Scalable Log-ratio Lasso Regression Enhances Microbiome Feature Selection for Predictive Models

Teng Fei*[1], Tyler Funnell[2], Nicholas R. Waters[2], Sandeep S. Raj[3], Sean M. Devlin[1], Anqi Dai[2], Oriana Miltiadous[4], Roni Shouval[5,6], Meng Lv[7], Jonathan U. Peled[5,6], Doris M. Ponce[5,6], Miguel-Angel Perales[5,6], Mithat Gönen[1], and Marcel R. M. van den Brink†[2,5,6]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center

[2]Department of Immunology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center

[3]Department of Medicine, Memorial Sloan Kettering Cancer Center

[4]Department of Pediatrics, Memorial Sloan Kettering Cancer Center

[5]Adult Bone Marrow Transplantation Service, Department of Medicine, Memorial Sloan Kettering Cancer Center

[6]Department of Medicine, Weill Cornell Medical College

[7]Institute of Hematology, Peking University People's Hospital

*Corresponding author, email: feit1@mskcc.org

†Corresponding author, email: vandenbm@mskcc.org

Abstract

Identifying predictive biomarkers of patient outcomes from high-throughput microbiome data is of high interest in contemporary cancer research. We present `FLORAL`, an open-source computational tool to perform scalable log-ratio lasso regression modeling and microbial feature selection for continuous, binary, time-to-event, and competing risk outcomes. The proposed method adapts the augmented Lagrangian algorithm for a zero-sum constraint optimization problem while enabling a two-stage screening process for extended false-positive control. In extensive simulation studies, `FLORAL` achieved consistently better false-positive control compared to other lasso-based approaches and better variable selection F_1 score over popular differential abundance approaches. We demonstrate the practical utility of the proposed tool with a real data application on an allogeneic hematopoietic-cell transplantation cohort. The R package is available at https://github.com/vdblab/FLORAL.

1 Introduction

Advances in computational approaches for analyzing metagenomic data have substantially improved our understanding of the relationships between the human microbiota and environmental exposures, health conditions, treatment responses, and patient survival. Discovery of microbiome compositions predictive of human disease or treatment outcomes provide opportunities for therapeutic intervention. For example, large-scale collection efforts have reported thousands of associations between gut microbes and human health conditions using samples from thousands of individuals (Gacesa et al., 2022). At the same time, investigators have been increasingly transitioning from high-specificity, low-depth 16S ribosomal RNA sequencing to high-sensitivity, high-depth shotgun sequencing platforms, improving both taxonomic resolution and offering functional insights in a sequencing experiment. The aggressively evolving technology and quickly accumulating amount of available microbiome data over the past decade have motivated computational biologists and biostatisticians to develop robust analytical approaches to detect associations between microbiota and factors of interest, while avoiding false-positives (Mallick et al., 2021).

In the field of clinical cancer research, allogeneic hematopoietic cell transplants (allo-

HCT) provides a paradigm for understanding the importance of microbiome composition on clinical outcomes. While provided to patients with curative intent, high dose chemotherapy prior to the allograft stem cell infusion causes severe damage to gut microbiota, which further results in a prolonged period of immune system recovery, during which patients are at risk for life-threatening gut inflammation, opportunistic infections, and malnutrition. Therefore, it is of high interest to monitor and study the association between the microbial profiles of allo-HCT patients and the corresponding patient outcomes, such as overall survival, progression-free survival, and graft-vs-host disease (GVHD), which are commonly coded as binary, time-to-event, or competing risks outcomes (Copelan et al., 2007). Recent studies of patients undergoing allo-HCT have found that gut domination by *Enterococcus* bacteria significantly increases the risk of acute GVHD or death (Stein-Thoeringer et al., 2019; Peled et al., 2020). In a subsequent study, allo-HCT patients with GVHD were found to have lower abundance of *Clostridia* bacteria than those without GVHD (Burgos da Silva et al., 2022).

To fulfill the research questions in mircobiome-related studies as discussed above, one common objective is to investigate whether the microbial profile differs across two patient populations, such as GVHD versus non-GVHD groups in allo-HCT cohort studies, treatment versus control groups in clinical trials, and exposure versus non-exposure groups in epidemiological studies. Summary metrics such as α-diversity (Shannon, 1948; Simpson, 1949), β-diversity (Bray and Curtis, 1957), and global tests (Anderson, 2001; Zhao et al., 2015) are usually applied to provide a community-level comparison across groups. For the purpose of understanding the biological mechanism of how microbiota interacts within the human microenvironment, taxonomic-level comparisons for individual taxon measured from microbiome sequencing, such as genus or species, are also of high interest. The most widely applied strategy to investigate taxa-population association is to conduct differential abundance testing, which is a similar concept to the differential gene expression analysis for studying gene-population associations.

Researchers have blazed a unique path of developing statistical methods for differential taxa abundance analysis due to the special data characteristics of microbiome data. Microbiome sequencing data is compositional, meaning that the achieved sequencing depth and resulting profiles are relative due to the nature of the sequencing technology (Gloor et al., 2017). As the sequencing depth may heavily vary across samples, one should

account for both observed count and sequencing depth (i.e. total count) to facilitate a standardized quantification of a provided taxon across microbial samples. One naive approach is to perform the Wilcoxon rank-sum test for the observed relative taxon abundance (count divided by sequencing depth) across two groups. Using similar rank-based taxa filtering criteria while leveraging patient subgroup information, Segata et al. (2011) presented the linear discriminant analysis effect size (LEfSe) method with high popularity. To account for the high presence of zero counts, zero-inflated Gaussian regression models (Paulson et al., 2013, `metagenomeSeq`) and procedures of distinguishing structural zeros from experimental zeros were proposed (Lin and Peddada, 2020, `ANCOM-II`, `ANCOM-BC`). A Dirichlet multinomial model was also presented to address differential abundance testing across more than two groups (Fernandes et al., 2013, `ALDEx2`). More recently, a beta-binomial model `corncob` was proposed to simultaneously capture differential taxon abundance and differential taxon variability to address overdispersion in microbial counts (Martin et al., 2020). To combine the utility of global testing and taxa-level comparison, a robust linear decomposition model `LDM` was proposed and has been extended for paired microbiome data and incorporating survival outcomes in a series of development (Hu and Satten, 2020; Zhu et al., 2021; Hu et al., 2022).

While the aforementioned differential abundance (DA) testing methods gained high popularity in taxonomic-level data analysis, there are some important limitations. Typically, the DA methods perform multiple tests for individual taxa followed by p-value adjustment, where taxa selection is determined by the choice of a p-value threshold. In practice, there is no clear guidance on a widely accepted threshold, such that differentially abundant taxa are inconsistently defined for different studies. For example, Derosa et al. (2022) and Lee et al. (2022) adapted the criterion of FDR adjusted p-value < 0.2, while Wallen et al. (2022) applied multiple cutoffs of 0.05 and 0.1 when reporting differentially abundant taxa based on `ANCOM-BC` results. Such arbitrary choices of p-value threshold is not ideal for the reproducibility of microbiome research. Moreover, the majority of DA methods were developed for two-sample comparison of microbial abundances, which makes it challenging to incorporate survival outcomes, such as the patient overall survival or competing risks that are important endpoints in allo-HCT studies. To apply the DA methods for survival outcomes, one common practice is to perform tests between the event group and the non-event group (Worsley et al., 2021, for example), which over-

looks time-to-event information and causes severe information loss. While a remedy has been proposed by Hu et al. (2022) by utilizing the Martingale residual as a surrogate of time-to-event information, the method still regards the microbial features as the outcome variable, such that the features cannot be effectively incorporated in an at-risk process as in the partial likelihood of the standard Cox proportional hazards model (Cox, 1972).

As an alternative approach to modeling the association between taxa and patient characteristics, penalized regression models have also been under development in the past decade, though having gained lower attentions than the DA methods. Unlike the DA methods, penalized regression models treat patient characteristics or outcomes as outcome variables and taxa as predictors. Adapting the ideas from classic linear regression models for compositional data (Aitchison, 1982), Lin et al. (2014) proposed a log-ratio ℓ_1 regularization linear regression model, or log-ratio lasso, with a fixed reference feature for calculating the ratios. One important result from Lin et al. (2014) is that the estimator can be reformulated as a linear lasso problem with a zero-sum constraint on regression coefficients. To tackle the same zero-sum constraint optimization problem, Altenbuchinger et al. (2017) developed R package `zeroSum` with extensions to generalized linear models and Cox models. As an alternative way of interpreting the model, Bates and Tibshirani (2019) presented the covariate space of $\binom{p}{2}$ unique pairwise log-ratios out of p taxa, and showed that the optimization problem is equivalent to what was presented in Lin et al. (2014). The corresponding algorithm was implemented in R package `logratiolasso` which utilized `glmnet` (Hastie et al., 2015) and a weighting scheme to account for the zero-sum constraint. Moreover, Bates and Tibshirani (2019) proposed a two-step filtering procedure which performs pairwise log-ratio modeling for features selected by the constraint lasso problem, such that specific ratios of taxa pairs can be identified. Recently, (Calle et al., 2023) implemented the idea of creating the $\binom{p}{2}$ pairwise log-ratio covariate space via `glmnet` in R package `coda4microbiome`. Compared to the DA methods, log-ratio lasso methods conduct variable selection using a more objective criterion of cross-validated prediction performance. In addition, survival information can be naturally incorporated into the lasso framework as a regularized Cox model (Simon et al., 2011), which makes it more appealing to utilize the log-ratio lasso methods to tackle the survival analysis with microbiome data.

