Title
Spatiotemporal Modeling of Microbial Communities

Author
Shenhav, Liat

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Spatiotemporal Modeling of Microbial Communities

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Computer Science

by

Liat Shenhav

2020

ABSTRACT OF THE DISSERTATION

Spatiotemporal Modeling of

Microbial Communities

by

Liat Shenhav

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2020

Professor Eran Halperin, Chair

Microbial communities can undergo rapid changes, that can both cause and indicate host disease, rendering longitudinal microbiome studies key for understanding microbiome-associated disorders. However, most standard statistical methods, based on random samples, are not applicable for addressing the methodological and statistical challenges associated with repeated, structured observations of a complex ecosystem. Therefore, to elucidate how and why our microbiome varies in time, and whether these trajectories are consistent across humans, we developed new methods for modeling the temporal and spatial dynamics of microbial communities. We developed a method to identify 'time-dependent' microbes (Shenhav et al., PLoS Computational Biology 2019) and showed that their temporal patterns differentiate between the developing microbial communities of infants and those of adults. We also developed models to deconvolute the dynamics of microbial community formation. Using these methods, we found significant differences between vaginally- and cesarean-delivered infants in terms of initial colonization and succession of their gut microbial community (Shenhav et al., Nature Methods 2019) as well as the trajectories of these communities in the first years of life (Martino*, Shenhav* et al., Nature Biotechnology). These models, designed to identify and predict time-dependent patterns, will help researchers better understand the temporal nature of the human microbiome from the time of its formation at birth and throughout life.

The dissertation of Liat Shenhav is approved.

Eleazar Eskin

Sriram Sankararaman

Nandita Garud

Eran Halperin, Committee Chair

University of California, Los Angeles

2020

Table of content

List of Figures and Tables

Acknowledgements

I would first and foremost want to thank my advisor, Prof. Eran Halperin, for the hours he spent in teaching the secrets of the trade, and in brainstorming, planning and executing amazing research. For his ability to move mountains to do great science, and for doing this so seamlessly. And for the tenacity, determination and confidence that make such a winner.

The work presented here would not have been possible without the amazing collaboration with Prof. Itzik Mizrahi and Dr. Ori Furman who introduced me to the fascinating world of microbiome science and worked hand in hand with us in developing these methods and algorithms. I would also like to thank Prof. Itsik pe'er for his genius and patience and for his support throughout my PhD, and to Tyler Joseph for his collaborative spirit and brilliant ideas.

Special recognition should go Cameron Martino for a fruitful collaboration and partnership. And to Mike Thompson, Leah Briscoe and Ulzee An for their work and contribution.

Thanks also to my friends and colleagues from the Halperin lab: Regev, Elior, Nadav, Brian, Brandon, Johnson, Jeff, and Misagh. It was a great pleasure working alongside you.

Finally, I want to thank my family, for providing me with the tools and environment necessary to succeed and for not seeming bored when I go on about some algorithm or problem.

And to the love of my life, Raphael, Jenny and Judith, I could not have done this without you.

Vita/Biographical sketch

Education

M.Sc. in statistics, Tel-Aviv University, Israel. 2014 – 2016

Advisors: Prof. Yoav Benjamini and Prof. Ruth Heller.

Graduated magna cum laude.

B.Sc. in Mathematics, Tel-Aviv University, Israel. 2010 – 2014

Honors, awards and scholarships

2017 - 2020	Doctoral Fellowship, Computer Science Department University of California, Los Angeles.
2016/17	Fellow, Edmond J. Safra Program in Bioinformatics Tel-Aviv University.
2016	Received magna cum laude honors (M.Sc.) Faculty of Exact Sciences, Tel-Aviv University.
2015	Certificate of Excellence, Teaching Faculty of Exact Sciences, Tel-Aviv University.

Publications

*equal contribution

Published/accepted

1. Liat Shenhav*, David Zeevi*, Resource conservation manifests in the genetic code, bioRxiv, doi: https://doi.org/10.1101/790345, *Science, in press.*
2. Cameron Martino*, Liat Shenhav*, George Armstrong, Daniel McDonald, Yoshiki Vázquez-Baeza, James T Morton, Lingjing Jiang, Austin D Swafford, Eran Halperin, Rob Knight, Context-aware dimensionality reduction deconvolutes dynamics of gut microbial community development, *Nature Biotechnology, 2020.*
3. Ori Furman*, Liat Shenhav*, Goor Sasson, Fotini Kokou, Hen Honig, Shamay Jacoby, Tomer Hertz, Otto X Cordero, Eran Halperin, Itzhak Mizrahi, Stochasticity framed by deterministic effects

of diet and age drive rumen microbiome assembly dynamics, *Nature communications 11(1),1-13, 2020.*

4. Tyler Joseph, Liat Shenhav, Joao B Xavier, Eran Halperin, Itsik Pe'er, Compositional Lotka-Volterra describes microbial dynamics in the simplex, *PLoS Computational Biology 16(5), 2020.*
5. Liat Shenhav, Mike Thompson, Tyler A Joseph, Leah Briscoe, Ori Furman, David Bogumil, Itzhak Mizrahi, Itsik Pe'er, Eran Halperin, FEAST: fast expectation-maximization for microbial source tracking, *Nature Methods 16 (7), 627-632 , 2019.*
6. Liat Shenhav*, Ori Furman*, Leah Briscoe, Mike Thompson, Justin D Silverman, Itzhak Mizrahi, Eran Halperin, Modeling the temporal dynamics of the gut microbial community in adults and infants, *PLoS Computational Biology 15(6), 2019.*
7. Justin D. Silverman*, Liat Shenhav*, Eran Halperin, Sayan A Mukherjee, Lawrence A David, Statistical considerations in the design and analysis of longitudinal microbiome studies, *bioRxiv, doi: https://doi.org/10.1101/448332*
8. Iman Jaljuli, Yoav Benjamini, Liat Shenhav, Orestis Panagiotou, Ruth Heller, Quantifying replicability and consistency in systematic reviews, *arXiv:1907.06856, 2019.*
9. Elior Rahmani, Regev Schweiger, Liat Shenhav, Theodora Wingert, Ira Hofer, Eilon Gabel, Eleazar Eskin, Eran Halperin, BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference, *Genome biology, 19(1), 2019.*
10. Regev Schweiger, Eyal Fisher, Elior Rahmani, Liat Shenhav, Saharon Rosset, Eran Halperin, Using stochastic approximation techniques to efficiently construct confidence intervals for heritability, *Journal of Computational Biology 25(7), 2018.*
11. Elior Rahmani, Liat Shenhav, Regev Schweiger, Paul Yousefi, Karen Huen, Brenda Eskenazi, Celeste Eng, Scott Huntsman, Donglei Hu, Joshua Galanter, Sam S Oh, Melanie Waldenberger, Konstantin Strauch, Harald Grallert, Thomas Meitinger, Christian Gieger, Nina Holland, Esteban G Burchard, Noah Zaitlen, Eran Halperin, Genome-wide methylation data mirror ancestry information, *Epigenetics and Chromatin 10(1), 2017.*
12. Elior Rahmani, Regev Schweiger, Liat Shenhav, Eleazar Eskin, Eran Halperin, A Bayesian framework for estimating cell type composition from DNA methylation without the need for methylation reference, *International Conference on Research in Computational Molecular Biology (RECOMB), 2017.*
13. Elior Rahmani, Reut Yedidim, Liat Shenhav, Regev Schweiger, Omer Weissbrod, Noah Zaitlen, Eran Halperin, GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data, *Bioinformatics, 33 (12) 1870-1872, 2017.*
14. Zohar Meiri, Shay Keren, Amir Rosenblatt, Tal Sarig, Liat Shenhav, David Varssano, Efficacy of corneal collagen cross-linking for the treatment of keratoconus: a systematic review and meta-analysis, *Cornea 35(3), 2016.*
15. Liat Shenhav, Ruth Heller, Yoav Benjamini, Quantifying replicability in systematic reviews: the r-value, a*rXiv:1502.00088, 2015.*

Introduction

There is increasing recognition that the human microbiome is a contributor to many aspects of human physiology and health including obesity, non-alcoholic fatty liver disease, inflammatory diseases, cancer, metabolic diseases, aging, and neurodegenerative disorders[1–9]. This suggests that the human microbiome may play important roles in the diagnosis, treatment, and ultimately prevention of human disease. These applications require an understanding of the temporal variability of the microbiota over the lifespan of an individual particularly since we now recognize that our microbiota is highly dynamic, and that the mechanisms underlying these changes are linked to ecological resilience and host health[10–12].

Due to the lack of data and insufficient methodology, we currently have major gaps in our understanding of fundamental mechanisms related to the temporal behavior of the microbiome. Critically, we currently do not have a clear characterization of how and why our microbiome varies in time, and whether these dynamics are consistent across humans. It is also unclear whether we can define ‘stable’ or ‘healthy’ dynamics as opposed to ‘abnormal’ or ‘unhealthy’ dynamics, which could potentially reflect an underlying health condition or an environmental factor affecting the individual, such as antibiotics exposure or diet. Moreover, there is no consensus as to whether the microbial community structure varies continuously or jumps between discrete community states, and whether or not these states are shared across individuals [13,14].

The need for understanding the temporal dynamics of the microbiome and its interaction with host attributes have led to a rise in longitudinal studies that record the temporal variation of microbial communities in a wide range of environments, including the human gut microbiome. These time series studies are enabling increasingly comprehensive analyses of how the microbiome changes over time, which are in turn beginning to provide insights into fundamental questions about microbiome dynamics[10,11,15].

One of the most fundamental questions that still remains unanswered is to what degree a microbial community deterministically depends on its initial composition (e.g., microbial composition at birth). More generally, it is unknown to what degree the microbial composition at a given time determines the microbial composition at a later time. Additionally, there is only preliminary evidence of the long-term effects of early life events on the gut microbial community composition, and it is currently unclear whether these long-term effects traverse through a predefined set of potential trajectories [15,16].

To address these questions, we first developed a mixed effects model with variance components for the prediction and inference of microbial community temporal dynamics (Shenhav et al., PLoS Computational Biology 2019). This model identifies 'time-dependent' microbes and infers their trajectory over time. Using this model, we characterized temporal patterns differentiating between the developing microbial communities of infants and those of adults. We further showed that closely related species interact with similar partners, a known phylogenetic structure, indicating ecological interactions are phylogenetically conserved[17]. In parallel, I took part in a collaborative project (Furman*, Shenhav* et al., Nature Communications, 2020) in which I used this method to uncover latent temporal patterns that provided data-driven support to ecological theories of microbial community assembly.

In a second project, we sought to directly identify the meaningful, low-rank microbial community dynamics that differentiate host phenotypes. To this end, we developed a multi-dimensional decomposition algorithm which factorizes a tensor into a linear combination of its components (e.g., subjects, microbial abundances and time) and allows clustering of phenotypes by microbial community dynamics (Martino*, Shenhav* et al., Nature Biotechnology, 2020). We demonstrated the utility of this method by identifying microbial niches unique to different modes of delivery (vaginal vs. cesarean delivery).

Finally, in a third project, we sought to quantify the origins of complex microbial communities. We modeled a given microbial community as a mixture of potential ecosystems, and developed a scalable deconvolution method for "microbial source tracking" (Shenhav et al., Nature Methods 2019)[18]. Our method provides a computationally efficient tool that can simultaneously evaluate hundreds to thousands of potential source

environments, as well as the contribution of an unknown, uncharacterized source, and outperforms state-of-the-art methods. We demonstrated the utility of our method by modeling initial colonization and succession in the gut microbiome of infants and revealed a significantly larger maternal contribution in vaginally-delivered infants over cesarean-delivered ones.

Overall, this thesis depicts the development of tools for microbiome analysis and the application of these tools to analyze the human microbiome in clinical settings. Predicting the temporal dynamics of the microbiome, as well as extracting its low-rank representation presents new opportunities in harnessing this vast ecosystem to study and diagnose human diseases.

Methods

Modeling the temporal dynamics of the gut microbial community in adults and infants

MTV-LMM algorithm

MTV-LMM uses a linear mixed model (see [19] for a detailed review), a natural extension of standard linear regression, for the prediction of time series data. We describe the technical details of the linear mixed model below. We assume that the relative abundance levels of focal taxa j at time point t depend on a linear combination of the relative abundance levels of the microbial community at previous time points. We further assume that temporal changes in relative abundance levels, in taxa j, are a time-homogeneous high-order Markov process. We model the transitions of this Markov process using a linear mixed model, where we fit the p previous time points of taxa j as fixed effects and the q previous time points of the rest of the microbial community as random effects. p and q are the temporal parameters of the model. For simplicity of exposition, we present the generative linear mixed model that motivates the approach taken in MTV-LMM in two steps. In the first step we model the microbial dynamics in one individual host. In the second step we extend our model to N individuals, while accounting for the hosts' effect.

We first describe the model assuming there is only one individual. Consider a microbial community of m taxa measured at T equally spaced time points. We get as input an $m \times T$ matrix M, where M_{jt} represents

the relative-abundance levels of taxa j at time point t. Let $y_j = (M_{j(p+1)}, \ldots, M_{jt})$ t be a (T − p) × 1 vector of taxa j relative abundance, across T − p time points starting at time point p + 1 and ending at time point T. Let X_j be a (T − p) × (p + 1) matrix of p + 1 covariates, comprised of an intercept vector as well as the first p time lags of taxa j (i.e., the relative abundance of taxa j in the p time points prior to the one predicted). Formally, for k = 1 we have ${X_{tk}}^j = 1$, and for $1 < k \leq p+1$we have ${X_{tk}}^j = M_{j,t-k+1}$ for $t \geq k$. For simplicity of exposition and to minimize the notation complexity, we assume for now that p = 1. Let W be an $(T-q) \times qm$ normalized relative abundance matrix, representing the first q time lags of the microbial community. For simplicity of exposition we describe the model in the case q = 1, and then $W_{tj} = M_{jt}$(in the more general case, we have $W_{tj} = M_{j/q,t-(j \bmod q)}$, where p, q ≤ $T-1$).

With these notations, we assume the following linear model:

$$y^j = X^j\beta^j + Wu^j + \epsilon^j, (1)$$

where u^j and ϵ^j are independent random variables distributed as $u^j \sim N(0_m, {\sigma^2}_{u^j} I_m)$ and $\epsilon^j \sim N(0_m, {\sigma^2}_{\epsilon^j} I_{T-1})$.The parameters of the model are β_j(fixed effects), ${\sigma^2}_{u^j}$, and ${\sigma^2}_{\epsilon^j}$. We note that environmental factors known to be correlated with taxa abundance levels (e.g., diet, antibiotic usage [11,20]) can be added to the model as fixed linear effects (i.e., added to the matrix X_j).

Given the high variability in the relative abundance levels, along with our desire to efficiently capture the effects of multiple taxa in the microbial community on each focal taxa j, we represent the microbial community input data (matrix M) using its quantiles. Intuitively, we would like to capture the information as to whether a taxa is present or absent, or potentially introduce a few levels (i.e., high, medium, and low abundance). To this end, we use the quantiles of each taxa to transform the matrix M into a matrix $\widetilde{M}$, where $\widetilde{M} \in \{0,1,2\}$ depending on whether the abundance level is low (below 25% quantile), medium, or high (above 75% quantile). We also tried other normalization strategies, including quantile normalization, which is typically used in gene expression eQTL analysis [21,22], and the results were qualitatively similar (see

Appendix I Fig. S6). We subsequently replace the matrix W by a matrix $\widetilde{W}$, which is constructed analogously to W, but using $\widetilde{M}$ instead of M. Notably, both the fixed effect (the relative abundance of y_j at previous time points) and the output of MTV-LMM are the continuous relative abundance. The random effects are quantile-binned relative abundance of the rest of the microbial community at previous time points (matrix $\widetilde{W}$). Thus, our model can now be described as

$$y^j = X^j\beta^j + \widetilde{W}u^j + \epsilon^j, (2)$$

So far, we described the model assuming we have time series data from one individual. We next extend the model to the case where time series data is available from multiple individuals. In this case, we assume that the relative abundance levels of m taxa, denoted as the microbial community, have been measured at T time points across N individuals. We assume the input consists of N matrices,$M^1, \ldots, M^N$, where matrix M^i corresponds to individual i, and it is of size m × T. Therefore, the outcome vector y_j is now an n × 1 vector, composed of N blocks, where n = (T − 1)N, and block i corresponds to the time points of individual i. Formally, $y_j = M_{j,(k \bmod (T-1))}{}^{k/(T-1)}$. Similarly, we define X_j and $\widetilde{W}$ as block matrices, with N different blocks, where corresponds to individual i.

When applied to multiple individuals, Model (2) may overfit to the individual effects (e.g., due to the host genetics and or environment). In other words, since our goal is to model the changes in time, we need to condition these changes in time on the individual effects that are unwanted confounders for our purposes. We therefore construct a matrix H by randomly permuting the rows of each block matrix i in $\widetilde{W}$, where the permutation is conducted only within the same individual. Formally, we apply permutation $\pi_i \in S_{T-1}$on the rows of each block matrix i, M^i, corresponding to individual i, where S_{T-1} is the set of all permutations of (T − 1) elements. In each π_i , we are simultaneously permuting the entire microbial community. Hence, matrix H corresponds to the data of each one of the individuals, but with no information about the time (since the data was shuffled across the different time points). With this addition, our final model is given by

$$y^j = X^j\beta^j + \widetilde{W}u^j + Hr + \epsilon^j, (3)$$

where $u^j \sim N(0_m, {\sigma^2}_{u^j} I_m)$, $\epsilon^j \sim N(0_m, {\sigma^2}_{\epsilon^j} I_{T-1})$ and $r \sim N(0_m, {\sigma^2}_r I_m)$.

It is easy to verify that an equivalent mathematical representation of model 3 can be given by

$$y^j \sim N(X^j \beta^j , {\sigma^2}_{AR^j} K_1 + {\sigma^2}_{ind^j} K_2 + {\sigma^2}_{\epsilon^j} I), \qquad (4)$$

where ${\sigma^2}_{AR^j} = m{\sigma^2}_{u^j}$,$K_1 = \frac{1}{m} \widetilde{W}\, \widetilde{W}^T$,${\sigma^2}_{ind^j} = m{\sigma^2}_r$, $K_2 = \frac{1}{m} H\, H^T$. We will refer to K_1 as the temporal kinship matrix, which represents the similarity between every pair of samples across time (i.e., represents the cross-correlation structure of the data).

We note that for the simplicity of exposition, we assumed so far that each sample has the same number of time points T, however in practice the number of samples may vary between the different individuals. It is easy to extend the above model to the case where individual i has T_i time points, however the notations become cumbersome; the implementation of MTV-LMM, however takes into account a variable number of time points across the different individuals. Once the distribution of y_j is specified, one can proceed to estimate the fixed effects β_j and the variance of the random effects using maximum likelihood approaches. One common approach for estimating variance components is known as restricted maximum likelihood (REML). We followed the procedure described in the GCTA software package [23], under 'GREML analysis', originally developed for genotype data, and re-purposed it for longitudinal microbiome data. GCTA implements the restricted maximum likelihood method via the average information (AI) algorithm. Specifically, we performed a restricted maximum likelihood analysis using the function "–reml" followed by the option "–mgrm" (reflects multiple variance components) to estimate the variance explained by the microbial community at previous time points. To predict the random effects by the BLUP (best linear unbiased prediction) method we use "–reml-predrand". This option is actually to predict the total temporal effect (called "breeding value" in animal genetics) of each time point attributed by the aggregated effect of the taxa used to estimate the temporal kinship matrix. In both functions, to represent y_j (the abundance of taxa j at the next time point), we use the option "–pheno".

Time-explainability

We define the term time-explainability, denoted as χ, to be the temporal variance explained by the microbial community in the previous time points. Formally, for taxa j we define

$$\chi_j = \frac{{\sigma^2}_{AR^j}}{{\sigma^2}_{AR^j} + {\sigma^2}_{ind^j} + {\sigma^2}_{\epsilon^j}}$$

The time-explainability was estimated with GCTA, using the temporal kinship matrix. In order to measure the accuracy of time-explainability estimation, the average confidence interval width was estimated by computing the confidence interval widths for all autoregressive taxa and averaging the results. Additionally, we adjust the time-explainability P-values for multiple comparisons using the Benjamini-Hochberg method [24].

Best linear unbiased predictor

We now turn to the task of predicting ${y_t}^j$ using the taxa abundance in time t − 1 (or more generally in the last few time points). Using our model notation, we are given x_j and $\widetilde{w}$, the covariates associated with a newly observed time point t in taxa j, and we would like to predict ${y_t}^j$ with the greatest possible accuracy. For a simple linear regression model, the answer is simply taking the covariate vector x and multiplying it by the estimated coefficients $\widehat{\beta}: {y_t}^j = x^T\hat{\beta}$. This practice yields unbiased estimates. However, when attempting prediction in the linear mixed model case, things are not so simple. One could adopt the same approach, but since the effects of the random components are not directly estimated, the vector of covariates w~ will not contribute directly to the predicted value of ${y_t}^j$, and will only affect the variance of the prediction, resulting in an unbiased but inefficient estimate. Instead, one can use the correlation between the realized values of $\widetilde{Wu}$, to attempt a better guess at the realization of $\widetilde{wu}$for the new sample. This is achieved by computing the distribution of the outcome of the new sample conditional on the full dataset, by using the following property of the multivariate normal distribution. Assume we sampled t − 1 time points from taxa j, but the relative abundance level for the next time point t, ${y_t}^j$, is held out from the

algorithm. The conditional distribution of ${y_t}^j$ given the relative abundance levels at all previous time points, y^j, is given by

$$ {y_t}^j | \, y^j \sim N(x^T \beta^j \; + \; \Sigma_{t,-t} {\Sigma_{-t,-t}}^{-1} (y^j - X^j \beta^j), \Sigma_{t,-t} {\Sigma_{-t,-t}}^{-1} \Sigma_{-t,-t}), \qquad (5) $$

where $\Sigma = \widetilde{W} \, \widetilde{W}^T {\sigma_{u^j}}^2 \; + \; HH^T \sigma_r^2 \; + \; I {\sigma^2}_{\epsilon^j}$ and positive/negative indices indicate the extraction/ removal of rows or columns, respectively. Intuitively, we use information from the previous time points that have a high correlation with the new time point, to improve its prediction accuracy. The practice of using the conditional distribution is known as BLUP (Best Linear Unbiased Predictor). Therefore, MTV-LMM could be used to learn taxa effects in a train set (taxa abundance at time points 1, . . ., t), and subsequently use these learned taxa effects to predict the temporal-community contribution in the next time point in a test set (taxa j at t + 1). We will define the association matrix U (m × m) using BLUP, where u_{ij} is the effect of taxa i on taxa j.

Prediction Accuracy

The predictive ability of a model is commonly assessed using the prediction error variance, $PEV = Var(y^j - \hat{y}^j)$, where $\hat{y}^j$ is the Best Linear Unbiased Predictor of y^j. The proportional reduction in relative abundance variance accounted for by the predictions (referred to as R^2 in this paper) can be quantified using

$$ R^2 \; = \frac{Var(y^j) \; - \; Var(\hat{y}^j)}{Var(y^j)} \; = \frac{Cov(y^j, \hat{y}^j)^2}{Var(y^j) Var(\hat{y}^j)} $$

Notably, this definition is equivalent to the squared Pearson correlation. For every $t \; \in \{p+1, \ldots, T\}$, we calculate ${\hat{y}^j}_t$, where $p \; \geq q$ and the microbial community composition at time t was held out from the algorithm. We next compute R^2 between ${y^j}_{\{p+1,\ldots,T\}}$ and ${\hat{y}^j}_{\{p+1,\ldots,T\}}$.

Model selection

Given that the model presented in Eq (3) can be extended to any arbitrary p and q, we tested four different variations of this model: 1. $p = 0$ and $q = 1$ (no fixed effect, random effects based on 1-time lag), 2. $p = 1$ and $q = 1$ (one fixed effect based on 1-time lag, random effects based on 1-time lag), 3. $p = 0$ and $q = 3$ (no fixed effect, random effects based on 3-time lags) and 4. $p = 1$ and $q = 3$ (one fixed effect based on 1-time lag, random effects based on 3-time lags). We divide each dataset into three parts—training, validation, and test, where each part is approximately 1/3 of the time series (sequentially). We train all four models presented above and use the validation set to select a model for each taxa j based on the highest correlation with the observed relative abundance. We then compute sequential out-of-sample predictions on the test set with the selected model. Based on this metric, we found $p = 1$ and $q = 1$ to be the best model for most taxa. We use these parameters when comparing with the other methods such as sVAR and ARIMA-Poisson. There are three main justifications for the use of multiple time points in the model. First, Gibbons et al. [25] empirically performed a time-lag analysis and found that for most taxa the autocorrelation disappeared after 3 or 4 days, whereas for some taxa the autocorrelation disappeared after 1 or 2 days. Second, previous studies [26–28] found that the human microbiome reaches equilibrium within 10 days following small perturbations to the community. It is imperative to model the different taxa in a manner that will fit their temporal patterns. Third, allowing for the use of multiple previous time points increases flexibility so that the model can select the correct time window required for each taxa.

Phylogenetic analysis

We performed the following phylogenetic analysis. First, in order to test the hypothesis that both autoregressive and non-autoregressive dynamics carry a taxonomic signal, we fitted a linear mixed model, where the kinship matrix is now the phylogenetic distance between pairs of taxa and the outcomes are the time-explainability measurement for each taxa. Second, in order to test the hypothesis that only non-autoregressive dynamics carry a non-random taxonomic signal, we conducted a permutation test by shuffling the taxonomic order assigned to each taxa—generating new random "orders" using 100, 000

iterations. We counted the number of non-autoregressive orders in each iteration, thereby generating a null distribution, which we then used to calculate an exact P-value for the dataset in each iteration.

