Time-Informed Dimensionality Reduction for Longitudinal Microbiome Studies

Pixu Shi[1#], Cameron Martino[2,3,4#], Rungang Han[5], Stefan Janssen[6], Gregory Buck[7,8], Myrna Serrano[7,8], Kouros Owzar[1], Rob Knight[2,3,9,10], Liat Shenhav[11,12,13&] & Anru R. Zhang[1,14&]

[1]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA
[2]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA
[3]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA
[4]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA, USA
[5]Department of Statistical Science, Duke University, Durham, NC, USA
[6]Algorithmic Bioinformatics, Department of Biology and Chemistry, Justus Liebig University Giessen, Giessen, Germany
[7]Center for Microbiome Engineering and Data Analysis, Virginia Commonwealth University, Richmond, Virginia, USA
[8]Department of Microbiology and Immunology, Virginia Commonwealth University, Richmond, Virginia, USA
[9]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
[10]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
[11]The Center for Studies in Physics and Biology, The Rockefeller University, New York, NY, USA
[12]Institute for systems genetics, New York Grossman School of Medicine, New York University, New York, NY, USA
[13]Department of Computer Science, New York University, New York, NY, USA
[14]Department of Computer Science, Duke University, Durham, NC, USA
[#]These authors contributed equally to the paper
[&]These authors contributed equally to the paper

Corresponding authors: Emails: pixu.shi@duke.edu, robknight@ucsd.edu, liat.shenhav@nyulangone.org, anru.zhang@duke.edu

Abstract

Complex dynamics of microbial communities underlie their essential roles in health and disease, but our understanding of these dynamics remains incomplete**.** To bridge this gap, longitudinal microbiome data are being rapidly generated, yet their power is limited by technical challenges in design and analysis, such as varying temporal sampling, complex correlation structures over feature and time, and high dimensionality. Here, we present TEMPoral TEnsor Decomposition

(TEMPTED), the only time-informed dimensionality reduction method that extracts the underlying microbial dynamics while overcoming the statistical challenges posed by this type of data. TEMPTED facilitates beta-diversity analysis at both sample and subject levels and promotes reproducibility by enabling the transfer of the learned low-dimensional representation from training data to unseen test data. In data-driven simulations, TEMPTED enables host phenotype classification at 90% accuracy, a large margin of improvement from existing methods. In real data, TEMPTED identifies robust patterns in the vaginal microbiome during pregnancy, allowing for the detection of microbial signatures associated with term and preterm births that are reproducible across datasets.

Main Text

Introduction

Given the highly dynamic and complex nature of microbial communities, identifying and predicting their time-dependent patterns is crucial to understanding their structure and function. The collection of longitudinal microbiome samples provides a unique opportunity to capture the dynamics of microbial communities and their associations with host phenotypes. However, the nature of longitudinal microbiome data poses several analytical challenges. First, microbiome data is high-dimensional, making dimensionality reduction key in guiding analysis and interpretation. Second, the pattern of intra-host variation may change over time and vary across hosts, making it challenging to extract robust temporal patterns of microbial features[1]. Third, due to inherent practical limitations of longitudinal studies (e.g., missed patient follow-ups or inconsistent sample collection), multiple hosts often have missing temporal samples which translates into irregular temporal sampling across hosts[2–5].

To model longitudinal microbiome data, many studies first use dimensionality reduction methods, such as principal coordinate analysis (PCoA); however, this method analyzes data at the sample level and does not utilize or account for within-subject correlation and temporal structures. In recent years, unsupervised tensor-methods designated to model longitudinal microbiome data have been developed. CTF[1], microTensor[6], TCAM[7], and EMBED[8] format temporal microbiome data into tabular tensors and apply tensor decomposition to identify low-dimensional structures. However, these tensor-based methods assume all hosts have the same sampling time points with low-level or no missingness, which is often unrealistic in clinical settings. Moreover, CTF and microTensor do not account for the continuity in temporal structure, TCAM does not provide dimension reduction for time or samples, while EMBED aims at characterizing temporal structures without offering dimension reduction for hosts or samples. In addition, a large majority of these methods can not transfer the learned low-dimensional representation from training data to unseen test data. Another relevant class of models is the multivariate functional models that can depict feature trajectories. However, they are unsuitable

for dimensionality reduction or managing unknown structures in hosts[9]. An alternative to analyzing longitudinal microbiome data is by using supervised methods, which are focused on generative models inferring the dynamics of these communities (e.g., generalized Lotka Volterra)[10–13]. Another example is the mixed-effect type models which are widely used to quantify intra-host variation, but typically analyze one microbial feature at a time with limited ability to model temporal patterns[14]. While these methods account for the correlation structure induced by repeated measures as well as for sparsity and compositionality, their output does not directly allow the clustering of phenotypes by microbial community dynamics.

Here, we introduce TEMPoral TEnsor Decomposition (TEMPTED), an unsupervised dimensionality reduction tool for high-dimensional temporal data with flexible temporal sampling. TEMPTED formats longitudinal microbiome data into an order-3 temporal tensor with subject, feature, and continuous time as its three dimensions. The tensor is then decomposed into a summation of low-dimensional components, each consisting of a subject loading vector, a feature loading vector, and a temporal loading function (**Fig. 1**). These loadings provide time-informed dimension reduction and beta-diversity analysis at both the sample and subject levels, and identify corresponding microbial signatures whose temporal trends can aid in discerning host phenotypes. TEMPTED also enables the transfer of the learned low-dimensional representation from training data to unseen test data, thus facilitating research reproducibility. TEMPTED is unique in that it accounts for the temporal order of samples as a continuous variable, without assuming identical temporal sampling across hosts. It is able to retain hosts with missing time points and prevents bias introduced by imputing missing samples or binning time points. It is the only dimensionality reduction method currently available for temporal data that offers this flexibility. These unique properties enable accurate host phenotype classification in data-driven simulations, where otherwise classification was impossible. Further, using TEMPTED we uncover previously undetectable microbial dynamics separating mice with leukemia from healthy ones and pregnancies ending in preterm and term birth.

Results

We evaluated TEMPTED's ability to perform phenotype discrimination using two data-driven simulations. The first simulation is based on the ECAM dataset[2], which sampled the gut microbiome of infants delivered vaginally versus by C-section during the first two years of life (**Fig. S1**). This dataset was chosen due to previously observed differences in longitudinal trajectories between delivery modes[1,2]. The second simulation utilizes the FARMM[15] dataset that comprises daily fecal microbiome samples collected over 15 days from 30 individuals equally divided into three dietary categories—vegan, omnivore, and exclusive enteral nutrition (EEN) without dietary fiber—with all subjects receiving antibiotic treatment during days 6 to 8 (**Fig. S2**). This dataset was chosen due to its use of metagenomics sequencing, and previously observed differences between EEN diet and the other two in the recovery of microbiome after

antibiotic treatment[6,15]. We evaluated TEMPTED and alternative computational methods on their ability to differentiate host phenotypes based on microbial dynamics at the subject level through precision recall (PR) of classification and the sample level through PERMANOVA F-statistic (**Methods**).

These evaluations were performed across random subsets of samples in order to simulate sparse and varying temporal sampling to assess the impact of different sampling densities. First, at the subject level, we evaluated TEMPTED as compared to CTF, microTensor, and TCAM, since methods like PCoA, including Bray-Curtis, Unifrac, and Weighted Unifrac are limited to sample-level analysis. TEMPTED outperforms all methods and reduces the AUC-PR error of host-phenotype classification by more than 50% compared to CTF and microTensor, and this superiority is maintained even when other methods utilize more time points (**Fig. 2a,c**). Among these methods, TEMPTED and TCAM are the only ones capable of transferring the learned low-dimensional representation from training to test data and performing out-of-sample prediction, yet TCAM cannot handle missingness in time points (**Fig. 2a,c; Methods**). Second, at the sample level, TEMPTED outperforms all existing methods in phenotype differentiation across all sampling densities (**Fig. 2b**), while TCAM does not provide sample-level dimension reduction.

To examine the performance of TEMPTED on real data, we applied it to three publicly available datasets. First, we used a mouse study of Acute Lymphoblastic Leukemia (ALL), the most common form of childhood cancer with high genetic predisposition[3]. Fecal microbiome samples were longitudinally sampled from wildtype and predisposed ($Pax5^{+/-}$ genotype) mice that were raised in a specific pathogen-free environment and transferred to a conventional facility at early adulthood to resemble children's transition into kindergarten[3] (**Fig. S3**). While both TEMPTED and existing methods could successfully recover mouse genotype from the microbiome (wild type vs. $Pax5^{+/-}$), existing methods failed to consistently predict the onset of ALL from microbial profiles (**Fig. S6-8**). In contrast, TEMPTED was the only method that clearly distinguished healthy mice from those that developed ALL while associating the difference to specific ASVs (**Fig. 3a-c, S9, Table S1, S2, Supp Table 1**). Additionally, TEMPTED identified ASVs that clearly separate wildtype and $Pax5^{+/-}$ mice (**Fig. 3b-c, S10**). As there is currently no reliable biomarker for ALL onset, this finding might facilitate lead time for therapy. Furthermore, TEMPTED could correlate disease onset to just 12 out of 1065 ASVs. Several of the ALL onset associated ASVs have been associated previously with leukemia and other cancers in mice, in particular those classified in the *Acetatifactor* genus[16–18].

