Enhanced Feature Selection for Microbiome Data using FLORAL: Scalable Log-ratio Lasso Regression

Teng Fei*[1], Tyler Funnell[2], Nicholas R. Waters[2], Sandeep S. Raj[3], Keimya Sadeghi[2], 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†[8]

[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

[8]City of Hope Los Angeles and City of Hope National Medical Center

*Corresponding author, email: feit1@mskcc.org

†Corresponding author, email: mvandenbrink@coh.org

Abstract

Identifying predictive biomarkers of patient outcomes from high-throughput microbiome data is of high interest, while existing computational methods do not satisfactorily account for complex survival endpoints, longitudinal samples, and taxa-specific sequencing biases. We present `FLORAL` (https://vdblab.github.io/FLORAL/), an open-source computational tool to perform scalable log-ratio lasso regression and microbial feature selection for continuous, binary, time-to-event, and competing risk outcomes, with compatibility of longitudinal microbiome data as time-dependent covariates. 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 and real-data analyses, `FLORAL` achieved consistently better false-positive control compared to other lasso-based approaches, and better sensitivity over popular differential abundance testing methods for datasets with smaller sample size. In a survival analysis in allogeneic hematopoietic-cell transplant, we further demonstrated considerable improvement by `FLORAL` in microbial feature selection by utilizing longitudinal microbiome data over only using baseline microbiome data.

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[1]. At the same time, the rapidly 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 [2, 3].

Allogeneic hematopoietic cell transplants (allo-HCT) provides a paradigm for understanding the importance of microbiome composition in clinical outcomes. While provided to patients with curative intent, high-dose chemotherapy prior to the transplant causes severe damage to gut microbiota, which further increases the risk of life-threatening gut

inflammation, opportunistic infections, and malnutrition. Therefore, it is of high interest to monitor and study the association between the microbial profiles and the corresponding patient outcomes, which are commonly coded as continuous, binary, time-to-event, or competing risks outcomes [4].

To fulfill the need of identifying microbial biomarkers, differential taxa abundance analysis approaches have been applied to compositional microbiome data [5]. As the sequencing depth may heavily vary across samples, one should account for both observed count and sequencing depth (i.e. total read count per sample) to facilitate a standardized quantification of a taxon of interest across samples. One naive approach is to perform the two-sample Wilcoxon rank-sum test for the observed relative taxon abundance (count divided by sequencing depth). More sophisticated strategies include applying multi-stage Wilcoxon tests and linear discriminant analysis (`LEfSe` [6]), modeling high presence of zero counts (`metagenomeSeq` [7], `ANCOM-II` [8], `ANCOM-BC` [9]), direct modeling of count data (`ALDEx2` [10], `corncob` [11]), and performing permutation tests (`LDM` [3, 12, 13]).

While the above differential abundance (DA) testing methods are useful, there are some important limitations. Typically, the DA methods perform multiple hypothesis testing followed by p-value adjustment, where taxon-outcome associations are assessed in a univariable manner without accounting for other taxa, which tends to inflate the number of selected taxa. In addition, taxa selection is determined by a chosen p-value threshold, where the choices of 0.2, 0.1 and 0.05 have been widely reported without consensus [14–16], potentially contributing to reproducibility issues in microbiome research [**citations**]. Moreover, the majority of DA methods lack utilities of handling time-to-event response variables and longitudinal microbiome data, compromising the best use of data by performing comparisons across binary "event" and "non-event" groups without accounting for follow-up [17]. Furthermore, taxon-specific sequencing bias may disrupt the rank of relative abundances across samples [18], suggesting methods based on relative abundance or sequencing depths may suffer potential performance loss.

As an alternative approach to identifying the taxa-outcome association, penalized log-ratio regression (or log-ratio lasso) models were derived from classic compositional data regression [19], treating ratios between microbial features as predictors, with linear [20–23], binary [21–23], or time-to-event [21] outcome variables. Since there are $\binom{p}{2}$ unique pairwise ratios out of p taxa, computationally efficient algorithms with zero-sum

constrained loss functions [20–22] were widely established, avoiding direct enumeration of ratios [23]. In addition, a two-step variable selection scheme was proposed to further reduce the false discovery rate [22]. Unlike the DA methods, log-ratio lasso regression assesses taxa-outcome associations in multivariable models, conducts variable selection using more objective criteria based on cross validations, naturally incorporates various types of response variables including time-to-event, and effectively circumvents the potential issues introduced by taxa-specific sequencing biases [18]. Nevertheless, currently available software packages (`zeroSum` [21], `logratiolasso` [22]) have not comprehensively implemented all previously developed features for various outcome types or variable selection strategies. Moreover, the existing methods were not developed to incorporate complex outcomes such as competing risks [4, 24], or time-dependent microbial predictors, which have already been widely available in large-scale longitudinal clinical studies.

Here we propose `FLORAL` to perform linear, logistic, Cox proportional hazards [25], and Fine-Gray proportional subdistributional hazards [26] log-ratio lasso regression and subsequent feature selection for high-dimensional compositional data (**Fig.1**). We develop a unified loss function framework that can easily adapt various types of outcome variables (**Fig.1A**). Instead of enumerating $\binom{p}{2}$ possible pairs of taxa, the proposed algorithm works on the p-dimensional covariate space as facilitated by the zero-sum constraint, which only requires affordable computing memory. To accommodate longitudinal microbiome data, `FLORAL` enables time-dependent covariates in the Cox and Fine-Gray models. Furthermore, `FLORAL` is featured with built-in multi-step variable selection with further enhanced false discovery control and model interpretability (**Fig.1B**).

We conducted extensive real-data and simulation studies to assess our method's performance and compare with various benchmark methods. We demonstrate that `FLORAL` achieves reasonable sensitivity and high specificity in publicly available microbiome datasets from 39 studies with binary comparison groups [27]. Using a 16S rRNA sequencing dataset of 8,967 longitudinal stool samples from a cohort of 1,415 allo-HCT patients from Memorial Sloan Kettering Cancer Center (MSKCC), we illustrate that incorporating longitudinal microbiome data can provide much richer information compared to only using baseline microbiome data, where we successfully identified *Enterococcus*, *Blautia*, *Erysipelatoclostridium*, and *Staphylococcus* as predictive features of patient overall survival, which have been previously reported [28–31].

2 Results

2.1 Simulations Showed Superior Variable Selection Performance of FLORAL Among Lasso-based Methods and Beyond

Extensive simulations based on the log-ratio models were performed to evaluate the sensitivity, specificity, and overall variable selection performance (F_1 score) for different methods with various types of simulated outcomes, including continuous, binary, survival and competing-risks outcomes. Here, F_1 score is defined as $F_1 = 2(\text{precision}^{-1}+\text{recall}^{-1})^{-1}$ on a range between 0 and 1, where a higher F_1 score indicates a better overall performance of precision and recall. We considered simulation scenarios with varying sample sizes (n), effect sizes (u), number of features (p), feature correlations (ρ), and feature sparsity levels (s), aiming to conduct a comprehensive method evaluation. In each simulation run, the outcome was generated based on log-ratios formed by 10 underlying "true" features. See Online Methods for detailed simulation configurations and descriptions of compared methods and evaluation metrics.

Our simulations demonstrated superior variable selection performance of FLORAL (**Fig.2, S1-S5**). As shown in **Fig.2**, FLORAL achieved the highest median F_1 score out of 100 simulations in most scenarios with different sample sizes and types of outcomes, with big performance advantages for binary and survival outcomes under moderate sample sizes ($n = 100, 200$). Similar performance advantages were also observed under different effect sizes (**Fig.S2**), number of features (**Fig.S3**), correlation levels (**Fig.S4**) and sparsity levels (**Fig.S5**), with a few exceptions where FLORAL also reached comparable performances compared to other methods.

The better performance of FLORAL can be explained by the effective control of false positive features via its two-step feature selection mechanism (**Fig.S1A**) and the high sensitivity as an intrinsic characteristic of lasso-based method (**Fig.S1B**). Like other lasso-based methods, FLORAL obtained better overall F_1 scores at $\lambda = \lambda_{1\text{se}}$ than at $\lambda = \lambda_{\min}$ in most simulated scenarios, where a sparser selected feature set offered much fewer false positive features. Due to its stricter feature selection process, FLORAL's sensitivity was slightly compromised when the effect size was very weak ($u = 0.1$, equivalent to odds ratio or hazard ratio of $e^{0.1} = 1.1$) or the sample size was small ($n = 50$) (**Fig.S1,S2**), where the setting $\lambda = \lambda_{1\text{se}}$ could obtain zero selected features while $\lambda = \lambda_{\min}$ might reach a

better F_1 score. Nevertheless, `FLORAL` still achieved reasonable improvements over other lasso-based methods at fairly moderate effect sizes ($u = 0.25, 0.5$), larger sample sizes ($n \geq 100$) and various other settings.

Compared to `FLORAL` and other lasso-based methods, the DA methods showed generally lower false-positive rates but also much lower sensitivity at smaller sample sizes and moderate effect sizes (**Fig.S1B,S2C**), resulting in lower overall F_1 scores (**Fig.2,S2A**). As sample size increased, the DA methods gained higher power to recognize true signals, gradually reaching or exceeding `FLORAL`'s F_1 scores at sample size $n = 500$ (**Fig.2**). Notably, `metagenomeSeq` appeared to over-select features with a higher sensitivity but also much higher false-positive rates compared to other methods, while `ANCOM-BC` tended to have slightly inflated false positive rates at smaller sample sizes and smaller effect sizes (**Fig.S1A,S2B**). Moreover, `LDM` and `LEfSe` showed high robustness at reasonably large effect size ($u = 0.5$) and sample size ($n = 200$), where both methods maintained the best median F_1 scores across all DA methods for binary and survival outcomes (**Fig.S3-S5**), outperforming `FLORAL` at smaller numbers of features ($p \leq 200$, **Fig.S3**) or at higher sparsity levels ($s = 0.95$, **Fig.S5**). This demonstrated the robustness of methods based on permutation test (`LDM`) and non-parametric test (`LEfSe`).

