MAFFIN: Metabolomics Sample Normalization Using Maximal Density Fold Change with High-Quality Metabolic Features and Corrected Signal Intensities

Huaxu Yu¹, Tao Huan^{1,*}

¹ Department of Chemistry, Faculty of Science, University of British Columbia, Vancouver Campus, 2036 Main Mall, Vancouver, V6T 1Z1, BC, Canada

* Author to whom correspondence should be addressed:

Dr. Tao Huan

Tel: (+1)-604-822-4891

E-mail: thuan@chem.ubc.ca

Internet: <https://huan.chem.ubc.ca/>

Abstract

Sample normalization is a critical step in metabolomics to remove differences in total sample amount or concentration of metabolites between biological samples. Here, we present MAFFIN, an accurate and robust post-acquisition sample normalization workflow that works universally for metabolomics data collected by mass spectrometry (MS)-based platforms. The most important design of MAFFIN is the calculation of normalization factor using maximal density fold change (MDFC) value computed by a kernel density-based approach. MDFC is more accurate than traditional median FC-based normalization, especially when the numbers of up- and down-regulated metabolic features are different. In addition, we showcase two essential steps that are overlooked by conventional normalization methods, and incorporated them into MAFFIN. First, instead of using all detected metabolic features, MAFFIN automatically extracts and uses only the high-quality features to calculate FCs and determine the normalization factor. In particular, multiple orthogonal criteria are proposed to pick up the high-quality features. Second, to guarantee the accuracy of the FCs, the MS signal intensities of the high-quality features are corrected using serial quality control (QC) samples. Using simulated data and urine metabolomics datasets, we demonstrated the critical need of high-quality feature selection, MS signal correction, and MDFC. We also show the superior performance of MAFFIN over other commonly used post-acquisition sample normalization methods. Finally, a biological application on a human saliva metabolomics study shows that MAFFIN provides robust sample normalization, leading to better data separation in principal component analysis (PCA) and the identification of more significantly altered metabolic features.

Introduction

Liquid chromatography-mass spectrometry (LC-MS)-based untargeted metabolomics enables high-throughput and high-coverage detection and quantification of thousands of metabolic features in various biological sample types.¹⁻³ However, many biological samples types, such as urine, saliva, sweat, and fecal, have large differences in total sample amount or concentration of metabolites. For instance, urine samples can have a greater than 15-fold concentration difference because of water intake or other physiological factors.⁴ For human fecal samples, its water content can commonly change from 60% to 85% even within a week.^{5, 6} Besides this biology-induced dilution effect, inappropriate analytical procedure can also induce significant sample-to-sample variation. For example, the mass of a small amount of frozen tissue (e.g., less than 1 mg) can be significantly affected by the moisture absorbed from air during the weighing process.⁷ Likewise, cell counts performed before scraping cells off of a petri dish can be inaccurate if some cells are left behind.⁸ These analytical and biological variations cause different total sample amount between comparative groups, leading to larger variation and/or systematic biases in quantitative metabolomics comparisons.

The metabolomics community has recognized the importance of sample normalization. Various normalization strategies have been proposed to assure accurate and precise quantitative comparison and statistical analysis. These normalization methods can usually be classified into pre- or post-acquisition sample normalization.⁹ Pre-acquisition sample normalization is applied prior to MS data collection. It measures a certain quantity that reflects the total sample amount or concentrations of metabolites. The samples are then reconstituted to appropriate final volumes based on the measured quantities to make the total concentration consistent between the samples.

For example, creatinine levels generally reflect the urine concentration and is commonly used for normalizing urine samples.^{4, 10, 11} For cell samples, total DNA or protein concentration is expected to agree with the quantity of cells and thus are often used to normalize cell samples when cell counts are not accurate.⁸ Recent work has used dansylation labeling to quantify the total amount of amine- and phenol-containing metabolites for sample normalization.^{12, 13} The underlying assumption of that work is that the total amount of amine- and phenol-containing metabolites is a good reflection of the total metabolome, and it should be consistent in a biological sample type that is involved in a metabolomics study. In general, pre-acquisition sample normalization is often favored over post-acquisition sample normalization. One key benefit of pre-acquisition sample normalization is that researchers can estimate the total amount of metabolites and then optimize sample loading amount on the LC-MS system for better analytical performance.¹⁴ In addition, pre-acquisition sample normalization also enables comparisons across different sample batches and sample quality checks. As such, pre-acquisition sample normalization is preferred if a proper quantity that indicates the total metabolome concentration is available. Unfortunately, for many biological sample types, such quantity either does not always work or does not exist. For instance, the urine creatinine normalization was reported to have bias, especially when glomerular filtration rate is changed due to kidney injury¹⁰. On the other side, many samples types, including but not limited to saliva, sebum, and sweat, still lack reliable quantities for pre-acquisition sample normalization.