Alpha diversity measures

To measure the alpha diversity, we used Shannon-Wiener index, which is defined as $H = -\sum p_j \, ln(p_j)$, where pj is the relative abundance of species j. Shannon-Wiener index accounts for both abundance and evenness of the species present. Additionally, we computed the 'effective number of species' (also known as true diversity), the number of equally-common species required to give a particular value of an index. The 'effective number of species' associated with a specific Shannon-Wiener index a is equal to $exp(a)$.

Preliminary taxa screening according to temporal presence-absence patterns

To calculate the temporal kinship matrix we included taxa using the following criteria. A taxa is present in at least 10% of the time points (removes dominant zero abundance taxa). In the David et al. dataset we included 1051 (out of 2804), in the Caporaso et al. dataset we included 922 (out of 3436) and in the DIABIMMUNE dataset we included 1440 (out of 7244) taxa.

Methods comparison

We compared MTV-LMM to two existing methods: sVAR suggested by [25] and Poisson regression suggested by [29]. In the sVAR method, we followed the procedure described in [25] , while running the model and computing the prediction for each individual separately, since it can only handle one individual at a time. We then computed an aggregated prediction accuracy score for each taxa, by averaging the prediction accuracy of each individual. In the Poisson regression method, we followed the procedure described in [29], while running the model for all the individuals simultaneously and calculating prediction accuracy for each taxa. We used the taxa that passed the screening suggested in [29] (eliminating any taxa in the data for which there were a small number (< 6) of average reads per sample). In both models, the training set was 0.67 of the data and the test set was the remaining 0.33 of the data. In both cases we used the code supplied by the authors.

Datasets

We evaluated the performance of MTV-LMM using three real longitudinal datasets with 16S rRNA gene sequencing. All data sets are publicly available. The first data set was collected and studied by David et al. (2014) [11](2 adult donors). The next data set was collected and studied by Caporaso et al. (2011) [10] (2 adult donors). The third data set was collected by the 'DIABIMMUNE' project and studied by Yassour et al. (2016) [15] (39 infant donors). In order to compare across studies and reduce technical variance between studies, closed reference OTUs were clustered at 99% identity against the Greengenes database 13_8 [30]. Open reference OTU picking was also run [31], in order to look for non-database OTUs that might contribute substantially to community dynamics. OTU tables were normalized by random sub-sampling to contain 10, 000 reads per sample. David et al. (2014) dataset [11]. Stool samples from 2 healthy American adults were collected (donor A = DA and donor B = DB). DA collected gut microbiota samples between days 0 and 364 of the study (total 311 samples). DB primarily collected gut microbiota samples between study days 0 and 252 (total 180 samples). The V4 region of the 16S ribosomal RNA gene subunit was used to identify bacteria in a culture-independent manner. DNA was amplified using custom barcoded primers and sequenced with paired-end 100 bp reads on an Illumina GAIIx according to a previously published protocol [32]. 'OTU picking' and 'quality control' were performed essentially as described [11]. In this work, we used the OTUs shared across donors (2, 804 OTUs). Caporaso et al. (2011) dataset [10] . Two healthy American adults, one male (M3) and one female (F4), were sampled daily at three body sites (gut (feces), mouth, and skin (left and right palms)). M3 was sampled for 15 months (total 332 samples) and F4 for 6 months (total 131 samples). Variable region 4 (V4) of 16S rRNA genes present in each community sample were amplified by PCR and subjected to multiplex sequencing on an Illumina Genome Analyzer IIx according to a previously published protocol [32]. 'OTU picking' and 'quality control' were performed essentially as described [10] . In this work, we used the OTUs shared across donors (3, 436 OTUs). DIABIMMUNE dataset [15]. Monthly stool samples collected from 39 Finnish infants aged 2 to 36 months. To analyze the composition of the microbial communities in this cohort, DNA from stool samples was isolated and amplified and the V4 region of the 16S rRNA gene was sequenced. Sequences were sorted into

OTUs. 16S rRNA gene sequencing was performed essentially as previously described in [15]. In this work, we used all the OTUs in the sample (7, 244 OTUs).

Code availability

Code is available in https://github.com/cozygene/MTV-LMM.

Context-aware dimensionality reduction deconvolutes gut microbial community dynamics

Preprocessing with robust-clr

Prior to running tensor factorization, we use the robust centered log-ratio transformation (robust-clr) to center the data around zero and approximate a normal distribution[33]

$$rclr(x) = \left[log\ log\ \frac{x_1}{g_r(x)}, \ldots, log\ log\ \frac{x_D}{g_r(x)}\right] \quad (1)$$

$$g_r(x) = \left(\prod_{i \in \Omega_x} x_i\right)^{1/|\Omega_x|} \quad (2)$$

where x_i is the abundance of microbe i, Ω_x is the set of observed microbes in sample x and $g_r(x)$ is the geometric mean only defined on microbes with abundance > 0. Unlike the traditional clr transformation, the robust-clr handles the high level of sparsity found in microbial datasets without requiring imputation. Furthermore, this transformation has shift invariant properties that allow the restructuring of the matrix into tensor form.

Tensor factorization via alternating least squares minimization

Here we follow the tensor notations of Lim[34] and Anandkumar et al. [35], for a full notation see the supplemental methods. To perform tensor factorization on sparse data we followed a procedure introduced by Jain and Oh[36]. Due to the high level of sparsity in microbiome datasets we would like to find the

minimum rank representation of T that best explains only *observed* values defined as Ω. We use the projection $P_\Omega(T)_{ijt}$

$$P_\Omega(T)_{ijt} = f(x) = \{T_{ij}, \quad if(i,j,t) \in \Omega \; 0, \quad Otherwise \quad (3)$$

The objective function being optimized through alternating least squares minimization (ALS) is given by

$$\| P_\Omega(T) - P_\Omega\left(\sum_{i=1}^{r} \sigma_i(a_i \otimes b_i \otimes c_i)\right) \|_F^2 \quad (4)$$

where a, b, and c are unstructured, orthogonal, and have a Euclidean norm of 1. The low rank representations a, b, and c correspond to loadings for the first, second and third tensor modes respectively. It is important to note that this factorization is permutation invariant, meaning the order of time or space is not a factor in the subsequent loadings of c.

Factorization trajectories

Here, we focus on the interpretation of tensor factorization for biological data. We are primarily concerned with 3rd-order tensors from studies following multiple subjects over several timepoints. In this tensor the first mode is the subjects or environments sampled. The second mode is biological features such as microbes, metabolites, or genes. The third mode is timepoints where subjects/environments were sampled repeatedly. Of utmost interest is the relation between subject or features and the third mode of time. To obtain easily interpretable loadings we introduce trajectories given by

$$Subject\ Trajectory = a \odot c = [a_1 \otimes c_1, \ldots, a_r \otimes c_r] \in R^{d^2 \times r}$$

$$Feature\ Trajectory = b \odot c = [b_1 \otimes c_1, \ldots, b_r \otimes c_r] \in R^{d^2 \times r}$$

where $\odot$ represents the Khatri-Rao product. These trajectories are of the shape (subjects $\times$ time, rank) or (features $\times$ time, rank) where each rank-1 column has an accompanying singular value σ_r.

Log-ratio feature selection

In order to explore how feature rankings in b or b $\odot$ c partitioned subjects we used log-ratios between highly (positive) and lowly (negative) ranked features along the first axis of variation. To avoid the use of pseudo-counts we explore the sum of the minimum number of highly and lowly ranked features summed across all samples, such that no log-ratio contains a zero value. For ECAM 1400 and DIABIMMUNE 750 total features were used and split between numerator and denominator evenly such that no samples were dropped due to zero values (Appendix II Fig S5). We then used a Linear Mixed Effects (LME) model via statsmodels (v. 0.11.0) to test if the log-ratio changed over time and in response to birth mode for ECAM and DIABIMMUNE separately. The LME model produced residual R^2 values of 0.976 and 0.986 for DIABIMMUNE and ECAM respectively. The resulting p-values from the LME were significant ($P < .05$) by birth mode, time in days, and the interaction of the two (Appendix II Table S4). To produce the microbial birth-mode signature, we used only sequences shared among ECAM, DIABIMMUNE, and the American Gut Project (1,064 features total). We used the ranking structure inferred from ECAM and DIABIMMUNE to evenly divide these shared features into vaginal or cesarean-associated taxa (532 each in the numerator and denominator, respectively). A t-test via SciPy (v. 1.4.1) was used on the microbial birth-mode signature (i.e., log-ratio) to test for significance between birth modes stratified by age or time point for both data sets, respectively.

Data driven simulation benchmarks

Data driven simulations were designed to benchmark different characteristics of data without making assumptions about microbial dynamics. The IBD dataset was chosen due to its high temporal resolution and two-group (low-rank) comparison. Simulations were generated using a procedure from Äijö et al.[37] modified to use a Poisson-lognormal distribution (PLN) [38] as opposed to a Poisson-Multinomial

distribution. This simulation was repeated for different levels of dispersion, subsampling (i.e. sparsity), sampling density (i.e. number of timepoints) and percentage of randomly missing samples.

Case Study Sequence Processing

Raw sequences were quality controlled, trimmed at 100 nucleotides, and clustered as amplicon sequence variants (sOTUs) using QIIME 2 release 2019.7 and Deblur (v. 1.1.0)[39]. The phylogenetic tree was created using SEPP sequence insertion with the Greengenes tree 13.8 release as the reference tree[30,40]. Taxonomy assignments were made using a Naive Bayes classifier as implemented in QIIME2 (v. 2019.7). All data preprocessing was conducted on Qiita[32] where all the data used here is freely available. All other visualizations were plotted through Matplotlib.

Quantitative comparison of metrics

All comparisons were made between Jaccard, Bray-Curtis, Weighted UniFrac, Unweighted UniFrac, Aitchison, and CTF distances. All distance metrics were calculated through QIIME2 (v. 2019.7). PERMANOVA on distances between subject groupings (i.e. vaginal vs. caesarean birth mode) was performed through scikit-bio (v. 0.5.5). Dimensionality reduction on distances was performed through PCoA via scikit-bio (v. 0.5.5). The first three components of each dimensionality reduction were evaluated through k-nearest neighbors (KNN) classification via scikit-learn (v. 0.21.2). To assess the classification accuracy, KNN classification was performed with 100-fold 40:60 cross-validation evaluating AUC and APR prediction accuracy at each fold-iteration via scikit-learn (v. 0.21.2).

Basis for simulations

Halfvarson et al. The IBD cohort used as the introduction example is a previously published dataset by Halfvarson et al. (Qiita ID 1629)[41]. The dataset consists, after filtering as described below, of 23 subjects (14 Crohn's disease (CD), 9 Control) each with one to eight samples for a total of 134 samples. Samples were filtered from the original data for only CD and Control. For the data-driven simulations, only the first 6 time points were retained to reduce the missing time points across subjects. The resulting data was then run through the data-driven simulation protocol described above for a sequencing depth of 500, 1000, and

10000 mean reads per sample. CTF was performed on each simulated data set through gemelli (v. 0.0.5) with a set rank of 2.

Case study: ECAM

The ECAM dataset published by Bokulich et al. followed 43 infants (19 c-section, 24 vaginally delivered) from birth over the first year of life with monthly fecal sampling (Qiita ID 10249)[42]. Three months (month 6, 15, and 19) were removed for a lack of subjects represented and CTF analysis was run with a set rank of 2. Features with < 5 total counts across samples were filtered. Samples with < 2000 reads per sample were removed.

Case study: DIABIMMUNE

The DIABIMMUNE dataset, published by Yassour et al., followed 39 infants (4 c-section, 35 vaginally delivered) from the 2nd month after birth over the first three years of life with monthly fecal sampling (Qiita ID 11884)[15]. Two months (month 28 and 30) were removed for a lack of subjects represented and CTF analysis was run with a set rank of 4. Features with < 5 total counts across samples were filtered. Samples with < 2000 reads per sample were removed.

Case study: American Gut

The American Gut Project data and metadata tables were acquired from ftp://ftp.microbio.me/AmericanGut/manuscript-package/ which was provided in McDonald et al. [43]. From this data the combined ECAM and DIABIMMUNE log-ratio feature set was used on the subset of the data with age and birth-mode labels provided (8,436 total samples).

Data availability

The sequences and biom tables for the IBD, ECAM, DIABIMMUNE, and AGP datasets can be found on Qiita (http://qiita.microbio.me) under study IDs 1629, 10249, 11884, and 10317 and at EBI or BioProject under ERP020401, ERP016173, PRJNA290381, and ERP012803.

Code availability

The CTF codebase named Gemelli is a fully unit tested open-source python package, and is installable through pip or conda. Additionally, CTF is wrapped in a QIIME2 plugin: https://github.com/biocore/gemelli; All the code and analyses are available in the 'Code Ocean' capsule: https://dx.doi.org/10.24433/CO.5938114.v1.

FEAST: fast expectation-maximization for microbial source tracking

The FEAST probabilistic model

Consider a single sink sample represented by a vector x, where x_j corresponds to the abundance of taxa j, $1 \leq j \leq N$. Let K be the number of known sources. Each known source is represented by a vector y_i, where y_{ij} is the observed abundance of taxa j in source i $(1 \leq i \leq K)$. Additionally, we assume there is an unobserved source $(K+1)$. Let $C_i = \sum_{j=1}^{N} y_{ij}$ and $C = \sum_{j=1}^{N} x_j$ be the total taxa counts of the known sources and sink respectively. With this notation, the generative model is as follows: we assume that there are mixture proportions α—a vector of length $K+1$—where α_i corresponds to the fraction of source i in the sink, hence $\sum_{i=1}^{K+1} \alpha_i = 1$. We also assume that there is an unknown relative abundance for each of the sources. For each source, $1 \leq i \leq K+1$, we have a vector γ_i, where $\sum_{j=1}^{N} \gamma_{ij} = 1$. Each γ_{ij} represents the true relative abundance of taxa j in source i.

$$\beta_j = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij}$$

$$y_i \sim Multinomial\big(C_i, (\gamma_{i1}, \ldots, \gamma_{iN})\big)$$

$$x \sim Multinomial(C, (\beta_1, \ldots, \beta_N))$$

Importantly, α and γ are not observed and are parameters of the model.

Fast inference via Expectation-Maximization

FEAST uses an Expectation-Maximization approach[44] to infer the model parameters. The likelihood is given by

$$p(\alpha,\gamma) = \left(\frac{C}{x_1,\ldots,x_N}\right)\prod_{j=1}^{N}\beta_j^{x_j}\prod_{i=1}^{K}\left(\frac{C}{y_{i1},\ldots,y_{iN}}\right)\prod_{j=1}^{N}\gamma_{ij}^{y_{ij}}$$

$$= \left(\frac{C}{x_1,\ldots,x_N}\right)\prod_{j=1}^{N}\left(\sum_{i=1}^{K+1}\alpha_i\,\gamma_{ij}\right)^{x_j}\prod_{i=1}^{K}\left(\frac{C}{y_{i1},\ldots,y_{iN}}\right)\prod_{j=1}^{N}\gamma_{ij}^{y_{ij}}$$

E step The log likelihood is given by

$$log\ log\ p(\alpha,\gamma) = \sum_{j=1}^{N} x_j\ log\ log\left(\sum_{i=1}^{K+1}\alpha_i\gamma_{ij}\right) + \sum_{i=1}^{K}\sum_{j=1}^{N} y_{ij}\ log\ log\left(\gamma_{ij}\right) + const$$

The expected complete log likelihood (Q) is given by

$$Q = \sum_{i=1}^{K+1}\sum_{j=1}^{N} x_j\,p(j)\cdot\ log(\alpha_i\,\gamma_{ij}) + \sum_{i=1}^{K}\sum_{j=1}^{N} y_{ij}\ log(\gamma_{ij}) + const$$

where

$$p(\,i\mid j) = \frac{\alpha_i^{(t)}\,\gamma_{ij}^{(t)}}{\sum_{i=1}^{K+1}\alpha_i^{(t)}\,\gamma_{ij}^{(t)}}$$

A more detailed derivation can be found in Appendix III.

M step Since the γ_{ij} are required to sum to 1, we use Lagrange multipliers δ_i to constrain the γ_{ij} values.

The Lagrangian is given by

$$L = \sum_{i=1}^{K+1}\sum_{j=1}^{N} x_j\, p(j) \cdot log(\alpha_i\, \gamma_{ij}) \quad + \sum_{i=1}^{K}\sum_{j=1}^{N} y_{ij}\, log(\gamma_{ij}) \quad - \sum_{i=1}^{K} \delta_i \left(\sum_{j=1}^{N} \gamma_{ij} \; - \; 1 \right)$$

Taking partial derivatives of L and solving gives the optimal update

$$\gamma_{ij}^{(t+1)} = \frac{x_j\, p(j) + y_{ij}}{\sum_{j=1}^{N} x_j\, p(j) + y_{ij}}$$

The update for the mixing probabilities is given by

$$\alpha_i^{(t+1)} = \sum_{j=1}^{N} \frac{x_j\, p(j)}{C} = \sum_{j=1}^{N} \frac{x_j}{C} \frac{\alpha_i^{(t)}\, \gamma_{ij}^{(t)}}{\sum_{i=1}^{K+1} \alpha_i^{(t)}\, \gamma_{ij}^{(t)}}$$

FEAST has two hyperparameters: the convergence threshold and the maximum number of iterations. In all our experiments we set these to default values of 10^{-6} and 1000 respectively. We used the multinomial distribution to model the data generating process since it is particularly relevant when analyzing microbiome datasets. Specifically, it addresses count uncertainty rather than directly transforming counts to relative abundances, and also models the competition to be counted (between taxa) instead of treating the counts of each taxon as independent[45].

Simulation studies

Parameters and settings

To construct realistic simulation scenarios, we used real microbiome data as sources, and simulated sinks as convex combinations thereof. Therefore, our simulations are representative of the abundance, over-dispersion of zeros, and technical noise mostly observed in real microbiome data. We designed our simulation parameters to reflect the wide range of Jensen-Shannon divergences and

potential sources observed across the real datasets we investigated. For a detailed description of the parameters and settings in each simulation study, see Appendix III.

Main simulation study In order to examine the accuracy of *FEAST*, we used multiple source environments with varying degrees of overlap in their distribution by randomly sampling from the Earth Microbiome Project. Each source environment was sub-sampled to contain 10,000 reads. In each iteration of the simulation we sampled K+1 known environments and used them to build a synthetic sink, with different mixing proportions. To simulate an unknown source, only K source environments are designated as known sources. We used 30 mixing proportions (corresponding to 30 simulated sinks) and $K = 20$ known sources in each iteration. For a detailed description of the simulation, see Appendix III.

Sequencing depth simulations In order to examine the robustness of FEAST to varying levels of sequencing depth, we used multiple source environments from the Earth Microbiome Project while varying their sequencing depth. In each iteration of our simulation we sampled environments (with median Jensen-Shannon divergence of 0.95) and used them to build a synthetic sink, with different mixing proportions and a set sequencing depth ranging from 100 through 10,000. Notably, by choosing a median Jensen-Shannon divergence of 0.95 we wanted to emphasize that even under the scenario in which the sources are non-overlapping and thus trivial to disambiguate, the sequencing depth will have an effect. Additionally, in these simulations, we only varied the sequencing depth of the sources. However, since the sink samples are a linear combination of the sources, these samples are also, indirectly, affected. To simulate an unknown source, only K source environments are designated as known sources. We used 30 mixing proportions (corresponding to 30 simulated sinks) and $K = 20$ known sources in each iteration. For a detailed description of the simulation, see Appendix III.

Unknown source simulations In order to evaluate FEAST's ability to estimate the contribution of the unknown source, we used real source environments from Lax et al. (2014) [46] and created synthetic sink

communities. Given that any source not sampled should, theoretically, be accounted for in the unknown source, realistic values of the unknown source can therefore span the range of percentages occupied by the observed sources. Specifically, there are scenarios in which the known sources comprise the entirety of the sink (unknown source contribution = 0), or on the other hand, scenarios in which the known sources did not contribute any taxa to the sink (unknown source contribution = 1). Therefore, the unknown source contribution values in our simulation ranges from 0 to 1. As a measure of accuracy, we used the squared Pearson correlation between the estimated mixing proportions and the true mixing proportions for each individual source across repeated simulation runs for the same scenario as the measure of accuracy. We used 30 mixing proportions (corresponding to 30 simulated sinks) and 5 sources (4 known sources) in each iteration. For a detailed description of the simulation, see Appendix III.

Noisy samples among sources As source assignment is discretionary (i.e., multiple samples can be pooled to a single source or considered as individual sources), we sought to examine the robustness of *FEAST* in the case where we have noisy realizations of the sources and their effect on prediction accuracy, we used K+1 distinct source environments by randomly sampling from the Earth Microbiome Project (i.e., soil, fresh water, feces, sebum, etc.), where each source was represented by 10 different samples (e.g., $soil_1, soil_2$, etc.). We then amalgamated these 10 samples (per source environment) and used the amalgamation of each source to build simulated sinks, with 30 different mixing proportions (corresponding to 30 simulated sinks). In each iteration of our simulation, we aggregated $s \in \{1, \dots, 10\}$ samples from the representative samples of each source environment to estimate the different mixing proportions.

Prediction accuracy

To measure accuracy, we used the squared Pearson correlation coefficient between the estimated and true mixing proportions, for each individual source across repeated simulation runs (i.e., different mixing proportions) for the same Jensen Shannon divergence value. In each iteration we varied the degree of similarity of the source environments.

Running time measurements

In each iteration, we used K randomly selected source environments from the Earth Microbiome Project, where $K \in \{5, 10, 50, 100, 500, 100\}$. Each source environment was down-sampled to contain 10,000 reads. We recorded the run-time of each method, for each number of source environments, each iteration. Notably, the running time of the hundreds of samples using the random forest classifier is relatively short. However, given that both SourceTracker and *FEAST* substantially improve accuracy over the random forest approach, we focused on these two methods for all subsequent benchmarks shown.

Comparing model performance

We evaluated the performance of our model against common approaches widely used for microbial source tracking—namely, SourceTracker[47] and the random forest classifier [48]. Both methods use community structure to measure the similarity between sink samples and potential source environments. The statistical model used by *FEAST* shares many similarities with the model proposed by SourceTracker[47], namely that both models assume each sink is a convex combination of the known and unknown sources. Additionally, in both methods, source assignment is discretionary (i.e., multiple samples can be pooled to a single source or considered as individual sources). Thus, the main difference between the methods lies in their optimization procedure. *FEAST* uses an Expectation-Maximization algorithm to evaluate the proportions of source contribution, whereas SourceTracker uses a Gibbs Sampler (MCMC). Notably, in other fields in genomics it has been demonstrated that such optimization can be critical in terms of the reduction of running time. For example, in statistical genetics, the original method for the inference of population structure, STRUCTURE [49], uses a Markov chain Monte Carlo for the parameter estimation, while other methods such as FRAPPE [50] and ADMIXTURE [51] uses Expectation-Maximization and quasi-Newton optimization techniques respectively to reach similar accuracy, but considerably more efficiently. This improvement in running time eventually may translate to improvement in accuracy. Particularly, the accuracy achieved by SourceTracker may be improved by increasing the number of burn-in iterations, however this comes at the expense of additional running time.

Distinguishing ICU patients from healthy adults

The objective of this set of experiments is to classify each sink (ICU patient or a healthy adult) using its overall dissimilarity to all sources (healthy adults). The dependent variable (y) is a binary vector of cases (ICU patients) and controls (healthy adults) $y_i \in \{0,1\}, i = \{1, \ldots, N\}$ where N is the number of sink samples. When classifying using *FEAST* or SourceTracker, we designate the proportion of the unknown source as a predictor for each sink's class label. When classifying using Jensen-Shannon, and UniFrac, we designate the average of the dissimilarity measurements between the sink and all the other sources as the predictor.

FEAST We applied *FEAST* to every sink sample (ICU or healthy), where the known sources are 100 distinct healthy individuals from the American Gut Project. We next used the estimated proportions of the unknown source as the input to the classifier.

SourceTracker We applied SourceTracker to every sink sample (ICU or healthy), where the known sources are 100 distinct healthy individuals from the American Gut Project. We next used the estimated proportions of the unknown source as the input to the classifier.

Jensen Shannon divergence We calculated the Jensen Shannon divergence value between each sink sample (ICU or healthy) and the known source samples used in *FEAST* and SourceTracker (e.g., 100 distinct healthy individuals from the American Gut Project). We next used the average Jensen Shannon divergence value (across known sources) as the input to the classifier.

UniFrac We calculated the Weighted UniFrac distance between each sink sample (ICU or healthy) and the known source samples used in *FEAST* and SourceTracker (e.g., 100 distinct healthy individuals from the American Gut Project). We next used the average Weighted UniFrac distance (across known sources) as the input to the classifier.

Data distribution

Throughout the paper, the box-plot elements are: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

Data sets

We evaluated the performance of *FEAST* using five data sets collected using both 16S rRNA gene and whole metagenome shotgun sequencing.

The first data set was collected and studied by Backhed et al. 2015[16] (accession number ERP005989), which characterizes the temporal gut microbiome of 98 Swedish infants, each sampled at birth, four months after birth, and 12 months after birth. This data set also contains gut microbiome samples collected from the infants' corresponding mothers during the first few days after delivery. 83 infants were delivered vaginally and the remaining 15 by C-section. In this dataset, shotgun sequencing reads were assembled into contigs using SOAPdenovo2 [52]. The contigs were binned according to their abundance variations across samples and GC-depth pattern for further assembly into draft genomes. The draft genomes were then clustered into MetaOTUs based on MUMi[53] and the Spearman distance [54], and their taxa were determined in relation to the NCBI genomes.

