TEMPTED: time-informed dimensionality reduction for longitudinal microbiome studies

Pixu Shi[1,2*†], Cameron Martino[3,4,5†], Rungang Han[6], Stefan Janssen[7], Gregory Buck[8,9], Myrna Serrano[8,9], Kouros Owzar[1], Rob Knight[3,4,10,11,12*], Liat Shenhav[13,14,15*‡], Anru R. Zhang[1,16*‡]

[1]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA.
[2]Duke Microbiome Center, Duke University, Durham, NC, USA.
[3]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA.
[4]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA.
[5]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA.
[6]Department of Statistical Science, Duke University, Durham, NC, USA.
[7]Algorithmic Bioinformatics, Department of Biology and Chemistry, Justus Liebig University Giessen, Giessen,Germany.
[8]Center for Microbiome Engineering and Data Analysis, Virginia Commonwealth University, Richmond, VA, USA.
[9]Department of Microbiology and Immunology, Virginia Commonwealth University, Richmond, VA, USA.
[10]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA.
[11]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA.
[12]Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA.
[13]Institute for Systems Genetics, New York Grossman School of Medicine, New York University, New York, NY, USA.
[14]Department of Microbiology, New York Grossman School of Medicine, New York University, New York, NY, USA.

[15]Department of Computer Science, New York University, New York, NY, USA.
[16]Department of Computer Science, Duke University, Durham, NC, USA.

*Corresponding author(s). E-mail(s): pixu.shi@duke.edu; robknight@ucsd.edu; liat.shenhav@nyulangone.org; anru.zhang@duke.edu;
Contributing authors: cameronmartino@gmail.com; hrg.stat@gmail.com; stefan.janssen@computational.bio.uni-giessen.de; gregory.buck@vcuhealth.org; Myrna.Serrano@vcuhealth.org; kouros.owzar@duke.edu;
†P. Shi and C. Martino contributed equally to this work.
‡L. Shenhav and A. R. Zhang contributed equally to this work.

Abstract

Longitudinal studies are crucial for understanding complex microbiome dynamics and their link to health. We introduce TEMPoral TEnsor Decomposition (TEMPTED), a time-informed dimensionality reduction method for high-dimensional longitudinal data that treats time as a continuous variable, effectively characterizing temporal information and handling varying temporal sampling. TEMPTED captures key microbial dynamics, facilitates beta-diversity analysis, and enhances reproducibility by transferring learned representations to new data. In simulations, it achieves 90% accuracy in phenotype classification, significantly outperforming existing methods. In real data, TEMPTED identifies vaginal microbial markers linked to term and preterm births, demonstrating robust performance across datasets and sequencing platforms.

Background

Given the highly dynamic and complex nature of microbial communities, identifying and predicting their time-dependent patterns are 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 are 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-up visits or inconsistent sample collection), multiple hosts often have missing temporal samples, translating into irregular temporal sampling across hosts [2–5].

To investigate 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 have been developed to model longitudinal microbiome data. CTF [1], microTensor [6], TCAM [7], FTSVD [8], and EMBED [9] 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 of 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, most of these methods can not transfer the learned low-dimensional representation from training data to independent testing 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 [10]. 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) [11–14]. Another example is the mixed-effect type models widely used to quantify intra-host variation, but typically analyze one microbial feature at a time with limited ability to model temporal patterns [15]. 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 testing data, thus facilitating research reproducibility. TEMPTED is unique in that it can handle varying temporal sampling and missing time points, a prevalent issue in longitudinal microbiome studies, without the need of time discretization, sample removal, or sample imputation. Treating time as a continuous variable allows adjacent time points to borrow information from each other, thus reducing the impact of noises and enhancing signals. It is the only dimensionality reduction method currently available for temporal data that offers this flexibility. These unique properties enable TEMPTED to have superior performance in extracting key information in the dataset in data-driven simulations. 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

An overview of TEMPTED

Let $i = 1, \ldots, n$ denote subjects, $j = 1, \ldots, p$ denote features, and $\boldsymbol{\mathcal{Y}}_{ijt}$ denote the value of feature j from subject i at time point $t \in T_i = \{t_{i1}, \ldots, t_{im_i}\}$. Here, T_i is a subset of interval T containing m_i time points and can be different across subjects to accommodate varying temporal sampling and missing time points. TEMPTED allows users to choose their own preferred data normalization and transformation to obtain $\boldsymbol{\mathcal{Y}}_{ijt}$. As illustrated in Fig. 1b, we adopt the model setting proposed in [8], which decomposes the temporal tensor formed by $\boldsymbol{\mathcal{Y}}_{ijt}$ using an approximately CANDECOMP/PARAFAC (CP) low-rank structure:

$$\boldsymbol{\mathcal{Y}}_{ijt} = \sum_{\ell=1}^{r} \lambda_\ell a_i^{(\ell)} b_j^{(\ell)} \xi^{(\ell)}(t) + \boldsymbol{\mathcal{Z}}_{ijt}, \tag{1}$$

