NAR Genomics and Bioinformatics, 2024, **6**, lqae038
https://doi.org/10.1093/nargab/lqae038
Advance access publication date: 25 April 2024
Methods Article

Microbiome compositional data analysis for survival studies

Meritxell Pujolassos[1], **Antoni Susín**[2] **and M.Luz Calle**[1,3,*]

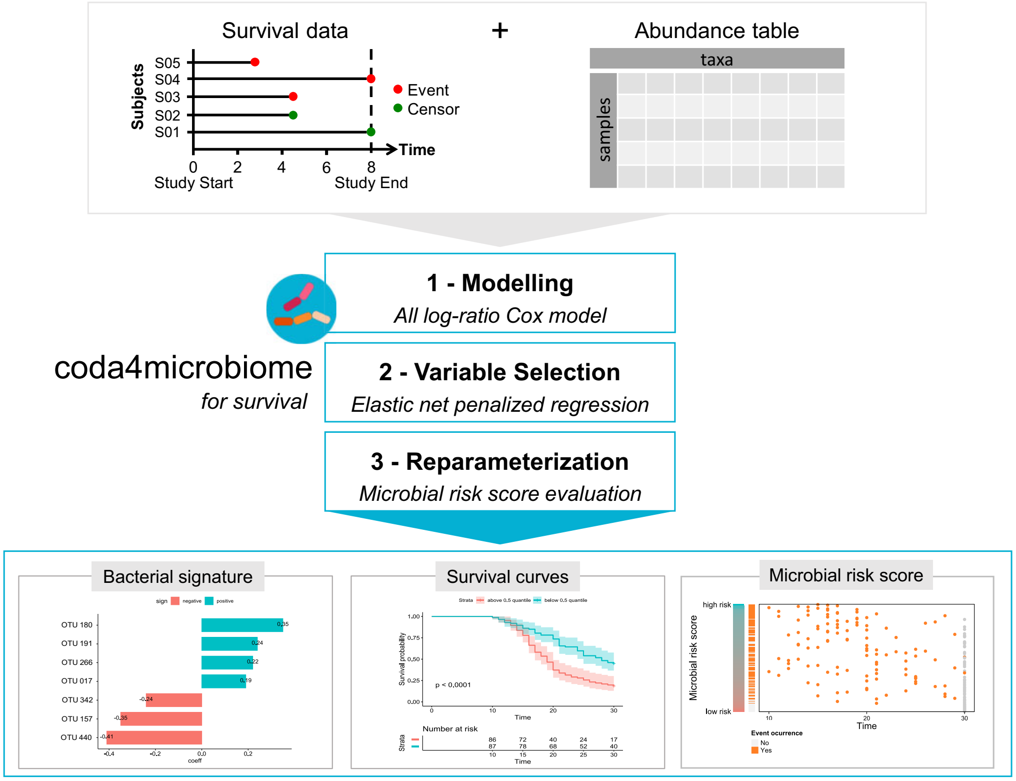
[1]Bioscience Department, Faculty of Sciences, Technology and Engineering, University of Vic – Central University of Catalunya, Vic 08500, Spain
[2]Mathematical Department, UPC-Barcelona Tech, Barcelona 08034, Spain
[3]Institut de Recerca i Innovació en Ciències de la Vida i de la Salut a la Catalunya Central (IRIS-CC), Vic 08500, Spain

[*]To whom correspondence should be addressed. Tel: +34 938861222; Email: malu.calle@uvic.cat

Abstract

The growing interest in studying the relationship between the human microbiome and our health has also extended to time-to-event studies where researchers explore the connection between the microbiome and the occurrence of a specific event of interest. The analysis of microbiome obtained through high throughput sequencing techniques requires the use of specialized Compositional Data Analysis (CoDA) methods designed to accommodate its compositional nature. There is a limited availability of statistical tools for microbiome analysis that incorporate CoDA, and this is even more pronounced in the context of survival analysis. To fill this methodological gap, we present *coda4microbiome* for survival studies, a new methodology for the identification of microbial signatures in time-to-event studies. The algorithm implements an elastic-net penalized Cox regression model adapted to compositional covariates. We illustrate *coda4microbiome* algorithm for survival studies with a case study about the time to develop type 1 diabetes for non-obese diabetic mice. Our algorithm identified a bacterial signature composed of 21 genera associated with diabetes development. *coda4microbiome* for survival studies is integrated in the R package *coda4microbiome* as an extension of the existing functions for cross-sectional and longitudinal studies.

Graphical abstract



Received: November 8, 2023. **Revised:** January 23, 2024. **Editorial Decision:** March 27, 2024. **Accepted:** April 8, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

The advent of high throughput sequencing techniques, including sequencing of 16S ribosomal rRNA and shotgun metagenomics, has greatly facilitated the study of the role of the human microbiome in health and disease. Although specific mechanisms of interaction are still uncertain, the growing body of evidence connecting the human microbiome and our health underscores the relevance of microbiome research in identifying novel biomarkers for disease diagnosis and prognosis, as well as for improving treatments for specific diseases (1,2).

While microbiome studies are promising, its analysis still faces many experimental and computational challenges (3). One of them is the compositional nature of microbiome data, which adds complexity to its statistical analysis. Compositions are commonly defined as vectors of positive real numbers constrained to a total sum (4) but they have a broader definition as vectors of positive real numbers with parts or components carrying relative information with respect to each other. The quantification of microbial compositions from high throughput sequencing techniques is limited by the sequencing depth which entails compelled dependencies between the observed abundances of each taxon in the sample. This means that a change in the relative abundance of one microbe drives changes in the observed relative abundances of the others just to fulfil the total sum constraint (5). In 1982, Aitchison laid the foundations of Compositional Data Analysis (CoDA) and suggested the log-ratio approach. It consists of analysing logarithms of ratios between components to extract their relative information instead of analysing each component separately (4).

In recent years, microbiome analyses are being included in time-to-event studies, also called survival studies, where researchers analyse the time until a given event occurs, e.g., the actual death of an individual, onset or reemission of a disease or response to a treatment, to understand its relationship with other features. Including microbiome analysis in representative, well-phenotyped population cohort studies, with sufficient follow-up time, allowed the identification of microbial signatures associated to overall mortality, and particularly highly related to gastrointestinal and respiratory causes (6); as well as the definition of a microbiome 'uniqueness' index associated to healthy ageing and predictive for all-cause mortality risk (7). Microbiome has also been used for assessing chemoradiation performance in cervical cancer patients' survival (8), and in colorectal cancer prognosis (9).