Next, we applied TEMPTED to two independent studies that longitudinally sample the vaginal microbiome during pregnancy: one study sequenced shotgun metagenomics and used metagenome-assembled genomes (MAGs)[4,19] and while the second sequenced 16S rRNA genes and used ASVs[4,5] (**Fig. S4, S5**). Remarkably, TEMPTED consistently captured differences in the

vaginal microbiome trajectories of pregnancies that resulted in term vs. preterm birth in both datasets. Notably, this differentiation between term and preterm births is based on the dynamics of 8 leading microbial features in each study with a significant overlap (**Fig. 3e, 3h**), which can potentially serve as a biomarker to detect preterm birth early in pregnancy (**Fig. 3f, 3i**). Of note, the dynamics of these features are required in order to differentiate between term and preterm birth as this separation was not found when performing differential abundance analysis using ALDEx2[20] at any single time point (**Supp Table 2-3; Methods**). The identified leading ASVs and MAGs both associate *Lactobacillus spp.* and specifically *Lactobacillus crispatus* with term birth, and associate *Gardnerella, Megasphaera, and Atopobium vaginae* with preterm birth (**Fig. 3e, 3h**), consistent with existing literature[4,19,21]. Overall, our results highlight the robustness of TEMPTED in multiple studies and sequencing technologies in uncovering microbial dynamics underlying host phenotypes.

Discussion

TEMPTED was designed to address an important unmet need in the rapidly evolving field of microbiome research – namely unsupervised dimensionality reduction for longitudinal microbiome data that account for the temporal order of samples and accommodates varying temporal sampling schemes and missing values. By leveraging temporal structures shared across hosts, TEMPTED is a computationally efficient tool that facilitates beta-diversity analysis at both sample and subject levels and allows the transfer of the learned low-dimensional representation from training data to unseen datasets.

We demonstrated the utility of TEMPTED using data-driven simulations and real data applications. First, using data-driven simulations based on data sequenced by two technologies (ECAM by 16S and FARMM by shotgun metagenomics) we showcase that TEMPTED outperforms existing methods by a large margin in the ability to discern phenotype differences. Second, using three publicly available datasets on different ecosystems (vaginal microbiome and gut microbiome), we showcase that TEMPTED uncovers biological signals that go undetected by existing methods and produce reproducible results across studies and data types (e.g. 16S and metagenomic sequencing). The microbial features identified by TEMPTED are strong indicators of phenotype differentiation and thus can be leveraged in the design of microbiome-based biomarkers. Finally, its flexibility in accommodating varying temporal sampling, and ability in transferring the learned low-dimensional representation from training to testing data, make it a highly practical tool for microbiome research, promoting research reproducibility.

Despite its many strengths, users of TEMPTED should also be aware of several limitations. Like other unsupervised dimensionality reduction methods such as PCA, PCoA, CTF, microTensor, and TCAM, TEMPTED does not guarantee the capture of the phenotype of interest in its leading components or a single component. Nevertheless, TEMPTED has a flexible model setting that

accommodates a wide range of high-dimensional temporal data, making it a valuable and powerful tool for research beyond longitudinal microbiome studies.

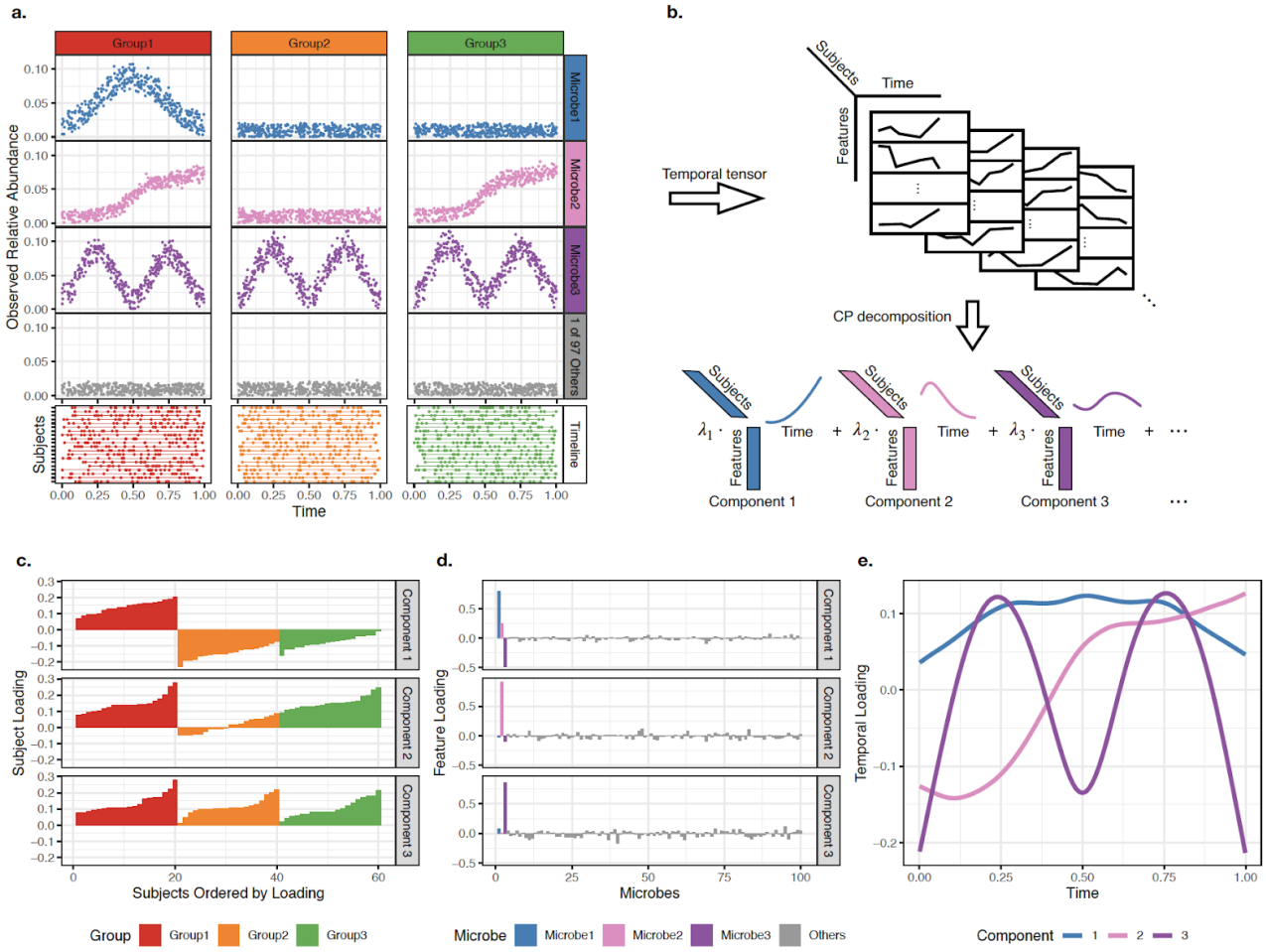

Figure 1. Overview of the TEMPTED algorithm for analyzing multi-subject multi-feature temporal microbiome data. (a) Simulated microbiome count data is transformed into relative abundance and plotted for four representative microbes and three groups of hosts. Microbe 1 has a unique temporal pattern for Group 1, Microbe 2 has a temporal pattern shared by Groups 2 and 3, Microbe 3's temporal pattern is shared by all groups, and Microbes 4-100 have no temporal patterns. Sampling time points are uniformly distributed and used as is without binning. **(b)** The observed multi-subject multi-feature temporal data is formatted into a temporal tensor with three modes representing subject, feature (microbe), and time, respectively. TEMPTED reduces the dimension of the temporal tensor by decomposing it into a small number of components, each of which contains a subject loading vector, a feature loading vector, and a temporal loading function. **(c-e)** TEMPTED loadings of the first three components from the simulated data. The first component captures the bell-shaped trend of Microbe 1 and separates Group 1 from Groups 2 and 3. The second component captures the increasing trend of Microbe 2 and separates Group 2 from Groups 1 and 3. The third component captures the m-shaped trend of Microbe 3 and does not separate any of the groups. Microbes 4-100 have low feature loadings in all components.

