Metabolomics (2025) 21:98
https://doi.org/10.1007/s11306-025-02297-1

ORIGINAL ARTICLE

Evaluation of normalization strategies for mass spectrometry-based multi-omics datasets

Chi Yen Tseng[1] · Jessica A. Salguero[1] · Joshua D. Breidenbach[1] · Emilia Solomon[1] · Claire K. Sanders[2] · Tara Harvey[1] · M. Grace Thornhill[1] · Salvator J. Palmisano[1] · Zachary J. Sasiene[1] · Brett R. Blackwell[1] · Ethan M. McBride[1] · Kes A. Luchini[1] · Erick S. LeBrun[3] · Marc Alvarez[1] · Phillip M. Mach[1] · Emilio S. Rivera[1] · Trevor G. Glaros[1]

Received: 28 April 2025 / Accepted: 20 June 2025 / Published online: 1 July 2025
This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025

Abstract

Introduction Data normalization is crucial for multi-omics integration, reducing systematic errors and maximizing the likelihood of discovering true biological variation. Most studies assess normalization for a single omics type or use datasets from separate experiments. Few address time-course data, where normalization might bias temporal differentiation. In this study, we compared common normalization methods and a machine learning approach, Systematical Error Removal using Random Forest (SERRF), using multi-omics datasets generated from the same experiment—even from the same cell lysate.

Objectives To develop a straightforward process to assess normalization effects and identify the most robust methods across multi-omics datasets.

Methods We analyzed metabolomics, lipidomics, and proteomics datasets from primary human cardiomyocytes and motor neurons exposed to acetylcholine-active compounds over time. Normalization effectiveness was evaluated based on improvement in QC features consistency and observing the change in treatment and time-related variance.

Results Probabilistic Quotient Normalization (PQN) and Locally Estimated Scatterplot Smoothing (LOESS) QC were identified as optimal for metabolomics and lipidomics, while PQN, Median, and LOESS normalization excelled for proteomics. These methods consistently enhanced QC feature consistency in metabolomics and lipidomics, and preserved time-related variance or treatment-related variance in proteomics, demonstrating their effectiveness and robustness. SERRF normalization, applied only to metabolomics in this study, outperformed other methods in some datasets but inadvertently masked treatment-related variance in others.

Conclusion Our evaluation identified PQN and LoessQC as the top methods for metabolomics and lipidomics, and PQN, Median, and Loess normalization for proteomics, in multi-omics integration in a temporal study.

Keywords Normalization · Multi-omics · Time-course · Metabolomics · Lipidomics · Proteomics

1 Introduction

Data normalization is a critical preprocessing step in the analysis of mass spectrometry-based omics datasets, including metabolomics, proteomics, and lipidomics. Its primary purpose is to maximize the discovery of meaningful biological differences by reducing systematic technical variation arising from discrepancies in sample preparation, extraction, digestion, and instrumental noise, which are often uncontrollable in an experiment (Katajamaa and Orešič, 2007; Mertens, 2017). Despite its importance, many studies evaluating normalization methods often overlook how these methods alter the underlying data structure or remove biological variation (Chua et al., 2023). Typically,

✉ Chi Yen Tseng
chiyen_tseng@lanl.gov

[1] Biochemistry and Biotechnology Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 84545, USA

[2] Microbial and Biome Sciences Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[3] LMI, Colorado Springs, CO 80919, USA

evaluations focus on improving feature consistency in quality control (QC) samples and enhancing differentiation across treatments (Chen et al., 2017; Wang et al., 2013). However, the performance of a normalization strategy depends on the data structure (Callister et al., 2006; Chen et al., 2017; Misra, 2020) and inappropriate normalization strategies can obscure genuine biological signals, leading to inaccurate findings (Chen et al., 2017; Cuevas-Delgado et al., 2020; Dubois et al., 2020; Wang et al., 2021). This issue is particularly evident when sophisticated algorithms, which make rigid assumptions about the nature of bias or involve machine learning, overfit the data, introduce bias, and misinterpret biological phenomena (Fan et al., 2019; Liebal et al., 2020; Välikangas et al., 2018).

Time-course exposure datasets present additional challenges because both time and treatment factors contribute to variance, and these datasets naturally exhibit time-dependent variations in the data structure (Hejblum et al., 2015; Le et al., 2015; Rubingh et al., 2009). Limited studies have systematically addressed how different normalization methods affect such datasets—specifically, which methods can reduce systematic variation without distorting the underlying longitudinal data structure (Sindelar et al., 2021). Avoiding normalization for a longitudinal study to keep pure biological responses has also been reported (Deda et al., 2017). We aim to identify a robust and effective normalization method by evaluating its impact on the variance in feature space attributable to time and treatment. Datasets should preserve time or treatment-related variance following normalization.

In addition to time-course challenges, most normalization evaluations for mass spectrometry-based omics focus on one or two types of omics data rather than including multiple omics platforms (Chua et al., 2023; Li et al., 2016; Välikangas et al., 2018). Each omics type possesses distinct characteristics that influence mass spectral identification (Smith et al., 2014) and quantification (Khoury et al., 2018). Well-known factors that affect omics data analysis include the variation in fragmentation, ionization efficiency, enzyme digestion rate, ion suppression, and spectral counting (Smith et al., 2014). These technical factors affect omics feature identification, quantification, and missing value generation, varying by omics type and complicate the selection of normalization methods (Brombacher et al., 2025; Flores et al., 2023; Kong et al., 2022).

Additionally, normalization methods differ depending on their underlying assumptions and the use of pooled QC samples, which are created by mixing small amounts of multiple individual samples from a study. These methods assume datasets have similar distribution or consistent trends to from which optimal scaling factors can be derived (Callister et al., 2006; Veselkov et al., 2011). Some advanced normalization approaches utilize pooled QC samples to learn feature correlations and address variation in sample injection order, e.g., Systematical Error Removal using Random Forest (SERRF) (Fan et al., 2019). However, it remains uncertain whether these assumptions or machine learning strategies perform robustly in multi-omics experiments or if a single normalization method can perform reliably across diverse omics datasets.

Comparing normalization methods for multi-omics data introduces several challenges. Omics data are often measured from separate experiments and samples, complicating the separation of biological variation from instrumental variations, experimental conditions, and individual sample differences (Conesa & Beck, 2019). A robust normalization method would perform equally well for omics datasets with higher or lower variation in total intensity (Välikangas et al., 2018; Zhao et al., 2020) and identifying such a method requires careful evaluation.

In this study, we address these challenges by systematically evaluating common normalization methods and machine learning-based SERRF normalization across metabolomics, proteomics, and lipidomics datasets derived from the same biological samples, even from the same lysate. Omics measurements were taken from primary human cardiomyocytes and motor neurons exposed to acetylcholine-active compounds to evaluate the performance of different normalization methods in datasets with higher variability in feature intensity across samples (cardiomyocytes) versus lower variability (neurons). To determine robust normalization strategies suitable for different omics datasets, we focus on examining how these methods affect variance explained by time and treatments while enhancing feature consistency in QC samples. Furthermore, we investigate how the inherent characteristics of each omics dataset impact the effectiveness of normalization methods. Our goal is to provide a framework that guides the selection of appropriate normalization techniques in multi-omics studies, ultimately improving data integration and the reliability of biological interpretations.

2 Materials and methods

2.1 Cell culture, exposure, lysis, sample preparation, and data acquisition

Human iPSC-derived motor neurons and commercially sourced cardiomyocytes were cultured and maintained as detailed in the recent publication (Rivera et al., 2024). Neurons and cardiomyocytes were exposed to carbaryl (Sigma Aldrich, product number 32055, CAS 63-25-2) and chlorpyrifos (Supelco, product number 45395, CAS 2921-882) at

0.1 µM with acetonitrile (ACN) as a vehicle control. Cells were collected at 5, 15, 30, 60, 120, 240, 480, 720, 1440 min post-exposure. Information about cell lysis, sample extraction, data collection for metabolomics, lipidomics, and proteomics is described in Supplementary Document 1 and in previously published work (Rivera et al., 2024). Metabolomics datasets were acquired using reverse-phase (RP) and hydrophilic interaction chromatography (HILIC) in both positive and negative ionization modes. Lipidomics datasets were acquired in positive and negative modes. Proteomics datasets were acquired using RP chromatography in positive mode. Detailed information regarding cell culture, exposure, lysis, sample preparation, and data acquisition is also available in recent work by our group (Breidenbach et al., 2024; Rivera et al., 2024). The multi-omics datasets were generated from the same samples and lysates.