2.2 `FLORAL` Demonstrated Effective False Positive Control on 39 Publicly Available 16S rRNA Amplicon Sequencing Datasets

We applied various lasso-based regression methods and differential abundance testing methods on publicly available 16S microbiome datasets for 39 studies [32–67] as reported by Nearing et al. [27]. The 39 datasets contain a variety of research contexts including human, mouse, and environmental studies, where for each specific study there were two groups with hypothetical differences in their corresponding taxa abundance profiles. The distribution of sample size, number of features and the ratio between the comparison group sizes had a wide range, where both the sample size and number of features ranged from less than 50 to several thousands (**Fig.3A**). For lasso-based methods, we treated the identity of the binary comparison groups as a binary outcome, such that logistic regression with lasso penalty was performed.

Due to the lack of gold standard definition of truly differentially abundant taxa, it is challenging to assess methods' sensitivity. Therefore, we mainly focused on evaluating

the specificity of the methods by randomly shuffling the group labels for each data set then running the methods. Theoretically, the differential abundance signals will be fully eliminated after random shuffling, such that any selected taxon can be treated as a false-positive feature. In parallel, we also applied the same methods on the original datasets without random group label shuffling, which offered descriptive statistics such as number of selected taxa, computing time, and median area under the receiver operating characteristic curve (AUC) of all selected taxa. See Methods section for detailed descriptions on the datasets and the configurations of different methods.

Fig.3B displays the numbers of selected taxa by various methods for the 39 public 16S datasets with shuffled group labels. As described above, larger numbers of selected taxa indicated higher false-positive rates. As observed, the lasso-based methods obtained reasonably low false-positive rates with the penalty parameter $\lambda = \lambda_{1se}$, while there was an inflation of false positives when using $\lambda = \lambda_{min}$. Thanks to its two-step variable selection strategy, `FLORAL` showed consistently lower numbers of false-positives than `zeroSum` while selecting zero taxa for all but three datasets with $\lambda = \lambda_{1se}$. In terms of the DA methods, like observed in the simulations (Section 2.1), most methods selected zero taxa for most datasets and showed good false positive control. However, `metagenomeSeq` failed to control false positive findings, with false-positive rates up to 20% for most datasets. In addition, `ANCOM-BC` also had fairly high false-positive rates for datasets with relatively low sample sizes. The above observations were highly consistent with our simulation results (**Fig.S1A**), which further demonstrated `FLORAL`'s satisfactory protection against false positive findings.

The same analysis procedure was repeated for the original group labels without shuffling. The lasso-based approaches tended to select fewer genera than the DA methods (**Fig.S6**). This is expected as the DA methods perform comparisons for independent taxa then using multiple testing adjustment, such that many highly correlated features may be selected simultaneously. In contrast, lasso-based methods perform feature selection from multivariable regression models, such that the selected features are conditioned on all other feature values, resulting in a sparser set of selected taxa. Notably, `ANCOM-BC` and `metagenomeSeq` selected more genera than other methods for most datasets, which can be explained by their high false-positive rates as observed in **Fig.3B**. In addition, `FLORAL` achieved high median AUC (**Fig.S7**) and reasonable computing time (**Fig.S8**),

showing good practical utility for datasets of diverse characteristics.

2.3 FLORAL Achieves Robust Signal Detection in Time-dependent Microbiome Samples

Allogeneic hematopoietic bone marrow transplant (allo-HCT) patients from Memorial Sloan Kettering Cancer Center (MSKCC) with eligible samples with 16S rRNA sequencing data between January 2009 and June 2021 were selected to investigate the associations between taxa abundance and patient overall survival (OS), transplant-related mortality (TRM) and graft versus host disease (GvHD)-related mortality (GRM). Here, TRM and GRM are defined as described by Copelan et al. [4] with relapse and progression of disease as competing risks. Two patient cohorts were derived, namely the peri-engraftment sample cohort and the longitudinal sample cohort (**Fig.S9**). The peri-engraftment sample cohort (912 patients, 912 samples) consisted of all patients with at least one sample collected between day 7 and 21 after transplant, where the latest collected sample was used as a peri-engraftment "baseline" sample, such that the microbial association with survival outcomes was investigated using only peri-engraftment samples. Accordingly, time to survival outcomes was landmarked at the sample collection day related to transplant. In contrast, the longitudinal sample cohort (1,415 patients, 8,967 samples) included all patients with samples available between day -30 to 730 relative to transplant, where the latest sample collected before the transplant day was regarded as the baseline (day 0) sample. Patients without available pre-transplant samples will enter the risk set of the survival models at days corresponding to their earliest available post-transplant samples. As listed in **Table S1**, patient characteristics of the two cohorts are largely similar, which created an ideal scenario to compare the strength of signals using peri-engraftment samples versus using longitudinal samples.

`FLORAL` was utilized to fit log-ratio lasso models with peri-engraftment samples and longitudinal samples for overall survival (Cox model), TRM (Fine-Gray model) and GRM (Fine-Gray model), where the penalty parameter was set as $\lambda = \lambda_{1\text{se}}$ to enhance false-positive protection. In addition, the optional step of variable selection for drawing taxa selection probability was also applied, for 100 repeated 5-fold cross validations, to evaluate signal detection efficiency from either peri-engraftment samples or longitudinal samples. The regression models were adjusted for covariates including patient disease type,

graft source, age, and conditioning intensity, where the lasso penalty was only applied to taxa features but not the covariates. We also applied other lasso-based methods and popular DA methods to investigate associations between genera and OS using the peri-engraftment and longitudinal sample cohort if compatible. See Online Methods for detailed description of the methods and cohorts used.

The taxa selection probabilities obtained from `FLORAL` demonstrated much stronger signal detection capability of longitudinal microbiome features compared to peri-engraftment microbiome features (**Fig.4**). Using the peri-engraftment sample cohort, the microbial feature detection rates were below 50% for all three considered survival endpoints, indicating feature detection was largely dependent on the fold split and was less reliable (**Fig.4A-C**). In contrast, The longitudinal sample cohort provided not only more samples per patient but also more patients with eligible samples, which helped identify genera with detection rates higher than 80% or even 100% (**Fig.4D-F**). In particular, genera *Enterococcus*, *Blautia*, *Erysipelatoclostridium* and *Staphylococcus* were selected from the longitudinal sample cohort with high selection probabilities. Specifically, *Enterococcus* and *Staphylococcus* showed consistently harmful associations with OS, TRM and GRM, *Blautia* were identified to be associated with better OS and lower GRM cumulative incidence, and *Erysipelatoclostridium* were found to be also associated with better OS, and lower TRM and GRM cumulative incidences. Such high selection probability for the above three genera was not seen from the models using the peri-engraftment sample cohort (**Fig.4A-C**). The above results from the longitudinal sample cohort were highly consistent with previous studies [28–31, 68], demonstrating powerful utilities of `FLORAL` in analyzing survival endpoints with longitudinal microbiome data, where the signal detection is much more robust than using a single-time microbiome sample for each patient.

Compared to `FLORAL`, other lasso-based methods and popular DA methods did not achieve as effective feature selection performances. Like `FLORAL`, `glmnet`-based lasso models can also incorporate longitudinal microbial features with different data transformation options. However, these methods reached much lower feature selection rates than `FLORAL` using the longitudinal sample cohort in 100 cross-validation runs (**Fig.S10**), where important genera such as *Enterococcus* were hardly detected. In addition, `zeroSum` and `glmnet` were not able to better detect important genera using the peri-engraftment sample cohort than `FLORAL` (**Fig.S11**), indicating weak signals when only using the peri-engraftment

microbiome samples. Unlike `FLORAL` and `glmnet`, the DA methods are incompatible with longitudinal microbiome samples, and thus were only applied for the peri-engraftment cohort. As shown in **Fig.S12**, many DA methods conservatively selected no features at the threshold of 0.05 for the adjusted p-value, while `LEfSe` and `metagenomeSeq` selected a large number of genera. Nevertheless, all DA methods failed to identify *Blautia* and *Erysipelatoclostridium* as detected by `FLORAL` using the longitudinal sample cohort. The above results suggest that `FLORAL`'s improvements in microbial feature selection from the peri-engraftment cohort to the longitudinal cohort are attributed not only to its flexibility of incorporating longitudinal microbial features as time-dependent covariates, but also to its infrastructure of utilizing log-ratio based regression models.

3 Discussion

In this work, we present `FLORAL` for fitting log-ratio lasso regression models powered by the augmented Lagrangian algorithm with a two-step variable selection procedure. Compared to existing log-ratio lasso methods, `FLORAL` maintains reasonable sensitivity in variable selection, shows better false positive control in real data analyses, and effectively improves signal detection by incorporating longitudinal microbial features as time-dependent covariates.

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 $R_{i,k}$. This challenge was earlier described as "subcomposition difficulty" [19], 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, taxa-specific bias is highly prevalent in microbiome data [18], such that only ratios between two taxa carry stable relative magnitudes across samples

or studies that is invariant to taxa-specific biases, which further supports the analysis based on ratios over relative abundance.

As demonstrated by simulation and real data studies, the two choices of the penalty parameter $\lambda_{\min}$ and λ_{1se} have different properties, where λ_{1se} achieved better F_1 scores in simulations and lower false positive rates in the analyses of 39 real datasets compared to $\lambda_{\min}$. Therefore, we recommend users to choose $\lambda = \lambda_{1se}$ for better control of false discoveries. In studies with smaller sample sizes, it is likely to detect zero features with $\lambda = \lambda_{1se}$ in a single two-step variable selection with cross-validation. For those studies with small scales, we recommend using multiple runs of cross-validation to rank the strength of the features by their selection probabilities, such that features with weak signals may still be captured and reported regarding their importance relative to other features.

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 microbial features, which appeared to be challenging for the DA methods. However, it is important to note that the DA methods may serve as more reasonable options if the research interest is to compare paired or correlated microbial features [12], where generalized estimating equation (GEE) extensions of log-ratio lasso regression are required to better account for the dependency across subjects.