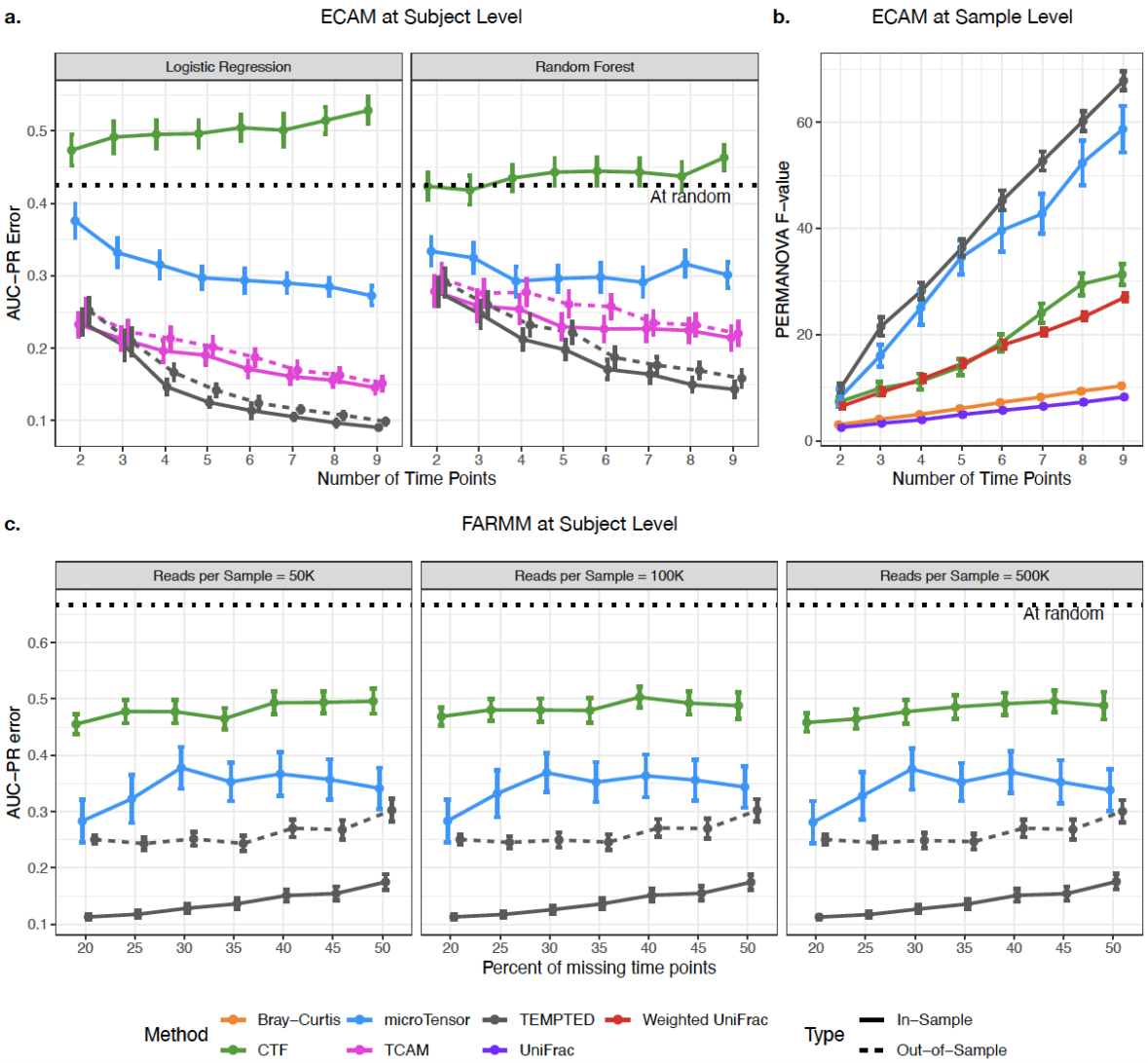

Figure 2. TEMPTED outperforms CTF, microTensor, TCAM and PCoA in identifying group structures. TEMPTED demonstrates superior performance in reducing host-level phenotype discriminatory error by more than 50% compared to methods capable of handling missing time points **(a, c)** and improves the sample-level group discriminatory power (**b**). Sample-level discriminatory power is quantified by PERMANOVA pseudo F-statistic (**b**). Host-level group classification error is quantified by AUC-PR (1 - area under the precision-recall curve). Both logistic regression and random forest classifiers were employed and shown in **(a)**, and the results from the better of the two classifiers were shown in **(c)**. Dimension reduction is performed in-sample and out-of-sample respectively, and group labels are predicted using leave-one-out for logistic regression and out-of-bag for random forest. The methods were applied to two datasets: the ECAM infant fecal microbiome data **(a, b)**, which distinguishes between infants delivered vaginally (N-subject=23) and by cesarean section (N-subject=17); and the FARMM dataset **(c)**, which distinguishes between EEN diet (N-subjects=10) and vegetarian or omnivore diet (N-subject=20). Error bars represent 1.96 standard errors. The first two components of each method were used. For ECAM-based simulation, we randomly choose a given number of samples from each subject such that CTF, microTensor and TCAM can use the order of the infant age as time variable to form a

tensor with no missing values, while TEMPTED uses the infant age as is. For FARMM-based simulation, we randomly drop samples from 15 time points to achieve different percent of missingness, which CTF and microTensor can manage but TCAM cannot. Different reads per sample is obtained by resampling reads in each sample.

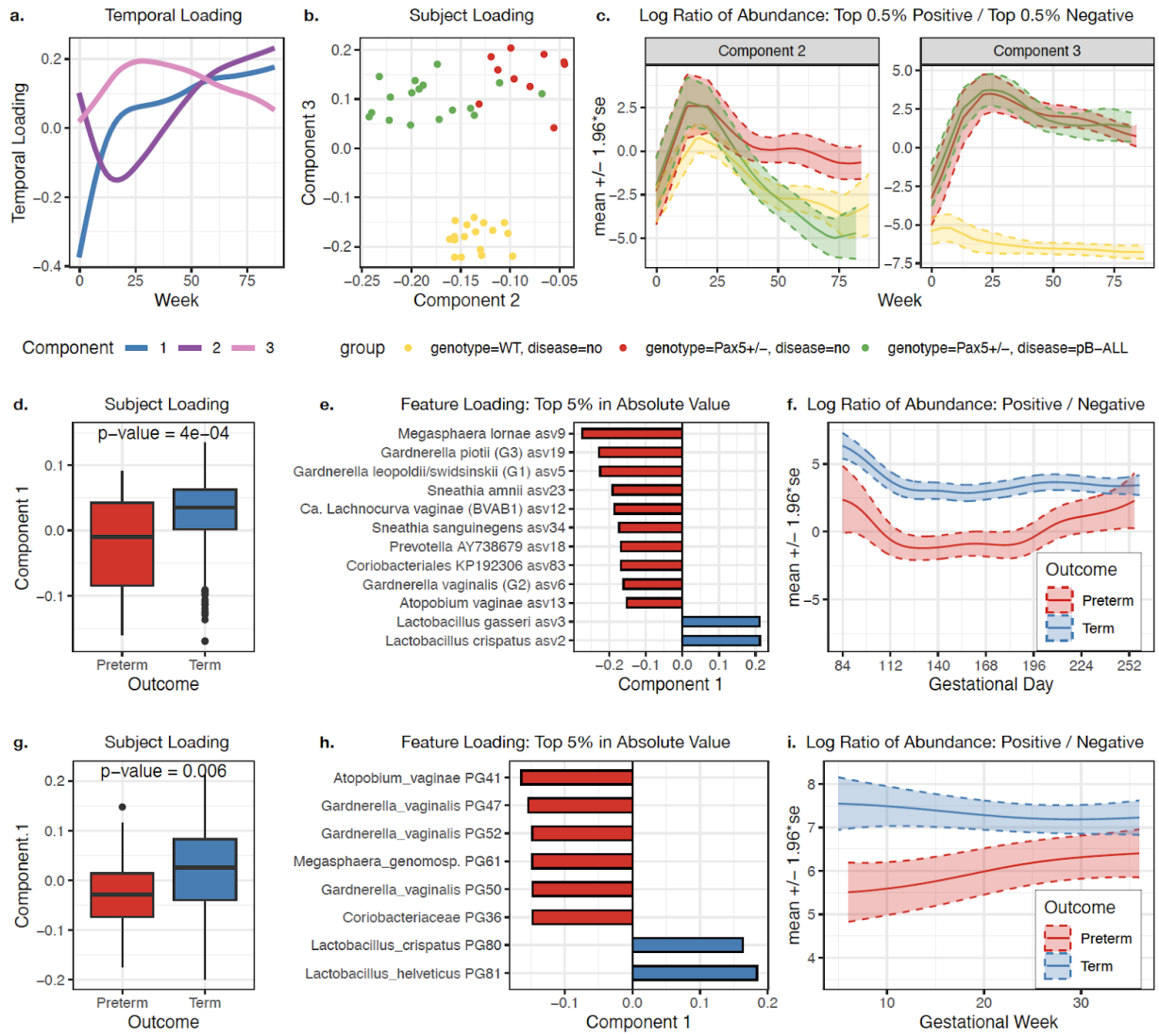

Figure 3. TEMPTED reveals latent trajectories distinguishing health phenotypes. (a)-(c) show the results of applying TEMPTED to a mouse study on leukemia (17 wild-type mice, 27 Pax5$^{+/-}$ mice of which 17 developed leukemia)[3]. **(a)** Temporal loadings capture major temporal patterns in the central-log-ratio-transformed microbial abundance. **(b)** Subject loadings separate genotypes in Component 3 and disease statuses in Component 2. **(c)** Log ratio of the total abundance of the top 0.5% ASVs over the bottom 0.5% ASVs, where 1065 ASVs are ranked by feature loadings of Components 2 (left) and Component 3 (right), respectively. The log ratio from Component 2 diverges between disease statuses for Pax5$^{+/-}$ mice at around 40 weeks after antibiotic treatment. The log ratio from Component 3 shows distinct temporal patterns between genotypes. **(d-f)** Results of applying TEMPTED to ASVs in the vaginal microbiome during pregnancy (141 term and 49 preterm)[4,5]. **(g-i)** Results of applying TEMPTED

to MAGs in the vaginal microbiome during pregnancy (111 term and 35 preterm)[4,19]. **(d,g)** Subject loadings separating pregnancies resulting in term and preterm deliveries in Component 1. The P-value is from the Wilcoxon rank sum test. **(e,h)** Top ASVs and MAGs ranked by the absolute value of feature loadings of Component 1. Features with negative and positive loadings are associated with preterm birth and term birth, respectively. **(f,i)** Log ratio of the abundance of the ASVs with top positive loadings and top negative loadings in (e,h), separating pregnancies resulting in term and preterm deliveries.

