Counting Missing Values in a Metabolite-Intensity Data Set for Measuring the Analytical Performance of a Metabolomics Platform

ARTICLE *in* ANALYTICAL CHEMISTRY · DECEMBER 2014

Impact Factor: 5.64 · DOI: 10.1021/ac5039994 · Source: PubMed

Tao Huan

The Scripps Research Institute

analytical chemistry

Article

pubs.acs.org/ac

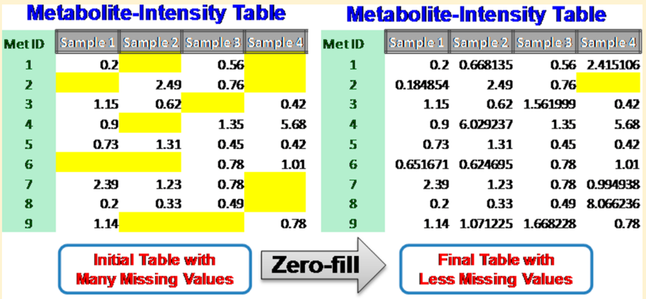
Counting Missing Values in a Metabolite-Intensity Data Set for Measuring the Analytical Performance of a Metabolomics Platform

Tao Huan and Liang Li*

Department of Chemistry, University of Alberta, Edmonton, Alberta T6G2G2, Canada

S Supporting Information

ABSTRACT: Metabolomics requires quantitative comparison of individual metabolites present in an entire sample set. Unfortunately, missing intensity values in one or more samples are very common. Because missing values can have a profound influence on metabolomic results, the extent of missing values found in a metabolomic data set should be treated as an important parameter for measuring the analytical performance of a technique. In this work, we report a study on the scope of missing values and a robust method of filling the missing values in a chemical isotope labeling (CIL) LC-MS metabolomics platform. Unlike conventional LC-MS, CIL LC-MS quantifies the concentration differences of individual metabolites in two comparative samples based on the mass spectral peak intensity ratio of a peak pair from a mixture of differentially labeled samples. We show that this peak-pair feature can be explored as a unique means of extracting metabolite intensity information from raw mass spectra. In our approach, a peak-pair peaking algorithm, IsoMS, is initially used to process the LC-MS data set to generate a CSV file or table that contains metabolite ID and peak ratio information (i.e., metabolite-intensity table). A zero-fill program, freely available from MyCompoundID.org, is developed to automatically find a missing value in the CSV file and go back to the raw LC-MS data to find the peak pair and, then, calculate the intensity ratio and enter the ratio value into the table. Most of the missing values are found to be low abundance peak pairs. We demonstrate the performance of this method in analyzing an experimental and technical replicate data set of human urine metabolome. Furthermore, we propose a standardized approach of counting missing values in a replicate data set as a way of gauging the extent of missing values in a metabolomics platform. Finally, we illustrate that applying the zero-fill program, in conjunction with dansylation CIL LC-MS, can lead to a marked improvement in finding significant metabolites that differentiate bladder cancer patients and their controls in a metabolomics study of 109 subjects.



Metabolite-Intensity Table

Met ID	Sample 1	Sample 2	Sample 3	Sample 4
1	0.2	🟨	0.56	🟨
2	🟨	2.49	0.76	🟨
3	1.15	0.62	🟨	0.42
4	0.9	🟨	1.35	5.68
5	0.73	1.31	0.45	0.42
6	🟨	🟨	0.78	1.01
7	2.39	1.23	0.78	🟨
8	0.2	0.33	0.49	🟨
9	1.14	🟨	🟨	0.78

Metabolite-Intensity Table

Met ID	Sample 1	Sample 2	Sample 3	Sample 4
1	0.2	0.668135	0.56	2.415106
2	0.184854	2.49	0.76	🟨
3	1.15	0.62	1.561999	0.42
4	0.9	6.029237	1.35	5.68
5	0.73	1.31	0.45	0.42
6	0.651671	0.624695	0.78	1.01
7	2.39	1.23	0.78	0.994938
8	0.2	0.33	0.49	8.066236
9	1.14	1.071225	1.668228	0.78

Initial Table with Many Missing Values
Zero-fill
Final Table with Less Missing Values

Missing intensity values is common in a multiple-sample data set generated by an "omics" analytical tool for genomics, proteomics, and metabolomics applications.[1−6] One of the major roles of an omics study is to find genes, proteins, or metabolites that have significant differences in different biological groups (e.g., healthy vs diseased samples). Analytical tools are used to generate a rectangular matrix or table containing an intensity (or quantity) value in each sample column that is associated with an individual gene, protein, or metabolite in a row. Missing values in the table can cause problems in performing a statistical calculation.[7] Genomics and proteomics researchers have devoted a considerable amount of effort to understand and develop appropriate methods to handle the missing data.[1−3,8−13] There is an increasing awareness of this problem in the field of metabolomics, and several papers have been published on this topic,[4−6,14−22] including the development of statistical tools to fill the missing values or simply disregard all the features with missing data. However, filling the missing values nonexperimentally needs to be carefully performed.[5,6,22] There are debates on whether missing values should be filled and, if so, how best the missing values are filled (e.g., should we use the lowest intensity or a mean of all the measured values in a data set to fill the missing values?).[18−20]

We echo the view of a growing number of researchers on the importance of dealing with missing values properly in metabolomics. In our view, an effective approach to tackle the problem is from the experimental side, i.e., developing and applying robust analytical tools to profile the metabolomes of many samples with the least number of missing values. In an ideal situation, there should be very few missing values if a metabolomic technique is capable of detecting and quantifying all the metabolites; missing values would only indicate the true absence of the metabolites due to biological reasons. However, because of technical limitations of current analytical methods, the extent of missing values can be quite large, even in a replicate data set of the same sample where metabolite concentrations should be the same. In the case of LC-MS based metabolomics research, low-concentration or not-easily ionizable metabolites may not be detected due to a detection

Received: October 26, 2014
Accepted: December 12, 2014
Published: December 12, 2014

DOI: 10.1021/ac5039994

sensitivity issue or ion suppression effect. In addition, data processing including peak picking may cause the loss of peak intensity information.[18,22,23] There are several metrics including detection sensitivity, technical precision, quantification accuracy, and the number of detectable metabolites that have been routinely used to measure the analytical performance of a metabolome profiling technique.[24−26] We feel that the extent of missing values should be considered as another important parameter for gauging the performance of a method. In other words, the number of missing values should be reported, like the number of metabolites profiled, as a criterion to gauge the quality of a data set.