Post-acquisition sample normalization is an alternative strategy that complements and addresses the limitations of pre-acquisition sample normalization. It is a data driven approach that adjusts the MS signal intensities of the collected data based on a certain criterion that is considered to be

unchanged across samples. Commonly used post-acquisition normalization methods include the use of mass spectrum total useful signal (MSTUS) or total useful peak area (TUPA), quantile normalization, probabilistic quotient normalization (PQN), and median fold change.¹⁵⁻¹⁹ However, these methods perform inadequately due to their inherent limitations.⁹ In particular, problem with total peak intensity, problem with median fold change. In this work, we developed MAFFIN, a universal post-acquisition normalization method that is adaptable to any sample type. Using both simulated and real metabolomics data sets, we demonstrated that the MAFFIN workflow outperforms four other normalization methods, including sum normalization, median normalization, probabilistic quotient normalization, and quantile normalization, in reducing the biological variation. The normalized data can largely improve the confidence of downstream statistical analysis and biological interpretations.

100

Methods

Chemicals. Methanol (MeOH), water (H₂O), acetonitrile (ACN), sodium formate (NaFA), and ammonium acetate (all LC–MS-grade) were purchased from Thermo Fisher Scientific (Waltham, MA). Ammonium hydroxide was obtained from MilliporeSigma (Burlington, MA).

Urine and Saliva Sample Preparation

The human urine and saliva samples were collected from four healthy volunteers, including 2 males and 2 females. For each participant, 5 samples were collected at random time points to mimic different sample concentrations. The collected biospecimens were stored at -80 °C prior to the metabolomics experiment. For metabolome extraction, 100 µL of urine was mixed with 1 mL of ice-cold methanol in a 1.5 mL Eppendorf vial and vortexed. The solution was incubated at -20 °C for 2 hrs, followed by centrifugation at 14,000 rpm for 15 min at 4 °C. The clear supernatant was dried in a SpeedVac at 20 °C and then reconstituted in 150 µL ACN and H₂O (1:1, v:v) mixed solvent for LC-MS analysis. The method blank was also prepared following the same protocol but without urine. The QC sample was prepared by pooling 20 µL aliquots from each sample. The saliva samples were prepared following the same workflow.

LC-MS Analysis. LC-MS analysis was performed on a UHR-QqTOF (ultra-high resolution Qq-time-of-flight) mass spectrometer Impact II (Bruker Daltonics, Bremen, Germany) coupled with a 1290 Infinity II LC system (Agilent Technologies, Palo Alto, CA). Hydrophilic interaction chromatography (HILIC) separation was performed on a SeQuant ZIC-pHILIC column (150 mm × 2.1 mm, 5 µm, 200 Å). Mobile phase A was 5% ACN in H₂O with 10 mM ammonium acetate (pH = 9.8, adjusted by ammonium hydroxide), and mobile phase B was 5% H₂O in ACN with no

buffer. The elution gradient was set as: 0 min, 95% B; 2 min, 95% B; 20 min, 5% B; 23 min, 5% B; 24 min, 95% B; 35 min, 95% B. The column temperature was 35 °C, and flow rate was 0.15 mL/min. The injection volumes were optimized to 1 µL and 6 µL for the urine and saliva samples, respectively. To correct the signal intensities, a series of QC samples were injected with different loading amounts in three technical replicates. For urine samples, the serial QC samples were injected at 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0 µL (11 volumes in total). For saliva samples, the serial QC samples were injected at 0.6, 1.2, 2.4, 3.6, 4.8, 6.0, 7.2, and 8.4 µL (8 volumes in total). The MS data was collected in negative ion and data-dependent acquisition mode. The MS settings were: capillary voltage, 3.0 kV; nebulizer gas, 1.6 bar; dry gas, 7 L/min; dry gas temperature, 220 °C; mass scan range, 65-1500 (*m/z*); spectra rate, 8.00 Hz; and cycle time, 3.0 s. For centroid spectra calculation, the peak summation width was 3 pts. The mass spectrometer was calibrated using sodium formate.

Data Processing and Statistical Analysis.

The raw LC-MS data was first converted to abf files using Reifycs Abf Converter (ver. 4.0.0). The converted files were processed in MS-DIAL (ver. 4.70) for peak detection, peak alignment, peak area integration, and metabolite annotation.²⁰ The detailed settings of MS-DIAL are available in SI Text S-1. Further metabolite annotation was performed on MS-FINDER (ver. 3.52).²¹ The feature annotation results were manually verified using HMDB.²²

To assess the intragroup variation, pooled relative standard deviation (PRSD) and pooled relative median absolute deviation (PRMAD) were calculated. In brief, when there are *m* biological groups and each group has at least two replicates, we use X_i to denote the intensity vector of a feature in

group t and SD_{Xt} to denote the standard deviation of intensities in group t . The n_{Xt} represents the number of samples within group t . The PRSD was calculated as Eq. 1

$$PRSD = \sqrt{\frac{(n_{X1}-1)SD_{X1}^2 + (n_{X2}-1)SD_{X2}^2 + \dots + (n_{Xn}-1)SD_{Xn}^2}{n_{X1} + n_{X2} + \dots + n_{Xn} - m}} \quad (1)$$

The PRMAD was calculated as Eq. 2

$$PRMAD = \sqrt{\frac{(n_{X1}-1)MAD_{X1}^2 + (n_{X2}-1)MAD_{X2}^2 + \dots + (n_{Xn}-1)MAD_{Xn}^2}{n_{X1} + n_{X2} + \dots + n_{Xn} - m}} \quad (2)$$

where MAD_{Xt} denotes the median absolute deviation of intensities in group t .