In large-scale follow up studies with longitudinal samples, one commonly encountered challenge is to utilize all of the microbiome data. Due to the limitation of the DA methods, it is usually only possible to perform two-sample comparisons for microbiome samples collected at a specified time window, such as the peri-engraftment period in the allo-HCT patient cohort. Although linear mixed-effect models (`MaAsLin2`, for example) have been proposed for longitudinal microbiome data analysis [2], the method can be regarded as an extended DA method in the sense that it requires a pre-specified significance threshold, clearly defined groups for comparison, and a data transformation scheme which

is usually based on relative abundance. When the comparison groups are well defined at baseline, such as treatment group versus control group, it is helpful to apply linear mixed-effect models to investigate the association between the comparison groups and microbial trajectories. On the other hand, if a survival endpoint is of interest, then a regression model with time-dependent covariates, like `FLORAL`, will be more appropriate to better incorporate time-to-event information.

There are several opportunities of further development for `FLORAL`. First, the regularization model can be extended beyond the scope of the lasso regression with ℓ_1-penalty, which facilitates subsequent fine-tuning of the models with potential utility of prediction. Our augmented Lagrangian algorithm can be easily modified to perform elastic-net regression [69], adaptive lasso [70], or other regularization forms. 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, Crestone Inc, 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). Memorial Sloan Kettering Cancer Center (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 and the V Foundation.

References

1. Gacesa, R. *et al.* Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604,** 732–739 (2022).
2. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLoS computational biology* **17,** e1009442 (2021).
3. Hu, Y.-J. & Satten, G. A. Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics* **36,** 4106–4115 (2020).
4. Copelan, E. *et al.* 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 (2007).
5. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology* **8,** 2224 (2017).
6. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome biology* **12,** 1–18 (2011).
7. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10,** 1200–1202 (2013).
8. Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology* **8,** 2114 (2017).
9. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nature communications* **11,** 3514 (2020).
10. Fernandes, A. D., Macklaim, J., Linn, T., Reid, G. & Gloor, G. ANOVA-like differential gene expression analysis of single-organism and meta-RNA-seq. *PLoS one* **8,** e67019 (2013).
11. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics* **14,** 94 (2020).
12. Zhu, Z., Satten, G. A., Mitchell, C. & Hu, Y.-J. 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 (2021).

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

421 14. Derosa, L. *et al.* Intestinal Akkermansia muciniphila predicts clinical response to PD-
422 1 blockade in patients with advanced non-small-cell lung cancer. *Nature medicine*
423 **28,** 315–324 (2022).

424 15. Lee, K. A. *et al.* Cross-cohort gut microbiome associations with immune checkpoint
425 inhibitor response in advanced melanoma. *Nature Medicine* **28,** 535–544 (2022).

426 16. Wallen, Z. D. *et al.* Metagenomics of Parkinson's disease implicates the gut micro-
427 biome in multiple disease mechanisms. *Nature Communications* **13,** 6958 (2022).

428 17. Worsley, S. F. *et al.* Gut microbiome composition, not alpha diversity, is associated
429 with survival in a natural vertebrate population. *Animal microbiome* **3,** 1–18 (2021).

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

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

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

436 21. Altenbuchinger, M. *et al.* Reference point insensitive molecular data analysis. *Bioin-*
437 *formatics* **33,** 219–226 (2017).

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

440 23. Calle, M. L., Pujolassos, M. & Susin, A. coda4microbiome: compositional data analy-
441 sis for microbiome cross-sectional and longitudinal studies. *BMC bioinformatics* **24,**
442 82 (2023).

443 24. Bakoyannis, G. & Touloumi, G. Practical methods for competing risks data: a re-
444 view. *Statistical methods in medical research* **21,** 257–272 (2012).

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

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

449 27. Nearing, J. T. *et al.* Microbiome differential abundance methods produce different
450 results across 38 datasets. *Nature Communications* **13,** 342 (2022).

451 28. Taur, Y. *et al.* Intestinal domination and the risk of bacteremia in patients undergo-
452 ing allogeneic hematopoietic stem cell transplantation. *Clinical infectious diseases*
453 **55,** 905–914 (2012).

454 29. Peled, J. U. *et al.* Microbiota as predictor of mortality in allogeneic hematopoietic-
455 cell transplantation. *New England Journal of Medicine* **382,** 822–834 (2020).

456 30. Miltiadous, O. *et al.* Early intestinal microbial features are associated with CD4
457 T-cell recovery after allogeneic hematopoietic transplant. *Blood, The Journal of the*
458 *American Society of Hematology* **139,** 2758–2769 (2022).

459 31. Nguyen, C. L. *et al.* High-resolution analyses of associations between medications,
460 microbiome, and mortality in cancer patients. *Cell* **186,** 2705–2718 (2023).

461 32. Chase, J. *et al.* Geography and location are the primary drivers of office microbiome
462 composition. *MSystems* **1,** 10–1128 (2016).

463 33. Ji, P., Parks, J., Edwards, M. A. & Pruden, A. Impact of water chemistry, pipe ma-
464 terial and stagnation on the building plumbing microbiome. *PloS one* **10,** e0141087
465 (2015).

466 34. Nearing, J. T. *et al.* Infectious complications are associated with alterations in the
467 gut microbiome in pediatric patients with acute lymphoblastic leukemia. *Frontiers*
468 *in Cellular and Infection Microbiology* **9,** 28 (2019).

469 35. Son, J. S. *et al.* Comparison of fecal microbiota in children with autism spectrum
470 disorders and neurotypical siblings in the simons simplex collection. *PloS one* **10,**
471 e0137725 (2015).

472 36. Schubert, A. M. *et al.* Microbiome data distinguish patients with Clostridium difficile
473 infection and non-C. difficile-associated diarrhea from healthy controls. *MBio* **5,** 10–
474 1128 (2014).

37. Dinh, D. M. *et al.* Intestinal Microbiota, Microbial Translocation, and Systemic Inflammation in Chronic HIV Infection. *Journal of Infectious Diseases* **211,** 19–27. ISSN: 1537-6613. http://dx.doi.org/10.1093/infdis/jiu409 (July 2014).

38. Goodrich, J. K. *et al.* Human Genetics Shape the Gut Microbiome. *Cell* **159,** 789–799. ISSN: 0092-8674. http://dx.doi.org/10.1016/j.cell.2014.09.053 (Nov. 2014).

39. Vincent, C. *et al.* Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. *Microbiome* **1.** ISSN: 2049-2618. http://dx.doi.org/10.1186/2049-2618-1-18 (June 2013).

40. Baxter, N. T., Ruffin, M. T., Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine* **8.** ISSN: 1756-994X. http://dx.doi.org/10.1186/s13073-016-0290-3 (Apr. 2016).

41. Singh, P. *et al.* Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome* **3.** ISSN: 2049-2618. http://dx.doi.org/10.1186/s40168-015-0109-2 (Sept. 2015).

42. Papa, E. *et al.* Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease. *PLoS ONE* **7** (ed Ravel, J.) e39242. ISSN: 1932-6203. http://dx.doi.org/10.1371/journal.pone.0039242 (June 2012).

43. Ross, M. C. *et al.* 16S gut community of the Cameron County Hispanic Cohort. *Microbiome* **3.** ISSN: 2049-2618. http://dx.doi.org/10.1186/s40168-015-0072-y (Mar. 2015).

44. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484. ISSN: 1476-4687. http://dx.doi.org/10.1038/nature07540 (Nov. 2008).

45. Mejía-León, M. E., Petrosino, J. F., Ajami, N. J., Domínguez-Bello, M. G. & de la Barca, A. M. C. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Scientific Reports* **4.** ISSN: 2045-2322. http://dx.doi.org/10.1038/srep03814 (Jan. 2014).

505 46. Frère, L. *et al.* Microplastic bacterial communities in the Bay of Brest: Influence
506 of polymer type and size. *Environmental Pollution* **242,** 614–625. ISSN: 0269-7491.
507 http://dx.doi.org/10.1016/j.envpol.2018.07.023 (Nov. 2018).

508 47. Hoellein, T. J. *et al.* Longitudinal patterns of microplastic concentration and bac-
509 terial assemblages in surface and benthic habitats of an urban river. *Freshwater*
510 *Science* **36,** 491–507. ISSN: 2161-9565. http://dx.doi.org/10.1086/693012
511 (Sept. 2017).

512 48. Alkanani, A. K. *et al.* Alterations in Intestinal Microbiota Correlate With Sus-
513 ceptibility to Type 1 Diabetes. *Diabetes* **64,** 3510–3520. ISSN: 1939-327X. http:
514 //dx.doi.org/10.2337/db14-1847 (June 2015).

515 49. Kesy, K., Oberbeckmann, S., Kreikemeyer, B. & Labrenz, M. Spatial Environmental
516 Heterogeneity Determines Young Biofilm Assemblages on Microplastics in Baltic Sea
517 Mesocosms. *Frontiers in Microbiology* **10.** ISSN: 1664-302X. http://dx.doi.org/
518 10.3389/fmicb.2019.01665 (Aug. 2019).

519 50. De Tender, C. A. *et al.* Bacterial Community Profiling of Plastic Litter in the Belgian
520 Part of the North Sea. *Environmental Science & Technology* **49,** 9629–9638. ISSN:
521 1520-5851. http://dx.doi.org/10.1021/acs.est.5b01093 (Aug. 2015).

522 51. Oberbeckmann, S., Osborn, A. M. & Duhaime, M. B. Microbes on a Bottle: Sub-
523 strate, Season and Geography Influence Community Composition of Microbes Col-
524 onizing Marine Plastic Debris. *PLOS ONE* **11** (ed Carter, D. A.) e0159289. ISSN:
525 1932-6203. http://dx.doi.org/10.1371/journal.pone.0159289 (Aug. 2016).

526 52. Rosato, A. *et al.* Microbial colonization of different microplastic types and biotrans-
527 formation of sorbed PCBs by a marine anaerobic bacterial community. *Science of*
528 *The Total Environment* **705,** 135790. ISSN: 0048-9697. http://dx.doi.org/10.
529 1016/j.scitotenv.2019.135790 (Feb. 2020).

