frontiers | Frontiers in Microbiomes

TYPE Original Research
PUBLISHED 15 October 2025
DOI 10.3389/frmbi.2025.1584516

OPEN ACCESS

EDITED BY
Yanni Sun,
City University of Hong Kong,
Hong Kong SAR, China

REVIEWED BY
Tahir Mahmood,
University of the West of Scotland,
United Kingdom
Jiajin Wei,
City University of Hong Kong,
Hong Kong SAR, China

*CORRESPONDENCE
Fentaw Abegaz
fentawabegaz@gmail.com

RECEIVED 27 February 2025
ACCEPTED 19 September 2025
PUBLISHED 15 October 2025

CITATION
Abegaz F, Abedini D, Dong L, Westerhuis JA, van Eeuwijk F, Bouwmeester H and Smilde AK (2025) Analysis of microbiome high-dimensional experimental design data using generalized linear models and ANOVA simultaneous component analysis. *Front. Microbiomes* 4:1584516. doi: 10.3389/frmbi.2025.1584516

Analysis of microbiome high-dimensional experimental design data using generalized linear models and ANOVA simultaneous component analysis

Fentaw Abegaz[1,2]*, Davar Abedini[1], Lemeng Dong[1], Johan A. Westerhuis[1], Fred van Eeuwijk[2], Harro Bouwmeester[1] and Age K. Smilde[1]

[1]Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands, [2]Biometris, Wageningen University Research, Wageningen, Netherlands

In microbiome studies, addressing the unique characteristics of sequence data—such as compositionality, zero inflation, overdispersion, high dimensionality, and non-normality—is crucial for accurate analysis. In addition, integrating experimental design elements into microbiome data analysis is important for understanding how factors such as treatment, time, and interactions affect microbial abundance. To achieve these objectives, we developed a new method that combines generalized linear models (GLMs) with ANOVA simultaneous component analysis (ASCA), which we term GLM-ASCA. This method aims to improve microbiome analysis by providing a more comprehensive understanding of differential abundance patterns in response to experimental conditions. GLM-ASCA models the unique characteristics of microbiome sequence data with GLMs and uses ASCA to effectively separate the effects of different experimental factors on microbial abundance. We evaluated GLM-ASCA using simulated data and subsequently applied it to real data to analyze the effect of nitrogen deficiency on root microbiome recruitment in tomato. Simulation studies demonstrated the effectiveness of GLM-ASCA in analyzing microbiome data in complex experimental designs, and the real-data application revealed valuable insights into the dynamics of microbial communities under nitrogen starvation, including the identification of beneficial bacterial species that promote tomato (*Solanum lycopersicum*) growth and health through nitrogen fixation.

KEYWORDS

generalized linear models, ANOVA simultaneous component analysis, experimental design, high dimensional microbiome data, differential abundance analysis, Tweedie model

1 Introduction

In microbiome research, high-throughput sequencing techniques such as amplicon sequencing (e.g., 16S rRNA gene sequencing) and whole-genome shotgun sequencing have become standard approaches for generating data from samples obtained from well-designed experiments aimed at understanding the mechanisms governing host–microbiome interactions (Zancarini et al., 2021). The microbiome count data produced through these sequencing methods typically exhibit distinct characteristics, including compositionality, zero inflation, overdispersion, high dimensionality, and non-normality. Various statistical tools have been developed to analyze microbiome data while addressing one or more of these characteristics, such as MaAsLin2: Multivariable Association with Linear Models 2 (Mallick et al., 2021) and LinDA: Linear Models for Differential Abundance (Zhou et al., 2022). While these tools implement univariate generalized linear models (GLMs) and are effective in identifying associations between individual features and covariates, they are limited in capturing the multivariate structure of the data or the joint effects of multiple factors across features.

It is also worthwhile to integrate treatment designs into statistical models to effectively address relevant research questions (Smilde et al., 2005). This facilitates precise accounting of the details of the treatment design, such as intervention time, treatment variations, and interactions between multiple factors, thereby allowing for an accurate assessment of how each factor and their combinations affect microbial abundance. However, in plant microbiome studies involving hundreds to thousands of correlated features and a limited number of samples, addressing experimental design elements such as treatment, time, and interactions, along with analyzing the specific characteristics of microbiome data, remains a considerable challenge.

In this context, several developments have incorporated techniques such as ANOVA-based partitioning of sources of variation using multivariate methods, with the aim of accounting for both the inherent data characteristics and the study design. One prominent method in this regard is ANOVA simultaneous component analysis (ASCA/ASCA+) (Smilde et al., 2005; Jansen et al., 2005; Thiel et al., 2017; Bertinetto et al., 2020; Martin and Govaerts, 2020; Thiel et al., 2023). ASCA/ASCA+ combines dimension reduction projection techniques with traditional linear statistical modeling to identify the main sources of variability in the resulting responses. It also provides visually interpretable representations of factor effects and their interactions, facilitating the interpretation of multivariate structures within the statistical model related to the experimental design (Smilde et al., 2005; Thiel et al., 2017). Moreover, ASCA/ASCA+ has been modified to cope with multivariate data in unbalanced multifactorial designs using weighted-effect ASCA (WE-ASCA) (Ali et al., 2020). ASCA+ has also been extended to analyze longitudinal data using linear mixed-effects models (Martin and Govaerts, 2020; Madssen et al., 2021; Jarmund et al., 2022). Furthermore, variable selection approaches have been implemented in VASCA (variable-selection ASCA) (Camacho et al., 2023) using permutation-based testing, and in GASCA (group-wise ASCA) (Saccenti et al., 2018) utilizing sparse group-wise principal component analysis (PCA).

Multivariate methodologies such as ASCA have proven useful in the analysis of metabolomics data, where linear models and PCA can be reasonably applied to continuous data. In the case of microbiome studies, however, the ASCA+ framework needs to be adapted to account for nonlinear and non-normally distributed data. In particular, extending the linear modeling approach in ASCA to GLMs, which assume an exponential family of probability distributions to accommodate data features such as counts, zero inflation, and overdispersion, would greatly enhance its applicability to microbiome data. In this work, we introduce GLM-ASCA (generalized linear models–ANOVA simultaneous component analysis), a novel approach for the analysis of microbiome data. GLM-ASCA incorporates the experimental design within a multivariate framework, providing a stronger statistical foundation for addressing the complexities of microbiome data analysis.

In linear models that employ ordinary least squares, orthogonal effect decomposition with ASCA/ASCA+ enables the separation and identification of distinct sources of variation in multivariate datasets, allowing for more precise and insightful analysis, particularly for data resulting from balanced experimental designs. In contrast, in GLMs, the iterative reweighted least squares (IRLS) algorithm is widely used to find maximum likelihood estimates rather than the ordinary least squares algorithm applied in linear models. As a result, parameter estimation in GLMs heavily relies on observation weights determined by the specific exponential family model under consideration. Including observation weights in GLMs complicates orthogonal effect decomposition, regardless of whether the treatment design is balanced. For this reason, it is crucial to appropriately extend ASCA within the framework of GLMs. In this context, we introduce methods for achieving orthogonal effect decomposition in GLMs for balanced design data to effectively utilize ASCA when integrated with GLMs. Thus, GLM-ASCA provides distinct advantages in well-structured experimental designs (e.g., full factorial designs, repeated measures) by decomposing variation attributable to main effects and interactions while accounting for the underlying multivariate structure. This integrated approach enables more transparent interpretation, improves the detection of biologically meaningful effects that may be missed by univariate methods, and enhances the identification of key features driving differential responses to experimental conditions.

To evaluate the performance of GLM-ASCA, we conducted a simulation study, which revealed that GLM-ASCA performs well, particularly in small-sample settings, motivating its application to microbiome experimental data. Subsequently, we applied GLM-ASCA to real microbiome data from tomato plants subjected to nitrogen starvation over time. Nitrogen is essential for plant fitness and productivity; however, non-legume crop plants mainly rely on chemical fertilizers. The application of these chemicals has severe environmental consequences, including water pollution from nutrient runoff, disruption of aquatic ecosystems through

eutrophication, and the release of greenhouse gases that contribute to climate change (Savci, 2012). Identifying beneficial microorganisms capable of fixing atmospheric nitrogen (N_2), reducing denitrification, and releasing inorganic nitrogen through mineralization—as well as the chemical communication that plants use to recruit them—can reduce the use of chemical fertilizers and pave the way toward sustainable agriculture and healthy ecosystems (Moreau et al., 2019; Mahmud et al., 2020; Abedini et al., 2021).

The paper is organized as follows: Section 2 introduces GLM-ASCA and demonstrates orthogonal decomposition in GLMs. Section 3 presents the performance of GLM-ASCA on simulated microbiome datasets and provides the analysis of plant microbiome data. Finally, Section 4 concludes the work with a discussion of the main results and future directions.

2 Methods and materials

2.1 GLM-ASCA

2.1.1 GLM-ASCA decomposition

Consider a model containing F main and interaction effects so that the design matrix $\mathbf{X}$ can be decomposed into $F+1$ blocks: $\mathbf{X} = (X_0,\mathbf{X}_1,\ldots,\mathbf{X}_F)$, where $\mathbf{X}_f$ is a matrix depending on the levels of factor f and X_0 is a column vector of ones to estimate the intercept. Let $\mathbf{Y} = (\mathbf{y}_1,\ldots,\mathbf{y}_p)$ represent the $n \times p$-dimensional response matrix. In order to accommodate various types of response data, including continuous, count, binary, and categorical variables, we extend the standard ASCA framework to be applied with Generalized Linear Models. That is, unlike ASCA, which uses ANOVA or linear regression, GLM-ASCA utilizes appropriate Generalized Linear Models to each column of the multivariate response matrix $\mathbf{Y}$. In particular, GLM-ASCA decomposes the working response matrix rather than the observed response matrix $\mathbf{Y}$ consistent with the standard extension of LMs to GLMs (Lovison, 2014).

In GLM-ASCA, univariate GLMs are first fitted to each column of the multivariate response matrix $\mathbf{Y} = (\mathbf{y}_1,\ldots,\mathbf{y}_p)$ using the same design matrix $\mathbf{X}$. Consider the j-th response variable $\mathbf{y}_j = (y_{j1},\ldots,y_{jn})$, representing a vector of n observations with mean μ_j. A GLM is specified for $\mathbf{y_j}$ with linear predictor $\eta_j = \mathbf{X}\beta_j$, link function $g(\mu_j) = \eta_j$, and variance function $V(\boldsymbol{\mu_j}) = \mathrm{diag}(V(\mu_{ij}))$. The maximum likelihood estimate of the regression coefficients (β_j) is given by (McCullagh, 2019):

$$\hat{\beta}_j = (\mathbf{X}^T\hat{\mathbf{W}}_j\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}_j\hat{\mathbf{z}}_j,$$

where $\hat{\mathbf{z}}_j = \hat{\eta}_j + \hat{\mathbf{r}}_j^w$ is the working response, with $\hat{\eta}_j = \mathbf{X}\hat{\beta}_j$, and $\hat{r}_j^w = \hat{\mathbf{D}}_j^{-1}(\mathbf{y}_j - \hat{\mu}_j)$ the working residuals. Here, $\hat{\mathbf{D}}_j$ is a diagonal matrix with entries $\left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)\Big|_{\hat{\eta}_{ij}}$, $i = 1,\ldots,n$, and $\hat{\mathbf{W}}_j$ is a diagonal matrix of weights with elements $\hat{w}_{ij} = \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\Big|_{\hat{\eta}_{ij}}\right)^2 / V(\hat{\mu}_{ij})$. Details on GLMs are given in the Supplementary Material Text S1 and Supplementary Table S1.