2.2 Data preparation

An overview of the multi-omics data pre-processing and normalization evaluation steps is shown in Fig. 1. We processed metabolomics datasets using Compound Discoverer 3.3 (Thermo Scientific, San Jose, CA), lipidomics datasets using MS-DIAL 5.1 (Tsugawa et al., 2020), and proteomics datasets using Proteome Discoverer 3.0 (Thermo Scientific, San Jose, CA). Additional details on data preprocessing, including filtering, missing value imputation and gap filling, are described in Supplementary Document 1.

2.3 Normalization methods

Normalization methods were designed to either match the total intensity of all features and/or the distribution of feature intensities across all samples. We selected six commonly used normalization methods based on different assumptions. Total Ion Current (TIC) normalization assumes that total feature intensity is consistent across samples. Locally Estimated Scatterplot Smoothing (LOESS) normalization assumes balanced proportions of upregulated and downregulated features. Median normalization assumes constant median feature intensity across samples. Quantile normalization assumes overall distribution of feature intensities is similar and can be mapped to the same percentile of the normal distribution. LOESS, Median, and Quantile normalization were performed using the limma package in R. (Ritchie et al., 2015). Probabilistic Quotient Normalization (PQN) also assumes overall distribution of feature intensities is similar across samples; however, instead of assuming normal distribution like Quantile normalization, PQN adjusts the distribution based on the ranking of a reference spectrum. A reference spectrum (the median spectrum from either pooled QC samples or all the samples) is used for estimating dilution factors based on the relative ratio for PQN (Callister et al., 2006; Veselkov et al., 2011). Variance Stabilizing Normalization (VSN) assumes feature variances are dependent on their mean and a transformation can make variance approximately constant and comparable across features. VSN is only selected method that transforms the data distribution. VSN was applied using the vsn package in R (Huber et al., 2002) for proteomics only. TICQC, medianQC, and LOESSQC normalization methods were executed similarly to their standard counterparts, but with each sample normalized individually against all QC samples at a time. Machine learning-based SERRF normalization uses correlated compounds in QC samples to correct systematic errors, including batch effects and injection order. SERRF was applied to the datasets from Compound Discoverer 3.3 software (Thermo Scientific) (Fan et al., 2019).

2.4 Evaluation: differentiation across treatments or time points

To assess the effectiveness of normalization methods in differentiating across treatments or time points, we calculated effect size (R^2 value) for each normalized dataset using PERMANOVA (adonis2 function, vegan package in R; Oksanen et al., 2022). This metric reflects the proportion of variance explained by differences in time points or treatments. Effect size was only reported for datasets with statistically significant differentiation across time points or treatments. Additionally, differentiation between treatments was verified using ANOSIM (also from the vegan package in R), serving as a complementary method to confirm observed distinctions in treatment groups.

Significantly altered metabolites in neuronal cells exposed to chlorpyrifos across time points, as identified in both RP positive and RP negative datasets after PQN or Quantile normalization, were uploaded to MetaboAnalyst 6.0 (Pang et al., 2024) to identify associated KEGG pathways. Further details on the data visualization are provided in Supplementary Document 1.

2.5 Evaluation: feature consistency in QC samples

The relative standard deviation (RSD) was calculated for each feature in normalized omics datasets. The number of features with an RSD less than 0.2, i.e., consistent features, was used to demonstrate if normalization methods improved feature consistency in QC samples (Cajka & Fiehn, 2016). To compare across different omics datasets, the number of features with an RSD less than 0.2 for each normalized dataset was compared to its pre-normalized dataset as a metric to quantify improvement following normalization.

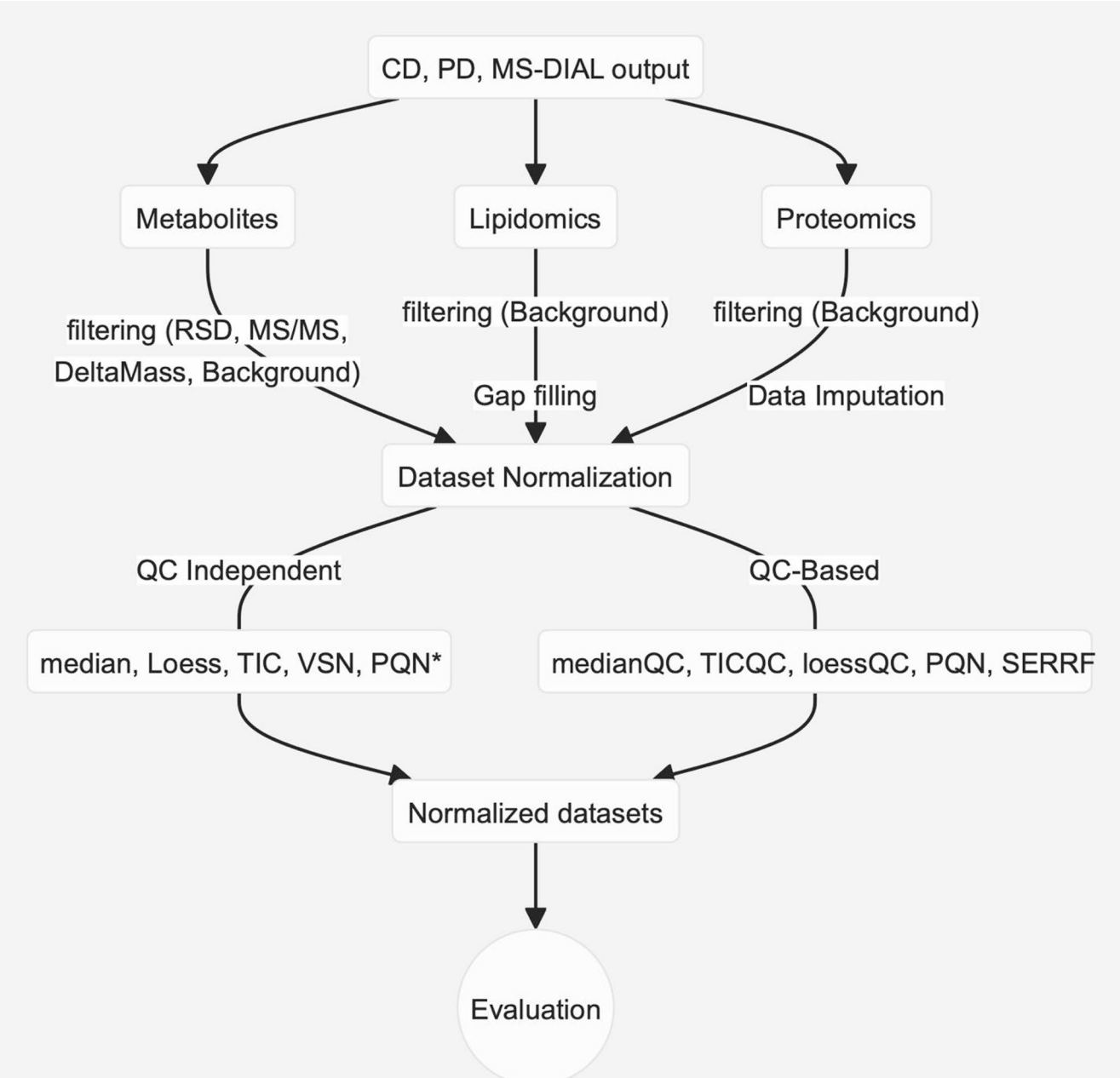


Fig. 1 Multi-omics data were preprocessed by applying filtering, imputation, and normalization. Two distinct strategies were used: one based on the average sample as the reference (QC-independent) and another relying solely on pooled QC samples (QC-based). The resulting normalized datasets were evaluated to assess the effectiveness of each normalization approach

2.6 Identification of significant features

The number of significant features was determined sequentially through each feature by building a separate generalized additive model (GAM) with consideration of time and time-treatment interaction as smoothing effects (mgcv package in R) (Wood, 2011). The false discovery rate (FDR) was calculated to correct for multiple comparisons using the p.adjust function, Benjamini & Hochberg adjustment method in R (Benjamini & Hochberg, 1995). FDR less than 0.1 was considered as significant altered features across treatments.