530 53. Lamoureux, E. V., Grandy, S. A. & Langille, M. G. I. Moderate Exercise Has Limited
531 but Distinguishable Effects on the Mouse Microbiome. *mSystems* **2** (ed Lozupone,
532 C.) ISSN: 2379-5077. http://dx.doi.org/10.1128/mSystems.00006-17 (Aug.
533 2017).

534 54. Dranse, H. J. *et al.* The impact of chemerin or chemokine-like receptor 1 loss on the
535 mouse gut microbiome. *PeerJ* **6,** e5494. ISSN: 2167-8359. http://dx.doi.org/10.
536 7717/peerj.5494 (Sept. 2018).

537 55. Douglas, G. M. *et al.* Multi-omics differentially classify disease state and treatment
538 outcome in pediatric Crohn's disease. *Microbiome* **6.** ISSN: 2049-2618. http://dx.
539 doi.org/10.1186/s40168-018-0398-3 (Jan. 2018).

540 56. McCormick, A. R. *et al.* Microplastic in surface waters of urban rivers: concentration,
541 sources, and associated bacterial assemblages. *Ecosphere* **7.** ISSN: 2150-8925. http:
542 //dx.doi.org/10.1002/ecs2.1556 (Nov. 2016).

543 57. Gevers, D. *et al.* The Treatment-Naive Microbiome in New-Onset Crohn's Disease.
544 *Cell Host & Microbe* **15,** 382–392. ISSN: 1931-3128. http://dx.doi.org/10.1016/
545 j.chom.2014.02.005 (Mar. 2014).

546 58. Lozupone, C. A. *et al.* Alterations in the Gut Microbiota Associated with HIV-1
547 Infection. *Cell Host & Microbe* **14,** 329–339. ISSN: 1931-3128. http://dx.doi.org/
548 10.1016/j.chom.2013.08.006 (Sept. 2013).

549 59. Schneider, D. *et al.* Gut bacterial communities of diarrheic patients with indications
550 of Clostridioides difficile infection. *Scientific Data* **4.** ISSN: 2052-4463. http://dx.
551 doi.org/10.1038/sdata.2017.152 (Oct. 2017).

552 60. Yurgel, S. N. *et al.* Variation in Bacterial and Eukaryotic Communities Associated
553 with Natural and Managed Wild Blueberry Habitats. *Phytobiomes Journal* **1,** 102–
554 113. ISSN: 2471-2906. http://dx.doi.org/10.1094/PBIOMES-03-17-0012-R (Jan.
555 2017).

556 61. Zhu, L. *et al.* Characterization of gut microbiomes in nonalcoholic steatohepatitis
557 (NASH) patients: A connection between endogenous alcohol and NASH. *Hepatology*
558 **57,** 601–609. ISSN: 0270-9139. http://dx.doi.org/10.1002/hep.26093 (Jan.
559 2013).

560 62. Scheperjans, F. *et al.* Gut microbiota are related to Parkinson's disease and clinical
561 phenotype. *Movement Disorders* **30,** 350–358. ISSN: 1531-8257. http://dx.doi.
562 org/10.1002/mds.26069 (Dec. 2014).

563 63. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced
564 susceptibility to arthritis. *elife* **2,** e01202 (2013).

64. Zupancic, M. L. *et al.* Analysis of the Gut Microbiota in the Old Order Amish and Its Relation to the Metabolic Syndrome. *PLoS ONE* **7** (ed Thameem, F.) e43052. ISSN: 1932-6203. http://dx.doi.org/10.1371/journal.pone.0043052 (Aug. 2012).

65. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10.** ISSN: 1744-4292. http://dx.doi.org/10.15252/msb.20145645 (Nov. 2014).

66. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nature Microbiology* **4,** 1183–1195. ISSN: 2058-5276. http://dx.doi.org/10.1038/s41564-019-0426-5 (May 2019).

67. Noguera-Julian, M. *et al.* Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* **5,** 135–146. ISSN: 2352-3964. http://dx.doi.org/10.1016/j.ebiom.2016.01.032 (Mar. 2016).

68. Stein-Thoeringer, C. *et al.* Lactose drives Enterococcus expansion to promote graft-versus-host disease. *Science* **366,** 1143–1149 (2019).

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

70. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101,** 1418–1429 (2006).

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

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

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

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

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

76. Nocedal, J. & Wright, S. J. Penalty and augmented Lagrangian methods. *Numerical Optimization,* 497–528 (2006).

77. Scheike, T. H., Zhang, M.-J. & Gerds, T. A. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95,** 205–220 (2008).

78. Nearing, J. 16S rRNA Microbiome Dataset https://figshare.com/articles/dataset/16S_rRNA_Microbiome_Datasets/14531724 (May 2021).

79. Liao, C. *et al.* Compilation of longitudinal microbiota data and hospitalome from hematopoietic cell transplantation patients. *Scientific data* **8,** 71 (2021).

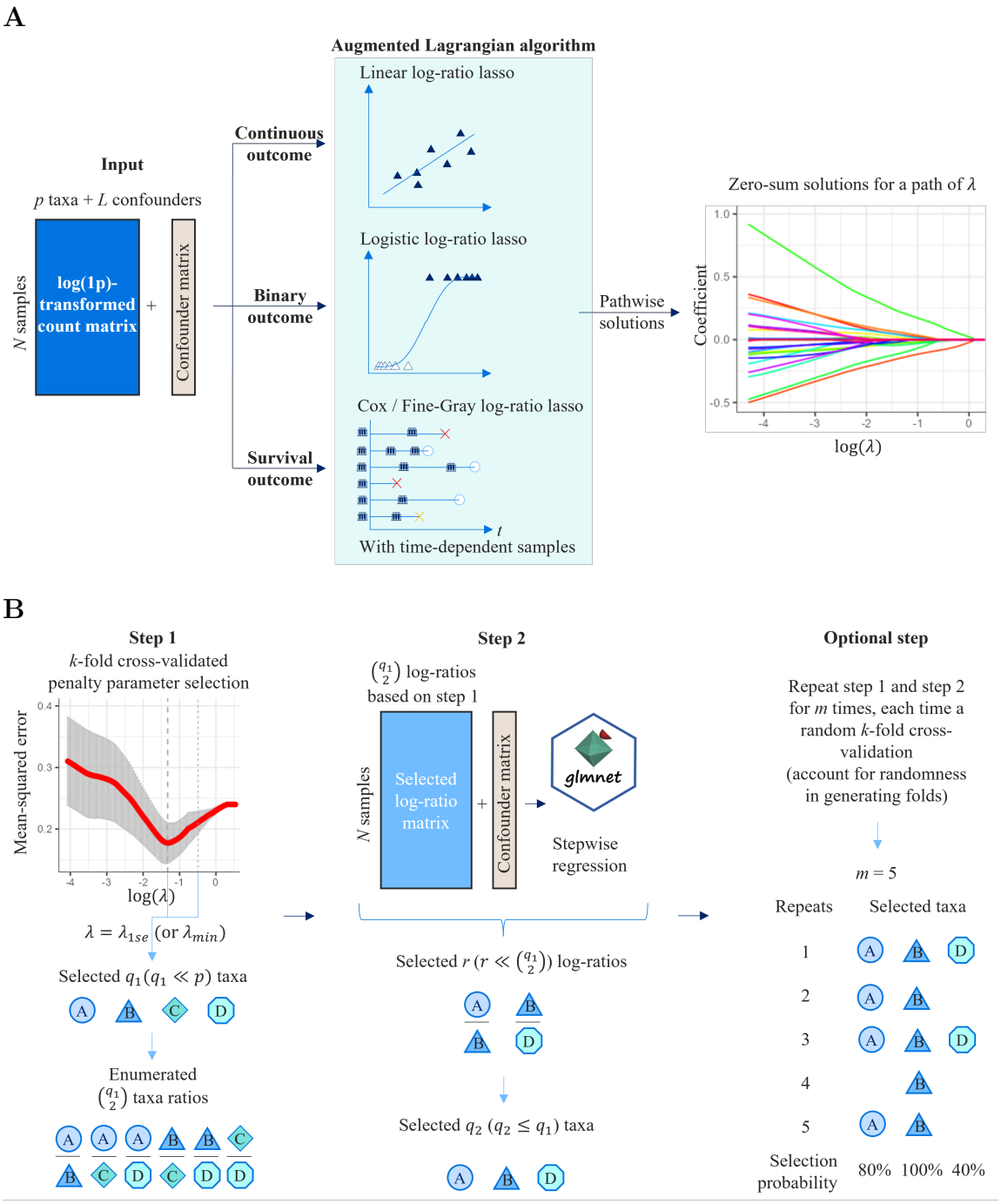

Fig. 1: `FLORAL` performs log-ratio lasso regression and stepwise feature selection. **A.** The log-ratio lasso regression is conducted by an augmented Lagrangian algorithm with a zero-sum constraint, which is compatible with continuous, binary, survival and competing-risk outcomes. Longitudinal microbiome samples can be incorporated in survival models as time-dependent covariates. The algorithm is applied on a pre-specified path of penalty parameter λ, which returns a path of coefficient estimates satisfying the zero-sum constraint. **B.** Variable selection starts with k-fold cross validation (Step 1) which selects the penalty parameter and corresponding taxa with non-zero coefficients. The log-ratios enumerated from the taxa selected in Step 1 will be filtered in Step 2 by lasso regression and stepwise regression, where the remaining ratios and corresponding taxa are reported. Optionally, Step 1 and 2 can be repeated with additional k-fold data splits and calculations of taxa selection probabilities.