Few statistical tools have been specifically developed for analysing microbiome data in survival studies and even less within the CoDA framework. Most of the microbial survival studies either ignore the compositionality of microbiome data or, at best, perform the centered log-ratio (clr) transformation of microbiome data using the geometric mean of the composition followed by standard survival analysis. While this is a handy option, we expose in the discussion section the limitations of this approach. Other CoDA methods that have been proposed for microbiome analysis, like *ANCOM-II* (10), *ANCOM-BC* (11), *ALDEx2* (12) or *Selbal* (13), are not suitable for survival studies.

We recently developed *coda4microbiome* (14), a new algorithm for microbiome analysis in the CoDA framework implemented for cross-sectional and longitudinal studies. In this work we present the extension of *coda4microbiome* for survival studies. As described in Calle *et al.* (14), *coda4microbiome* algorithm aims to identify a microbial signature, i.e. a model based on microbial abundances, that best predicts a given response variable. In survival studies, the goal will be to identify a microbial signature that is associated to the risk of developing an event of interest. As in the existing *coda4microbiome* algorithm, its extension to survival data follows three main steps that are briefly explained hereafter: modelling, variable selection and reparameterization.

Cox's proportional hazard regression model (15) is one of the most frequently used models for survival data and it is also the choice in our case. However, Cox's model cannot be directly applied to compositional covariates without some previous log-ratio transformation that map these features from the simplex to the real space. Instead of using the clr-transformation mentioned above, we consider the 'all-pairwise log-ratio' Cox's model, i.e., a Cox's proportional hazard model with all possible pairwise log-ratios of microbial features as regressors.

The stated Cox's model will most likely be high-dimensional since in microbiome studies the number of microbial taxa (K) is usually larger than the number of samples (n), and even more in this case that we are considering all pairs of taxa which increases the number of features from K to $K \cdot (K-1)/2$. In this setting, classical approaches for model fitting will be either infeasible or will likely induce overfitting (16). To address this high-dimensional problem we consider penalized regression (17) that simultaneously performs variable selection and model fitting. Penalized regression consists in adding a penalty term to the objective function, which shrinks the coefficient estimates toward zero and forces some of them to be exactly zero. This results in a parsimonious Cox's model containing the most relevant features (in our case, log-ratios of pairs of microbial taxa) for predicting the risk of developing the event of interest. We will call *microbial risk scores* to the linear predictions of the estimated Cox's model.

Since the interpretation of a model whose predictors are pairwise log-ratios is far from straightforward, we take advantage of the properties of logarithms to expand log-ratios. This reparameterization results in a Cox's model expressed in terms of the initial microbial features (log-transformed abundances) which is much more meaningful. Moreover, by construction, the sum of the coefficients of the expanded model is zero. This defines two groups of taxa, those with a positive coefficient that contribute to larger microbial risk scores and those with a negative coefficient that contribute to smaller risk scores. Hence, the risk of developing the event of interest is expressed in relation to the relative abundances between these two groups of taxa.

In this article, we describe the methodology of the new algorithm for microbiome survival studies in detail, along with its main functions in sections '*coda4microbiome* algorithm for survival data' and '*coda4microbiome* for survival main functions', respectively (Materials and Methods). We illustrate *coda4microbiome* for survival data assessing the relationship between mice gut microbiome and type 1 diabetes development rate in section 'Microbiome and type 1 diabetes onset' (Results). We further discuss the utility and main advantages of *coda4microbiome* algorithm in comparison to log-transformation approaches in the Discussion section.

The new functions developed for survival analysis have been added to the existing *coda4microbiome* R package, available at CRAN (https://cran.r-project.org/web/packages/coda4microbiome/index.html). A detailed tutorial about the new functions is available in *coda4microbiome* blog (https://malucalle.github.io/coda4microbiome/). The data and code for reproducing the analysis are available at DOI 10.5281/zenodo.10552383.

Materials and methods

coda4microbiome algorithm for survival data

Assume a survival study with n individuals where time-to-event for subject i is denoted as t_i. Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{iK})$ be the microbial composition for K taxa in the i-th subject. Microbial abundances (X) can be either raw counts or relative abundances.

The goal of *coda4microbiome* algorithm for survival data is to identify those microbial taxa whose relative abundances are associated to survival time.

We consider the Cox's proportional hazards regression model (15) with all possible pairwise log-ratios of taxa as covariates. This regression model states a possible relationship between pairs of microbial species (log-ratios) and the risk of the given event to occur.

Let $h(t|X)$ be the hazard function at time t for an individual with microbial composition $\boldsymbol{X}$. The all pairwise log-ratios Cox's proportional hazards model is given by:

$$h(t|\boldsymbol{X}) = h_0(t) \cdot \exp\left(\sum_{1 \leq \mathrm{j} < \mathrm{k} \leq \mathrm{K}} \beta_{jk} \cdot \log(X_j/X_k)\right)$$

where $h_0(t)$ is the baseline hazard of the model and β_{jk} is the regression coefficient for the log-ratio between components X_j and X_k.

This model can also be expressed as a generalized linear model where the logarithm of the hazard ratio is a linear combination of all pairwise log-ratios:

$$\log\left(\frac{h\,(t|\boldsymbol{X})}{h_0\,(t)}\right) = \sum_{1 \leq \mathrm{j} < \mathrm{k} \leq \mathrm{K}} \beta_{jk} \cdot \log\left(X_j/X_k\right). \qquad (1)$$

With the aim of identifying which taxa are associated with the outcome, variable selection is carried out by the estimation of the regression coefficients (β_{jk}) subjected to an elastic-net penalization (Equation (2)) where L is the log partial likelihood function for the Cox model (18).

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{L(\beta) + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1\} \qquad (2)$$

Such penalization can also be written in terms of λ and α, with $\lambda_1 = \lambda(1-\alpha)$ and $\lambda_2 = \lambda\alpha$, where λ controls the amount of penalization and α the mixing between norms. By default, α is set to 0.9 but this value is adjustable by the user.

With the appropriate penalization some coefficients are shrank to zero which leads to the elimination of those log-ratios less associated to the outcome. The optimal value for λ is selected after a cross-validation process performed with the function *cv.glmnet()* from *glmnet* R package (19) and considering the maximization of the Harrell's C-index statistic (20). The C-index is defined as the probability that a given sample with lower risk score outlives a sample with higher risk score. This measure reports how well the survival model predicts the observed sequence of events. The algorithm also allows for adjustment of non-compositional variables (e.g., age, sex, clinical variables, etc.) that are added as an offset into the *cv.glmnet()* function.