References

1. Martino, C. *et al.* Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* **39**, 165–168 (2021).
2. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).
3. Vicente-Dueñas, C. *et al.* An intact gut microbiome protects genetically predisposed mice against leukemia. *Blood* **136**, 2003–2017 (2020).
4. Fettweis, J. M. *et al.* The vaginal microbiome and preterm birth. *Nat. Med.* **25**, 1012–1021 (2019).
5. Costello, E. K. *et al.* Longitudinal dynamics of the human vaginal ecosystem across the reproductive cycle. *bioRxiv* 2022.11.20.517263 (2022) doi:10.1101/2022.11.20.517263.
6. Ma, S. & Li, H. A tensor decomposition model for longitudinal microbiome studies. *Ann. Appl. Stat.*
7. Mor, U. *et al.* Dimensionality reduction of longitudinal ’omics data using modern tensor factorizations. *PLoS Comput. Biol.* **18**, e1010212 (2022).
8. Shahin, M., Ji, B. & Dixit, P. D. EMBED: Essential MicroBiomE Dynamics, a dimensionality reduction approach for longitudinal microbiome studies. *npj Systems Biology and Applications* **9**, 1–11 (2023).
9. Happ, C. & Greven, S. Multivariate Functional Principal Component Analysis for Data

Observed on Different (Dimensional) Domains. *J. Am. Stat. Assoc.* **113**, 649–659 (2018).

10. Gibson, T. E. & Gerber, G. K. Robust and Scalable Models of Microbiome Dynamics. *arXiv [stat.ML]* (2018).
11. Shenhav, L. *et al.* Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLoS Comput. Biol.* **15**, e1006960 (2019).
12. Äijö, T., Müller, C. L. & Bonneau, R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics* **34**, 372–380 (2018).
13. Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S. & David, L. A. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* vol. 6 Preprint at https://doi.org/10.1186/s40168-018-0584-3 (2018).
14. Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. *Applied Longitudinal Analysis*. (John Wiley & Sons, 2012).
15. Tanes, C. *et al.* Role of dietary fiber in the recovery of the human gut microbiome and its metabolome. *Cell Host Microbe* **29**, 394–407.e5 (2021).
16. Buteyn, N. J. *et al.* Activation of the intracellular pattern recognition receptor NOD2 promotes acute myeloid leukemia (AML) cell apoptosis and provides a survival advantage in an animal model of AML. *The Journal of Immunology* **204**, 1988–1997 (2020).
17. Lee, C. *et al.* CD1d Modulates Colonic Inflammation in NOD2−/− Mice by Altering the Intestinal Microbial Composition Comprising Acetatifactor muris. *J. Crohns. Colitis* **13**, 1081–1091 (2019).
18. Chen, L. *et al.* Altering the Microbiome Inhibits Tumorigenesis in a Mouse Model of Oviductal High-Grade Serous CarcinomaMicrobiome Effects in an In Vivo Ovarian Cancer Model. *Cancer Res.* **81**, 3309–3318 (2021).

19. Liao, J. *et al.* Microdiversity of the Vaginal Microbiome is Associated with Preterm Birth. *bioRxiv* (2023) doi:10.1101/2023.01.13.523991.

20. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One* **8**, e67019 (2013).

21. Sun, S. *et al.* Race, the Vaginal Microbiome, and Spontaneous Preterm Birth. *mSystems* **7**, e0001722 (2022).

Methods

Data Preprocessing

Let $i = 1, \ldots, n$ denote subjects, $j = 1, \ldots, p$ denote features, and $\boldsymbol{\mathcal{Y}}_{ijt}$ denote the observed read count of feature j from subject i at time points $t \in T_i = \{t_{i1}, \ldots, t_{im_i}\}$, where T_i is a subset of interval T and can be different across subjects. `TEMPTED` allows users to choose their own preferred data normalization and transformation. Here, to account for the variation in sequencing depth across samples and the highly skewed distribution of microbiome sequencing data, we apply the centered-log-ratio (CLR) transformation to read counts added by .5 [1, 2]

$$\tilde{\boldsymbol{\mathcal{Y}}}_{ijt} = \log\left(\frac{\boldsymbol{\mathcal{Y}}_{ijt} + .5}{\left(\prod_{\tilde{j}=1}^{p}(\boldsymbol{\mathcal{Y}}_{i\tilde{j}t} + .5)\right)^{1/p}}\right).$$

Because some bacteria taxa are much more abundant than others across most time points, when such a time-invariant structure is not of interest, we offer an optional step to extract and remove this structure from $\tilde{\boldsymbol{\mathcal{Y}}}$ to improve the efficiency of our algorithm. Specifically, we average $\tilde{\boldsymbol{\mathcal{Y}}}_{ijt}$ across t, to form matrix $\bar{\boldsymbol{\mathcal{Y}}} = (Y_{ij\cdot})$, calculate the first singular component of matrix $\bar{\boldsymbol{\mathcal{Y}}}$, and subtract the resulting rank-1 approximation of $\bar{\boldsymbol{\mathcal{Y}}}$ from $\tilde{\boldsymbol{\mathcal{Y}}}_t$ for each t. This optional step reduces the effect of highly abundant bacteria taxa and other time-invariant factors that may confound downstream analysis.

`TEMPTED` algorithm

We adopt the model setting proposed in [3], which decomposes the temporal tensor formed by $\tilde{\boldsymbol{\mathcal{Y}}}_{ijt}$ using an approximately CP low-rank structure. However, while [3] assumes T_i to be identical across all subjects, TEMPTED's algorithm can accommodate the more common scenario of varying temporal sampling T_i across subjects. Our objective is to minimize the following function:

$$\min_{a_i, b_j, \xi} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{t \in T_i} \left(\tilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \sum_{\ell=1}^{r} \lambda_\ell \hat{a}_i^{(\ell)} \hat{b}_j^{(\ell)} \hat{\xi}^{(\ell)}(t)\right)^2$$

$$\text{subject to } \sum_{i=1}^{n} (a_i^{(\ell)})^2 = \sum_{j=1}^{p} (b_j^{(\ell)})^2 = \int \left(\xi^{(\ell)}(t)\right)^2 \mathrm{d}t = 1, ||\xi^{(\ell)}(t)||_{\mathcal{H}} \leq C_\xi,$$

where r is the number of low-rank components to approximate the data tensor $\tilde{\boldsymbol{\mathcal{Y}}}$, $a^{(\ell)} = (a_1^{(\ell)}, \ldots, a_n^{(\ell)})$ are subject loadings for subject-level beta analysis, $b^{(l)} = (b_1^{(l)}, \ldots, b_p^{(l)})$ are feature loadings that can be used to construct sample-level beta analysis and to identify features contributing to the beta analysis, $\xi^{(l)}(t)$ is the temporal loading that captures the shared temporal patterns among subjects and features, and λ quantifies the contribution of each component. Here, $\|\xi(t)\|_{\mathcal{H}^2}$ is the *reproducing kernel Hilbert space (RKHS) norm* of function $\xi(t)$ with the rescaled Bernoulli polynomial as the reproducing kernel,

$$\mathbb{K}(s,t) = 1 + k_1(s)k_2(t) + k_2(s)k_2(t) - k_4(|s-t|),$$

where $k_1(s) = s - .5$, $k_2(s) = (k_1^2(s) - 1/12)/2$, and $k_4(s) = (k_1^4(s) - k_1^2(s)/2 + 7/240)/24$. This kernel guarantees $\xi(t)$s to be absolutely continuous and squared-integrable in its second-order derivative (see [3] for details). C_ξ is a tuning parameter to control the smoothness of $\xi(t)$s.

For each component $\ell = 1, \ldots, r$ sequentially, we perform the following Steps 1 to 3 to estimate a, b, and $\xi(t)$.

Step 1: (Initialization) Initialize $\hat{a} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$. Set $\hat{b}$ as the first left singular vector of mode-2 matricization of $\tilde{\boldsymbol{\mathcal{Y}}}$: i.e., the p-by-$(\sum_{i=1}^n |T_i|)$ matrix with $\{\tilde{\boldsymbol{\mathcal{Y}}}_{i\cdot}(t) \in \mathbb{R}^p\}_{1=i\ldots,n,t\in T_i}$ as its columns.

Step 2: (Estimation of loadings) To estimate the loadings, we minimize the following function by iteratively updating $\hat{\xi}$, $\hat{a}$, and $\hat{b}$ respectively until convergence:

$$\sum_{i=1}^n \sum_{j=1}^p \sum_{t\in T_i} \left(\tilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \eta a_i b_j \xi(t)\right)^2 + C_\lambda \|\xi\|_{\mathcal{H}}^2. \tag{1}$$

(a) Update $\hat{\xi}$ by applying kernel ridge regression [3] to solve

$$\hat{\xi} = \arg\min_{\xi\in\mathcal{H}} \sum_{i=1}^n \sum_{j=1}^p \sum_{t\in T_i} \left(\tilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \hat{a}_i \hat{b}_j \xi(t)\right)^2 + C_\lambda \|\xi\|_{\mathcal{H}}^2, \tag{2}$$

and normalize $\hat{\xi}$ to $\hat{\xi} = \hat{\xi} / \left(\int \hat{\xi}(t)^2 \mathrm{d}t\right)^{1/2}$.