In this work, we report an investigation of the issue of missing values in a chemical isotope labeling (CIL) LC-MS metabolomics platform. In high-performance CIL LC-MS, the isotope labeling reagents are rationally designed to improve both LC separation efficiency and MS detection sensitivity significantly.[27−34] For example, dansylation, targeting the amine/phenol submetabolome, allows the detection of labeled metabolites with a sensitivity improvement of 10- to 1000-fold over the unlabeled counterparts.[27] With the ability of detecting thousands of putative metabolites from an individual sample (e.g., human urine) by using this platform, an important question rises as to how well we can profile them consistently in multiple samples, as metabolomics requires analyzing many samples of usually the same type, not just one sample. To this end, we have developed a data processing workflow that explores a unique feature of peak-pair picking from mass spectra generated by differential CIL LC-MS in order to fill the missing values in a multiple-sample data set. This method allows a significant reduction of missing values, enabling determination of a greater number of significant metabolites that separates different groups of samples, a common goal of many metabolomics studies in disease biomarker discovery and systems biology. To facilitate method comparison in terms of missing values, we propose a standardized approach of counting missing values in a replicate data set as a way of gauging the extent of missing values for a given analytical method.

EXPERIMENTAL SECTION

Dansylation Labeling. ^{12}C-dansyl chloride for metabolite labeling was purchased from Sigma-Aldrich Canada (Markham, ON, Canada). ^{13}C-dansyl chloride was synthesized in our lab.[27] The labeling reaction was performed according to a protocol reported previously.[34]

LC-MS. The ^{12}C- and ^{13}C-labeled samples were mixed and centrifuged at 20 800*g* for 10 min before injecting into a Bruker Maxis Impact QTOF mass spectrometer (Billerica, MA, USA) linked to an Agilent 1100 HPLC system (Palo Alto, CA, USA). A reversed-phase Zorbax Eclipse Plus C18 column (2.1 mm × 100 mm, 1.8 μm particle size, 95 Å pore size) from Agilent was used. Solvent A was 0.1% (v/v) formic acid in water with 5% (v/v) ACN, and solvent B was 0.1% (v/v) formic acid in ACN. The gradient elution profile was as follows: t = 0.0 min, 20% B; t = 3.5 min, 35% B; t = 18.0 min, 65% B; t = 24 min, 99% B; t = 28 min, 99% B. The flow rate was 180 μL/min. The sample injection volume was 2 μL.

Zero-Fill Program. The LC-MS data generated were first processed using a peak-pair picking software, IsoMS.[35] The level 1 peak pairs[35] were aligned from multiple runs by retention time match within 30 s and accurate mass match within 5 ppm to produce a CSV file. The zero-fill program was then used to fill the missing values in the CSV file. This program was written in R and is freely available from www.mycompoundid.org.[36]

In zero-fill, finding the missing value of a peak pair in the raw data of a sample uses information on retention time (rt), *m/z* value (mz), and absolute intensity (int) of the ^{13}C-peak of the pair. The ^{13}C-peak is from a controlled sample (e.g., a ^{13}C-labeled pooled sample) that is spiked into all the ^{12}C-labeled individual samples. Thus, the absolute intensity of this peak for a given labeled metabolite should be theoretically the same in mass spectra of all the samples. A matching score is used to find the peak pair based on similarities of these three parameters. It is defined as

$$\text{score} = \left(1 - \frac{\text{rt.diff}}{\text{rt.tol}}\right)/4 + \left(1 - \frac{\text{mz.diff}}{\text{mz.tol}}\right)/2 + (1 - 2 \times \text{int.diff})/4$$

where

$$\text{rt.diff} = \text{abs}(\text{rt.}^{13}\text{C.peak} - \text{rt.rawdata.peak})$$

$$\text{mz.diff} = 1 \times 10^{6} \times \frac{\text{abs}(\text{mz.}^{13}\text{C.peak} - \text{mz.rawdata.peak})}{\text{mz.}^{13}\text{C.peak}}$$

$$\text{int.diff} = \text{abs}\left(\log\left(\frac{\text{int.}^{13}\text{C.peak}}{\text{int.rawdata.peak}}\right)\right)$$

The default rt tolerance (tol) is 30 s, and the default mz tolerance is 5 ppm. A different weight (divided by 2 or 4) is assigned to each of the similarity equations in the score function; mz is deemed to be more important than rt and int and therefore given more weight. We tested different weighing factors and found these were most appropriate to generate the optimal results. If the matching score is larger than 0.6, it will be considered as a match. This scoring algorithm was developed using several metabolomic data sets where missing values in metabolite-intensity tables had been manually picked from the raw data.

Statistical Analysis. Multivariate statistical analysis was carried out using SIMCA-P+ 12 (Umetrics AB, Umea, Sweden). Volcano plot was plotted using Origin 8.5.

RESULTS AND DISCUSSION

IsoMS and Missing Values. Figure 1 shows the workflow for processing CIL LC-MS data. IsoMS is used to perform peak picking, peak pairing, peak-pair filtering, and peak intensity ratio calculation.[35] Using IsoMS-align script, information on the peak pair IDs and their peak intensity ratios from multiple LC-MS runs is extracted to produce a CSV file. In picking the peak pairs, IsoMS classifies the peak pairs into three groups, namely, level 1, 2, or 3.[35] Level 1 peak pairs are the most confident pairs where the ^{13}C-natural-isotope peaks are accompanied by the light- and heavy-chain labeled metabolite peaks within a pair. Level 2 peak pairs miss one of the ^{13}C-natural-isotope peaks. Level 3 peak pairs are the least confident pairs with both ^{13}C-natural-isotope peaks missing. To reduce the extent of false positive peak pairs found by IsoMS, only level 1 peak pairs are retained in the metabolite-intensity table. In doing so, the false positive rate (FPR) is usually less than 5%.