After modelling and variable selection, the result is a Cox model composed by the logarithms of pairs of taxa with non-zero coefficient.

$$\log\left(\frac{h\,(t|\boldsymbol{X})}{h_0\,(t)}\right) = \sum_{1 \leq j < k \leq K} \hat{\beta}_{jk} \cdot \log\left(X_j/X_k\right). \qquad (3)$$

The linearity of logarithms permits the reparameterization of (Equation (3)) into single taxa, instead of pairs of taxa, which makes interpretation of results more meaningful:

$$\log\left(\frac{h\,(t|\boldsymbol{X})}{h_0\,(t)}\right) = \sum_{1 \leq j \leq K} \hat{\theta}_j \cdot \log\left(X_j\right) \qquad (4)$$

where $\hat{\theta}_j$ is the sum of the coefficients $\hat{\beta}$ that correspond to a log-ratio involving component j.

The linear predictor of the model, i.e., the right part of Equation ((4)), provides a numerical value that is related to the risk of developing the event. In our context, we call it *microbial risk score*, M, since it is obtained as the combination of microbial abundances. For each individual $i \in \{1, \ldots, n\}$, its microbial risk score is given by:

$$M_i = \sum_{1 \leq j \leq K} \hat{\theta}_j \cdot \log(X_{ij}) \qquad (5)$$

It can be proved that this final microbial signature is a log-contrast function, *i.e.*,$\sum_{j=1}^{K} \hat{\theta}_j = 0$ (21). The zero-sum constraint ensures the scale invariance CoDA principle required for compositional data analysis (22). It also provides a convenient interpretation of the signature as a weighted balance between two groups of taxa, those with positive coefficient vs those with negative coefficient, as illustrated in the example below. See (23) for a formal definition of weighted balance.

Interpretability of the model through a toy example

We illustrate with a toy example the main features of the proposed methodology, emphasizing the reparameterization step that ensures the model is a log-contrast and the interpretation of the resulting microbial signature as a balance between two groups of taxa.

Let's consider a microbial community of 5 taxa whose abundance composition is given by $X = (X_1, X_2, X_3, X_4, X_5)$. The algorithm aims to find the optimal combination of these five taxa abundances, or a subset of them, that can accurately predict the survival time of interest. This is accomplished through three main steps:

(1) Modelling. The Cox proportional hazards model with all pairwise log-ratios is considered. Given 5 initial variables, the number of pairwise log-ratios is equal to $\binom{5}{2} = 10$. Thus, the Cox model has 10 coefficients, $\beta_{ij}, 1 \leq i < j \leq 5$:

$$\log\left(\frac{h(t|\boldsymbol{X})}{h_0(t)}\right) = \beta_{12}\log\left(X_1/X_2\right) + \beta_{13}\log\left(X_1/X_3\right)$$
$$+ \ldots + \beta_{15}\log\left(X_1/X_5\right) + \beta_{23}\log\left(X_2/X_3\right)$$
$$+ \ldots + \beta_{25}\log\left(X_2/X_5\right) \; + \beta_{34}\log\left(X_3/X_4\right)$$
$$+\beta_{35}\log\left(X_3/X_5\right) \; + \beta_{45}\log\left(X_4/X_5\right)$$

(2) Variable selection. Penalized regression (elastic-net) is applied to estimate the coefficients of the above Cox model.
Let's assume that after elastic-net penalization only four coefficients are different from zero and their values are: $\beta_{13} = 1, \quad \beta_{15} = 5, \ \beta_{23} = 2$ and $\beta_{25} = 4$. Thus, the estimated Cox model is given by:

$$\log\left(\frac{h(t|\mathbf{X})}{h_0(t)}\right) = \log(X_1/X_3) + 5\log(X_1/X_5) + 2\log(X_2/X_3) + 4\log(X_2/X_5) \quad (6)$$

(3) Reparameterization:

By expanding the logarithm of a ratio as the difference of logarithms, we can rewrite the above Cox model as:

$$\log\left(\frac{h(t|\mathbf{X})}{h_0(t)}\right) = \log(X_1) - \log(X_3) + 5\log(X_1) - 5\log(X_5) + 2\log(X_2) - 2\log(X_2) + 4\log(X_2) - 4\log(X_5)$$

After aggregating the terms corresponding to the same variables, it reduces to a model with coefficients that sum to zero, confirming that the model is log-contrast:

$$\log\left(\frac{h(t|\mathbf{X})}{h_0(t)}\right) = 6\log(X_1) + 4\log(X_2) - \log(X_3) - 9\log(X_5) \quad (7)$$

Model 1 (Equation 6) and model 2 (Equation 7) are equivalent, but model 1 is expressed with log-ratios, making it more challenging to interpret. Instead, model 2 is a linear combination of (log-transformed) variables that we are much more familiar with.

The right part of (Equation 7), referred as *microbial risk score*, and denoted by M, provides the combination of microbial abundances that best predicts survival time. Large values of M are associated to large hazard ratios, *i.e.*, larger risks than the baseline.

In this example, the microbial risk score is given by the combination of four out of the five initial taxa: $M = 6\log(X_1) + 4\log(X_2) - \log(X_3) - 9\log(X_5)$. As mentioned above, M can be interpreted as a weighted balance between two groups of taxa: those that contribute positively to the risk of developing the event of interest (taxa X_1 and X_2 with weights 6 and 4) and those that contribute negatively to the risk (taxa X_3 and X_5 with weights 1 and 9). Taxa X_4 appears to be not related to the survival time, since it is not part of the model. Let's consider two individuals with microbial compositions (1, 2, 10, 5, 30) and (12, 25, 15, 10, 17), respectively. For the first subject, the microbial risk score M is equal to −1.3, which means that the balance tilts towards the variables X_3 and X_5, resulting in a lower risk of developing the event of interest than the second subject that has a microbial risk score of 9.7, corresponding to a balance that leans towards taxa X_1 and X_2. To be noticed that these risk scores can be calculated without concern that the total abundance of the two individuals is different (48 and 79, respectively). This is because the microbial risk score is a log-contrast function, ensuring scale invariance. Indeed, the same risk scores would be obtained if the abundances are normalized to relative abundances beforehand.

A graphical representation of the contribution of each taxon to the microbial risk score is provided by the signature plot (Figure 1).

coda4microbiome for survival main functions

The algorithm for time-to-event data is implemented in the *coda_coxnet()* function within the R package *coda4microbiome*. Other two functions, *plot_survcurves()* and *plot_prediction_surv()*, are added into the package for a graphical representation of the results. We briefly describe these functions below and their implementation is illustrated in a case study in section 'Microbiome and type 1 diabetes onset in mice'.

