

Review

# A Review on Differential Abundance Analysis Methods for Mass Spectrometry-Based Metabolomic Data

Zhengyan Huang <sup>1,\*</sup> and Chi Wang <sup>2,\*</sup><sup>1</sup> Everest Clinical Research Corporation, Little Falls, NJ 07424, USA<sup>2</sup> Markey Cancer Center, Department of Internal Medicine, University of Kentucky, Lexington, KY 40536, USA

\* Correspondence: xanthehuang@gmail.com (Z.H.); chi.wang@uky.edu (C.W.)

**Abstract:** This review presents an overview of the statistical methods on differential abundance (DA) analysis for mass spectrometry (MS)-based metabolomic data. MS has been widely used for metabolomic abundance profiling in biological samples. The high-throughput data produced by MS often contain a large fraction of zero values caused by the absence of certain metabolites and the technical detection limits of MS. Various statistical methods have been developed to characterize the zero-inflated metabolomic data and perform DA analysis, ranging from simple tests to more complex models including parametric, semi-parametric, and non-parametric approaches. In this article, we discuss and compare DA analysis methods regarding their assumptions and statistical modeling techniques.

**Keywords:** differential abundance; mass spectrometry; metabolomics; zero-inflated data



**Citation:** Huang, Z.; Wang, C. A Review on Differential Abundance Analysis Methods for Mass Spectrometry-Based Metabolomic Data. *Metabolites* **2022**, *12*, 305. <https://doi.org/10.3390/metabo12040305>

Academic Editor: Georgios Theodoridis

Received: 28 February 2022

Accepted: 27 March 2022

Published: 30 March 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Metabolomics has become a mature science, with over 20 years since it was first coined in 1998 [1–3]. It is the study of small molecules, known as metabolites, of chemical reactions within a biological system, which directly reflects the biochemical activity and provides insights into the underlying status of the system [4]. As a key component of the omics cascade, metabolomics best represents the molecular phenotype [5,6].

Even though the diverse nature of metabolites remains a challenge in compound identification and reliable quantification, metabolomics is routinely applied to multiple disciplines in life science with the advances in Mass Spectrometry (MS) [7]. Together with its various techniques, MS has high sensitivity, high mass resolution and accuracy, and the capability to detect and quantify numerous metabolites simultaneously [7–9]. The common applications of MS-based metabolomics include but are not limited to metabolite structure elucidation [10–12], metabolic profiling [10,13–15], and metabolite identification [16–19].

Despite the advances that have been achieved, MS-based approaches still have detection limits, which can complicate metabolite identification and quantification [7,9,20]. The diversity of metabolites, including varied chemical structure, unclear scope of metabolic network, and dynamic range of abundance, can cause those detection limits [7,21]. One frequently seen characteristic of high-throughput MS-based metabolomics data is zero inflation, where the zero values are due to either the absent of the metabolites, abundance levels below the detection limits, or both. The zero values are referred to as point mass values (PMVs) and non-zero values are referred to as non-PMVs [22]. To distinguish the zero values caused by the two different reasons, PMVs are further classified as biological point mass values (BPMVs) and technical point mass values (TPMV). BPMVs exist if metabolites are absent in the experimental sample for a biological reason, and TPMVs exist if metabolites present in the sample but the signal is below the detection limit for a technical reason [22,23].

The proportion of PMVs can be very large. Do et al. (2018) reported an overall missing rate of 19.41%, with 80.6% metabolites that had at least one PMV. Among those

metabolites, about 10% had a rate of PMV over 70%. The average missing rate per observation is 19.6% [24]. In the study conducted by Faquih et al. (2020), the authors reported 58.6% metabolites had at least one PMV with an average PMV rate per observation at 38% [25]. Taylor et al. (2013) summarized the PMV rate in metabolomic, proteomic, and glycomic studies. The overall PMV rate for metabolomics data sets ranges from 14.63% to 28.53% [26,27]. In addition to the large proportion of PMVs, studies have also confirmed that MS-based omics data can be missing not at random (MNAR), which is caused by the censored values due to detection limits [26,27].

The large proportion of PMVs has a substantial impact on the downstream analysis as ignoring the PMVs can lead to biased results. In addition, the two types of PMVs are hard to separate during the experimental process due to detection limits. Appropriate statistical methods are required to characterize PMVs and distinguish BPMVs and TPMVs to ensure unbiased and efficient inference.

Another important issue for downstream statistical analysis is how to model the non-PMVs. Li et al. (2019) found that the non-PMVs of many metabolites in a metabolomic dataset were not normally distributed even after log-transformation [28]. As many parametric models require data normality assumption, this finding raises cautions about the choice of statistical models for robust analysis.

A major type of downstream statistical analysis for metabolomic data is the DA analysis, which identifies differentially abundant metabolic features between samples from different experimental groups. In this review, we focus on statistical methods for DA analysis and discuss the pros and cons of each method regarding their assumptions and statistical modeling techniques.