Inspecting the metabolite-intensity table generated by IsoMS, it is apparent that there are many missing values in the table

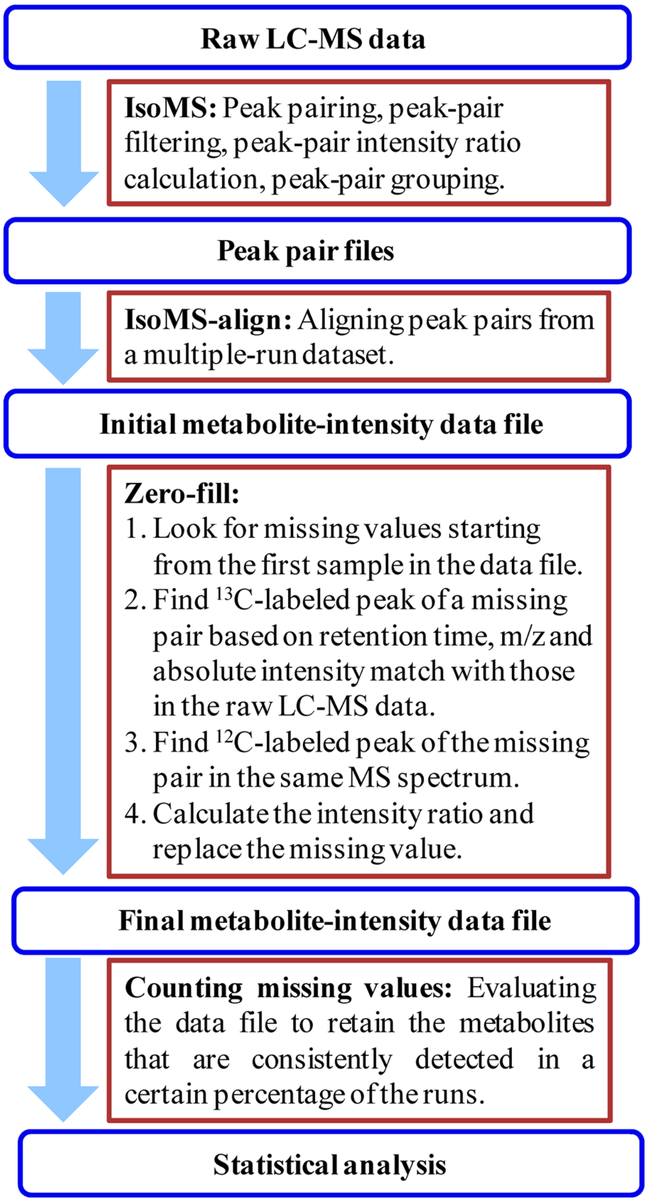

Raw LC-MS data
IsoMS: Peak pairing, peak-pair filtering, peak-pair intensity ratio calculation, peak-pair grouping.
Peak pair files
IsoMS-align: Aligning peak pairs from a multiple-run dataset.
Initial metabolite-intensity data file
Zero-fill:
1. Look for missing values starting from the first sample in the data file.
2. Find 13C-labeled peak of a missing pair based on retention time, m/z and absolute intensity match with those in the raw LC-MS data.
3. Find 12C-labeled peak of the missing pair in the same MS spectrum.
4. Calculate the intensity ratio and replace the missing value.
Final metabolite-intensity data file
Counting missing values: Evaluating the data file to retain the metabolites that are consistently detected in a certain percentage of the runs.
Statistical analysis

Figure 1. Workflow for processing CIL LC-MS data that incorporates the zero-fill program.

from a multiple-run data set, even in replicate runs of the same sample. As an example, Figure 2A shows a distribution of the number of peak pairs found in ^{12}C-/^{13}C-dansyl labeled human urine samples (i.e., experimental triplicate runs of the same urine). Among the 1549 peak pairs found in run 1 and run 2, 960 pairs or 62% are in common. Comparing run 1 and run 3, 944 out of 1540 pairs (61%) are in common. There are 971 common pairs out of 1516 pairs (64%) found in run 2 and run 3. As the sample number increases, the number of commonly detected metabolites decreases (see below). In metabolomics work, it is common to use a criterion such as the 50% rule to retain the metabolites with missing intensity values in no more than 50% of the samples for statistical analysis. Currently, there is no consensus on what this percentage limit should be.[5,6,22]

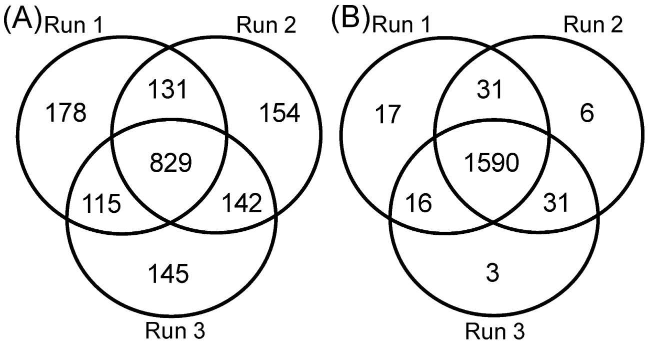


Figure 2. Venn diagrams of the number of peak pairs detected from experimental triplicate analysis of ^{13}C-/^{12}-C-dansyl labeled human urine samples: (A) without zero-fill and (B) with zero-fill.

Missing values in replicate runs are mainly caused by technical and data processing limitations. To reduce the number of missing values, measurement should be done using a technique that gives very high reproducibility. However, even for a very reproducible technique, data processing can be the limiting factor. In processing LC-MS data (with or without CIL), because of the need to balance the sensitivity and specificity in peak picking and intensity measurement, some low-abundance peaks or other peaks not meeting a set of criteria in the peak picking algorithm are missing in the metabolite-intensity table. Reanalyzing the original LC-MS data may help fill in the missing values in the table. This can be done manually by inspecting the original spectrum or chromatogram. Because this is a time-consuming process, manually filling the missing values is best done for selected metabolites that have already been found to be significant in statistical analysis of the initial metabolite-intensity data file. However, this approach will not alter the initial metabolite-intensity table used to perform statistical analysis for finding the significant metabolites in the first place. Alternatively, an algorithm may be developed to automate the reanalysis process to detect and fill the missing values (i.e., zero-fill). However, this is not easy to implement due to the fact that it is often difficult to differentiate the metabolite peaks from the background peaks when the signal intensity is very low, even with a high resolution instrument. Solvents, impurities, salts, etc., and their multimers and clusters can produce many peaks at the low mass region (m/z < 300) where a large portion of metabolite ions are detected.