(b) Update $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_n)$ by

$$\hat{a}_i = \frac{\sum_{j=1}^p \sum_{t\in T_i} \tilde{\boldsymbol{\mathcal{Y}}}_{ijt} \hat{b}_j \hat{\xi}(t)}{\sum_{t\in T_i} \hat{\xi}(t)^2}, \quad \text{and } \hat{a} = \hat{a}/\|\hat{a}\|_2. \tag{3}$$

(c) Update $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_p)$ by

$$\hat{b}_j = \frac{\sum_{i=1}^n \sum_{t\in T_i} \tilde{\boldsymbol{\mathcal{Y}}}_{ijt} \hat{a}_i \hat{\xi}(t)}{\sum_{i=1}^n \sum_{t\in T_i} (\hat{a}_i \hat{\xi}(t))^2}, \quad \text{and } \hat{b} = \hat{b}/\|\hat{b}\|_2. \tag{4}$$

Step 3: (Subtracting previous components) Update $\tilde{\boldsymbol{\mathcal{Y}}}$ by

$$\tilde{\boldsymbol{\mathcal{Y}}}_{ijt} = \tilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \hat{\eta} \hat{a}_i \hat{b}_j \hat{\xi}(t),$$

where $\hat{\eta} \in \mathbb{R}$ is obtained by solving the following least squares problem:

$$\arg\min_{\eta} \sum_{i=1}^n \sum_{j=1}^p \sum_{t\in T_i} \left(\tilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \eta \hat{a}_i \hat{b}_j \hat{\xi}(t)\right).$$

Step 4: $\lambda = (\lambda_1, \ldots, \lambda_r)$ is estimated via the following least squares problem:

$$\widehat{\lambda} = \operatorname{argmin}_{\lambda} \sum_{i=1}^n \sum_{j=1}^p \sum_{t\in T_i} \left(\boldsymbol{\mathcal{Y}}_{ijt} - \sum_{\ell=1}^r \lambda_\ell \hat{a}_i^{(\ell)} \hat{b}_j^{(\ell)} \hat{\xi}^{(\ell)}(t)\right)^2.$$

Sample-Level Beta Diversity and Subject-Specific Trajectories

The feature loading $b_j^{(\ell)}$ quantifies the contribution of Feature j to Component ℓ. They can be used as the weights vector to aggregate the trajectory of all p features into the following subject-specific trajectory corresponding to Component ℓ:

$$\boldsymbol{\mathcal{B}}_{it}^{(\ell)} = \sum_{j=1}^{p} b_j^{(\ell)} \tilde{\boldsymbol{\mathcal{Y}}}_{ijt}. \tag{5}$$

Here, the vector $\left(\boldsymbol{\mathcal{B}}_{it}^{(1)}, \cdots, \boldsymbol{\mathcal{B}}_{it}^{(r)}\right)$ also serves as an r-dimensional representation of Sample t from Subject i, which can be used for sample-level beta analysis.

For microbiome studies specifically, subject-specific trajectories can be constructed using the log-ratio abundance of top over bottom ranking features sorted by feature loading:

$$\boldsymbol{\mathcal{B}}_{it}^{(\ell)} = \log\left(\left(0.5 + \sum_{j:b_j^{(\ell)} \text{ in top quantile}} \boldsymbol{\mathcal{C}}_{ijt}\right) / \left(0.5 + \sum_{j:b_j^{(\ell)} \text{ in bottom quantile}} \boldsymbol{\mathcal{C}}_{ijt}\right)\right).$$

where $\boldsymbol{\mathcal{C}}_{ijt}$ is the read count of Feature j in the tth sample of Subject i.

Transfer of Dimension Reduction to New Data

Let $\boldsymbol{\mathcal{Y}}_{\text{train}}$ and $\boldsymbol{\mathcal{Y}}_{\text{test}}$ be two datasets consisting of the same features measured within the same time frame $[0, T]$. Suppose that `TEMPTED` decomposes $\boldsymbol{\mathcal{Y}}_{\text{train}}$ into r components, where the ℓth component has subject loading $a_{\text{train}}^{(\ell)}$, feature loading $b_{\text{train}}^{(\ell)}$, and temporal loading $\xi_{\text{train}}^{(\ell)}$. Assuming that $\boldsymbol{\mathcal{Y}}_{\text{train}}$ and $\boldsymbol{\mathcal{Y}}_{\text{test}}$ share the same feature loading and temporal loading, we can estimate the subject loading of $\boldsymbol{\mathcal{Y}}_{\text{test}}$ (i.e., $a_{\text{test}}^{(\ell)}$) through one iteration of Step 2 of the TEMPTED algorithm, with $b_{\text{train}}^{(\ell)}$ and $\xi_{\text{train}}^{(\ell)}$ plugged in. The phenotype of the testing subjects will be predicted by applying classifiers trained by $a_{\text{train}}^{(\ell)}$ to $a_{\text{test}}^{(\ell)}$. Since the dimension reduction is performed without any knowledge of $\boldsymbol{\mathcal{Y}}_{\text{test}}$, classifiers trained by $a_{\text{train}}^{(\ell)}$ have no information from the testing data. The prediction of phenotype for the testing data by such classifiers is purely out-of-sample, and its accuracy is evaluated through data-driven simulation (Figure 2 dashed lines).

Case Study: Pax5 Mice Leukemia Data

The Pax5 dataset was published by [4] and deposited at `https://qiita.ucsd.edu/study/description/11953`, artifact ID 75878. The dataset we used here consists of fecal samples collected from 17 wild type and 27 Pax5 heterozygous mice that were treated with antibiotics at the beginning of the experiment, and 17 Pax5 mice developed leukemia during the study. Samples were subjected to 16S V4 short-read Illumina sequencing. Raw reads are deposited at ENA with accession PRJEB34720. Please refer to [4] for detailed sequence processing methods. Note that the original publication contains further amplicon sequence data and PacBio long read data not

used here. Samples with < 35000 reads are removed. Mice with < 2 time points after filtering are removed. Time points were recorded as days from the end of antibiotic treatment and were divided by 7 to obtain weeks. `TEMPTED` uses ASVs appearing in $> 5\%$ of all samples and `CTF` uses ASVs appearing in $> 10\%$ of all samples.

Case Study: Shotgun Metagenomic Vaginal Data

The dataset used in this study was collected and published by [5] and deposited under dbGaP (study no. 20280; accession ID phs001523.v1.p1). It comprises a total of 705 vaginal samples obtained from 175 pregnant women visiting maternity clinics in Virginia and Washington. Among them, 40 women had preterm pregnancies, and the remaining 135 women had term pregnancies. To ensure data quality, subjects with less than two time points and MAGs that appeared in less than 5% of all the samples were removed from the analysis. For the time-specific analysis we used ALDEx2 with gestational weeks 5-36. A test was performed on weeks with at least 2 term and 2 preterm subjects.

Case Study: 16S Vaginal Data

The 16S data used in this study were published by [6] and [5], and were deposited under BioProjects PRJNA393472 (subjects enrolled at the University of Alabama, Birmingham) and PRJNA821262 (subjects enrolled at Stanford University). Preprocessed data were obtained from `https://purl.stanford.edu/pz745bc9128`. We focus specifically on the second- and third-trimester pre-delivery vaginal samples from 141 term pregnancies and 49 pre-term pregnancies. Samples with < 40000 reads were removed, and ASVs (amplicon sequence variants) appearing in $\leq 5\%$ of all the samples were removed as well. For time-specific analysis using ALDEx2, to ensure enough sample size at each time point, gestational days were floored to gestational weeks. No test was performed on Week 36 because it only contains three preterm subjects.

Data Driven Simulation

The ECAM dataset was published by [7] (Qiita ID 10249). Preprocessed data were obtained from `https://codeocean.com/capsule/6494482/tree/v1`. Months 15 and 19 were removed due to their large amount of missingness. Operational taxonomic units (OTUs) appearing in $\leq 5\%$ of the remaining samples were removed, as were samples with < 2000 reads. Subjects with fewer than nine time points were also removed, leaving 23 vaginally delivered infants and 17 Cesarean-delivered infants in the analysis. For $m = 2, \ldots, 9$, m samples were randomly chosen from each subject to form the simulated dataset. Time points were recorded in days and used as is for `TEMPTED`. For `CTF`, `microTensor`, and `TCAM`, methods that demand input as a tabular tensor low-level or no missing time points, the order of the m samples were used as the time variable.

The FARMM dataset was published by [8] (BioProject ID PRJNA675301). Preprocessed relative abundance data were obtained from `https://github.com/syma-research/microTensor/tree/main/data/FARMM`. Taxa appearing in fewer than 5 samples were

removed. Samples with fewer than 5 taxa were also removed. Time point zero was removed because no subject in the vegan group has samples at time zero, making the missingness not at random. The remaining 15 time points were used as is for `TEMPTED`. For `CTF`, and `microTensor`, while `TCAM` cannot be run with such missingness. Simulated read counts are generated from multinomial distribution based on the observed relative abundance, which is equivalent to rarefication on the observed read counts.