Although suitable and promising for microbiome data analysis, the log-ratio lasso

method is not widely applied in literature due to limited software development and lack of systematic comparison with the DA methods. For instance, Lin et al. (2014)'s initial work was well developed both theoretically and algorithmically, but no easy access is available to its augmented Lagrangian algorithm. Bates and Tibshirani (2019)'s implementation enables log-ratio lasso linear and logistic regression, but it did not solve Lin et al. (2014)'s optimization problem in a generalizable infrastructure. It is unknown how to use a similar data manipulation technique as used by `logratiolasso` to handle more complex data, such as time-to-event data. While Altenbuchinger et al. (2017)'s `zeroSum` package is a very effective tool for the lasso problem with zero-sum constraint, the lack of two-step filtering may result in unwanted high false positive hits and inconvenience in identifying specific taxa ratios for model interpretation. The recent work from Calle et al. (2023) attempted to work directly on the $\binom{p}{2}$-dimensional covariate space; however, contemporary microbiome sequencing data may contain thousands of taxonomic features, making creation of the $\binom{p}{2}$-dimension space infeasible with limited computing memory.

Motivated by the practical needs of microbiome data analysis in clinical cancer research and the insufficient development of statistical tools, we present `FLORAL`, an R package written with `RcppArmadillo` (Eddelbuettel and Sanderson, 2014) which supports linear, logistic, Cox proportional hazards, and Fine-Gray proportional subdistributional hazards log-ratio lasso regression for high-dimensional compositional data. We implemented `FLORAL` using a generalized augmented Lagrangian algorithm (Lin et al., 2014), such that it can easily adapt continuous, binary, survival and other patient outcomes. Instead of enumerating $\binom{p}{2}$ possible pairs of taxa, `FLORAL` works on the p-dimensional covariate space as facilitated by the zero-sum constraint, which only requires affordable computing memory. Furthermore, `FLORAL` is featured with built-in two-stage variable selection based on Bates and Tibshirani (2019)'s pairwise log-ratio framework to further reduce potential false positive findings and enhance model interpretability. We conducted extensive simulation studies to assess and compare our method's performance with other lasso methods, as well as popular DA methods. To better understand the utility of the log-ratio lasso approaches, detailed comparisons were performed under various settings of sample size, feature size, effect size, taxa sparsity level and taxa correlation structures. We demonstrate that `FLORAL` achieves reasonable sensitivity and highly reliable specificity. Using 16S rRNA gene sequencing data from a cohort of allogeneic hematopoietic bone

marrow transplant patients from Memorial Sloan Kettering Cancer Center (MSKCC) (Peled et al., 2020), we successfully identified *Enterococcus* as a predictive feature of patient overall survival, which is a known mediator of peri-transplant mortality through antibiotic-resistance blood stream infections and is associated with dysbiosis.

2 Methods

2.1 Log-ratio Regression Models

Let $\boldsymbol{X}_{n\times p}$ denote the absolute count matrix of microbiome data with p taxa and n patients. For a scalar outcome such as continuous or binary outcome, we denote the corresponding response vector as $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}^T$. For survival outcome, we denote $\tilde{\boldsymbol{T}} = \{\tilde{T}_1, \ldots, \tilde{T}_n\}^T$ and $\boldsymbol{\Delta} = \{\Delta_1, \ldots, \Delta_n\}^T$ as observed survival time subject to right censoring and the censoring indicator, respectively. We denote the observed version of the above quantities from the ith patient as $\boldsymbol{X}_i, \boldsymbol{Y}_i, \tilde{\boldsymbol{T}}_i, \boldsymbol{\Delta}_i$, respectively. For a scalar outcome $\boldsymbol{Y}$, we model the association between $\boldsymbol{Y}$ and $\boldsymbol{X}$ via a log-ratio regression model

$$g\{E(Y_i|\boldsymbol{X}_i)\} = \beta_0 + \sum_{1\le j<k\le p} \theta_{j,k} \log\left(\frac{X_{i,j}}{X_{i,k}}\right), \tag{1}$$

where $g(\cdot)$ is a link function accounting for the data characteristic of $\boldsymbol{Y}$, β_0 is an unknown intercept term, and $\theta_{j,k}$ are unknown coefficients corresponding to the paired log-ratio covariates $\log(X_{i,j}/X_{i,k})$. Here, we adapt the notion of pairwise log-ratio as described in Bates and Tibshirani (2019), where there are $\binom{p}{2}$ unknown $\theta_{j,k}$ for $1 \le j < k \le p$. For survival outcome $(\tilde{\boldsymbol{T}}, \boldsymbol{\Delta})$, we consider a log-ratio proportional hazards model

$$h(t|\boldsymbol{X}_i) = h_0(t) \exp\left\{\sum_{1\le j<k\le p} \theta_{j,k} \log\left(\frac{X_{i,j}}{X_{i,k}}\right)\right\}, \tag{2}$$

where $h(t|\boldsymbol{X}_i)$ denotes the hazard function conditioned on microbial features $\boldsymbol{X}_i$, $h_0(t)$ is the baseline hazard function, and $\theta_{j,k}$ are unknown regression coefficients associated with $\log(X_{i,j}/X_{i,k})$. Note that model (2) can be naturally extended for longitudinal microbiome data $\boldsymbol{X}(t)$, such that

$$h(t|\boldsymbol{X}_i) = h_0(t) \exp\left\{\sum_{1\le j<k\le p} \theta_{j,k} \log\left(\frac{X_{i,j}(t)}{X_{i,k}(t)}\right)\right\}.$$

In practice, longitudinal stool samples are only available at a finite number of time points, where the last value carried forward (LVCF) strategy is widely applied (Tsiatis and

Davidian, 2004). Moreover, as pointed out by Therneau et al. (2020), the Fine-Gray subdistributional proportional hazards model (Fine and Gray, 1999) can be equivalently estimated by a weighted Cox model, which offers a convenient pathway of implementing competing risks modeling under the same framework.

Both models (1) and (2) can be simplified as a more concise form as demonstrated in Bates and Tibshirani (2019) and Calle et al. (2023) by rewriting the log of ratios as differences of log-counts. Let $\beta_k = \sum_{j=1}^{k-1} -\theta_{j,k} + \sum_{j=k+1}^{p} \theta_{k,j}$ and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$, one can show by algebra that

$$\sum_{1\le j<k\le p} \theta_{j,k} \log\left(\frac{X_{i,j}}{X_{i,k}}\right) = \sum_{k=1}^{p}\left\{\sum_{j=1}^{k-1} -\theta_{j,k} + \sum_{j=k+1}^{p} \theta_{k,j}\right\} \log X_{i,k} = \sum_{k=1}^{p} \beta_k \log X_{i,k}$$

and

$$\sum_{k=1}^{p} \beta_k = \sum_{k=1}^{p}\left\{\sum_{j=1}^{k-1} -\theta_{j,k} + \sum_{j=k+1}^{p} \theta_{k,j}\right\} = \sum_{j=1}^{p}\sum_{k=j+1}^{p} -\theta_{j,k} + \sum_{k=1}^{p}\sum_{j=k+1}^{p} \theta_{k,j} = 0.$$

Therefore, models (1) and (2) can be rewritten as

$$g\{E(Y_i|\boldsymbol{X}_i)\} = \beta_0 + \sum_{k=1}^{p} \beta_k \log X_{i,k}, \text{ subject to } \sum_{k=1}^{p} \beta_k = 0 \tag{3}$$

and

$$h(t|\boldsymbol{X}_i) = h_0(t) \exp\left\{\sum_{k=1}^{p} \beta_k \log X_{i,k}\right\}, \text{ subject to } \sum_{k=1}^{p} \beta_k = 0, \tag{4}$$

correspondingly. In modern microbiome studies, the number of taxa p can reach the scale of thousands. Compared to models (1) and (2) which impose $\binom{p}{2}$ log-ratio features, models (3) and (4) show appealing computational benefits of having a much lower dimensional covariate space as p increases. To address commonly encountered zero counts in microbiome data, we suggest using $\{\log(\boldsymbol{X}+1)\}$ to replace $\{\log(\boldsymbol{X})\}$ as an approximate covariate space which keeps zero counts as zeros after log transformation.