The second data set was collected and studied by Lax et al. 2014[46] (accession number ERP005806). This study used the V4 region of the 16S rRNA gene to evaluate the microbial contamination from seven groups of individuals in their respective residences over the course of six weeks. In our analysis, we investigated one house, where the inhabitants were genetically related. We used skin samples of inhabitants from several body parts (hand, foot, and nose) as sources, and indoor house surfaces (e.g. kitchen floor, kitchen counter) as sinks.

The third dataset was collected and studied by Knights et al. 2011 [47] (data from this study are stored in https://github.com/danknights/sourcetracker). This study used datasets of bacterial 16S rRNA [55,56] (V2 region of the 16S rRNA gene) to investigate contamination in settings such as office buildings, hospitals, and research laboratories. As potential contaminants, human skin, oral cavities, feces and temperate soils were considered.

The fourth dataset was collected and studied by McDonald et al. 2016 [57] (accession number ERP012810), the American Gut Project [43] (EBI project number PRJEB11419), Using the V4 region of the 16S rRNA gene, McDonald et al. characterized a cohort of patients from an intensive care unit (ICU). The study collected samples from the skin, mouth and feces (gut) of 115 American and Canadian ICU patients at time of admission (within 48 hours) to the ICU as well as at time of discharge from the ICU.

The fifth dataset was collected and studied by Taur et al. 2012 [58] (data from this study are stored in http://www.ncbi.nlm.nih.gov/sra). In this study by Taur et al. 2012[58], fecal specimens were collected longitudinally from 94 patients undergoing allo-HSCT from before treatment up to 35 days after treatment. This study used the V1-V3 region of bacterial 16S ribosomal RNA genes.

Data Availability

All of the datasets analyzed in this paper are public and can be referenced at the following accession numbers: The first data set was collected and studied by Backhed et al. 2015[16] (accession number ERP005989). The second data set was collected and studied by Lax et al. 2014[46] (accession number ERP005806). The third dataset was collected and studied by Knights et al. 2011[47] (data from this study are stored in https://github.com/danknights/sourcetracker). The fourth dataset was collected and studied by McDonald et al. 2016[57] (accession number ERP012810) and the American Gut Project [43] (EBI project number PRJEB11419). The fifth dataset was collected and studied by Taur et al. 2012[58] (data from this study are stored in http://www.ncbi.nlm.nih.gov/sra). In our simulations we used the Earth microbiome project (ftp://ftp.microbio.me/emp/release1/otu_tables/closed_ref_greengenes/).

Code availability

Code is available in https://github.com/cozygene/FEAST

Results

Modeling the temporal dynamics of the gut microbial community in adults and infants

A brief description of MTV-LMM

We begin with an informal description of the main idea and utility of MTV-LMM. A more comprehensive description can be found in the Methods. MTV-LMM is motivated by our assumption that the temporal changes in the abundance of taxa are a time-homogeneous high-order Markov process. MTV-LMM models the transitions of this Markov process by fitting a sequential linear mixed model (LMM) to predict the relative abundance of taxa at a given time point, given the microbial community composition at previous time points. Intuitively, the linear mixed model correlates the similarity between the microbial community composition across different time points with the similarity of the taxa abundance at the next time points. MTV-LMM is making use of two types of input data: (1) continuous relative abundance of focal taxa j at previous time points and (2) quantile-binned relative abundance of the rest of the microbial community at previous time points. The output of MTV-LMM is prediction of continuous relative abundance, for each taxon, at future time points. In order to apply linear mixed models, MTV-LMM generates a temporal kinship matrix, which represents the similarity between every pair of samples across time, where a sample is a normalization of taxa abundances at a given time point for a given individual (see Methods). When predicting the abundance of taxa j at time t, the model uses both the global state of the entire microbial community in the last q time points, as well as the abundance of taxa j in the previous p time points. The parameters p and q are determined by the user, or can be determined using a cross-validation approach; a more formal description of their role is provided in the Methods. MTV-LMM has the advantage of increased power due to a low number of parameters coupled with an inherent regularization mechanism, similar in essence to the widely used ridge regularization, which provides a natural interpretation of the model.

Model evaluation

We evaluated MTV-LMM by testing its accuracy in predicting the abundance of taxa at a future time point using real time series data. Such evaluation will mitigate overfitting, since the future data points are held

out from the algorithm. To measure accuracy on real data, we used the squared Pearson correlation coefficient between estimated and observed relative abundance along time, per taxon. In addition we validated MTV-LMM using synthetic data, illustrating realistic dynamics and abundance distribution, as suggested by Aijo et al. 2018 [37]. Following [37], we evaluate the performance of the model using the 'estimation-error', defined to be the Euclidean distance between estimated and observed relative abundance, per time point (see Appendix I Note S1). We used real time series data from three different datasets, each composed of longitudinal abundance data. These three datasets are David et al. [11] (2 adult donors—DA, DB—average 250 time points per individual), Caporaso et al. [10] (2 adult donors—M3, F4—average 231 time points per individual), and the DIABIMMUNE dataset [15] (39 infant donors—average 28 time points per individual). In these datasets, the temporal parameters p and q were estimated using a validation set, and ranged from 0 to 3. See Methods for further details. We compared the results of MTV-LMM to common approaches that are widely used for temporal microbiome modeling, namely the AR(1) model (see Methods), the sparse vector autoregression model sVAR [25] , the ARIMA Poisson regression [29] and TGP-CODA [37]. Overall, MTV-LMM's prediction accuracy is higher than AR's (Appendix I S1 Table) and significantly outperforms both the sVAR method and the Poisson regression across all datasets, using real time-series data (Fig. 1). In addition, since TGP-CODA can not be fully applied to these real datasets (due to scalability limitations), we used synthetic data, considering a scenario of 200 taxa and 70 time points with realistic dynamics and abundance distribution, as suggested by the authors of this method. Similarly to the real data, MTV-LMM significantly outperforms all the compared methods (Appendix I Fig. S1).

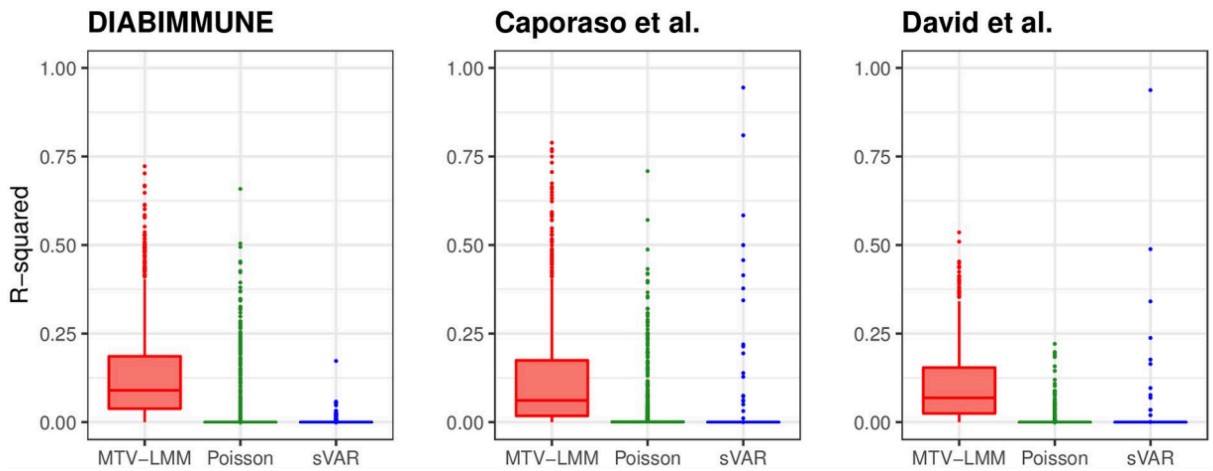

Figure 1. Model comparison. *MTV-LMM* outperforms commonly used methods in prediction accuracy (R^2) and detection of autoregressive dynamics. *MTV-LMM* predictions are in red, ARIMA Poisson regression in green, and sVAR in blue.

Inference on the estimated association matrix

We applied MTV-LMM to the DIABIMMUNE infant dataset and estimated the species-species association matrix across all individuals, using 1440 taxa that passed a preliminary screening according to temporal presence-absence patterns (see Methods). We found that most of these effects are close to zero, implying a sparse association pattern. Next, we applied a principal component analysis (PCA) to the estimated species-species associations and found a strong phylogenetic structure (PerMANOVA P-value = 0.001) suggesting that closely related species have similar association patterns within the microbial community (Fig. 2). These findings are supported by Thompson et al. [17], who suggested that ecological interactions are phylogenetically conserved, where closely related species interact with similar partners. Gomez et al. [59] tested these assumptions on a wide variety of hosts and found that generalized interactions can be evolutionary conserved. We note that the association matrix estimated by MTV-LMM should be interpreted with caution since the number of possible associations is quadratic in the number of species, and it is, therefore, unfeasible to infer with high accuracy all the associations. However, we can still aggregate information across species or higher taxonomic levels to uncover global patterns of the microbial composition dynamics (e.g., principal component analysis).

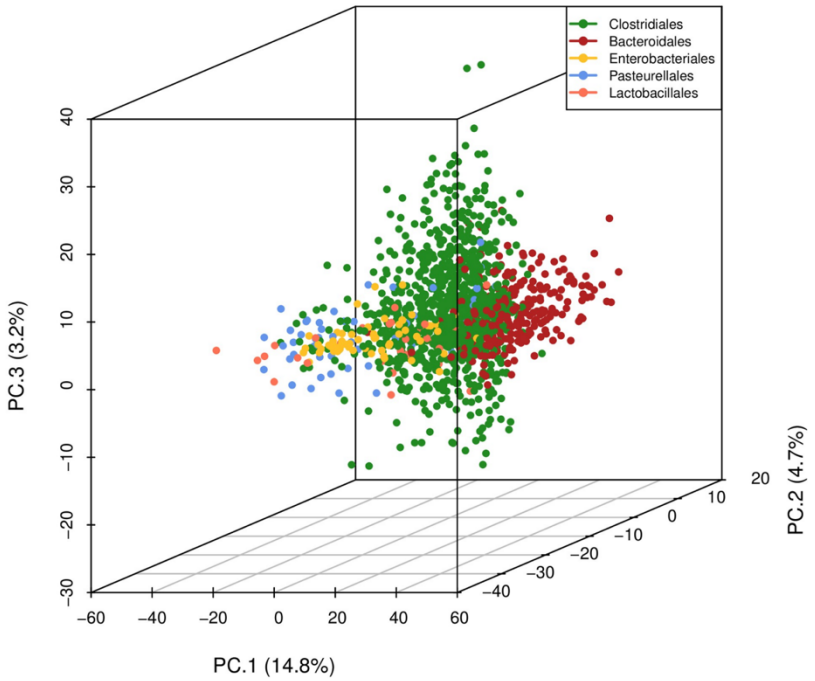

Figure 2. The first three principle components of the inferred association matrix recover known phylogenetic structure. Closely related species, in the DIABIMUNE dataset, have similar association patterns within the microbial community. Shown on each axis is the percentage of variance explained by each principal component for the top five orders in the data.

Time-explainability as a measure of the autoregressive component in the microbial community

In order to address the fundamental question regarding the gut microbiota temporal variation, we quantify its autoregressive component. Namely, we quantify to what degree the abundance of different taxa can be inferred based on the microbial community composition at previous time points. In statistical genetics, the fraction of phenotypic variance explained by genetic factors is called heritability and is typically evaluated under an LMM framework [60]. Intuitively, linear mixed models estimate heritability by measuring the correlation between the genetic similarity and the phenotypic similarity of pairs of individuals. We used MTV-LMM to define an analogous concept that we term time-explainability, which corresponds to the fraction of temporal variance explained by the microbiome composition at previous time points. In order to highlight the effect of the microbial community, we next estimated the time explainability of taxa in each dataset, using the parameters q = 1, p = 0. The resulting model corresponds to the formula: taxa(t) = microbiome community(t−1) + individual effect(t−1) + unknown effects. Of the taxa we examined, we

identified a large portion of them to have a statistically significant time-explainability component across datasets. Specifically, we found that over 85% of the taxa included in the temporal kinship matrix are significantly explained by the time-explainability component, with estimated time-explainability average levels of 23% in the DIABIMMUNE infant dataset (sd = 15%), 21% in the Caporaso et al. (2011) dataset (sd = 15%) and 14% in the David el al. dataset (sd = 10%) (Fig. 3, Appendix I Fig. S2). Notably, we found that higher time explainability is associated with higher prediction accuracy (Appendix I Fig. S3).

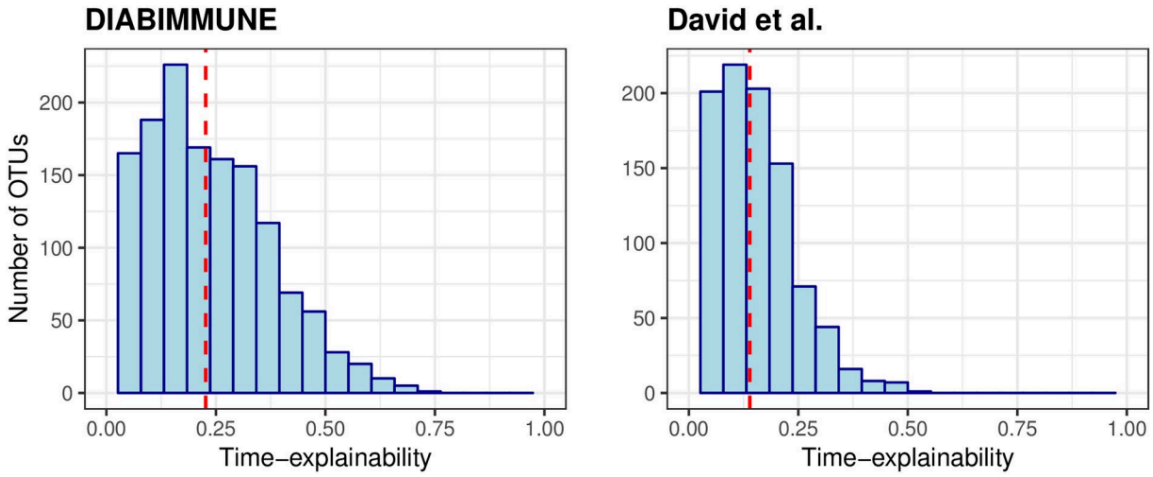

Figure 3. Time-explainability distribution. Time-explainability distribution in the DIABIMMUNE infant dataset (left) and David et al. adult dataset (right). The average time-explainability (denoted by a dashed line) in the DIABIMMUNE cohort is 23% and in David et al. is 14%.

Non-autoregressive dynamics contain phylogenetic structure

As a secondary analysis, we aggregated the time-explainability by taxonomic order, and found that in some orders (non-autoregressive orders) all taxa are non-autoregressive, while in others (mixed orders) we observed the presence of both autoregressive and non-autoregressive taxa (Appendix I Fig. S4), where an autoregressive taxa have a statistically significant time-explainability component. Particularly, in the DIABIMMUNE infant data set, there are 7244 taxa, divided into 55 different orders. However, the taxa recognized by MTV-LMM as autoregressive (1387 out of 7244) are represented in only 19 orders out of the 55. The remaining 36 orders do not include any autoregressive taxa. Unlike the autoregressive organisms, these non-autoregressive organisms carry a strong phylogenetic structure (t-test p-value <

10−16), that may indicate a niche/habitat filtering. This observation is consistent with the findings of Gibbons et al. [25] , who found a strong phylogenetic structure in the non-autoregressive organisms in the adult microbiome. Notably, across all datasets, there is no significant correlation between the order dominance (number of taxa in the order) and the magnitude of its time-explainability component (median Pearson r = 0.12). For example, in the DIABIMMUNE data set, the proportion of autoregressive taxa within the 19 mixed orders varies between 2% and 75%, where the average is approximately 20%. In the most dominant order, Clostridiales (representing 68% of the taxa), approximately 20% of the taxa are autoregressive and the average time-explainability is 23%. In the second most dominant order, Bacteroidales, approximately 35% of the taxa are autoregressive and the average time-explainability is 31%. In the Bifidobacteriales order, approximately 75% of the taxa are autoregressive, and the average time-explainability is 19% (Fig 4). We hypothesize that the large fraction of autoregressive taxa in the Bifidobacteriales order, specifically in the infants dataset, can be partially attributed to the finding made by [61], according to which some sub-species in this order appear to be specialized in the fermentation of human milk oligosaccharides and thus can be detected in infants but not in adults. This emphasizes the ability of MTV-LMM to identify taxa that have prominent temporal dynamics that are both habitat and host-specific.

As an example of MTV-LMM's ability to differentiate autoregressive from non-autoregressive taxa within the same order, we examined Burkholderiales, a relatively rare order (less than 2% of the taxa in the data) with 76 taxa overall, where only 19 of which were recognized as autoregressive by MTV-LMM. Indeed, by examining the temporal behavior of each non-autoregressive taxa in this order, we witnessed abrupt changes in abundance over time, where the maximal number of consecutive time points with abundance greater than 0 is very small. On the other hand, in the autoregressive taxa, we witnessed a consistent temporal behavior, where the maximal number of consecutive time points with abundance greater than 0 is well over 10 (Appendix I Fig. S5).

The autoregressive component of an adult versus infant microbiome

The colonization of the human gut begins at birth and is characterized by a succession of microbial consortia [62–65], where the diversity and richness of the microbiome reach adult levels in early childhood. A longitudinal study has recently been used to show that infant gut microbiome begins transitioning towards an adult-like community after weaning [66]. This observation is validated using our infant longitudinal data set (DIABIMMUNE) by applying PCA to the temporal kinship matrix (Fig. 4). Our analysis reveals that the first principal component (accounting for 26% of the overall variability) is associated with time. Specifically, there is a clear clustering of the time samples from the first nine months of an infant's life and the rest of the time samples (months $10-36$) which may be correlated to weaning. As expected, we find a strong autoregressive component in an infant microbiome, which is highly associated with temporal variation across individuals. By applying PCA to the temporal kinship matrix, we demonstrate that there is high similarity in the microbial community composition of infants at least in the first 9 months. This similarity increases the power of our algorithm and thus helps MTV-LMM to detect autoregressive taxa.

In contrast to the infant microbiome, the adult microbiome is considered relatively stable [10,67], but with considerable variation in the constituents of the microbial community between individuals. Specifically, it was previously suggested that each individual adult has a unique gut microbial signature [68,69], which is affected, among others factors, by environmental factors [20] and host lifestyle (i.e., antibiotics consumption, high-fat diets [11] etc.). In addition, David et al. [11] showed that over the course of one year, differences between individuals were much larger than variation within individuals. This observation was validated in our adult datasets (David et al. and Caporaso et al.) by applying PCA to the temporal kinship matrices.

In both David et al. and Caporaso et al., the first principal component, which accounts for 61% and 43% of the overall variation respectively, is associated with the individual's identity (Fig. 5). Using MTV-LMM we observed that despite the large similarity within adult individuals, there is also a non-negligible autoregressive component in the adult microbiome. The fraction of variance explained by time across individuals can range from 6% up to 79% for different taxa. These results shed more light on the temporal

behavior of taxa in the adult microbiome, as opposed to that of infants, which are known to be highly affected by time [66].

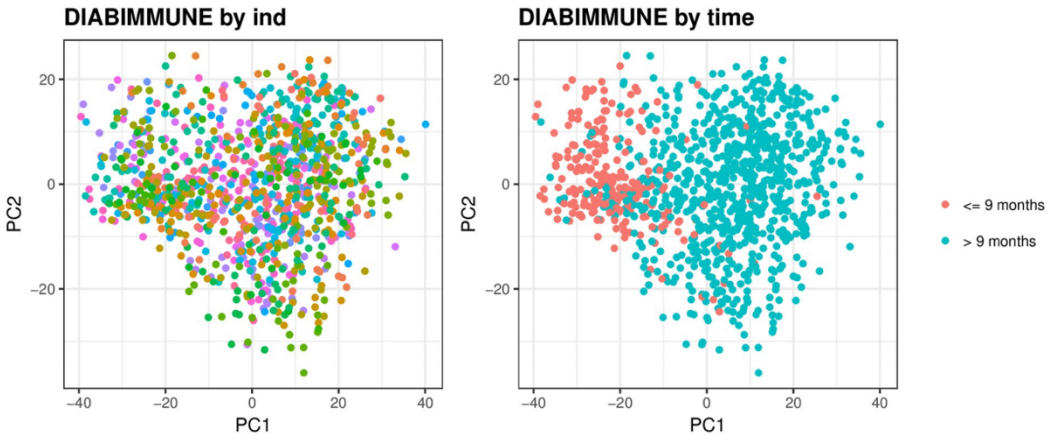

Figure 4. The first two principal components of the temporal kinship matrix in infants. The first two principal components of the temporal kinship matrix color coded by individual (left; 39 infant donors) and by time (right; before and after nine months) using the DIABIMMUNE data.

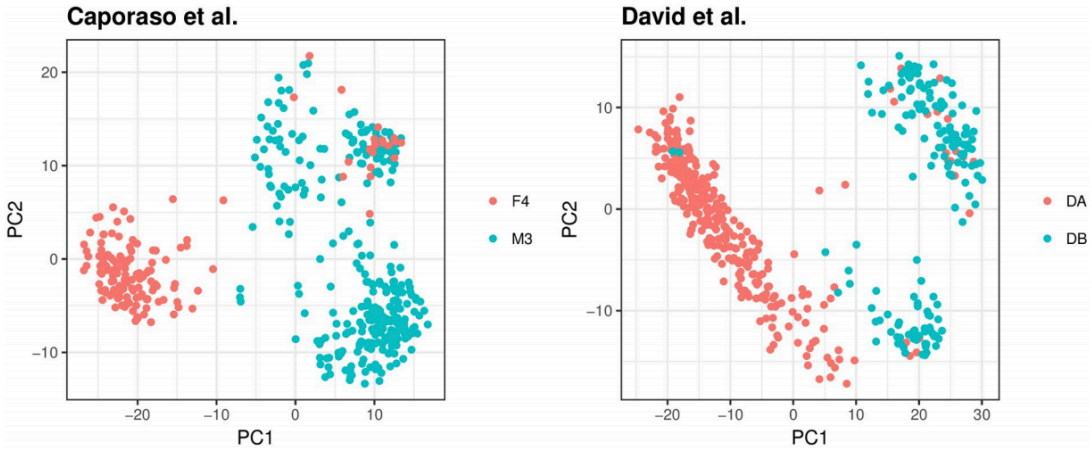

Figure 5. The first two principal components of the temporal kinship matrix in adults. The first two principal components of the temporal kinship matrix color coded by individual. Caporaso et al. [10] (left; 2 adult donors: M3, F4) and David et al. [11] (right; 2 adult donors: DA, DB).

Context-aware dimensionality reduction deconvolutes gut microbial community dynamics

Host-associated microbiomes are often host-specific, with the subject driving most of the variation. This host-specific variation can obscure microbial changes that are broadly associated with a given phenotype. Collecting multiple samples from the same participant, either longitudinally or from different body sites (that is, 'repeated measures'), is a valid experimental approach to control for interindividual variation. However, there are multiple challenges to leveraging this type of experimental design due to the nature of microbiome sequencing datasets.

One common way to explore microbiome sequencing data is by performing dimensionality reduction on a distance matrix (for example, principal coordinates analysis (PCoA)), which describes the relationship among samples, allowing global differences across a dataset to be observed. Nonetheless, when applied to repeated measures, this approach does not account for the inherent temporal or spatial correlation structure. An alternative to analyze repeated measures microbiome data is by using supervised methods, which are focused on generative models inferring the dynamics of these communities (for example, generalized Lotka Volterra)[37,70–72]. Although these methods account for the correlation structure induced by repeated measures, as well as for sparsity and compositionality, their output does not directly allow clustering of phenotypes by microbial community dynamics.

To address these challenges simultaneously, we developed compositional tensor factorization (CTF), which allows an unsupervised dimensionality reduction for repeated measures data, producing both a traditional beta-diversity analysis as well as a differential feature abundance assessment. In the first step, a two-dimensional matrix is transformed using the robust, centered log-ratio technique[33] to account for the inherent sparse and compositional nature of next-generation sequencing datasets[73] (Fig. 6a). Next, this

transformed matrix is restructured into a three-dimensional tensor, which relates microbial sequences, sampled host and time or space (Fig. 6b). Decomposition (that is, factorization) of this tensor provides distinct vectors for subjects (U), microbial features (V) and timepoints (W) (Fig. 6c). Analogous to the concept of reference frames[74], these vectors are unit scaled and therefore can be ordered, where their ranking indicates their association to the underlying phenotypic groups. From here on, we will refer to the ordering of these vectors as 'rankings' (that is, 'feature rankings'). Notably, CTF assumes the data harbors an underlying low-rank structure, where only a few phenotypic factors explain the majority of the variance[33] (Fig. 6d–g).

