and resolution of MS. A complementary tool such as MetTailor will therefore be of great importance from the data quality control perspective.

However, we remark that the DBS algorithm is inherently limited by the fact that its working material is the preprocessed data, which is a significantly compressed form of information for each peak feature as centroiding, deisotoping, and compression of adducts is applied during the first-round processing. With the increasing computation power, a more advanced alignment algorithm with a local alignment procedure matching individual peak features can be developed, directly utilizing various raw peak features such as isotopic distributions and adduct information without a severe data compression step.

5.2 Missing data imputation

Another important point of discussion is on the use of machine learning methods to impute missing data (Gromski *et al.*, 2014). One may argue that the DBS procedure can be replaced by using other imputation procedures such as *k*-nearest neighbor or random forest imputation (Stekhoven and Bühlmann, 2012; Troyanskaya *et al.*, 2001). To put the utility of the two approaches in perspective, it is important to recognize the difference in the source of missing data addressed by each method. The DBS algorithm is addressing the problem of misalignment or variation in the elution profiles of the same analytes across different GC-MS/LC-MS experiments. Since this approach addresses the missing data problem incurred during the data tabulation process, it only deals with the source of missing data that can be corrected within the limit of detection of mass spectrometers.

By contrast, the aforementioned methods perform imputation by inference, which do not differentiate between different sources of missing data and primarily aim to fill in the missing values based on the pattern matching with other compounds. In practice, this group of methods should be used carefully in highly variable systems as GC-MS/LC-MS, as a last resort after all available remedies to correct the errors in the data extraction stage have been exhausted. This is because the algorithms depend heavily on the availability of informative neighbor compounds in the same dataset. Moreover, although these methods are widely used at present, they were evaluated primarily on gene expression microarrays that quantify predetermined molecular targets. Hence whether the same procedures will be applicable and equally efficient on MS-based metabolomics datasets, plagued by the ambiguity of isomers in small molecules, remains to be evaluated.

5.3 Data normalization

Finally, we have also proposed a novel normalization algorithm that removes temporally systematic variations frequently observed in the GC-MS and LC-MS pipeline. Retention time is clearly one of the most important anchors of experimental variations in the GC-MS and LC-MS platforms (Rudnick *et al.*, 2014), which is often ignored in conventional computational normalization strategies. It is also important to note that some of the distribution equalizing methods such as IQR-based normalization or quantile normalization can introduce over-adjustments, since these procedures are applicable only when a certain assumptions are met. For example, the procedure based on mean centering and standard deviation scaling requires that the overall log-intensity distribution be normally distributed. The quantile normalization is not applicable when missing data are prevalent and can be risky when there are only a handful of metabolites (e.g. a few hundreds) since the procedure can easily over-correct intensities of very high or very low abundance compounds due to the granularity of percentile points.

5.4 Conclusion

Overall, our contribution in this work is to provide an open source software package implementing two data processing steps, as a complimentary tool to remedy some gaps unaddressed by the popular data extraction tools in the context of large sample experiments. There are a number of experimental factors that are unique to MS platforms and the two proposed methods are different from the existing alternatives that had been developed for other -omics platforms such as gene expression microarrays. We provide these tools in a popular R programming environment, and will continue to adapt the tools for constantly evolving instrumentation in the future.

Acknowledgments

The authors are grateful to Scott Walmsley and Damian Fermin for helpful discussion, and to Hiromi Koh for carefully reading the manuscript.

Funding

This work was supported in part Singapore Ministry of Education Tier 2 grant (R-608-000-088-112).

Conflict of Interest: none declared.

References

Cui,L. *et al.* (2013) Serum metabolome and lipidome changes in adult patients with primary dengue infection. *PLoS Neglected Trop. Dis.*, **7**, e2373.

Dettmer,K. *et al.* (2007) Mass spectrometry-based metabolomics. *Mass Spectr. Rev.*, **26**, 51–78.

Griffiths,W.J. and Wang,Y. (2009) Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chem. Soc. Rev.*, **38**, 1882–1896.

Gromski,P.S. *et al.* (2014) Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, **4**, 433–452.

Horai,H. *et al.* (2010) Massbank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectr.*, **45**, 703–714.

Kuhl,C. *et al.* (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography / mass spectrometry datasets. *Anal. Chem.*, **84**, 283–289.

Lange,E. *et al.* (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, **9**, 375.

Manna,S.K. *et al.* (2013) Metabolomics reveals aging-associated attenuation of noninvasive radiation biomarkers in mice: potential role of polyamine catabolism and incoherent DNA damage-repair. *J. Proteome Res.*, **12**, 2269–2281.

Newton,M. *et al.* (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Pluskal,T. *et al.* (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

Rudnick,P. *et al.* (2014) Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data. *Mol. Cell. Proteomics*, **13**, 1341–1351.

Smith,C.A. *et al.* (2005) METLIN: a metabolite mass spectral database. *Therap. Drug Monitoring*, **27**, 747–751.

Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Smith,R. *et al.* (2013) LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief. Bioinf.*, 16, 104–117.

Stekhoven,D. and Bühlmann,P. (2012) MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–118.

Sysi-Aho,M. *et al.* (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, **8**, 93.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Wei,Z. and Li,H. (2007) A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**, 1537–1544.

Wishart,D.S. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.

Wishart,D.S.S., *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.