2.2 The Log-ratio Lasso Estimator and the Augmented Lagrangian Algorithm

Let $\mathcal{L}(\boldsymbol{\beta})$ denote the log-likelihood of model (3) or the log-partial likelihood of model (4). We define the log-ratio lasso estimator as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n}\mathcal{L}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \text{ subject to } \sum_{k=1}^{p} \beta_k = 0, \tag{5}$$

where λ is a regularization parameter and $\|\cdot\|_1$ denotes ℓ_1 norm. We adapt the similar treatment in `glmnet` (Friedman et al., 2010) to approximate $\mathcal{L}(\boldsymbol{\beta})$ by its second-order Tyler expansion centered at $\tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is either an initial value or the estimate from a previous iteration. Define $\boldsymbol{Z} = \log(\boldsymbol{X})$, $\tilde{\boldsymbol{\eta}} = \boldsymbol{Z}\tilde{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) = \tilde{\boldsymbol{\eta}} - \{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}^{-1}\dot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})$, where $\dot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})$ and $\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})$ denote the gradient and Hessian matrix of $\mathcal{L}(\tilde{\boldsymbol{\eta}})$, respectively. Then we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) \approx& \mathcal{L}(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}(\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \frac{\partial^2}{\partial \boldsymbol{\beta}^2} \mathcal{L}(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \\
=& \mathcal{L}(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{Z}\boldsymbol{\beta} - \tilde{\boldsymbol{\eta}})^T \dot{\mathcal{L}}(\tilde{\boldsymbol{\eta}}) + \frac{1}{2}(\boldsymbol{Z}\boldsymbol{\beta} - \tilde{\boldsymbol{\eta}})^T \ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})(\boldsymbol{Z}\boldsymbol{\beta} - \tilde{\boldsymbol{\eta}}) \\
=& \frac{1}{2}\{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{Z}\boldsymbol{\beta}\}^T \ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{Z}\boldsymbol{\beta}\} + C(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\beta}}),
\end{aligned}$$

where the first term in the formula on the last row is a weighted quadratic form of $\boldsymbol{Z}\boldsymbol{\beta}$ and the second term $C(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\beta}})$ is independent of $\boldsymbol{\beta}$. To alleviate computational burdens for the $n \times n$ matrix $\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})$, we follow Hastie et al. (2015) and Simon et al. (2011) to substitute $\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})$ by its diagonal elements $\mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}$. That is, the working loss function is defined as

$$\tilde{\mathcal{L}}(\boldsymbol{\beta}) = \{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{Z}\boldsymbol{\beta}\}^T \mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}\{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{Z}\boldsymbol{\beta}\}, \tag{6}$$

which is a standard weighted least squares form with continuous response $\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}})$ and predictors $\boldsymbol{Z}$. It is also straightforward to show that $\tilde{\mathcal{L}}(\boldsymbol{\beta})$ is equivalent to the standard least squares form $\tilde{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2$ when $\boldsymbol{Y}$ is continuous. This also forms the working optimization problem for the proposed lasso estimator:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{Z}\boldsymbol{\beta}\}^T \mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}\{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \boldsymbol{Z}\boldsymbol{\beta}\} + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \text{ subject to } \sum_{k=1}^{p} \beta_k = 0. \tag{7}$$

With a unified formula of working loss function (6) for different choices of $\mathcal{L}(\boldsymbol{\beta})$, the corresponding lasso optimization problem with constraint (7) can be conveniently defined for either scalar or survival outcomes if the first- and second-order differentiation with respect to $\tilde{\boldsymbol{\eta}}$ are well defined for the (partial) log-likelihood function $\mathcal{L}(\boldsymbol{\beta})$.

We adapted the augmented Lagrangian approach (Bertsekas, 2014) used by Lin et al. (2014) to solve the constrained optimization problem (7). Specifically, the constraint

$\sum_{k=1}^{p} \beta_k = 0$ is incorporated in the following target function

$$\begin{aligned}\tilde{L}_\mu(\boldsymbol{\beta}, \gamma) &= \frac{1}{n}\tilde{\mathcal{L}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 + \gamma\sum_{k=1}^{p}\beta_k + \frac{\mu}{2}\left(\sum_{k=1}^{p}\beta_k\right)^2 \\ &= \frac{1}{n}\tilde{\mathcal{L}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\mu}{2}\left(\sum_{k=1}^{p}\beta_k + \alpha\right)^2,\end{aligned} \tag{8}$$

where γ is the Lagrange multiplier, term $\mu(\sum_{k=1}^{p}\beta_k)^2/2$ is deployed to avoid the situation where $\sum_{k=1}^{p}\beta_k$ goes to negative infinity, and $\alpha = \gamma/\mu$ enables merging the Lagrange multiplier $\gamma\sum_{k=1}^{p}\beta_k$ and the penalty term $\mu(\sum_{k=1}^{p}\beta_k)^2/2$ into a single term.

Given λ, μ, and an initial value $\hat{\boldsymbol{\beta}}^{(0)} = \tilde{\boldsymbol{\beta}}$ which can be obtained by a warm start, estimation of $\hat{\boldsymbol{\beta}}$ is conducted by a coordinate gradient descent algorithm with iteratively updated value of α as proposed in Lin et al. (2014), where the initial value of α at the first iteration, $\alpha^{(0)}$, is zero. In the ith iteration, the corresponding estimate $\hat{\boldsymbol{\beta}}^{(i)}$ is updated by minimizing $\tilde{L}_\mu(\boldsymbol{\beta}, \gamma)$ with fixed values of $\lambda, \mu, \alpha^{(i)}$ and an initial value $\tilde{\boldsymbol{\beta}}$ from the previous iteration. This step of updating $\hat{\boldsymbol{\beta}}^{(i)}$ can be performed by an inner loop of standard coordinate descent algorithm where the kth component of $\hat{\boldsymbol{\beta}}^{(i)}$ is updated by

$$\hat{\beta}_k^{(i)} = \frac{1}{\boldsymbol{Z}_k^T\boldsymbol{Z}_k/n + \mu} S_\lambda\left\{\frac{1}{n}\boldsymbol{Z}_k^T \mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}\{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \sum_{l\neq k}\hat{\beta}_l^{(i)}\boldsymbol{Z}_l\} - \mu\left(\sum_{l\neq k}\hat{\beta}_l^{(i)} + \alpha^{(i)}\right)\right\},$$

where $\boldsymbol{Z}_k$ denotes the kth column of matrix $\boldsymbol{Z}$ and $S_\lambda(x) = \mathrm{sgn}(|x| - \lambda)_+$ is the soft thresholding operator. After each inner-loop coordinate descent from feature 1 to feature p, we update $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\eta}}$, $\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}})$, and $\mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}$ for the next inner-loop coordinate descent. The updates of $\hat{\boldsymbol{\beta}}^{(i)}$ stops if the loss function $\tilde{L}_\mu(\hat{\boldsymbol{\beta}}^{(i)}, \gamma)$ is converged at a tolerance parameter δ'. With updated $\hat{\boldsymbol{\beta}}^{(i)}$ from the inner loop, the penalty parameter α is updated as

$$\alpha^{(i+1)} = \alpha^{(i)} + \sum_{k=1}^{p}\beta_k^{(i)},$$

such that a larger penalty will be imposed in the $(i+1)$th iteration for $\sum_{k=1}^{p}\beta_k^{(i+1)}$ if $\sum_{k=1}^{p}\beta_k^{(i)}$ deviates from zero in the ith iteration. The algorithm stops if $\|\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(i-1)}\|_1 < \delta$ for a pre-specified tolerance parameter δ. Detailed implementation of the algorithm is reported as Algorithm 1. In actual implementation, we calculate $p \times p$ matrix $\boldsymbol{A}$ and p-vector $\boldsymbol{B}$, as defined in Algorithm 1, prior to the coordinate descent loop to save computational cost. We also specify the maximum iteration number u' for the inner loop and u for the outer loop to bring an early stop if the convergence is not reached. In our analysis, we constantly use $\mu = 1, \delta = \delta' = 10^{-7}$ and $u = u' = 100$, where the choices of $\mu = 1$ is suggested by Lin et al. (2014).

Algorithm 1 Iterative optimization algorithm for (8) with given λ and μ.

Input: Initial value of $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$; $n \times p$ matrix $\boldsymbol{Z}$; parameters λ, μ; tolerance parameter δ, δ'; maximum inner iteration number u, u'

Set $\hat{\boldsymbol{\beta}}^{(0)} = \tilde{\boldsymbol{\beta}}$, $\alpha^{(1)} = 0$, $i = 0$, $d_{\boldsymbol{\beta}} = 1$

while $d_{\boldsymbol{\beta}} > \delta$ and $i \leq u$ **do**

  Set $i = i + 1$

  Set $\tilde{\boldsymbol{\eta}} = \boldsymbol{Z}\tilde{\boldsymbol{\beta}}$, $d = 1$, $j = 0$

  while $d > \delta'$ and $j \leq u'$ **do**

    Set $j = j + 1$, idx = which($\tilde{\boldsymbol{\beta}} > 0$). Initialize $\check{\boldsymbol{\beta}}$

    Compute $\dot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})$, diag$\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}$, $\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}})$

    Set $\boldsymbol{A}_{p\times p} = \boldsymbol{Z}^T \text{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}\boldsymbol{Z}$, $\boldsymbol{B}_{p\times 1} = \boldsymbol{Z}^T[\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) \odot \text{vec}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}]$

    for k = i,. . . ,p **do**

      Update $\check{\beta}_k = \frac{1}{\boldsymbol{A}_{k,k}+\mu} S_\lambda \left\{ \frac{1}{n}(\boldsymbol{B}_k - A_{k,\text{idx}} \odot \tilde{\boldsymbol{\beta}}_{\text{idx}} + A_{k,k} \odot \tilde{\beta}_k) - \mu(\sum_{l\neq k, l\in \text{idx}} \tilde{\boldsymbol{\beta}}_l + \alpha^{(i)}) \right\}$

    end for

    Set $d = |\tilde{L}_\mu(\check{\boldsymbol{\beta}}, \gamma) - \tilde{L}_\mu(\tilde{\boldsymbol{\beta}}, \gamma)|$, $\tilde{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\eta}} = \boldsymbol{Z}\tilde{\boldsymbol{\beta}}$

  end while

  Set $\hat{\boldsymbol{\beta}}^{(i)} = \check{\boldsymbol{\beta}}$, $\alpha^{(i+1)} = \alpha^{(i)} + \sum_{k=1}^{p} \hat{\beta}_k^{(i)}$, $d_{\boldsymbol{\beta}} = \|\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}}^{(i-1)}\|_1$

end while

Output: $\hat{\boldsymbol{\beta}}^{(i)}$

2.3 Pathwise Solution and Cross Validation

To have a global picture on how model sparsity is governed by different choices of λ, we solve the optimization problem (7) by Algorithm 1 on a decreasing path $\boldsymbol{\lambda}$ of λ. By default, the path $\boldsymbol{\lambda}$ starts with

$$\lambda_{(1)} = \max_k \frac{1}{n} |\boldsymbol{Z}_k^T \text{diag}\{\ddot{\mathcal{L}}(\boldsymbol{0})\}\tilde{\boldsymbol{Z}}(\boldsymbol{0})|$$

which acquires $\hat{\boldsymbol{\beta}} = \boldsymbol{0}$ (Simon et al., 2011). Then a sequence of length m, $\lambda_{(1)}, \ldots, \lambda_{(m)}$ is generated with equal distance on log scale, where $\lambda_{(m)}$ is typically selected as $0.01\lambda_{(1)}$ if $n < p$ and $0.0001\lambda_{(1)}$ if $n \geq p$.