In addition, the relative log abundance (RLA) plot was used to compare the data before and after normalization.²³ The RLA plot was made using R package *metabolomics* (ver. 0.1.4). Principal component analysis (PCA) was performed on MetaboAnalyst 5.0.²⁴ Log-transformation and autoscaling was used to pretreat the data before PCA analysis. To select the significantly altered metabolic features, the F -test was first applied to each feature to compare the group variance (unequal variance: $p < 0.05$). Two-sided unpaired Student's t test was performed, and the F -test result was used to determine whether equal or unequal variance should be used. The p values from the t test were further corrected using Benjamini & Hochberg correction method. One-sided Mann–Whitney U test was used to compare PRSD and PRMAD between different normalization methods. The p value correction and U test was performed in R using package *stats* (ver. 4.1.0).

High-Quality Feature Selection

MAFFIN extracts high-quality metabolic features using the following criteria: (1) their averaged intensities in biological samples should be higher than their intensities in the method blank (default:

sample average intensity $> 2 \times$ blank intensity); (2) eluted during the analytical gradient; (3) low relative standard deviation (RSDs) of signal intensities in the replicate QC injections (default: RSDs $< 25\%$, according to FDA's guideline²⁵); (4) their MS intensities are highly correlated with their loading amount in serial QC samples (default: Pearson correlation > 0.9). The high-quality metabolic features can be further refined by examining their chromatographic peak shapes and grouping degenerate features, which are features arising from the same metabolite.²⁶ In this work, chromatographic peak shapes were manually checked, and degenerate features were grouped using the MS-DIAL curation result (SI Text S-2).

MS Signal Intensity Correction

MS signal intensity correction was performed to convert the MS signal intensity to QC loading amount. In brief, the serial diluted QC samples with different loading amounts are first analyzed alongside biological samples. The feature intensities will be corrected if they have enough data points in serial QC samples and their MS intensities are highly correlated with serial QC loading amounts (Pearson correlation > 0.9). For each feature fulfilling these requirements, a calibration curve was built using the serial diluted QC data, and the signal intensities of the metabolic feature in the biological samples were then converted to QC loading amounts. The R code for data processing was adapted from a previous publication.²⁷

Maximal Density Fold Change

The maximal density FC (MDFC) between two samples is calculated from the FC distribution. First, the FCs of all high-quality features are calculated and log2 transformed. A kernel density function is used to estimate the probability density of the transformed FCs. The kernel density

computation was achieved using the kernel “gaussian” in the *stats* R package. In particular, the bandwidth is critical to kernel density calculation, which determines the impact of each data point on the estimated density function. We proposed a method to automatically optimize the bandwidth to ensure normalization accuracy. That method iterates over a range of bandwidths and chooses the bandwidth that achieves the best sample normalization result (detailed in **SI Text S-3**). For the simulated data set where only two samples are involved, the bandwidth was optimized using “SJ” method embedded in the *density()* R function.²⁸

Simulated Data Set

A simulated data set was generated to demonstrate that MDFC can accurately estimate the normalization factor (k) between two samples, even when the percentage of up-regulated features is different from that of down-regulated features (i.e., imbalanced data). To initiate the simulation, three parameters were required, including a normalization k , percentage of unchanged features, and percentage of up-regulated features (the percentage of down-regulated features can then be calculated). Two samples were simulated with 5,000 features in total. For the unchanged features, their intensities in sample 1 and sample 2 were generated from the same normal distribution. In this case, the intensities of all features in sample 1 (IS_1) have

$$IS_1 \sim N(\mu, \sigma^2)$$

The intensities of unchanged features in sample 2 (IS_{2-un}) have

$$IS_{2-un} \sim N(k\mu, \sigma^2)$$

For the up- or down-regulated features, their intensities in sample 1 and sample 2 were generated from two normal distributions with a certain FC of mean values (f). Compared to sample 1, the intensities of up- (IS_{2-up}) and down- (IS_{2-down}) regulated features in sample 2 have

$$IS_{2-up}, IS_{2-down} \sim N(k\mu f, \sigma^2)$$

where f is higher than 3/2 when up-regulated, and lower than 2/3 when down-regulated. To obtain a comprehensive evaluation, we tested different percentages of unchanged and up-regulated features. The percentage of unchanged features varies from 30 to 100% in steps of 1%. The percentage of up-regulated features varies from 0 to 100% of the unchanged features in steps of 1%.