Zero-Fill Program. CIL LC-MS offers an opportunity to overcome the difficulty of implementing an automated zero-fill process. In CIL LC-MS, the metabolite ion mass is shifted to a higher mass (m/z > 300) by adding the labeling group (e.g., +234.0583 Da for a dansyl labeled metabolite). This reduces the extent of background interference. More importantly, all the metabolite peaks in differential CIL LC-MS are detected in pairs and thus can be distinguished from the singlet background peaks. In addition, a ^{13}C-labeled control sample is spiked to all ^{12}C-labeled individual samples. As a consequence, the absolute intensity of the ^{13}C-peak of a metabolite peak-pair should be similar for all the samples, providing another differentiator. We have developed a zero-fill program to reanalyze the CIL MS data after the initial generation of the metabolite-intensity data file by IsoMS.

As Figure 1 shows, the zero-fill program first reads the metabolite-intensity file and then looks for missing values starting from the first sample run. Once a missing value is found, it goes back to the raw MS file. On the basis of matching retention time, m/z, and ^{13}C-labeled-peak intensity of the missing-value peak-pair with those in the raw MS file, the program finds the correct ^{13}C-peak. In the case that the ^{13}C-peak is not present in the raw data, the program stops the

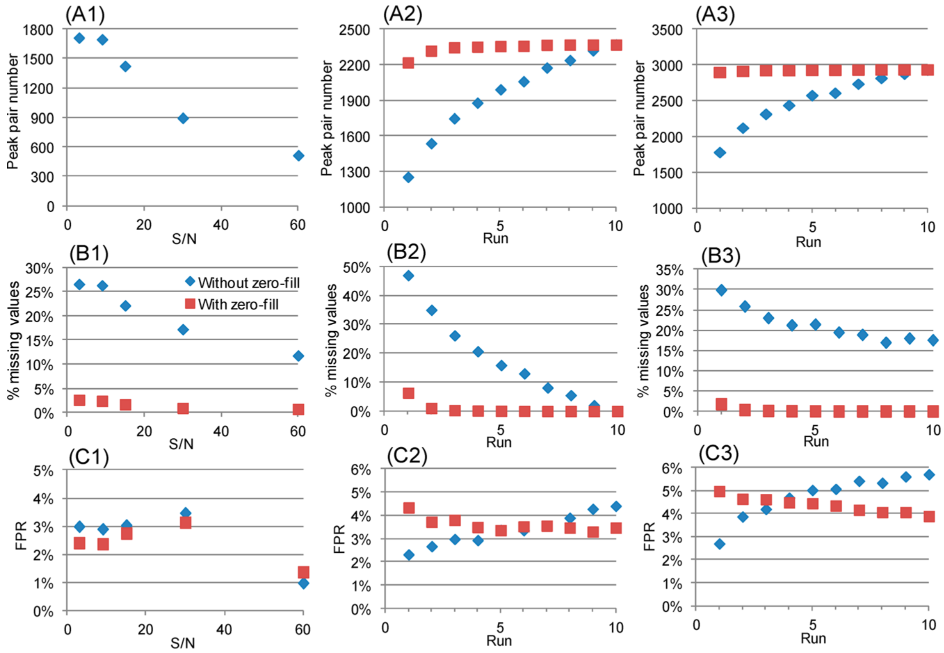


Figure 3. (A) Number of peak pairs detected, (B) percentage of missing values, and (C) FPR as a function of S/N used for IsoMS data processing of (1) the experimental triplicate data set of labeled urine, (2) the 10-run replicate injection data set, and (3) the 30-run data set.

search for the corresponding ^{12}C-labeled peak to avoid generating any false positive result. If the ^{13}C-peak exists in the raw data, the program would go ahead and search for the ^{12}C-peak based on retention time, m/z, and intensity as well as the fact that the ^{12}C-peak must exist in the same mass spectrum as the ^{13}C-peak. Once both peaks are picked, the zero-fill program calculates the peak intensity ratio. This ratio is entered into the metabolite-intensity table to replace the missing value. To distinguish the ratios determined by IsoMS and zero-fill, 2 and 9 decimal places are kept for the ratios from IsoMS and zero-fill, respectively. This helps manual checking if needed.

Performance of Zero-Fill. We have systematically evaluated the performance of the zero-fill program with an objective of extracting a maximal number of peak pairs from a multiple-run data set within an acceptable level of FPR (i.e., <5%). In the workflow shown in Figure 1, IsoMS is first used to process the data set using a chosen S/N threshold for extracting the peak pairs. The value of this threshold has a large effect on the number of peak pairs picked by this program. Figure 3(A1—C1) shows the total number of level 1 peak pairs, the percentage of missing values, and FPR as a function of threshold value used for peak-pair picking (i.e., S/N 3, 9, 15, 30 and 60). These results were obtained from an experimental triplicate data set of dansyl labeled human urine. Figure 3(A1) shows an overall decrease in the peak pair number as the S/N threshold increases. The FPR level (see Figure 3(C1), without zero-fill) does not change significantly except that it is lower at the threshold of S/N 60 from which only the very high abundance peaks are picked. These results indicate that IsoMS is able to pick the level 1 peak pairs with FPR of <4% even at a very low threshold (S/N 3). However, the numbers of peak pairs detected using S/N 3 and 9 thresholds are similar, suggesting that lowering the threshold from 9 to 3 cannot increase the peak pair number anymore. Manual inspection of the results indicates that many of the peak pairs with S/N < 9 are not belonging to the level 1 group. The plot in Figure 3(B1) (without zero-fill) shows that the percentage of missing values in each run decreases as the threshold increases. This is consistent with the notion that the high abundance peaks are more reproducible. Considering that the performance of using S/N 9 is similar to that of S/N 3 and IsoMS data processing is faster with S/N 9 (i.e., 5 min per run using S/N 9 vs 20 min per run using S/N 3), we choose a threshold of S/N 9 to carry out the IsoMS data processing to generate the initial metabolite-intensity data file.