We used the same central-log-transformed data as input for `TEMPTED` and `TCAM`, while `CTF` and `microTensor` take count data as input. The log-fold-change over baseline transformation used in the `TCAM` paper yields worse results for `TCAM`.

Software Usage

PERMANOVA was implemented using R package `vegan`. Logistic regression and random forest were implemented using R package `stats` and `randomForest`, respectively. AUC-ROC was calculated using R package `PRROC`. ALDEx2 was performed using R package `ALDEx2` (1.32.0). CTF was performed using Python plugin gemelli (0.0.8).

Code and Data Availability

`TEMPTED` is compiled into an R package that can be installed from `https://github.com/pixushi/tempted`. All data analysis and simulation are performed using R and Python scripts through R package `reticulate`. All processed data and codes used for this paper are available on `https://github.com/pixushi/TEMPTED_paper`. The TEMPTED method has been implemented through the R package "tempted" on CRAN (tutorial at https://github.com/pixushi/TEMPTED) and the Python package "gemelli" (tutorial at https://github.com/biocore/gemelli/tree/jointrpca).

References

1. Aitchison, J. & Aitchison, J. *The Statistical Analysis of Compositional Data* ISBN: 9780412280603. `https://books.google.com/books?id=RHKmAAAAIAAJ` (Springer Netherlands, 1986).
2. Shi, P., Zhou, Y. & Zhang, A. R. High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis. *Biometrika* **109,** 405–420 (2022).
3. Han, R., Shi, P. & Zhang, A. R. Guaranteed Functional Tensor Singular Value Decomposition. *Journal of the American Statistical Association* **to appear** (2023).
4. Vicente-Dueñas, C. *et al.* An intact gut microbiome protects genetically predisposed mice against leukemia. *Blood* **136,** 2003–2017 (2020).
5. Fettweis, J. M. *et al.* The vaginal microbiome and preterm birth. *Nature medicine* **25,** 1012–1021 (2019).

6. Costello, E. K. *et al.* Longitudinal dynamics of the human vaginal ecosystem across the reproductive cycle. *bioRxiv,* 2022–11 (2022).

7. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science translational medicine* **8,** 343ra82–343ra82 (2016).

8. Tanes, C. *et al.* Role of dietary fiber in the recovery of the human gut microbiome and its metabolome. *Cell host & microbe* **29,** 394–407 (2021).

Supplementary Information for:

Time-Informed Dimensionality Reduction for Longitudinal Microbiome Studies

Details of kernel ridge regression to update ξ(t)

By the presenter theorem[1], the solution of formula (2) in the Methods section must be a linear combination of $\mathbb{K}(\cdot, s)$ for $s \in \cup_{i=1}^{n} T_i$, and the weight of this linear combination can be solved by kernel ridge regression. Specifically, we introduce the concatenated observation vector and kernel matrix:

$$\tilde{y} = (y_{1,1,T_1}, y_{2,1,T_2}, \ldots, y_{n,1,T_n}, \ldots, y_{1,p,T_1}, y_{2,p,T_2}, \ldots, y_{n,p,T_n})^\top \in \mathbb{R}^{p(|T_1|+\cdots+|T_n|)};$$

$$\tilde{K} = \begin{bmatrix} \mathbb{K}(T_1, T_1) & \cdots & \mathbb{K}(T_1, T_n) \\ \vdots & \ddots & \vdots \\ \mathbb{K}(T_n, T_1) & \cdots & \mathbb{K}(T_n, T_n) \end{bmatrix} \in \mathbb{R}^{(|T_1|+\cdots+|T_n|)\times(|T_1|+\cdots+|T_n|)};$$

$$\widetilde{AB} = \begin{bmatrix} b_1^{(t)} \cdot \mathrm{diag}\left(\overbrace{a_1^{(t)}}^{|T_1|\text{times}}, \ldots, \overbrace{a_n^{(t)}}^{|T_n|\text{times}}\right) \\ b_2^{(t)} \cdot \mathrm{diag}\left(\overbrace{a_1^{(t)}}^{|T_1|\text{times}}, \ldots, \overbrace{a_n^{(t)}}^{|T_n|\text{times}}\right) \\ \vdots \\ b_p^{(t)} \cdot \mathrm{diag}\left(\overbrace{a_1^{(t)}}^{|T_1|\text{times}}, \ldots, \overbrace{a_n^{(t)}}^{|T_n|\text{times}}\right) \end{bmatrix} \in \mathbb{R}^{p(|T_1|+\cdots|T_n|)\times(|T_1|+\cdots+|T_n|)}.$$

The solution to formula (2) in the Methods section is equivalent to

$$\min_{\alpha \in \mathbb{R}^{|T_1|+\cdots+|T_n|}} \left\| \tilde{y} - \widetilde{AB}\tilde{K}\alpha \right\|^2 + \lambda \alpha^\top \tilde{K} \alpha,$$

which can be solved by

$$\hat{\alpha} = (\widetilde{AB}^\top \widetilde{AB}\tilde{K} + \lambda I)^{-1} \widetilde{AB}^\top \tilde{y}, \quad \hat{\xi}(s) = \alpha^\top \left(\mathbb{K}(T_1, s), \ldots, \mathbb{K}(T_n, s)\right)^\top.$$

Here, I is the identity matrix, and $b^{(t)}$ will cancel out in $\widetilde{AB}^\top \widetilde{AB}$ because $||b^{(t)}||_2 = 1$.

Toy Example

We simulated a dataset with 100 features and three groups of subjects, each containing 20 subjects. For each subject, we randomly selected 20 time points from uniform distribution between 0 and 1. We considered three temporal trends: $f_1(t) = \frac{1}{2}\sin(2\pi t - \frac{1}{2}\pi) + \frac{1}{2}$, $f_2(t) = \mathrm{atan}(10t - 5)/\pi + \mathrm{atan}(5)/\pi$, and $f_3(t) = \frac{1}{2}\sin(4\pi t - \frac{1}{2}\pi) + \frac{1}{2}$. For Group 1, we simulated the absolute abundance of the first feature from $f_1(t)$ perturbed by a uniform distribution on (0, 0.5). For Groups 1 and 3, we simulated the absolute abundance of the second feature from $f_2(t)$ perturbed by a uniform distribution on (0, 0.25). For Groups 1, 2, and 3, we simulated the absolute abundance of the third feature from $f_3(t)$ perturbed by a uniform distribution on (0, 0.5). The absolute abundance of the remaining features was simulated from a uniform distribution on (0, 0.25). We then calculated the relative abundance and used them as proportion parameters in a multinomial distribution to generate counts. The sequencing depth was generated from an exponentially transformed normal distribution with a mean of 10 and standard deviation of 0.5, resulting in minimum, median, and maximum depths of 5K, 22K, and 113K, respectively.