Model evaluation using data-driven simulations

To demonstrate the use of CTF, we applied it to a simulated longitudinal dataset with two phenotypic groups. Simulations were generated based on distributions in real longitudinal 16S data from Halfvarson et al.[41] while varying the sequencing depth and temporal sampling densities as described by Äijö et al.[37] This dataset was chosen because there were strong differences in microbial composition and beta diversity between subjects with and without Crohn's disease[41]. We compared CTF to state-of-the-art beta-diversity metrics through PCoA including Jaccard[75], Bray–Curtis[76], Aitchison[77], unweighted UniFrac[78] and weighted UniFrac[79]. *K*-nearest neighbor (KNN) classification by disease state in each of our simulations revealed that CTF exhibited higher accuracy than existing methods regardless of sequencing depth or the number of longitudinally collected samples (Fig. 7, Appendix II Table S1 and Fig. S1). CTF also exhibited higher discriminatory power by PERMANOVA *F* statistic across all levels of sequencing depth and at higher sampling densities (≥3 timepoints, Fig. 7).

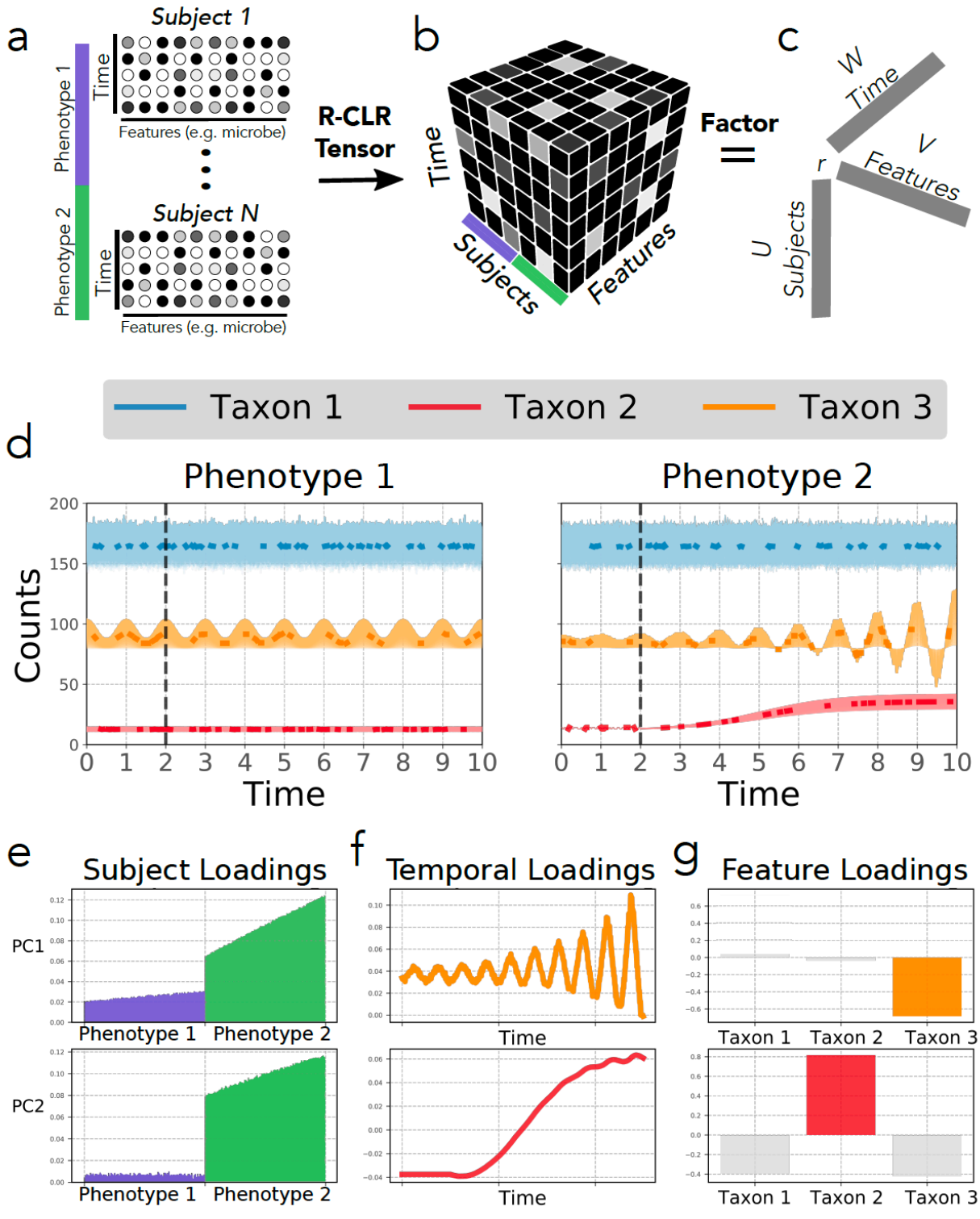

Figure 6. Overview of the CTF algorithm. (**a**) CTF utilizes feature abundance matrices for subjects over time. For each subject with a phenotype of interest, the data is represented as relative abundances of features (abundance gradient represented in grayscale) over time. (**b**) The matrices are concatenated, robust-centered log-ratio transformed

(R-CLR) and structured into a tensor format with modes corresponding to subjects, features and time. (**c**) The resulting tensor is then factored based only on observed data into loading vectors for each dimension (i.e. subject, timepoint, and feature). (**d**) Simulated count data is plotted on the y-axis for three taxa with the mean counts in bold and missing values absent from the bold line. Standard deviation of distributions are shaded behind. Two phenotypes are compared; a control unchanging in time (left) and a dynamic phenotype with a perturbation at time point 2 (right). Taxon 1 (blue) is highly abundant and noisy, taxon 2 (red) is lowly abundant but growing exponentially in phenotype 2, and taxon 3 (orange) is oscillatory with increasing amplitude in phenotype 2. The first two principal component axes (i.e. loadings) from CTF (PC1 (top) and PC2 (bottom)) are plotted on the y-axis with the corresponding sample (**e**), time (**f**), and feature loadings (**g**). In PC1, phenotype 2 is linked to the unstable oscillatory waveform of highly loaded taxon 3 (orange, top). Similarly, in PC2, phenotype 2 is linked to the sigmoidal waveform of highly loaded taxon 2 (red, bottom).

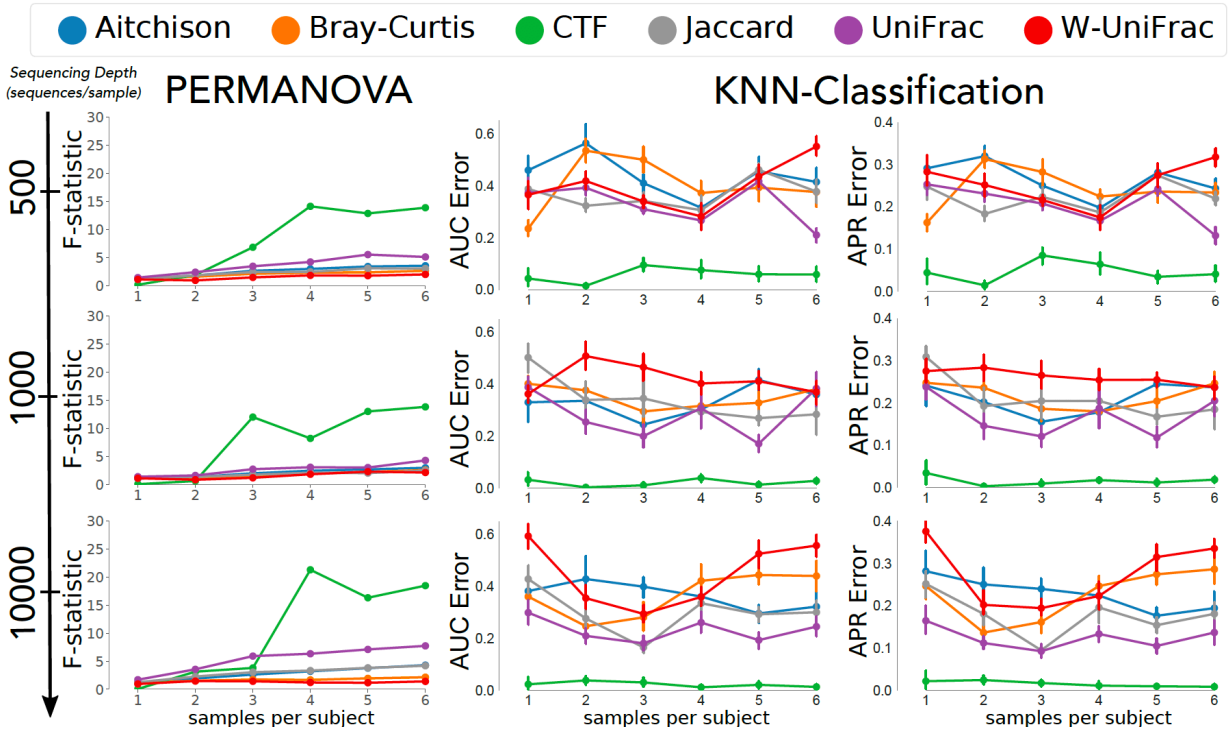

Figure 7. CTF outperforms popular distance metrics in longitudinal in silico data-driven simulations. Increasing sequencing depth (500 - 10,000; rows) over differing temporal sampling densities (x-axis) evaluated for PERMANOVA F-statistic as a measure of discriminatory power (left column), in addition to KNN-classification cross-validation by AUC (n=100; middle column), and APR (n=100; right column). Compared among CTF (green) and popular distance metrics Aitchison (blue), Bray-Curtis (orange), Jaccard (grey), unweighted (purple), and weighted (red) UniFrac. Error bars represent standard error of the mean.

Tracking infant gut development

We next applied CTF to two published datasets that tracked infant gut development over time. The datasets abbreviated to ECAM (*n* subjects, 43)[42] and DIABIMMUNE (*n* subjects, 39)[15] followed infants for the first 2 and 3 years of life, respectively. Both studies observed that birth mode (that is, vaginal delivery or cesarean section (c-section)) differentiated microbial community composition. Similar to our results from the simulated data, CTF is tenfold better at discriminating vaginally from cesarean born infants compared to state-of-the-art beta-diversity metrics (Appendix II Figs. S2a,b, S3a,b and Table S2).

We sought to examine CTF's ability to reproducibly identify differentially abundant microbes in an unsupervised manner. To this end, we compared the feature rankings between the ECAM and DIABIMMUNE datasets along the first axis of variation and found they were significantly correlated (Pearson correlation, $R2 = 0.974$, $P < 10{-}10$) (Appendix II Fig. S2). While these two datasets had <50% overlap at the sOTU level (Appendix II Fig. S2d), high- and low-ranked sOTUs, grouped at the genus level, were similar across both datasets (Appendix II Fig. S2e). We note that although these datasets were collected and processed using distinct protocols and by different laboratories, CTF identified the same taxa driving gut microbiome differentiation by birth mode, suggesting a robust microbial structure across infants.

We constructed a birth-mode log ratio of vaginally delivered to-cesarean features using the sOTUs most associated with vaginal and cesarean birth in each dataset (Appendix II Fig. S4 and Methods). Samples were substantially separated by birth mode in both datasets along time (Appendix II Fig. S5 and Table S3). We note that these birth-mode microbial signatures are not confounded by established differentiators such as antibiotics usage or feeding mode (Appendix II Fig. S5). Nonetheless, we cannot rule out the possibility of unmeasured confounders. We next combined those sOTUs common to both ECAM and DIABIMMUNE birth-mode ratios to create a 'microbial birth-mode signature'.

To examine the robustness of this microbial birth-mode signature, we tested its discriminatory ability in data from the American Gut Project (AGP) ($n = 8{,}099$), a large cross-sectional dataset[43]. We found that this signature significantly differentiated participants under the age of four by birth mode (t-test $P = 0.042$, Appendix II Fig. S6), consistent with our previous findings. The robustness of this microbial signature, across multiple datasets, highlights the ability of CTF to identify differentially abundant features reproducibly associated with a phenotype.

In both the ECAM and DIABIMMUNE datasets we observed that throughout infant development samples from vaginally versus cesarean-born infants became less distinct as time progressed (Appendix II Fig. S2a,b). Similarly, the microbial birth-mode signature no longer differentiated participants by birth mode in samples from participants above the age of four in the AGP dataset (Appendix II Fig. S6).

FEAST: fast expectation-maximization for microbial source tracking

A brief description of FEAST

FEAST is a highly efficient Expectation-Maximization-based method that takes as input a microbial community (called the sink) as well as a separate group of potential source environments (called the sources) and estimates the fraction of the sink community that was contributed by each of the source environments. By virtue of these mixing proportions often summing to less than the entire sink, FEAST also reports the potential fraction of the sink attributed to other origins, collectively referred to as the unknown source. The statistical model used by FEAST assumes each sink is a convex combination of known and unknown sources. FEAST is agnostic to the sequencing data type (i.e., 16s rRNA or shotgun sequencing), and can efficiently estimate up to thousands of source contributions to a sample.

Model evaluation using data-driven synthetic mixtures

We compared the accuracy of FEAST to both SourceTracker[47], and the random forest classifier used in previous source-tracking work[48]. We source communities based on distributions in real source environments from the Earth Microbiome Project[80], while varying the level of divergence between

sources (see Methods). In each of our simulations, FEAST exhibited higher accuracy than SourceTracker and the random forest classifier across all levels of divergence (Fig. 8a; Appendix III Fig. S1). Since both SourceTracker and FEAST substantially improve accuracy over the random forest approach, we focused on these two methods for all subsequent benchmarks shown. Next, we examined the robustness of FEAST and SourceTracker through varying levels of sequencing depth, when disambiguation between sources is trivial (high divergence). As expected, the accuracy of both algorithms increased as sequencing depth increased. Nonetheless, we observed that FEAST still compared favorably across all levels of sequencing depth (Appendix III Fig. S2). Finally, as it may be nearly impossible to obtain sequencing data for all potential sources in a study, we sought to evaluate FEAST's ability to estimate the contribution of the unknown source. To this end, we used real source environments from Lax et al. (2014)[46], while varying the unknown source contribution from absent to exclusive. Across these experiments, FEAST was significantly more accurate in estimating the unknown source proportion (two-sided t-test p-value $< 10^{-14}$). Notably, by properly adjusting its estimates for the unknown source, FEAST also produces more accurate mixing proportions for the observed sources as well as low variance (Fig. 8b; Appendix III Fig. S3, S4).

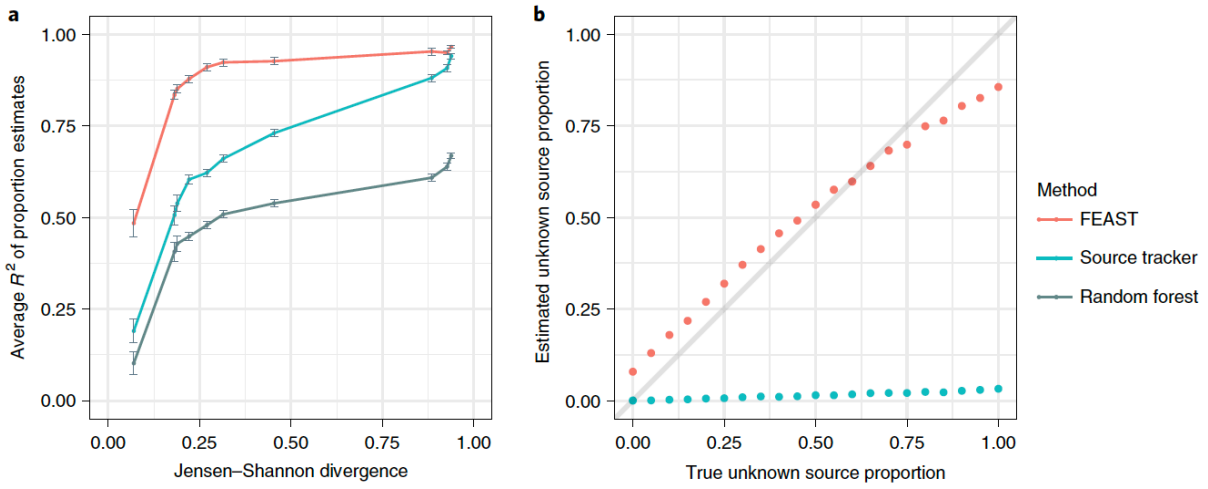

Figure 8. Methods comparison. (a) The accuracy of FEAST, the random forest classifier and SourceTracker on simulated data. Each simulation was performed using 20 real source environments and simulated sinks. The x axis is average Jensen–Shannon divergence value across known sources (that is, the degree of overlap between the sources

from completely identical to completely non-overlapping). The y axis represents correlation across all source environments between true and estimated mixing proportions; error bars show the standard error of the mean (n = 30). (b) Evaluation of FEAST and SourceTracker through varying levels of unknown source proportions.

Running time

One of FEAST's distinct advantages over other methods is its speed (Fig. 9; Appendix III Table S1). Specifically, across all experiments, FEAST reduced running time by a factor of 30-300 compared to SourceTracker, while maintaining and even improving the accuracy. Consequently, FEAST can simultaneously estimate thousands of potential source environments on the order of minutes to hours, wherein SourceTracker may take upwards of days (Appendix III Table S1). We note that SourceTracker's accuracy may potentially be improved by increasing the number of burn-in iterations or otherwise increasing the number of iterations of the Markov chain, however, this comes at the expense of additional running time (See Methods for a comprehensive discussion of the tradeoff between time and accuracy in Markov Chain Monte Carlo).

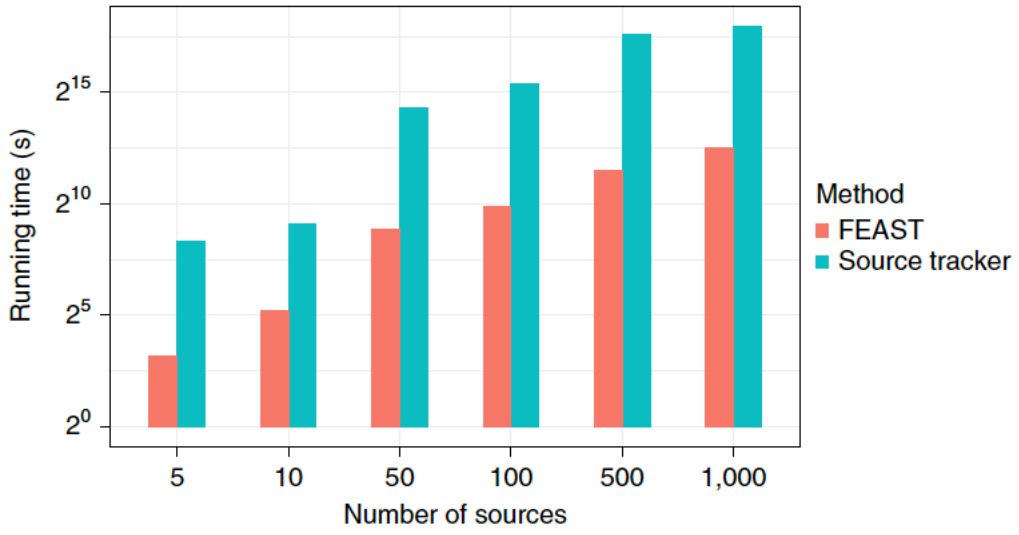

Figure 9. Running time comparison to current state-of-the-art. Running time (log scale, seconds) comparison across all simulation studies, using a sequencing depth of 10,000 reads per source.

Real data applications

We applied FEAST to five real datasets in order to demonstrate the utility of microbial source tracking methods across different contexts. We first use FEAST as it was originally intended—to quantify the contribution of sources to specific sink environments.

Succession and initial colonization in infants

Using FEAST for time-series analysis offers a quantitative way to characterize developmental microbial populations, such as the infant gut. In this context, we can leverage previous time-points and external sources to understand the origins of a specific, temporal community state. For instance, we can estimate if taxa in the infant gut originate from the birth canal, or if they are derived from some other external source at a later time point. To demonstrate this capability, we used longitudinal data from Backhed et al. 2015[16], which contains gut microbiome samples from infants as well as from their corresponding mothers. In this analysis, we treated samples taken from the infants at age 12 months as sinks, considering respective earlier time points and maternal samples as sources. In these settings, FEAST revealed a significantly larger maternal contribution in vaginally-delivered infants over caesarean-delivered infants (Fig. 10), where other methods did not (Appendix III Fig. S5). These results are consistent with the results of Backhed et al. 2015. We further explored whether biological mothers were more likely to be identified as sources of their infant's microbiome than other potential source communities. We considered all maternal and early infant samples as potential sources and found that for over 83% of the sink samples, the top contributing sources were from the same family (Appendix III).

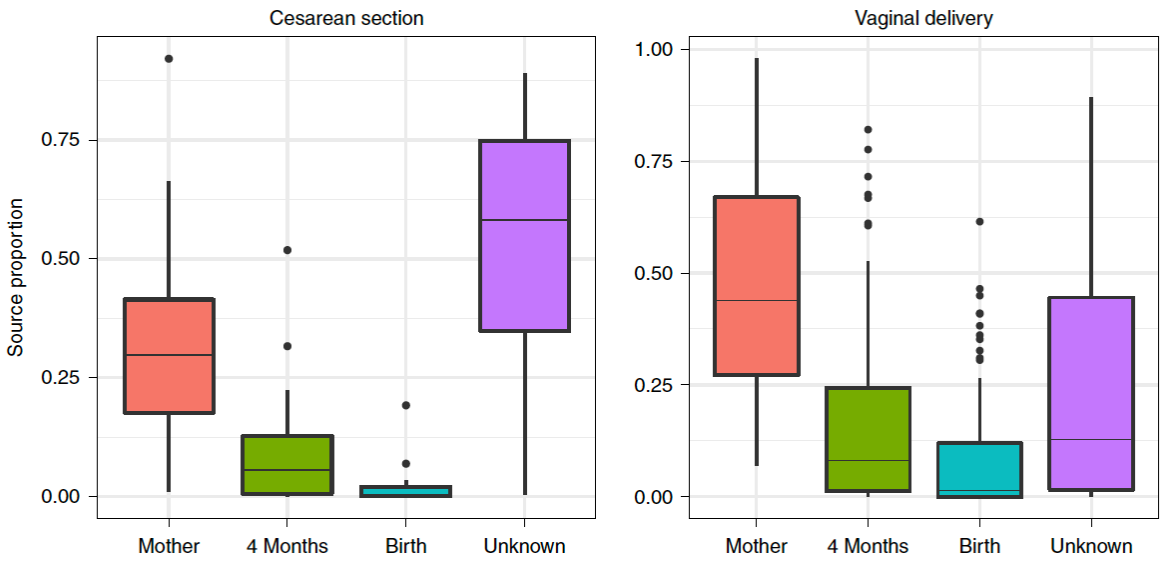

Figure 10. FEAST estimations of source contribution to the sink; that is, gut microbiome of focal infant at 12-months of age. Box plots indicate the median (central lines), IQR (hinges) and the 5th and 95th percentiles (whiskers). Sources: gut microbiome of mother, focal infant at 4 months and focal infant at birth. (n = 98 sinks).

Detecting contamination

To validate FEAST's utility in detecting contamination, we first replicated the analysis of Knights et al.[47] who investigated contamination in settings such as office buildings, hospitals, and research laboratories. In these settings, where the disambiguation between sources was relatively easy, FEAST estimated source contributions consistent with those reported by Knights et al.[47], despite minor discrepancies (Appendix III Fig. S6). Next, we analyzed longitudinal data collected by Lax et al.[46]. In this analysis, we investigated one household, where the inhabitants were genetically related. We used skin samples of inhabitants from several body parts as sources and indoor house surfaces as sinks. Our analysis using FEAST shows that surfaces in home-settings are more diverse than their human sources, and might not be entirely composed of bacteria originated from humans (Fig. 11). Our results stand in qualitative contrast to those of Lax et al.[46], where they found that an overwhelming majority of microbial communities on these surfaces originated from humans. We believe that the difference stems from an underestimation of the unknown source by SourceTracker, which was used in the original analysis of Lax et al. Such underestimation is exacerbated in cases like this, when disambiguation of sources is challenging, i.e., due to all individuals

living in the same house. We further investigated whether we could elucidate the composition of these unknown sources, at the first time point, by including additional source environments from the Earth Microbiome Project. Indeed, in addition to the contribution of the four inhabitants, we find potential evidence for contributions from avian egg product (8%), freshwater fish (8%) and soil (1%). As a consequence, the unknown source contribution was reduced to 5.8% (from approximately 25%; Fig. 11).

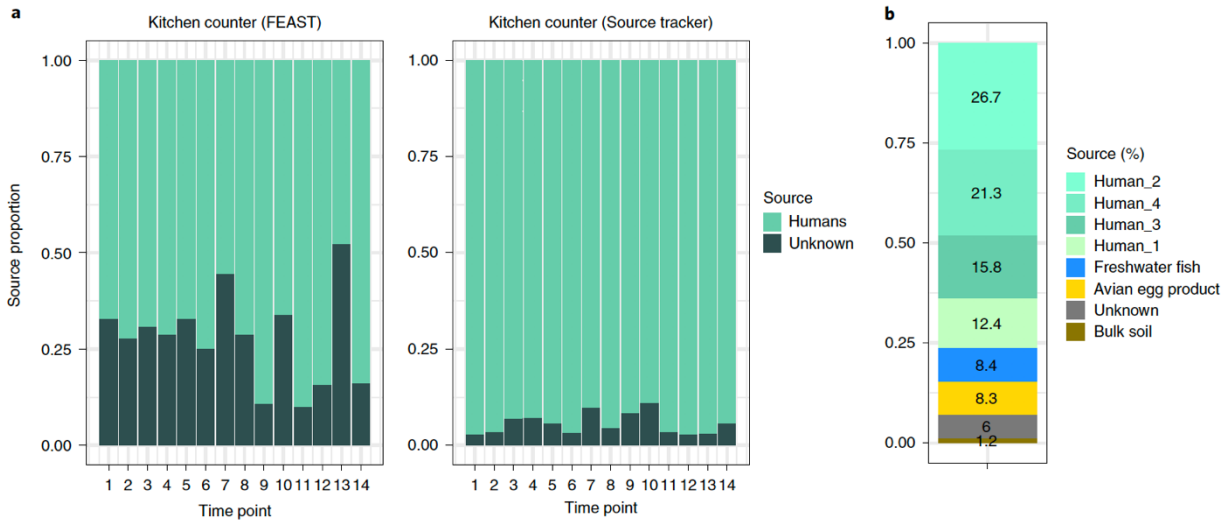

Figure 11. The proportion of the unknown sources in kitchen counter samples using FEAST and SourceTracker. (a) Source estimates considering 12 known human sources (hand, foot and nose across four inhabitants) using data from Lax et al.15 (b) FEAST estimations of source contribution in one house kitchen counter, at the first time point, using additional sources from the Earth Microbiome Project.