Cross validation is used to determine the optimal choice of λ which maximizes the cross-validated predictive performance. Standard criteria, such as mean-squared error, deviance, or concordance index, are used to evaluate prediction errors. Two choices of λ are reported, namely λ_{min} which minimizes cross-validated prediction error, and λ_{1se}

which provides a sparser solution than $\lambda_{\min}$ but still obtains cross-validated prediction error within one standard error of that of $\lambda_{\min}$. Taxa with non-zero coefficient estimates $\hat{\boldsymbol{\beta}}$ at $\lambda_{\min}$ or $\lambda_{1\text{se}}$ are selected and interpreted as predictive for the clinical outcomes.

Algorithm 1 with pathwise solution and cross validation utility is implemented by `RcppArmadillo` (Eddelbuettel and Sanderson, 2014) which achieves reasonable computing speed for large p in the scale of thousands of taxa.

2.4 Two-step Model Selection

In the previous sections we derive the algorithm to efficiently find a sparse set of predictive taxa. However, the log-ratio lasso algorithm cannot identify specific pairs of log-ratios for further interpretation of the model. One natural extension is to perform exhaustive search on all possible pairs of log-ratios for the selected features from the log-ratio lasso algorithm (Bates and Tibshirani, 2019). Since the selected taxa set is much sparser than the full set of p taxa, the corresponding number of pairwise combinations is also much fewer than $\binom{p}{2}$, which enables faster computation via standard lasso models and stepwise regression models. In our implementation, we first perform a standard lasso regression via `glmnet` over the enumeration of log-ratios from the selected feature set to filter out log-ratios not contributing to a better prediction. Then we apply a stepwise regression model for the selected log-ratios to further exclude log-ratios that do not substantially improve model fitting. The two-stage feature selection aims to keep the strongest signals in the model while obtaining meaningful interpretations for specific ratios of microbes.

2.5 Simulation Studies

We performed extensive simulation studies to assess various methods' behaviour under various scenarios. Let n be the sample size and p be the feature size. For each simulated sample i, $i = 1, \ldots, n$, we first simulate the underlying taxa composition $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{ip})^T$, where $c_{ik} \leq 0$ for all k and $\sum_{k=1}^{p} c_{ik} = 1$. To get $\boldsymbol{c}_i$, we simulate a p-vector $\boldsymbol{x}_i$ which follows a p-variate normal distribution $N_p(\boldsymbol{\xi}, \boldsymbol{\Sigma})$, where $\boldsymbol{\xi}_k = \log p$ for $k = 1, 2, 3, 5, 6, 8$ and otherwise $\boldsymbol{\xi}_k = 0$. This choice of $\boldsymbol{\xi}$ makes features 1, 2, 3, 5, 6, 8 more abundant than others. Correspondingly, the variance parameters $\sigma_k^2 = \boldsymbol{\Sigma}_{k,k}$ satisfies $\sigma_k^2 = \sqrt{\log p/2}$ for $k = 1, 2, 3, 5, 6, 8$ and otherwise $\sigma_k^2 = 1$, which makes highly abundant features of higher variation. We let $\Sigma_{j,k} = \rho^{|j-k|}$, $\rho \in (0, 1)$ be the correlation

between features j and k, where features of adjacent indices are more correlated than features of distant indices. To generate sparsity in counts, we then specify a sparsity level $s \in (0,1)$ and randomly force $s \times p$ many elements in $\boldsymbol{x}_i$ to be $-\infty$. Then we calculate $c_{ik} = \exp(\boldsymbol{x}_{ik}) / \sum_d \exp(\boldsymbol{x}_{id})$ for all k to obtain $\boldsymbol{c}_i$.

With $\boldsymbol{c}_i$ generated, we further generate the outcome variables. The continuous response variable Y_i^c is generated by

$$Y_i^c = 0.25u\left\{ \log \frac{c_{i1}}{c_{i2}} + \log \frac{c_{i3}}{c_{i4}} \right\} + 0.5u\left\{ \log \frac{c_{i5}}{c_{i6}} + \log \frac{c_{i7}}{c_{i8}} + \log \frac{c_{i9}}{c_{i10}} \right\} + \epsilon_i.$$

Here u controls the effect sizes, where the first two ratios have half of the effect sizes of the latter three ratios. The error term ϵ_i follows independent standard normal distribution for each i. Similarly, the binary outcome variable Y_i^b is simulated from a $Bernoulli(q_i)$ distribution, where

$$q_i = \text{expit}\left[0.25u\left\{ \log \frac{c_{i1}}{c_{i2}} + \log \frac{c_{i3}}{c_{i4}} \right\} + 0.5u\left\{ \log \frac{c_{i5}}{c_{i6}} + \log \frac{c_{i7}}{c_{i8}} + \log \frac{c_{i9}}{c_{i10}} \right\}\right].$$

Here, $\text{expit}(x) = 1/(1+e^{-x})$. For survival outcomes, the event time T_i is simulated from the distribution function $F_{T_i}(t) = 1 - \exp\{0.1(1-e^t)\exp(h_i)\}$, where

$$h_i = 0.25u\left\{ \log \frac{c_{i1}}{c_{i2}} + \log \frac{c_{i3}}{c_{i4}} \right\} + 0.5u\left\{ \log \frac{c_{i5}}{c_{i6}} + \log \frac{c_{i7}}{c_{i8}} + \log \frac{c_{i9}}{c_{i10}} \right\}.$$

It can be shown that the associated hazard function is equal to $0.1e^t \exp(h_i)$ which belongs to the family of proportional hazards models. Then a random censoring time V_i is generated as the minimum of an $Exponential(0.1)$ and a $Uniform(5,6)$ distribution. Then the observable survival time $\tilde{T}_i = \min(T_i, V_i)$ and event indicator $\Delta_i = I(T_i < V_i)$ are obtained.

To simulate the actually observable count data $\boldsymbol{X}_i$, we first randomly generate the sequencing depth D_i as the largest integer smaller than a random variable following a $Unif(5000, 50000)$ distribution, where 5000 to 50000 is a reasonable range for high-quality microbiome 16S rRNA sequencing depths. Then the count data $\boldsymbol{X}_i$ is generated from a multinomial distribution with D_i instances and the probability vector $\boldsymbol{c}_i$. The simulated outcome ($\boldsymbol{Y}^c, \boldsymbol{Y}^b$ or $(\tilde{\boldsymbol{T}}, \boldsymbol{\Delta})$) and count matrix $\boldsymbol{X}$ were the input for lasso or DA methods.

For each of the three types of response variables, we investigated the performance of methods based on a reference scenario where $n = 200$, $p = 500$, $s = 0.8$, $\rho = 0$, and $u = 1$. Controlling other parameters as fixed, we compared $n = 50, 100, 200, 500$, $p = 100, 200, 500, 1000$, $s = 0.8, 0.95$, $\rho = 0, 0.5$, and $u = 0.1, 0.5, 1, 2$. This serves as a comprehensive survey in understanding the behavior of methods under various settings.

2.6 Real Data Applications

Microbiome sequencing data with patient outcomes were utilized to assess the empirical performance of a variety of microbial feature selection approaches. We focused on an allo-HCT cohort from MSKCC (Peled et al., 2020), where patient stool samples were collected pre- and post- transplantation for 16S microbiome sequencing. As reported in Peled et al. (2020), a higher gut microbiota α-diversity (Simpson's reciprocal index) between day 7 and day 21 after transplantation was significantly associated with better overall survival.

In our analyses, we defined day 0 as the date of transplantaion (i.e. cell infusion). To keep the time window of stool samples consistent with the previous study, we selected the last available microbiome sequencing absolute count data at genus level between day 7 and day 21 for each patient as the feature matrix $\boldsymbol{X}$. Only samples with sequencing depths larger than 5000 amplicon sequence variant (ASV) counts were included for further investigation. In addition, singleton ASVs in each sample were excluded. After filtering, microbiome data from 663 samples were selected for further analyses, where the sequencing depth ranges between 5097 and 98787 with a median of 19526. The patient overall survival data $(\tilde{\boldsymbol{T}}, \boldsymbol{\Delta})$ was then landmarked at the corresponding stool sample collection time and censored at 24 months after transplantation. To filter out rare taxa, we only kept genera with larger than 100 counts in at least 10 samples, such that 82 genera were included in the model.