For all p responses in $\mathbf{Y}$, the estimated parameter vectors $\hat{\beta}_j$ are collected as columns of the matrix $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} = \begin{pmatrix} \hat{\beta}_{01} & \hat{\beta}_{02} & \cdots & \hat{\beta}_{0p} \\ \hat{\beta}_{11} & \hat{\beta}_{12} & \cdots & \hat{\beta}_{1p} \\ \vdots & \vdots & & \vdots \\ \hat{\beta}_{F1} & \hat{\beta}_{F2} & \cdots & \hat{\beta}_{Fp} \end{pmatrix}. \quad (1)$$

Equivalently, we may write (Equation 1) as

$$\hat{\mathbf{B}} = \begin{pmatrix} \hat{\mathbf{B}}_0 \\ \hat{\mathbf{B}}_1 \\ \vdots \\ \hat{\mathbf{B}}_F \end{pmatrix},$$

where the f-th row vector $\hat{\mathbf{B}}_f$ contains the estimated parameters for $\mathbf{X}_f$ in the design matrix $\mathbf{X} = (X_0,\mathbf{X}_1,\ldots,\mathbf{X}_F)$ across all p responses. The working responses can then be written in matrix form as

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{B}} + \hat{\mathbf{R}}^w, \quad (2)$$

where $\hat{\mathbf{Z}} = (\hat{\boldsymbol{z}}_1,\ldots,\hat{\boldsymbol{z}}_p)$ is the matrix of working responses, and $\hat{\mathbf{R}}^w = (\hat{\mathbf{r}}_1^w,\ldots,\hat{\mathbf{r}}_p^w)$ contains the corresponding working residuals as column vectors.

By decomposing the design matrix according to the main and interaction effects: $\mathbf{X} = (X_0,\mathbf{X}_1,\ldots,\mathbf{X}_F)$, the working response matrix in Equation 2 can be rewritten as

$$\begin{aligned} \hat{\mathbf{Z}} &= X_0\hat{\mathbf{B}}_0 + \mathbf{X}_1\hat{\mathbf{B}}_1 + \cdots + \mathbf{X}_F\hat{\mathbf{B}}_F + \hat{\mathbf{R}}^w \\ &= \mathrm{M}_0 + \mathrm{M}_1 + \cdots + \mathrm{M}_\mathbf{F} + \hat{\mathbf{R}}^w, \end{aligned} \quad (3)$$

where $\mathrm{M_f} = \mathbf{X}_f\hat{\mathbf{B}}_f, f = 0,1,\ldots,F$, are the effect matrices for the different main and interaction terms. To apply the standard ASCA framework, Section (2.2) demonstrates, under certain conditions, the orthogonal decomposition of the sum of squares of the working response matrix into the sum of squares of effect matrices and the sum of squares of working residuals matrix as follows using Equation 3:

$$\|\hat{\mathbf{Z}}\|^2 = \|\mathrm{M}_0\|^2 + \|\mathrm{M}_1\|^2 + \|\mathrm{M}_2\|^2 + \cdots + \|\mathrm{M}_\mathbf{F}\|^2 + \|\hat{\mathbf{R}}^w\|^2, \quad (4)$$

where $\|\cdot\|$ is the Frobenius norm of a matrix. In the GLM-ASCA decomposition similar to ASCA, Principal Component Analysis is subsequently applied to the various effect matrices. For the f-th effect matrix, we obtain

$$\mathrm{M_f} = \mathbf{T}_f\mathbf{P}_f^T, \quad (5)$$

where $\mathbf{T}_f$ are the scores and $\mathbf{P}_f$ are the loadings for the effect matrix $\mathrm{M_f}$. Note that when the first few principal components are considered sufficient, a residual term is needed to be added on the right-hand side of Equation 5.

Then, the GLM-ASCA decomposition of all effect matrices is given by

$$\hat{\mathbf{Z}} = \mathrm{M}_0 + \mathbf{T}_1\mathbf{P}_1^T + \cdots + \mathbf{T}_F\mathbf{P}_F^T + \hat{\mathbf{R}}^w.$$

As a result, each main and interaction effect is evaluated using score and loading plots. For example, a plot of the first principal component loadings of effect f, $\mathbf{P}_f$, shows which responses are most

affected by the model effect f. Similarly, the score plots of $\mathbf{T}_f$, show how the effect levels are located with respect to one another. In addition, projecting the augmented effect matrix helps to show the variability between the observations in the projected score plot (Zwanenburg et al., 2011). Given the augmented matrix,

$$\mathrm{M}_{\mathrm{f}}^{a} = \mathrm{M}_{\mathbf{f}} + \hat{\mathbf{R}}^{w},$$

which is analogous to partial residuals in the univariate GLM used for effect display (see Supplementary Table S1, Supplementary Material), the projected scores are then calculated as

$$\mathbf{T}_f^a = \mathrm{M}_f^a \mathbf{P}_f^T, \quad f = 1, 2, \ldots, F. \tag{6}$$

2.1.2 Percentages of variation

For the multivariate case, the ASCA literature offers a modified version of the classical ANOVA approach to calculate the percentage of variance explained by each effect. In ASCA, the square of the Frobenius norm is used to compute the sums of squares of the effect matrices (Vis et al., 2007). In the GLM-ASCA approach, for balanced experimental designs that provide an orthogonal decomposition, as mentioned above (Equation 4), the sum of squares of the working response matrix can be decomposed into the sum of squares of the effect matrices and the sum of squares of the working residual matrix as follows:

$$\| \hat{\mathbf{Z}} \|^2 = \| \mathrm{M}_0 \|^2 + \| \mathrm{M}_1 \|^2 + \| \mathrm{M}_2 \|^2 + \cdots + \| \mathrm{M}_{\mathbf{F}} \|^2 + \| \hat{\mathbf{R}}^{w} \|^2 .$$

These sums of squares expressed in Frobenius norms can be used to quantify the importance of effects.

That is, the importance of a given effect f is determined by the percentage of the total working response variance explained by the effect f:

$$\%\, Var_f = \frac{\| \mathrm{M}_{\mathbf{f}} \|^2}{\| \hat{\mathbf{Z}} \|^2 - \| \mathrm{M}_0 \|^2} \times 100, \quad f = 1, 2, \ldots, F,$$

where $\%Var_f$ denotes the percentage of variance explained by effect f or importance of effect f.

2.1.3 Permutation-based global effect tests

We also explored testing the statistical significance of main and interaction effects across the response variables to support the quantified effect importance. It is possible to determine whether main and interaction effects have a globally significant influence on all response variables by obtaining p-values using permutation testing (Zwanenburg et al., 2011). The advantage of permutation-based tests is that they provide a robust, non-parametric alternative to traditional parametric methods, avoiding reliance on assumptions such as normality, which are often violated in high-dimensional microbiome data. However, this advantage comes at the cost of increased computational burden, particularly when a large number of permutations are required for accurate inference across many features. In the permutation-based global test, we first generate N_p random permutations ($Y^{(k)}$, $k = 1,2,\ldots,N_p$) of the rows of the response data matrix Y. For each permuted dataset, a GLM is fitted, the effect matrices are retrieved, PCA is applied to the effect matrices, and the score and loading matrices are obtained for both the observed data Y and all permuted datasets $Y^{(k)}$. The first q principal components can be used for permutation testing, visualization, and further analysis. The number of principal components (q) retained plays a critical role in test sensitivity and is generally chosen based on criteria such as cumulative explained variance (e.g., ≥80%) or through visual inspection of scree plots. Formal statistical tests may also be used to determine q (Camargo, 2022, and references therein). Importantly, q does not need to be the same for all effect matrices, as each effect may explain a different proportion of total variance and require separate consideration.

A global statistic is defined based on the square Frobenius norm of the score matrix of the first q principal components, $\mathrm{T}_{\mathrm{f_q}}$. For testing effect f, the statistic for the observed data is then computed as

$$SS_f = \| \mathrm{T}_{\mathrm{f_q}} \|^2,$$

and for all permuted datasets under the null distribution as

$$SS_f^{(k)} = \| \mathrm{T}_{\mathrm{f_q}}^{(\mathbf{k})} \|^2, \quad k = 1, \ldots, N_p .$$

Finally, the p-value for effect of f is calculated as

$$\mathrm{p-value}_f = \frac{\#\left\{ SS_f^{(k)} \geq SS_f, \quad k = 1, \ldots, N_p \right\} + 1}{N_p + 1}, \quad f = 1, \cdots, F,$$

and the null hypothesis of no effect is rejected if the resulting p-value is less than a specified significance level (e.g., 0.05).

2.1.4 Feature selection in GLM-ASCA

When an effect is found to be significant using the permutation-based global effect test, the next step is to identify features that contribute to the significant effect. In classical PCA, a rule of thumb can be used to select features with high absolute loadings satisfying a specified threshold criterion. Alternatively, sparse PCA can be used to set the loadings of unimportant features to zero (Saccenti et al., 2018). In the ASCA context, for example, group-wise ASCA (GASCA) incorporates sparsity based on group-wise principal component analysis, where sparsity is defined in terms of groups of correlated variables identified in the correlation matrices computed from the effect matrices (Saccenti et al., 2018). In addition, several permutation-based significance tests have been implemented in the ASCA framework. For example, tests based on leverages and squared prediction errors are discussed in ASCA-genes (Tarazona et al., 2012; Nueda et al., 2007).

In this work, we used scaled leverages that measure the importance of features in a PCA model, computed as

$$\mathbf{h}_f = \mathrm{diag}\left(\mathbf{C}_f \mathbf{C}_f^T\right), \quad f = 1, 2, \ldots, F, \tag{7}$$

where $\mathbf{h}_f$ is a vector of scaled leverages corresponding to each feature in the PCA model for effect f, diag denotes the diagonal entries of a matrix, and $\mathbf{C}_f$ is a matrix of scaled loadings obtained by multiplying the loadings $\mathbf{P}_f$ of the first q principal components by the square roots of the variances explained by the respective principal components. When there are two or more principal components, this adjustment to the unscaled leverages ($\mathrm{diag}(\mathbf{P}_f P_f^T)$) (Tarazona et al.,

2012; Nueda et al., 2007), takes into account the variance contribution of each principal component to feature selection.

Based on the permutation procedure described above for global effect tests and the computed PCA loadings, we assess the importance of each feature for a given effect. To test the contribution of feature j to effect f, we compute the scaled leverage statistic from the observed data, given by the j-th element of $\mathbf{h}_f$ in Equation 7 and denoted by $\mathbf{h}_{fj}$.

Under the null distribution, for each permuted dataset $k = 1,\ldots, N_p$, we compute the permuted scaled leverage statistics:

$$\mathbf{h}_f^{(k)} = \text{diag}\left(\mathbf{C}_f^{(k)}\mathbf{C}_f^{(k)^T}\right), \quad f = 1, 2, \ldots, F,$$

where $\mathbf{C}_f^{(k)}$ is the matrix of scaled loadings obtained by multiplying the permutation-based loadings $\mathbf{P}_f^{(k)}$ (from the first q principal components) by the square roots of the variances they explain. The permuted scaled leverage statistic for feature j is the j-th element, $\mathbf{h}_{fj}^{(k)}$, of $\mathbf{h}_f^{(k)}$.

The p-value for testing the contribution of feature j to effect f is then calculated as:

$$\text{p} - \text{value}_{fj} = \frac{\#\left\{\mathbf{h}_{fj}^{(k)} \geq \mathbf{h}_{fj}\right\} + 1}{N_p + 1}, \quad f = 1, \ldots, F, \text{ and } j = 1, \ldots p\,.$$

Finally, to account for multiple testing of p features on effect f, we apply the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR) (Benjamini and Hochberg, 1995). Features with BH adjusted p-values below a predefined significance level (e.g., 0.05) are identified as significantly contributing to the corresponding effect.