To perform a survival analysis with *coda_coxnet()* function, three essential inputs are needed: the taxa abundance table (either relative or absolute abundances), the survival time, and the event occurrence for each sample. It is possible to adjust by any non-compositional variable introducing a dataset containing covariates. Other parameters editable by the user are the number of variables to use in the variable selection step, level of mixing between L1 and L2 norms in elastic net penalization, number of folds in cross-validation process and the minimum absolute value of the coefficient for a variable to be included in the final model.

coda_coxnet() function returns three different results: (i) the *coda4microbiome* model in terms of the selected variables and their respective coefficients, which are also graphically represented in a bar plot called the 'signature plot'; (ii) the microbial risk score for each sample and its graphical representation in a heatmap, called the 'risk score plot', which displays samples sorted by their microbial risk score together with an adjacent scatterplot of survival times and (iii) a summary of the model accuracy that includes the C-index value of the signature applied to the same data used to generate the model (apparent C-index) and the mean C-index value and its standard deviation from *cv.glmnet()* output.

The risk score plot can also be independently generated with *plot_riskscore()* function using the obtained microbial risk score from *coda_coxnet()* output.

We developed a third function for an ultimately visualization of results: *plot_survcurves()*. The function plots the survival curves of samples stratified in two groups according to their microbial risk score. Stratification threshold is set by default to the median value of the overall microbial risk, but it can be adjusted by the user. The plot also displays the P-value of the log-rank test (24) between the survival curves, and a table with the number of individuals at risk at every time.

Results

Microbiome and type 1 diabetes onset in mice

We illustrate *coda4microbiome* algorithm for survival studies evaluating the association between the time to develop type 1 diabetes (T1D) and gut microbiome with data from a non-obese diabetic (NOD) mice study (25). T1D is an autoimmune disease that is gaining incidence worldwide, also among paediatric population. Early life exposure to antibiotics is critical for immune system development and it might lead to an acceleration of T1D onset (26). In that line, Zhang *et al.* (25) showed the effect of a single use of antibiotic on T1D development rate in NOD mice. We illustrate our methodology using survival data from Zhang's study, which was processed

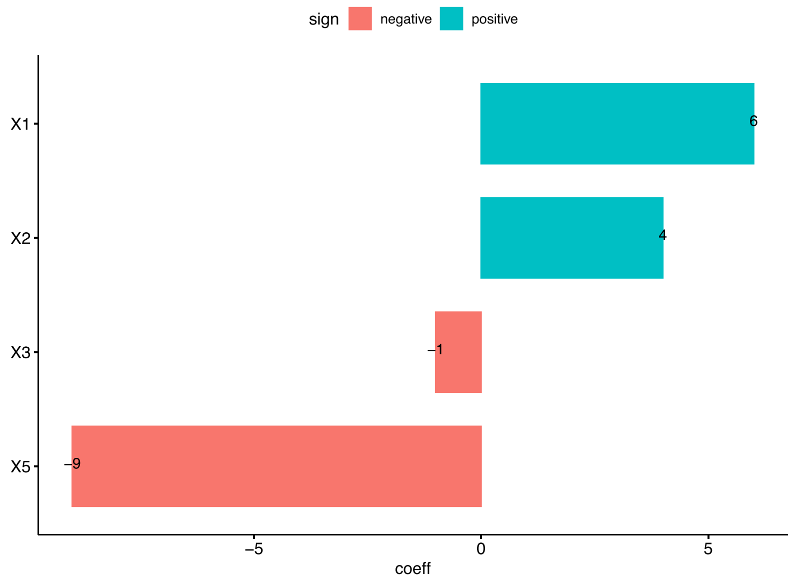


Figure 1. Signature plot: bar plot representing the contribution of each variable to the survival risk, *i.e.,* their coefficient in the model. Negative coefficients in red and positive coefficients in blue.

by Gu *et al.* (27). Our analysis was performed at the genus level, keeping taxa present in more than 5% of samples, *i.e.*, we removed those taxa with 95% zeros or more, for not being statistically informative. Doing this, we are not assuming that these low abundant taxa are not biologically important, we are just removing them because, with the resolution of the sequencing process, they cannot contribute to the statistical model. We used the genus level for illustrative purposes, but the algorithm can be applied to any other taxonomic level. The analysis at different taxonomic levels can reveal specific aspects and contribute to a more comprehensive understanding of the problem.

The processed data, available at *coda4microbiome* package, includes metagenomic information of 30 taxa for 173 mice (55 T1D free and 118 that developed T1D). The dataset also includes whether mice developed T1D or not, the time to T1D onset (survival time), as well as sex and antibiotic administration.

To identify the bacterial signature associated to the time of T1D development we implemented *coda4microbiome::coda_coxnet()* function. We adjusted the model by sex and antibiotic administration to obtain a bacterial signature that is not affected by these possible confounders. The new function performs variable selection through a cross-validation of the penalized regression by implementing *cv.glmnet()* from glmnet package. The function sets a sequence of lambda values and cross-validates every penalized model. Red dots in Figure 2 correspond to the mean cross-validation measures (*C*-index) for every lambda value, and their standard deviations are represented with the upper and lower error bars. The two vertical dashed lines in Figure 2 correspond to values of 'lambda.min' and 'lambda.1se'. On the left, 'lambda.min' is the degree of penalization that provides minimum mean cross-validated error or, in this case, maximum *C*-index. On the right, 'lambda.1se' is the value of lambda that provides the most parsimonious model with a cross-validated *C*-index within one standard error of the maximum. By default, *coda4microbiome* uses 'lambda.1se'. In our example, the fitted model with 'lambda.1se' results in 21 log-ratios with non-zero coefficient. Through reparameterization, the log-ratios are expanded, and the model is expressed as a log-contrast of 21 different taxa, each one with a specific contribution to the balance: 10 taxa with positive coefficient and 11 with negative coefficient (Figure 3).

The model provides a combination of microbial abundances that determines the risk of developing T1D in each subject; we denote such risk as the *microbial risk score*. A graphical representation of microbial risk scores and survival times is given by the 'risk score plot' (Figure 4). The graphic displays samples vertically ordered according to their microbial risk score (left column: from low to high risk of developing the event of interest) and their observed times (horizontal axis: from the beginning to the end of the experiment). Samples that developed the event of interest are plotted in orange, and censored samples in grey. A vertical bar at the right plot indicates whether the individual has experienced the event or not.

In our example, 'Event occurrence' refers to the development (or not) of T1D; and 'Time' corresponds to time until the development of T1D or the duration of the experiment. The risk score plot (Figure 4) is useful to graphically explore the association between microbial risk scores and the time to development of T1D. By comparing the left bar of the plot (microbial risk score) with the right bar of the plot (occurrence of the event), we can see a higher presence of T1D development events for those individuals with higher microbial risk scores. When focusing on the distribution of survival times, we observe that the times to development T1D for individuals with higher risk scores are slightly shorter than for individuals with lower risk scores. Moreover, censored samples are more abundant among those individuals with lower microbial risk scores.