2.7 Applied Methods and Assessment

We tested the proposed method and a number of existing approaches in simulation studies and real data applications. First, we applied a family of related lasso estimators, including the proposed log-ratio lasso estimator `FLORAL`; standard lasso estimators using log-counts, relative abundances, or centered log-ratio (CLR) transformed counts; the zero-sum constraint lasso method `zeroSum` (Altenbuchinger et al., 2017); as well as the $\binom{p}{2}$ log-ratio lasso estimator `coda4microbiome` (Calle et al., 2023). For each simulated dataset, we controlled the cross-validation folds to be identical for all lasso methods. The optimal lasso penalty parameters $\lambda_{\min}$ and $\lambda_{1\text{se}}$ were determined by a 5-fold cross-validated prediction error.

In addition to the lasso approaches, popular DA methods were applied for feature selection, including `ALDEx2` (Fernandes et al., 2013), `ANCOM-II` and `ANCOM-BC` (Lin and Peddada, 2020), `LDM` (Hu and Satten, 2020), `LEfSe` (Segata et al., 2011), linear models (LM) for taxa abundance, and the two-sample Wilcoxon rank-sum test. Scalar outcomes $\boldsymbol{Y}^c$ or $\boldsymbol{Y}^b$ were treated as covariates for the DA methods. For survival outcome $(\tilde{\boldsymbol{T}}, \boldsymbol{\Delta})$, we used the Martingale residual as the covariate for `LDM`, while the censoring indicator $\boldsymbol{\Delta}$ was used as the patient group indicator for other methods. We did not test `ALDEx2`, `LEfSe`, nor the Wilcoxon test for continuous outcomes, nor `coda4microbiome` for survival outcomes due to lack of compatibility. We also did not apply `ANCOM-II`, `ANCOM-BC`, nor `coda4microbiome` for scenarios with $p = 1000$ due to slow computation speed. FDR adjusted p-value cutoff 0.05 was applied to define differentially abundant taxa.

In simulation studies, we evaluated predictive performance, variable selection ability, and estimation biases for the lasso-based methods. We also assessed variable selection for the DA methods. To evaluate variable selection performance, we summarized the number of false negatives (FN), false positives (FP), and calculated the F_1 score $F_1 = 2\text{TP}/(2\text{TP}+\text{FP}+\text{FN})$, where TP represents the number of true positives. For lasso methods, we utilized independently generated datasets to assess prediction performances by mean-squared error (continuous outcome), area under the ROC curve (binary outcome), and concordance index (survival outcome) for the lasso estimators at $\lambda_{\min}$ and $\lambda_{1\text{se}}$ fitted from the training data. In addition, $L_q, q = 1, 2$ losses of the estimator, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q$, are calculated to examine how biased is the proposed estimator compared to the underlying true parameter values.

For real applications, we mainly focused on assessing the variable selection stability for lasso-based methods. Specifically, we ran lasso-based methods under randomly generated 5 cross-validation folds for 100 repeats, and then compared the selection frequencies of certain taxa by different lasso-based methods. Taxa selected from the lasso-based methods and the DA methods were interpreted using the domain knowledge of HCT. For the specific ratios selected by the two-stage process of `FLORAL`, we further constructed a multivariable Cox regression model to validate the association between the identified log-ratios and overall patient survival adjusted for relevant patient characteristics.

3 Results

3.1 FLORAL Achieved Superior Variable Selection Performance Among Lasso-based Methods

Our simulations demonstrate consistently better F_1 score of FLORAL and zeroSum over the $\binom{p}{2}$ log-ratio estimator coda4microbiome and glmnet with various choices of count transformations, where FLORAL showed consistently better false positive control while zeroSum attained better sensitivity under limited sample size or weak effect size. In addition, different choices of the penalty term λ, $\lambda_{1\text{se}}$ and $\lambda_{\min}$, attained different performances under various simulation settings.

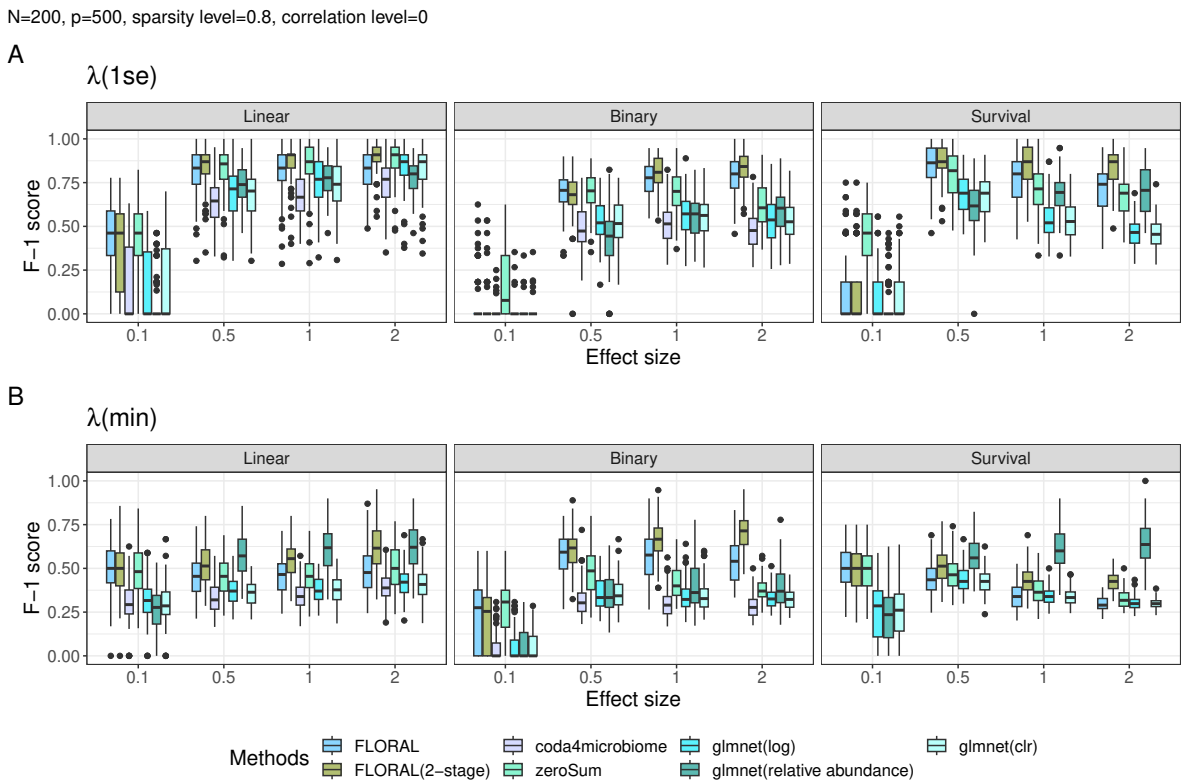

Fig. 1: Variable selection F_1 score obtained by lasso approaches for linear, logistic and Cox log-ratio models with penalty parameter (A) $\lambda_{1\text{se}}$ and (B) $\lambda_{\min}$ with $n = 200, p = 500, s = 0.8, \rho = 0$ and different choices of effect sizes $u = 0.1, 0.5, 1, 2$.

Overall, the variable selection F_1 score for linear, binary, or survival outcomes improves when effect size or sample size increases (**Fig.1, Fig.S1**), and when number of

features or sparsity level decreases (**Fig.S2, S3**). The correlation structure of the simulated underlying relative composition, however, appears not to cause obvious perturbations of the F_1 scores (**Fig.S4**). In addition, $\lambda_{1\text{se}}$ obtained better F_1 score than $\lambda_{\min}$ under larger effect size ($u = 0.5, 1, 2$) or sample size ($n = 100, 200, 500$), while $\lambda_{\min}$ showed better score under the smallest effect size $u = 0.1$ and similar scores as $\lambda_{1\text{se}}$ with the smaller sample size $n = 50$ (**Fig.1A,B and Fig.S1A,B**). These observations can be explained by the higher sensitivity of $\lambda_{\min}$ under limited sample size and weak signals (**Fig.S5C,D and S6C,D**), and the higher specificity of $\lambda_{1\text{se}}$ when sample size is reasonably large and the signals are stronger (**Fig.S5A,B and S6A,B**). Moreover, higher numbers of features caused an inflation of false positives obtained by $\lambda_{\min}$ while $\lambda_{1\text{se}}$ was more robust to larger numbers of features (**Fig.S7A,B**). Higher sparsity levels also caused higher false positive rates and false negative rates especially for binary outcomes (**Fig.S8**). Correlation level, however, appeared not to largely affect the variable selection performances of $\lambda_{\min}$ and $\lambda_{1\text{se}}$ (**Fig.S9**).