## 2. Statistical Methods for DA Analysis

Naïve approaches for DA analysis include ignoring the PMVs or imputing the PMVs with non-zero values. Specifically, one approach is to delete the PMVs and apply standard methods, such as two-sample *t*-test [29] or moderated *t*-test [22,30], to the non-PMVs. However, ignoring the zero values changes the distribution of abundance level under consideration so that the results can be biased. The other approach is data imputation, which is frequently used to handle missing data including the zero-inflation issue. There are some normalization and imputation methods developed for MS data [25–27,31,32]. Once the zero values are imputed, the data can be analyzed using standard statistical methods such as two-sample *t*-tests. However, as we have mentioned above, due to the complex mechanisms and MNAR nature of the data, imputation methods need to be applied case by case. It is difficult to identify a suitable imputation method for a given dataset, and an inappropriate method could induce unreliable results and inferences [27,33,34].

Statistical models that can account for zero values without the need of imputation have been developed to handle different types of zero-inflated data, where zero-inflation presents not only in metabolomic studies but also in many other medical, health care, and economical studies [35–38]. Two types of zero-inflated data are frequently seen in practice; one is zero-inflated count data and the other is zero-inflated nonnegative continuous data. A recent review summarized zero-inflated count models and their applications [39]. Reviews on zero-inflated nonnegative continuous data are also available [40,41].

In this review, we focus on statistical models that have been used to handle MS-based metabolomics data. Based on the strategy of modeling PMVs and non-PMVs, these methods can be classified into three categories: one-part tests, two-part tests, and mixture models [22]. In the following sub-sections, we summarize the methods in each category. For convenience, we first introduce the following notations. Let  $Y_{ij}$  be the log-transformed abundance level and  $\delta_{ij}$  be the PMV indicator ( $\delta_{ij} = 1$  if PMV or  $\delta_{ij} = 0$  if non-PMV) for the  $j$ th metabolite from the  $i$ th subject, respectively,  $\lambda_j$  be the detection limit for the  $j$ th metabolite, and  $X_i$  be a vector of covariates for the  $i$ th subject.

### 2.1. One-Part Tests

A one-part test considers the whole distribution of metabolite data that does not separately model PMVs and non-PMVs. It uses a single test statistic that accounts for both PMVs and non-PMVs to compare a metabolite's abundance level between experimental groups.

#### 2.1.1. Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test was first introduced by Wilcoxon in 1945 [42] for two-group comparison problems. It is often applied when the distribution of continuous measures is not normal as an alternative non-parametric option of the two-sample *t*-test. Let  $n_1$  and  $n_2$  be the number of subjects in groups 1 and 2, respectively. The test statistic for comparing the abundance of metabolite  $j$  between groups is

$$W_j = \frac{|U_j - \mu_U| - 0.5}{\sigma_U} \quad (1)$$

where  $U_j = n_1 n_2 + n_1(n_1 + 1)/2 - \sum_{i \in \text{Group } 1} r(Y_{ij})$ ,  $r(Y_{ij})$  is the rank of  $Y_{ij}$  among all observations of metabolite  $j$ ,  $\mu_U = (n_1 n_2)/2$  is the mean of  $U_j$  under the null hypothesis of no difference between groups, and  $\sigma_U = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$  is the standard deviation. For MS-based metabolomics data, since there are tied ranks largely due to PMVs,  $\sigma_U$  needs to be adjusted as follows:

$$\sigma'_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum_{k=1}^{K_j} (t_{kj}^3 - t_{kj})}{12(n_1 + n_2)(n_1 + n_2 - 1)}} \quad (2)$$

where  $K_j$  is the total number of unique ranks and  $t_{kj}$  is number of ties for the  $k$ th rank for the  $j$ th metabolite.

#### 2.1.2. Truncated Wilcoxon-Test

The truncated Wilcoxon-test was proposed by Hallstrom in 2010 to handle zero-inflated data for two group comparison with equal sample size [43]. The Wilcoxon rank-sum test is performed after an equal and maximal amount of zeros are removed from each group to gain power. The method was extended to data with unequal sample size by Wang et al. (2021) [44]. Assuming the equal and maximal amount of zero observations are removed from each group,  $n'_1$  and  $n'_2$  observations are left. The test statistic is calculated using equations in Section 2.1.1 with  $n_1$  and  $n_2$  to be replaced by  $n'_1$  and  $n'_2$ .

#### 2.1.3. Tobit-Model

The Tobit-model [22] assumes PMVs are TPMVs caused by left censoring at the detection limit. It models data by a left-censored normal distribution. The log likelihood function for metabolite  $j$  is:

$$\log L(\mu_j, \sigma_j) = \sum_{i: \delta_{ij}=0} \log \left\{ \frac{1}{2\pi\sigma_j} \varphi \left( \frac{Y_{ij} - \mu_{ij}}{\sigma_j} \right) \right\} + \sum_{i: \delta_{ij}=1} \log \left\{ \Phi \left( \frac{\lambda_j - \mu_{ij}}{\sigma_j} \right) \right\} \quad (3)$$

where  $\mu_{ij} = \beta_{0j} + I(i \in \text{Group } 2)\beta_{1j}$ ,  $\sigma_j$  is the standard deviation, and  $\varphi(\cdot)$  and  $\Phi(\cdot)$  are density and cumulative distribution functions of the standard normal distribution, respectively. A likelihood ratio test is applied to test the hypothesis of  $\beta_{1j} = 0$  for DA analysis.