Comparison to Other Normalization Methods

MAFFIN workflow was compared to four commonly used post-acquisition sample normalization methods, including sum normalization, median normalization, PQN normalization, and quantile normalization. These four normalization methods were performed on MetaboAnalyst 5.0.²⁴

Results and Discussion

MAFFIN Workflow

The schematic workflow of MAFFIN is shown in **Figure 1**. To perform MAFFIN, a set of serial diluted QC samples (or QC samples with different loading amounts) need to be analyzed alongside the real metabolomics samples. The input of MAFFIN is a metabolic feature intensity table, which is generated by conventional metabolomics data processing, including metabolic feature extraction and retention time alignment.^{20, 29} Before performing MDFC-based sample normalization, two data preprocessing steps are needed. First, high-quality features were selected because many detected features are not real metabolites or metabolites with poor quantification performance. These low-quality features include but are not limited to features that: 1) are background noise; 2) have low technical reproducibility in LC-MS analysis; 3) have low Pearson correlation between sample loading amount and measured signal intensity. These metabolic features do not represent real metabolites in biological samples. Therefore, their intensity values should not be considered in sample normalization. However, in conventional post-acquisition sample normalization methods, these low-quality features are not excluded from calculating normalization factors, which might cause biased sample normalization. It is important to note that the concept of choosing high-quality metabolic features for sample normalization should also be applicable in nuclear magnetic resonance (NMR)-based metabolomics.

Another important feature of MAFFIN is to correct MS signal intensities before FC calculation. Our previous works show that the MS signal FCs can be inflated or compressed for most metabolic features even in the linear response range of electrospray ionization.^{27, 30} In MAFFIN, we adapted the previously published MRC workflow to convert MS signals to QC loading amounts.²⁷ The FCs

calculated using QC loading amounts represent the real concentration FCs and benefit the following estimation of maximal density FC. It is important to note that both high-quality feature extraction and MS signal intensity correction are optional in the MAFFIN workflow, but we highly recommend these two steps, as they can significantly improve the normalization accuracy.

After high-quality metabolic feature selection and MS signal intensity correction, the refined intensity table is used to calculate MDFC for normalization. Its fundamental theory is that for two samples, the unchanged features should have very similar FCs due to proportional total sample amount or concentration differences, while the up- or down-regulated features have different FCs. If we calculate the density of FC distribution of all features, the FC at the highest density (i.e., MDFC) will be caused by unchanged features, thus represents the sample concentration ratio. To calculate MDFC, the sample with the most features detected is first selected as the reference sample. FCs of metabolic feature intensities are then calculated between the test sample and the reference sample. Next, the FCs are log-transformed to make them symmetric around the nonchanged metabolites. After that, Gaussian kernels are used to estimate a density function of the FC distribution, and the FC at the maximal density is used as the normalization factor. Compared to histogram-based discrete estimation, kernel density simulation is continuous and thus has a more accurate estimation of normalization factor.

Selection of High-Quality Metabolic Features

We first demonstrated the importance of selecting high-quality metabolic features for post-acquisition sample normalization. First, we randomly picked two urine samples from the human urine data set and applied high-quality feature selection. As we can see in **Figure 2A**, the low-

quality features have a FC distribution distinct from that of the high-quality features. In particular, the MDFC of low-quality features is 1.34, while the MDFC of high-quality features is 2.05. This demonstrates that if the low-quality features are not excluded, the estimated normalization factor will be biased. Therefore, it is important to select the high-quality metabolic features for post-acquisition sample normalization.

We further evaluated the impact of high-quality feature selection on normalization using the complete urine metabolomics data set. In **Figure 2B**, the metabolic feature number decreased from 4042 to 1490 after all filters were applied. The drop in feature number indicates that the features with good analytical performance might only account for a small portion of all detected features, and we believe that only these features should be selected for sample normalization.³¹ After applying each filter, we corrected the MS signal intensities and normalized the data using MDFC. We then calculated the intragroup variation for each feature using PRSD and PRMAD, as both are measures of the variability of feature intensities. In more specific terms, PRSD uses the square root of the deviation from the mean, while PRMAD uses the median of the absolute deviations from mean. As they capture different data information, we used both to achieve a comprehensive evaluation. Since each feature provides a PRSD and a PRMAD, we examined the median of PRSD and PRMAD of all features and plotted them in **Figures 2C** and **2D**. We observed that the filter of “QC RSD” reduced intragroup variation the most. This suggests that it is critical to remove features with large RSDs in technical replicates prior to normalization factor calculation. Furthermore, the normalization results were evaluated using PCA, as a better normalization method leads to better data separation in the PCA score plot.³²⁻³⁴ As shown in **Figure 2E**, a clear trend of improved data separation is observed with more filters applied to select high-quality features.

It is also important to note that the two last steps, selecting features by peak shape and feature grouping, can also help reduce the intragroup variation. For features with bad peak shapes, properly determining their peak areas is a challenge for feature extraction software.³⁵ Therefore, features with poor chromatographic peak shapes should be excluded from the calculation of normalization factor. In addition, feature grouping clusters features from the same metabolite and keeps only one feature for normalization, ensuring that all unique features have the same weight for normalization. However, the PRSD results indicate that these two steps do not have a strong effect on variation reduction. Therefore, these two steps are not included in MAFFIN. If needed, users can incorporate other existing bioinformatics tools to select features with good chromatographic peak shapes and/or conduct feature grouping. For example, *peakonly*³⁶ or EVA³⁷ can be used to remove features with poor chromatographic peak shapes. On the other side, mz.unity,³⁸ ISFrag,³⁹ or MS-CleanR⁴⁰ can be used for feature grouping.