Compared to other lasso-based approaches, the `FLORAL` estimator and its two-stage procedure obtained higher F_1 scores under larger effect size and sample size, especially for binary and survival outcomes (**Fig.1, Fig.S1**). Additionally, `FLORAL` showed consistently reliable false positive control under various scenarios (**Fig.S5A-S9A**) with $\lambda = \lambda_{1\text{se}}$, where the two-stage filtering procedure further reduced the number of false positives. Under weaker effect sizes and smaller sample sizes, `zeroSum` showed strong F_1 scores due to its superior sensitivity (**Fig.S5C,S6C**). However, `FLORAL` was able to reach similar or better F_1 score for its less inflated false positive findings compared to `zeroSum`. Compared to `FLORAL` and `zeroSum`, other lasso-based methods showed clear gaps in variable selection performance for binary and survival regression models. The $\binom{p}{2}$ lasso regression method `coda4microbiome` showed heavily inflated false positive hits due to its large covariate space (**Fig.S5A,B, S6A,B**, for example). Similarly, `glmnet` with log-transformed or CLR-transformed counts also suffered high false positive rates. In contrast, `glmnet` with relative abundance showed better false positive control than other lasso-based methods, while it was also the least sensitive method. As sample size increased, `glmnet` with relative abundance reached similar F_1 scores as `FLORAL` achieved under $n = 500$ (**Fig.S1A**) due to the well controlled false positive rate and better sensitivity under a large sample size (**Fig.S6A,C**). These results indicate the better utility of `FLORAL` for the variable selection

due to its nice balance between a reliable false positive control and reasonable sensitivity. The results also demonstrate the better sensitivity of `zeroSum`, which can serve as a decent method for studies with limited sample size.

Apart from variable selection performances, we also evaluated prediction performances (**Fig.S10-S14**) and estimation biases (**Fig.S15-S19**) for the lasso-based methods under various simulation settings. Unlike variable selection where λ_{1se} usually obtained better performances, for prediction and estimation biases λ_{min} tended to achieve better results. Moreover, the performance gaps between lasso-based methods were not as large as seen in variable selection for binary and survival outcomes, where the only obviously under-performing approach was `glmnet` with relative abundance covariate space, which was largely due to the incorrect model specification.

3.2 `FLORAL` Outperforms Differential Abundance Testing Methods in Variable Selection in Simulation Settings

As the leading log-ratio lasso methods, `FLORAL` and `zeroSum` also outperformed popular DA methods under the majority of simulation settings with linear, binary and survival outcomes. As shown in **Fig.2** and **Fig.S20-S23**, the main advantage of `FLORAL` and `zeroSum` was their clearly better sensitivity over the DA methods under various scenarios. On the other hand, `FLORAL` and `zeroSum` in general had higher false positives than the other DA methods, while `FLORAL` with the two-stage taxa filtering demonstrated better false positive control than `zeroSum`.

For survival outcomes, `FLORAL` clearly outperformed other DA methods as most DA methods are not able to properly handle survival information. As the only DA approach which account for censoring data, `LDM` showed reasonably well performance for stronger signals (**Fig.2A**) and larger sample sizes (**Fig.S20A**). However, the DA methods showed lack of sensitivity in scenarios with a smaller sample size or weaker signals. In contrast to survival outcomes, many DA methods performed well for binary outcome due to their intrinsic feature of comparing microbial abundance across two patient groups. In particular, `LEfSe` reached similar level of F_1 scores as `FLORAL` and `zeroSum` under modest effect size (**Fig.2A**) and sample size (**Fig.S20A**). As sample size increased to 500, most DA methods performed similarly well, although still outperformed by `FLORAL` (**Fig.S20**). In terms of continuous outcomes, `ANCOM-BC` showed impressive sensitivity and false positive

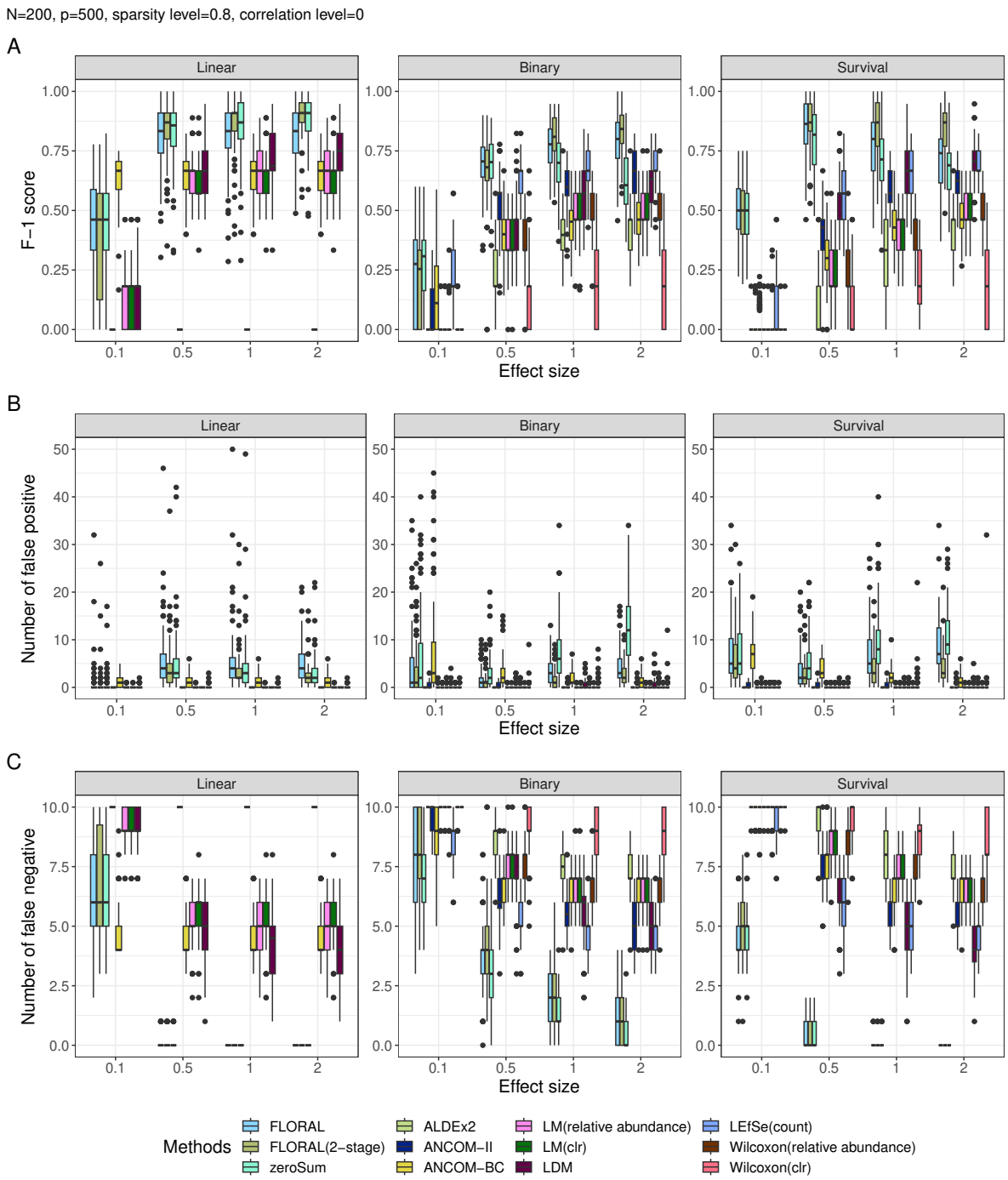

Fig. 2: F_1 score (A), number of false positive features (B) and number of false negative features (C) obtained by `FLORAL`, `zeroSum`, and various DA approaches with $n = 200, p = 500, s = 0.8, \rho = 0$ and $u = 0.1, 0.5, 1, 2$.

control at the smallest effect size considered (**Fig.2A**). However, `ANCOM-BC`'s advantage did not preserve for larger effect sizes and was not seen in other scenarios.

The above comprehensive simulation studies demonstrated the consistent variable selection performance of `FLORAL` for different types of outcome variables with reasonable sensitivity and proper false positive control. Consistent with our previous observations, `zeroSum` was more advantageous over other DA methods under settings with smaller sample sizes and effect sizes for its better sensitivity (**Fig.2C, Fig.S20C**). However, `zeroSum` was outperformed by a few DA methods for binary and survival outcomes with larger sample sizes or effect sizes, which was mainly attributed to the lack of false positive control (**Fig.2B, Fig.S20B**).

3.3 `FLORAL` Identified Enterococcus as a Marker for Overall Survival of Allogeneic HCT Patients

Lasso-based methods and the DA approaches were applied to identify potential microbial markers for overall survival (OS) using the MSKCC allo-HCT cohort data. The majority of the DA methods, including `ANCOM-II`, `ANCOM-BC`, `ALDEx2`, `LDM`, and Wilcoxon rank-sum test, did not find any signals using the FDR adjusted p-value cutoff of 0.05. `LEfSe` identified *Pediococcus*, *Granulicatella*, *Faecalibacterium*, *Rothia*, *Schaalia*, *Terrisporobacter*, *Roseburia*, and *Scardovia* as differentially abundant across patient groups dichotomized by the 2-year survival indicator. The results from the DA methods indicate that the effect sizes of taxa were small, such that the majority of methods could not identify a single hit even with a reasonably large sample size of 663.