2.7 Script availability

Detailed R scripts for normalization, evaluation of treatment/time differentiation, QC feature consistency, significant features identification, and plotting are provided in Supplementary Document 1.

3 Results

3.1 QC feature consistency and differentiation across timepoints or treatments in normalized metabolomics datasets

The effectiveness of each normalization method on metabolomics datasets was evaluated based using QC feature consistency and the relative variance explained by treatment or time following normalization (Fig. 2a), assuming the better normalization methods could improve QC feature consistently across datasets. Normalized datasets from four primary human cardiomyocyte and four primary human neuronal samples were included as datapoints in the comparison. In cardiomyocytes and neurons, comparison of metabolomics datasets before normalization showed that QC feature was more consistent in neurons (median QC RSD≈0.19) than in cardiomyocytes (median QC RSD≈0.35) (Table

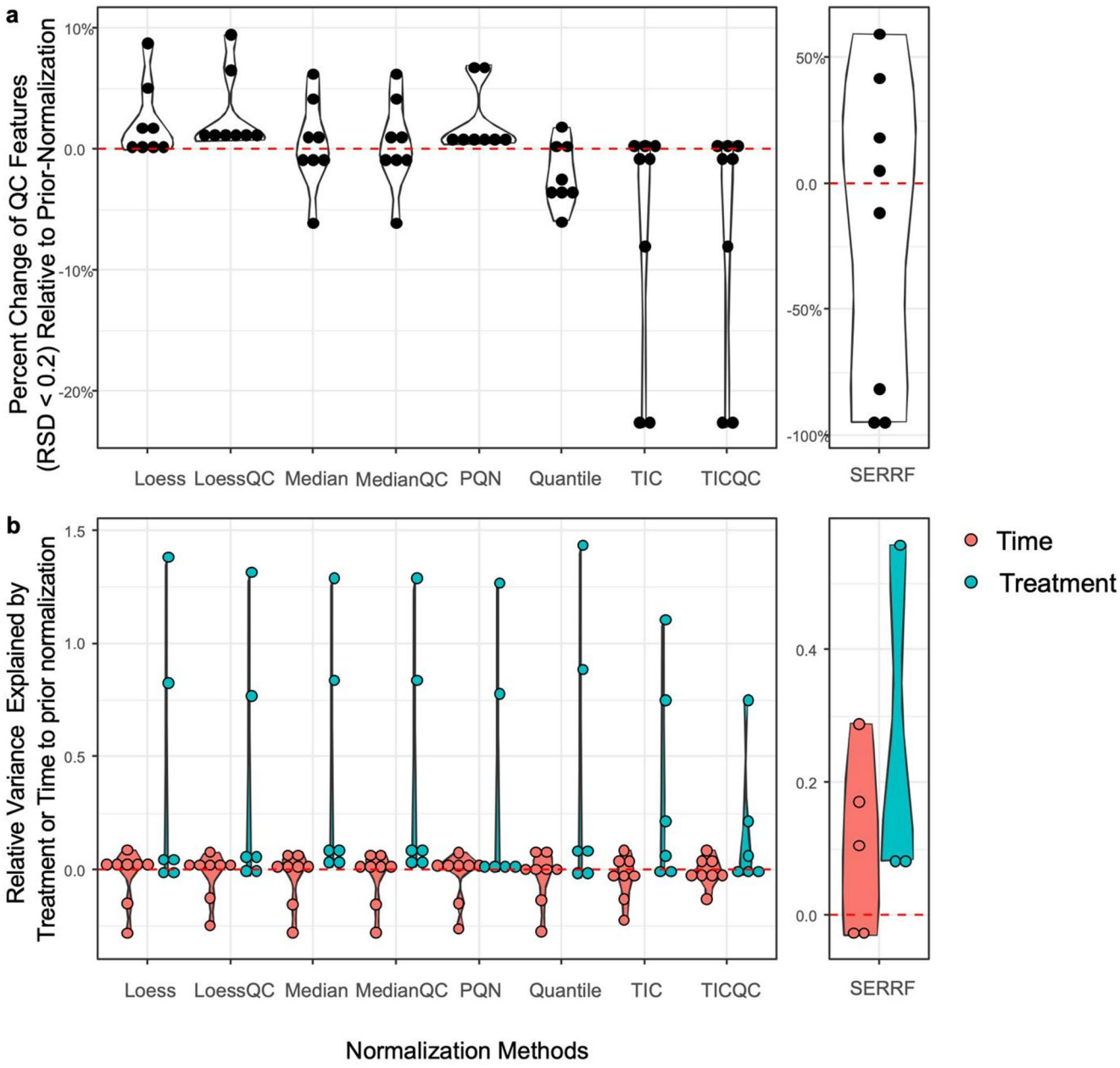


Fig. 2 Metabolomics datasets. The effectiveness of each normalization method was evaluated based on two criteria: **a** feature consistency in QC samples, defined as the percent change in consistent QC features (RSD < 0.2) following normalization; and **b** the relative variance explained by treatment or time after normalization. Individual datapoints represent the neuron and cardiomyocyte datasets. Datasets in which variance could not be significantly explained by time or treatment variation (PERMANOVA, $p \geq 0.5$) were excluded. Note that a different scale was used for the SERRF normalization method to emphasize its distinct performance relative to the other methods

S1). Following normalization, Loess, LoessQC, and PQN improved QC feature consistency, as indicated by reducing median RSD values in QC samples and an increase in the number of features with RSD < 0.2 relative to pre-normalization levels across datasets, supporting their effectiveness in reducing systematic variability (Table 1).

Time accounted for over 20% of variance, while treatment typically explained less than 5% variance in feature space (Table S1). The relative variance explained by time or treatment in each dataset following normalization, compared to pre-normalization, is shown as individual data points (Fig. 2b). Datasets in which variance was not significantly explained by time or treatment ($p < 0.05$) were excluded from the comparison. In most metabolomic datasets, normalization increased the variance explained by either time or treatment. However, in two datasets, normalization substantially increased the variance explained by treatment, while concurrently reducing the variance explained by time (Fig. 2b).

However, SERRF normalization failed to improve the number of consistent QC features in 50% of datasets (Fig. 2a). In addition, variance was not significantly explained by treatment in 62.5%, indicating that SERRF is a more dataset-dependent normalization method for these metabolomics datasets. Nevertheless, when effective, SERRF normalization increased consistent QC features by up to 50%, compared to a maximum of 10% improvement observed with other normalization methods.

3.2 QC feature consistency and differentiation across timepoints or treatments in normalized lipidomics datasets

Lipidomics datasets, measured in both positive and negative modes from the same cell lysate as metabolomics, showed improved QC feature consistency following normalization with Loess, LoessQC, Median, MedianQC, and PQN (Fig. 3a; Table 1). Among these five normalization methods, PQN and LoessQC yielded the lowest median RSD values (Table 1). Cardiomyocytes exhibited slightly lower QC feature variability (median QC RSD ≈ 0.11) compared to neurons (median QC RSD ≈ 0.17) (Table S2).