MS Signal Correction

Next, we used the serial diluted QC samples in the urine data set to demonstrate that the FCs calculated from signal intensities should not be used directly without signal correction. To demonstrate this, we evaluated whether the metabolic feature FCs are a good estimation of the dilution factor. In principle, the dilution factor should be easily estimated using the feature FC at the maximal count. As shown in the light gray histograms in **Figure 3A**, our results of dilution factors 3, 6, and 10 indicate that the FCs calculated directly from signal intensities might not be a good representation of the true dilution factor. In particular, at higher dilution factor, the maximal

count disappears. In comparison, after applying signal intensity correction, the FC at the maximal count can always be easily recognized (the black gray histograms in **Figure 3A**).

Next, we studied how MS signal intensity correction improves the normalization results. We used the high-quality features selected and corrected their MS intensities according to the MAFFIN workflow. Both corrected and non-corrected data were normalized using the MDFC-based method. We plotted the PRSD and PRMAD of normalized data, including corrected and non-corrected data, as well as the raw intensity data (no normalization applied) results. As shown in **Figures 3C** and **3D**, the intragroup variation was clearly reduced after intensity correction. In addition, a clearer data separation was observed on the PCA score plot (**Figure 3E**), which confirms the improved data quality after MS signal correction.

MDFC vs. Median FC

We further compared MDFC with conventionally used median FC to show its advantages in estimating normalization factors. The median FC-based normalization was reported to correct the NMR-based metabolomics data.¹⁸ Although easy to calculate, median is not a robust estimator of normalization factors. To demonstrate this, we first simulated a two-sample data set of imbalanced percentages of up- and down-regulated features with a true dilution factor of 1.5. In this simulation, we first assumed three types of features, which are unchanged, up-, and down-regulated. We then fixed the percentage of unchanged features (50%) and simulated different percentages of up-regulated features from 5 to 45% in steps of 5%. As shown in **Figure 4A**, MDFC shows a stable and accurate estimation of normalization factor (1.5001 ± 0.003) across the testing percentages. In comparison, the median FC estimated normalization factors ranging from 1.39 to 1.63, when up-

regulated features account for 5% and 45%. This demonstrates that median FC has poor normalization factor estimates when the data is imbalanced. We made a more comprehensive evaluation by testing different percentages of unchanged, up-, and down-regulated features at two other true dilution factors (2.0 and 5.0). In each case, the differences between the estimated normalization factors and the true dilution factors were calculated and presented in color gradient. As shown on the left side of **Figure 4B**, when the number of up-regulated features are different from the number of down-regulated features, positive and negative errors were observed in median FC-based method. In comparison, on the right side, the MDFC-based method is robust and never shows any bias across the testing range. Furthermore, comparing the results of different dilution factors, the median FC bias gets bigger (the colored area gets larger) as the true dilution factor increases, which suggests that median FC is less accurate for biological samples of significantly different total sample amount or metabolite concentrations.

Performance of MAFFIN

The performance of the MAFFIN workflow was evaluated and compared to four other post-acquisition normalization methods using the same human urine data set. The RLA plot was first used to visualize the overall sample concentration difference. To make the RLA plot, all feature intensities were log-transformed, and for each feature, its transformed intensities were subtracted by the median value, log transformed. The scaled intensities were visualized using box plot.²³ Our results show that before sample normalization, a clear biological dilution effect exists in urine samples (**Figure 5A**). Interestingly, after sample normalization by MAFFIN, the urine metabolomics data show little overall concentration variance (**Figure 5B**), indicating that MAFFIN can effectively reduce biological dilution effect.

Next, we compared MAFFIN to other normalization methods. In this work, sum, median, PQN, and quantile normalization were compared, all of which are post-acquisition sample normalization methods. After normalizing the urine metabolomics data by the different methods, their intragroup variations, including PRSD and PRMAD, were calculated and plotted in violin plots as shown **Figures 5C and 5D**, respectively. In both plots, the MAFFIN algorithm provides the lowest intragroup variation. We further performed an one-tailed U test, which examines whether MAFFIN results have statistically lower PRSD or PRMAD than the other methods. As presented in **Figures 5C and 5D**, the data processed by MAFFIN have significantly lower intragroup variation than the other methods ($p < 0.001$). We further compared MAFFIN to other methods using PCA analysis. As presented in **Figure 5E**, MAFFIN shows the best data separation in PCA score plot. There are many explanations for the lower normalization powers of the commonly used algorithms. The sum normalization result is severely influenced by features with high abundance.¹⁸ The median normalization assumes that the median signal intensities should be same in different samples. However, the median signal intensity does not mean median metabolite concentration, as different metabolites have different ionization efficiencies. The PQN method uses median FC to estimate the dilution factor, which is biased with imbalanced data. The quantile normalization assumes all data has the same distribution, which is too strict and not in line with reality.¹⁷ Besides the inherent limitations, the lack of feature selection and MS signal intensity correction in these four methods also contribute to their poor performance. Collectively, these results suggest that MAFFIN can effectively reduce the intragroup variation, serving as a robust and universal normalization method to remove variation.