2.2 Orthogonal decomposition in GLM

This section presents the technical details of the orthogonal decomposition in GLMs provided in (4). We demonstrate the orthogonal decomposition of the sum of squares of the working response variable by showing that the working residuals are orthogonal to the fitted values in balanced designs where the design matrix $\mathbf{X}$ is orthogonal with respect to effects. It has been shown that in Generalized Linear Models, the adjusted working residuals ($\mathbf{r}^{w\star} = \hat{\mathbf{W}}^{1/2}\hat{\mathbf{r}}^{w}$), like the linear model ordinary residuals, provide an exact orthogonal decomposition of the sum of squares of the adjusted working response ($\hat{\mathbf{z}}^{\star} = \hat{\mathbf{W}}^{1/2}\hat{\mathbf{z}}^{w}$) (Lovison, 2014):

$$\hat{\mathbf{z}}^{\star T}\hat{\mathbf{z}}^{\star} = \hat{\boldsymbol{\eta}}^{\star T}\hat{\boldsymbol{\eta}}^{\star} + \hat{\mathbf{r}}^{w\star T}\hat{\mathbf{r}}^{w\star}\,. \tag{8}$$

The main challenge in extending ASCA models directly to GLMs, similar to their use in LMs, is the difficulty in further orthogonally decomposing the linear predictor ($\hat{\boldsymbol{\eta}}^{*}$) in Equation 8 to specific effect sources of variation. That is, in GLM, unless the observation weights are one or constant, further orthogonal effect decomposition of $\hat{\eta}^{\star} = \hat{\mathbf{W}}^{1/2}\hat{\eta} = \hat{\mathbf{W}}^{1/2}\mathbf{X}\hat{\beta}$ according to orthogonal columns of a design matrix $\mathbf{X}$ is not straightforward (Hosmer et al., 2013). To address this problem, we considered two orthogonal decomposition issues: decomposing the sum of squares of the working response and decomposing the sum of squares of the linear predictor.

Decomposing the sum of squares of the working response: we establish an exact orthogonal decomposition using the unscaled or working response $\hat{\mathbf{z}}$ rather than the scaled or adjusted working response ($\hat{\mathbf{z}}^{\star}$). That is, the sum of squares of the working response can be decomposed into the sum of squares of the linear predictor and the sum of squares of the working residuals:

$$\hat{\mathbf{z}}^T\hat{\mathbf{z}} = \hat{\eta}^T\hat{\eta} + \hat{\mathbf{r}}^{wT}\hat{\mathbf{r}}^{w}\,. \tag{9}$$

This can be demonstrated as follows. Using definitions of the linear predictor we have

$$\begin{aligned}
\hat{\eta} &= \mathbf{X}\hat{\beta} \\
&= \mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}\hat{z} \\
&= \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{1/2}\hat{\mathbf{W}}^{1/2}\hat{z} \\
&= \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{1/2}\hat{\mathbf{z}}, \text{ where the Hat matrix } \hat{\mathbf{H}} \\
&= \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{1/2} \\
&= \mathbf{G}\hat{\mathbf{z}},
\end{aligned} \tag{10}$$

Where

$$\mathbf{G} = \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{1/2}\,. \tag{11}$$

Similarly, for the working residuals

$$\begin{aligned}
\hat{\mathbf{r}}^{w} &= \hat{\mathbf{z}} - \hat{\eta} \\
&= \hat{\mathbf{z}} - \mathbf{G}\hat{\mathbf{z}} \\
&= (\mathbf{I} - \mathbf{G})\hat{\mathbf{z}}\,.
\end{aligned} \tag{12}$$

In general, the sum of squares of the working response decomposes as

$$\hat{\mathbf{z}}^T\hat{\mathbf{z}} = \hat{\eta}^T\hat{\eta} + \hat{\mathbf{r}}^{wT}\hat{\mathbf{r}}^{w} + 2\hat{\eta}^T\hat{\mathbf{r}}^{w}\,.$$

So that, orthogonality of $\hat{\eta}$ and $\hat{\mathbf{r}}^{\mathbf{w}}$ can be achieved if $\hat{\eta}^T\hat{\mathbf{r}}^{w} = 0$. Using Equations 10, 12 we obtain

$$\begin{aligned}
\hat{\eta}^T\hat{\mathbf{r}}^{w} &= (\mathbf{G}\hat{\mathbf{z}})^T(\mathbf{I} - \mathbf{G})\hat{\mathbf{z}} \\
&= \hat{\mathbf{z}}^T(\mathbf{G}^T - \mathbf{G}^T\mathbf{G})\hat{\mathbf{z}}\,.
\end{aligned} \tag{13}$$

We note that the orthogonality property can be satisfied if the matrix $\mathbf{G}$ is idempotent and symmetric like the Hat matrix, $\hat{\mathbf{H}}$. However, from the properties of $\hat{\mathbf{H}}$, it is clear that $\hat{\mathbf{G}}$ is idempotent but not symmetric. That is

$$\mathbf{GG} = \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{1/2}\hat{\mathbf{W}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{1/2} = \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{1/2} = \mathbf{G}\,.$$

demonstrates that $\mathbf{G}$ is idempotent but

$$\mathbf{G}^T = (\hat{\mathbf{W}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{1/2})^T = \hat{\mathbf{W}}^{1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{-1/2},$$

differs from $\mathbf{G}$ implying that $\mathbf{G}$ is not symmetric under the given general setting of GLMs with observation weights $\hat{\mathbf{W}}$, posing a challenge to achieving orthogonality. However, here two approaches are introduced to ensure orthogonality in GLMs. The first approach utilizes the properties of the Hat matrix derived from

balanced treatment designs and saturated model formulations. The second approach involves choosing or deriving a link function in GLMs that provides observation weights equal to one or a constant value.

2.2.1 Balanced designs and saturated models

First, we consider the properties of the Hat matrix to establish an orthogonal decomposition of the sum of squares of the working response as the sum of squares of the working residuals and the fitted linear predictor values. As shown in Equation 11, given a special structure of the Hat matrix $\hat{\mathbf{H}}$ and weight matrix $\hat{\mathbf{W}}$ that allow these matrices to be commutative under multiplication, then $\mathbf{G} = \hat{\mathbf{H}}$ can be established.

With some algebraic simplifications using the full rank property of the design matrix arising from balanced designs and saturated models for one replication per experimental unit (see examples in the Supplementary Material Text S2), the Hat matrix turns out to be the identity matrix of dimension $n \times n$. That is, using the design matrix $\mathbf{X}^{(1)}$ for one replication per experimental unit:

$$\begin{aligned}\hat{\mathbf{H}} &= \hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)}(\mathbf{X}^{(1)T}\hat{\mathbf{W}}_n\mathbf{X}^{(1)})^{-1}\mathbf{X}^{(1)T}\hat{\mathbf{W}}_n^{1/2}\\ &= \hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)}(\mathbf{X}^{(1)T}\hat{\mathbf{W}}_n\mathbf{X}^{(1)})^{-1}\mathbf{X}^{(1)T}\hat{\mathbf{W}}_n^{1/2}\\ &\quad\times(\hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)})(\hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)})^{-1}\\ &= \hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)}(\mathbf{X}^{(1)T}\hat{\mathbf{W}}_n\mathbf{X}^{(1)})^{-1}(\mathbf{X}^{(1)T}\hat{\mathbf{W}}_n\mathbf{X}^{(1)})(\hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)})^{-1}\\ &= \hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)}\mathbf{I}_{n\times n}(\hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)})^{-1}\\ &= \mathbf{I}_{n\times n}\,.\end{aligned} \tag{14}$$

In this case, it follows that the matrix $\mathbf{G}$ also equals the identity matrix,

$$\mathbf{G} = \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{W}}^{1/2} = \hat{\mathbf{W}}^{-1/2}\mathbf{I}_{n\times n}\hat{\mathbf{W}}^{1/2} = \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{W}}^{1/2} = \mathbf{I}_{n\times n}\,.$$

In general, for R replications included in the experiment, the design matrix data structure can be expressed by vertically concatenating the $\mathbf{X}^{(1)}$ coding matrices as

$$\mathrm{X} = \begin{matrix}B_1\\B_2\\\vdots\\B_R\end{matrix}\begin{pmatrix}\mathrm{X}^{(1)}\\\mathrm{X}^{(1)}\\\vdots\\\mathrm{X}^{(1)}\end{pmatrix},$$

where $B_i$is used to indicate a block of n experimental units for the i-th replication. Similarly, the weights can be expressed as a block diagonal matrix as

$$\hat{\mathrm{W}} = \begin{matrix}B_1\\B_2\\\vdots\\B_R\end{matrix}\begin{pmatrix}\hat{\mathrm{W}}_n & 0 & 0 & \cdots & 0\\0 & \hat{\mathrm{W}}_n & 0 & \cdots & 0\\\vdots & & & & \vdots\\0 & 0 & 0 & \cdots & \hat{\mathrm{W}}_n\end{pmatrix}.$$

Letting

$$\mathbf{Q} = \hat{\mathbf{W}}^{1/2}\mathbf{X} = (\hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)}, \hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)}, \cdots, \hat{\mathbf{W}}_n^{1/2}\mathbf{X}^{(1)})^T,$$

the Hat matrix is rewritten as

$$\hat{\mathbf{H}} = \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T,$$

and using Equation 14 obtained for one replication per experimental unit, we have the final structure of the Hat matrix which is a matrix of identical identity block matrices multiplied by the fraction of replications $\frac{1}{R}$ (detail derivation is given in the Supplementary Information, Text S3)

$$\hat{\mathbf{H}} = \begin{pmatrix}\frac{1}{R}\mathbf{I}_{n\times n} & \cdots & \frac{1}{R}\mathbf{I}_{n\times n}\\\vdots & & \vdots\\\frac{1}{R}\mathbf{I}_{n\times n} & \cdots & \frac{1}{R}\mathbf{I}_{n\times n}\end{pmatrix}.$$

Similar Hat matrix structures are also described for ANOVA fixed effect models (Orenti et al., 2012). Examples of Hat matrices in balanced and saturated designs in GLMs are given in Text S4 (Supplementary Material).