The lasso-based methods showed better sensitivity in detecting predictive taxa. To examine the impact of random cross-validation folds on determining the penalty parameter λ, we summarized the frequency of taxa being selected by various lasso-based methods across 100 5-fold cross-validation processes with randomly generated folds in **Fig.3**. As observed, taxa selected by `FLORAL` and `zeroSum` were minimally influenced by the choice of cross validation folds, where *Enterococcus*, *Staphylococcus*, *Schaalia*, *Erysipelatoclostridium* were consistently selected with frequency above 95%. In addition, the two-stage filtering by `FLORAL` identified two specific ratios with strongest associations, namely the ratio between *Schaalia* and *Enterococcus* and the ratio between *Schaalia* and *Staphylococcus*. It it worth noting that *Erysipelatoclostridium* was dropped by the two-stage procedure, indicating that ratios related to *Erysipelatoclostridium* did not show as strong association with OS as the other candidate ratios did. In contrast to `FLORAL`

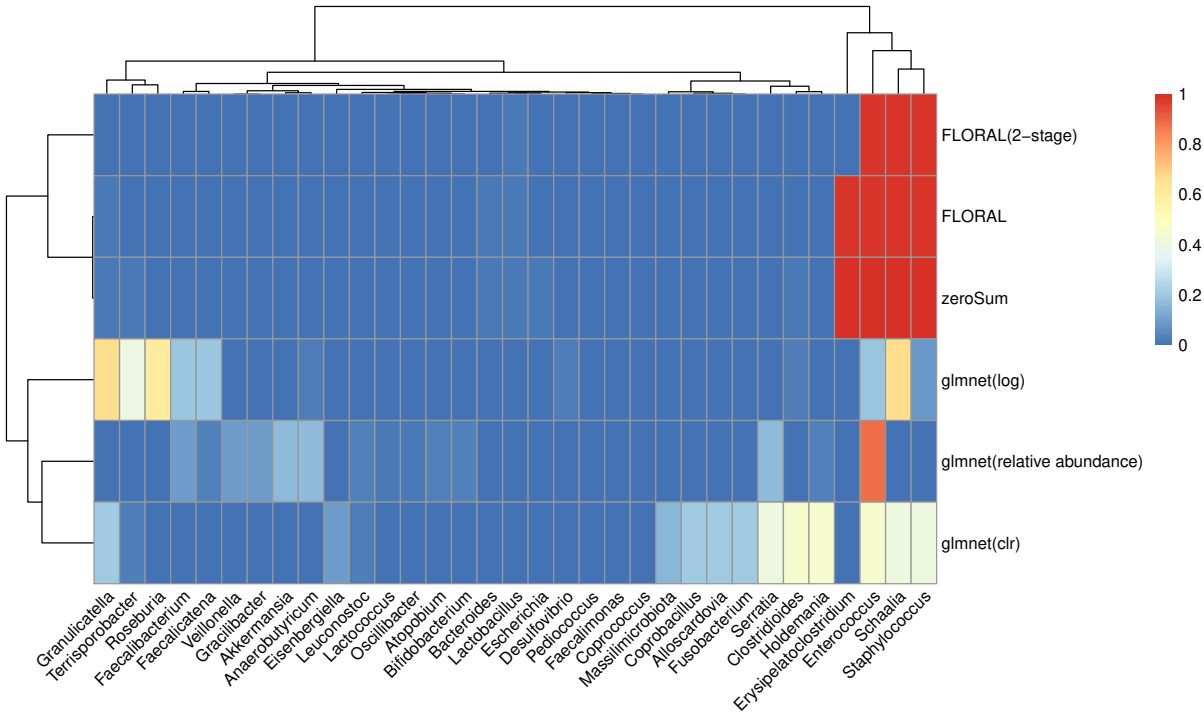

Fig. 3: Heatmap showing the selection frequency of taxa out of 100 random 5-fold cross validations for the six lasso-based methods. Only taxa selected at least once out of 100 5-fold cross validations were included.

and `zeroSum`, taxa selected by other lasso-based methods were heavily influenced by the choice of folds. Although `glmnet` using relative abundance could identify *Enterococcus*, the frequency of selection is only 80%, which still indicates a reasonably high chance of missing *Enterococcus* when running the algorithm only once in practice.

We further investigated the association between the selected two log-ratios and OS after adjusting for patient characteristics. In a multivariable model adjusted for age, disease type, graft source and conditioning intensity, both log-ratios were significantly associated with overall survival (Table 1). As observed, a higher ratio between *Schaalia* and *Enterococcus* is significantly associated with a lower hazard ratio thus better overall survival. This indicates that high abundance level or domination of *Enterococcus* was associated with worse survival, which is consistent with previous findings that *Enterococcus* domination is associated with highly increased risk of antiboitic-resistent bloodsteam infection (Stein-Thoeringer et al., 2019), which further increases the mortality rate.

Characteristic	**HR**[1]	**95% CI**[1]	**p-value**
Log-ratio of *Schaalia* and *Enterococcus*	0.96	0.93, 1.00	**0.042**
Log-ratio of *Schaalia* and *Staphylococcus*	0.95	0.91, 0.99	**0.017**
Age	1.03	1.02, 1.04	**<0.001**
Disease			
AML/ALL/MDS/MPN[2]	—	—	
Others	1.25	0.90, 1.74	0.2
Graft source			
T-cell Depleted	—	—	
Unmodified	1.11	0.76, 1.63	0.6
Cord	1.21	0.69, 2.13	0.5
Conditioning intensity			
Ablative	—	—	
Nonablative	0.65	0.35, 1.19	0.2
Reduced Intensity	0.90	0.59, 1.38	0.6

[1] HR = Hazard Ratio, CI = Confidence Interval

[2] AML: acute myeloid leukemia, ALL: acute lymphoblastic leukemia MDS: myelodysplastic syndrome, MPN: myeloproliferative neoplasm

Table 1: Multivariable model of overall survival with selected log-ratios adjusted for age, disease, graft source and conditioning intensity.

4 Discussion

In this work, we present `FLORAL` for fitting log-ratio lasso regression models powered by the augmented Lagrangian algorithm. Compared to existing log-ratio lasso methods, `FLORAL` maintains reasonable sensitivity in variable selection, shows better false positive control, facilitates a two-stage procedure to enhance the interpretability of the model. We also extended the utility of the method to time-dependent covariates and competing risk outcomes, which aims to better utilize longitudinal microbiome sample information and tackle real world survival analysis challenges due to competing risks outcomes.

Compared to the widely applied microbiome data transformation based on relative abundance $R_{i,k} = X_{i,k}/\sum_{d=1}^{p} X_{i,d}, k = 1, \ldots, p$, the log-ratio model better fits the compositional nature of microbiome data and provides several conveniences in handling the data

and interpreting microbial associations. First, relative abundance $R_{i,k}$ depends on the absolute counts of the collection of p taxa measured from the sequencing process. Given different sequencing depths across samples or studies, the detectable taxa features vary, which may affect the consistency in quantifying $R_{i,k}$. Such difficulty was described as "subcomposition difficulty" in Aitchison (1982), which leads to different analysis results due to the varying definition of the entire feature set. In contrast, the ratio $X_{i,j}/X_{i,k}$ between two taxa j and k is a stable quantity that is invariant of subcomposition changes across samples caused by technical variations, which can potentially enhance the reproducibility of analyses. Moreover, as illustrated in McLaren et al. (2019), taxa-specific bias is highly prevalent in microbiome data, such that only ratios between two taxa carry stable meaning across samples or studies that is invariant to taxa-specific biases, which further supports the analysis based on ratios over relative abundance. In addition, relative abundances $R_{i,1}, \ldots, R_{i,p}$ share the common denominator, which may induce strong colinearity of the covariate space that is not ideal for variable selection. The log-ratio models, on the other hand, can be simplified as regression models (3) and (4) with a constraint, such that the covariate space is still intact for each individual taxon.

As demonstrated by simulation studies, two choices of the penalty parameter $\lambda_{\min}$ and $\lambda_{1\text{se}}$ have different properties. Compared to $\lambda_{\min}$, $\lambda_{1\text{se}}$ offers higher sparsity at the cost of prediction performance. Therefore, we recommend users choosing $\lambda = \lambda_{1\text{se}}$ for variable selection purposes while using $\lambda = \lambda_{\min}$ for prediction purposes. In real data scenario, $\lambda = \lambda_{1\text{se}}$ may return an empty set due to high sparsity of signals. In that case, $\lambda = \lambda_{\min}$ can also be used for variable selection. For better model interpretation, we offer the option of two-stage feature selection which helps to identify specific log taxa ratios which are predictive of clinical outcomes. Based on the simulation results, the two-stage procedure usually further reduced the feature set such that false positives were better controlled.

While aiming to optimize the same loss function as considered by the `zeroSum` package (Altenbuchinger et al., 2017), our augmented Lagrangian algorithm showed different properties from `zeroSum` as illustrated in simulation studies. Under larger sample size or effect size, `zeroSum` showed less controlled false positives. This may be explained by the fact that `zeroSum` randomly chooses pairs of coordinates to update in each cycle (Altenbuchinger et al., 2017), such that a certain parameters may not be updated and shrunk

to zero by chance. In contrast, each coefficient is updated in our algorithm without any random sampling processes, which better avoids false positive findings. However, our simulation does imply that `zeroSum` shows slightly better sensitivity under small effect size and limited sample size.

The proposed method offers an effective alternative to the popular differential abundance testing approaches. Unlike the DA approaches where users are required to specify cutoffs for adjusted p-values, `FLORAL` conducts variable selection based on cross-validated prediction error or model fitting criteria, such that the selected taxa have direct contribution to better prediction performances and are not determined by arbitrary p-value thresholds. Moreover, the log-ratio lasso regression method better addresses the association between survival outcomes and microbiome data and offers a natural framework for incorporating longitudinal covariates, which appeared to be challenging for the DA methods. However, it is important to notice that the DA methods may serve as more reasonable options if the research interest is to investigate how microbiota is varied due to certain exposures. We recommend users considering whether it is more suitable to treat microbiome data as outcome or covariates, and then determining the appropriate analysis tools to apply.