### 2.2. Two-Part Tests

A two-part test first uses two independent test statistics, one for assessing the difference in non-PMVs and the other for assessing the difference in PMVs, and then combines the two test statistics to determine the overall difference between experimental groups [22,45].

A two-part test explicitly compares the proportion of PMVs between groups, although it does not further separate PMVs into BPMVs and TPMVs.

### 2.2.1. Two-Part t-Test

For PMVs, a Pearson's Chi-square test statistic is applied to compare the zero proportion between the two groups. For non-PMVs, a *t*-test is applied on non-zero values to get the test statistic. The test statistics for PMVs and non-PMVs both follow the chi-square distribution with 1 degree of freedom (d.f.). Assuming the proportion of PMVs is not 0 and not 1 in both groups, the pooled test statistic, the Pearson's Chi-square test statistic plus the square of the *t*-test statistic, follows a chi-square distribution with 2 d.f.s [22].

### 2.2.2. Two-Part Wilcoxon Test

The two-part Wilcoxon test is constructed similarly to the two-part *t*-test, except that it uses a Wilcoxon rank-sum test instead of a *t*-test for non-PMVs [22].

### 2.2.3. SDA

Li et al. (2019) [28] proposed a semi-parametric approach named semi-parametric differential abundance analysis (SDA), which applies a logistic regression for the PMVs (Equation (4)) and a semi-parametric model (Equation (5)) for the non-PMVs:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_{0j} + \gamma_j \mathbf{X}_i, \quad (4)$$

$$Y_{ij} = \beta_j \mathbf{X}_i + \varepsilon_{ij}, \quad (5)$$

where  $\gamma_j$  and  $\beta_j$  are the covariates' effects for *j*th metabolite for the PMVs and non-PMVs, respectively. In Equation (5), the distribution of the independent error term  $\varepsilon_{ij}$  is unspecified, which allows the metabolite abundance level to be arbitrarily distributed that can deviate from the normal distribution. SDA considers the following kernel-smoothed likelihood for parameter estimation:

$$L(\beta_j, \gamma_j, \gamma_{0j}) = \prod_{i=1}^N \left[ \frac{\exp(\gamma_{0j} + \gamma_j \mathbf{X}_i)}{1 + \exp(\gamma_{0j} + \gamma_j \mathbf{X}_i)} \right]^{I(\delta_{ij}=1)} \times \left[ \frac{\frac{1}{Nh} \sum_{i^*=1}^N K\left\{ \frac{Y_{i^*j} - \beta_j \mathbf{X}_{i^*} - (Y_{ij} - \beta_j \mathbf{X}_i)}{h} \right\}}{\log(Y_{ij}) \{1 + \exp(\gamma_{0j} + \gamma_j \mathbf{X}_i)\}} \right]^{I(\delta_{ij}=0)} \quad (6)$$

where  $\frac{1}{Nh} \sum_{i^*=1}^N K\left\{ \frac{Y_{i^*j} - \beta_j \mathbf{X}_{i^*} - (Y_{ij} - \beta_j \mathbf{X}_i)}{h} \right\}$  is the kernel density estimator with  $K(\cdot)$  as a one dimensional kernel function,  $h$  as the bandwidth, and  $N$  as the sample size. For DA analysis on the effect of a covariate, SDA assesses whether the corresponding model coefficients in  $\gamma_j$  and  $\beta_j$  are equal to zero based on a likelihood ratio test.

## 2.3. Mixture Models

The mixture model considers PMVs as a mixture of BPMVs and TPMVs, where the TPMVs component is quantified by the left censoring probability from a parametric model on non-BPMVs (including both TPMVs and non-PMVs). As the mixture model clearly separates BPMVs and TPMVs, it provides sufficient flexibility for comparing the proportion of BPMVs, proportion of TPMVs, and mean of non-BPMVs between groups, although a parametric model assumption is required to characterize the distribution of non-BPMVs.

### 2.3.1. Left-Inflated Mixture Likelihood Ratio Test (LIM-LRT)

The left-inflated mixture model (LIM) combines a Bernoulli distribution and a left-censored normal distribution. It has been applied to many studies including

omics [22,26,46–49]. Specifically, the distribution of abundance of metabolite  $j$  for subject  $i$  from group  $g$  ( $g = 1$  or  $2$ ) has the following density function:

$$f(Y_{ij}|p_{jg}, \mu_{jg}, \sigma_{jg}) = \begin{cases} p_{jg} + (1 - p_{jg})\Phi\left(\frac{\lambda_j - \mu_{jg}}{\sigma_{jg}^2}\right), & \text{if } \delta_{ij} = 1 & (7) \\ (1 - p_{jg})\varphi\left(\frac{Y_{ij} - \mu_{jg}}{\sigma_{jg}^2}\right), & \text{if } \delta_{ij} = 0 & (8) \end{cases}$$

where  $\mu_{jg}$  is the mean,  $\sigma_{jg}$  is the standard deviation, and  $p_{jg}$  and  $(1 - p_{jg})\Phi(\lambda_j|\mu_{jg}, \sigma_{jg})$  are the proportions of BPMVs and TPMVs, respectively, for metabolite  $j$  from group  $g$ . Based on Equation (8), non-PMVs follow a truncated normal distribution:

$$f(Y_{ij}|\delta_{ij} = 0, \mu_{jg}, \sigma_{jg}) = \frac{\varphi\left((Y_{ij} - \mu_{jg})/\sigma_{jg}^2\right)}{\sigma_{jg}\left(1 - \Phi\left((\lambda_j - \mu_{jg})/\sigma_{jg}^2\right)\right)} \quad (9)$$