Applying the zero-fill program to reanalyze the triplicate data set, the percentage of missing values drops dramatically from 26.4% to 2.5%. This can be more clearly seen in Figure 2B where the distribution of the number of peak pairs found in the three runs is shown. The common peak pairs found in the three runs increases from 829 (48.9%) to 1590 (93.9%). The average run-to-run reproducibility determined from the individual values of 97.51% (run 1 vs run 3; 1606/1647), 97.51% (run 1 vs run 2), and 98.30% (run 2 vs run 3) is 98%, compared to 67% ± 1% without using zero-fill. Many of the retrieved values can be manually confirmed by inspecting the peak pairs in the raw mass spectra. In fact, with zero-fill, the FDR drops from 2.9% to 2.4% (see Figure 3(C1) at S/N 9). Thus, the zero-fill program can retrieve missing values from the raw data very effectively.

We have studied the performance of zero-fill in a data set containing 10 replicate injections of the same dansyl urine sample. Figure 3(A2) shows the number of peak pairs detected with and without zero-fill as a function of cumulative injection number. Without zero-fill, the cumulative number of peak pairs increases gradually and then reaches a near-plateau after 9 injections. The percentage of missing values also gradually reduces as more replicate data are included in the combined runs (Figure 3(B2)). However, with zero-fill, both the total number of peak pairs detected and the percentage of missing values reach the plateau much faster. In fact, the results of duplicate injections with zero-fill are similar to those of 9 or 10 injections without zero-fill (see Figure 3(A2)). Even using one injection, 2217 peak pairs can be detected, compared to 2368 peak pairs from the combined results of duplicate injections. As Figure 3(C2) shows, with zero-fill, the FPR decreases as more replicate data are included, while without zero-fill, the FPR increases.

We have also analyzed the performance of zero-fill on a data set of 30 LC-MS runs from experimental triplicate of dansyl labeled samples with 10 injections for each sample. The results of experimental triplicate measure the overall experimental variations, not just instrumental variation which is gauged by repeat injections of the same sample. Figure 3(A3−C3) shows the plots where the *x*-axis represents the injection number. The combined results of three triplicate samples from each injection are used. For example, for injection 1, the total number of peak pairs detected in the three samples from the first injection is used (i.e., three LC-MS runs). For injection 2, the combined total number of peak pairs detected in the three samples from the first and second injections is plotted (i.e., 6 LC-MS runs). As Figure 3(A3−C3) shows, the trends of changes in the number of peak pairs, percentage of missing values, and FPR are similar to the injection replicate data set shown in Figure 3(A2−C2). However, in the experimental triplicate results, even after 10 replicate runs for each sample, there are still about 17% missing values (∼510 peak pairs) if zero-fill is not performed (see Figure 3(B3)). These are the peak pairs with variations caused by the sample handling process. For example, some low abundance metabolites might be labeled with slightly different efficiencies in the triplicate samples, which can result in signal intensity reduction in one of the ^{13}C-natural-isotope peaks to a level that the peak pair is no longer belonging to the level 1 group. In contrast, with zero-fill, the percentage of missing values drops much faster and reaches almost zero after two injections of each sample. Even with one injection, most of the peak pairs from the combined results are detected (see Figure 3(B3)).

The above results indicate that with zero-fill the number of peak pairs detected in each run can reach a near-maximal number even without performing replicate runs for each sample. However, the maximal number of peak pairs detectable within a data set is dependent on the number of runs present in the data set. Comparing the total maximal number of peak pairs detected in the 3-run data set (Figure 3(A1)) to the 10-run data set (Figure 3(A2)) and the 30-run data set (Figure 3(A3)), it is clear that the maximal number increases as the number of LC-MS runs increases. This is understandable considering the fact that each run adds some unique peak pairs to the total. However, there appears to be a diminished return as the number of runs increases beyond a certain value. For example, using 10 runs, instead of 3 runs, the peak pair number increases from 1700 to 2350 (i.e., 38% with a net gain of 650 pairs). However, using 30 runs, instead of 10 runs, the pair number increases from 2350 to 2900 (i.e., 23% with a net gain of only 550 pairs). Thus, performing replicates merely for the purpose of increasing the peak pair number in a data set needs to be considered within the context of instrumental time available. In a clinical metabolomics study involving the profiling of hundreds of samples, one may choose not to perform replicate runs in order to save instrument time. On the other hand, for a cellular metabolomics work where only a few samples are profiled, it may be well justified to perform replicate runs. In any case, with zero-fill, we can recover the missing values in a data set very effectively and efficiently.

It should be noted that the zero-fill program reported in this work was specifically developed for CIL LC-MS where a true metabolite is detected as a peak pair, not a singlet peak. In traditional LC-MS without differential isotope labeling, a low abundance metabolite peak initially not picked by data processing software in a mass spectrum is very difficult to be differentiated from the noise and chemical background peaks. Picking any peak of similar *m*/*z* and retention time for zero-fill would run into a risk of increasing FPR.

Characterization of Missing Values. As indicated earlier, the source of missing values in a replicate run data set is mainly from the measurement and data processing processes which can be influenced much more by the low abundance peaks than the high abundance ones. We have characterized the missing values in terms of signal intensity in the 10-run data set. While peak ratio is used to measure the relative concentration in CIL LC-MS, the absolute intensity of a peak is related to abundance and detection sensitivity of the metabolite. It should be noted that detection sensitivity of different metabolites becomes more uniform after dansylation labeling. For example, the difference in MS signal intensity for 17 dansyl amino acid standards is within 1 order of magnitude, compared to more than 3 orders of magnitude for unlabeled amino acids.[27] Thus, the absolute intensity of labeled metabolites is a good indication of analyte abundance in a sample. Figure 4 shows a histogram of the peak pair distribution as a function of the absolute intensity measured by S/N. The S/N values are binned in $\log_9$ to distribute the number of peak pairs found in each bin evenly across the *x*-axis. In the low S/N bins, there are significantly more pairs detected with zero-fill. For example, at S/N around 9−15 (i.e., 1-1.25 in $\log_9$), about 300 pairs are detected with zero-fill, compared to 150 pairs detected without zero-fill. In the high S/N bins, the number of peak pairs found with and without zero-fill is similar. Thus, the zero-fill process recovers mainly the low intensity or low abundance metabolites that fail to detect in the first path of data analysis by IsoMS.