In these lipidomics datasets, time accounted for over 20% of the variance, while treatment typically explained less than 5% (Table S2). Changes in variance was visualized as individual data points, excluding datasets where time or treatment effects were not significant ($p < 0.05$) (Fig. 3b). Overall, normalization reduced the variance explained by time while increasing the variance explained by treatment (Fig. 3b; Table 1).

3.3 Differentiation across timepoints or treatments in normalized proteomics datasets

QC feature consistency was not assessed in the proteomics datasets, as pooled QC samples were not included in the experimental design. Instead, we evaluated the effect of normalization on datasets from the two cell types. Comparing feature intensity, cardiomyocytes showed much higher feature intensity variability than neuron datasets (Fig. S1).

Time initially accounted for a substantially larger portion of variance (≈17% R^2) compared to treatment (≈3% R^2) (Table S3). In cardiomyocyte datasets, normalization reduced the variance explained by time while greatly increasing the variance explained by treatment (Fig. 4). Conversely, in neuron datasets, normalization increased the variance explained by time while reducing that explained by treatment. Among the normalization methods, PQN, Median, and Loess were the least likely to reduce variance explained by time in cardiomyocyte datasets or by treatment in neuron datasets (Fig. 4; Table 1).

3.4 Downstream effects of normalization methods: significant feature size and related pathways

The number of significant features between treatments and their associated functions depends on the chosen normalization method. To illustrate this, we compared two reverse-phase metabolomic neuron datasets to examine how significantly differentiated features ($FDR < 0.1$), and their associated pathways varied between PQN normalization—which consistently improved QC feature consistency—and Quantile normalization, which was least effective. Quantile normalization identified 15% more significant metabolites than PQN normalization across all time points (Fig. 5a). Although the pathways associated with these altered metabolites were largely consistent, PQN uniquely enabled the detection of three KEGG pathways related to amino acid metabolism that were not observed in the Quantile-normalized datasets (Fig. 5b).

4 Discussion

4.1 Observing normalization effects across datasets

Normalization enhances feature consistency in QC samples, reduces analytical variability, and facilitates the identification of biologically significant features. Common metrics for evaluating normalization effectiveness include PCA dispersion, median RSD in QC samples, and the total number of consistent features (Kirwan et al., 2013; Wang et al., 2013). In this study, we used the number of consistent

Table 1 Summary comparison of normalization methods performance across metabolomics, lipidomics, and proteomics datasets

Norm. Method	# Sig-nificant by Treat.	Median RSD	Relative Δ # RSD < 0.2	Avg. $\Delta R^2/R^2$ (Pre-Norm) by Treat.	Avg. $\Delta R^2/R^2$ (Pre-Norm) by Time
Metabolomics					
Pre-normalization	6/8	0.277 (0.038)	0 (0)	0 (0)	0 (0)
SERRF	3/8	0.207 (0.084)	− 0.198 (0.221)	0.24 (0.159)	0.102 (0.06)
LoessQC	6/8	0.272 (0.037)	0.027 (0.012)	0.366 (0.226)	− 0.024 (0.038)
Loess	6/8	0.273 (0.037)	0.022 (0.011)	0.38 (0.241)	− 0.025 (0.044)
PQN	6/8	0.273 (0.038)	0.022 (0.01)	0.355 (0.221)	− 0.028 (0.04)
Median	6/8	0.275 (0.037)	0.004 (0.013)	0.395 (0.22)	− 0.029 (0.043)
MedianQC	6/8	0.275 (0.037)	0.004 (0.013)	0.395 (0.22)	− 0.029 (0.043)
Quantile	6/8	0.276 (0.036)	− 0.021 (0.009)	0.411 (0.249)	− 0.028 (0.042)
TICQC	6/8	0.282 (0.039)	− 0.067 (0.036)	0.169 (0.122)	− 0.006 (0.023)
TIC	6/8	0.282 (0.04)	− 0.068 (0.036)	0.354 (0.191)	− 0.034 (0.035)
Lipidomics					
Pre-normalization	3/4	0.15 (0.019)	0 (0)	0 (0)	0 (0)
LoessQC	3/4	0.14 (0.024)	0.044 (0.003)	0.083 (0.107)	− 0.139 (0.013)
PQN	3/4	0.141 (0.021)	0.04 (0.007)	0.123 (0.032)	− 0.147 (0.027)
Loess	2/4	0.142 (0.021)	0.027 (0.016)	0.121 (0.16)	− 0.16 (0.063)
Median	2/4	0.146 (0.019)	0.02 (0.005)	0.045 (0.115)	− 0.102 (0.054)
MedianQC	2/4	0.146 (0.019)	0.02 (0.005)	0.045 (0.115)	− 0.102 (0.054)
TIC	3/4	0.155 (0.033)	− 0.062 (0.104)	0.175 (0.074)	− 0.115 (0.03)
TICQC	3/4	0.155 (0.033)	− 0.062 (0.104)	0.175 (0.074)	− 0.115 (0.03)
Quantile	2/4	0.162 (0.022)	− 0.096 (0.032)	0.076 (0.092)	− 0.138 (0.014)

	Treatment Effect (p)	$\Delta R^2/R^2$ (Pre-Norm) by Treat.	$\Delta R^2/R^2$ (Pre-Norm) by Time
Proteomics			
Cardiomyocyte			
Prior Normalization	0.52	0	0
Median	0.013	0.712	− 0.008
PQN	0.014	0.701	− 0.011
Loess	0.01	0.671	− 0.015
VSN	0.016	0.713	− 0.016
Quantile	0.013	0.708	− 0.02
TIC	0.031	0.698	− 0.029
Neuron			
Prior Normalization	0.027	0	0
Loess	0.039	− 0.03	0.077
PQN	0.04	− 0.05	0.058
Median	0.043	− 0.058	0.066
TIC	0.041	− 0.067	0.021
Quantile	0.045	− 0.069	0.06
VSN	0.055	− 0.081	0.126

The table summarizes the number of datasets with significant treatment effects (# significant by Treat.), the average of median QC feature relative standard deviations (Median RSD) across datasets, and the average relative change in variance explained by treatment or time following normalization (Avg. $\Delta R^2/R^2$ (Pre-Norm) by Treat., Avg. $\Delta R^2/R^2$ (Pre-Norm) by Time) in metabolomics and lipidomics. All comparisons except for proteomics were performed on combined datasets from neurons and cardiomyocytes. For proteomics, we assessed the statistical significance of the treatment effect (Treatment Effect (*p*)) and the change in variance explained by treatment or time following normalization ($\Delta R^2/R^2$ (Pre-Norm) by Treat., $\Delta R^2/R^2$ (Pre-Norm) by Time) separately for neuronal and cardiomyocyte datasets