Finally, MAFFIN was demonstrated on a human saliva metabolomics work composed of 10 male and 10 female saliva samples. LC-MS analysis of these saliva samples detected 2092 metabolic features. Among them, 847 were selected as high-quality metabolic features in MAFFIN (**Figure 6A**), and the sample with the most detected features was used as the reference sample (**Figure 6B**). After MS signal correction, the normalization factors for each sample were determined (**Figure 6B**). It is interesting to see that the reference sample, which has the most detected features, also has the largest normalization factor (factor = 1, **Figure 6B**). This is reasonable as samples with more metabolic features are usually more concentrated. From the normalization results, we found that the normalization factors have an RSD of 24.2%, and the highest sample concentration was 2.72-fold of the lowest one. These results suggest there is a large dilution effect in the saliva samples and that sample normalization is necessary to remove the intrasample variance.

The benefit of MAFFIN can be clearly seen, as post-normalization separation of two groups on a PCA score plot is more distinct (**Figure 6C**). Furthermore, the differential metabolic analysis was applied on the data before and after normalization, and *t* test was utilized to select the significantly altered metabolic features (adjusted $p < 0.05$, FC < 0.667 or > 1.5). Among the 2092 total features, there are 159 newly discovered significant features after applying the MAFFIN algorithm, while 37 features became insignificant after normalization (**Figure 6D**). We further manually annotated all the newly discovered significant features by matching their experimental MS² spectra to publicly available databases²⁰ and in-silico MS² databases²¹. The annotations were also manually verified using HMDB.²² The feature annotations and *t* test results are detailed in **SI Table S-1**. As shown in **Figure 6E**, we presented the concentration levels of two annotated metabolites, which became significant after MAFFIN-based sample normalization. Among them, salicylic acid was

409 reported to appear in human saliva and used in clinical pharmacokinetic studies.^{41, 42} In addition,
410 phenylalanine has also been confirmed to be critical to oral health.^{43, 44} In summary, MAFFIN is a
411 robust normalization method to help reveal metabolome differences in case studies with high
412 accuracy and confidence.
413

Conclusion

In this work, we propose MAFFIN, a novel sample normalization strategy that normalizes untargeted metabolomics data collected from MS-based platforms. Our results show that by incorporating high-quality feature selection, MS signal intensity correction, and MDPC, MAFFIN outperforms commonly used post-acquisition sample normalization strategies in terms of reducing intragroup variation and improving separation in PCA plots. MAFFIN was also applied on a saliva metabolomics study and identified 159 additional significantly altered metabolic features related to sex difference. As a post-acquisition normalization method, MAFFIN requires no extra measurements on samples and shows accurate and robust normalization results for both balanced and imbalanced data. MAFFIN can be extremely useful for biological samples that lack a quantity that reflects the total sample amount or concentrations of metabolites. These include saliva, sweat, fecal content, and many others. Furthermore, MAFFIN can also be adapted for non-biological samples such as waste water or other environmental samples used for exposome research. We believe that the development of MAFFIN addresses the critical challenge of sample normalization in untargeted metabolomics. It advances the standardization of untargeted metabolomics workflows and broadens metabolomics applications in biomarker discovery, mechanistic understanding, and beyond.

432 **Supplementary Information**

433 **Table S-1.** Annotation and t test results of newly discovered significant metabolic features; **Text**
434 **S-1.** Parameter settings in MS-DIAL; **Text S-2.** Feature grouping based on MS-DIAL curation
435 results; **Text S-3.** Bandwidth optimization.

436

437 **Acknowledgment**

438 This study was funded by University of British Columbia Start-up Grant (F18-03001), Canada
 439 Foundation for Innovation (CFI 38159), New Frontiers in Research Fund/Exploration (NFRFE-
 440 2019-00789), National Science and Engineering Research Council (NSERC) Discovery Grant
 441 (RGPIN-2020-04895). We also thank Alisa Hui for proof reading this manuscript.