A likelihood ratio test (LIM-LRT) for the hypothesis of  $\mu_{j1} = \mu_{j2}$  and  $p_{j1} = p_{j2}$  is used to assess whether metabolite  $j$  is differentially abundant between groups.

### 2.3.2. DASEV

Huang et al. (2020) noticed that the variance estimation from LIM could be unstable in presence of a large proportion of zero values, which affected the DA analysis results [50]. To address this issue, they adapted the variance shrinkage approach proposed by Smyth (2004) for microarray data to the mixture model setting, where data from the ensemble of metabolites were borrowed to achieve a more robust variance estimation of each individual metabolite [30]. Specifically, the variances of all metabolites,  $\sigma_j^2$ 's, are assumed to have the following common prior distribution:

$$\sigma_j^2 \sim \text{Inv} - \text{Gamma}\left(\frac{d_0}{2}, \frac{d_0 s_0^2}{2}\right), \quad (10)$$

where  $d_0/2$  and  $d_0 s_0^2/2$  are the shape and scale parameters for the inverse-gamma distribution, respectively. The  $d_0$  and  $s_0$  are specified as follows:

$$d_0 = 2m^2/v + 4, \quad (11)$$

$$s_0 = \sqrt{m(d_0 - 2)/d_0}, \quad (12)$$

where  $m$  and  $v$  are the sample mean and variance for the initial estimate of  $\sigma_j^2$  across all metabolites. After the shape and scale parameters are determined, iterations are done until convergence to obtain estimates of  $\hat{p}_{jg}$  and  $\hat{\mu}_{jg}$  by maximizing the likelihood:

$$L(p_{j1}, p_{j2}, \mu_{j1}, \mu_{j2}|\sigma_j) = \prod_{g=1}^2 \left\{ \prod_{i \in \text{Group } g} f(Y_{ij}|p_{jg}, \mu_{jg}, \sigma_j) \right\}, \quad (13)$$

and  $\hat{\sigma}_j^2$  by maximizing the posterior:

$$p\left(\sigma_j^2 | \text{Data}\right) \propto L(p_{j1}, p_{j2}, \mu_{j1}, \mu_{j2}|\sigma_j) \left(\frac{d_0 s_0^2}{2}\right)^{\frac{d_0}{2}} \frac{\sigma_j^{2(-1-\frac{d_0}{2})}}{\Gamma\left(\frac{d_0}{2}\right)} \exp\left(-\frac{d_0 s_0^2}{2\sigma_j^2}\right). \quad (14)$$

After all model estimates are obtained until convergence, a likelihood ratio test is applied for DA analysis. Huang et al. (2020) also extended LIM to allow covariate adjustment, where a logistic regression model was used to characterize the association between

covariates and the proportion of BPMVs and a linear model was used to characterize the association between covariates and the mean of non-BPMVs [50].

#### 2.4. Model Comparison

Simulation studies that compared the performance of different methods were conducted in the literature. Gleiss et. al. compared models including Wilcoxon rank-sum test, truncated Wilcoxon test, Tobit-model, two-part *t*-test, two-part Wilcoxon test, and LIM-LRT, Huang et al. compared LIM-LRT and DASEV, and Li et al. compared two-part *t*-test, two-part Wilcoxon test, and SDA [22,28,50]. In summary, if the proportion of BPMVs are similar between the two groups, one-part tests generate acceptable results. Two-part tests have more reliable estimates comparing to one-part tests especially when the TPMVs proportions are not too high [22]. Two-part Wilcoxon test shows good performance if TPMVs can be ruled out [22]. SDA is able to handle both normally and non-normally distributed features simultaneously, and outperforms two-part *t*-test and two-part Wilcoxon test for non-normally distributed features [28]. Mixture models, especially DASEV, can provide less biased estimates on both proportions of TPMVs and BPMVs when the TPMV proportion is high [22,50]. LIM-LRT, DASEV, and SDA all yield good true positive rates when the PMV proportion is not very high [22,28,50].

### 3. Practical Guidelines

Table 1 summarizes the modeling technique and assumption for each DA analysis method. In practice, the choice of appropriate methods to use depends on the characteristics of the specific dataset. The following factors need to be considered.