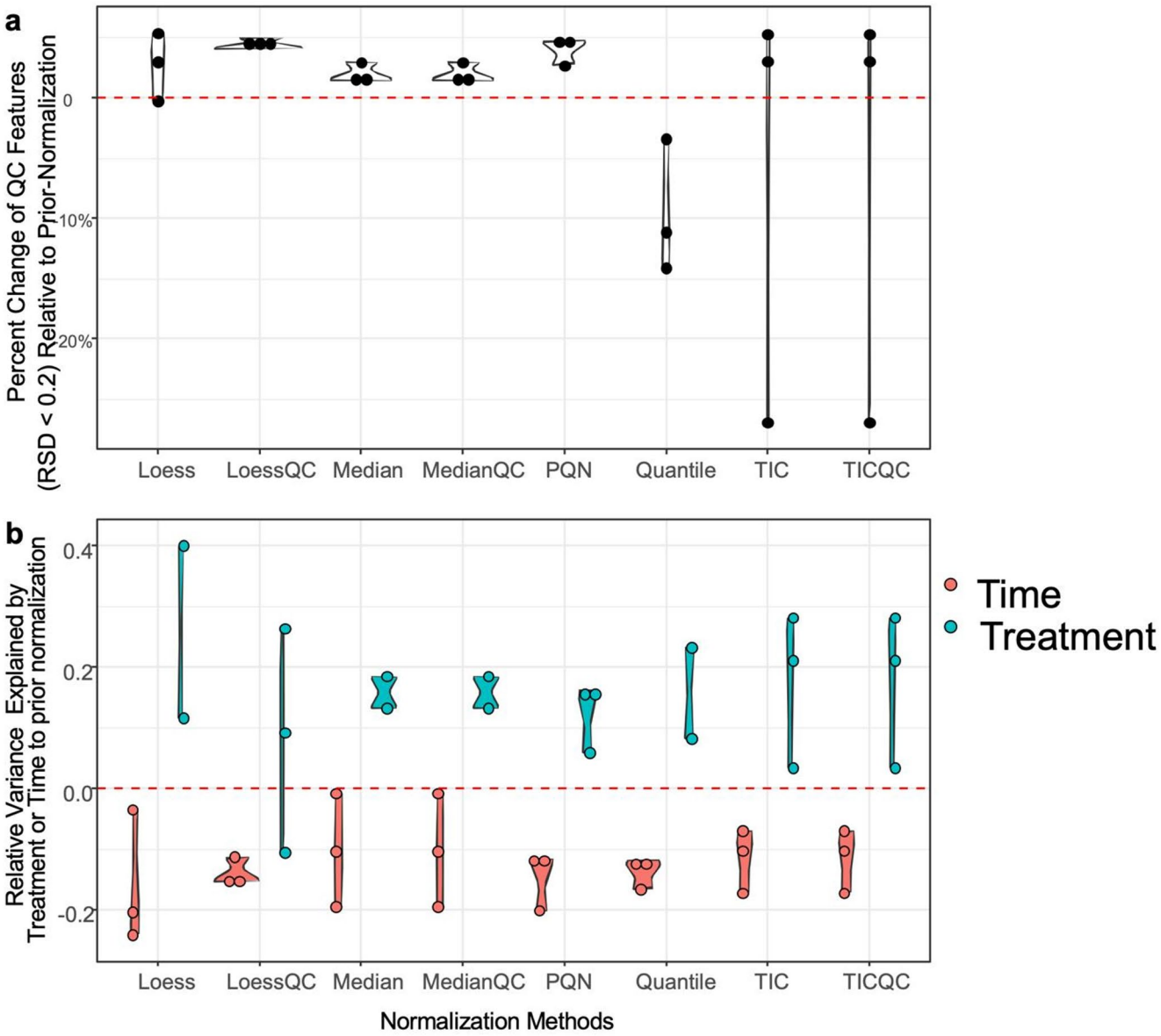


Fig. 3 Lipidomics datasets. The effectiveness of each normalization method was evaluated using two criteria: **a** feature consistency in QC samples, defined as the percent change in consistent QC features (RSD < 0.2) following normalization; and **b** the relative variance explained by treatment or time following normalization. Individual datapoints represent the neuron and cardiomyocyte datasets. Datasets in which variance was not explained by time or treatment (PERMANOVA, $p \geq 0.5$) were excluded

features following normalization as a key indicator of reduced systematic error, and observed changes in the variance explained by time or treatment across datasets following normalization. The multi-omics datasets originated from the same samples and lysates (Breidenbach et al., 2024), enabling a fair evaluation of normalization approaches by minimizing inter-sample variability.

Chlorpyrifos (an organophosphate pesticide) and carbaryl (an N-methyl carbamate pesticide), both known acetylcholine-active compounds (Eaton et al., 2008; Gunasekara et al., 2008), induced strong biological effects at tested exposure levels (Rivera et al., 2024). Among the two cell types studied, cardiomyocyte datasets exhibited greater variability, especially in metabolomics and proteomics, suggesting higher noise levels, and the noisy mass spectra could mask peaks of biological importance (Du et al., 2008). To ensure transparency and robustness, normalization methods were evaluated using predefined metrics with the assumption that the benchmark outcome varies in terms of data characteristics (Brombacher et al., 2025). Given the limited sample size, we did not perform cross-validation, which can bias results toward methods that emphasize increasing variance, i.e., over-fitting the data (Xi et al., 2014). For omics data, meaningful cross-validation for classifying treatment effects

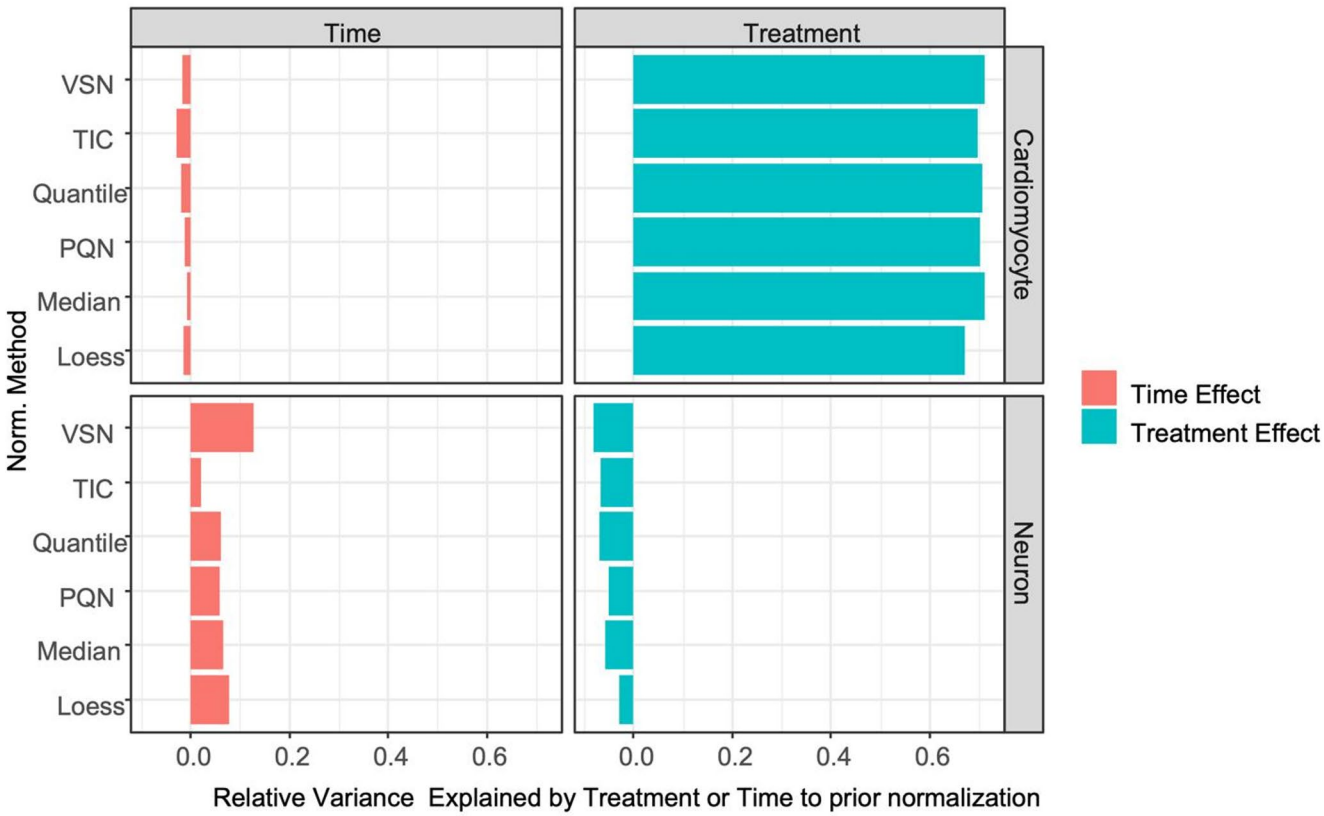


Fig. 4 The effectiveness of each normalization method in proteomics datasets, acquired from cardiomyocyte and neuron samples, was evaluated based on the relative variance explained by treatment or time following normalization. Individual bars represented the neuron and cardiomyocyte datasets, variance in time and treatment effects after normalization is shown in orange and blue, respectively

typically requires a substantially larger number of biological replicates (Rodríguez-Pérez et al., 2018; Zhang et al., 2019). However, because we examined variance explained by time, treatment, and their interaction, we effectively had only three replicates per condition. Instead of performing cross-validation, we prioritized methods that enhance QC feature consistency and observed how different normalization affected variance explained by time or treatment.