We now finalize the orthogonality property in GLMs. For balanced designs and saturated models, we carefully exploit the block identity structure of the Hat matrix derived above and the diagonal form of the weight matrix to do commutative multiplication of the matrices in $\mathbf{G}$:

$$\begin{aligned}\mathbf{G} &= \hat{\mathbf{W}}^{-1/2}\hat{H}\hat{\mathbf{W}}^{1/2}\\ &= \hat{\mathbf{H}}\hat{\mathbf{W}}^{-1/2}\hat{\mathbf{W}}^{1/2}, \text{ since } \hat{\mathbf{H}} \text{ includes block identity or block diagonal matrices of ones and } \hat{\mathbf{W}} \text{ is a diagonal block matrix}\\ &= \hat{\mathbf{H}}\,.\end{aligned}$$

It follows that, for saturated models and balanced designs, $\mathbf{G}$ is idempotent and symmetric. Next, we simplify the orthogonal condition in Equation 13

$$\begin{aligned}\hat{\eta}^T\hat{\mathbf{r}}^w &= \hat{\mathbf{z}}^T\mathbf{G}^T(\mathbf{I}-\mathbf{G})\hat{\mathbf{z}}\\ &= \hat{\mathbf{z}}^T(\mathbf{G}-\mathbf{G})\hat{\mathbf{z}}, \quad G \text{ is symmetric and idempotent}\\ &= 0.\end{aligned}$$

Thus, $\hat{\eta}$ and $\mathbf{r}^w$ are orthogonal. As a result, we obtain an exact orthogonal decomposition of the sum of squares of the working response:

$$\hat{\mathbf{z}}^T\hat{\mathbf{z}} = \hat{\eta}^T\hat{\eta} + \hat{\mathbf{r}}^{wT}\hat{\mathbf{r}}^w\,.$$

Introducing new link functions with weights equal to one or a constant

The second approach to ensure orthogonality is to set the observation weights to 1 or a constant value (Dossou-Gbete and Tinsson, 2005). Under this constraint, $\mathbf{G} = \hat{\mathbf{H}}$ and orthogonality follows as described above, or the scaled version of the orthogonal decomposition in Equation 8 simplifies to the unscaled orthogonal decomposition in Equation 9. The constraint can be met by introducing a new link function with observation weights $w_i = w$, $i = 1,\ldots, n$ where $w = 1$ or a constant. Using the definition of observation weights which can be expressed as

$$w = \frac{1}{\left(\frac{\partial g(\mu)}{\partial \mu}\right)^2 V(\mu)}, \tag{15}$$

With weights equal to a constant, a new link function can be derived (Dossou-Gbété and Tinsson, 2005) by simplifying Equation 15 as

$$g(\mu) = \frac{1}{w^{1/2}} \int V(\mu)^{-1/2}\, d\mu\,.$$

In general, we rewrite the resulting orthogonal decomposition of the squared norm of $\hat{\mathbf{z}}$ as

$$\| \hat{\mathbf{z}} \|^2 = \| \hat{\eta} \|^2 + \| \hat{\mathbf{r}}^w \|^2\,. \tag{16}$$

Decomposing the sum of squares of the linear predictor ($\hat{\eta}$): we now address the second issue of orthogonality, which is an orthogonal effect decomposition of the linear predictor $\hat{\eta}$ into specific effect sources of variation. For a balanced design with a model containing all F main and interaction effects, and using sum coding of factor levels, the design matrix $\mathbf{X}$ can be decomposed into $F + 1$ orthogonal blocks that include the constant term and one for each model effect: $\mathbf{X} = (X_0 \mid \mathbf{X}_1 \mid \ldots \mid \mathbf{X}_F)$. The design matrix $\mathbf{X}$ being orthogonal leads to further orthogonal decomposition of $\hat{\eta} = X_0\hat{\beta}_0 + \mathbf{X}_1\hat{\beta}_1 + \cdots \mathbf{X}_F\hat{\beta}_F$ as

$$\| \hat{\eta} \|^2 = \| X_0\hat{\beta}_0 \|^2 + \| \mathbf{X}_1\hat{\beta}_1 \|^2 + \| \mathbf{X}_2\hat{\beta}_2 \|^2 + \cdots + \| \mathbf{X}_F\hat{\beta}_F \|^2\,. \tag{17}$$

When the link function corresponds to observation weights equal to one or a constant, a balanced design alone is a sufficient condition to ensure this orthogonal decomposition, consistent with the classical regression partitioning of the sum of squares (Montgomery et al., 2021; Radhakrishna Rao and Toutenburg, 1999). For completeness of presentation, details of the derivations under our framework are presented in Text S5 (Supplementary Material). Consequently, in the context of GLM, a full orthogonal decomposition of the sum of squares of the working response is obtained by substituting Equation 17 into Equation 16 as

$$\begin{aligned}\| \hat{\mathbf{z}} \|^2 = {} & \| X_0\hat{\beta}_0 \|^2 + \| \mathbf{X}_1\hat{\beta}_1 \|^2 + \| \mathbf{X}_2\hat{\beta}_2 \|^2 \\ & + \cdots + \| \mathbf{X}_F\hat{\beta}_F \|^2 + \| \hat{\mathbf{r}}^w \|^2\,.\end{aligned}$$

For example, for a two-factor model with main effects A and B and interaction effect AB the GLM decomposition is

$$\begin{aligned}\| \hat{\mathbf{z}} \|^2 = {} & \| X_0\hat{\beta}_0 \|^2 + \| X_A\hat{\beta}_A \|^2 + \| X_B\hat{\beta}_B \|^2 \\ & + \| X_{AB}\hat{\beta}_{AB} \|^2 + \| \hat{\mathbf{r}}^w \|^2\,.\end{aligned}$$

In this study, we work under the assumption of a balanced design and a saturated model, which enables orthogonal decomposition and facilitates interpretable effect estimation within the GLM-ASCA framework described in Section 2.1.1. Although this assumption is strong, it is frequently satisfied in well-controlled experimental settings, particularly in randomized factorial designs, where balance is intentionally maintained to ensure equal representation of treatment combinations, minimize confounding, and support orthogonal design structures that allow precise and independent estimation of factor effects. We also observed that incorporating an offset term in the GLM violates the orthogonal decomposition property. As part of our ongoing research, we are developing extensions to the GLM-ASCA framework to accommodate unbalanced designs and non-orthogonal decompositions, with the goal of increasing its applicability to more diverse and complex experimental settings.

2.3 GLM for microbiome data analysis

Analyzing microbiome datasets from high-throughput sequencing presents challenges due to overdispersion, zero inflation, non-normality, and compositionality. We addressed these issues using generalized linear models (GLMs) with a Tweedie family of distributions—one of the more flexible and general families for count and positive continuous data analysis. Compared with negative binomial–based models (Love et al., 2014b; Robinson et al., 2010), which are sensitive to zero inflation, and zero-inflated models (Zhang et al., 2016b), which require separate components for modeling excess zeros and abundance, the Tweedie GLM offers an integrated approach that inherently accommodates both zero and nonzero values within a single framework. This eliminates the need for an additional zero inflation term or arbitrary data imputations. Furthermore, the Tweedie model avoids the necessity of adding pseudocounts (Xia, 2023), which is typically required when applying log transformation in Gaussian-based models.

The Tweedie distribution is parameterized by the mean (μ), dispersion (ϕ), and power parameter (ρ), and it yields several well-known distributions for specific values of ρ, such as Gaussian (ρ = 0), Poisson (ρ = 1), gamma (ρ = 2), inverse Gaussian (ρ = 3), and compound Poisson–gamma ($1 < \rho < 2$). In particular, the Tweedie compound Poisson–gamma distribution has a point mass at zero and a skewed distribution on the positive real line, making it suitable for modeling count and positive continuous data with excess zeros, such as microbiome sequence count data (Mallick et al., 2022). This distribution has been applied for differential expression analysis of single-cell RNA sequencing (scRNA-seq) data (Mallick et al., 2022). The Tweedie compound Poisson–gamma distribution (Dunn and Smyth, 2005; Lian et al., 2023) is given by

$$f(y;\mu,\phi,\rho) = a(y;\phi,\rho)\exp\left\{\frac{1}{\phi}\left(\frac{y\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\right)\right\},$$

where the form of $a(y,\phi,\rho)$ is found in (Dunn and Smyth, 2005).

We considered GLM with a Tweedie distribution for each response, $Y \sim$ Tweedie(μ,ϕ,ρ), with logarithm link function $g(\mu) = \log(\mu) = \eta = \mathbf{x}^T\beta$, then for given ϕ and ρ, the mean and variance to be used in the GLM setting are given by Equation 18

$$E(Y) = \mu = g^{-1}(\eta)$$

$$Var(Y) = \phi\mu^{\rho}\,. \tag{18}$$

Furthermore, the compositional nature of microbiome data—resulting from differences in total sequence read counts (sequencing depth or library size) across samples due to the sequencing process—is addressed through normalization methods. Normalization is

an important step in microbiome sequencing data analysis used to remove bias caused by compositional effects or differences in sequencing depths or library sizes between samples. Several forms of normalization have been introduced (Xia, 2023) for microbiome data, including scaling-based normalization, zero-inflation–based normalization, and compositionally aware normalization. Some commonly used scaling-based normalization procedures, originally adopted from RNA sequencing (RNA-seq) data analysis, include the median-of-ratios method (Love et al., 2014b) and the trimmed mean of M-values (TMM) (Robinson et al., 2010). In addition, methods such as the geometric mean of pairwise ratios (Chen et al., 2018), Wrench normalization (Kumar et al., 2018), and the geometric mean of positive counts (poscounts) (McMurdie and Holmes, 2013) have been extended to account for zero inflation. Scaling-based normalization involves obtaining a scaling factor that adjusts raw counts to produce normalized counts or normalized library sizes. Normalized library sizes, for example, are used as offsets in generalized linear models (GLMs) to remove biases caused by differences in sequencing depths (Love et al., 2014b; Robinson et al., 2010; Mallick et al., 2022). On the other hand, compositionally aware normalization methods commonly used include centered log-ratio (CLR) transformation (Fernandes et al., 2014) and additive log-ratio (ALR) transformation (Mandal et al., 2015).

In this work, the raw microbiome count data were normalized using either the "poscounts" option in the *DESeq2* R package (Love et al., 2014a) or transformed using the modified centered log-ratio transformation (mCLR) from the *SPRING* R package (Yoon et al., 2019). By using normalized or transformed counts that account for biases caused by compositional effects or variations in library size, including an offset term is not necessary in the GLM. This approach maintains the orthogonal decomposition outlined above within the Tweedie GLM framework.

In general, the Tweedie GLM, which is appropriate for modeling positive data with many zeros, can effectively address the key characteristics of microbiome data. The Tweedie compound Poisson–gamma model ($1 < \rho < 2$), implemented in the R packages tweedieverse (Mallick et al., 2022) for the analysis of overdispersed and zero-inflated single-cell RNA-seq count data and mcglm (Bonat and Jørgensen, 2016), was used to estimate the model parameters.

2.4 Plant microbiome data: experimental setup and microbial DNA extraction

To assess the impact of nitrogen availability on the bacterial community composition of tomato roots, tomato seeds (Solanum lycopersicum cv. Moneymaker) were grown in an aeroponic system following the methodology outlined by Abedini et al. (manuscript in preparation[1]). Briefly, the seeds underwent surface sterilization and were pre-germinated at 25°C for 3 days. The pre-germinated seeds were then transplanted into small baskets filled with greenhouse soil. These baskets were placed in a large bucket equipped with an aeroponic system (Supplementary Figure S1). The aeroponic system utilized one-quarter–strength Hoagland solution, with a spraying duration of 15 s and a 10 min interval between sprays. The greenhouse environment was maintained at 22°C with 60% relative humidity and an 8 h dark/16 h light photoperiod. After 10 days of growth under standard control conditions, nitrogen starvation was initiated. The plants were randomly assigned to two groups: the control group, which received one-quarter–strength Hoagland solution containing 5.6 mM nitrogen, and the nitrogen-starved group, which received one-quarter–strength Hoagland solution without NH_4NO_3. Sampling occurred at 4, 8, 12, and 16 days after the start of nitrogen starvation, with five replicates for both the control and nitrogen-starved conditions. Because control and nitrogen-starved plants were grown and harvested at the same time for each sampling point, observed differences in the root microbiome between the two conditions can be attributed to nitrogen deprivation rather than to variations in growth stage.

Afterward, 45 mL of the collected Hoagland solution was vacuum filtered through a 0.2 μm membrane filter. The resulting filter, which retained the microbes, was placed into a PowerSoil kit bead tube and processed using a bead mill Tissuelyser for 5 min. Microbial DNA was then extracted following the PowerSoil protocol. The extracted genomic DNA was amplified for bacterial 16S rDNA targeting the V3 and V4 regions using primers 341F (5′-CCTACGGGNGGCWGCAG-3′) and 805R (5′-GACTACHVGGGTATCTAATCC-3′). Sequencing was performed on the Illumina NovaSeq6000 SP platform at Genome Quebec in Montreal, Quebec. Sample demultiplexing was carried out at the Genome Quebec facility. The resulting sequences underwent trimming, quality assessment, merging, and taxonomic classification using the dadasnake pipeline (Weißbecker et al., 2020). Taxonomic classification was performed with mothur using SILVA SSU v138 as the reference database. After preprocessing, the 16S dataset retained a total of 5300 amplicon sequence variants (ASVs) across five replicates for each of the four time points (4, 8, 12, and 16 days) under both control and nitrogen-starved conditions, resulting in a total of 40 samples.