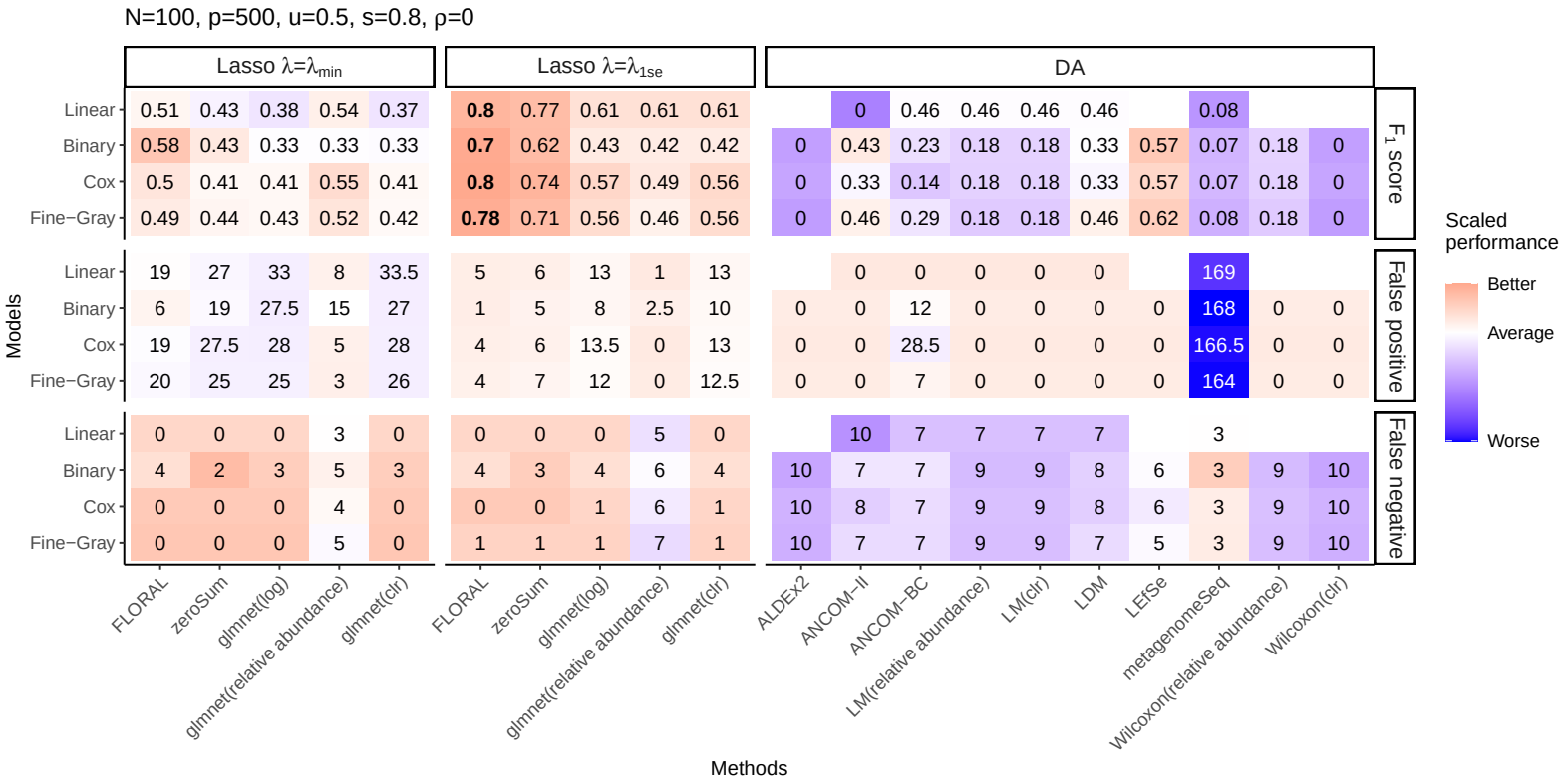

Fig. 2: Median F_1 score, median number of false positive features, and median number of false negative features obtained by lasso and DA methods for linear, binary, survival, and competing risk models out of 100 simulations with $n = 100, u = 0.5, p = 500, s = 0.8, \rho = 0$, where there were 10 true features out of $p = 500$ features in each simulation run. For each type of regression model, metrics across all methods were scaled to mean zero and standard deviation one for color visualization. The highest median F_1 scores across all methods were printed in bold fonts. For the DA methods, the censoring indicator of the survival or competing risks outcomes were used to define patient groups except for LDM, where the Martingale residual was first computed then correlated with taxa abundances. Part of the DA methods were not evaluated for continuous outcome due to incompatibility. The adjusted p-value cutoff was set as 0.05 for all DA methods. log: log-transformation; clr: centered log-ratio transformation.

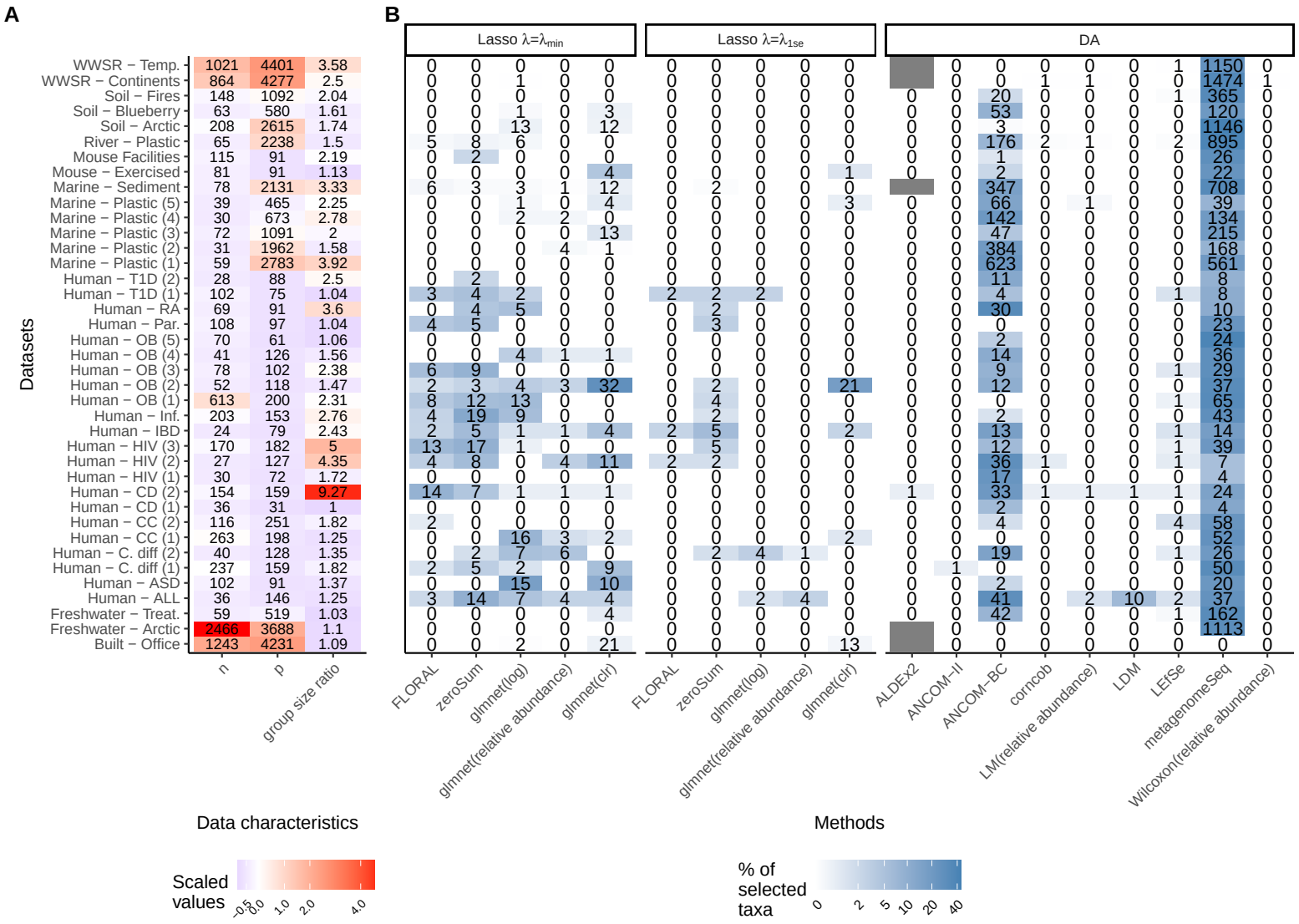

Fig. 3: **A.** Data characteristics of the 39 publicly available 16S microbiome datasets, including sample size (n), number of genera (p), and ratio between the sizes of comparison groups. The color scheme represents scaled characteristics across all datasets. **B.** Number of selected taxa from the 39 publicly available 16S microbiome datasets by feature selection methods, with comparison group labels randomly shuffled. Part of data were unavailable for ALDEx2 due to memory overflow. The color scheme represents the percentage of selected taxa out of all taxa in a certain data set.