**Table 1.** Comparison of statistical methods for DA analysis.

Category	Method	Able to Distinguish TPMVs and BPMVs	Free of Data Normality Assumption	Available R Function/Package	References
One-part test	Wilcoxon rank-sum test	N	Y	wilcox.test	[42]
	Truncated Wilcoxon test	N	Y	<a href="https://rdrr.io/github/chvlyl/ZIR/">https://rdrr.io/github/chvlyl/ZIR/</a>	[43,44]
	Tobit-model	N	N	VGAM ( <a href="https://cran.r-project.org/web/packages/VGAM/index.html">https://cran.r-project.org/web/packages/VGAM/index.html</a> )	[22]
Two-part test	Two-part <i>t</i> -test	N	N	t.test binom.test	[22]
	Two-part Wilcoxon test	N	Y	wilcox.test binom.test	[22]
	SDA	N	Y	SDAMS ( <a href="https://bioconductor.org/packages/release/bioc/html/SDAMS.html">https://bioconductor.org/packages/release/bioc/html/SDAMS.html</a> )	[28]
Mixture Model	LIM-LRT	Y	N	<a href="https://cemsii.meduniwien.ac.at/en/kb/science-research/software/statistical-software/limlrt/">https://cemsii.meduniwien.ac.at/en/kb/science-research/software/statistical-software/limlrt/</a>	[22,26,46,47]
	DASEV	Y	N	<a href="http://sweb.uky.edu/~cwa236/DASEV.html">http://sweb.uky.edu/~cwa236/DASEV.html</a>	[50]

Y: Yes; N: No. All the hyperlinks were accessed on 25 March 2022.

The composition of PMVs. As different methods model PMVs in different ways, we would suggest first investigating the composition of PMVs before performing DA analysis. One can draw a histogram to investigate the empirical distribution of abundance level. If the observed PMV proportion is substantially higher than the extrapolation of the distribution of non-PMVs, it would indicate the presence of BPMVs. Under such situation, the Tobit-model, which assumes PMVs are all from TPMVs, may not be appropriate. Further, if one wants to separate the proportions of BPMVs and TPMVs, the mixture model-based approaches, LIM-LRT and DASEV, would be preferred.

Data normality. We would also suggest checking data normality by using the Q-Q plot, Kolmogorov-Smirnov test, and Shapiro-Wilk test. If the data substantially deviate from normal distributions, non-parametric and semi-parametric methods that do not require the normal assumption would be preferred. Those methods include Wilcoxon rank-sum test, truncated Wilcoxon test, two-part Wilcoxon test, and SDA.

Sample size. Although non-parametric and semi-parametric methods are robust to distributional assumptions, they typically require larger sample sizes compared to parametric methods. For example, the Wilcoxon rank sum test requires a sample size of at least 16 [51,52]. Therefore, if the experiment only has a few replicates per treatment group, using a parametric method is more feasible.

Confounder adjustment. Adjusting for confounders, e.g., age and sex, is allowed for some parametric and semi-parametric methods such as DASEV and SDA. Therefore, for studies with a complex design and/or presence of confounders, those methods would be preferred.

Finally, it is always a good practice to consider more than one method and compare the results to make more robust inference.

#### 4. Discussion

Handling zero inflation is an important task for analyzing MS-based metabolomic data. The characteristics of zero-inflated data need to be carefully assessed in order to choose appropriate statistical methods for data analysis, which will impact analysis results and interpretation. In this paper, we have reviewed a variety of statistical methods to model zero-inflated data for DA analysis. By comparing these methods in the aspects of assumptions and statistical modeling techniques, we have provided guidelines for choosing appropriate methods in practical situations. Our review focuses on cross-sectional studies. For the more complex longitudinal metabolomics studies on the progression of diseases [53–55], current approaches consider mixed effect models [56–58]. New method developments to handle the zero-inflation issue are highly desired to achieve more robust performance and increase the predictability of such studies.

In addition to DA analysis, the zero inflation issue also broadly affects many other types of downstream analysis of metabolomic data such as cluster analysis [59], disease diagnostic modeling [60], pathway analysis [61–63], and multi-omics analysis [64]. For example, a common approach for pathway analysis is the overrepresentation analysis [61,63], which identifies enrichment of a metabolic pathway by assessing the overrepresentation of metabolites from the pathway in a list of metabolites of interest compared to the background. The overrepresentation analysis is based on an input of a list of metabolites of interest, which is usually the list of differentially abundant metabolites from a DA analysis. Thus, the strategy of handling PMVs in DA analysis will have an impact on the results of pathway analysis.