For an additional assessment of the possible association between the microbial risk score and the time to development of T1D (survival time), we analysed the survival curves of NOD mice stratified according to their microbial risk score

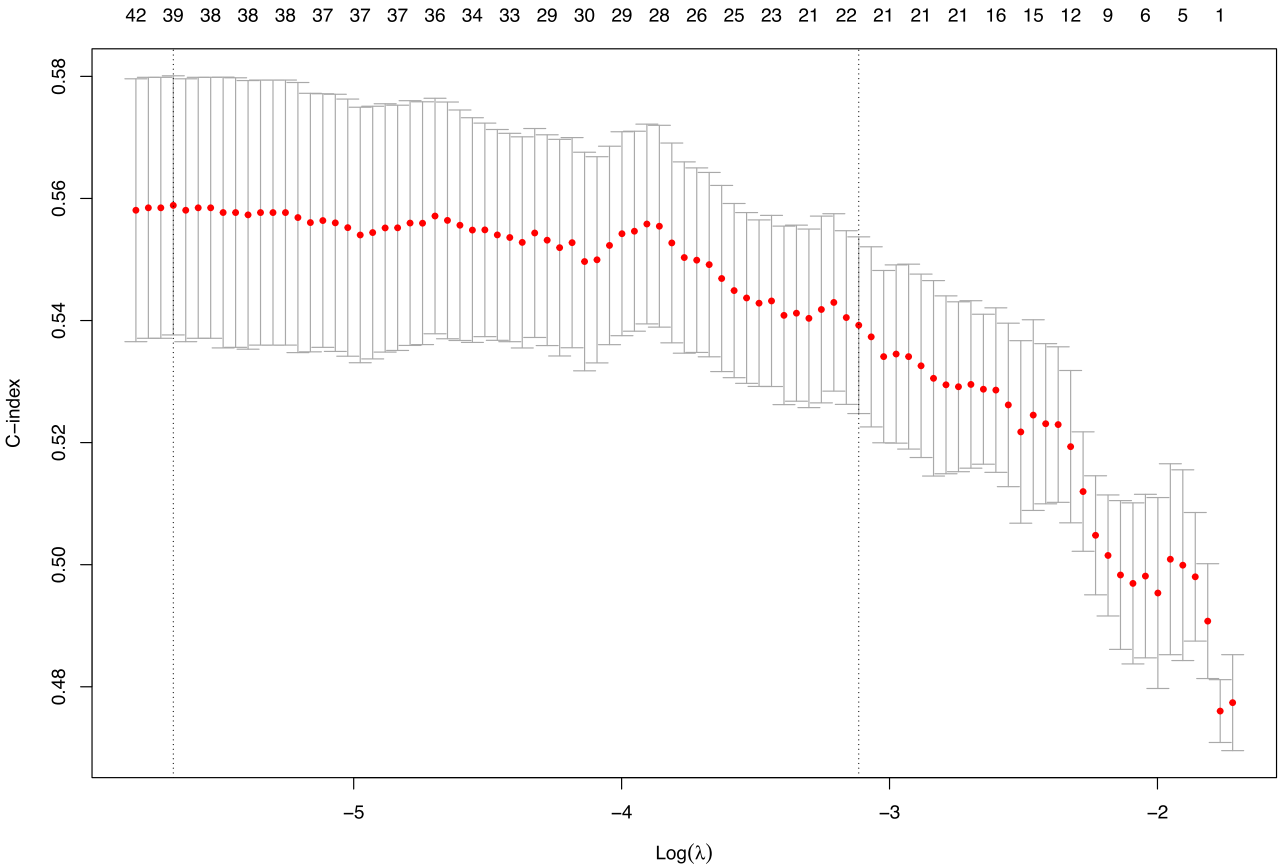


Figure 2. Cross-validation accuracy curve for different degrees of penalization. Horizontal axis reports log-transformed penalization parameter (λ) while cross-validation *C*-index is reported along the vertical axis. Vertical dashed-lines highlight lambda.min and lambda.1se values of penalization. On top, the number of selected log-ratios for each degree of penalization.

(Figure 5). The plot, that was generated with plot_survcurves() function, shows that individuals with higher microbial risk scores (above the median, red line) have shorter times of T1D development compared to individuals presenting lower values of the risk score (below the median, blue line), which develop T1D later in time. Differences in T1D development times appear to be statistically significant according to the log-rank test (24) (*P*-value < 0.0001).

Model assessment: proportional hazards assumption

Our algorithm implements the Cox's regression model (18), which assumes proportionality of hazards, *i.e.*, the hazard ratio between one group of individuals and the baseline group is constant over time. It also assumes linearity between the log hazard ratio and each covariate. Both, the proportional hazards assumption and the linearity assumption can be assessed by statistical tests or graphical representations that test each explanatory variable individually. However, it is not clear how to test these hypotheses in high-dimensional settings involving penalized regression. Instead, in this study we tested the Cox model obtained after variable selection with *coda4microbiome* algorithm. Specifically, we considered the Cox's proportional hazard model with the obtained T1D microbial risk score, *M*:

$$h(t|M) = h_0(t) \cdot \exp(\beta \cdot M) \tag{8}$$

This model was then tested for the proportional hazard assumption using *survival::cox.zph()* which implements Grambsch and Therneau test (28). The proportional hazards assumption was not rejected with a *P*-value $= 0.3$. Accordingly, the graphical inspection of the Schoenfeld residuals in Figure 6A does not show any pattern along time since proportional hazards assumes that $\hat{\beta}$ do not vary over time. Regarding the linear relationship between predictors and the outcome, we graphically tested residuals deviance of the Cox model (Figure 6B). Though this does not prove the assumptions of the initial Cox model (Equation (1)), it provides an additional interpretation of results: $\exp(\hat{\beta})$ is the estimated hazard ratio between two individuals whose microbial risk scores differ in 1 unit, and this ratio remains constant over time. In our example, taking as the reference group those individuals with balanced composition between the two bacterial groups (*i.e.*,M_i=0), the risk of T1D for an individual with M_i=1 is increased by a factor of $\exp(\hat{\beta}) = 5.48$.