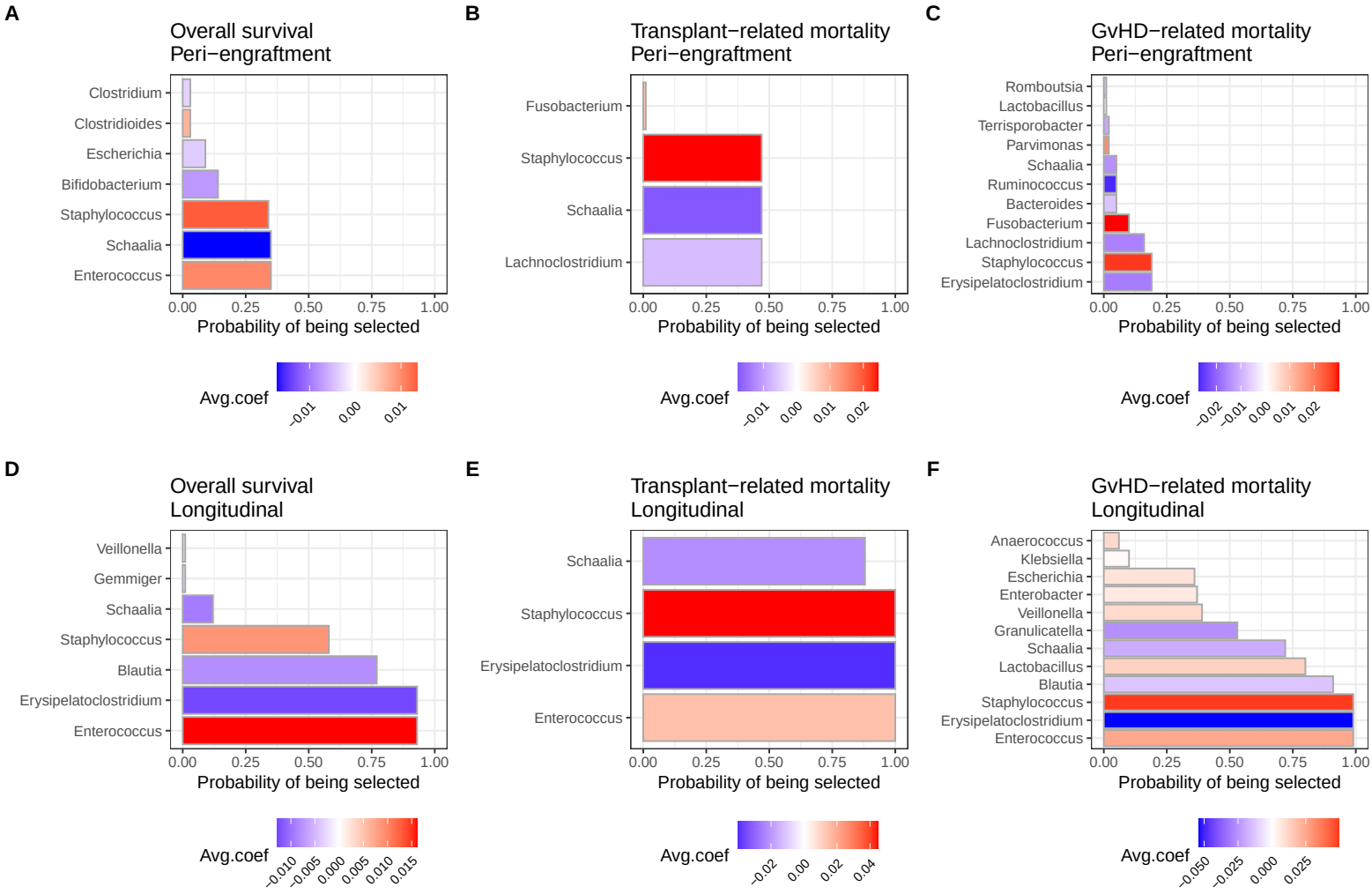

Fig. 4: Probabilities of genera being selected from 100 repeats of 5-fold cross-validation with random fold split for Cox model of overall survival with **A.** peri-engraftment samples or **D.** longitudinal samples; Fine-Gray model of transplant-related mortality with **B.** peri-engraftment samples or **E.** longitudinal samples; Fine-Gray model of GvHD-related mortality with **C.** peri-engraftment samples or **F.** longitudinal samples. The color scheme represents the average lasso coefficient estimates of the corresponding genus at $\lambda^{(i)} = \lambda_{1\text{se}}$ over $i = 1, \ldots, 100$ repeats.

4 Methods

4.1 Overview of FLORAL

Given p microbial features, L confounding factors (if applicable), and the corresponding outcome of interest, FLORAL performs log-ratio lasso regression and subsequent variable selection for continuous, binary, and survival or competing risk outcomes (**Fig.1**), where longitudinal microbial features are incorporated in time-to-event models as time-dependent covariates. The regression model assumes that only a sparse set of the $\binom{p}{2}$ possible ratios between two microbial features are associated with the outcome of interest, which can be achieved by ℓ_1-regularization on a $\binom{p}{2}$-dimensional unknown parameter space. An augmented Lagrangian algorithm with a zero-sum constraint, which effectively reduces the covariate space from $\binom{p}{2}$ dimensions to p dimensions, was developed to conduct pathwise estimation of the log-ratio lasso models under pre-specified values of the penalty parameter λ. Subsequently, the step 1 variable selection is based on a k-fold cross validation, where a cross-validated predictive model assessment metric (such as mean-squared error or deviance) helps to identify a value of λ which achieves the best prediction ($\lambda_{\min}$) or a sparser feature set with reasonable model fitting ($\lambda_{1\text{se}}$). Given selected λ, the q_1 taxa ($q_1 << p$) with non-zero regression coefficients will be selected as taxa contributing to better prediction performances. With q_1 selected taxa, FLORAL enumerates all $\binom{q_1}{2}$ possible ratio configurations, then performs the step 2 variable selection by running lasso regression followed by stepwise regression on the $\binom{q_1}{2}$-dimensional log-ratio features, which further selects r ratios ($r << \binom{q_1}{2}$) with strongest signals. Subsequently, q_2 taxa ($q_2 \leq q_1$) forming the r selected ratios can be obtained as selected set of predictive taxa. As an optional step, variable selection steps 1 and 2 can be repeated for m times, such that the variable selection can be replicated under multiple random configurations of folds for cross validations. This optional step can help assess the probability of a certain taxon being selected after accounting for the uncertainties in defining folds.

4.2 Log-ratio Regression Models

For a given sample, let $\boldsymbol{X}$ denote the absolute count vector for p microbial taxa. Let $\boldsymbol{W}$ denote the confounder vector with L features. For a scalar outcome such as continuous

or binary outcome, we denote the corresponding response variable as Y. For survival outcome, we denote $(\tilde{T}, \Delta)$ as observed survival time subject to right censoring and the censoring indicator, respectively. We denote the realization of the above random quantities for the ith patient as $\boldsymbol{X}_i, \boldsymbol{W}_i, Y_i, \tilde{T}_i, \Delta_i$, respectively. For a scalar outcome Y_i, we model the association between Y_i and $\boldsymbol{X}_i$, $\boldsymbol{W}_i$ via a log-ratio generalized linear regression model (GLM):

$$g\{E(Y_i|\boldsymbol{X}_i, \boldsymbol{W}_i)\} = \theta_0 + \sum_{1\le j<k\le p} \theta_{j,k} \log\left(\frac{X_{i,j}}{X_{i,k}}\right) + \sum_{1\le l\le L} \omega_l W_{i,l}, \tag{1}$$

where $g(\cdot)$ is a link function accounting for the distribution of $\boldsymbol{Y}$, θ_0 is an unknown intercept term, $A_{i,j}$ represents the jth element of the vector $\boldsymbol{A}_i$, and $\theta_{j,k}$, ω_l are unknown coefficients corresponding to the paired log-ratios $\log(X_{i,j}/X_{i,k})$ and the patient characteristics $W_{i,l}$, respectively. Here, we adapt the notion of pairwise log-ratio [22], where there are $\binom{p}{2}$ unknown $\theta_{j,k}$ for $1 \le j < k \le p$. Similarly, for survival outcome $(\tilde{T}_i, \Delta_i)$, we consider a log-ratio proportional hazards model

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

where $h(t|\boldsymbol{X}_i, \boldsymbol{W}_i)$ denotes the hazard function conditioned on microbial features $\boldsymbol{X}_i$ and patient characteristics $\boldsymbol{W}_i$, and $h_0(t)$ is the baseline hazard function. Note that model (2) can be naturally extended for longitudinal microbiome data $\boldsymbol{X}(t)$, such that

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

where $\boldsymbol{X}(t)$ can be updated at different times of sample collection. In practice, longitudinal microbiome samples are only available at a finite number of time points, where the last value carried forward (LVCF) strategy is applied [71]. Moreover, the Fine-Gray subdistributional proportional hazards model [26] can be equivalently estimated by a weighted Cox model [72], 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 by rewriting the log of ratios as differences of log-counts [22, 23]. Let $\beta_k = \sum_{j=1}^{k-1} -\theta_{j,k} + \sum_{j=k+1}^{p} \theta_{k,j}$, 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

$$g\{E(Y_i|\boldsymbol{X}_i,\boldsymbol{W}_i)\} = \theta_0+\sum_{k=1}^{p}\beta_k\log X_{i,k}+\sum_{l=1}^{L}\omega_l W_{i,l},\ \text{subject to}\ \sum_{k=1}^{p}\beta_k=0 \tag{3}$$

and

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

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.

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

Denote $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$, $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_L)^T$, and $\boldsymbol{\zeta}=(\theta_0,\boldsymbol{\beta}^T,\boldsymbol{\omega}^T)^T$ for the GLM model or $\boldsymbol{\zeta}=(\boldsymbol{\beta}^T,\boldsymbol{\omega}^T)^T$ for the proportional hazards model. Let $\mathcal{L}(\boldsymbol{\zeta})$ denote the log-likelihood of model (3) or the log-partial likelihood of model (4). We define the log-ratio lasso estimator as

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

where λ and ξ are regularization penalty parameters and $\|\cdot\|_1$ denotes ℓ_1-norm. Here, we consider different regularization parameters for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ to facilitate higher flexibility in real practice, where investigators may set $\xi=\lambda$ or $\xi=0$ to conduct microbial feature selections with or without penalizing the confounding covariate effects.