Microbial source tracking as a metric of similarity

In the following experiments we used FEAST in a different context—as a metric of similarity. To the best of our knowledge this is a novel application of microbial source tracking. In these experiments we focused on the human gut microbiome, but rather than seeking among sources the contributors to a sink sample (e.g., gut microbiome), we seek to represent each sink as a mixture of `characteristic environments,' — source environments that are similar in composition to the sink and therefore capture its characteristics. We then quantify the similarities between the sink and its characteristic environments using mixing proportions reported by FEAST.

FEAST distinguishes ICU patients from healthy adults

To demonstrate FEAST's utility in distinguishing and characterizing bacteria-related health conditions, we first replicated the analysis of McDonald et al. 2016[57] (Appendix III Fig. S7) in which they characterized a cohort of patients from an intensive care unit (ICU). Indeed, we find that our results using FEAST are consistent with the analysis of McDonald et al., i.e., gut samples from ICU patients are markedly different than those of healthy individuals. Next, we performed an additional analysis that was not included in the original study of McDonald et al.: we used a bidirectional approach, randomly assigning gut samples from the American Gut Project (healthy controls) as either sources or sinks, in addition to the ICU patient gut microbiome sinks (see Methods for complete description). In doing so, we aimed to quantify the similarity between the gut microbiome of ICU patients and healthy controls by comparing their source composition. Using FEAST, we found significant differences in the source composition between the two sink types (p-value = .02551; Appendix III Fig. S8). To verify our findings, we used UniFrac distance[78], Jensen Shannon divergence and the Bray-curtis dissimilarity (Fig. 12, Appendix III Fig. S8), which also captured the differences between the ICU patients and healthy controls (i.e. healthy sources are more similar to healthy sinks). However, we note that there is a large variance in the microbiome similarities among healthy controls, whether they are sources or sinks. We hypothesize that this variance stems from differences between individuals' microbiomes unrelated to their health (e.g., diet). We note that these results should be interpreted with caution, since the healthy and ICU patients are not matched and therefore batch effects or other confounders may affect the results. Nevertheless, if indeed the prediction accuracy is driven by confounders, these results demonstrate that FEAST can capture such confounder information better than existing methods.

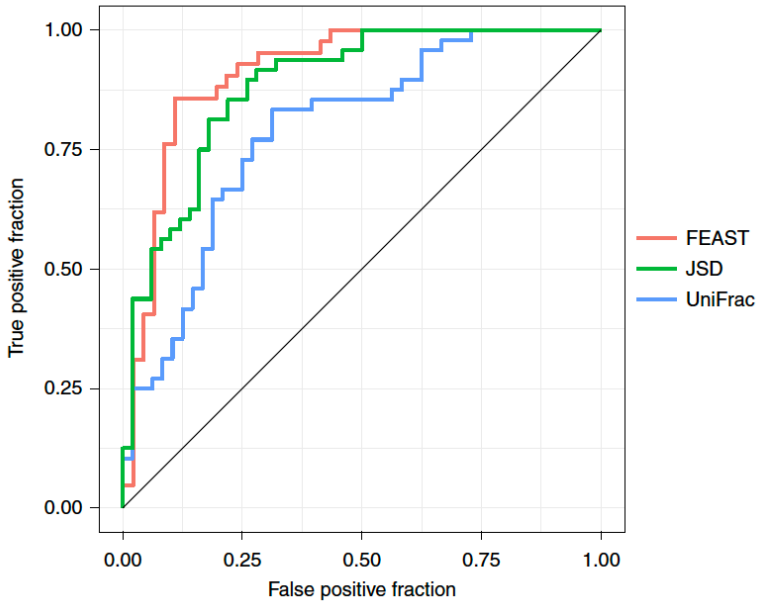

Figure 12. The receiver operating characteristic curve using FEAST, weighted UniFrac and Jensen–Shannon divergence to classify healthy individuals and patients in ICU with dysbiosis. FEAST area under curve (AUC), 0.91; weighted UniFrac AUC, 0.78 and Jensen–Shannon divergence (JSD) AUC, 0.87.

FEAST implicates time-related compositional shifts in a cancer longitudinal study

Considering the utility of FEAST as a method for classifying phenotypes, we sought to also characterize a cohort of cancer patients undergoing allogeneic hematopoietic stem cell transplantation (allo-HSCT). In a study by Taur et al.[58], it was suggested that assessing the gut microbiome of patients undergoing allo-HSCT may identify those at high risk for bloodstream infection (i.e., bacteremia). Notably, many of the patients were found to have intestinal domination, a condition in which at least 30% of the microbiome is comprised of a single bacterial taxon. As the exact nature of the association between compositional shifts in the microbiome and bacteremia is unclear, it is crucial to elucidate the dynamics of microbial community composition in patients undergoing allo-HSCT. This led us to examine whether FEAST can be used as a tool for such an assessment. To this end, we labeled the two consecutive samples from before and during the first event of intestinal domination as sinks, and all corresponding samples from earlier time points as sources (per patient). FEAST revealed a significantly larger proportion of the unknown source in the sink samples with intestinal domination in comparison to the sink samples before intestinal domination (two-

sided t-test p-value < 0.001; Fig. 13, Appendix III Fig. S9). This is expected, as bacterial domination is defined in terms of abundance fractions, so by definition would be reflected in mixture proportions. Nonetheless, this result was not significant using other methods (two-sided t-test p-value = 0.09). We therefore demonstrated FEAST's ability to capture shifts in microbial community composition that may underlie differences between pathogenic and neutral phenotypes.

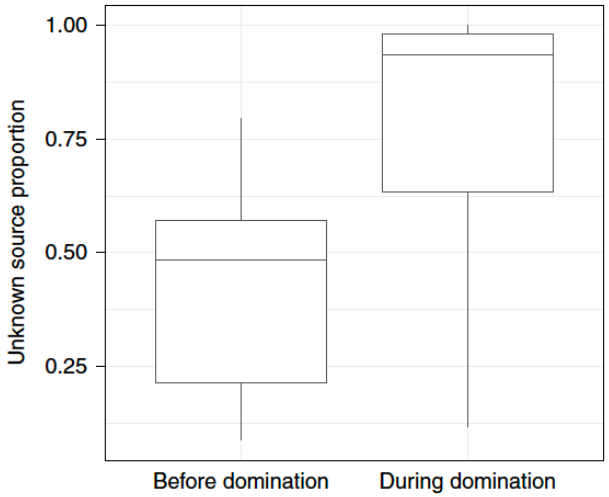

Figure 13. Significant differences in the distribution of the unknown source between sink samples before and during the first event of intestinal domination across 94 patients undergoing allo-HSCT. Box plots indicate the median (central lines), IQR (hinges) and the 5th and 95th percentiles (whiskers).

Discussion

In this work we presented three algorithms, tailored to the spatiotemporal nature of microbial communities, aimed to elucidate how and why microbial communities vary across time and space, and whether these trajectories are consistent across host phenotypes.

First, we presented MTV-LMM, a flexible and computationally efficient tool, which can be easily adapted by researchers to select the core time-dependent taxa, quantify their temporal effects and predict their future abundance. Using MTV-LMM we find that in contrast to previous reports, a considerable portion of microbial taxa in both infants and adults display temporal structure that is predictable using the previous composition of the microbial community. In reaching this conclusion we have adopted a number of

concepts common in statistical genetics for use with longitudinal microbiome studies. We introduce concepts such as time-explainability and the temporal kinship matrix, which we believe will be of use to other researchers studying longitudinal microbiota dynamics, through the framework of linear mixed models. Time-explainability can be informative for selecting autoregressive taxa that are essential to understanding the temporal behavior of the microbiome in longitudinal studies. In particular, such taxa can be used to characterize the temporal trajectories of the microbial community. The temporal kinship matrix can be used to uncover low-rank temporal structure. Using *MTV-LMM*, we have demonstrated that taxa autoregressiveness is a spectrum where certain taxa are almost entirely determined by the community composition at previous time points, some are somewhat dependent on the previous time points, and others are completely independent of previous time points. We further show that *MTV-LMM* can identify autoregressive taxa in both 'evolving' (i.e., infant's gut) and 'stable' (i.e., adult gut) ecosystems.

Next, we presented CTF, an unsupervised method that allows full use of repeated measures and specifically longitudinal data, while accounting for the inherent properties of microbiome sequencing datasets, namely high-dimensionality, sparsity and compositionality. In both simulated and real datasets, CTF outperformed the current state-of-the-art beta-diversity metrics. Although CTF can reveal robust microbial signatures, several considerations are necessary when applying this tool. First, CTF relies on an assumption that the underlying data is of low rank. This assumption can be violated, making CTF inappropriate to use, such as when the data are driven by a gradient rather than discrete groupings (for example, the 88 Soils dataset[56]). Our implementation of CTF estimates the underlying rank and informs the user if the data does not meet this requirement[81]. Second, CTF, like other beta-diversity metrics, does not directly account for the presence of confounders that may affect downstream clustering, requiring additional validations. Finally, although CTF leverages repeated measures to account for interindividual variation and is optimal in the case of a synchronization event (for example, treatment, diet), it is permutation invariant and does not take into account the ordering of longitudinal data. In addition to longitudinal datasets as benchmarked here, CTF could also be used for spatially repeated measurements. This includes studies where samples are collected contemporaneously; for example, where multiple body sites are measured (such as skin and saliva)

or sites with different phenotypes (such as lesioned versus adjacent non lesioned skin). Furthermore, CTF could be used to analyze other types of data with high interindividual variation, such as metabolomics or proteomics. In summary, CTF leverages the power of repeated measures study design to elucidate biological changes while accounting for interindividual variability. We propose the use of this tool both for the reanalysis of existing datasets and for future microbial community research.

Finally, we presented FEAST, a method designed to address an important need in the rapidly evolving field of microbiome research—namely, to quantify the fraction of each source environment in a target microbial community (sink), through a natural, scalable statistical model. As a result, it provides a computationally efficient tool that can simultaneously evaluate hundreds to thousands of potential source environments, as well as the contribution of an unknown, uncharacterized source, outperforming state-of-the-art methods in terms of both speed and accuracy. The utility of FEAST is established in two different contexts. First, we used FEAST as it was originally intended—to quantify the contribution of different source environments to a target microbial community. In this context, we were able to address questions surrounding succession and initial colonization of microbial species. Specifically, using FEAST we quantitatively reaffirmed the findings of Backhed et al., [16] who demonstrated that gut microbiota of infants delivered by cesarean section showed significantly less resemblance to their mothers' compared to vaginally delivered infants. Second, we used FEAST as a metric of similarity. In this context, FEAST can help researchers better understand the compositional characteristics of the human microbiome. We showed the ability of FEAST to differentiate between the gut microbiome of ICU patients experiencing dysbiosis and that of healthy controls as well as in patients experiencing intestinal domination. These results suggest that FEAST may be useful in distinguishing and characterizing phenotypes or conditions related to microbial injury. Furthermore, by highlighting novel differences among source composition, FEAST may contribute insight to downstream analyses aiming to implicate differences between healthy and diseased phenotypes at the taxa level.

We expect these methods to drive research into the role of microbiome dynamics in health and disease.

Appendix I

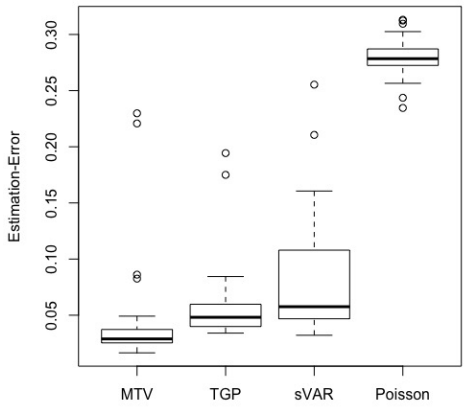

Figure. S1 Estimation errors of *MTV-LMM*, TGP-CODA, sVAR and ARIMA Poisson models. Estimation errors calculated using synthetic data, illustrating realistic dynamics and abundance distribution, with 200 taxa and 70 time points, as suggested by Aijo et al. 2018[37]. Estimation error is defined to be the Euclidean distance between estimated relative abundance and the true ones per time point (Wilcoxon test P-value *MTV-LMM* vs. TGP-CODA = 0.01501, *MTV-LMM* vs. sVAR P-value = $2.224e-08$).

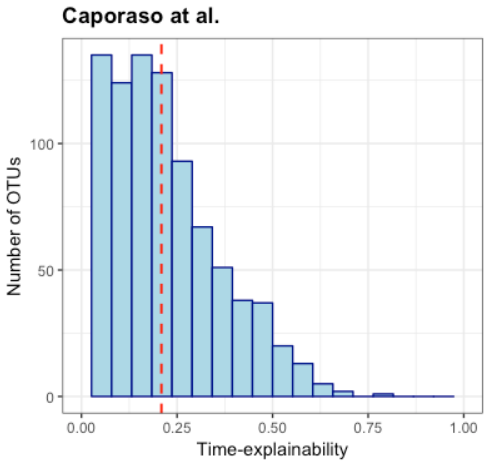

Figure. S2 Time-explainability distribution. Time-explainability distribution in Caporaso et al. dataset. The average time-explainability in this cohort is 0.2 (denoted by a dashed line).

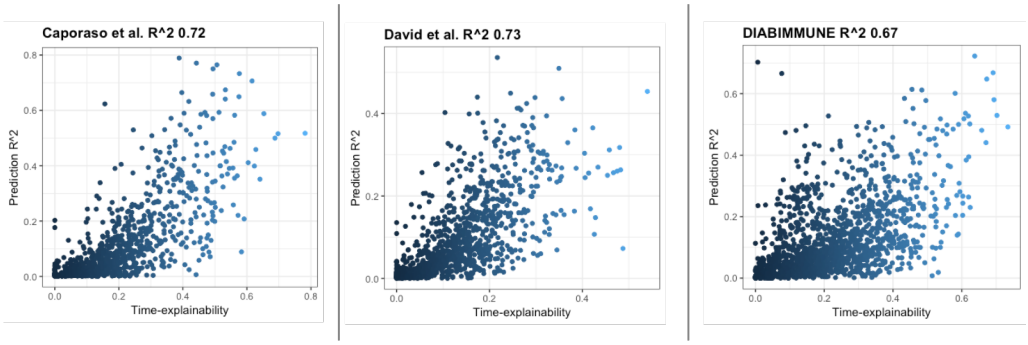

Figure. S3 Prediction accuracy (R2) as a function of time-explainability.

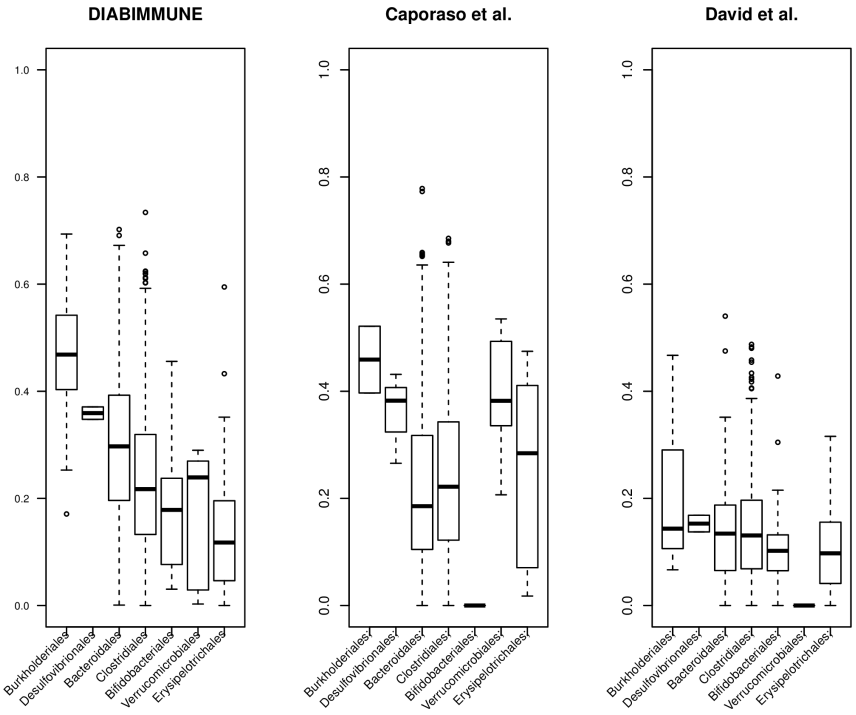

Figure. S4 Time-explainability distribution differ by taxonomic order across datasets. Boxplots illustrate the time-explainability distribution across all datasets. Presented are the top seven orders in the DIABIMMUNE dataset.

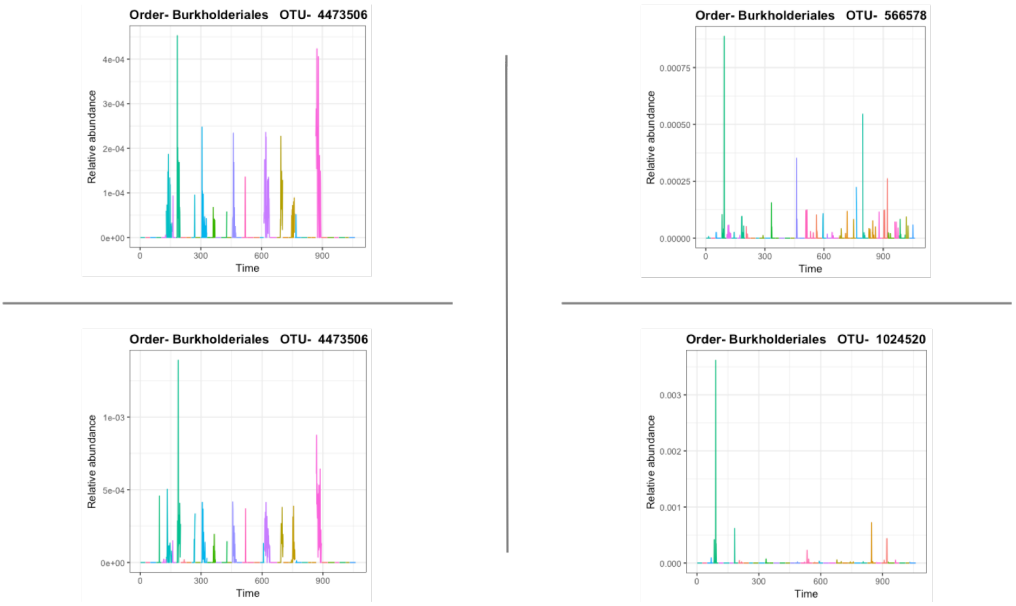

Figure. S5 Relative abundance of taxa from order Burkholderiales in the DIABIMMUNE dataset, colored by individual. Right hand-side, the autoregressive taxa, taxa with a significant time-explainability component (top and bottom: time-explainability = 0.49, 0.35, 95% CI = [0.4, 0.58], [0.33, 0.36]). Left hand-side are the non-autoregressive taxa.

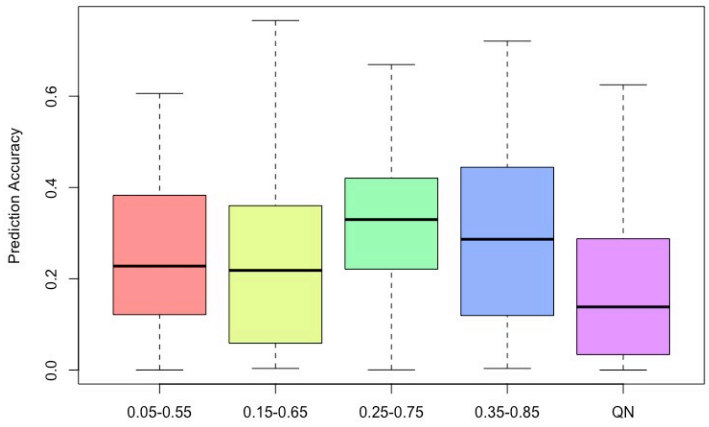

Figure. S6 Sensitivity analysis of the binning parameters used to normalize microbial abundance. Each boxplot corresponds to the prediction accuracy distribution under different binning parameters, i.e., a 25% lower quantile and a 75% upper quantile compared to 5% and 55%, 15% and 65%, 35% and 85%, and quantile normalization. This analysis was conducted on a simulated microbial community composed of 50 species over 50 time points (data was generated as described in the simulation section).

	David et al.	Caporaso et al.	DIABIMMUNE
AR(1)	$< 2.210^{-16}$	10^{-7}	10^{-4}
sVAR	$< 2.210^{-16}$	$< 2.210^{-16}$	$< 2.210^{-16}$
ARIMA Poisson-regression	$< 2.210^{-16}$	$< 2.210^{-16}$	$< 2.210^{-16}$

S1 Table Predictive accuracy comparison. P-values of the Wilcoxon test comparing the prediction accuracy ($R2$) of *MTV-LMM* with the prediction accuracy of the AR(1) model, the sVAR model and the ARIMA (1, 0, 0)-Poisson regression model.

Supplementary Note 1 Simulation study.

To further test the capabilities of MTV-LMM to accurately estimate the temporal dynamics and predict the future abundance of microbes within a community, we used synthetic data, illustrating realistic dynamics and abundance distribution, as suggested by Aijo et al. 2018 [37]. Specifically, we consider the Subject A time series from David et al. (2014) [11] and match the relative abundances and dynamics of taxa in synthetic data using real data:

1. We filter the proportion estimates series using a running median filter of length 15; $y_{\{filt,t\}} = median\,(y_{\{t-7\}}, y_{\{t-6\}}, \dots, y_{\{t-1\}}, y_{\{t\}}, y_{\{t+1\}}, y_{\{t+2\}}, \dots, y_{\{t+6\}}, y_{\{t+7\}})$ in order to reduce the amount of noise present in estimates. The filtered estimates are re-normalized to ensure that they sum up to one at each time point.
2. We discard those bacterial species that are lowly abundant (average proportion is less than a threshold) followed by a re-normalization step leaving us noise-free relative abundances of 200 bacterial species.
3. We transform the simplex-valued estimates to real space using the inverse softmax function to add noise and sampling zeros.
4. We add Gaussian distributed noise with zero-mean and standard deviation (SD) $\sigma = 0.5$ and impose a predefined number of sampling zeros by setting corresponding log odds ratios to -10, i.e., to a value that is much smaller than the other values.

5. Noisy relative abundances are obtained by projecting the values onto the simplex using the softmax function.
6. Noisy (over dispersed and zero-inflated) count data (N_t is sampled from the Poisson distribution with the rate $\lambda = 10{,}000$ are generated from Multinomial distribution using the noisy relative abundances of the part of the Subject A time series (days from 60 to 140) that is highly dynamic David et al. (2014).

Following [37] we evaluate the performance of the model using the 'estimation-error', defined to be the Euclidean distance between estimated relative abundance and the true ones per time point. The 'estimation-error' was calculated on a held-out test set that was kept hidden from the algorithm.