Discussion

The intricate interpretability of CoDA methods and the limited availability of specialized software pose a significant obstacle in tackling compositionality in microbiome data analysis. This is even more evident in the case of survival studies where the existence of specific CoDA algorithms for this setting is really scarce. Among those who are aware of the

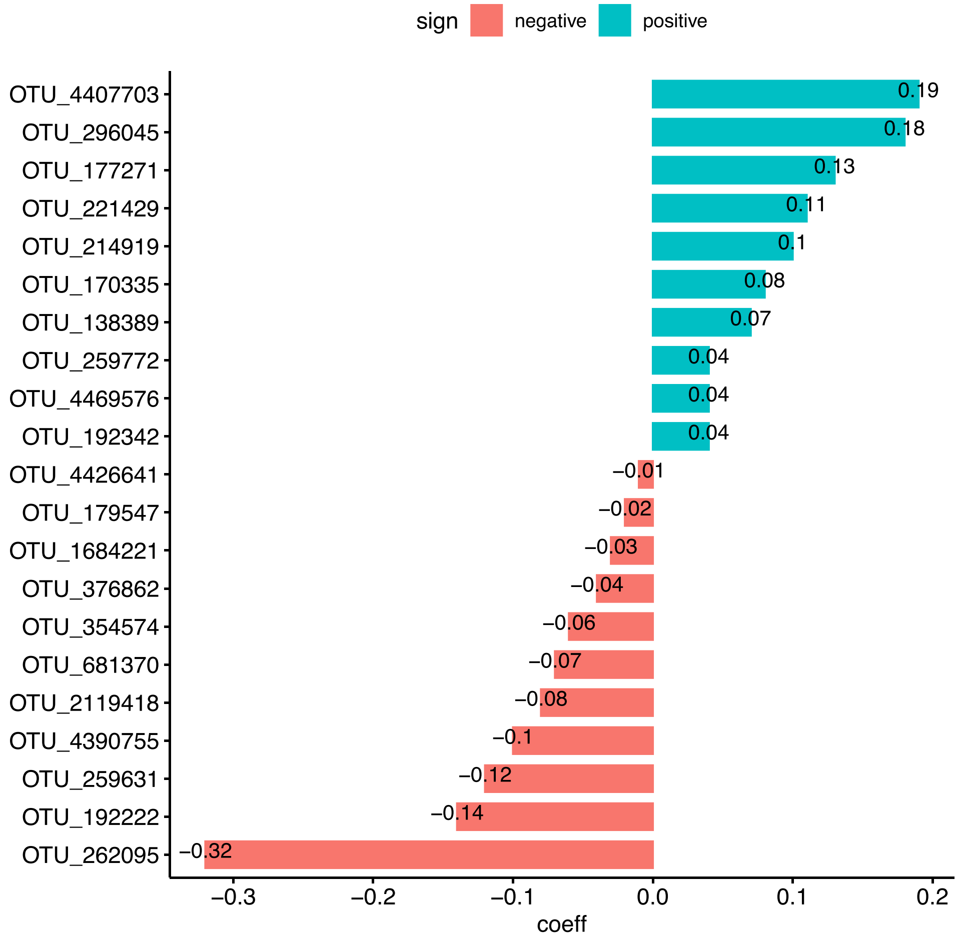


Figure 3. Microbial signature for T1D onset. The final model for T1D onset risk prediction is composed of a balance between two groups of taxa. Those that contribute to the microbial signature with a positive coefficient ($\hat{\boldsymbol{\theta}}$) are plotted in blue and those with negative coefficient in red.

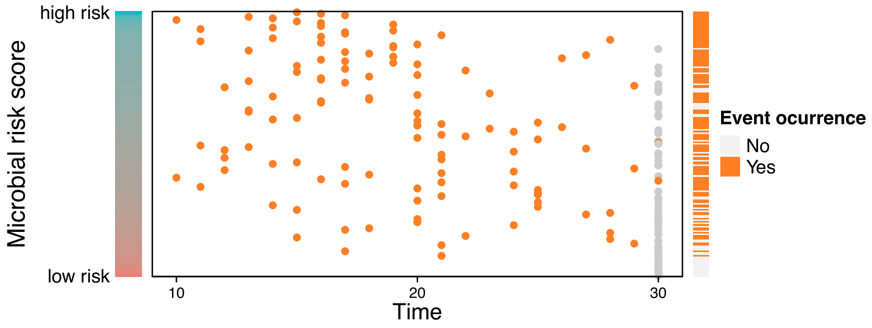


Figure 4. Microbial risk score plot. Samples are ordered by their microbial risk score (vertical left axis) and plotted along time (horizontal axis). In orange, samples that developed T1D (event occurrence = 'Yes'); in grey censored samples (event occurrence = No).

need to consider data compositionality in the analysis, the most common approach is to normalize microbiome data by the geometric mean of the composition (clr transformation) and then apply the Cox regression model to look for associations between features and the risk of developing a given event. In this case, one should be cautious about interpreting the transformed variables as if they were the original variables. In general, any method that rely on log-ratio transformations should be carefully interpreted since its results depend on the reference used. In particular, one should not interpret clr-transformed variables as single features without taking into account their dependence on the geometric mean (29). A problem of this approach when the goal is variable selection is that irrelevant features are never completely removed from the analysis since the normalization term (the geometric mean) of the selected clr-transformed variables contains all components. Simulation studies showed that the power of selecting important variables is reduced by the clr transformation (23) for small compositions due to the high variability of the geometric mean, whereas when the number of variables is large, the effect on variable selection performance is negligible (23). Finally, the clr transformation is not sub-compositionally

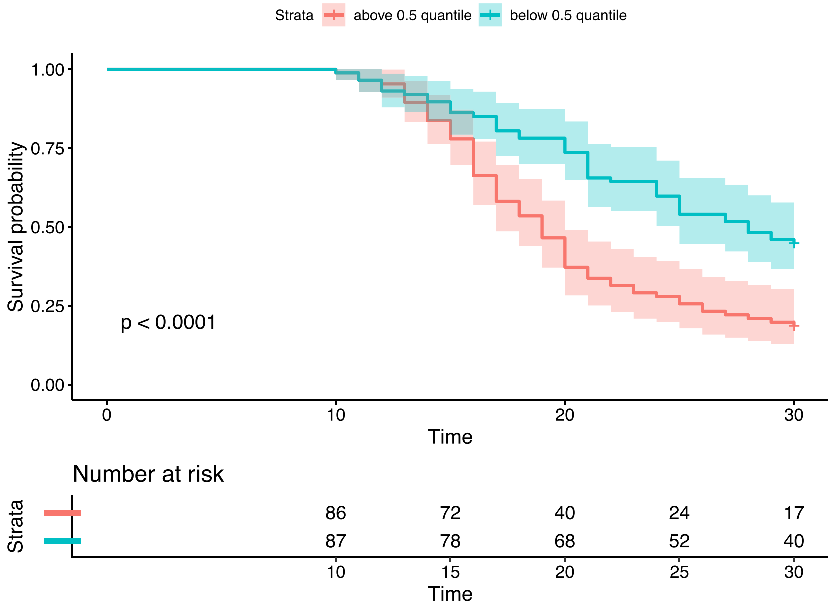


Figure 5. Survival curves stratified by microbial risk score. Sample stratification is based on the median risk score value (default): survival curve for samples presenting higher microbial risk scores than the median in red, and survival curve for samples with microbial risk scores below the median in blue. The *P*-value of the log-rank test is shown in the plot. On the bottom, a table with the number of samples at risk over time.