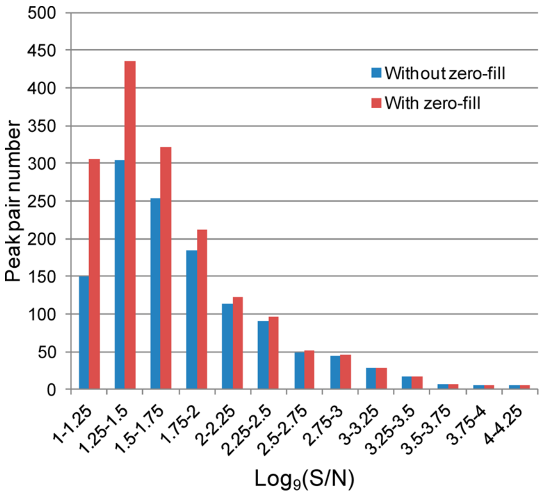


Figure 4. Number of peak pairs as a function of $\log_9$(S/N).

In terms of the reproducibility of peak ratio values in the 10-run data set, the median and average CVs for the data set without zero-fill were found to be 8% and 12%, respectively, compared to 12% and 13% with zero-fill. For the filled values alone, the median and average CVs were 14% and 15%. Thus, zero-fill only resulted in a very small reduction in reproducibility for relative quantification of metabolites.

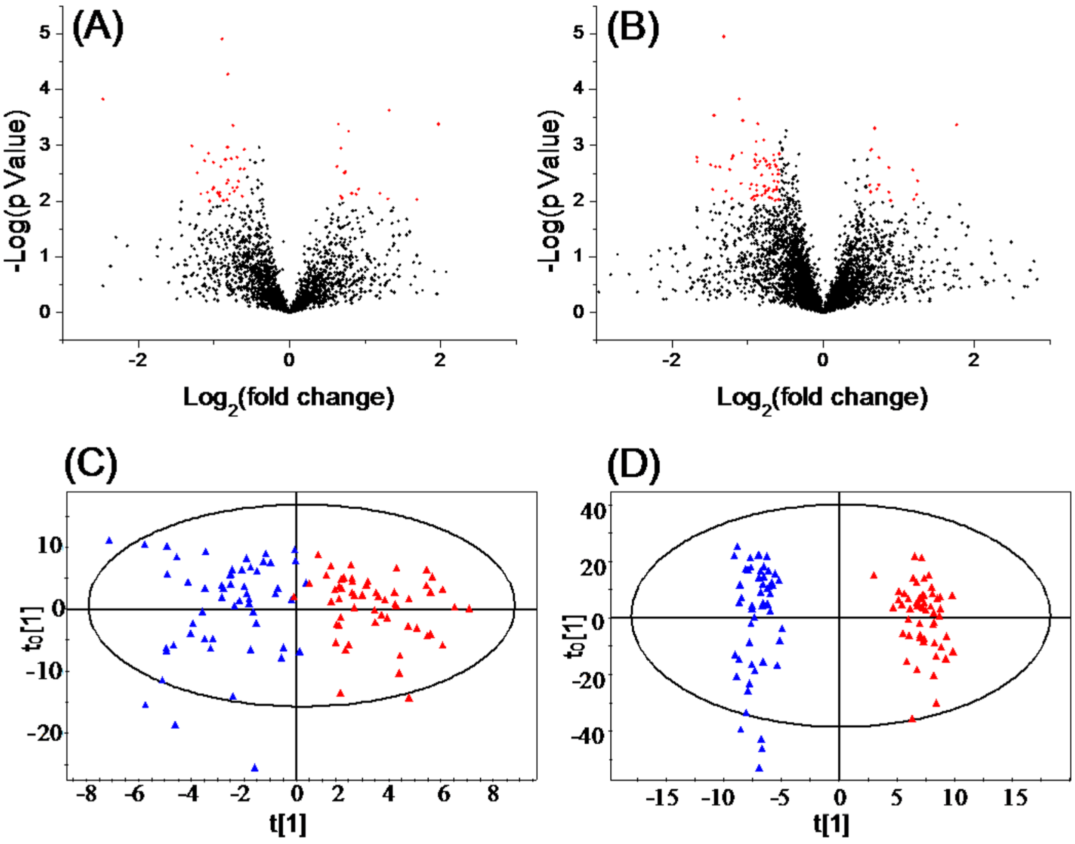


Figure 5. Volcano plots of the 109-sample data set from a bladder cancer biomarker discovery study: (A) without zero-fill and (B) with zero-fill. The red dots represent a metabolite with a fold change of ≥1.5 and p-value of ≤0.01. OPLS-DA plots of the 109-sample data set: (C) without zero-fill and (D) with zero-fill.

Standardization of Counting Missing Values. Because missing values are mainly from the low abundance peaks which are more difficult to detect reproducibly, the extent of missing values in a data set should be a good indicator to judge the overall analytical performance of a metabolome profiling method. We propose to use an experimental triplicate data set (e.g., the data shown in Figure 2) and a 10-run injection replicate data set (e.g., the data shown in Figure 3(A2)) of the same sample to measure the performance of a method regarding the missing values. Although using data of different samples would have the benefit of evaluating how well a method quantifies the same metabolites of different concentrations in different samples, it requires a set of standard samples available for method evaluation. Replicate data of the same sample is readily generated in a lab. Using the same type of sample (e.g., human urine), the performance of different methods in terms of missing values can still be compared, at least within the context of performing a metabolomics study using this type of sample. Recent development of standard samples such as NIST serum standard should facilitate future work of comparing different methods, if such a standard is used across different platforms and methods.[37]

Using the replicate data set, we propose that the performance indicators be (1) number of peak pairs detected per run and the total number of peak pairs detected within a data set (triplicate or 10-run replicate), (2) intensity dynamic range from the lowest absolute signal intensity giving a quantity result to the highest absolute intensity giving a quantity result, and (3) number of missing values and percentage of missing values in triplicate and 10-run replicate data sets. Supplemental Tables T1 and T2, Supporting Information, show the summary of the results for the triplicate and 10-run data sets obtained by the dansylation CIL LC-MS method, respectively.