Appendix II

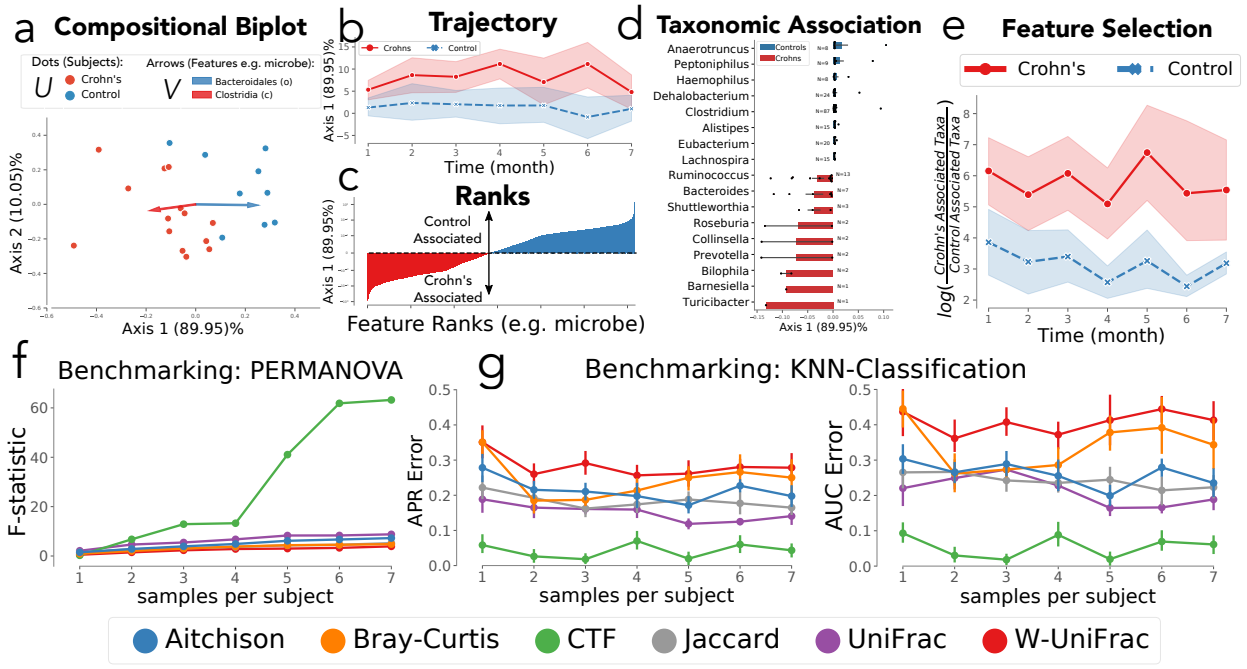

Figure S1. IBD dataset benchmarking. CTF was applied to longitudinal 16S data from Halfvarson et al.[8]. Multiple distinct downstream analyses were generated from the CTF output: **(a)** The subject (U) and feature (V) loadings can be visualized as compositional biplot ordinations, using the top two ordination axes where each point represents a subject's time series and the arrows represent the top-ranked microbial features differentiating the subjects. **(b)** The compositional biplot axis driving phenotypically relevant separation was used to track sample distance over time (referred to here as a 'trajectory'). This trajectory demonstrates that Crohn's disease samples are compositionally distinct from healthy control samples across the entire sampling timeline. **(c)** Differentially abundant microbes most associated with the phenotypes were identified by plotting the feature rankings by the major axis of separation. The highest ranked sOTUs revealed by CTF are associated with Crohn's-disease while the lowest are associated with the control group. **(d)** sOTU feature rankings averaged by genus colored by control (n=9 subjects; 950 sOTUs; blue) and Crohn's (n=14 subjects; 950 sOTUs; red; number of sOTUs in each genus annotated on plot). **(e)** These phenotype-associated taxa were chosen based on the feature ranks to identify reference frames (log-ratios of sequencing counts) that can differentiate phenotypes, distinguishing healthy (n=9 subjects) from Crohn's (n=14 subjects) disease samples. These trajectories and differentially abundant taxa are supported by previous findings in the IBD literature[8,19–21] **(f)** Disease status PERMANOVA F-statistic among different distance metrics. (**g**) K-Nearest Neighbor classification compared by APR (left) and AUC (right) among different distance metrics. Error bars represent standard error of the mean.

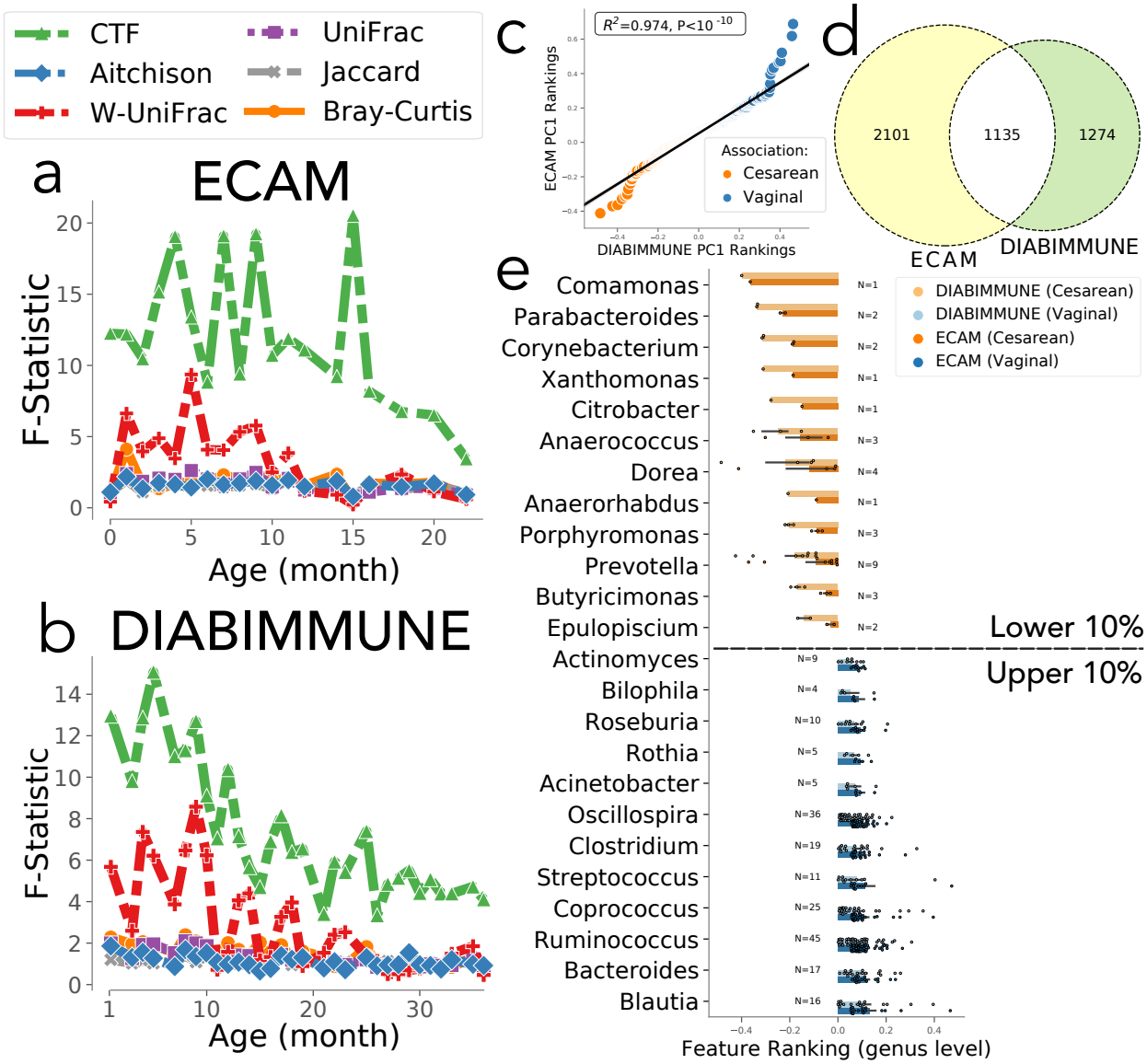

Figure S2. Feature rankings distinguishing birth-modes across the ECAM and DIABIMMUNE datasets are tightly correlated. (a & b) PERMANOVA F-statistic (y-axis) separating vaginal vs cesarean birth-mode colored by distance metric for ECAM (top) and DIABIMMUNE (bottom). (**c**) Regression plot between sOTUs ranked in ECAM and DIABIMMUNE datasets (R^2=0.974, $P<10^{-10}$); Pearson correlation shown with two-tailed p-value. (**d**) Venn diagram of the number of disjoint and shared sOTUs between datasets. (**e**) The top and bottom 10% ranked sOTUs averaged by genus in ECAM and DIABIMMUNE colored by vaginal (blue) and cesarean (orange) birth modes (number of sOTUs in each genus annotated on plot). Error bars represent standard error of the mean.

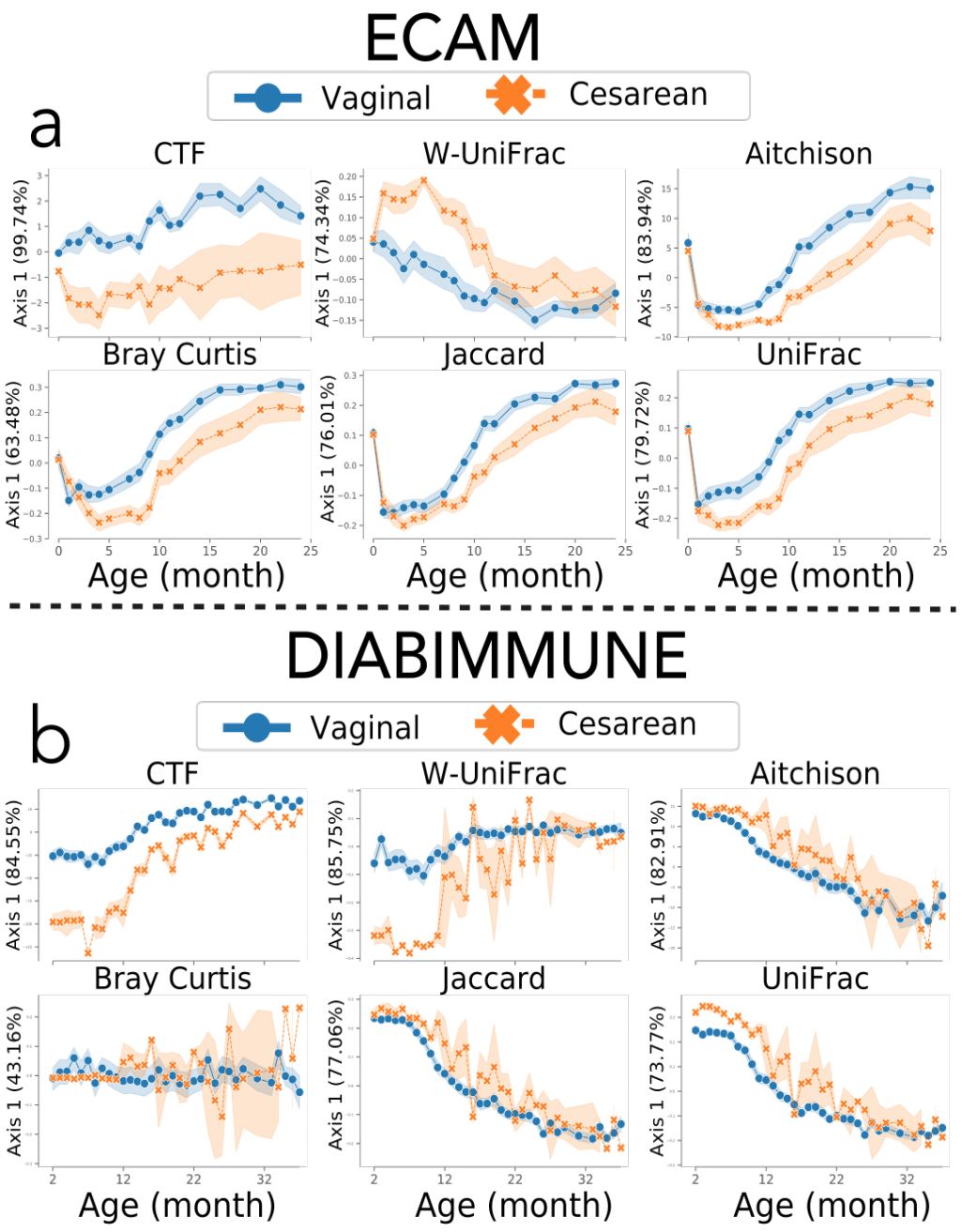

Figure S3. CTF outperforms traditional distance metrics in distinguishing samples by birth-mode over time. (a & b) Comparison between the ECAM (top) and DIABIMMUNE (bottom) infant development studies with the first principal component (y-axes) of various distance metrics over time (x-axes) colored by vaginal (blue) and cesarean (orange) birth-modes. The relative percent explained variance is the fraction of the first component divided by the top 3 components to normalize eigenvalues among methods. Error bars represent standard error of the mean.

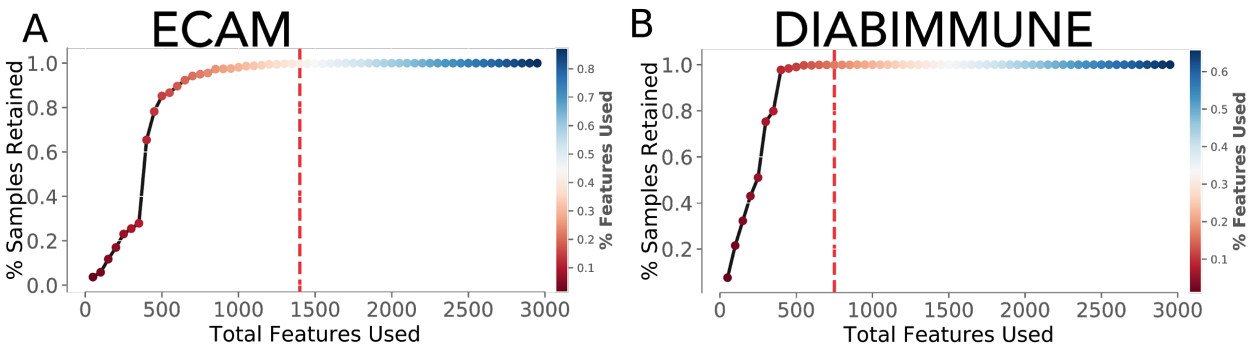

Figure S4. Selecting the number of features used in the log-ratio to prevent sample dropouts from zeros. The percent of samples retained (y-axis) when including a number of ranked features (x-axis) in the log-ratio colored by the percent of features used from the total dataset for ECAM (**A**) and DIABIMMUNE (**B**). The red line represents the number of features used in the final log-ratio for each dataset.

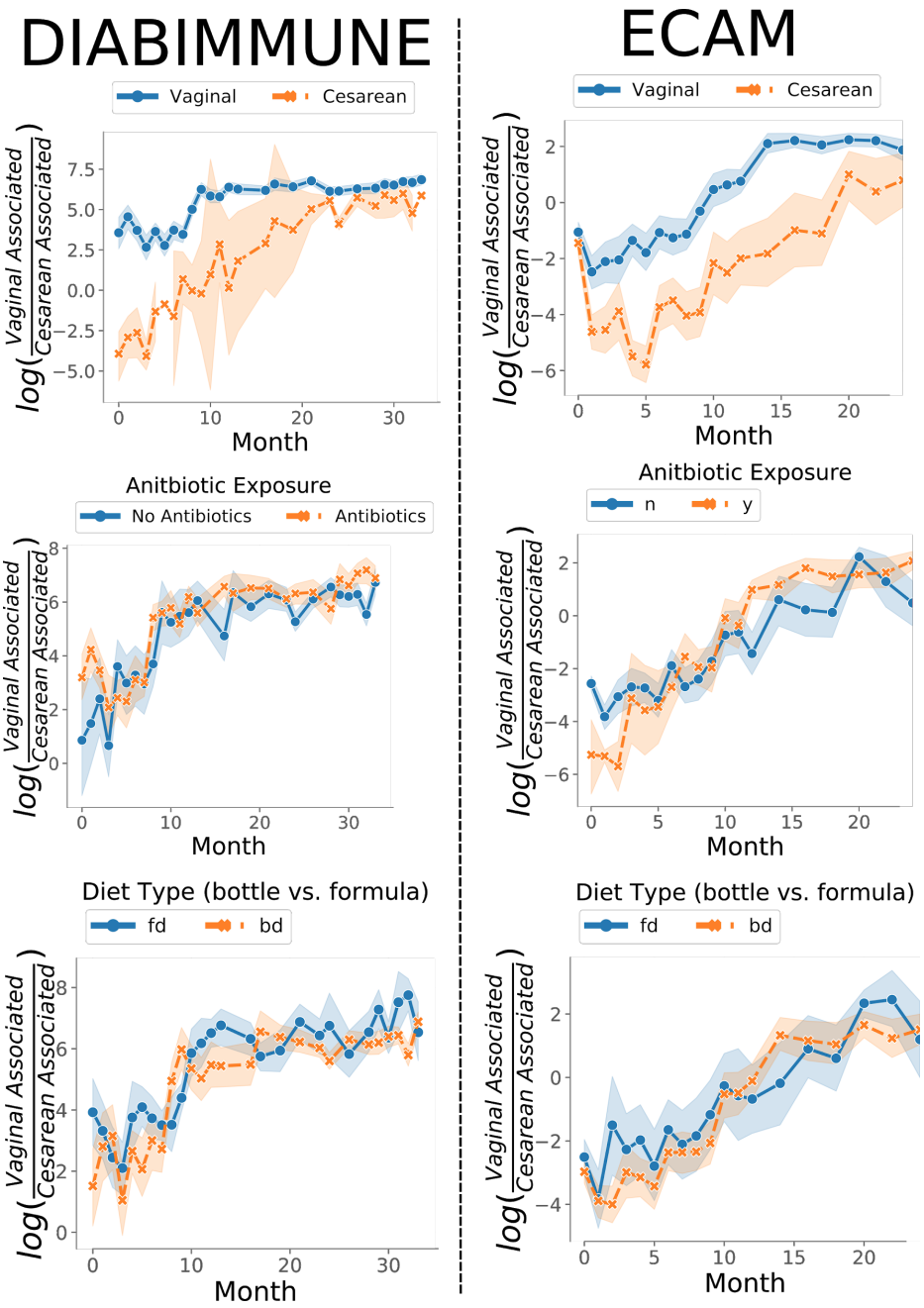

Figure S5. Birth-mode ratios designed from CTF feature rankings distinguish samples by birth-mode over time. The log-ratios for the ECAM (1400 sOTUs) and DIABIMMUNE (750 studies are plotted on the y-axis over time (x-axis) showing separation by birth-mode using these ratios. This grouping of subjects is not confounded by antibiotics exposure (yes/no) or by diet. Error bars represent standard error of the mean.

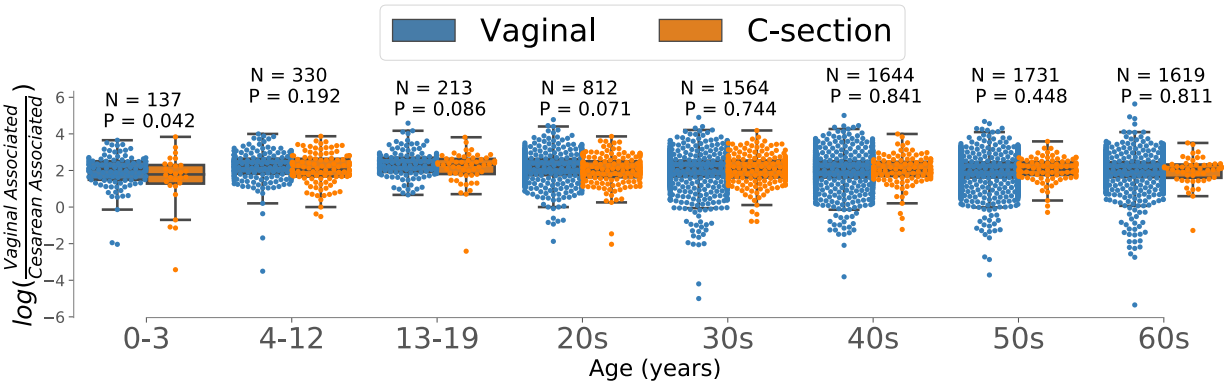

Figure S6. Birth-mode microbial signature in AGP dataset. Birth-mode microbial signature (log-ratio of the 532 most highly and 532 most lowly ranked features that were all identified in the DIABIMMUNE, ECAM, and AGP datasets) is plotted on the y-axis in all sub-panels. Birth-mode microbial signature over age groups (years; x-axis) colored by cesarean (orange) and vaginal (blue) birth modes in the AGP dataset. The box plots represent the minimum, maximum, median, first and third quartile values (shaded region). Significance was evaluated by a two-sided t-test.

Sequencing Depth (seq/sample)	Comparison Method	CTF Fold-Increase	CTF Percent-Increase	
		F-stat Fold Increase	APR	AUC
500	**Aitchison**	3.90 ± 1.93	23.04 ± 6.86	33.48 ± 8.92
	Bray-Curtis	4.78 ± 2.42	24.49 ± 6.66	42.61 ± 8.32
	Jaccard	4.69 ± 2.36	18.55 ± 7.24	26.74 ± 12.31
	UniFrac	3.01 ± 1.48	18.99 ± 4.94	28.48 ± 5.89
	W-UniFrac	5.55 ± 3.08	18.91 ± 7.63	38.70 ± 15.35
1000	**Aitchison**	3.20 ± 1.28	23.62 ± 6.50	33.48 ± 10.28
	Bray-Curtis	4.23 ± 1.80	33.04 ± 7.58	51.09 ± 12.69
	Jaccard	3.66 ± 1.61	23.77 ± 7.24	33.48 ± 10.14
	UniFrac	2.23 ± 0.85	17.39 ± 1.73	24.57 ± 2.63
	W-UniFrac	5.69 ± 2.12	26.01 ± 4.76	46.74 ± 8.80
10000	**Aitchison**	3.67 ± 1.94	22.97 ± 2.90	30.22 ± 3.60
	Bray-Curtis	6.80 ± 4.11	24.49 ± 7.02	42.17 ± 12.89
	Jaccard	3.54 ± 1.97	14.64 ± 3.66	19.57 ± 5.79
	UniFrac	1.91 ± 1.02	11.59 ± 4.54	18.48 ± 2.48
	W-UniFrac	9.94 ± 6.33	12.90 ± 5.55	20.00 ± 10.35

Table S1. CTF shows improvement over traditional distance metrics in simulations across different sequencing depth. Fold increase in PERMANOVA f-statistic (left) or percent increase in K-Nearest Neighbor classification (right) by CTF over other distance metrics in simulated dataset.

AUC and APR percent increase mean ± s.d. across all time points for mean 100-fold cross-validation at each time point.

PERMANOVA F-statistic fold-increase mean ± s.d. across all time points.

	APR		PERMANOVA F-statistic CTF Fold-Increase	
Comparison Method	**DIABIMMUNE**	**ECAM**	**DIABIMMUNE**	**ECAM**
CTF	0.983 ± 0.001	0.768 ± 0.007	1.0 ± 0.0	1.0 ± 0.0
Aitchison	0.885 ± 0.003	0.552 ± 0.004	6.13 ± 0.39	8.11 ± 1.17
Bray-Curtis	0.87 ± 0.002	0.589 ± 0.006	5.00 ± 0.24	8.88 ± 2.53
Jaccard	0.88 ± 0.002	0.592 ± 0.006	6.40 ± 0.48	8.66 ± 1.05
UniFrac	0.874 ± 0.001	0.552 ± 0.005	5.32 ± 0.22	7.79 ± 1.09
W-UniFrac	0.864 ± 0.003	0.582 ± 0.007	3.94 ± 0.45	10.41 ± 4.94

Table S2. CTF improves over existing methods across all time and increases the number of significant time points. Comparison of KNN-classification and PERMANOVA quantitative benchmarking between CTF and existing methods for DIABIMMUNE and ECAM datasets.

APR mean ± s.d. across all time points for mean 100-fold cross-validation at each time point.

PERMANOVA F-statistic fold-increase mean ± s.e. across all time points.

		Intercept	birth-mode	month	birth-mode:month	Group Var
DIABIMMUNE	**Coef.**	-2.491	6.362	0.306	-0.204	1.791
	Std.Err.	0.785	0.832	0.023	0.025	0.215
	z	-3.173	7.644	13.306	-8.259	-
	P>|z|	0.002	<.001	<.001	<.001	-
	[0.025	-4.03	4.731	0.261	-0.252	-
	0.975]	-0.952	7.993	0.351	-0.156	-
ECAM	**Coef.**	-4.362	2.097	0.16	0.067	3.279
	Std.Err.	0.483	0.641	0.025	0.032	0.335
	z	-9.037	3.272	6.395	2.131	-
	P>|z|	0	0.001	0	0.033	-
	[0.025	-5.308	0.841	0.111	0.005	-
	0.975]	-3.416	3.353	0.209	0.129	-

Table S3. Linear mixed-effects model results on birth mode associated log-ratios is significant by birth mode for both ECAM and DIABIMMUNE. P>|z| represents two-tailed p-value of the approximated z-statistic for a given parameter estimate.

Appendix III

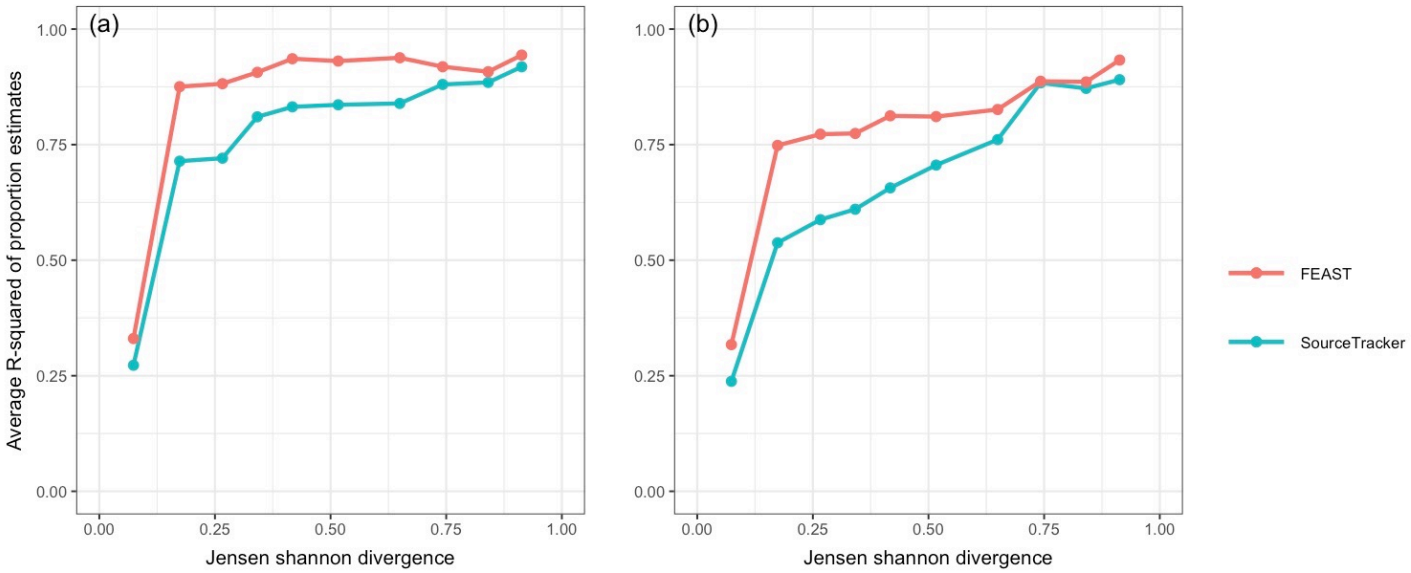

Figure S1. The accuracy of FEAST and SourceTracker using data-driven synthetic mixtures. The accuracy of FEAST and SourceTracker on simulated data. Each simulation was performed using 10 real source environments and simulated sinks. The x-axis is average Jensen-Shannon divergence value across known sources. The y-axis represents correlation across all source environments between true and estimated mixing proportions, measured by (a) the squared Pearson correlation coefficient averaged across sources, and (b) the squared Spearman correlation coefficient averaged across sources.