Most normalization methods assume central tendency, constant distribution, and similar systematic biases across feature intensities (Callister et al., 2006; Veselkov et al., 2011). Key differences among normalization methods include whether they alter feature rankings (e.g., VSN), assume a normal distribution or use a reference distribution (Quantile vs. PQN), normalize based on equal total intensity (TIC), and whether they rely solely on QC samples or incorporate all samples. We included widely used normalization methods based on distinct underlying assumptions, but excluded others that shared similar assumptions.

However, these assumptions may not hold because of the complex nature of time-course datasets, which can introduce diverse and unpredictable variation (De Livera et al., 2012; Yang et al., 2020). While few studies evaluated the effectiveness of normalization methods by assessing their classification performance, none have included variance explained by time (Yang et al., 2020). Time effects account for the majority of variance in longitudinal datasets and interact with treatment effects (Ren et al., 2015). Therefore, preserving time-related variance is a key metric for assessing whether normalization introduces additional variation.

In addition to time-course challenges, normalization effects on QC feature consistency varied across omic layers. In metabolomics and lipidomics, PQN and LoessQC yielded consistent improvement in QC feature consistency across datasets while quantile, TIC, and TICQC performed worse in QC consistency. The current proteomics datasets did not include QC samples; QC feature consistency was not evaluated. Although SERRF normalization is gaining popularity for bridging batch effects (Bergmann et al., 2024; Fan et al., 2019; Huang et al., 2021), we found that SERRF substantially improve QC consistency in some metabolomics datasets but was ineffective in others. In several cases, treatment-related variance missed by SERRF was captured by other methods, suggesting that SERRF may obscure genuine biological signals.

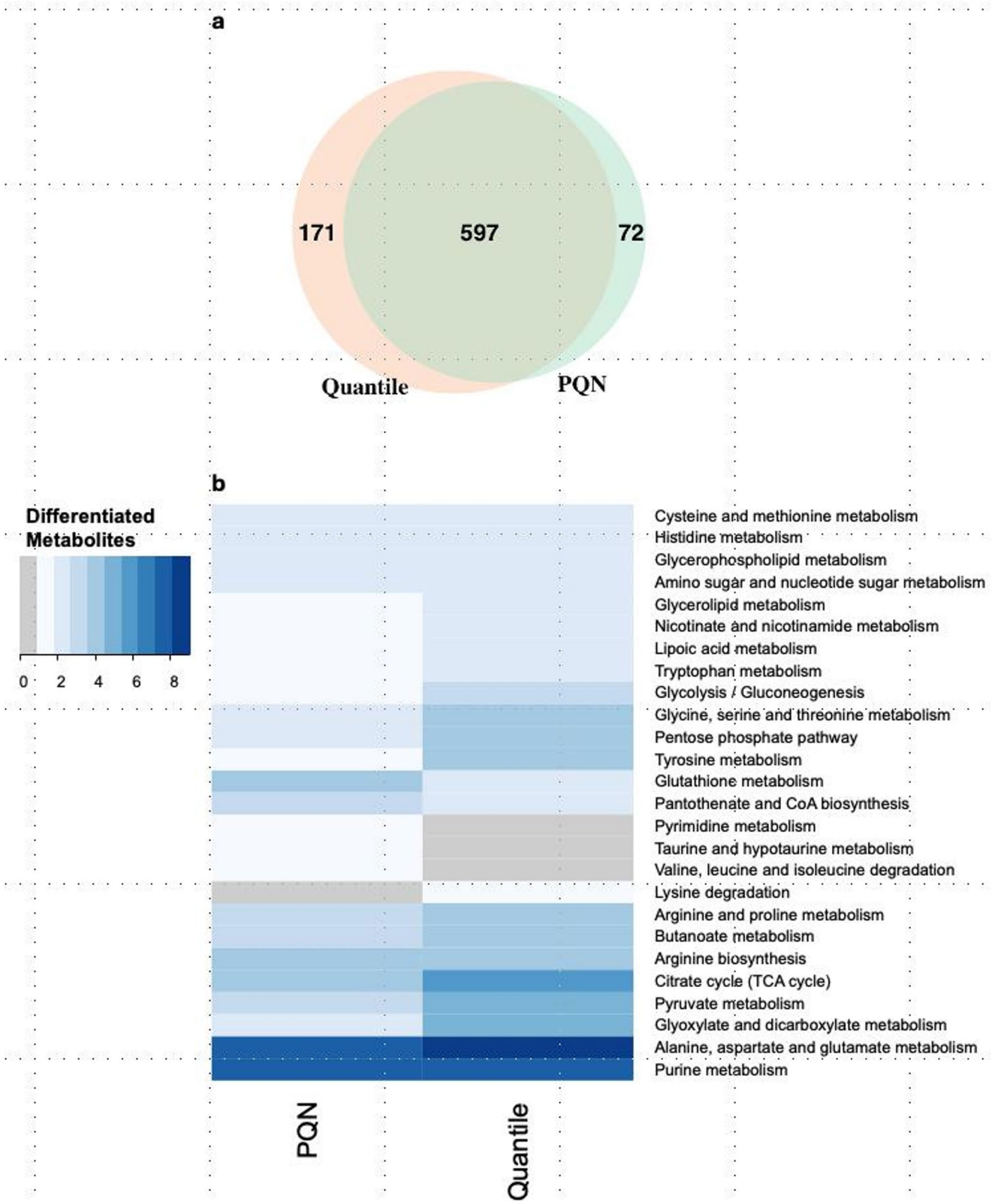


Fig. 5 We compared significantly altered metabolites identified from PQN-normalized and Quantile-normalized reverse-phase (RP) positive and RP negative datasets in neuron cells exposed to chlorpyrifos. PQN normalization represented the better-performing method, while Quantile normalization was one of the least effective. **a** Venn diagram showing that Quantile normalization identified 15% more significant metabolites than PQN normalization across all time points. **b** Heatmap comparing the KEGG pathways associated with the altered metabolites, colored by the number of annotated metabolites in each pathway. Although both methods identified largely overlapping KEGG pathways, PQN uniquely detected three amino acid metabolism pathways not identified in the Quantile-normalized datasets

SERRF has been shown to be unsuitable for smaller datasets, e.g., < 500 samples, as it has shown signs of overfitting and varied performance (Fan et al., 2019; Zhang et al., 2023).

In metabolomics, normalization generally increased the variance explained by both time and treatment, whereas in lipidomics, it tended to decrease the variance explained by time while concurrently increasing the variance explained by treatment. In proteomics, the effect of normalization was mixed, either increasing or decreasing the variance explained by time depending on the cell types. The decrease of variance explained by time or treatment could suggest a biased normalization result, or over-normalization, which may related to median feature intensity and sample-dependent detection limits (Brombacher et al., 2025). The introduction of biases could obscure meaningful biological variation, leading to misleading conclusions (Wehrli et al., 2019)

4.2 Downstream effects following normalization

We identified PQN and LoessQC as the most robust normalization methods across all omic layers while Quantile normalization was often the least effective in terms of QC feature consistency. When comparing PQN and Quantile, we observed that both these methods identified similar KEGG pathways, suggesting that the choice of normalization method did not substantially alter overall pathway annotations, consistent with a recent study (Carvalho et al., 2024). However, despite this similarity, PQN normalization uncovered three dysregulated biochemical pathways related to amino acid degradation that Quantile normalization did not. This indicates that PQN is overall more effective than Quantile normalization in capturing biologically relevant responses.

5 Conclusion

Our study highlights the importance of selecting appropriate normalization methods across omics datasets before multi-omics integration. By systematically evaluating commonly used methods using multi-omics datasets derived from a single experiment—including those from the same cell lysate, we identified several key assessment criteria. These criteria emphasize maximizing QC feature consistency and preserving critical data structures as variance explained by time or treatment.