442

References

1. Patti, G. J.; Yanes, O.; Siuzdak, G., Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology* **2012**, *13* (4), 263-269.
2. Lei, Z.; Huhman, D. V.; Sumner, L. W., Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry* **2011**, *286* (29), 25435-25442.
3. Gowda, G. N.; Djukovic, D., Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Mass Spectrometry in Metabolomics* **2014**, 3-12.
4. Warrack, B. M.; Hnatyshyn, S.; Ott, K.-H.; Reily, M. D.; Sanders, M.; Zhang, H.; Drexler, D. M., Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B* **2009**, *877* (5), 547-552.
5. Bliss, D. Z.; Savik, K.; Jung, H.; Jensen, L.; LeMoine, M.; Lowry, A., Comparison of Subjective Classification of Stool Consistency and Stool Water Content. **1999**, *26* (3), 137-141.
6. Karu, N.; Deng, L.; Slae, M.; Guo, A. C.; Sajed, T.; Huynh, H.; Wine, E.; Wishart, D. S., A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. *Analytica Chimica Acta* **2018**, *1030*, 1-24.
7. Yu, H.; Villanueva, N.; Bittar, T.; Arsenault, E.; Labonté, B.; Huan, T., Parallel metabolomics and lipidomics enables the comprehensive study of mouse brain regional metabolite and lipid patterns. *Analytica Chimica Acta* **2020**, *1136*, 168-177.
8. Silva, L. P.; Lorenzi, P. L.; Purwaha, P.; Yong, V.; Hawke, D. H.; Weinstein, J. N., Measurement of DNA concentration as a normalization strategy for metabolomic data from adherent cell lines. *Analytical chemistry* **2013**, *85* (20), 9536-9542.
9. Wu, Y.; Li, L., Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A* **2016**, *1430*, 80-95.
10. Waikar, S. S.; Sabbiseti, V. S.; Bonventre, J. V., Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate. *Kidney International* **2010**, *78* (5), 486-494.
11. Jatlow, P.; McKee, S.; O'Malley, S. S., Correction of Urine Cotinine Concentrations for Creatinine Excretion: Is It Useful? *Clinical Chemistry* **2003**, *49* (11), 1932-1934.
12. Wu, Y.; Li, L., Determination of total concentration of chemically labeled metabolites as a means of metabolome sample normalization and sample loading optimization in mass spectrometry-based metabolomics. *Analytical chemistry* **2012**, *84* (24), 10723-10731.
13. Wu, Y.; Li, L., Dansylation metabolite assay: a simple and rapid method for sample amount normalization in metabolomics. *Analytical chemistry* **2014**, *86* (19), 9428-9433.
14. Zheng, J.; Dixon, R. A.; Li, L., Development of isotope labeling LC-MS for human salivary metabolomics and application to profiling metabolome changes associated with mild cognitive impairment. *Analytical chemistry* **2012**, *84* (24), 10802-10811.
15. Blaise, B. J.; Correia, G. D. S.; Haggart, G. A.; Surowiec, I.; Sands, C.; Lewis, M. R.; Pearce, J. T. M.; Trygg, J.; Nicholson, J. K.; Holmes, E.; Ebbels, T. M. D., Statistical analysis in metabolic phenotyping. *Nature Protocols* **2021**, *16* (9), 4299-4326.
16. 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.; Pelczar, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. *Analytical Chemistry* **2011**, *83* (15), 5864-5872.
17. Bolstad, B. M.; Irizarry, R. A.; Åstrand, M.; Speed, T. P., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **2003**, *19* (2), 185-193.

18. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H., Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ¹H NMR Metabonomics. *Analytical Chemistry* **2006**, *78* (13), 4281-4290.
19. Nam, S. L.; de la Mata, A. P.; Dias, R. P.; Harynuk, J. J., Towards standardization of data normalization strategies to improve urinary metabolomics studies by GC× GC-TOFMS. *Metabolites* **2020**, *10* (9), 376.
20. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods* **2015**, *12* (6), 523-526.
21. Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihiro, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M., Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical chemistry* **2016**, *88* (16), 7946-7958.
22. Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.-A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; MacInnis, G. D.; Weljie, A. M.; Dowlatbadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L., HMDB: the Human Metabolome Database. *Nucleic Acids Research* **2007**, *35* (suppl_1), D521-D526.
23. De Livera, A. M.; Dias, D. A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P., Normalizing and Integrating Metabolomics Data. *Analytical Chemistry* **2012**, *84* (24), 10768-10776.
24. Pang, Z.; Chong, J.; Zhou, G.; de Lima Moraes, D. A.; Chang, L.; Barrette, M.; Gauthier, C.; Jacques, P.-É.; Li, S.; Xia, J., MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research* **2021**, *49* (W1), W388-W396.
25. FDA., Bioanalytical Method Validation - Guidance for Industry. *Bioanalytical Method Validation - Guidance for Industry*. **2018**.
26. Mahieu, N. G.; Patti, G. J., Systems-Level Annotation of a Metabolomics Data Set Reduces 25000 Features to Fewer than 1000 Unique Metabolites. *Anal Chem* **2017**, *89* (19), 10397-10406.
27. Yu, H.; Huan, T., Patterned Signal Ratio Biases in Mass Spectrometry-Based Quantitative Metabolomics. *Analytical Chemistry* **2021**, *93* (4), 2254-2262.
28. Sheather, S. J.; Jones, M. C., A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. **1991**, *53* (3), 683-690.
29. Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G., XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry* **2012**, *84* (11), 5035-5039.
30. Yu, H.; Xing, S.; Nierves, L.; Lange, P. F.; Huan, T., Fold-change compression: an unexplored but correctable quantitative bias caused by nonlinear electrospray ionization responses in untargeted metabolomics. *Analytical Chemistry* **2020**, *92* (10), 7011-7019.
31. Sindelar, M.; Patti, G. J., Chemical discovery in the era of metabolomics. *Journal of the American Chemical Society* **2020**, *142* (20), 9097-9105.
32. Li, B.; Tang, J.; Yang, Q.; Li, S.; Cui, X.; Li, Y.; Chen, Y.; Xue, W.; Li, X.; Zhu, F., NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Research* **2017**, *45* (W1), W162-W170.
33. 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.; Pelczar, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal Chem* **2011**, *83* (15), 5864-72.
34. Chawade, A.; Alexandersson, E.; Levander, F., Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. *Journal of Proteome Research* **2014**, *13* (6), 3114-3120.