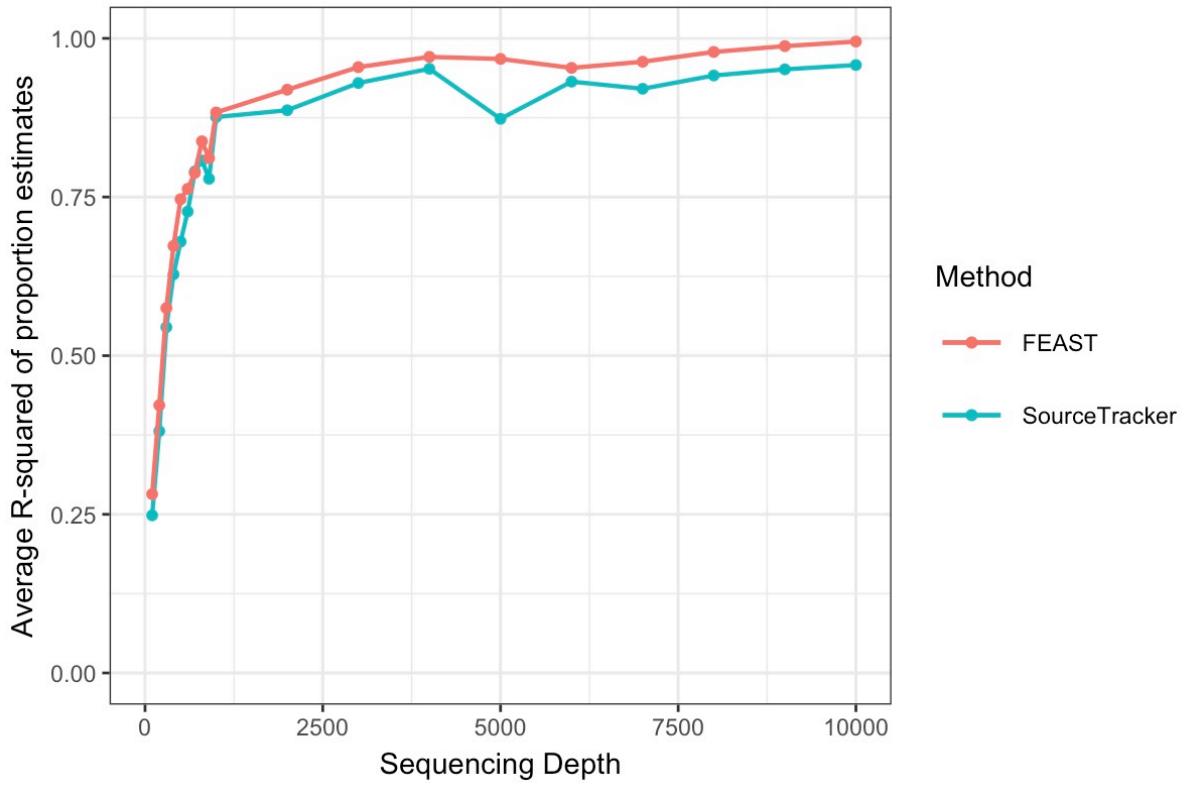

Figure S2. Evaluation of FEAST and SourceTracker through varying levels of sequencing depth. Evaluation of FEAST and SourceTracker through varying levels of sequencing depth. Similarity of sequences remained constant (Jensen-Shannon divergence = 0.95, trivial to disambiguate), while sequencing depth was set to vary in the range 100-10,000.

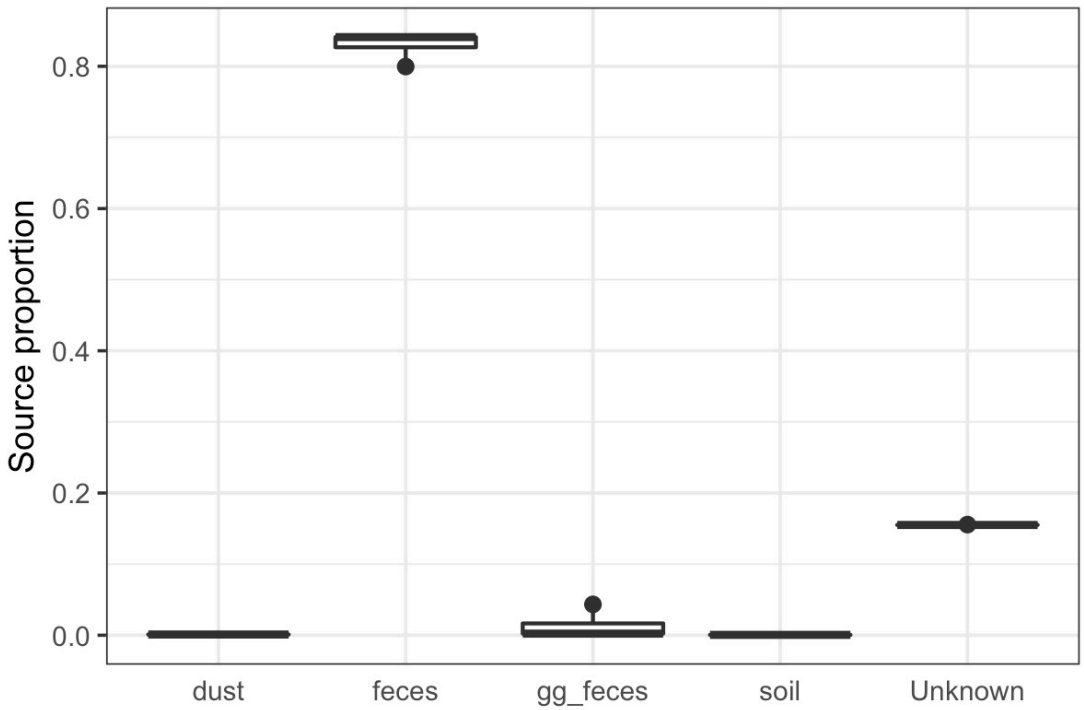

Figure S3. The expected variance in FEAST's output. The expected variance in FEAST's output using the dataset from McDonald et al. We used the gut microbiome of one, randomly selected, ICU patient as a sink, and the sources considered by McDonald et al. : 126 healthy controls, 126 samples of mammalian corpse decomposition, 126 samples of the gut from healthy children, and 126 samples from indoor house surfaces. By repeating this analysis 100 times and calculating the standard deviation of each source we demonstrate that the variance in FEAST's output is very small (i.e., sd(dust) = 7.7e-05, sd(healthy adults' feces) = 0.01, sd(healthy children's feces) = 0.01,sd(soil) = 5e-05, sd(unknown) = 8.5e-05).

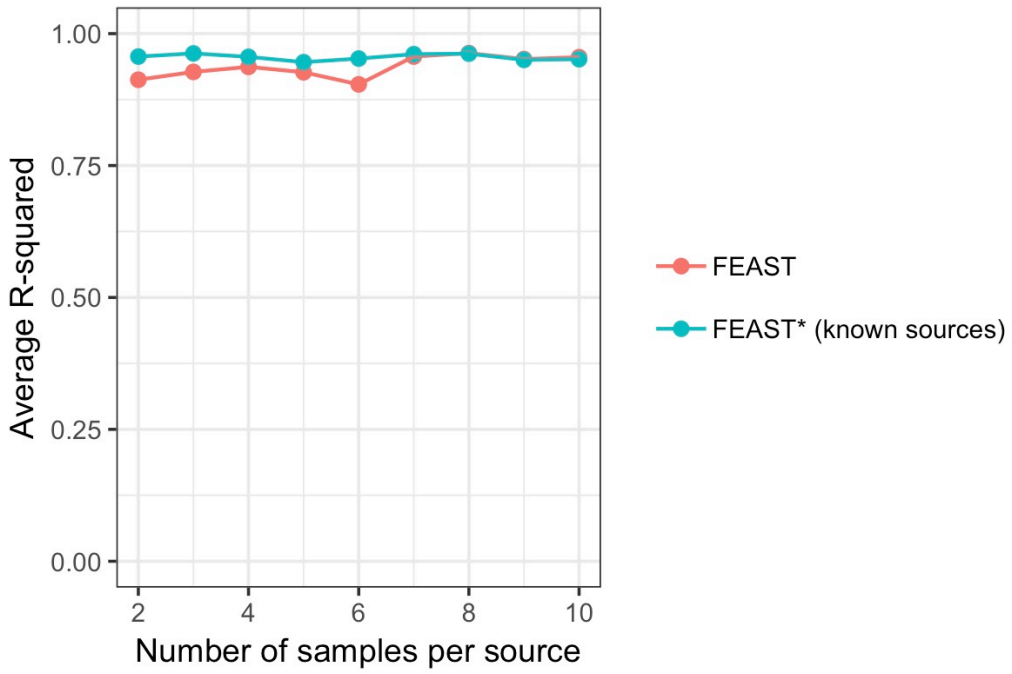

Figure S4. The effect of noisy samples among sources on prediction accuracy. As we increase the number of samples per source, FEAST's prediction accuracy improves, however this effect is moderate (squared Pearson correlation ranges from 0.9 - 0.99, Jensen-Shannon divergence values range from 0.87-0.92).

Figure S5. The source proportions using SourceTracker. SourceTracker estimations of source contribution (the gut microbiome of mother, infant at 4 months and infant at birth) to the gut microbiome of 12-month-old infants. According to SourceTracker differences between C-section (n = 15) and Vaginally-delivered (n = 83) infants in terms of maternal contribution are not significant (two-sided t-test p-value = 0.6408). Box plots indicate the median (central lines), interquartile range (hinges), and the 5th and 95th percentiles (whiskers).

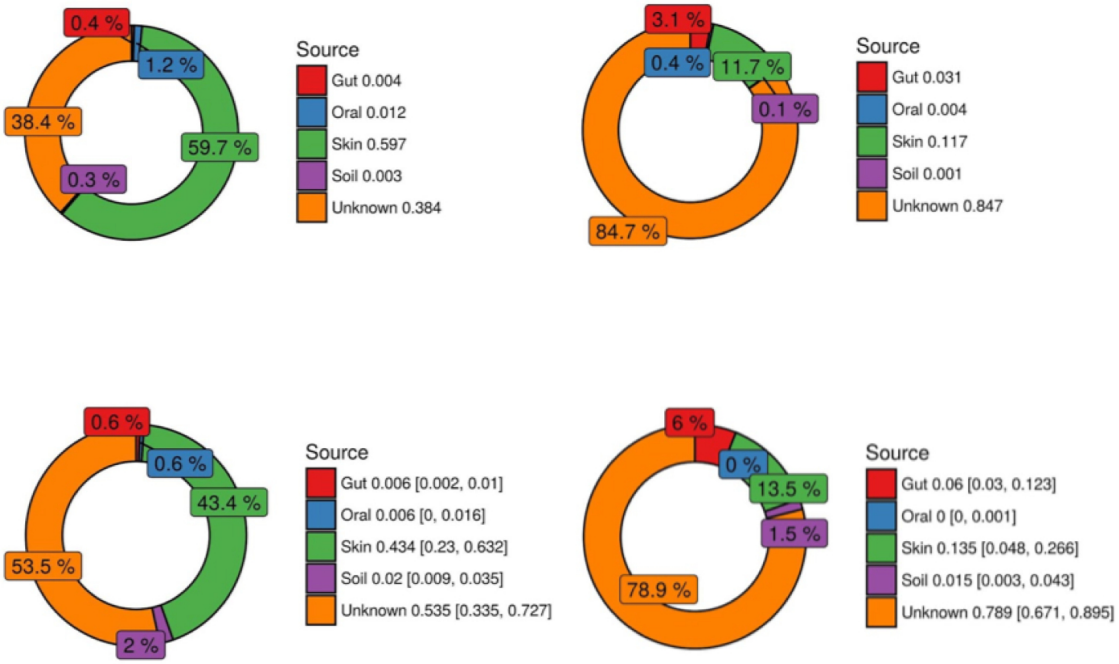

Figure S6. Detecting contamination in lab-settings. FEAST and SourceTracker report consistent proportions of contamination, despite minor discrepancies in a lab-setting (left: keyboard, right: Counter). Estimates on the top row were reported by SourceTracker and estimates on the bottom row were reported by FEAST.

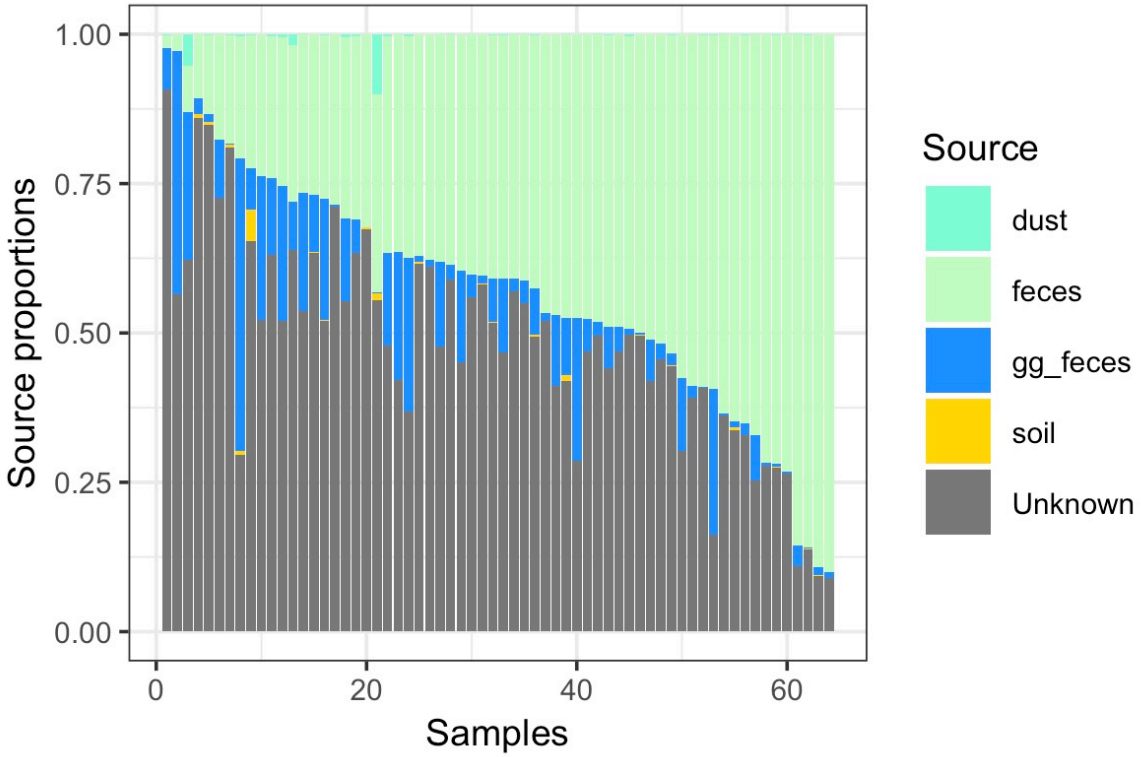

Figure S7. Gut microbiome samples from ICU patients are not reminiscent of gut samples from healthy individuals. Gut samples from ICU patients are not reminiscent of gut samples from healthy individuals. We used the gut microbiome of each ICU patient (at discharge or after 10 days) as a sink, and the sources considered by the original study (McDonald et al. 2016): 126 samples from the American Gut Project (healthy controls), 126 samples of mammalian corpse decomposition, 126 samples of the gut from healthy children (Global Gut study) , and 126 samples from indoor house surfaces.

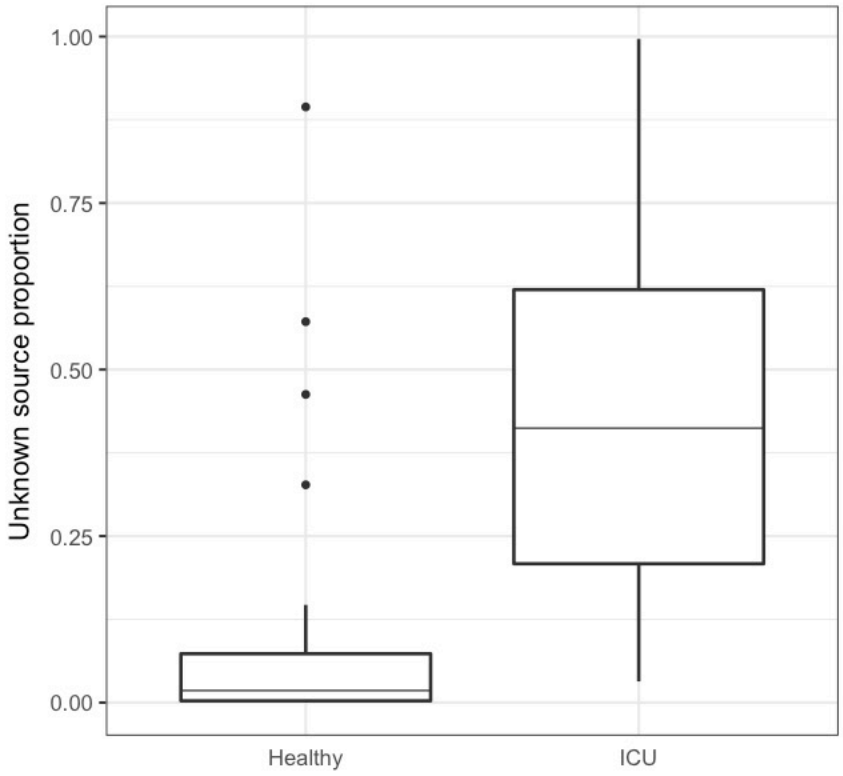

Figure S8. Unknown source distribution across sink samples (ICU patients vs. healthy individuals). The distribution of the unknown source across sink samples - healthy individuals and ICU patients (n = 100).

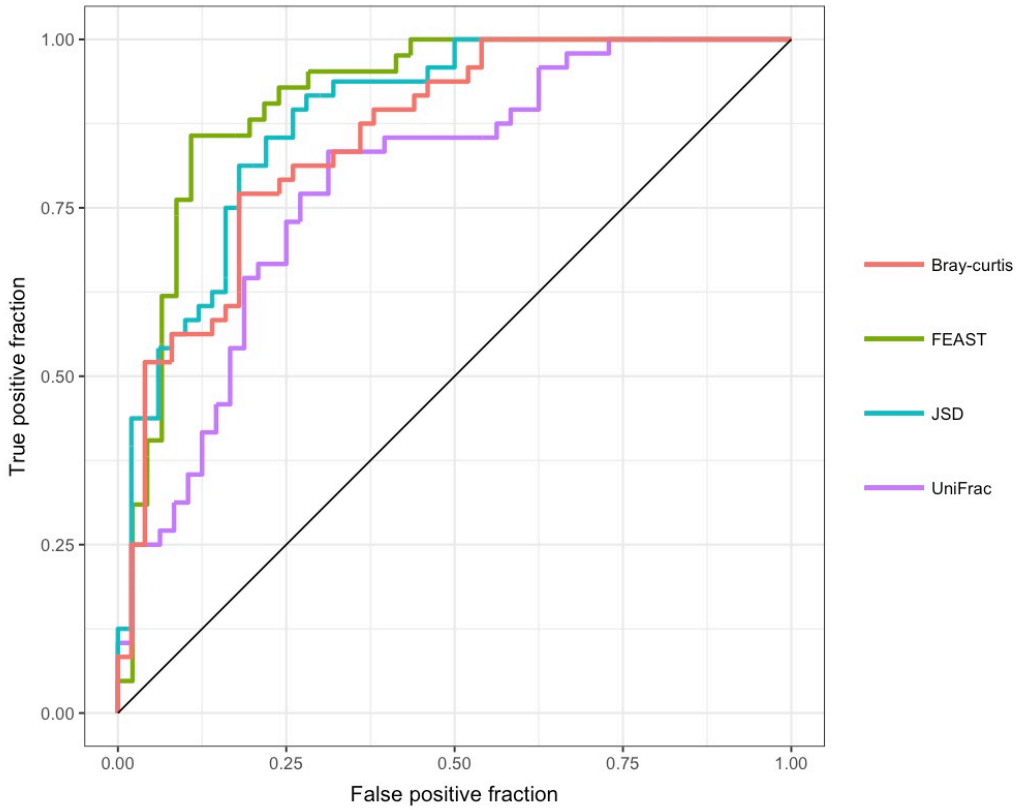

Figure S9. Distinguishing between ICU patients and healthy individuals. The receiver operating characteristic curve (ROC curve) using FEAST, Weighted UniFrac, Bray-curtis and Jensen Shannon divergence to classify healthy individuals and ICU patients with dysbiosis. FEAST AUC = 0.91, Weighted UniFrac AUC = 0.78, Jensen Shannon divergence AUC = 0.87, Bray-curtis AUC = 0.86.

.

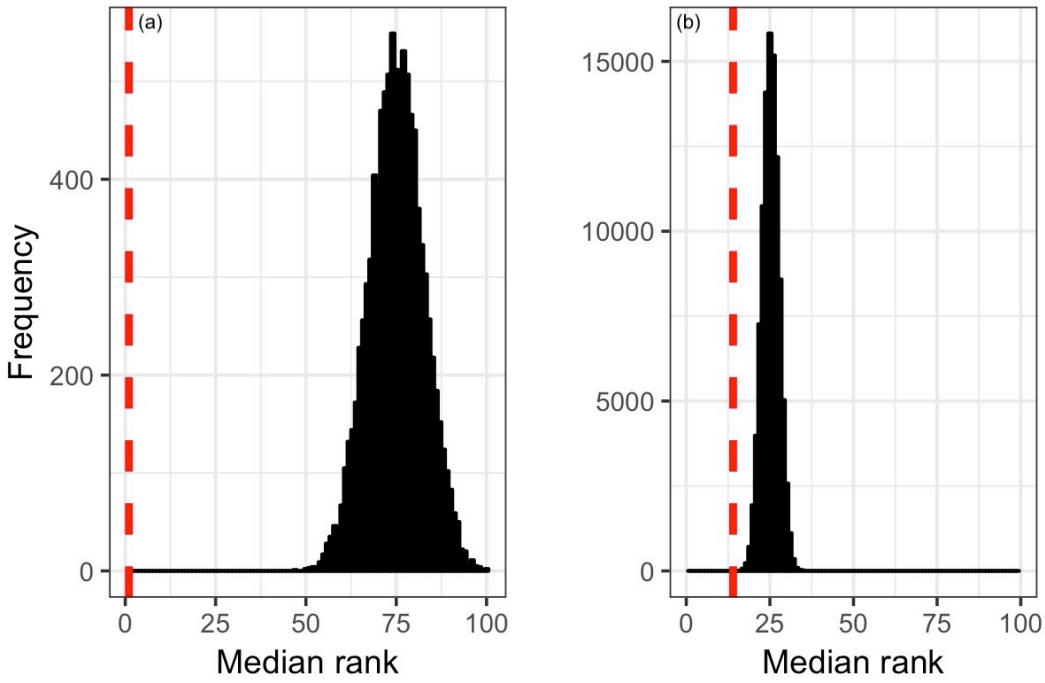

Figure S10. The source contribution across maternal samples. Distribution of the median random maternal rank in two scenarios: (a) all maternal and early infant samples (from all the infants in the study) were considered as potential sources (n = 293 sources), and (b) only the maternal samples were considered as potential sources (n = 98 sources). In both scenarios samples taken from infants at age 12 months were considered as sinks (n = 98 sinks). The red vertical line in each figure corresponds to the actual median rank of the maternal contribution.

Supplementary Table: Running time comparison

Number of sources	5	10	50	100	500	1000
SourceTracker	00 : 05 : 18	00 : 09 : 06	05 : 39 : 00	11 : 43 : 02	54 : 34 : 02	71 : 07 : 00
FEAST	00 : 00 : 09	00 : 00 : 36	00 : 07 : 42	00 : 15 : 54	00 : 47 : 46	01 : 35 : 30

Table S1. Running time (hh:mm:ss) comparison across multiple source environments, randomly sampled from the Earth Microbiome Project. Sequencing depth is 10,000 reads per source.

Supplementary Note: Main simulation study In order to examine the accuracy of *FEAST*, we used multiple source environments with varying degrees of overlap in their distribution by randomly sampling from the Earth Microbiome Project. Each source environment was sub-sampled to contain $10,000$ reads. In each iteration of our simulation we sampled $K+1$ known environments and used them to build a synthetic sink, with different mixing proportions. In order to simulate an unknown source, we use only K source environments as our sources.