We adapt the similar treatment in `glmnet` [73] to approximate $\mathcal{L}(\boldsymbol{\zeta})$ by its second-order Taylor expansion centered at $\tilde{\boldsymbol{\zeta}}$, which is either a vector of initial values or the estimates from a previous iteration. Let $\tilde{\boldsymbol{X}}=\{\log(\boldsymbol{X}_1+1),\ldots,\log(\boldsymbol{X}_n+1)\}^T$ and

$\tilde{\boldsymbol{W}} = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n\}^T$ denote the $n \times p$ log-transformed microbiome count matrix and the $n \times L$ confounding covariate matrix, respectively. Define $\boldsymbol{Z} = \{\boldsymbol{1}_n, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{W}}\}$ (GLM) or $\boldsymbol{Z} = \{\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{W}}\}$ (proportional hazards model), $\tilde{\boldsymbol{\eta}} = \boldsymbol{Z}\tilde{\boldsymbol{\zeta}}$ as the linear predictor with $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$, 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 by second-order Taylor expansion we have

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

where the first term in the formula on the last row is a weighted quadratic form of $\boldsymbol{Z}\boldsymbol{\zeta}$ and the second term $C(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\zeta}})$ is independent of $\boldsymbol{\zeta}$. To alleviate computational burdens for the $n \times n$ matrix $\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})$, we follow [69] and [74] 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

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

which is a standard weighted least squares form with continuous response vector $\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}})$ and predictor matrix $\boldsymbol{Z}$. It is also straightforward to show that $\tilde{\mathcal{L}}(\boldsymbol{\zeta})$ is equivalent to the standard least squares form $\tilde{\mathcal{L}}(\boldsymbol{\zeta}) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\zeta}\|_2^2$ when $\boldsymbol{Y}$ is continuous. Based on the working loss function, we obtain the working optimization problem for the proposed lasso estimator:

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

With a unified formula of working loss function (6) for different choices of $\mathcal{L}(\boldsymbol{\zeta})$, 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 log-likelihood, or partial log-likelihood function $\mathcal{L}(\boldsymbol{\zeta})$.

We adapted the augmented Lagrangian approach [75] to solve the constrained optimization problem (7). Specifically, the constraint $\sum_{k=1}^{p} \beta_k = 0$ is incorporated in the

following target function

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

where γ is the Lagrange multiplier, $\mu(\sum_{k=1}^{p}\beta_k)^2/2$ is the standard term used in the penalty method. Following Lin et al.'s approach [20], we define $\alpha = \gamma/\mu$ 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. In practice, the augmented Lagrangian method is able to achieve the constraint without using a overly large value of μ, which avoids ill-conditioning caused by having large μ [76]. We typically let $\mu = 1$ as fixed in our algorithm.

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

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

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. Similarly, if the mth component of $\hat{\boldsymbol{\zeta}}^{(i)}$ belongs to $\hat{\boldsymbol{\omega}}^{(i)}$, then it is updated by

$$\hat{\zeta}_m^{(i)} = \frac{1}{\boldsymbol{Z}_m^T\mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}\boldsymbol{Z}_m/n} S_\xi\Big(\frac{1}{n}\boldsymbol{Z}_m^T\mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}\{\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}}) - \sum_{l\neq m}\hat{\zeta}_l^{(i)}\boldsymbol{Z}_l\}\Big).$$

As observed from the above update formula for $\hat{\boldsymbol{\beta}}^{(i)}$ and $\hat{\boldsymbol{\omega}}^{(i)}$, the main difference is that the zero-sum constraint is only applied for $\hat{\boldsymbol{\beta}}^{(i)}$, but not for $\hat{\boldsymbol{\omega}}^{(i)}$. After each inner-loop coordinate descent for each feature, we update $\tilde{\boldsymbol{\zeta}}$, $\tilde{\boldsymbol{\eta}}$, $\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\eta}})$, and $\mathrm{diag}\{\ddot{\mathcal{L}}(\tilde{\boldsymbol{\eta}})\}$ for the next inner-loop coordinate descent. The updates of $\hat{\boldsymbol{\zeta}}^{(i)}$ stops if the loss function $\tilde{L}_\mu(\hat{\boldsymbol{\zeta}}^{(i)},\gamma)$ 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 $\|\boldsymbol{\zeta}^{(i)} - \boldsymbol{\zeta}^{(i-1)}\|_1 < \delta$ 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 $u = u' = 100$.

4.4 Pathwise Solution and Cross Validation

To have a global picture on how feature 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

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

which acquires $\hat{\boldsymbol{\zeta}} = \mathbf{0}$ [74]. 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 $n \geq p$. Here we consider $\xi = \lambda$ or $\xi = 0$, such that ξ follows the same path as λ does or is fixed as a constant.

k-fold 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 and deviance are used to evaluate prediction errors. Two choices of λ are reported, namely $\lambda_{\min}$ which minimizes cross-validated prediction error, and $\lambda_{1\mathrm{se}}$ which provides a sparser solution than $\lambda_{\min}$ but still obtains cross-validated prediction error within one standard error of that of $\lambda_{\min}$. The cross validation serves as the first step of variable selection in `FLORAL` (**Fig.1B**), where taxa with non-zero coefficient estimates $\hat{\boldsymbol{\beta}}$ at $\lambda_{\min}$ or $\lambda_{1\mathrm{se}}$ are selected for step 2 variable selection.

4.5 Step 2 Variable Selection

In the previous sections we derive the algorithm to efficiently find a sparse set of predictive taxa. However, specific pairs of log-ratios are not identifiable via cross validation based on the estimates obtained by Algorithm 1. To facilitate a sparser feature selection with interpretability for specific ratios, one natural extension is to perform exhaustive search on all possible pairs of the log-ratios for the selected features obtained from the Step 1 cross validation [22]. Since the number of selected taxa q_1 from the Step 1 cross validation is much smaller than p, the corresponding number of pairwise combinations $\binom{q_1}{2}$ is also much smaller than $\binom{p}{2}$ (**Fig.1B**), which only requires standard memory usage in standard R packages for lasso models and stepwise regression. 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.

4.6 An Optional Step for Feature Selection Probabilities

The result of the cross-validated variable selection and the subsequent second step selection depends on how subjects are split into folds, such that different fold splits may select different taxa or taxa ratios. In real data analysis where sample size is small or signals are weak, it is helpful to repeat the cross validation for more objective evaluations of feature selection.

Thus, we developed an optional step to assess the reliability of variable selection, which repeats the k-fold cross-validated 2-step variable selection procedure by m times (**Fig.1B**). In each of the m repeats, the cross validation folds will be randomly generated, such that the corresponding penalty parameter λ will correspond to different sets of selected features. This optional step allows investigators to assess how robustly a certain microbial feature is selected based on different fold split schemes, where a higher selection probability indicates higher confidence of association between the feature and the outcome.

4.7 Simulation Studies

4.7.1 Data Generation

We performed extensive simulation studies to assess various methods' performances under different scenarios. Let n be the sample size and p be the number of features. 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 $c_{ik} \geq 0$ 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 $\boldsymbol{c}_i$.

Four types of outcomes, namely continuous, binary, time-to-event, and competing risk, are considered in our simulations conditioned on $\boldsymbol{c}_i$. Given $\boldsymbol{c}_i$, we first generate a "true count" vector $\boldsymbol{C}_i$ following a multinomial distribution with 10^6 counts and probability vector $\boldsymbol{c}_i$. Note the $\boldsymbol{C}_i$ facilitates defining the log-ratios by $\log(1+\cdot)$ transformation, which mitigates the arbitrary choice of increments for proportions $\boldsymbol{c}_i$. Then the corresponding underlying true linear predictor is generated as

$$l_i = 0.5u\left\{ \log \frac{C_{i1}+1}{C_{i2}+1} + \log \frac{C_{i3}+1}{C_{i4}+1} \right\} + u\left\{ \log \frac{C_{i5}+1}{C_{i6}+1} + \log \frac{C_{i7}+1}{C_{i8}+1} + \log \frac{C_{i9}+1}{C_{i10}+1} \right\},$$

such that the first ten simulated features are true features associated with the outcome in the form of log-ratios. Here u controls the effect sizes, where the first two ratios have half of the effect sizes of the latter three ratios. Given l_i, the continuous response variable Y_i^c is generated by

$$Y_i^c = l_i + \epsilon_i,$$

where 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 $q_i = \text{expit}(l_i)$ and $\text{expit}(x) = 1/(1+e^{-x})$. For the time-to-event outcome,, the event time T_i is simulated from the distribution function $F_{T_i}(t) = 1 - \exp\{0.1(1 - e^t)\exp(l_i)\}$.

It can be shown that the associated hazard function is equal to $0.1e^t \exp(l_i)$ 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. For the competing risk outcomes, we follow Scheike et al.'s simulation approach [77]. Specifically, two failure types are assumed, where the cumulative incidence of the first and second failure types satisfy $F_{i,1}(t) = 1 - \{1 - 0.66(1 - e^{-t})\}^{l_i}$ and $F_{i,2}(t) = 1 - 0.34^{l_i}\{1 - \exp(-tl_i)\}$, respectively. The failure type $\epsilon_i \in \{1, 2\}$ can then be generated by the failure type probabilities defined by $F_{i,1}(\infty)$ and $F_{i,2}(\infty)$. Given failure type ϵ_i, the failure time T_i is generated from the conditional distribution function $F_{i,\epsilon_i}(t)/F_{i,\epsilon_i}(\infty)$. An Independent censoring time V_i is independently generated from a $Unif(0.19, 10)$. Then the observable survival time $\tilde{T}_i^c = \min(T_i, V_i)$ and failure type indicator $\Delta_i^c = \epsilon_i I(T_i < V_i)$ are obtained. In data analysis, we focus on investigating the association between features and the first of the two failure types.

Based on the underlying true taxa composition $\boldsymbol{c}_i$, we further simulate the observable count data $\boldsymbol{X}_i$ with different sequencing depths. First, the sequencing depth D_i is generated 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 $\boldsymbol{c}_i$.

For each of the four types of outcome variables, we investigated the performance of methods based on a reference scenario where $n = 200$, $p = 500$, $s = 0.8$, $\rho = 0$, and $u = 0.5$. 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 $u = 0.1, 0.25, 0.5$. This serves as a comprehensive survey in understanding the behavior of methods under various settings. For each simulation run, the simulated outcome $[\boldsymbol{Y}^c, \boldsymbol{Y}^b, (\tilde{\boldsymbol{T}}, \boldsymbol{\Delta})$, or $(\tilde{\boldsymbol{T}}^c, \boldsymbol{\Delta}^c)]$ and observable count matrix $\tilde{\boldsymbol{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ were the input data to various methods.