**Author Contributions:** Conceptualization, methodology, writing—review and editing: Z.H. and C.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Cancer Institute (R03CA211835) and the Biostatistics and Bioinformatics Shared Resource Facility of the University of Kentucky Markey Cancer Center (P30CA177558).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Oliver, S.G.; Winson, M.K.; Kell, D.B.; Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **1998**, *16*, 373–378. [[CrossRef](#)]
2. Alseekh, S.; Fernie, A.R. Metabolomics 20 years on: What have we learned and what hurdles remain? *Plant J.* **2018**, *94*, 933–942. [[CrossRef](#)] [[PubMed](#)]
3. Trivedi, D.K.; Hollywood, K.A.; Goodacre, R. Metabolomics for the masses: The future of metabolomics in a personalized world. *New Horiz. Transl. Med.* **2017**, *3*, 294–305. [[CrossRef](#)] [[PubMed](#)]
4. Liu, X.; Locasale, J.W. Metabolomics: A Primer. *Trends Biochem. Sci.* **2017**, *42*, 274–284. [[CrossRef](#)]
5. Guijas, C.; Montenegro-Burke, J.R.; Warth, B.; Spilker, M.E.; Siuzdak, G. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat. Biotechnol.* **2018**, *36*, 316–320. [[CrossRef](#)]
6. Sinem, N.; Abdullah, K. *Introductory Chapter: Insight into the OMICS Technologies and Molecular Medicine*; Sinem, N., Hakima, A., Eds.; Molecular Medicine; IntechOpen: London, UK, 2019.
7. Alseekh, S.; Aharoni, A.; Brotman, Y.; Contrepolis, K.; D’Auria, J.; Ewald, J.; Ewald, J.C.; Fraser, P.D.; Giavalisco, P.; Hall, R.D.; et al. Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. *Nat. Methods* **2021**, *18*, 747–756. [[CrossRef](#)]
8. Dunn, W.B. Mass spectrometry in systems biology an introduction. *Methods Enzym.* **2011**, *500*, 15–35.
9. Aretz, I.; Meierhofer, D. Advantages and Pitfalls of Mass Spectrometry Based Metabolome Profiling in Systems Biology. *Int. J. Mol. Sci.* **2016**, *17*, 632. [[CrossRef](#)]
10. Saghatelian, A.; Trauger, S.A.; Want, E.J.; Hawkins, E.G.; Siuzdak, G.; Cravatt, B.F. Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* **2004**, *43*, 14332–14339. [[CrossRef](#)]
11. Boiteau, R.M.; Hoyt, D.W.; Nicora, C.D.; Kinmonth-Schultz, H.A.; Ward, J.K.; Bingol, K. Structure Elucidation of Unknown Metabolites in Metabolomics by Combined NMR and MS/MS Prediction. *Metabolites* **2018**, *8*, 8. [[CrossRef](#)]
12. Levsen, K.; Schiebel, H.M.; Behnke, B.; Dötzer, R.; Dreher, W.; Elend, M.; Thiele, H. Structure elucidation of phase II metabolites by tandem mass spectrometry: An overview. *J. Chromatogr. A* **2005**, *1067*, 55–72. [[CrossRef](#)]
13. Dunn, W.B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J.D.; Halsall, A.; Haselden, J.N.; et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **2011**, *6*, 1060–1083. [[CrossRef](#)] [[PubMed](#)]
14. Shao, Y.; Li, T.; Liu, Z.; Wang, X.; Xu, X.; Li, S.; Xu, G.; Le, W. Comprehensive metabolic profiling of Parkinson’s disease by liquid chromatography-mass spectrometry. *Mol. Neurodegener.* **2021**, *16*, 4. [[CrossRef](#)] [[PubMed](#)]
15. Clarke, C.J.; Haselden, J.N. Metabolic profiling as a tool for understanding mechanisms of toxicity. *Toxicol. Pathol.* **2008**, *36*, 140–147. [[CrossRef](#)] [[PubMed](#)]
16. Lapainis, T.; Rubakhin, S.S.; Sweedler, J.V. Capillary electrophoresis with electrospray ionization mass spectrometric detection for single-cell metabolomics. *Anal. Chem.* **2009**, *81*, 5858–5864. [[CrossRef](#)] [[PubMed](#)]
17. Prasad, B.; Garg, A.; Takwani, H.; Singh, S. Metabolite identification by liquid chromatography-mass spectrometry. *TRAC Trends Anal. Chem.* **2011**, *30*, 360–387. [[CrossRef](#)]
18. Xiao, J.F.; Zhou, B.; Ransom, H.W. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Anal. Chem. TRAC* **2012**, *32*, 1–14. [[CrossRef](#)]
19. Dahal, U.P.; Jones, J.P.; Davis, J.A.; Rock, D.A. Small molecule quantification by liquid chromatography-mass spectrometry for metabolites of drugs and drug candidates. *Drug Metab. Dispos.* **2011**, *39*, 2355–2360. [[CrossRef](#)]
20. Easterling, L.F.; Yerabolu, R.; Kumar, R.; Alzarzeni, K.Z.; Kenttämaa, H.I. Factors Affecting the Limit of Detection for HPLC/Tandem Mass Spectrometry Experiments Based on Gas-Phase Ion-Molecule Reactions. *Anal. Chem.* **2020**, *92*, 7471–7477. [[CrossRef](#)]
21. Lu, W.; Su, X.; Klein, M.S.; Lewis, I.A.; Fiehn, O.; Rabinowitz, J.D. Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *Annu. Rev. Biochem.* **2017**, *86*, 277–304. [[CrossRef](#)]
22. Gleiss, A.; Dakna, M.; Mischak, H.; Heinze, G. Two-group comparisons of zero-inflated intensity values: The choice of test statistic matters. *Bioinformatics* **2015**, *31*, 2310–2317. [[CrossRef](#)] [[PubMed](#)]
23. Dakna, M.; Harris, K.; Kalousis, A.; Carpentier, S.; Kolch, W.; Schanstra, J.P.; Haubitz, M.; Vlahou, A.; Mischak, H.; Girolami, M. Addressing the challenge of defining valid proteomic biomarkers and classifiers. *BMC Bioinform.* **2010**, *11*, 594. [[CrossRef](#)] [[PubMed](#)]
24. Do, K.T.; Wahl, S.; Raffler, J.; Molnos, S.; Laimighofer, M.; Adamski, J.; Suhre, K.; Strauch, K.; Peters, A.; Gieger, C.; et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **2018**, *14*, 128. [[CrossRef](#)] [[PubMed](#)]
25. Faquih, T.; van Smeden, M.; Luo, J.; le Cessie, S.; Kastenmüller, G.; Krumsiek, J.; Noordam, R.; Van Heemst, D.; Rosendaal, F.R.; Vlieg, A.V.H.; et al. A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. *Metabolites* **2020**, *10*, 486. [[CrossRef](#)] [[PubMed](#)]
26. Taylor, S.L.; Leiserowitz, G.S.; Kim, K. Accounting for undetected compounds in statistical analyses of mass spectrometry ‘omic studies. *Stat. Appl. Genet. Mol. Biol.* **2013**, *12*, 703–722. [[CrossRef](#)] [[PubMed](#)]
27. Hrydziusko, O.; Viant, M.R. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. *Metabolomics* **2011**, *8*, 161–174. [[CrossRef](#)]