The simulation procedure was as follows. For each $l = 1 : T_1$ (different Jensen Shannon divergence values):

1. Draw $K+1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.
2. Draw noisy realization of $S_1, \ldots, S_{K+1}$ from the Multinomial distribution (denoted $\tilde{S_k}$).
3. For each $i = 1 : T_2$ (different mixing proportions):
   (a) Generate random mixing $m \sim Pareto(\alpha > 0)$, where $\sum m = 1$.
   (b) Set the sink sample abundances to $\sum_{k=1}^{K+1} m_k S_k$ per taxa.
   (c) Estimate the known source proportions in the sink using $\tilde{S_1}, \ldots, \tilde{S_K}$.
   (d) Estimate the unknown source proportions in the sink.
4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions per source and average across sources.

5. Calculate the average Jensen-Shannon divergence of m (based on the pairwise Jensen-Shannon divergence).

In the simulations presented we used $T_1 = 10$, $T_2 = 30$, $K = 20$.

Supplementary Note: Sequencing depth simulations In order to examine the robustness of *FEAST* to varying levels of sequencing depth, we used multiple source environments from the Earth Microbiome Project while varying their sequencing depth. In each iteration of our simulation we sampled environments (with median Jensen-Shannon divergence of 0.95) and used them to build a synthetic sink, with different mixing proportions and a set sequencing depth ranging from 100 through 10,000. Notably, by choosing a median Jensen-Shannon divergence of 0.95 we wanted to emphasize that even under the scenario in which the sources are non-overlapping and thus trivial to disambiguate, the sequencing depth will have an effect. Additionally, in these simulations, we only varied the sequencing depth of the sources. However, since the sink samples are a linear combination of the sources, these samples are also, indirectly, affected. To simulate an unknown source, only K source environments are designated as known sources.

The simulation procedure was as follows. For each $l = 1 : D_1$ (different sequencing depth values):

1. Draw $K + 1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.
2. Draw noisy realization of $S_1, \ldots, S_{K+1}$ from the Multinomial distribution (denoted $\tilde{S_k}$).
3. For each $i = 1 : D_2$ (different mixing proportions) :
   (a) Generate random mixing $m \sim Pareto(\alpha > 0)$, where $\sum m = 1$.
   (b) Set the sink sample abundances to $\sum_{k=1}^{K+1} m_k S_k$ per taxa.
   (c) Estimate the known source proportions in the sink using $\tilde{S_1}, \ldots, \tilde{S_K}$.
   (d) Estimate the unknown source proportions in the sink.

4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions per source and average across sources.

5. Calculate the average Jensen-Shannon divergence of m (based on the pairwise Jensen-Shannon divergence).

In the simulations presented we used $D_1 = 19$, $D_2 = 30$, $K = 20$.

Supplementary Note: Unknown source simulations In order to evaluate *FEAST*'s ability to estimate the contribution of the unknown source, we used real source environments from Lax et al. (2014) [1] where disambiguation of sources is challenging, and created synthetic sink communities. Given that any source not sampled should, theoretically, be accounted for in the unknown source, realistic values of the unknown source can therefore span the range of percentages occupied by the observed sources. Specifically, there are scenarios in which the known sources comprise the entirety of the sink (unknown source contribution = 0), or on the other hand, scenarios in which the known sources did not contribute any taxa to the sink (unknown source contribution = 1). Therefore, the unknown source contribution values in our simulation ranges from 0 to 1. As a measure of accuracy, we used the squared Pearson correlation between the estimated mixing proportions and the true mixing proportions for each individual source across repeated simulation runs for the same scenario as the measure of accuracy.

The simulation procedure was as follows. For each $l = 1 : U_1$ (different unknown source proportions):

1. Set the unknown proportion u to $U_1[l]$.

2. Generate random mixing $m - 1 \sim Pareto(\alpha > 0)$, where $\sum m - 1 = 1 - u$.

3. For each $i = 1 : U_2$ (different Jensen-Shannon divergence $\in (0.5 + \epsilon, 0.5 - \epsilon)$:

   (a) Draw $K + 1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.

(b) Draw noisy realization of $S_1, \ldots, S_{K+1}$ from the Multinomial distribution (denoted $\tilde{S}_k$).

(c) Set the sink sample abundances to $\sum_{k=1}^{K} m_k S_k + S_k$ per taxa Draw $K+1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.

(d) Estimate the unknown source proportions in the sink.

4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions of the unknown source.

In the simulations presented we used $U_1 \in (0, 1)$, $T_2 = 30$, $K = 4$, $\epsilon = 0.2$

Supplementary Note: The effect of noisy samples among sources on prediction accuracy We used $K+1$ distinct source environments randomly sampled from the Earth Microbiome Project (i.e., soil, fresh water, feces etc.), where each source was represented by 10 different samples (e.g., $soil_1$, $soil_2$, etc). We then amalgamated these 10 samples (per source environment) and used them to build a synthetic sink, with different mixing proportions. In each iteration of our simulation we sampled $k \in 1, \cdots, 10$ samples from each source environment in order to estimate the corresponding mixing proportions of the amalgamated sources. To simulate an unknown source, we use only K source environments as our known sources. Indeed, we observed that as we increase the number of samples per source, *FEAST*s prediction accuracy improves, however this effect is moderate (squared Pearson correlation ranges from $0.9 - 0.99$, Jensen-Shannon divergence values range from $0.87 - 0.92$).

The simulation procedure was as follows. Draw 11 sources $S_1, \ldots, S_K$, from the Earth Microbiome Project. From each source S_i draw 10 different samples.

1. Draw K sources $S_1, \ldots, S_{K+1}$, from the Earth Microbiome Project. From each source S_i draw 10 different samples.
2. Amalgamate the 10 samples per source environment and create new sources $\tilde{S}_1, \ldots, \tilde{S}_{K+1}$
3. Generate random mixing $m \sim Pareto(\alpha > 0)$, where $\sum m = 1$.

4. Set the sink sample abundances to $\sum_{k=1}^{K+1} m_k \tilde{S}_k$ per taxa.

For each $L = 1 : 10$ (different number of samples representing the sources):

1. Draw L samples from each source $S_{L1}, \ldots, S_{L(K+1)}$,
2. Draw noisy realization of $S_{L1}, \ldots, S_{L(K+1)}$ from the Multinomial distribution (denoted $\tilde{S}_{Lk}$).
3. For each $i = 1 : T_2$ (different mixing proportions) :
   (a) Estimate the known source proportions in the sink using $\tilde{S}_{L1}, \ldots, \tilde{S}_{LK}$.
   (b) Estimate the unknown source proportions in the sink.
4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions per source and average across sources.
5. Calculate the average Jensen-Shannon divergence of m (based on the pairwise Jensen-Shannon divergence).

In the simulations presented we used $K = 10, T_2 = 30$;

Supplementary Note: Using all maternal and early infant samples as potential sources

In this analysis we used the infants at their last time point as sink samples i.e., infant $i \in \{1, \ldots, 98\}$ at 12 months of age. First, we considered all maternal and early infant samples (from all the infants in the study) as potential sources. We used *FEAST* to rank the contribution of each source as compared to all other sources and found that the median contribution of the corresponding maternal sample across all sinks is 1. We performed a permutation test in which the ranks are randomly assigned for each sink, and the p-value is calculated as the number of permutations in which the median of the maternal contributions rank is smaller than the original median. We used 100,000 iterations and obtained a p-value < 0.0001 (Figure S10 (a)). Notably, the top 5 contributing sources included the corresponding infants family 83% of the time (43% of the cases,

the corresponding family ranked 1st, in 21% it ranked 2nd, 4% 3rd, 10% 4th and in 5% it ranked 5th). Next, We repeated these experiments by considering only the maternal samples as potential sources. In this set of sinks (i.e., infants at 12 months of age), the median maternal contribution was 14, and a similar permutation test as the one described above shows that this finding is statistically significance (p-value = 0.00017, Figure S10 (b)). Notably, the gut microbiome of healthy individuals is relatively similar. We therefore removed the samples with low Jensen Shannon divergence value to reduce noise in our estimations. To do this, for each sink_j, we calculated the Jensen Shannon divergence values (1) between mother_j and all other mothers (2) infant-at-birth_j and all other infants at birth (3) $\text{infant-at-4-months}_j$ and all other infants at 4 months, and calculated the median Jensen Shannon divergence for each of these source environments. We then removed samples whose Jensen Shannon divergence fell below their respective median.

Supplementary Note: Expectation-Maximization - derivation Here we derive the full EM algorithm for *FEAST* in detail. Recall that the observed data consist of the sink vector $x = (x_1, x_2..., x_N)$, and source vectors $y_i = (y_{i1}, y_{i2}, ..., y_{iN})$. for $1 \leq i \leq K$. The j-th component of each vector denotes the observed abundance of taxa j in the sink and sources respectively. Denote the total number of observations in each source by $C_i := \sum_{j=1}^{N} y_{ij}$ and total number of observations in the sink by $C := \sum_{j=1}^{N} x_j$. For each source, we have a vector $\gamma_i = (\gamma_{i1}, \gamma_{i2}, ..., \gamma_{iN})$ denoting the unobserved relative abundances of each source y_i. Further, there is assumed to be one unknown, unsampled, source–say $K+1$–with relative abundances $\gamma_{K+1} = (\gamma_{(K+1)1}, \gamma_{(K+1)2}, ..., \gamma_{(K+1)N})$.

Based on the source proportions, each source observation is assumed to have been generated by drawing a random sample from the source with replacement. Thus,

$$y_{ij} \sim \text{Multinomial}(C_i, \gamma_i) \tag{1}$$

For the sink we assume the following generative model. We draw C observations. For each observation $c = 1, ..., C$, we pick a source z^c with the probability of choosing source i given by α_i.

The vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{K+1})$ gives the proportion of the sink derived from each source. Once the source is chosen, we pick taxa x^c from source z^c based on the relative abundances γ_{z^c}. Hence

$$z^c \sim \text{Multinomial}(1, \alpha) \tag{2}$$

$$x^c | z^c \sim \text{Multinomial}(1, \gamma_{z^c}) \tag{3}$$

where we denote $z^c = i$ as having drawn sample c from source i, indicating that the multinomial observation $z^c = (0, ..., 1, ..., 0)$ has 1 in its i-th component and 0s elsewhere. If we marginalize out source assignments z^c, we obtain

$$p(x^c = j) = \sum_{i=1}^{K+1} p(x^c = j | z^c = i) p(z^c = i) = \sum_{i=1}^{K+1} \gamma_{ij} \alpha_i.$$

Hence the marginal distribution of x^c is Multinomial$(1, (\beta_1, ..., \beta_N))$, where $\beta_j = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij}$.

We can therefore rewrite the model as:

$$\beta_j = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij} \quad \text{for } j = 1, ..., N \tag{4}$$

$$y_i \sim \text{Multinomial}(C_i, (\gamma_{i1}, \ldots, \gamma_{iN})) \quad \text{for } i = 1, ..., K \tag{5}$$

$$x \sim \text{Multinomial}(C, (\beta_1, \ldots, \beta_N)) \tag{6}$$

The expected complete log likelihood As demonstrated above, the log likelihood is given by

$$\log p(x, y_1, y_2, ..., y_K | \alpha, \gamma) = \sum_{j=1}^{N} x_j \log \left(\sum_{i=1}^{K+1} \alpha_i \gamma_{ij} \right) + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \log(\gamma_{ij}) + \text{const} \tag{7}$$

Using the notation separating each draw from the sink, the complete log likelihood is given by

$$\log p(x^1, ..., x^C, z^1, ..., z^C, y_1, ..., y_K | \alpha, \gamma) = \sum_{c=1}^{C} \sum_{i=1}^{K+1} z_i^c (\log \gamma_{ix^c} + \log \alpha_i) + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \log(\gamma_{ij}) + \text{const} \tag{8}$$

where $x^c = j$ denotes that observation c corresponds to taxa j. Taking expectations and collecting terms, the expected complete log likelihood is given by

$$Q = \sum_{i=1}^{K+1} \sum_{j=1}^{N} x_j p(i|j) \cdot \log(\alpha_i \gamma_{ij}) + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \log(\gamma_{ij}) + \text{const} \tag{9}$$

where

$$p(i|j) = \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{i=1}^{K+1} \alpha_i^{(t)} \gamma_{ij}^{(t)}} \tag{10}$$

The remainder of the derivation follows the main text.

Table S2. An example of *FEAST*'s output, using the infants dataset from Bäckhed et al. 2015 [2], which includes the top 50 pairs of taxa shared between a vaginally-delivered infant at 12 months of age (sample ERR525717, sink) and its corresponding maternal sample (sample ERR525720, source) (an optional setting)

Class	Order	Family	Genus	Species	Sink	Source
Acidimicrobiia	Acidimicrobiales	AKIW874	NA	NA	0.19639	0.06038
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	coleohominis	0.17838	0.07551
Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	equorum	0.11066	0.01407
Gammaproteobacteria	Alteromonadales	211ds20	NA	NA	0.10158	0.01238
Bacilli	Bacillales	Planococcaceae	Lysinibacillus	odysseyi	0.06719	0.10009
Bacilli	Bacillales	Bacillaceae	Bacillus	horneckiae	0.04117	0.10399
Bacilli	Bacillales	Planococcaceae	Planococcus	maitriensis	0.02739	0.0308
Bacilli	Lactobacillales	Enterococcaceae	Melissococcus	plutonius	0.0245	0.11209
Actinobacteria	Actinomycetales	Actinosynnemataceae	Actinokineospora	diospyrosa	0.02243	0.00341
Actinobacteria	Actinomycetales	NA	NA	NA	0.01857	0.00605
Bacilli	Bacillales	Sporolactobacillaceae	Bacillus	racemilacticus	0.01553	0.02153
Actinobacteria	Actinomycetales	Actinomycetaceae	NA	NA	0.01425	0.01886
Acidimicrobiia	Acidimicrobiales	koll13	NA	NA	0.01308	0.00929
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	mucosae	0.01271	0.04955
Bacilli	Bacillales	Thermoactinomycetaceae	Mechercharimyces	mesophilus	0.01264	0.00328
Bacilli	Bacillales	Listeriaceae	Brochothrix	NA	0.01232	0.00968
Gammaproteobacteria	Oceanospirillales	Oleiphilaceae	NA	NA	0.01165	6.00E-05

Bacilli	Bacillales	[Exiguobacteraceae]	Exiguobacterium	NA	0.01011	0.01914
Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium	variabile	0.00981	0.00441
Actinobacteria	Actinomycetales	Actinosynnemataceae	NA	NA	0.00662	0.00128
Bacilli	Bacillales	Staphylococcaceae	Jeotgalicoccus	NA	0.00641	0.02238
Solibacteres	Solibacterales	Solibacteraceae	CandidatusSolibacter	NA	0.00609	0.00239
Actinobacteria	Actinomycetales	Frankiaceae	NA	NA	0.00558	0.00175
Actinobacteria	Actinomycetales	Dermabacteraceae	Dermabacter	NA	0.00521	0.00388
Actinobacteria	Actinomycetales	Brevibacteriaceae	Brevibacterium	aureum	0.00331	0.00185
Gammaproteobacteria	Alteromonadales	Alteromonadaceae	ND137	NA	0.00255	0.00028
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	pontis	0.0025	0.00418
Acidimicrobiia	Acidimicrobiales	Microthrixaceae	NA	NA	0.00209	0.00043
Clostridia	Clostridiales	Peptococcaceae	Desulfosporosinus	NA	0.00202	0.01059
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	delbrueckii	0.00198	0.01541
Actinobacteria	Actinomycetales	Brevibacteriaceae	Brevibacterium	casei	0.00182	0.00058
Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	Nitrincola	NA	0.00172	4.00E-05
Actinobacteria	Actinomycetales	Actinospicaceae	NA	NA	0.00142	0.00083
Bacilli	Bacillales	Bacillaceae	Geobacillus	NA	0.0014	0.00512
Bacilli	Lactobacillales	Leuconostocaceae	Weissella	NA	0.00136	0.0013
Clostridia	Clostridiales	Clostridiaceae	Caminicella	NA	0.00136	0.00034
Acidimicrobiia	Acidimicrobiales	Acidimicrobiaceae	NA	NA	0.00135	0.00017
Bacilli	Bacillales	Planococcaceae	Planomicrobium	NA	0.00133	0.00341
Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Thioalkalivibrio	NA	0.00133	0.00015
Bacilli	Bacillales	Bacillaceae	Bacillus	marisflavi	0.00129	0.01611
Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	benthica	0.00127	9.00E-05
Clostridia	Clostridiales	Ruminococcaceae	Oscillospira	guilliermondii	0.0012	0.00124
Actinobacteria	Actinomycetales	Actinomycetaceae	Mobiluncus	NA	0.00106	6.00E-05
Clostridia	Clostridiales	Clostridiaceae	Caloranaerobacter	NA	0.00101	4.00E-05
Bacilli	Bacillales	Bacillaceae	Bacillus	badius	0.00099	0.05367
Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Marinimicrobium	NA	0.00096	0.00015
Actinobacteria	Actinomycetales	Dietziaceae	Dietzia	NA	9.00E-04	0.00047
Bacilli	Lactobacillales	Aerococcaceae	Lacticigenium	naphtae	0.00089	0.00543
Bacilli	Bacillales	[Exiguobacteraceae]	NA	NA	0.00081	0.00661
Bacilli	Bacillales	Planococcaceae	Solibacillus	NA	0.00078	0.00077

References

1. Turnbaugh PJ, Gordon JI. The core gut microbiome, energy balance and obesity. J Physiol. 2009;587: 4153–4158.

2. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006;444: 1027–1031.

3. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A. 2005;102: 11070–11075.

4. Koren O, Spor A, Felin J, Fåk F, Stombaugh J, Tremaroli V, et al. Human oral, gut, and plaque microbiota in patients with atherosclerosis. Proc Natl Acad Sci U S A. 2011;108 Suppl 1: 4592–4598.

5. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148: 1258–1270.

6. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500: 541–546.

7. Jeffery IB, Quigley EMM, Öhman L, Simrén M, O'Toole PW. The microbiota link to irritable bowel syndrome: an emerging story. Gut Microbes. 2012;3: 572–576.

8. Marchesi JR, Dutilh BE, Hall N, Peters WHM, Roelofs R, Boleij A, et al. Towards the human colorectal cancer microbiome. PLoS One. 2011;6: e20447.

9. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490: 55–60.

10. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biol. 2011;12: R50.

11. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014;15: R89.

12. Gerber GK. The dynamic microbiome. FEBS Letters. 2014. pp. 4131–4139. doi:10.1016/j.febslet.2014.02.037

13. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473: 174–180.

14. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking "Enterotypes." Cell Host Microbe. 2014;16: 433–437.

15. Yassour M, Vatanen T, Siljander H, Hämäläinen A-M, Härkönen T, Ryhänen SJ, et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med. 2016;8: 343ra81.

16. Backhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. 2015;17: 690–703.

17. Thompson JN. The Geographic Mosaic of Coevolution. University of Chicago Press; 2005.

18. Shenhav L, Thompson M, Joseph TA, Briscoe L, Furman O, Bogumil D, et al. FEAST: fast expectation-maximization for microbial source tracking. Nature Methods. 2019. pp. 627–632. doi:10.1038/s41592-019-0431-x

19. Sánchez J. Shayle R. Searle, George Casella And Charles E. Mcculloch: Variance Components. John Wiley and Sons, New York 1992, XIII, 501 pp., $56.00. Biometrical Journal. 1994. pp. 76–76. doi:10.1002/bimj.4710360109

20. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment

dominates over host genetics in shaping human gut microbiota. Nature. 2018;555: 210–215.

21. Amaratunga D, Cabrera J. Analysis of Data From Viral DNA Microchips. null. 2001;96: 1161–1170.

22. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19: 185–193.

23. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88: 76–82.

24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 1995.

25. Gibbons SM, Kearney SM, Smillie CS, Alm EJ. Two dynamic regimes in the human gut microbiome. PLoS Comput Biol. 2017;13: e1005364.

26. Bucci V, Tzen B, Li N, Simmons M, Tanoue T, Bogart E, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome Biol. 2016;17. doi:10.1186/s13059-016-0980-6

27. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol. 2013;9: e1003388.

28. White JR, Navlakha S, Nagarajan N, Ghodsi M-R, Kingsford C, Pop M. Alignment and clustering of phylogenetic markers--implications for microbial diversity studies. BMC Bioinformatics. 2010;11: 152.

29. Ridenhour BJ, Brooker SL, Williams JE, Van Leuven JT, Miller AW, Dearing MD, et al. Modeling time-series data from microbial communities. ISME J. 2017;11: 2526–2537.

30. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6: 610–618.

31. Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ. 2014;2: e545.

32. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7: 335–336.

33. Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, et al. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. mSystems. 2019;4. doi:10.1128/mSystems.00016-19

34. Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. in 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005. 129–132 (2005).

35. Anandkumar, A., Ge, R. & Janzamin, M. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. arXiv [cs.LG] (2014).

36. Jain, P. & Oh, S. Provable Tensor Factorization with Missing Data. in Advances in Neural Information Processing Systems 27 (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 1431–1439 (Curran Associates, Inc., 2014).

37. Äijö T, Müller CL, Bonneau R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. Bioinformatics. 2018;34: 372–380.

38. Aitchison J, Ho CH. The multivariate Poisson-log normal distribution. Biometrika. 1989. pp. 643–

653. doi:10.1093/biomet/76.4.643

39. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. 2017. doi:10.1128/msystems.00191-16

40. Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang L, Xu ZZ, et al. Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. mSystems. 2018. doi:10.1128/msystems.00021-18

41. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol. 2017;2: 17004.

42. Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Science Translational Medicine. 2016. pp. 343ra82–343ra82. doi:10.1126/scitranslmed.aad7121

43. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems. 2018;3. doi:10.1128/mSystems.00031-18

44. Moon TK. The expectation-maximization algorithm. IEEE Signal Process Mag. 1996;13: 47–60.

45. Silverman JD, Shenhav L, Halperin EA, Mukherjee SA, David LA. Statistical Considerations in the Design and Analysis of Longitudinal Microbiome Studies. bioRxiv. 2018. p. 448332. doi:10.1101/448332

46. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science. 2014;345:

1048–1052.

47. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods. 2011;8: 761–763.

48. Smith A, Sterba-Boatwright B, Mott J. Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. Water Res. 2010;44: 4067–4076.

49. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155: 945–959.

50. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: Analytical and study design considerations. Genetic Epidemiology. 2005. pp. 289–301. doi:10.1002/gepi.20064

51. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009. pp. 1655–1664. doi:10.1101/gr.094052.109

52. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31: 1674–1676.

53. Deloger M, El Karoui M, Petit M-A. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol. 2009;191: 91–99.

54. Leung HCM, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z, et al. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics. 2011;27: 1489–1495.

55. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science. 2009;326: 1694–1697.

56. Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol. 2009;75: 5111–5120.

57. McDonald D, Ackermann G, Khailova L, Baird C, Heyland D, Kozar R, et al. Extreme Dysbiosis of the Microbiome in Critical Illness. mSphere. 2016;1. doi:10.1128/mSphere.00199-16

58. Taur Y, Xavier JB, Lipuma L, Ubeda C, Goldberg J, Gobourne A, et al. Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. Clin Infect Dis. 2012;55: 905–914.

59. Gómez JM, Verdú M, Perfectti F. Ecological interactions are evolutionarily conserved across the entire tree of life. Nature. 2010. pp. 918–921. doi:10.1038/nature09113

60. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42: 565–569.

61. Chaplin AV, Efimov BA, Smeianov VV, Kafarskaia LI, Pikina AP, Shkoporov AN. Intraspecies Genomic Diversity and Long-Term Persistence of Bifidobacterium longum. PLoS One. 2015;10: e0135658.

62. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A. 2010;107: 14691–14696.

63. Azad MB, Konya T, Maughan H, Guttman DS, Field CJ, Chari RS, et al. Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. CMAJ. 2013;185: 385–394.

64. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet.

2012;13: 260–270.

65. Mueller NT, Bakacs E, Combellick J, Grigoryan Z, Dominguez-Bello MG. The infant microbiome development: mom matters. Trends Mol Med. 2015;21: 109–117.

66. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A. 2011;108 Suppl 1: 4578–4585.

67. Antonopoulos DA, Huse SM, Morrison HG, Schmidt TM, Sogin ML, Young VB. Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. Infect Immun. 2009;77: 2367–2375.

68. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A. 2010;107: 6477–6481.

69. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, et al. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A. 2015;112: E2930–8.

70. Gibson TE, Gerber GK. Robust and Scalable Models of Microbiome Dynamics. arXiv [stat.ML]. 2018. Available: http://arxiv.org/abs/1805.04591

71. Shenhav L, Furman O, Briscoe L, Thompson M, Mizrahi I, Halperin E. Modeling the temporal dynamics of the gut microbial community in adults and infants. 2017. doi:10.1101/212993

72. Silverman JD, Durand HK, Bloom RJ, Mukherjee S, David LA. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. Microbiome. 2018;6: 202.

73. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology. 2017. doi:10.3389/fmicb.2017.02224

74. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing microbial composition measurement standards with reference frames. Nat Commun. 2019;10: 2719.

75. Jaccard P. The distribution of the flora in the alpine zone. 1. New Phytol. 1912;11: 37–50.

76. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol Monogr. 1957;27: 325–349.

77. Aitchison J. Principal component analysis of compositional data. Biometrika. 1983;70: 57–65.

78. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71: 8228–8235.

79. McDonald D, Vázquez-Baeza Y, Koslicki D, McClelland J, Reeve N, Xu Z, et al. Striped UniFrac: enabling microbiome analysis at unprecedented scale. Nat Methods. 2018;15: 847–848.

80. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551: 457–463.

81. Keshavan RH, Montanari A, Oh S. Low-rank matrix completion with noisy observations: A quantitative comparison. 2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton). 2009. pp. 1216–1222.