2.5 Simulation study

We conducted a simulation study to evaluate the effectiveness of GLM-ASCA in identifying true positive taxa associated with main and interaction effects. To achieve this, we generated synthetic microbiome data based on experimental conditions consisting of four time points, two treatment conditions, and eight levels for the time-treatment interaction.

Our simulation approach used the R package SparseDOSSA2 (Ma et al., 2021), which operates independently of the distributional assumptions underlying the Tweedie GLM. SparseDOSSA2 is a statistical simulation framework that can be adapted to analyze plant microbiome datasets, effectively capturing plant microbial dynamics, as demonstrated previously in human microbiome studies (Ma et al., 2021). SparseDOSSA2 generates realistic

1 Abedini, D., White, F., Jain, R., Guerrieri, A., Schram, R., Dong, L., et al. (2025). Multi-omics data analysis revealed a novel beneficial role for strigolactones in tomato.

simulated data by parameterizing real-world template microbial datasets to reflect key microbiome characteristics such as zero inflation and overdispersion. For our study, we used the experimental plant microbiome data described in Section 2.4 as a template. First, SparseDOSSA2 was used to estimate taxon-specific parameters with a Bayesian hierarchical model, including the means and variances of nonzero log abundances, as well as the probabilities of zeros ($\tilde{\mu}_j, \sigma_j^2, \pi_j j = 1,\ldots,p$). These estimated parameters were then used to generate synthetic features from a zero-inflated, truncated log-normal distribution. Accordingly, we produced null (baseline) data consisting of 206 features and 20 samples under control (nitrogen-rich) conditions. Strict filtering criteria were applied to retain features with at least 10 counts in a minimum of five samples.

Next, to simulate taxa with differential effects, we incorporated the plant microbiome experimental conditions, which included the main effects growth condition (control and N-starvation) and time (4, 8, 12, and 16 days), as well as the interaction effect between growth condition and time. We then selected 45 of the 206 null taxa with a relative abundance of at least 20% to be spiked-in as having "true" differential main and interaction effects with known effect sizes (log-fold differences). Of the 45 taxa with "true" differential effects, 20 were assigned growth condition main effects (β_1), 15 were assigned time effects represented by three parameters (β_2, β_3, β_4), and 10 were assigned growth condition × time interaction effects represented by three parameters (β_5, β_6, β_7). The effect size parameter values were varied: half of the spiked features were assigned positive effect sizes ($\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$) = (2,3,1.5,1.5,2.8,2.4,1.4), and the other half were assigned negative effect sizes (−2,−3,−1.5,−1.5,−2.8,−2.4,−1.4). To account for these effects on the data generation by SparseDOSSA2, the taxon-specific mean log abundances ($\tilde{\mu}_j$) used to generate null features were modified to mean log abundances $\tilde{\mu}_{ij}$ across samples for the spiked-in features using

$$\underbrace{\tilde{\mu}_{ij}}_{\substack{\text{Mean log abundance}\\ \text{across samples}}} = \underbrace{\tilde{\mu}_j}_{\substack{\text{Taxon specific}\\ \text{mean log abundance}}} + \underbrace{\beta_1 x_{1i}}_{\text{Growth conditon effect}} + \underbrace{\beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}}_{\text{Time effect}} + \underbrace{\beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i}}_{\text{Interaction effect}}$$

where x represents values assigned through sum or deviation coding for the factors: growth condition, time, and their interaction. The spiked-in taxa are then generated based on zero-inflated truncated log-normal distribution with ($\tilde{\mu}_{ij}, \sigma_j^2, \pi_j, i = 1,\ldots,n; j = 1,\ldots,p$).

The sample sizes were varied at 40, 80, and 160, with 5, 10, and 20 replications, respectively. We generated 100 simulated abundance datasets, each containing 45 spiked-in taxa ("true positives") and 161 null taxa with no differential effect ("true negatives"). Of the 45 spiked taxa, 20, 15, and 10 were true positives for growth condition, time, and growth condition × time interaction effects, respectively. The efficacy of Tweedie GLM-ASCA in identifying taxa with differential effects in high-dimensional microbiome data was evaluated using these synthetic datasets with known ("true") effect sizes. We compared GLM-ASCA with two recently developed methods for microbiome data analysis, MaAsLin2 (Mallick et al., 2021) and LinDA (Zhou et al., 2022). Both accommodate multiple continuous and categorical covariates, unlike many differential abundance methods that consider only a single categorical covariate. MaAsLin2 utilizes generalized linear models to identify multivariable associations between microbial features and metadata, whereas LinDA applies linear models for differential abundance analysis, accounting for compositional bias and inflated zeros in microbiome data.

3 Results

3.1 Simulation results

To evaluate the performance of GLM-ASCA in microbiome data analysis, we conducted a simulation study generating 100 simulated microbiome datasets including 206 taxa with varying total sample sizes (40, 80, and 160), as described in Section 2.5. After fitting GLM-ASCA, BH-adjusted p-values were used to classify taxa as true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Performance was evaluated based on sensitivity (statistical power), specificity, FDR, area under the curve (AUC), F1-score (F-score), and Matthews correlation coefficient (MCC).

3.1.1 Performance of GLM-ASCA with different normalization methods

We first evaluated the performance of GLM-ASCA using two normalization methods: "poscounts" from *DESeq2* and mCLR from *SPRING*, incorporating permutation-based feature selection with scaled leverages. Performance was assessed based on FDR control and statistical power (sensitivity) to detect relationships between spiked taxa and the effects of growth condition, time, and their interaction. As shown in Figure 1, GLM-ASCA with DESeq2-based normalization and scaled-leverage feature selection exhibited superior statistical power compared with mCLR normalization, particularly in scenarios with small sample sizes. However, both normalization methods effectively controlled FDR at the nominal 5% level across all sample size settings. Detailed performance measures of GLM-ASCA for individual main and interaction effects are presented in Supplementary Figure S2.

3.1.2 Comparison of GLM-ASCA with alternative methods

To further assess GLM-ASCA, we compared its performance with MaAsLin2 and LinDA. In these comparisons, GLM-ASCA was implemented using "poscounts" normalization, while MaAsLin2 was applied with its default settings except for normalization, which was adjusted to CSS (cumulative sum scaling) to address zero inflation in microbiome data. LinDA was used with all default settings.

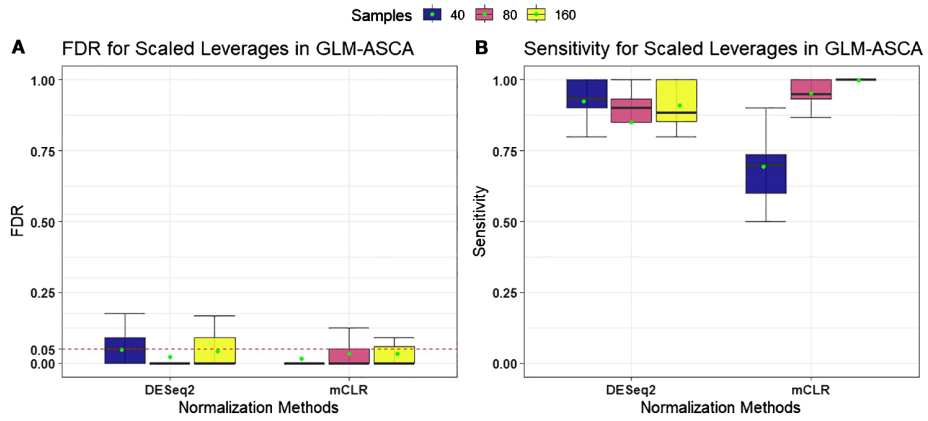

FIGURE 1
Performance measures [false discovery rate (FDR) and statistical power (sensitivity)] for GLM–ASCA using DESeq2 (poscounts) and mCLR normalizations, evaluated on simulated data derived from a template plant microbiome dataset. Boxplots are colored by total sample size. **(A)** With both normalizations, GLM-ASCA with scaled-leverage–based permutation feature selection demonstrated mean FDR (green dots) close to the nominal 5% level. **(B)** GLM-ASCA with DESeq2 (poscounts) achieved higher power in small-sample scenarios, whereas GLM-ASCA with mCLR achieved higher power in large-sample scenarios.

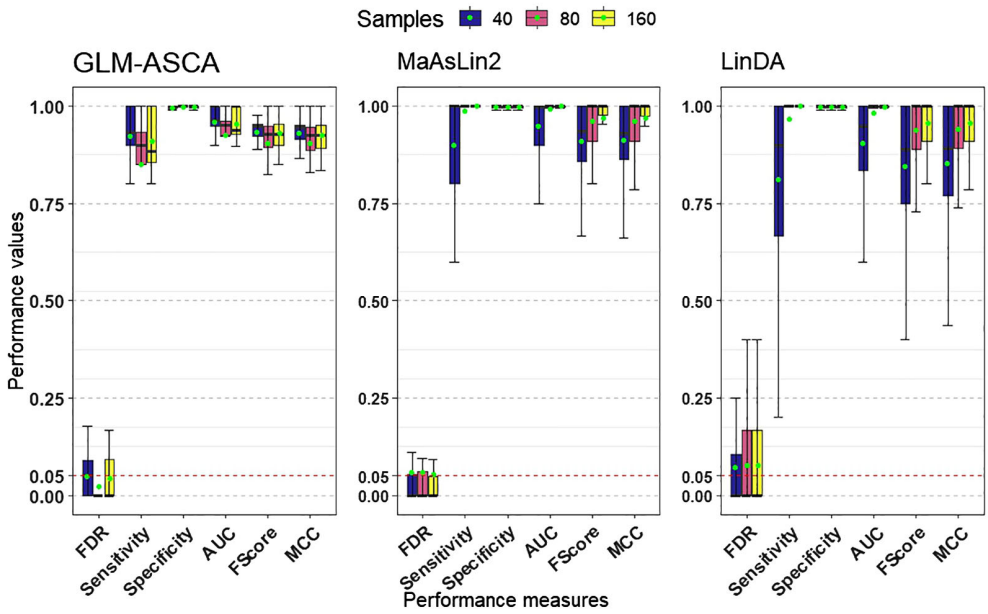

FIGURE 2
Performance measures (FDR, statistical power [sensitivity], specificity, AUC, F1-score, and MCC) for GLM–ASCA, MaAsLin2, and LinDA, evaluated on simulated data derived from a template plant microbiome dataset. Boxplots are colored by total sample size. GLM-ASCA, MaAsLin2, and LinDA demonstrated mean FDR (green dots) close to the nominal 5% level while maintaining high statistical power, AUC, F1-score, and MCC. GLM–ASCA achieved higher power in small-sample scenarios.

Figure 2 presents the simulation results across multiple performance measures. In our simulations, all three methods—GLM-ASCA, MaAsLin2, and LinDA—demonstrated high statistical power (Figure 2) when the sample size was large (e.g., $n = 80$ or $n = 160$), suggesting that each method can reliably detect true effects. However, in small-sample settings, GLM-ASCA exhibited greater power than MaAsLin2 and LinDA, indicating improved performance with limited sample sizes. This improved performance is likely due to its multivariate modeling framework, which captures shared patterns across features and leverages the joint data structure to detect effects even with few samples. Despite this difference in power, all methods performed comparably on measures such as specificity, AUC, F1-score, and MCC. Moreover, all three methods effectively maintained FDR control at the nominal 5% level.