28. Li, Y.; Fan, T.W.M.; Lane, A.N.; Kang, W.Y.; Arnold, S.M.; Stromberg, A.J.; Wang, C.; Chen, L. SDA: A semi-parametric differential abundance analysis method for metabolomics and proteomics data. *BMC Bioinform.* **2019**, *20*, 501. [[CrossRef](#)]
29. Zhang, D.; Fan, C.; Zhang, J.; Zhang, C.H. Nonparametric methods for measurements below detection limit. *Stat. Med.* **2009**, *28*, 700–715. [[CrossRef](#)]
30. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 1–25. [[CrossRef](#)]
31. Wang, P.; Tang, H.; Zhang, H.; Whiteaker, J.; Paulovich, A.G.; McIntosh, M. Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Biocomputing* **2006**, *11*, 315–326.
32. Hughes, G.; Cruickshank-Quinn, C.; Reisdorph, R.; Lutz, S.; Petrache, I.; Reisdorph, N.; Bowler, R.; Kechris, K. MSPrep-summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. *Bioinformatics* **2014**, *30*, 133–134. [[CrossRef](#)] [[PubMed](#)]
33. Webb-Robertson, B.J.; Wiberg, H.K.; Matzke, M.M.; Brown, J.N.; Wang, J.; McDermott, J.E.; Smith, R.D.; Rodland, K.D.; Metz, T.O.; Pounds, J.G.; et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.* **2015**, *14*, 1993–2001. [[CrossRef](#)]
34. Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **2016**, *15*, 1116–1125. [[CrossRef](#)] [[PubMed](#)]
35. Liaqat, M.; Kamal, S.; Fischer, F.; Zia, N. Zero-inflated and hurdle models with an application to the number of involved axillary lymph nodes in primary breast cancer. *J. King Saud Univ.-Sci.* **2022**, *34*, 101932. [[CrossRef](#)]
36. Zhang, P.; Pitt, D.; Wu, X. A New Multivariate Zero-Inflated Hurdle Model with Applications in Automobile Insurance. *ASTIN Bull.* **2022**, 1–24. [[CrossRef](#)]
37. Lam, K.F.; Xue, H.; Bun Cheung, Y. Semiparametric Analysis of Zero-Inflated Count Data. *Biometrics* **2006**, *62*, 996–1003. [[CrossRef](#)]
38. Neelon, B.; O'Malley, A.J.; Smith, V.A. Modeling zero-modified count and semicontinuous data in health services research part 2: Case studies. *Stat. Med.* **2016**, *35*, 5094–5112. [[CrossRef](#)]
39. Young, D.S.; Roemmele, E.; Yeh, P. Zero inflated modeling part I: Traditional zero inflated count regression models, their applications, and computational tools. *WIREs Comput. Stat.* **2020**, *14*, e1541. [[CrossRef](#)]
40. Liu, L.; Shih, Y.-C.T.; Strawderman, R.L.; Zhang, D.; Johnson, B.A.; Chai, H. Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review. *Stat. Sci.* **2019**, *34*, 253–279. [[CrossRef](#)]
41. Min, Y.; Agresti, A. Modeling Nonnegative Data with Clumping at Zero: A Survey. *J. Iran. Stat. Soc.* **2002**, *1*, 7–33.
42. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [[CrossRef](#)]
43. Hallstrom, A.P. A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Stat. Med.* **2010**, *29*, 391–400. [[CrossRef](#)] [[PubMed](#)]
44. Wang, W.; Chen, E.Z.; Li, H. Truncated Rank-Based Tests for Two-Part Models with Excessive Zeros and Applications to Microbiome Data. *arXiv* **2021**, arXiv:2110.05368.
45. Taylor, S.; Pollard, K. Hypothesis tests for point-mass mixture data with application to 'omics data with many zero values. *Stat. Appl. Genet. Mol. Biol.* **2009**, *8*, 8. [[CrossRef](#)] [[PubMed](#)]
46. Yang, Y.; Simpson, D.G. Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data. *Stat. Methods Med. Res.* **2012**, *21*, 393–408. [[CrossRef](#)] [[PubMed](#)]
47. Moulton, L.H.; Halsey, N.A. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* **1995**, *51*, 1570–1578. [[CrossRef](#)]
48. Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J.N.; Ansong, C.; Heffron, F.; Metz, T.O.; Qian, W.-J.; Yoon, H.; et al. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **2009**, *25*, 2028–2034. [[CrossRef](#)]
49. Wu, S.H.; Black, M.A.; North, R.A.; Atkinson, K.R.; Rodrigo, A.G. A statistical model to identify differentially expressed proteins in 2D PAGE gels. *PLoS Comput. Biol.* **2009**, *5*, e1000509. [[CrossRef](#)]
50. Huang, Z.; Lane, A.N.; Fan, T.W.M.; Higashi, R.M.; Weiss, H.L.; Yin, X.; Wang, C. Differential Abundance Analysis with Bayes Shrinkage Estimation of Variance (DASEV) for Zero-Inflated Proteomic and Metabolomic Data. *Sci. Rep.* **2020**, *10*, 876. [[CrossRef](#)]
51. Dwivedi, A.K.; Mallawaarachchi, I.; Alvarado, L.A. Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Stat. Med.* **2017**, *36*, 2187–2205. [[CrossRef](#)]
52. Mundry, R.; Fischer, J. Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: Examples from Animal Behaviour. *Anim. Behav.* **1998**, *56*, 256–259. [[CrossRef](#)] [[PubMed](#)]
53. Tsonaka, R.; Signorelli, M.; Sabir, E.; Seyer, A.; Hettne, K.; Aartsma-Rus, A.; Spitali, P. Longitudinal metabolomic analysis of plasma enables modeling disease progression in Duchenne muscular dystrophy mouse models. *Hum. Mol. Genet.* **2020**, *29*, 745–755. [[CrossRef](#)] [[PubMed](#)]
54. Overmyer, K.A.; Shishkova, E.; Miller, I.J.; Balnis, J.; Bernstein, M.N.; Peters-Clarke, T.M.; Meyer, J.G.; Quan, Q.; Muehlbauer, L.K.; Trujillo, E.A.; et al. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Syst.* **2021**, *12*, 23–40.e7. [[CrossRef](#)]
55. 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.; et al. Longitudinal metabolomics of human plasma reveals prognostic markers of COVID-19 disease severity. *Cell Rep. Med.* **2021**, *2*, 100369. [[CrossRef](#)]
56. Jendoubi, T.; Ebbels, T.M.D. Integrative analysis of time course metabolic data and biomarker discovery. *BMC Bioinform.* **2020**, *21*, 11. [[CrossRef](#)] [[PubMed](#)]