Metabolomics Application. Finally, we have applied the zero-fill program in a metabolomics study to demonstrate the benefits of using zero-fill for disease biomarker discovery. In this case, we applied zero-fill to a set of LC-MS data generated from a human bladder cancer metabolomics study.[38] It consists of 109 LC-MS runs of dansyl labeled urine samples collected from 55 bladder cancer patients and 54 controls. Individual samples were separately labeled with ^{12}C-danylation and then mixed with ^{13}C-dansylated universal metabolome-standard of human urine. The individual ^{13}C-/^{12}C-labeled mixtures were separated and analyzed using reversed-phase LC and a Bruker 9.4-T Fourier transform ion cyclotron resonance mass spectrometer.[38] Supplemental Table T3, Supporting Information, shows the original metabolite-intensity table generated using IsoMS from the 109 runs. Supplemental Table T4, Supporting Information, shows the table after applying the zero-fill program to the data set. The volcano and Orthogonal Projections to Latent Structures-Discriminant Analysis (OPLS-DA) plots of the data sets with and without zero-fill are shown in Figure 5.

As Figure 5A,B shows, more significant metabolites (in red) are detected in the volcano plot of the zero-filled data. There are 81 metabolites with a fold change of ≥1.5 and p-value of ≤0.01 in the data set with zero-fill, compared to 65 metabolites without zero-fill. A similar observation is found in the OPLS-DA analysis. There are 385 significant metabolites (VIP score of ≥1.5) found from the zero-filled data, compared to 53 metabolites without zero-fill. Supplemental Tables T5−T8, Supporting Information, list the significant metabolites including 24 metabolites that were positively identified using a dansyl standard library consisting of 280 compounds by comparing the accurate mass and retention time of an unknown to those of the library compounds. Many other metabolites were putatively identified on the basis of accurate mass match against the Human Metabolome Database (HMDB) and the Evidence-Based Metabolome Library (EML) by using the MyCompoundID MS Search program.[36] As Figure 5C,D shows, a much better separation of the cancer and control groups is obtained with the zero-filled data (without zero-fill: $R^2X = 0.389$, $R^2Y = 0.745$, $Q^2 = 0.562$; with zero-fill: $R^2X = 0.366$, $R^2Y = 0.972$, $Q^2 = 0.621$).

To determine if there is any enhancement in discriminating power of individual metabolites for the separation of two groups, the top 50 metabolites ranked by fold change in the zero-filled data set were examined. The median and average fold changes are 1.55 and 1.78, respectively, compared to 1.26 and 1.35 for the same 50 metabolites found in the data set before applying zero-fill. The average p-value is 0.001 with a median of 0.01 for the zero-filled data set, compared to 0.14 with a median of 0.09 for the no-zero-filled data set. Thus, both fold changes and p-values are significantly improved after zero-fill.

The above results clearly show a significant improvement of the quality of statistical analysis after applying zero-fill to the 109-sample data set, enabling the detection of more and better discriminating metabolites to differentiate two cohorts of samples. To measure the quality of the metabolite-intensity data in terms of missing values, we plot the percentage of common peak pairs detectable in cumulative samples as a function of sample runs in a data set (see Figure 6). This plot is informative for determining the consistency of metabolite detection among all the runs. For example, 2858 peak pairs or about 60% of the total number of peak pairs found in the zero-filled data set (4761) can be consistently quantified in half of the samples (109/2), while without zero-fill only 395 or 8.3% of the total (4761) are commonly detected. In our view, this type of plot should be presented, along with the metabolite-intensity table, when reporting the metabolome profiling data in a metabolomics study. This would assist in judging the overall coverage of the metabolomic profiles in a study.

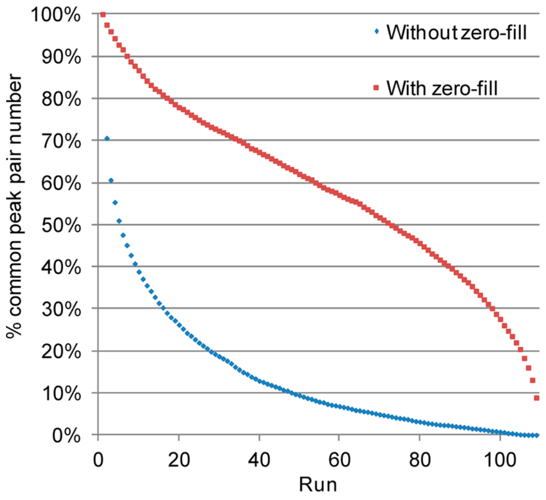


Figure 6. Percentage of common peak pairs detected in cumulative runs as a function of sample runs. The total number of pairs detected from 109 runs is 4761.

CONCLUSIONS

We report a detailed study on the issue of missing values in a chemical isotope labeling LC-MS metabolomics platform. A zero-fill program has been developed to retrieve missing values in the initial metabolite-intensity table generated by IsoMS. Missing values were found to be mainly from the low intensity metabolites. The zero-fill program allows significant reduction in missing values. This reduction affords the detection of more and better discriminating metabolites in a metabolomics study involving the metabolomic profiling of 109 samples for bladder cancer biomarker discovery.

Because the extent of missing values can have a profound effect on metabolomics results, we feel that counting missing values should be considered as one of the important metrics for measuring the analytical performance of a metabolomics platform. To facilitate method comparison, we proposed the use of two data sets, one from experimental triplicate and another one from 10 replicate injections of the same sample, to measure the extent of missing values. Finally, in reporting metabolomics data, we feel that it is important to include a summary of missing value analysis (e.g., a plot of number or percentage of common metabolites detected in cumulative samples as a function of sample runs). This analysis result, along with the metabolite-intensity table, measures the level of commonly quantifiable metabolites in a metabolomics study. At a chosen % threshold (e.g., metabolites commonly quantifiable in more than 50% of all the samples), the number of metabolites retained for statistical analysis should be reported. In this regard, future work is still needed to examine the issue of selecting the most appropriate % threshold for data inclusion in statistical analysis.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: liang.li@ualberta.ca.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Canadian Institutes of Health Research, the Canada Research Chairs program, Genome Canada, and Alberta Innovates. We thank Dr. Jun Peng and Dr. Yi-Ting Chen for generating the LC-MS data set of the bladder cancer study.

REFERENCES

(1) Liew, A. W. C.; Law, N. F.; Yan, H. *Briefings Bioinf.* **2011**, *12*, 498−513.