Normalization methods can have varying impacts on multi-omics datasets, influenced by factors such as cell type, signal strength, and inherent data variation. Considering all these factors, we identified that PQN and LoessQC normalization in metabolomics and lipidomics, and PQN, Median, and Loess normalization in proteomics were the top normalization methods. In each case, these methods consistently improved QC feature consistency without compromising underlying data structure. Selecting appropriate normalization methods enhances the reliability of multi-omics data integration and supports more accurate biological interpretations. We recommend a thorough evaluation of normalization effects on individual datasets prior to multi-omics integration in a time-course study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11306-025-02297-1.

Acknowledgements The authors would like to express their sincere gratitude to Dr. John Tipton for the insightful discussions and suggestions during the revision of this manuscript.

Author contributions C.T wrote the main manuscript text and prepared all figures. J.A and S.P contributed to data curation and processing. J.B, E.S, C.S, and T.H conducted experiments and exposure procedures. M.T, B.B, K.L and E.R extracted and acquired the omics datasets. J.B, C.S, Z.S, B.B, E.M, E.R, and T.G contributed to reviewing and editing. E.L, E.M, M.A, P.M supervised the research activities. E.R and T.G managed project administration and T.G. was responsible for funding acquisition.

Funding This work is released for publication in accordance with Los Alamos National Laboratory (LANL) LA-UR-25-24047 by Triad National Security, LLC (Los Alamos, NM), operator of the Los Alamos National Laboratory under Contract No. 89233218CNA000001 with the U.S. Department of Energy. The Los Alamos National Laboratory Directed R&D Fund, LDRD, grant number 20230084DR, funded this work.
This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under the Targeted Evaluation of Ionizing Radiation Exposure (TEI-REX) program contract #IARPA-20008-D2021-210803004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODIN, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Data availability Data that support the findings of this study have been deposited in the MassIVE database with the accession code MSV000097689 (Lipidomics), MSV000097682 (Proteomics), and MSV000097702 (Metabolomics). Normalization and evaluation R script are described in the supplementary document 1 and provided in the supplementary files.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the

source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bergmann, D., Matarrita-Rodríguez, J., & Abdulla, H. (2024). Toward a more comprehensive approach for dissolved organic matter chemical characterization using an orbitrap fusion tribrid mass spectrometer coupled with ion and liquid chromatography techniques. *Analytical Chemistry*, *96*(9), 3744–3753. https://doi.org/10.1021/acs.analchem.3c02599

Breidenbach, J. D., Rivera, E. S., Harvey, T., Mikolitis, A. S., Tseng, C. Y., Sanders, C. K., Solomon, E., Gleasner, C. D., Thornhill, M. G., Luchini, K. A., & McBride, E. M. (2024). The addition of transcriptomics to the bead-enabled accelerated monophasic multi-omics method: A step toward universal sample preparation. *Analytical Chemistry, 96*(46), 18343–18348. https://doi.org/10.1021/acs.analchem.4c02835

Brombacher, E., Schilling, O., & Kreutz, C. (2025). Characterizing the omics landscape based on 10,000 + datasets. *Scientific Reports*, *15*(1), 3189. https://doi.org/10.1038/s41598-025-87256-5

Cajka, T., & Fiehn, O. (2016). Toward merging untargeted and targeted methods in mass Spectrometry-Based metabolomics and lipidomics. *Analytical Chemistry*, *88*(1), 524–545. https://doi.org/10.1021/acs.analchem.5b04491

Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W. J., Webb-Robertson, B. J. M., Smith, R. D., & Lipton, M. S. (2006). Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of Proteome Research, 5*(2), 277–286. https://doi.org/10.1021/pr0503001

Carvalho, L. B., Teigas-Campos, P. A., Jorge, S., Protti, M., Mercolini, L., Dhir, R., Wiśniewski, J. R., Lodeiro, C., Santos, H. M., & Capelo, J. L. (2024). Normalization methods in mass spectrometry-based analytical proteomics: a case study based on renal cell carcinoma datasets. *Talanta, 266*, 124953. https://doi.org/10.1016/j.talanta.2023.124953

Chen, J., Zhang, P., Lv, M., Guo, H., Huang, Y., Zhang, Z., & Xu, F. (2017). Influences of normalization method on biomarker discovery in gas Chromatography–Mass Spectrometry-Based untargeted metabolomics: What should be considered?? *Analytical Chemistry*, *89*(10), 5342–5348. https://doi.org/10.1021/acs.analchem.6b05152

Chua, A. E., Pfeifer, L. D., Sekera, E. R., Hummon, A. B., & Desaire, H. (2023). Workflow for evaluating normalization tools for omics data using supervised and unsupervised machine learning. *Journal of the American Society for Mass Spectrometry, 34*(12), 2775–2784. https://doi.org/10.1021/jasms.3c00295

Conesa, A., & Beck, S. (2019). Making multi-omics data accessible to researchers. *Scientific Data*, *6*(1), 251. https://doi.org/10.1038/s41597-019-0258-4

Cuevas-Delgado, P., Dudzik, D., Miguel, V., Lamas, S., & Barbas, C. (2020). Data-dependent normalization strategies for untargeted metabolomics-a case study. *Analytical and Bioanalytical Chemistry*, *412*(24), 6391–6405. https://doi.org/10.1007/s00216-020-02594-9

De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M., & Speed, T. P. (2012). Normalizing and integrating metabolomics data. *Analytical Chemistry, 84*(24), 10768–10776. https://doi.org/10.1021/ac302748b

Deda, O., Gika, H. G., Taitzoglou, I., Raikos, N., & Theodoridis, G. (2017). Impact of exercise and aging on rat urine and blood metabolome. An LC-MS based metabolomics longitudinal study. *Metabolites*, *7*(1), 10. https://doi.org/10.3390/metabo7010010

Du, P., Stolovitzky, G., Horvatovich, P., Bischoff, R., Lim, J., & Suits, F. (2008). A noise model for mass spectrometry based proteomics. *Bioinformatics*, *24*(8), 1070–1077. https://doi.org/10.1093/bioinformatics/btn078

Dubois, E., Galindo, A. N., Dayon, L., & Cominetti, O. (2020). Comparison of normalization methods in clinical research applications of mass spectrometry-based proteomics. In 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1–10). Presented at the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). https://doi.org/10.1109/CIBCB48159.2020.9277702

Eaton, D. L., Daroff, R. B., Autrup, H., Bridges, J., Buffler, P., Costa, L. G., Coyle, J., McKhann, G., Mobley, W. C., Nadel, L., & Neubert, D. (2008). Review of the toxicology of chlorpyrifos with an emphasis on human exposure and neurodevelopment. *Critical Reviews in Toxicology, 38*(sup2), 1–125. https://doi.org/10.1080/10408440802272158

Fan, S., Kind, T., Cajka, T., Hazen, S. L., Tang, W. W., Kaddurah-Daouk, R., Irvin, M. R., Arnett, D. K., Barupal, D. K., & Fiehn, O. (2019). Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data. *Analytical Chemistry, 91*(5), 3590–3596. https://doi.org/10.1021/acs.analchem.8b05592

Flores, J. E., Claborne, D. M., Weller, Z. D., Webb-Robertson, B. J. M., Waters, K. M., & Bramer, L. M. (2023). Missing data in multi-omics integration: Recent advances through artificial intelligence. Frontiers in Artificial Intelligence, 6. Retrieved July 5, 2023, from https://www.frontiersin.org/articles/10.3389/frai.2023.1098308

Gunasekara, A. S., Rubin, A. L., Goh, K. S., Spurlock, F. C., & Tjeerdema, R. S. (2008). Environmental fate and toxicology of carbaryl. In D. M. Whitacre (Ed.), *Reviews of environmental contamination and toxicology.* Springer. https://doi.org/10.1007/978-0-387-78444-1_4

Hejblum, B. P., Skinner, J., & Thiébaut, R. (2015). Time-course gene set analysis for longitudinal gene expression data. *PLOS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004310

Huang, Y., Mu, R., Wen, D., Grimsby, J. S., Liang, M., & Rosenbaum, A. I. (2021). Differences in levels of phosphatidylinositols in healthy and stable Coronary Artery Disease subjects revealed by HILIC-MRM method with SERRF normalization. *PLOS ONE*, *16*(6), e0252426. https://doi.org/10.1371/journal.pone.0252426

Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, *18 Suppl*(1), S96–S104.