57. Berk, M.; Ebbels, T.; Montana, G. A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics* **2011**, *27*, 1979–1985. [[CrossRef](#)]
58. Mei, Y.; Kim, S.B.; Tsui, K.-L. Linear-mixed effects models for feature selection in high-dimensional NMR spectra. *Expert Syst. Appl.* **2009**, *36*, 4703–4708. [[CrossRef](#)]
59. Rusilowicz, M.J.; Dickinson, M.; Charlton, A.J.; O’Keefe, S.; Wilson, J. MetaboClust: Using interactive time-series cluster analysis to relate metabolomic data with perturbed pathways. *PLoS ONE* **2018**, *13*, e0205968. [[CrossRef](#)]
60. Gowda, G.A.N.; Zhang, S.; Gu, H.; Asiago, V.; Shanaiah, N.; Raftery, D. Metabolomics-based methods for early disease diagnostics. *Expert Rev. Mol. Diagn.* **2008**, *8*, 617–633. [[CrossRef](#)]
61. Wieder, C.; Frainay, C.; Poupin, N.; Rodríguez-Mier, P.; Vinson, F.; Cooke, J.; Lai, R.P.; Bundy, J.G.; Jourdan, F.; Ebbels, T. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. *PLoS Comput. Biol.* **2021**, *17*, e1009105. [[CrossRef](#)]
62. Xia, J.; Wishart, D.S. MetPA: A web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* **2010**, *26*, 2342–2344. [[CrossRef](#)] [[PubMed](#)]
63. Marco-Ramell, A.; Palau-Rodríguez, M.; Alay, A.; Tulipani, S.; Urpi-Sarda, M.; Sanchez-Pla, A.; Andres-Lacueva, C. Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinform.* **2018**, *19*, 1. [[CrossRef](#)] [[PubMed](#)]
64. Jiang, D.; Armour, C.R.; Hu, C.; Mei, M.; Tian, C.; Sharpton, T.J.; Jiang, Y. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Front. Genet.* **2019**, *10*, 995. [[CrossRef](#)] [[PubMed](#)]