The low feature–feature correlations observed in the template plant microbiome data (Supplementary Figure S15), particularly for

large sample sizes, appear to favor MaAsLin2 and LinDA, since these univariate methods perform optimally when features are weakly dependent and sufficient data support stable parameter estimation. Taken together, these findings suggest that while MaAsLin2 and LinDA are robust choices for univariate analysis in large-sample contexts, GLM-ASCA offers a notable advantage in small-sample, structured experimental designs by leveraging multivariate information.

We extended the simulation study using SparseDOSSA2 by increasing the number of features from 206 to 1000 with varying sparsity (proportion of zeros), while retaining the small-sample scenario (n = 40) and the same data generation procedure. This reflects common challenges in microbiome research, where datasets often involve limited samples and many sparse features. As shown in Figure 3, GLM-ASCA maintained robust performance under these conditions, effectively controlling FDR at the nominal 5% level while achieving moderately high statistical power. These findings highlight the effectiveness of GLM-ASCA in detecting true features and controlling false discoveries despite high dimensionality and small sample sizes. In terms of AUC, F1-score, and MCC, all three methods showed comparable performance in this small-sample scenario.

3.2 Microbiome data analysis results

The Tweedie-based GLM-ASCA was applied to tomato root microbiome data to identify microbes whose abundance was significantly affected by nitrogen starvation over time. The design matrix included growth condition with two levels (N-starvation: nitrogen starved; control: nitrogen rich), time with four levels (4, 8, 12, and 16 days), and their interaction (growth condition × days). The following generalized linear model was used to estimate the effect matrices:

$$log(\mu) = \beta_0 + \beta_G \cdot \text{GrowthCondition} + \beta_D \cdot \text{Days} + \beta_{GD} \cdot \text{GrowthCondition} \times \text{Days}, \quad (19)$$

with a logarithm link function relating mean microbial abundance to the experimental factors. Normalized counts were computed using *poscounts* normalization from the R package DESeq2. The design matrix was coded using sum coding. Out of 5300 ASVs, 1009 ASVs were retained after filtering for a minimum of 5 counts in at least 3 of the 5 replicates in each growth condition–time combination. The design was balanced, comprising 40 samples with 5 replicates for each condition and time point, and the GLM model (Equation 19) was saturated, including all main and interaction effects. Thus, two basic requirements for GLM-ASCA were satisfied: balanced design and saturated model specification.

For each filtered feature, univariate GLMs were fitted using the Tweedie distribution with the R packages tweedieverse and mcglm, which allow estimation of the dispersion (ϕ), power (ρ), and regression parameters (β). Estimates of the Tweedie power and dispersion parameters are shown in Supplementary Figure S3. The regression parameter estimates ($\hat{\beta}$) and design matrix were then used to calculate the effect matrices for the main effects of growth condition and time, as well as their interaction. PCA was applied to each effect matrix to obtain the score and loading matrices. Table 1 displays the percentages of explained variation due to main and interaction effects, calculated based on the adjusted response. The experimental conditions accounted for ~88% of the total variation. Table 1 also includes p-values from global tests of effects, computed using the Frobenius norm of principal component score matrices (see Section 2.1.3), which revealed significant main and interaction effects ($p < 0.05$).

One advantage of ASCA-based approaches is the ability to visualize effects using score and loading matrices. Results are shown

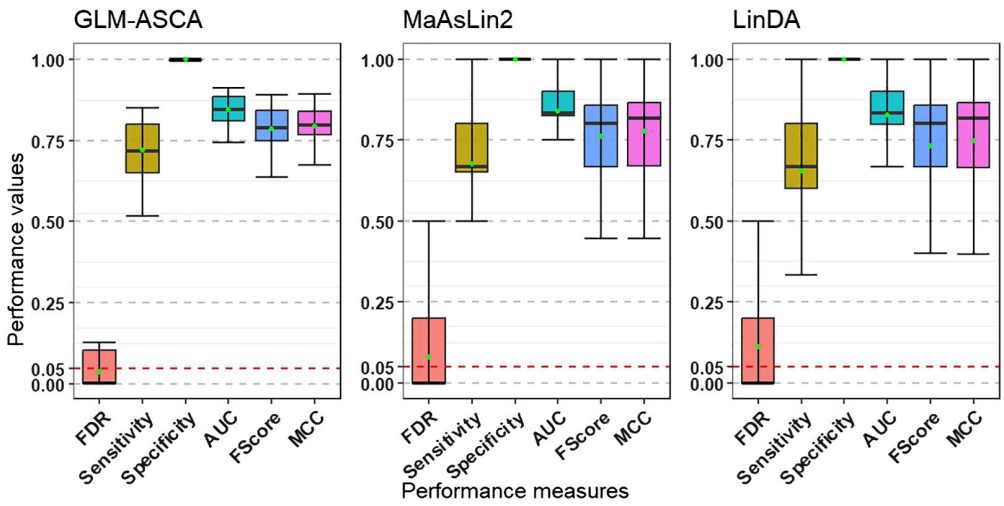

FIGURE 3
Performance measures (FDR, statistical power [sensitivity], specificity, AUC, F1-score, and MCC) for GLM–ASCA, MaAsLin2, and LinDA, evaluated on simulated data derived from a template plant microbiome dataset with 40 samples, 1000 features, and varying sparsity. Boxplots are colored by performance measures. Compared with MaAsLin2 and LinDA, GLM-ASCA demonstrated mean FDR (green dots) close to the nominal 5% level while maintaining moderately high statistical power.

TABLE 1 Percentage of explained variation in the adjusted abundance response of the tomato microbiome data, accounted for by experimental factors, using the Tweedie GLM-ASCA model.

Component	Explained variation (%)	Permutation p-values
Growth Condition	21.99	0.0001
Days	34.89	0.0001
Growth Condition x Days	30.87	0.0005
Residuals	12.25	
Total	100.00	

in Supplementary Figures S8-S14; Figures 4, 5. In these figures, points represent principal component scores computed for the two growth conditions at each time point. Lines connect scores across successive time points to illustrate temporal dynamics in microbial relative abundance. Error bars correspond to mean ± 1 standard deviation of the projected scores (Equation 6).

Because of the significant interaction effect between time and growth condition, the main effects (Supplementary Figures S8-S14) should not be interpreted separately. Thus, we combined the main effect of growth condition with the interaction effect of time (GrowthCondition + GrowthCondition × Days). This, in particular, allows for a more direct assessment of how the microbial abundance in the control and N-starvation groups changes over time during tomato growth. We applied PCA to the combined effect matrix (GrowthCondition + GrowthCondition × Days). Using the scaled leverage permutation test with 10,000 permutations, 121 ASVs that belong to 44 families were identified (Figure 4B, Supplementary Figure S4) to be significantly affected by nitrogen starvation over time (adjusted p-value < 0.05). The first two principal components from the combined effect matrix (Figures 4, 5) accounted for 75.3% of the total variation. On the first principal component, no significant differences were observed between the average principal scores of the N-starvation and control groups at day 4. However, beyond day 4, thedifference between groups increased over time (Figure 4A), indicating increasing divergence in microbial abundance between nitrogen-starved and nitrogen-rich conditions.

From loading plots (Figures 4B, Supplementary Figures S4, S5), families with positive loadings in the control group showed increasing microbial abundance over time, whereas in the N-starvation group, families with negative loadings showed increasing abundance. Families/genera enriched under nitrogen starvation included Acidobacteriaceae (Paludibaculum), Bdellovibrionaceae (Bedellovibrio), Burkholderiaceae (Polynucleobator), Caulobactraceae (Asticcacaulis), Chitinophagaceae (Terrimonas and Edaphobaculum), Comamonadaceae (Acidovorax, Aquabacterium and Methylibium), Gallionellacea (Candidatus Nitrotoga), Hydrogenophilaceae (Thiobacillus), Mycobacteriaceae (Mycobacterium), Nocardioidaceae (Aeromicrobium and Nocardioides), Opitutaceae (Opitutus), Pleomorphomonadaceae (Pleomorphomonas), Pseudomonadaceae (Pseudomonas), Reyranellaceae (Reyranella), Rhizobiaceae (AllorhizobiumNeorhizobiumPararhizobiumRhizobium, and Mesorhizobium), Solimonadaceae (Solimonas), and Sphingomonadaceae (Novosphingobium, Sphingobium and Sphingomonas) include one or more species that showed significantly increased abundance under nitrogen starvation, whereas families (genera) such as: Acetobacteraceae (Acidisoma), Alcaligenaceae (Bordetella), Burkholderiaceae (Burkholderia-CaballeroniaParaburkholderia, Robbsia

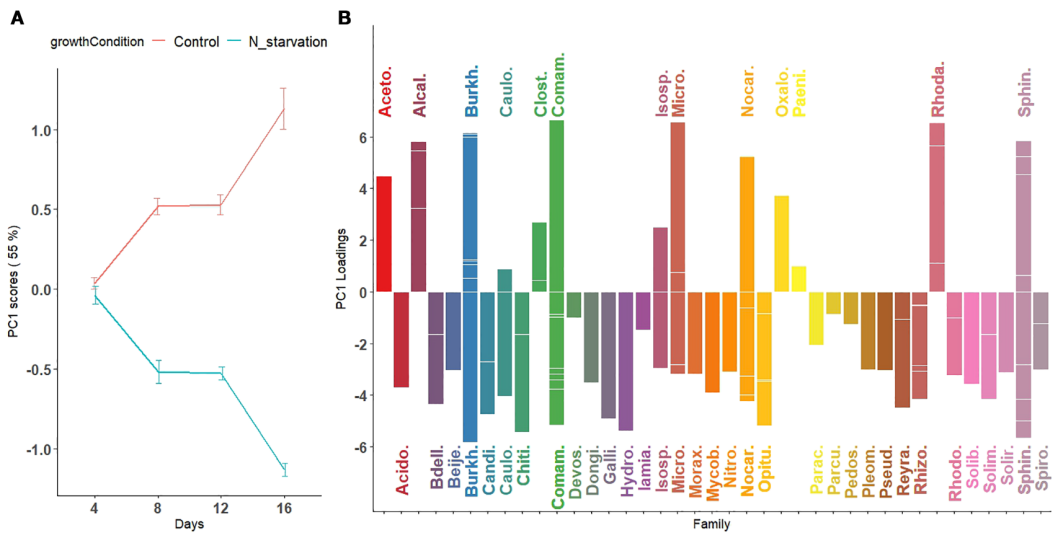

FIGURE 4
First principal component visualizing temporal variation in microbial abundance patterns across taxa under nitrogen starvation and control conditions. **(A)** Scores of the first principal component are plotted for nitrogen-starved (green line) and nitrogen-rich (control; red line) conditions over time (days). The trajectories represent overall trends in microbial abundance under each condition. **(B)** Loadings of the first principal component, representing the contribution of individual microbial species, are displayed as bars colored by their bacterial families. Downward-facing bars correspond to families with higher relative abundance under nitrogen-starved conditions (green trajectory in panel A), whereas upward-facing bars correspond to families with higher relative abundance under nitrogen-rich (control) conditions (red trajectory in panel A). For microbial families with multiple contributing species, individual species loadings are indicated by horizontal segments within the same vertical bar.