35. Yu, H.; Chen, Y.; Huan, T., Computational Variation: An Underinvestigated Quantitative Variability Caused by Automated Data Processing in Untargeted Metabolomics. *Analytical Chemistry* **2021**, *93* (25), 8719-8728.
36. Melnikov, A. D.; Tsentalovich, Y. P.; Yanshole, V. V., Deep Learning for the Precise Peak Detection in High-Resolution LC–MS Data. *Analytical Chemistry* **2020**, *92* (1), 588-592.
37. Guo, J.; Shen, S.; Xing, S.; Chen, Y.; Chen, F.; Porter, E. M.; Yu, H.; Huan, T., EVA: Evaluation of Metabolic Feature Fidelity Using a Deep Learning Model Trained With Over 25000 Extracted Ion Chromatograms. *Analytical Chemistry* **2021**.
38. Mahieu, N. G.; Spalding, J. L.; Gelman, S. J.; Patti, G. J., Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. *Analytical Chemistry* **2016**, *88* (18), 9037-9046.
39. Guo, J.; Shen, S.; Xing, S.; Yu, H.; Huan, T., ISFrag: De Novo Recognition of In-Source Fragments for Liquid Chromatography–Mass Spectrometry Data. *Analytical Chemistry* **2021**, *93* (29), 10243-10250.
40. Fraiser-Vannier, O.; Chervin, J.; Cabanac, G.; Puech, V.; Fournier, S.; Durand, V.; Amiel, A.; André, O.; Benamar, O. A.; Dumas, B.; Tsugawa, H.; Marti, G., MS-CleanR: A Feature-Filtering Workflow for Untargeted LC–MS Based Metabolomics. *Analytical Chemistry* **2020**, *92* (14), 9971-9981.
41. Silwood, C. J. L.; Lynch, E.; Claxson, A. W. D.; Grootveld, M. C., ¹H and ¹³C NMR Spectroscopic Analysis of Human Saliva. *Journal of Dental Research* **2002**, *81* (6), 422-427.
42. Khan, A. Z.; Aarons, L., A note on the use of salicylate saliva concentration in clinical pharmacokinetic studies. *Journal of Pharmacy and Pharmacology* **2011**, *41* (10), 710-711.
43. Liebsch, C.; Pitchika, V.; Pink, C.; Samietz, S.; Kastenmüller, G.; Artati, A.; Suhre, K.; Adamski, J.; Nauck, M.; Völzke, H.; Friedrich, N.; Kocher, T.; Holtfreter, B.; Pietzner, M., The Saliva Metabolome in Association to Oral Health Status. *Journal of Dental Research* **2019**, *98* (6), 642-651.
44. Sugimoto, M.; Wong, D. T.; Hirayama, A.; Soga, T.; Tomita, M., Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics* **2010**, *6* (1), 78-95.

564 TOC

565

566

Figure 1. Schematic workflow of MAFFIN.

Figure 2. High-quality feature selection in MAFFIN. (A) The FC distributions of low- and high-quality features from two urine samples. (B) Feature number; (C) median of PRSD; (D) median of PRMAD; and (E) PCA score plot after applying different feature selection criteria.

Figure 3. MS signal intensity correction in MAFFIN. (A) The FC distributions of technically diluted urine samples with different dilution factors. The histograms in light and dark gray are from measured MS signal intensity and corrected intensity, respectively. The red arrow indicates the position of the correct dilution factor. OOR: number metabolic features that are out of the plotting range. (B) Median of PRSD; (C) median of PRMAD; and (D) PCA score plot of normalization result before and after intensity correction.

Figure 5. Performance of MAFFIN and comparison to other methods. RLA plots of urine data (A) before and (B) after MAFFIN normalization. Violin plots show the (C) PRSD and (D) PRMAD distributions of different data normalization methods. (E) PCA score plots of data normalized by different methods.

Figure 6. Biological application of MAFFIN on human saliva samples. (A) The high-quality feature selection result. (B) The normalization factor calculation results. (C) PCA results before and after MAFFIN normalization. (D) Venn diagram shows the difference of significant metabolites before and after normalization (adjusted $p < 0.05$, $FC < 0.667$ or > 1.5). (E) Box plots show two newly discovered significant metabolites, salicylic acid and phenylalanine.