4.7.2 Method Configuration and Assessment

We tested lasso-based methods and different abundance (DA) testing methods in simulations as listed in **Fig.2**. For lasso-based methods, we considered methods with zero-sum constrained lasso (`FLORAL` and `zeroSum`) and standard `glmnet` models with relative abun-

dance, centered log-ratio transformed counts, and log-transformed counts. The same random fold split was used for all methods, where 10-fold cross-validated mean-squared error was used to identify $\lambda_{\min}$ and $\lambda_{1\text{se}}$ for scalar outcomes Y^c and Y^b, while 10-fold cross-validated log-likelihood deviance was used for survival outcomes $(\tilde{\boldsymbol{T}}, \boldsymbol{\Delta})$ and $(\tilde{\boldsymbol{T}}^c, \boldsymbol{\Delta}^c)$. Features with non-zero coefficients at chosen values of penalty parameter were regarded as selected features. For `FLORAL`, remaining features after the Step 2 variable selection were used for method assessment.

For the DA methods, we largely applied the methods with their default configurations as detailed in **Table S2**. In addition, the Benjamini-Hochberg approach was applied for p-value adjustments if applicable, where taxa with adjusted p-values smaller than 0.05 were defined as selected features. Scalar outcomes $\boldsymbol{Y}^c$ or $\boldsymbol{Y}^b$ were treated as covariates for the DA methods. We did not test `ALDEx2`, `LEfSe`, nor the Wilcoxon test for continuous outcomes $\boldsymbol{Y}^c$ due to incompatibility. For time-to-event outcome $(\tilde{\boldsymbol{T}}, \boldsymbol{\Delta})$ or $(\tilde{\boldsymbol{T}}^c, \boldsymbol{\Delta}^c)$, we used the Martingale residual as the covariate for `LDM`, while the censoring indicator $\boldsymbol{\Delta}$ or $I(\boldsymbol{\Delta}^c = 1)$ was used as the patient group indicator for other methods. Detailed versions of R packages used for each method are listed in **Table S2**.

We focused on evaluating the variable selection performance for each method. Given the knowledge of the ten underlying truly associated features, we summarized the number of false negatives (FN, ranging between 0 and 10), false positives (FP, ranging between 0 and $p-10$), and the F_1 score $2\text{TP}/(2\text{TP}+\text{FP}+\text{FN})$ (ranging between 0 and 1), for each method at each simulation run, where TP represents the number of true positives. Here, a smaller FN indicates better sensitivity, while a smaller FP indicates better specificity of the methods. Similarly, a higher F_1 score implies a better balance between precision and recall. To better visualize the simulation results, heatmaps for median F_1, median FN, and median FP were generated with colors scaled for each simulation scenario. Simulation results from `corncob` was omitted in the figures as we observed zero features being selected in all simulation scenarios, which implies that the data generating model does not satisfy `corncob`'s model assumption.

4.8 Real Data Applications

4.8.1 Publicly Available Datasets for Two-Sample Comparison

Publicly available 16S rRNA sequencing datasets from 39 studies [32–67] were retrieved from the online data repository [78]. We applied the same naming system as used by Nearing et al. [27] to annotate the datasets, which are presented in **Fig.3A**. For each dataset, sequencing counts from different amplicon sequencing variants (ASVs) but the same genus were aggregated to form the genera count table for subsequent analysis. All counts were included without pre-filtering. Appropriate transformations were applied if applicable to obtain data formats suitable for different methods. For linear regression model (LM) and Wilcoxon test, we used relative abundance data due to their better simulation performances observed over centered log-ratio transformed data. The binary group identity is treated as the outcome variable for the lasso-based methods and the covariate variable for the DA methods. To evaluate the false positive control of different methods, we additionally utilized randomly shuffled binary group identities in the analysis.

Similar to the simulations, we applied method configurations and feature selection criteria listed in **Table S2** to perform genera selection with the original binary group labels and the randomly shuffled labels. Same random fold splits were applied to different lasso-based methods. Selected genera and total running time were collected. For each selected genus from each method using the true binary labels, we calculated the area under the ROC curve (AUC) with respect to the true binary groups.

We assess the false positive control of various methods based on the selected number of taxa using the randomly shuffled labels. Due to random shuffling, no taxa are expected to be detected as associated with the groups. Thus, any selected features can be treated as false positive findings, where the percentage of selected genera can be interpreted as the false positive rate. To visualize the results, a heatmap was produced with colors representing the false positive rates for each dataset for each method. For selected taxa based on the true binary labels, we generate heatmaps to compare the number of selected taxa, the median taxon-specific AUC, and the running time as descriptive metrics. Due to the lack of gold standard genera for each study, no inferences were made about the sensitivity of the methods.

4.8.2 MSKCC allo-HCT Cohort

The 16S rRNA microbiome sequencing dataset of MSKCC patients receiving first allo-HCT between January 2009 and June 2021 was utilized to investigate the associations between genera and survival outcomes. The patient and fecal sample cohort has been partly described in past studies [29, 31, 68, 79], while the more recent samples between 2018 and 2021 were also included in analyses reported in this work. Detailed descriptions on sample collection and storage, DNA extraction, and bioinformatic pre-processing pipelines have been made available [31, 79]. Samples with sequencing depth < 5000 were excluded from the analysis. ASVs from the same genus were combined at genus level for subsequent analysis.

Two analysis cohorts were derived as illustrated in **Fig.S9**. We defined day 0 as the date of HCT. The peri-engraftment sample cohort consisted of the latest samples collected between day 7 and 21 relative to HCT for the 912 patients who had at least one sample collected between day 7 and 21. In contrast, the longitudinal sample cohort contained 8,967 samples from 1,415 patients, including the last sample collected prior to HCT and all samples post HCT for each patient.

Three survival endpoints of interest were defined, namely overall survival (OS), transplant-related survival (TRM), and graft-versus-host disease (GVHD)-related survival (GRM). Patients were censored at the time of last contact or at the time of second transplant, whichever occurred earlier. For TRM and GRM, we followed the hierarchical definition of competing risks [4]. Specifically, TRM or GRM will be censored by the competing risk of relapse or progression of disease. Patients who did not have recorded relapse or progression time, but with death due to relapse or disease progression, were also classified as having the endpoint of relapse or progression. For patients who did not experience relapse and progression, and also did not die due to relapse and progression, the causes of death would determine TRM and GRM. Here, TRM consists of all causes of death apart from relapse and progression, while GRM is a subset of TRM where patients died from GVHD or died after having GVHD. For the analysis associated with the peri-engraftment sample cohort, the time-to-event is landmarked at the sample collection time of the peri-engraftment sample. For the longitudinal sample cohort, the time origin is set as the time of transplant, while patients will enter the risk set at time 0 or the time of collection of the first stool sample, whichever happened earlier.

FLORAL was applied to investigate the association between genera and the survival endpoints defined above, adjusted for age, conditioning intensity, graft source, and disease type, using both the peri-engraftment sample cohort and the longitudinal sample cohort. The longitudinal microbial features were treated as time-dependent covariates, under the last-value-carried-forward assumption [71]. Cox proportional hazards model was applied for the OS, while Fine-Gray subdistributional proportional hazards model was applied for TRM and GRM. To assess how reliably FLORAL select microbial features using peri-engraftment samples versus longitudinal samples, the two-step variable selection procedure was repeated for 100 times under randomly generated 5-fold cross validation splits. For each survival endpoint, the percentages of times being selected using $\lambda = \lambda_{1\text{se}}$ out of 100 repeated runs were compared across the peri-engraftment and the longitudinal cohorts for taxa selected at least once.

Other methods listed in **Table S2** were also applied for feature selection for OS. For lasso-based methods, the same 100 5-fold splits used for FLORAL were used to generate taxa selection probabilities for `glmnet` and `zeroSum`, where `glmnet` with relative abundance, log-transformed counts, and centered log-ratio transformed counts were applied for both peri-engratment and longitudinal cohorts, while `zeroSum` was only applied for the peri-engraftment cohort due to its incompatibility with time-dependent covariates. Using the OS indicators as patient group labels, the DA methods were also applied to select differentially abundant genera across the two groups with the configurations listed in **Table S2**.

5 Data and Code Availability

Open-source R package FLORAL can be accessed via GitHub (https://vdblab.github.io/FLORAL) or CRAN (https://cran.r-project.org/package=FLORAL). R scripts used for analyses can be accessed via GitHub (https://github.com/vdblab/FLORAL-analysis/). 16S rRNA sequencing datasets for the 39 studies were retrieved from https://figshare.com/articles/dataset/16S_rRNA_Microbiome_Datasets/14531724 [78]. 16S rRNA sequencing dataset for the MSKCC allo-HCT cohort can be downloaded from https://doi.org/10.6084/m9.figshare.13584986 [79].

Algorithm 1 Iterative optimization algorithm for (8) with given λ and μ. Note that the following algorithm assumes no intercept term. The algorithm with intercept term can be derived similarly. $\odot$ denotes element-wise multiplication.

Input: Initial value of $\hat{\boldsymbol{\zeta}} = \tilde{\boldsymbol{\zeta}} = (\tilde{\boldsymbol{\beta}}^T, \tilde{\boldsymbol{\omega}}^T)^T$; $n \times (p+L)$ matrix $\boldsymbol{Z}$; parameters λ, μ; tolerance parameter δ, δ'; maximum inner iteration number u, u'

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

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

Set $i = i + 1$

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

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

Set $j = j+1$, idx $=$ which$(\tilde{\boldsymbol{\zeta}} > 0)$. Initialize $\check{\boldsymbol{\zeta}}$. 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 $k = 1, \ldots, p+L$ **do**

if ζ_k is an element of $\boldsymbol{\beta}$ **then**

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

end if

if ζ_k is an element of $\boldsymbol{\omega}$ **then**

Update $\check{\zeta}_k = \frac{1}{\boldsymbol{A}_{k,k}} S_\xi \left\{ \frac{1}{n}(\boldsymbol{B}_k - A_{k,\text{idx}} \odot \tilde{\boldsymbol{\zeta}}_{\text{idx}} + A_{k,k} \odot \tilde{\zeta}_k) \right\}$

end if

end for

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

end while

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

end while

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