(2) Albrecht, D.; Kniemeyer, O.; Brakhage, A. A.; Guthke, R. *Proteomics* **2010**, *10*, 1202−1211.

(3) Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D. *BMC Bioinf.* **2012**, *13*, 9.

(4) Bijlsma, S.; Bobeldijk, L.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K. *Anal. Chem.* **2006**, *78*, 567−574.

(5) Hrydziuszko, O.; Viant, M. R. *Metabolomics* **2012**, *8*, S161−S174.

(6) Gromski, P. S.; Xu, Y.; Kotze, H. L.; Correa, E.; Ellis, D. I.; Armitage, E. G.; Turner, M. L.; Goodacre, R. *Metabolites* **2014**, *4*, 433−452.

(7) Little, R.; Rubin, B. *Statistical Analysis with Missing Data*; Wiley: Hoboken, NJ, 2002.

(8) Anderle, M.; Roy, S.; Lin, H.; Becker, C.; Joho, K. *Bioinformatics* **2004**, *20*, 3575−3582.

(9) Torres-Garcia, W.; Brown, S. D.; Johnson, R. H.; Zhang, W. W.; Runger, G. C.; Meldrum, D. R. *Mol. BioSyst.* **2011**, *7*, 1093−1104.

(10) Valledor, L.; Jorrin, J. *J. Proteomics* **2011**, *74*, 1−18.

(11) Schwammle, V.; Leon, I. R.; Jensen, O. N. *J. Proteome Res.* **2013**, *12*, 3874−3883.

(12) Jung, K.; Dihazi, H.; Bibi, A.; Dihazi, G. H.; Beissbarth, T. *Bioinformatics* **2014**, *30*, 1424−1430.

(13) Koopmans, F.; Cornelisse, L. N.; Heskes, T.; Dijkstra, T. M. H. *J. Proteome Res.* **2014**, *13*, 3871−3880.

(14) Sangster, T. P.; Wingate, J. E.; Burton, L.; Teichert, F.; Wilson, I. D. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 2965−2970.

(15) Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M. *BMC Bioinf.* **2007**, *8*, 93.

(16) Dunn, W. B.; Broadhurst, D.; Brown, M.; Baker, P. N.; Redman, C. W. G.; Kenny, L. C.; Kell, D. B. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2008**, *871*, 288−298.

(17) Begley, P.; Francis-McIntyre, S.; Dunn, W. B.; Broadhurst, D. I.; Halsall, A.; Tseng, A.; Knowles, J.; Goodacre, R.; Kell, D. B.; Consortium, H. *Anal. Chem.* **2009**, *81*, 7038−7046.

(18) Veselkov, K. A.; Vingara, L. K.; Masson, P.; Robinette, S. L.; Want, E.; Li, J. V.; Barton, R. H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T. M.; Pelczer, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2011**, *83*, 5864−5872.

(19) Mattarucchi, E.; Guillou, C. *Anal. Chem.* **2011**, *83*, 9719−9720.

(20) Veselkov, K. A.; Vingara, L. K.; Masson, P.; Robinette, S. L.; Want, E.; Li, J. V.; Barton, R. H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T. M.; Pelczer, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2011**, *83*, 9721−9722.

(21) Mattarucchi, E.; Guillou, C. *Biomed. Chromatogr.* **2012**, *26*, 512−517.

(22) Mak, T. D.; Laiakis, E. C.; Goudarzi, M.; Fornace, A. J. *Anal. Chem.* **2014**, *86*, 506−513.

(23) Katajamaa, M.; Oresic, M. *BMC Bioinf.* **2005**, *6*, 12.

(24) Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L. *Chem. Soc. Rev.* **2010**, *40*, 387−426.

(25) Scalbert, A.; Brennan, L.; Fiehn, O.; Hankemeier, T.; Kristal, B. S.; van Ommen, B.; Pujos-Guillot, E.; Verheij, E.; Wishart, D.; Wopereis, S. *Metabolomics* **2009**, *5*, 435−458.

(26) Goodacre, R.; Broadhurst, D.; Smilde, A. K.; Kristal, B. S.; Baker, J. D.; Beger, R.; et al. *Metabolomics* **2007**, *3*, 231−241.

(27) Guo, K.; Li, L. *Anal. Chem.* **2009**, *81*, 3919−3932.

(28) Guo, K.; Li, L. *Anal. Chem.* **2010**, *82*, 8789−8793.

(29) Leng, J. P.; Wang, H. Y.; Zhang, L.; Zhang, J.; Wang, H.; Guo, Y. L. *Anal. Chim. Acta* **2013**, *758*, 114−121.

(30) Zhang, S. J.; You, J. M.; Ning, S. J.; Song, C. H.; Suo, Y. R. *J. Chromatogr., A* **2013**, *1280*, 84−91.

(31) Tayyari, F.; Gowda, G. A. N.; Gu, H. W.; Raftery, D. *Anal. Chem.* **2013**, *85*, 8715−8721.

(32) Mazzotti, F.; Benabdelkamel, H.; Di Donna, L.; Athanassopoulos, C. M.; Napoli, A.; Sindona, G. *J. Mass Spectrom.* **2012**, *47*, 932−939.

(33) Dai, W. D.; Huang, Q.; Yin, P. Y.; Li, J.; Zhou, J.; Kong, H. W.; Zhao, C. X.; Lu, X.; Xu, G. W. *Anal. Chem.* **2012**, *84*, 10245−10251.

(34) Zhou, R. K.; Guo, K.; Li, L. *Anal. Chem.* **2013**, *85*, 11532−11539.

(35) Zhou, R.; Tseng, C. L.; Huan, T.; Li, L. *Anal. Chem.* **2014**, *86*, 4675−4679.

(36) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.; Wishart, D. S.; Lin, G. *Anal. Chem.* **2013**, *85*, 3401−3408.

(37) Phinney, K. W.; Ballihaut, G.; Bedner, M.; Benford, B. S.; Camara, J. E.; Christopher, S. J.; et al. *Anal. Chem.* **2013**, *85*, 11732−11738.

(38) Peng, J.; Chen, Y. T.; Chen, C. L.; Li, L. *Anal. Chem.* **2014**, *86*, 6540−6547.