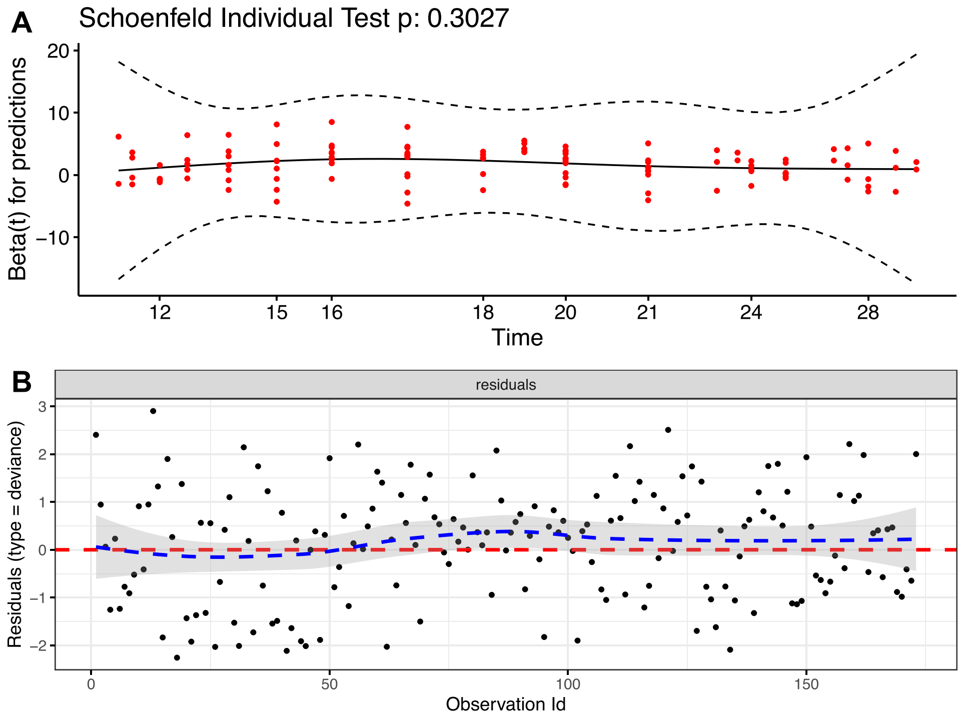


Figure 6. T1D Cox model assumptions assessment. (**A**) Graphical diagnostic of scaled Schoenfeld residuals over time. Dashed lines correspond to 2-standard-error around the fit (solid line). (**B**) Residuals deviance linearity check

coherent, which means that the results can differ substantially when the transformation is considered with the total composition or with a sub-composition. This complicates the transfer of results to new studies since the set of taxa in the original study may not align with the available taxa in the new study, rendering it impractical to apply the same clr transformation.

As an alternative to the clr transformation, McGregor *et al.* (30) introduced a new method for CoDA survival regression with compositional covariates: a Cox regression model involving the isometric log-ratio (ilr) transformation. They illustrated the methodology by analysing the association between mortality and a small composition involving only three components: sleeping time, physical activity, and sedentary behaviour. The main limitations of this approach are the difficulty of implementing the ilr transformation in high-dimensional contexts that involve compositions of hundreds of features, such as microbiome data, and the interpretation of results.

We believe that a key factor determining the utility and future adoption of a new algorithm for microbiome data analysis, in addition to its strong theoretical foundation, is the ease of interpreting the results it provides. With this intention in mind, we developed *coda4microbiome*. One of the most noteworthy aspects of *coda4microbiome* is that, even though it begins with an initial model using log-ratio transformed variables, the final model is expressed in terms of the original variables. By considering the model with all pairwise log-ratios, the constant-sum constraint of compositional data is removed, and the log-ratios are handled directly in the regression model, without any dependence restriction. After variable selection, the algorithm returns a log-contrast signature written in terms of the original single features. The final signature is compositionality coherent, ensuring the scale invariance principle (22).

As described above, the microbial risk score can be informally interpreted as a balance between two groups of taxa, measuring the contribution of one group with respect to the other. This should not be confused with amalgamation balances (31) that sum the abundances of each group of variables. The log-contrast function can be expressed as the log-ratio of two geometric means (23), thus involving the product of the abundances instead of their sum. The difference between the two types of balances is especially relevant regarding the effect of small abundance values. In a log-contrast function, the contribution of an abundance close to zero is very large since it involves the logarithm of the abundances. For instance, if a taxon is part of the log-contrast microbial signature with a positive coefficient and an individual has a very low abundance of this taxon, its microbial risk score will be highly negative (since the logarithm of small numbers are large negative values) meaning that the risk of developing the event is very low. Both models can provide different insights into the analysis of microbiome data.

The new functions of *coda4microbiome* for survival explore microbiome data with a principal focus on prediction. Unlike other differential abundance methods, *coda4microbiome* identifies the microbial signature with the minimum number of features that best predicts the risk of developing an event of interest (onset of a disease, response to a treatment or risk of death, for example). The likelihood of developing the event of interest, which we refer to as microbial risk score, is expressed in relation to the relative abundances of taxa that compose the bacterial signature. As mentioned before, we are deeply committed to making the results easily understandable, and especially in the context of survival data. For this reason, the package provides several functions for graphical representations of the taxa comprising the microbial signature, the predictive microbial risk score together with other variables and survival curves for different risk groups.

Our algorithm, *coda4microbiome* for survival, relies on the initial Cox regression model. It assumes constant hazards ratios over time for each of the pairwise log-ratios included as covariates. The proportional hazards assumption could be tested for each variable and globally using the Grambsch and Therneau test but in high dimensional microbiome studies it is not clear the practical utility of this verification. What should we do if one among hundreds of log-ratios does not satisfy the proportionality assumption? This is very likely to happen in a high-dimensional setting. Would this invalidate the Cox model? We don't think so. In fact, we are not interested in demonstrating a specific relationship between survival time and the pairwise log-ratios; we simply propose the Cox model as a tool or device for variable selection. Simulation studies showed that the lack of proportional hazards in penalized models may affect the selection of variables (32). However, since the focus of *coda4microbiome* is on prediction rather than inference, small departures from model assumptions could reduce the prediction accuracy of the model but should not have further implications in the analysis. On the other hand, testing the proportional hazards assumption of the identified signature in the final model can be useful for interpreting the risk of the event in relation to the microbial risk score.