Katajamaa, M., & Orešič, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, *1158*(1), 318–328. https://doi.org/10.1016/j.chroma.2007.04.021

Khoury, S., Canlet, C., Lacroix, M. Z., Berdeaux, O., Jouhet, J., & Bertrand-Michel, J. (2018). Quantification of lipids: Model, reality,

and compromise. *Biomolecules*, *8*(4), 174. https://doi.org/10.3390/biom8040174

Kirwan, J. A., Broadhurst, D. I., Davidson, R. L., & Viant, M. R. (2013). Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. *Analytical and Bioanalytical Chemistry*, *405*(15), 5147–5157. https://doi.org/10.1007/s00216-013-6856-7

Kong, W., Hui, H. W. H., Peng, H., & Goh, W. W. B. (2022). Dealing with missing values in proteomics data. *PROTEOMICS*, *22*(23–24), 2200092. https://doi.org/10.1002/pmic.202200092

Le, H., Jerums, M., & Goudar, C. T. (2015). Characterization of intrinsic variability in time-series metabolomic data of cultured mammalian cells. *Biotechnology and Bioengineering*, *112*(11), 2276–2283. https://doi.org/10.1002/bit.25646

Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., Cao, Q., Xue, W., Chen, N., & Zhu, F. (2016). Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Scientific Reports, 6*(1), 38881. https://doi.org/10.1038/srep38881

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., & Blank, L. M. (2020). Machine learning applications for mass Spectrometry-Based metabolomics. *Metabolites*, *10*(6), 243. https://doi.org/10.3390/metabo10060243

Mertens, B. J. A. (2017). Transformation, normalization, and batch effect in the analysis of mass spectrometry data for omics studies. In S. Datta & B. J. A. Mertens (Eds.), *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry* (pp. 1–21). Springer. https://doi.org/10.1007/978-3-319-45809-0_1

Misra, B. B. (2020). Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *European Journal of Mass Spectrometry*, *26*(3), 165–174. https://doi.org/10.1177/1469066720918446

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., et al. (2022). Vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan

Pang, Z., Lu, Y., Zhou, G., Hui, F., Xu, L., Viau, C., Spigelman, A. F., MacDonald, P. E., Wishart, D. S., Li, S., & Xia, J. (2024). MetaboAnalyst 6.0: Towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research, 52*(W1), W398–W406. https://doi.org/10.1093/nar/gkae253

Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. (2015). Computational and statistical analysis of metabolomics data. *Metabolomics*, *11*(6), 1492–1513. https://doi.org/10.1007/s11306-015-0823-6

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Rivera, E. S., LeBrun, E. S., Breidenbach, J. D., Solomon, E., Sanders, C. K., Harvey, T., Tseng, C. Y., Thornhill, M. G., Blackwell, B. R., McBride, E. M., & Luchini, K. A. (2024). Feature-agnostic metabolomics for determining effective subcytotoxic doses of common pesticides in human cells. *Toxicological Sciences, 202*(1), 85–95. https://doi.org/10.1093/toxsci/kfae101

Rodríguez-Pérez, R., Fernáández, L., & Marco, S. (2018). Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: A systematic study. *Analytical and Bioanalytical Chemistry*, *410*(23), 5981–5992. https://doi.org/10.1007/s00216-018-1217-1

Rubingh, C. M., Bijlsma, S., Jellema, R. H., Overkamp, K. M., van der Werf, M. J., & Smilde, A. K. (2009). Analyzing longitudinal microbial metabolomics data. *Journal of Proteome Research, 8*(9), 4319–4327. https://doi.org/10.1021/pr900126e

Sindelar, M., Stancliffe, E., Schwaiger-Haber, M., Anbukumar, D. S., Adkins-Travis, K., Goss, C. W., O'Halloran, J. A., Mudd, P. A., Liu, W. C., Albrecht, R. A., & García-Sastre, A. (2021). Longitudinal metabolomics of human plasma reveals prognostic markers of COVID-19 disease severity. *Cell Reports Medicine*. https://doi.org/10.1016/j.xcrm.2021.100369

Smith, R., Mathis, A. D., Ventura, D., & Prince, J. T. (2014). Proteomics, lipidomics, metabolomics: A mass spectrometry tutorial from A computer scientist's point of view. *Bmc Bioinformatics*, *15*(S7), S9. https://doi.org/10.1186/1471-2105-15-S7-S9

Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H., Okahashi, N., Yamada, Y., Tada, I., Bonini, P., Higashi, Y., Okazaki, Y., Zhou, Z., Zhu, Z. J., Koelmel, J., Cajka, T., Fiehn, O., Saito, K., Arita, M., & Arita, M. (2020). A lipidome atlas in MS-DIAL 4. *Nature Biotechnology*, *38*(10), 1159–1163. https://doi.org/10.1038/s41587-020-0531-2

Välikangas, T., Suomi, T., & Elo, L. L. (2018). A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics*, *19*(1), 1–11. https://doi.org/10.1093/bib/bbw095

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. (2011). Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry, 83*(15), 5864–5872. https://doi.org/10.1021/ac201065j

Wang, M., Jiang, L., Jian, R., Chan, J. Y., Liu, Q., Snyder, M. P., & Tang, H. (2021). RobNorm: Model-based robust normalization method for labeled quantitative mass spectrometry proteomics data. *Bioinformatics*, *37*(6), 815–821. https://doi.org/10.1093/bioinformatics/btaa904

Wang, S. Y., Kuo, C. H., & Tseng, Y. J. (2013). Batch normalizer: A fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid Chromatography/Time-of-Flight mass Spectrometry-Based metabolomics data and comparison with current calibration methods. *Analytical Chemistry*, *85*(2), 1037–1046. https://doi.org/10.1021/ac302877x

Wehrli, P. M., Michno, W., Blennow, K., Zetterberg, H., & Hanrieder, J. (2019). Chemometric strategies for sensitive annotation and validation of anatomical regions of interest in complex imaging mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, *30*(11), 2278–2288. https://doi.org/10.1007/s13361-019-02327-y

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood Estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Xi, B., Gu, H., Baniasadi, H., & Raftery, D. (2014). Statistical analysis and modeling of mass spectrometry-based metabolomics data. In D. Raftery (Ed.), *Mass spectrometry in metabolomics: Methods and protocols.* Springer. https://doi.org/10.1007/978-1-4939-1258-2_22

Yang, Q., Wang, Y., Zhang, Y., Li, F., Xia, W., Zhou, Y., Qiu, Y., Li, H., & Zhu, F. (2020). NOREVA: Enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Research, 48*(W1), W436–W448. https://doi.org/10.1093/nar/gkaa258

Zhang, X., Cui, X., & Zhu, F. (2019). The effects of sample size on omics study: From the perspective of robustness and diagnostic accuracy. *IOP Conference Series: Earth and Environmental Science*, *252*(2), 022127. https://doi.org/10.1088/1755-1315/252/2/022127

Zhang, Y., Fan, S., Wohlgemuth, G., & Fiehn, O. (2023). Denoising autoencoder normalization for Large-Scale untargeted

metabolomics by gas Chromatography-Mass spectrometry. *Metabolites*, *13*(8), 944. https://doi.org/10.3390/metabo13080944

Zhao, Y., Wong, L., & Goh, W. W. B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Scientific Reports*, *10*(1), 15534. https://doi.org/10.1038/s41598-020-72664-6

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.