Temporal Sampling of Studies

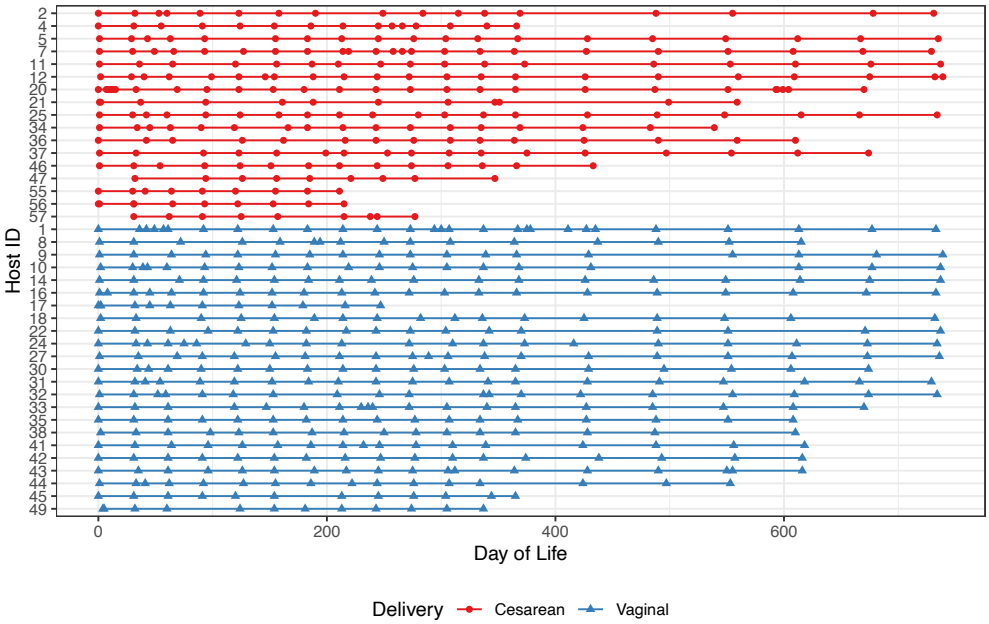

Figure 1: Temporal sampling of ECAM infant gut microbiome data

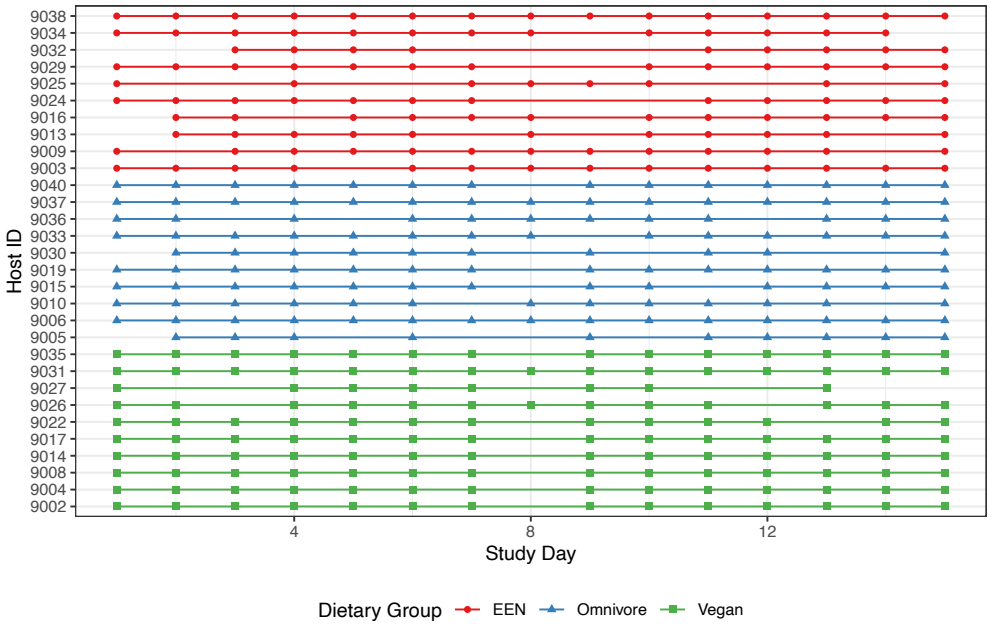

Figure 2: Temporal sampling of FARMM gut microbiome data

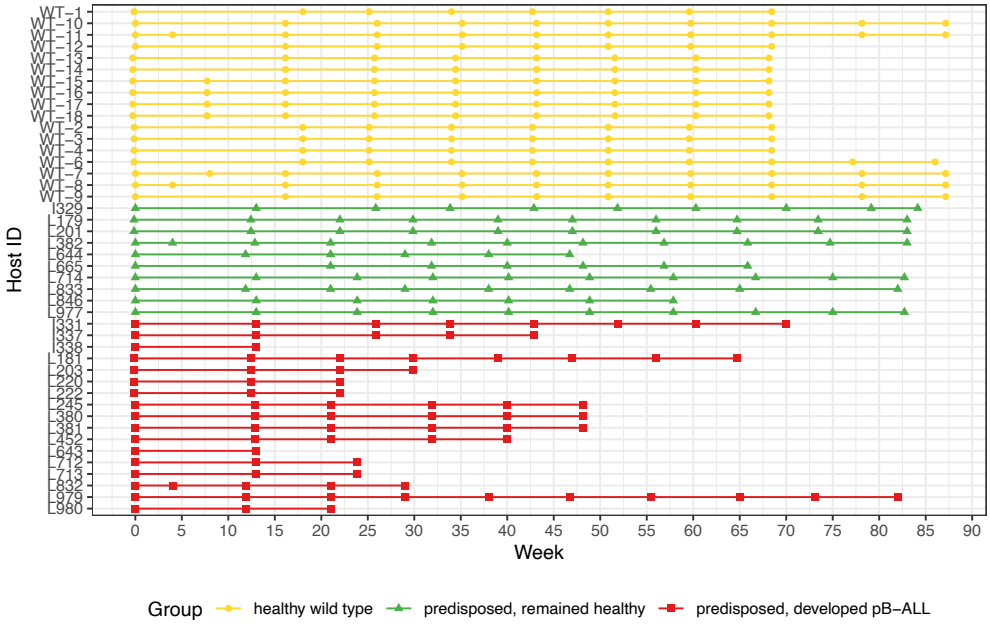

Figure 3: Temporal sampling of Pax5 mice leukemia data

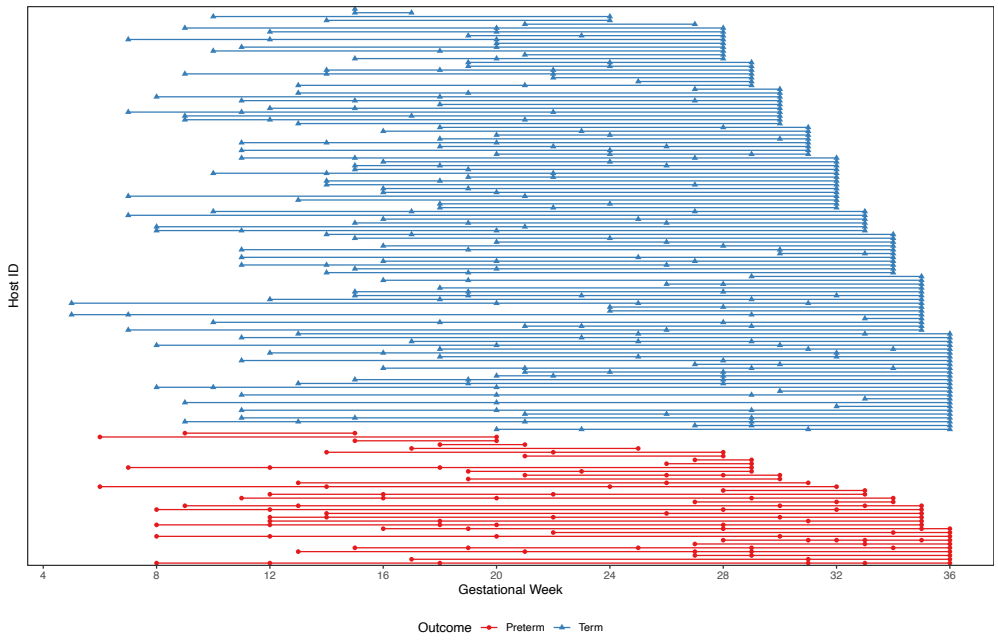

Figure 4: Temporal sampling of metagenomic vaginal microbiome data.

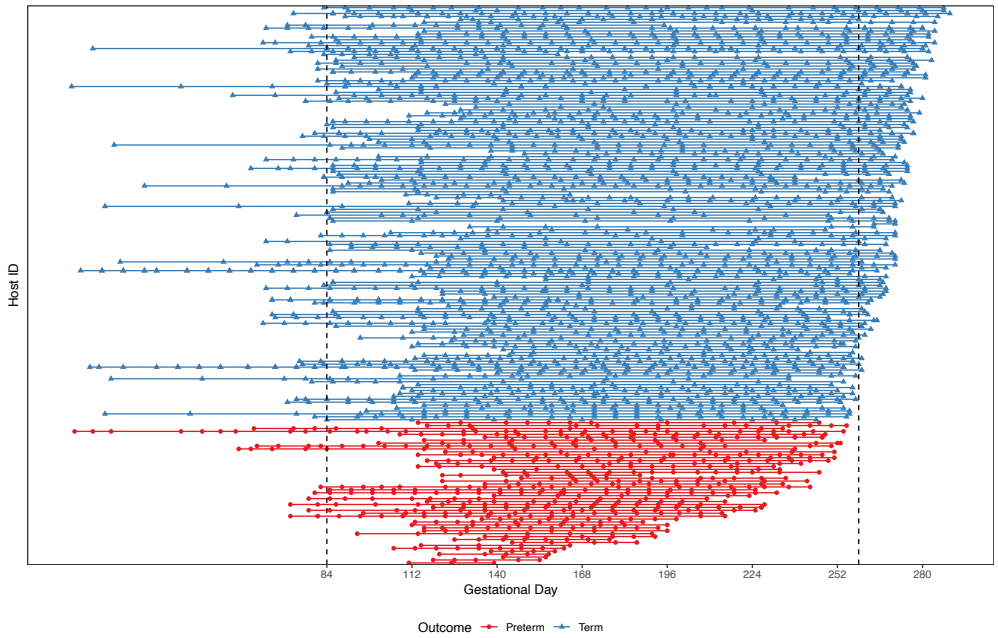

Figure 5: Temporal sampling of 16S vaginal microbiome data. Only second and third trimester samples, i.e. samples between the two dashed vertical lines, were used.

Analysis of Pax5 Mice Leukemia Data

Analysis with Other Methods

Figure 6: Trajectories of first three PCoA components for various distance metrics on the Pax5 mice leukemia data. Mean and error bands representing 1.96 standard errors of the mean are smoothed using kernel regression.