There are several areas for further development of `FLORAL`. First, our current implementation does not adjust for confounding clinical factors. Although this limitation can be addressed by treating the residual of a regression model of outcomes over clinical factors as outcome (Calle et al., 2023), it is unclear how to incorporate it for survival models. In contrast, a more effective way is to incorporate clinical factors as covariates in the lasso model framework without being regularized. Second, it is of high interest for medical researchers to perform inference on selected features, which motivates developments of post-selection inference procedures for the log-ratio lasso models. Last but not least, the application of `FLORAL` can also be extended to other compositional biomedical data, such as cell ratios from flow cytometry or single-cell sequencing experiments, nutrient ratios from dietary data, and metabolomics data.

Authors' Disclosures

J.U. Peled reports research funding, intellectual property fees, and travel reimbursement from Seres Therapeutics, and consulting fees from DaVolterra, CSL Behring, and from MaaT Pharma. He serves on an Advisory board of and holds equity in Postbiotics Plus Research. He has filed intellectual property applications related to the microbiome (reference numbers #62/843,849, #62/977,908, and #15/756,845). D.M. Ponce has served as advisory board member for Evive Biotechnology (Shanghai) Ltd (formerly Generon [Shanghai] Corporation Ltd), she served as advisory board member or consultant of Sanofi Corporation, CareDx, Ceramedix, Incyte, and receives research funding from Takeda Corporation and Incyte. M.-A. Perales reports honoraria from Adicet, Allovir, Caribou Biosciences, Celgene, Bristol-Myers Squibb, Equilium, Exevir, Incyte, Karyopharm, Kite/Gilead, Merck, Miltenyi Biotec, MorphoSys, Nektar Therapeutics, Novartis, Omeros, OrcaBio, Syncopation, VectivBio AG, and Vor Biopharma. He serves on DSMBs for Cidara Therapeutics, Medigene, and Sellas Life Sciences, and the scientific advisory board of NexImmune. He has ownership interests in NexImmune and Omeros. He has received institutional research support for clinical trials from Incyte, Kite/Gilead, Miltenyi Biotec, Nektar Therapeutics, and Novartis. M.R.M. van den Brink has received research support and stock options from Seres Therapeutics and stock options from Notch Therapeutics and Pluto Therapeutics; he has received royalties from Wolters Kluwer; has consulted, received honorarium from or participated in advisory boards for Seres Therapeutics, Vor Biopharma, Rheos Medicines, Frazier Healthcare Partners, Nektar Therapeutics, Notch Therapeutics, Ceramedix, Lygenesis, Pluto Therapeutics, GlaskoSmithKline, Da Volterra, Thymofox, Garuda, Novartis (Spouse), Synthekine (Spouse), Beigene (Spouse), Kite (Spouse); he has IP Licensing with Seres Therapeutics and Juno Therapeutics; and holds a fiduciary role on the Foundation Board of DKMS (a nonprofit organization). MSK has institutional financial interests relative to Seres Therapeutics.

Acknowledgments

This research was supported by National Cancer Institute award numbers, R01-CA228358, R01-CA228308, P30 CA008748 MSK Cancer Center Support Grant/Core Grant and P01-CA023766; National Heart, Lung, and Blood Institute (NHLBI) award number

R01-HL123340 and R01-HL147584; National Institute on Aging award number P01-AG052359, and Tri-Institutional Stem Cell Initiative. Additional funding was received from The Lymphoma Foundation, The Susan and Peter Solomon Family Fund, The Solomon Microbiome Nutrition and Cancer Program, Cycle for Survival, Parker Institute for Cancer Immunotherapy, Paula and Rodger Riney Multiple Myeloma Research Initiative, Starr Cancer Consortium, and Seres Therapeutics. OM reports funding from the Hyundai Hope on Wheels Young Investigator Award and Tow Center for Developmental Oncology Career Development Award. ML reports funding from Beijing Nova Program of Science and Technology (Z191100001119120). JUP reports funding from NHLBI NIH Award K08HL143189.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 139–160.

Altenbuchinger, M., Rehberg, T., Zacharias, H., Stämmler, F., Dettmer, K., Weber, D., Hiergeist, A., Gessner, A., Holler, E., Oefner, P. J., et al. (2017). Reference point insensitive molecular data analysis. *Bioinformatics* **33,** 219–226.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology* **26,** 32–46.

Bates, S. and Tibshirani, R. (2019). Log-ratio lasso: scalable, sparse estimation for log-ratio models. *Biometrics* **75,** 613–624.

Bertsekas, D. P. (2014). *Constrained optimization and Lagrange multiplier methods.* Academic press.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological monographs* **27,** 326–349.

Burgos da Silva, M., Ponce, D. M., Dai, A., M. Devlin, S., Gomes, A. L., Moore, G., Slingerland, J., Shouval, R., Armijo, G. K., DeWolf, S., et al. (2022). Preservation of the fecal microbiome is associated with reduced severity of graft-versus-host disease. *Blood, The Journal of the American Society of Hematology* **140,** 2385–2397.

Calle, M. L., Pujolassos, M., and Susin, A. (2023). coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC bioinformatics* **24,** 82.

Copelan, E., Casper, J. T., Carter, S. L., van Burik, J.-A. H., Hurd, D., Mendizabal, A. M., Wagner, J. E., Yanovich, S., and Kernan, N. A. (2007). A scheme for defining cause of death and its application in the t cell depletion trial. *Biology of Blood and Marrow Transplantation* **13,** 1469–1476.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202.

Derosa, L., Routy, B., Thomas, A. M., Iebba, V., Zalcman, G., Friard, S., Mazieres, J., Audigier-Valette, C., Moro-Sibilot, D., Goldwasser, F., et al. (2022). Intestinal akkermansia muciniphila predicts clinical response to pd-1 blockade in patients with advanced non-small-cell lung cancer. *Nature medicine* **28,** 315–324.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics & Data Analysis* **71,** 1054–1063.

Fernandes, A. D., Macklaim, J., Linn, T., Reid, G., and Gloor, G. (2013). Anova-like differential gene expression analysis of single-organism and meta-rna-seq. *PLoS one* **8,** e67019.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* **94,** 496–509.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33,** 1.

Gacesa, R., Kurilshikov, A., Vich Vila, A., Sinha, T., Klaassen, M., Bolte, L., Andreu-Sánchez, S., Chen, L., Collij, V., Hu, S., et al. (2022). Environmental factors shaping the gut microbiome in a dutch population. *Nature* **604,** 732–739.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology* **8,** 2224.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hu, Y., Li, Y., Satten, G. A., and Hu, Y.-J. (2022). Testing microbiome associations with survival times at both the community and individual taxon levels. *PLoS Computational Biology* **18,** e1010509.

Hu, Y.-J. and Satten, G. A. (2020). Testing hypotheses about the microbiome using the linear decomposition model (ldm). *Bioinformatics* **36,** 4106–4115.

Lee, K. A., Thomas, A. M., Bolte, L. A., Björk, J. R., de Ruijter, L. K., Armanini, F., Asnicar, F., Blanco-Miguez, A., Board, R., Calbet-Llopart, N., et al. (2022). Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nature Medicine* **28,** 535–544.

Lin, H. and Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature communications* **11,** 3514.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101,** 785–797.

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., Tickle, T. L., Weingart, G., Ren, B., Schwager, E. H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. *PLoS computational biology* **17,** e1009442.

Martin, B. D., Witten, D., and Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics* **14,** 94.

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *Elife* **8,** e46923.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10,** 1200–1202.

Peled, J. U., Gomes, A. L., Devlin, S. M., Littmann, E. R., Taur, Y., Sung, A. D., Weber, D., Hashimoto, D., Slingerland, A. E., Slingerland, J. B., et al. (2020). Microbiota as

predictor of mortality in allogeneic hematopoietic-cell transplantation. *New England Journal of Medicine* **382,** 822–834.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology* **12,** 1–18.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal* **27,** 379–423.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software* **39,** 1.

Simpson, E. H. (1949). Measurement of diversity. *nature* **163,** 688–688.

Stein-Thoeringer, C., Nichols, K., Lazrak, A., Docampo, M., Slingerland, A., Slingerland, J., Clurman, A., Armijo, G., Gomes, A., Shono, Y., et al. (2019). Lactose drives enterococcus expansion to promote graft-versus-host disease. *Science* **366,** 1143–1149.

Therneau, T., Crowson, C., and Atkinson, E. (2020). Multi-state models and competing risks. *CRAN-R (https://cran. r-project. org/web/packages/survival/vignettes/compete. pdf)* .

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* pages 809–834.

Wallen, Z. D., Demirkan, A., Twa, G., Cohen, G., Dean, M. N., Standaert, D. G., Sampson, T. R., and Payami, H. (2022). Metagenomics of parkinson's disease implicates the gut microbiome in multiple disease mechanisms. *Nature Communications* **13,** 6958.

Worsley, S. F., Davies, C. S., Mannarelli, M.-E., Hutchings, M. I., Komdeur, J., Burke, T., Dugdale, H. L., and Richardson, D. S. (2021). Gut microbiome composition, not alpha diversity, is associated with survival in a natural vertebrate population. *Animal microbiome* **3,** 1–18.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in microbiome-profiling studies

with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics* **96,** 797–807.

Zhu, Z., Satten, G. A., Mitchell, C., and Hu, Y.-J. (2021). Constraining permanova and ldm to within-set comparisons by projection improves the efficiency of analyses of matched sets of microbiome data. *Microbiome* **9,** 1–19.