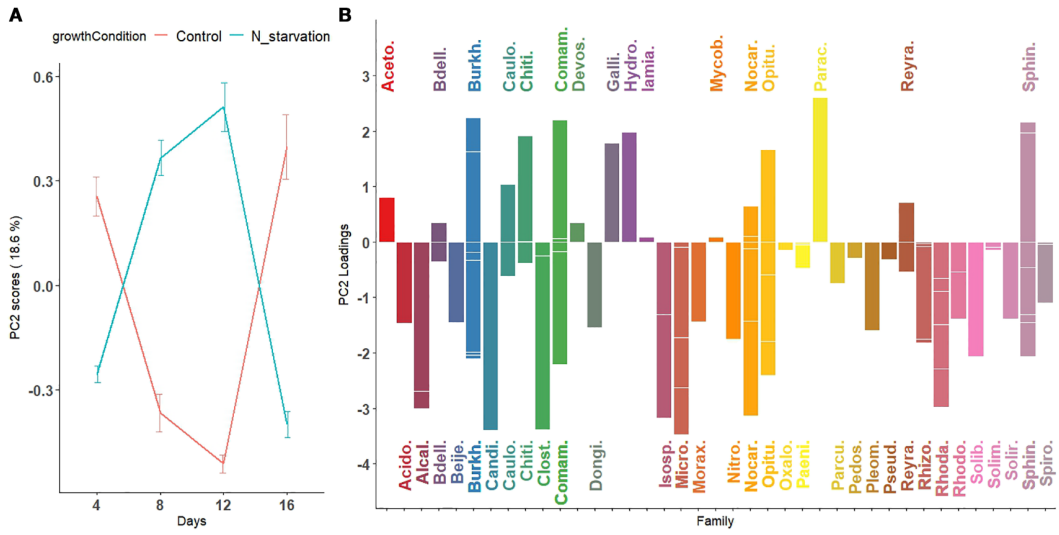

FIGURE 5
Second principal component visualizing nonlinear temporal patterns in microbial abundance across taxa under nitrogen starvation and control conditions. **(A)** Scores of the second principal component reveal a nonlinear pattern characterized by either a peak (for taxa with positive loadings) or a dip (for taxa with negative loadings) around day 12 in both nitrogen-starved and nitrogen-rich (control) conditions. **(B)** Loadings of the second principal component, representing the contribution of individual microbial species to this nonlinear pattern, are displayed as bars colored by their bacterial families. These trajectories capture temporal fluctuations in microbial abundance not explained by the first principal component, highlighting complex, taxa-specific responses to nitrogen availability throughout the experimental period.

and Pandoraea), Clostridiaceae (Clostridium), Comamonadaceae (Thiomonas), Microbacteriaceae (Leifsonia), Oxalobacteraceae, Rhodanobacteraceae (Rhodanobacter), and Sphingobacteriaceae (Mucilaginibacter) include one or more species that showed a significant increase in abundance under the control or nitrogen availability condition. When multiple species contributed significantly, loadings are represented by horizontal bars within each family.

In the second principal component (Figure 5), average scores followed a nonlinear (parabolic) pattern of microbial abundance over time. Under N-starvation, average abundance of Gallionellaceae, Hydrogenophilaceae, and Parachlamydiaceae increased until day 12, then sharply declined at day 16. Similarly, under control conditions, Alcaligenaceae, Candidatus Kaiserbacteria, Clostridiaceae, Microbacteriaceae, and Rhodanobacteraceae exhibited such curved abundance profiles. The third and fourth principal components, explaining ~26% of the remaining variation, are shown in Supplementary Figures S6-S7.

Enrichment of bacterial genera under nitrogen starvation highlights their potential roles in adapting to and mitigating nitrogen limitation. Several taxa identified here have previously been reported in nitrogen-related processes. For instance, species within Terrimonas (Guo et al., 2024), Thiobacillus (Li et al., 2023), Mycobacterium (Sellstedt and Richau, 2013), Pseudomonas (Wu et al., 2023; Sanow et al., 2023), Sphingomonas (Videira et al., 2009), Novosphingobium (Addison et al., 2007), Mesorhizobium (Menéndez et al., 2020), and Allorhizobium–Neorhizobium–Pararhizobium–Rhizobium (You et al., 2021) are known nitrogen fixers, with *nifH* genes detected in many studies. Similarly, Candidatus Nitrotoga (Lucker et al., 2015; Boddicker and Mosier, 2018), Aquabacterium (Zhang et al., 2016a), and Sphingobium (Boss et al., 2022) are implicated in nitrogen cycling. These taxa may enhance nitrogen availability to the plant either by directly fixing atmospheric nitrogen into plant-available forms such as ammonium, by contributing to nitrogen mineralization processes that convert organic nitrogen compounds into inorganic forms like ammonium and nitrate (Philippot et al., 2013), or by adapting plant development, such as root architecture (Abedini et al., manuscript in preparation[1]).

We also observed an increase in Bdellovibrio (Bratanis et al., 2020), a potential biocontrol agent. Increased Bdellovibrio abundance may reflect a regulatory mechanism suppressing pathogenic or competing bacteria, indirectly supporting beneficial taxa and plant health.

To validate our findings, for example, Sphingobium was identified as enriched under nitrogen deficiency. Isolation of a Sphingobium strain from nitrogen-starved tomato roots, followed by *in vitro* assays, confirmed that it can stimulate tomato growth under nitrogen-deficient conditions (Abedini et al., manuscript in preparation[1]). These findings suggest that nitrogen deficiency alters microbial community structure, and that recruited taxa support plant adaptation by enhancing nitrogen availability. Overall, the results highlight the effectiveness of GLM-ASCA in identifying key microbial taxa under specific experimental conditions, underscoring its potential as a powerful tool for microbiome data analysis.

4 Discussion

Statistical analysis of high-dimensional, non-normal, and non-linear data—such as those obtained from microbiome studies—and incorporating experimental design elements such as treatments, time, and interactions present challenges because traditional

statistical methods often assume normality and may not be appropriate for such datasets. Advanced statistical tools, such as ANOVA simultaneous component analysis (ASCA/ASCA+), have emerged as valuable approaches, providing insights into the main sources of variability and facilitating interpretation. However, adapting ASCA to count data with excess zeros, such as microbiome data, necessitates novel approaches. This led us to develop GLM-ASCA (generalized linear models–ANOVA simultaneous component analysis), which integrates treatment design elements and GLMs within a multivariate framework.

The simulation results demonstrated effective control over false discovery rates, highlighting the potential of GLM-ASCA as a robust feature selection tool. Application of GLM-ASCA to microbiome data to assess the effect of nitrogen starvation on tomato over time identified several bacterial families and genera that exhibited increased abundance under nitrogen deficiency, many of which have been implicated in nitrogen metabolism in previous studies. The observed changes in microbial abundance during nitrogen starvation suggest that plants modulate root exudation patterns to selectively recruit beneficial microbial taxa. These microbes contribute to nitrogen availability and support plant growth through multiple complementary mechanisms, including nitrogen cycling and mineralization, symbiotic and free-living nitrogen fixation, root colonization coupled with plant growth promotion, stress adaptation and stabilization of the rhizosphere under nutrient-limited conditions, and microbial community regulation and niche structuring.

For instance, the increased abundance of genera such as Sphingobium, Pseudomonas, and Mesorhizobium suggests potential mechanisms where these microbes enhance nitrogen availability either through biological nitrogen fixation or mineralization pathways. To further support and validate these results, whole-genome sequencing and co-culturing assays were conducted with Sphingobium sp. RS1, a strain isolated from nitrogen-starved tomato roots. These analyses revealed several plant growth–promoting traits, including the production of phytohormones such as indole-3-acetic acid (IAA) and the ability to mineralize organic nitrogen into plant-available forms (Abedini et al., manuscript in preparation[1]). Although this targeted validation experiment supported the role of specific taxa such as Sphingobium, many other microbial taxa identified in the present study as potentially involved in mitigating nitrogen deficiency require further validation. Rigorous experimental confirmation is necessary to determine their functional roles and assess their effectiveness in tomato, a nonlegume crop, under both controlled and field conditions. Once thoroughly validated, these results could enable the development of targeted inoculants or synthetic microbial consortia designed to improve plant growth and health in nitrogen-limited environments. Ultimately, such bio-based strategies have the potential to support sustainable agriculture by reducing dependence on chemical fertilizers and promoting more efficient nutrient use in crop production systems.

The results from both simulated and real data underscore the utility of the GLM-ASCA framework as an effective tool for identifying key microbial species responding to specific experimental conditions, treatments, or interactions. However, a crucial aspect of the current development of GLM-ASCA is its reliance on data from balanced experimental designs with model specifications that include all main and interaction effects. We are currently working on expanding the framework by incorporating different link functions and extending it to more general scenarios, including balanced designs without specification constraints and unbalanced designs.

Finally, with the growing importance of plant microbiome research in sustainable agriculture and human health, developing such statistical tools is crucial for identifying biologically important microbes that play key roles in enhancing agricultural practices and improving health outcomes. Moreover, the ability of GLM-ASCA to effectively handle complex experimental designs and accurately analyze microbial abundance patterns highlights its potential for broader applications beyond plant microbiome research. GLM-ASCA can be applied in various fields that require the analysis of high-dimensional, compositional, and zero-inflated data with complex experimental designs, including human microbiome studies, other omics applications involving high-throughput sequencing, and ecological studies.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author. All data sets generated and analyzed and the R-code used to analyze the data are available in Figshare public repository https://figshare.com/s/744f99a21afca4d6c002.

Author contributions

FA: Conceptualization, Data curation, Methodology, Validation, Writing – original draft, Writing – review & editing. DA: Data curation, Investigation, Validation, Writing – review & editing. LD: Data curation, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing – review & editing. JW: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. FE: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. HB: Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Resources, Validation, Writing – review & editing. AS: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. We acknowledge funding by the Dutch Research Council (NWO/OCW) for the MiCRop

Consortium program, Harnessing the second genome of plants (Grant number 024.004.014; to HB, LD, DA, AKS, JAW, FVE and FA), and the Dutch Research Council (NWO-TTW grant 16873 Holland Innovative Potato; to HB, LD and DA).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frmbi.2025.1584516/full#supplementary-material

References

Abedini, D., Jaupitre, S., Bouwmeester, H., and Dong, L. (2021). Metabolic interactions in beneficial microbe recruitment by plants. *Curr. Opin. Biotechnol.* 70, 241–247. doi: 10.1016/j.copbio.2021.06.015

Addison, S. L., Foote, S. M., Reid, N. M., and Lloyd-Jones, G. (2007). Novosphingobium nitrogenifigen ssp. nov., a polyhydroxyalkanoate-accumulating diazotroph isolated from a New Zealand pulp and paper wastewater. *Int. J. System Evolution Microbiol.* 57, 2467–2471. doi: 10.1099/ijs.0.64627-0

Ali, N., Jansen, J., van den Doel, A., Tinnevelt, G. H., and Bocklitz, T. (2020). WE-ASCA: the weighted-effect ASCA for analyzing unbalanced multifactorial designs—a Raman spectra-based example. *Molecules* 26, 66. doi: 10.3390/molecules26010066

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc Ser. B (Statistical Method.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bertinetto, C., Engel, J., and Jansen, J. (2020). ANOVA simultaneous component analysis: A tutorial review. *Analyt Chim Acta: X* 6, 100061. doi: 10.1016/j.acax.2020.100061

Boddicker, A. M., and Mosier, A. C. (2018). Genomic profiling of four cultivated Candidatus Nitrotoga spp. predicts broad metabolic potential and environmental distribution. *ISME J.* 12, 2864–2882. doi: 10.1038/s41396-018-0240-8

Bonat, W. H., and Jørgensen, B. (2016). Multivariate covariance generalized linear models. *J. R. Stat. Soc Ser. C (Applied Statistics)* 65, 649–675. doi: 10.1111/rssc.12145

Boss, B. L., Wanees, A. E., Zaslow, S. J., Normile, T. G., and Izquierdo, J. A. (2022). Comparative genomics of the plant-growth promoting bacterium Sphingobium sp. strain AEW4 isolated from the rhizosphere of the beachgrass Ammophila breviligulata. *BMC Genomics* 23, 508. doi: 10.1186/s12864-022-08738-8

Bratanis, E., Andersson, T., Lood, R., and Bukowska-Faniband, E. (2020). Biotechnological potential of Bdellovibrio and like organisms and their secreted enzymes. *Front. Microbiol.* 11, 662. doi: 10.3389/fmicb.2020.00662

Camacho, J., Vitale, R., Morales-Jimenez, D., and Gomez-Llorente, C. (2023). Variable-selection ANOVA simultaneous component analysis (VASCA). *Bioinformatics* 39, btac795. doi: 10.1093/bioinformatics/btac795

Camargo, A. (2022). Pcatest: testing the statistical significance of principal component analysis in r. *PeerJ* 10, e12967. doi: 10.7717/peerj.12967

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6, e4600. doi: 10.7717/peerj.4600

Dossou-Gbete, S., and Tinsson, W. (2005). Factorial experimental designs and generalized linear models. *SORT: Stat Operations Res. Trans.* 29, 0249–0268.