Log-ratios methodologies should be preceded by proper imputation methods because they do not allow the presence of zero among variables. To deal with the high sparsity of microbiome data, *coda4microbiome* algorithm implements a simple imputation approach to avoid zeros but other imputation methods can be applied externally before running the algorithm.

With this work, we provide the scientific community with a new CoDA algorithm that we hope will be useful for the analysis of microbiome data in survival studies.

Data availability

The data and code for reproducing the analysis is available at DOI 10.5281/zenodo.10552383.

Funding

Spanish Ministry of Economy, Industry and Competitiveness [PID2019-104830RB-I00 to M.L.C., PID2021-122136OB-C21 to A.S.].

Conflict of interest statement

None declared.

References

1. Manor,O., Dai,C.L., Kornilov,S.A., Smith,B., Price,N.D., Lovejoy,J.C., Gibbons,S.M. and Magis,A.T. (2020) Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.*, **11**, 5206.
2. Duvallet,C., Gibbons,S.M., Gurry,T., Irizarry,R.A. and Alm,E.J. (2017) Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.*, **8**, 1784.
3. Moreno-Indias,I., Lahti,L., Nedyalkova,M., Elbere,I., Roshchupkin,G., Adilovic,M., Aydemir,O., Bakir-Gungor,B., Pau,S., de,E.C., *et al.* (2021) Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol.*, **12**, 635781.
4. Aitchison,J. (1982) The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. C*, **44**, 139–177.
5. Calle,M.L. (2019) Statistical analysis of metagenomics data. *Genomics Inform*, **17**, e6.
6. Salosensaari,A., Laitinen,V., Havulinna,A.S., Meric,G., Cheng,S., Perola,M., Valsta,L., Alfthan,G., Inouye,M., Watrous,J.D., *et al.* (2021) Taxonomic signatures of cause-specific mortality risk in human gut microbiome. *Nat. Commun.*, **12**, 2671.

7. Wilmanski,T., Diener,C., Rappaport,N., Patwardhan,S., Wiedrick,J., Lapidus,J., Earls,J.C., Zimmer,A., Glusman,G., Robinson,M., *et al.* (2021) Gut microbiome pattern reflects healthy ageing and predicts survival in humans. *Nat. Metab.*, **3**, 274–286.
8. Sims,T.T., El Alam,M.B., Karpinets,T.V., Dorta-Estremera,S., Hegde,V.L., Nookala,S., Yoshida-Court,K., Wu,X., Biegert,G.W.G., Delgado Medrano,A.Y., *et al.* (2021) Gut microbiome diversity is an independent predictor of survival in cervical cancer patients receiving chemoradiation. *Commun. Biol.*, **4**, 237.
9. Debelius,J.W., Engstrand,L., Matussek,A., Brusselaers,N., Morton,J.T., Stenmarker,M. and Olsen,R.S. (2023) The local tumor microbiome is associated with survival in late-stage colorectal cancer patients. *Microbiol. Spectr.*, **11**, e0506622.
10. Kaul,A., Mandal,S., Davidov,O. and Peddada,S.D. (2017) Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.*, **8**, 2114.
11. Lin,H. and Peddada,S.D. (2020) Analysis of compositions of microbiomes with bias correction. *Nat. Commun.*, **11**, 3514.
12. Fernandes,A.D., Macklaim,J.M., Linn,T.G., Reid,G. and Gloor,G.B. (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One*, **8**, e67019.
13. Rivera-Pinto,J., Egozcue,J.J., Pawlowsky-Glahn,V., Paredes,R., Noguera-Julian,M. and Calle,M.L. (2018) Balances: a ew perspective for microbiome analysis. *Msystems*, **3**, e00053-18.
14. Calle,M.L., Pujolassos,M. and Susin,A. (2023) coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinf.*, **24**, 82.
15. Cox,D.R. (1972) Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B (Methodological)*, **34**, 187–220.
16. Salerno,S. and Li,Y. (2023) High-dimensional survival analysis: methods and applications. *Annu. Rev. Stat. Appl.*, **10**, 25–49.
17. Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B (Methodological)*, **58**, 267–288.
18. Cox,D. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.
19. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.
20. Harrell,F.E., Lee,K.L. and Mark,D.B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, **15**, 361–387.
21. Aitchison,J. and Bacon-Shone,J. (1984) Log contrast models for experiments with mixtures. *Biometrika*, **71**, 323–353.
22. Aitchison,J. (1994) Principals of compositional data analysis. *Multivariate Anal. Applic.*, **24**, 73–81.
23. Susin,A., Wang,Y., Cao,K.A.L. and Luz Calle,M. (2020) Variable selection in microbiome compositional data analysis. *NAR Genom. Bioinform.*, **2**, lqaa029.
24. Harrington,D.P. and Fleming,T.R. (1982) A class of rank test procedures for censored survival data. *Biometrika*, **69**, 553–566.
25. Zhang,X.S., Li,J., Krautkramer,K.A., Badri,M., Battaglia,T., Borbet,T.C., Koh,H., Ng,S., Sibley,R.A., Li,Y., *et al.* (2018) Antibiotic-induced acceleration of type 1 diabetes alters maturation of innate intestinal immunity. *eLife*, **7**, e37816.
26. Livanos,A.E., Greiner,T.U., Vangay,P., Pathmasiri,W., Stewart,D., McRitchie,S., Li,H., Chung,J., Sohn,J., Kim,S., *et al.* (2016) Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. *Nat. Microbiol.*, **1**16140.
27. Gu,W., Koh,H., Jang,H., Lee,B. and Kang,B. (2023) MiSurv: an integrative web cloud platform for user-friendly microbiome data analysis with survival responses. *Microbiol. Spectr.*, **11**, e0505922.
28. Grambsch,P.M. and Therneau,T.M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–541.
29. Quinn,T.P., Erb,I., Richardson,M.F. and Crowley,T.M. (2018) Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, **34**, 2870–2878.
30. McGregor,D.E., Palarea-Albaladejo,J., Dall,P.M., Hron,K. and Chastin,S.F.M. (2020) Cox regression survival analysis with compositional covariates: application to modelling mortality risk from 24-h physical activity patterns. *Stat. Methods Med. Res.*, **29**, 1447–1465.
31. Greenacre,M., Grunsky,E. and Bacon-Shone,J. (2021) A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Comput. Geosci.*, **148**, 104621.
32. Sheng,A. and Ghosh,S.K. (2020) Effects of proportional hazard assumption on variable selection methods for censored data. *Stat. Biopharm. Res.*, **12**, 199–209.

Received: November 8, 2023. **Revised:** January 23, 2024. **Editorial Decision:** March 27, 2024. **Accepted:** April 8, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com