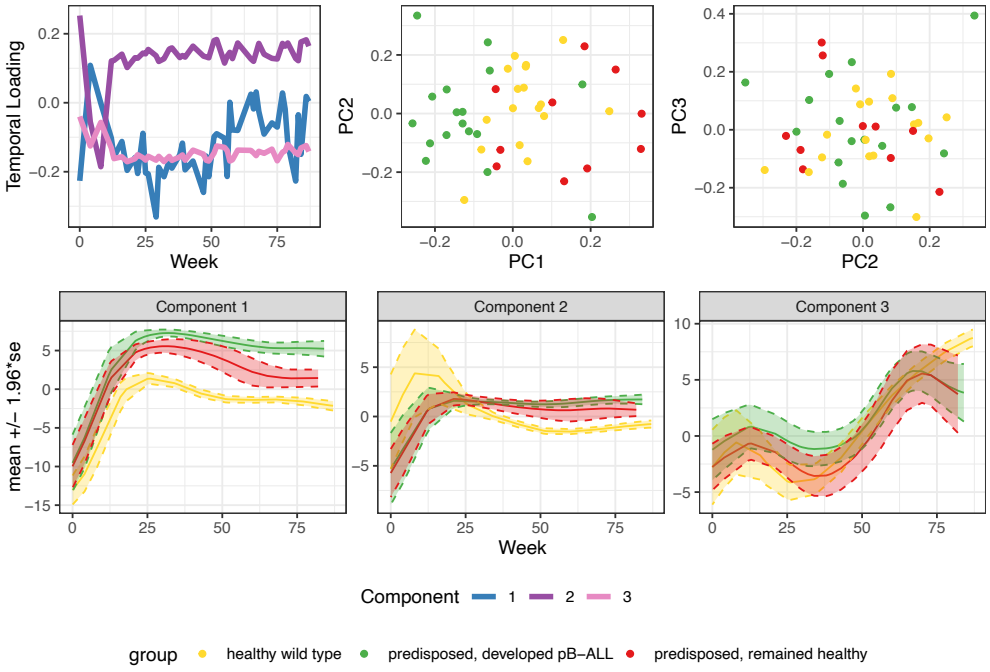

Figure 7: Results from `CTF` on the Pax5 mice leukemia data.

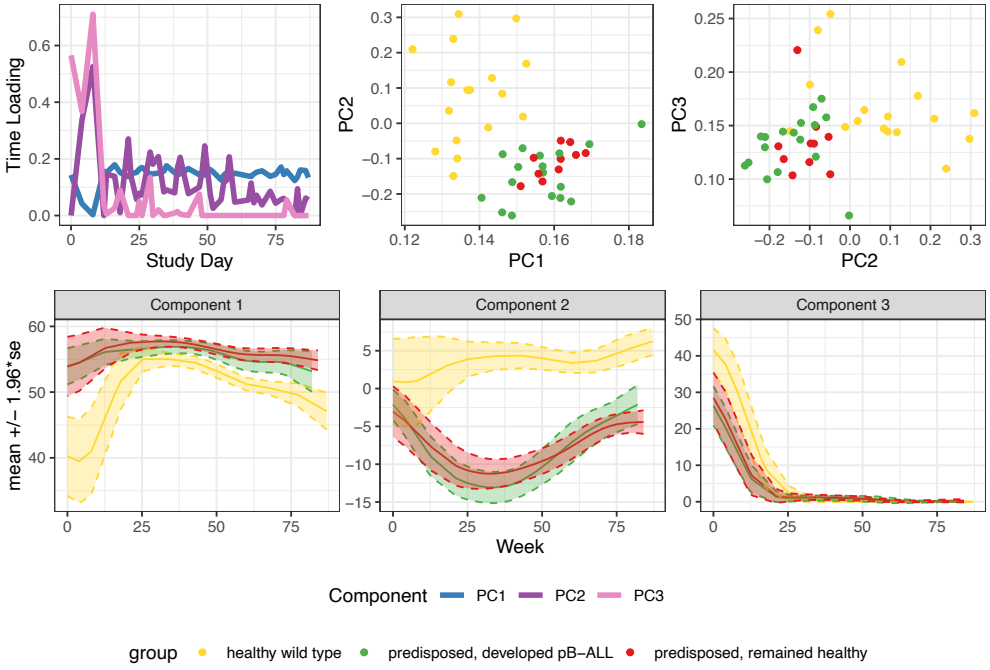

Figure 8: Results from `microTensor` on the Pax5 mice leukemia data.

	bray	jaccard	unifrac	wunifrac	TEMPTED	CTF
p-value	0.143	0.143	0.342	0.798	0.001	0.032
F-value	1.996	1.930	0.956	0.145	6.922	3.707

Table 1: PERMANOVA test between leukemia outcomes for 35-70 week samples among Pax5+/− mice

Component	bray	jaccard	unifrac	wunifrac	TEMPTED	CTF
1	0.007	0.006	0.135	0.672		0.029
2	0.613	0.523	0.131	0.613	1.7e-5	
3	0.491	0.430	0.849	0.759		

Table 2: Wilcoxon test between leukemia outcomes for 35-70 week samples among Pax5+/− mice for each component

Top ASVs identified by TEMPTED

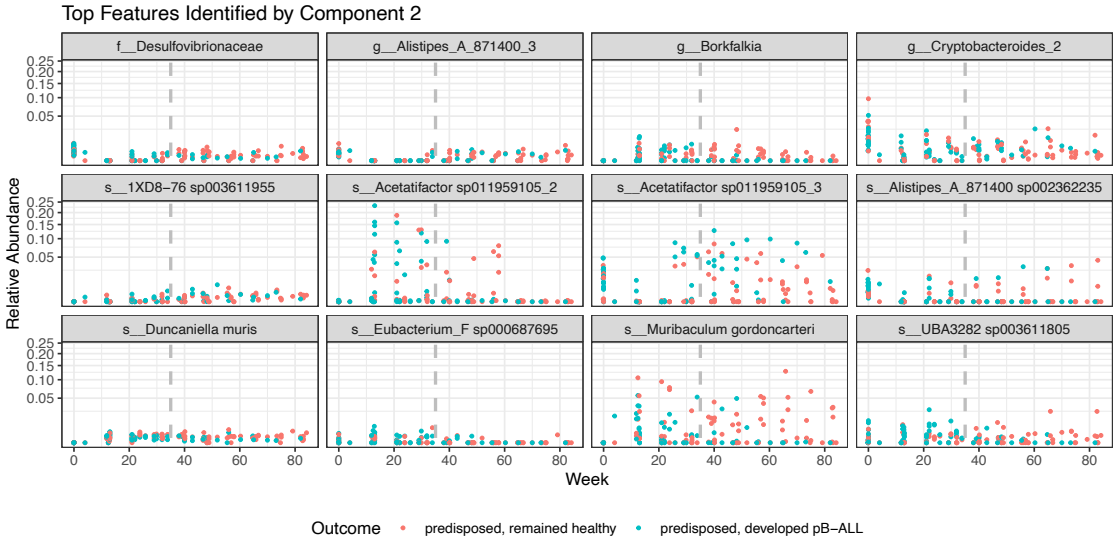

Figure 9: Relative abundance top ASVs identified by the second component of TEMPTED.

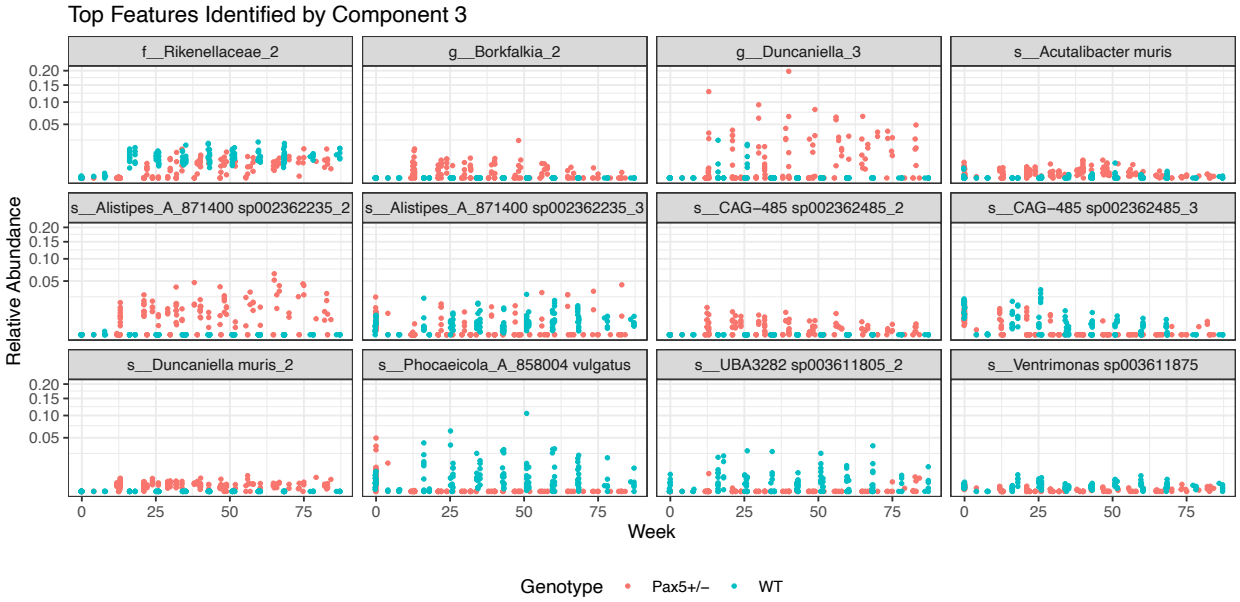

Figure 10: Relative abundance of top ASVs identified by the third component of TEMPTED.

References

1. Kimeldorf, G. & Wahba, G. Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications* **33,** 82–95 (1971).