Dunn, P. K., and Smyth, G. K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Stat Comput* 15, 267–280. doi: 10.1007/s11222-005-4070-y

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2, 1–13. doi: 10.1186/2049-2618-2-15

Guo, L., Liu, S., Zhang, P., Hakeem, A., Song, H., Yu, M., et al. (2024). Effects of different mulching practices on soil environment and fruit quality in Peach Orchards. *Plants* 13, 827. doi: 10.3390/plants13060827

Hosmer, J. D.W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression* Vol. 398 (Hoboken, New Jersey: John Wiley & Sons).

Jansen, J. J., Hoefsloot, H. C., van der Greef, J., Timmerman, M. E., Westerhuis, J. A., and Smilde, A. K. (2005). ASCA: analysis of multivariate data obtained from an experimental design. *J. Chemomet* 19, 469–481. doi: 10.1002/cem.952

Jarmund, A. H., Madssen, T. S., and Giskeødegard, G. F. (2022). ALASCA: An R package for longitudinal and cross-sectional analysis of multivariate data by ASCA-based methods. *Front. Mol. Biosci.* 9, 962431. doi: 10.3389/fmolb.2022.962431

Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S., and Corrada Bravo, H. (2018). Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* 19, 1–23. doi: 10.1186/s12864-018-5160-5

Li, Y., Guo, L., Yang, R., Yang, Z., Zhang, H., Li, Q., et al. (2023). Thiobacillus spp. and Anaeromyxobacter spp. mediate arsenite oxidation-dependent biological nitrogen fixation in two contrasting types of arsenic-contaminated soils. *J. Hazard Mater* 443, 130220.

Lian, Y., Yang, A. Y., Wang, B., Shi, P., and Platt, R. W. (2023). A tweedie compound poisson model in reproducing kernel hilbert space. *Technometrics* 65, 281–295. doi: 10.1080/00401706.2022.2156615

Love, M., Anders, S., and Huber, W. (2014a). Differential analysis of count data–the DESeq2 package. *Genome Biol.* 15, 10–1186. doi: 10.1186/s13059-014-0550-8

Love, M. I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21. doi: 10.1186/s13059-014-0550-8

Lovison, G. (2014). A note on adjusted responses, fitted values and residuals in Generalized Linear Models. *Stat. Model.* 14, 337–359. doi: 10.1177/1471082X13508263

Lucker, S., Schwarz, J., Gruber-Dorninger, C., Spieck, E., Wagner, M., and Daims, H. (2015). Nitrotoga-like bacteria are previously unrecognized key nitrite oxidizers in full-scale wastewater treatment plants. *ISME J.* 9, 708–720.

Ma, S., Ren, B., Mallick, H., Moon, Y. S., Schwager, E., Maharjan, S., et al. (2021). A statistical model for describing and simulating microbial community profiles. *PloS Comput. Biol.* 17, e1008913. doi: 10.1371/journal.pcbi.1008913

Madssen, T. S., Giskeødegard, G. F., Smilde, A. K., and Westerhuis, J. A. (2021). Repeated measures ASCA+ for analysis of longitudinal intervention studies with multivariate outcome data. *PloS Comput. Biol.* 17, e1009585. doi: 10.1371/journal.pcbi.1009585

Mahmud, K., Makaju, S., Ibrahim, R., and Missaoui, A. (2020). Current progress in nitrogen fixing plants and microbiome research. *Plants* 9, 97. doi: 10.3390/plants9010097

Mallick, H., Chatterjee, S., Chowdhury, S., Chatterjee, S., Rahnavard, A., and Hicks, S. C. (2022). Differential expression of single-cell RNA-seq data using Tweedie models. *Stat Med.* 41, 3492–3510. doi: 10.1002/sim.9430

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. *PloS Comput. Biol.* 17, e1009442. doi: 10.1371/journal.pcbi.1009442

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis.* 26, 27663. doi: 10.3402/mehd.v26.27663

Martin, M., and Govaerts, B. (2020). LiMM-PCA: Combining ASCA+ and linear mixed models to analyse high-dimensional designed data. *J. Chemomet* 34, e3232. doi: 10.1002/cem.3232

McCullagh, P. (2019). *Generalized linear models* (New York: Routledge).

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One* 8, e61217. doi: 10.1371/journal.pone.0061217

Menéndez, E., Perez-Yepez, J., Hernandez, M., Rodriguez-Perez, A., Velazquez, E., and Leon-Barrios, M. (2020). Plant growth promotion abilities of phylogenetically diverse mesorhizobium strains: effect in the root colonization and development of tomato seedlings. *Microorganisms* 8, 412. doi: 10.3390/microorganisms8030412

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis* (Hoboken, New Jersey: John Wiley & Sons).

Moreau, D., Bardgett, R. D., Finlay, R. D., Jones, D. L., and Philippot, L. (2019). A plant perspective on nitrogen cycling in the rhizosphere. *Funct. Ecol.* 33, 540–552. doi: 10.1111/1365-2435.13303

Nueda, M. J., Conesa, A., Westerhuis, J. A., Hoefsloot, H. C., Smilde, A. K., Talon, M., et al. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA–SCA. *Bioinformatics* 23, 1792–1800. doi: 10.1093/bioinformatics/btm251

Orenti, A., Marano, G., Boracchi, P., and Marubini, E. (2012). Pinpointing outliers in experimental data: the Hat matrix in Anova for fixed and mixed effects models. *Ital. J. Public Health* 9. doi: 10.2427/8663

Philippot, L., Raaijmakers, J. M., Lemanceau, P., and van der Putten, W. H. (2013). Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* 11, 789–799. doi: 10.1038/nrmicro3109

Radhakrishna Rao, C., and Toutenburg, H. (1999). *Linear models: least squares and alternatives* (New York: Springer).

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Saccenti, E., Smilde, A. K., and Camacho, J. (2018). Group-wise ANOVA simultaneous component analysis for designed omics experiments. *Metabolomics* 14, 1–18. doi: 10.1007/s11306-018-1369-1

Sanow, S., Kuang, W., Schaaf, G., Huesgen, P., Schurr, U., Roessner, U., et al. (2023). Molecular mechanisms of pseudomonas-assisted plant nitrogen uptake: Opportunities for modern agriculture. *Mol. Plant-Microbe Interact.* 36, 536–548. doi: 10.1094/MPMI-10-22-0223-CR

Savci, S. (2012). An agricultural pollutant: chemical fertilizer. *Int. J. Environ. Sci. Dev.* 3, 73. doi: 10.7763/IJESD.2012.V3.191

Sellstedt, A., and Richau, K. H. (2013). Aspects of nitrogen-fixing Actinobacteria, in particular free-living and symbiotic Frankia. *FEMS Microbiol. Lett.* 342, 179–186. doi: 10.1111/1574-6968.12116

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C., Lamers, R.-J. A., van der Greef, J., and Timmerman, M. E. (2005). Anova-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* 21, 3043–3048. doi: 10.1093/bioinformatics/bti476

Tarazona, S., Prado-Lopez, S., Dopazo, J., Ferrer, A., and Conesa, A. (2012). Variable selection for multifactorial genomic data. *Chemomet Intell Lab. Syst.* 110, 113–122. doi: 10.1016/j.chemolab.2011.10.012

Thiel, M., Benaiche, N., Martin, M., Franceschini, S., Van Oirbeek, R., and Govaerts, B. (2023). limpca: An R package for the linear modeling of high-dimensional designed data based on ASCA/APCA family of methods. *J. Chemomet* 37, e3482. doi: 10.1002/cem.3482

Thiel, M., Feraud, B., and Govaerts, B. (2017). ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J. Chemomet* 31, e2895. doi: 10.1002/cem.2895

Videira, S. S., De Araujo, J. L. S., da Silva Rodrigues, L., Baldani, V. L. D., and Baldani, J. I. (2009). Occurrence and diversity of nitrogen-fixing Sphingomonas bacteria associated with rice plants grown in Brazil. *FEMS Microbiol. Lett.* 293, 11–19. doi: 10.1111/j.1574-6968.2008.01475.x

Vis, D. J., Westerhuis, J. A., Smilde, A. K., and van der Greef, J. (2007). Statistical validation of megavariate effects in ASCA. *BMC Bioinf.* 8, 1–8. doi: 10.1186/1471-2105-8-322

Weißbecker, C., Schnabel, B., and Heintz-Buschart, A. (2020). Dadasnake, a Snakemake implementation of DADA2 to process amplicon sequencing data for microbial ecology. *GigaScience* 9, giaa135. doi: 10.1093/gigascience/giaa135

Wu, X., Wang, X., Meng, H., Zhang, J., Lead, J. R., and Hong, J. (2023). Pseudomonas fluorescens with nitrogen-fixing function facilitates nitrogen recovery in reclaimed coal mining soils. *Microorganisms* 12, 9. doi: 10.3390/microorganisms12010009

Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome cancer research. *Gut Microbes* 15, 2244139. doi: 10.1080/19490976.2023.2244139

Yoon, G., Gaynanova, I., and Muller, C. L. (2019). Microbial networks in SPRING-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.* 10, 449195. doi: 10.3389/fgene.2019.00516

You, Y., Aho, K., Lohse, K. A., Schwabedissen, S. G., Ledbetter, R. N., and Magnuson, T. S. (2021). Biological soil crust bacterial communities vary along climatic and shrub cover gradients within a sagebrush steppe ecosystem. *Front. Microbiol.* 12, 569791. doi: 10.3389/fmicb.2021.569791

Zancarini, A., Westerhuis, J. A., Smilde, A. K., and Bouwmeester, H. J. (2021). Integration of omics data to unravel root microbiome recruitment. *Curr. Opin. Biotechnol.* 70, 255–261. doi: 10.1016/j.copbio.2021.06.016

Zhang, X., Li, A., Szewzyk, U., and Ma, F. (2016a). Improvement of biological nitrogen removal with nitrate-dependent fe (ii) oxidation bacterium Aquabacterium parvum B6 in an up-flow bioreactor for wastewater treatment. *Bioresource Technol.* 219, 624–631. doi: 10.1016/j.biortech.2016.08.041

Zhang, X., Mallick, H., and Yi, N. (2016b). Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *J. Bioinf. Genomics* 2, 1–9.

Zhou, H., He, K., Chen, J., and Zhang, X. (2022). LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome Biol.* 23, 95. doi: 10.1186/s13059-022-02655-5

Zwanenburg, G., Hoefsloot, H. C., Westerhuis, J. A., Jansen, J. J., and Smilde, A. K. (2011). Anova–principal component analysis and ANOVA–simultaneous component analysis: a comparison. *J. Chemomet* 25, 561–567. doi: 10.1002/cem.1400