where r is the number of low-rank components to approximate the data tensor $\boldsymbol{\mathcal{Y}}$, λ_ℓ quantifies the contribution of each component, $a^{(\ell)} = (a_1^{(\ell)}, \ldots, a_n^{(\ell)})$ are subject loadings, $b^{(\ell)} = (b_1^{(\ell)}, \ldots, b_p^{(\ell)})$ are feature loadings, $\xi^{(\ell)}(t)$ is the temporal loading that captures the shared temporal patterns among subjects and features, and $\boldsymbol{\mathcal{Z}}_{ijt}$ includes unexplained remainder terms and measurement errors. Different from FTSVD [8], where T_i is assumed to be identical across all subjects, TEMPTED can accommodate the more common scenario of varying temporal sampling T_i across subjects. Our objective is to estimate λ_ℓ, $a^{(\ell)}$, $b^{(\ell)}$, and $\xi^{(\ell)}(t)$ while requiring $\xi^{(\ell)}(t)$ to be smooth. For details of the assumptions on $\xi(t)$ and the algorithm for estimation, see Methods.

The subject loadings $a^{(\ell)}$ can be used for subject-level beta analysis such as classifier training. 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 all p features into the following subject-specific trajectory corresponding to Component ℓ:

$$\boldsymbol{\mathcal{B}}_{it}^{(\ell)} = \sum_{j=1}^{p} b_j^{(\ell)} \widetilde{\boldsymbol{\mathcal{Y}}}_{ijt}. \tag{2}$$

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 to construct Euclidean distance and perform sample-level beta analysis. The definition of (2) is generic and can be applied to any longitudinal temporal data.

Since most microbiome data from 16S or shotgun metagenomic sequencing are compositional, researchers are often interested in the relative abundance of one group of microbes versus another. In this scenario, the users can zoom into a small number of features most relevant to each component by specifying a quantile cutoff for the

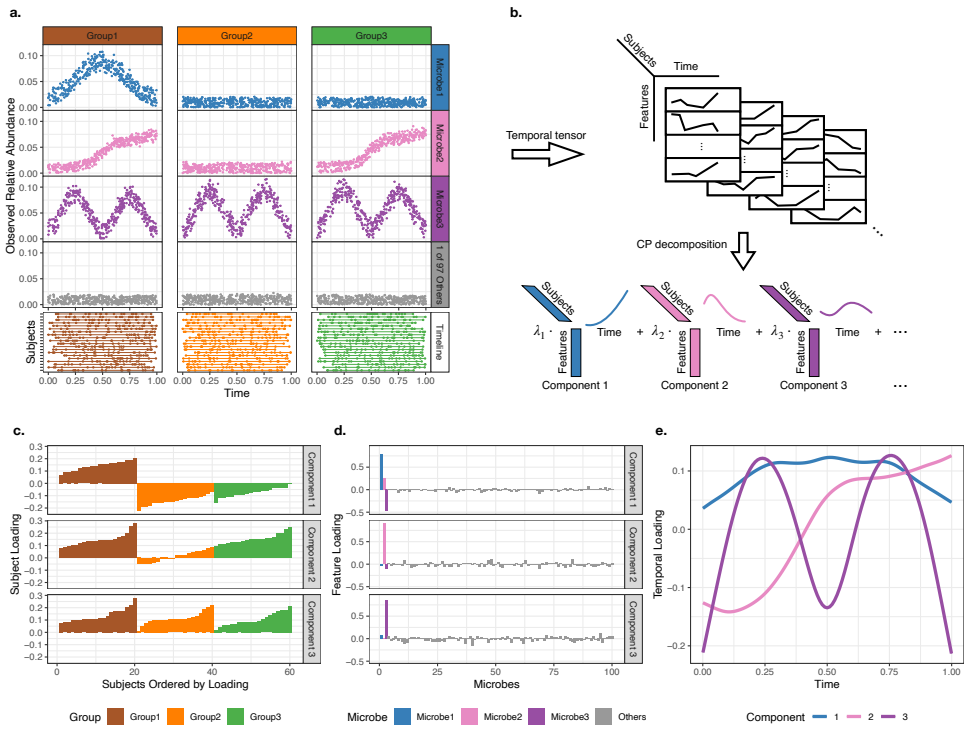

Fig. 1 Overview of the TEMPTED algorithm for analyzing multi-subject multi-feature temporal microbiome data. (a) Simulated microbiome count data (see Additional file 1) is transformed into relative abundance and plotted for four representative microbes and three groups of hosts. Microbe 1 has a unique 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 are 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 containing 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 groups. Microbes 4-100 have low feature loadings in all components.

feature loadings, and construct trajectories of log-ratio abundance of top over bottom ranking features:

$$\boldsymbol{\mathcal{B}}_{it}^{(\ell)} = \log\left(\left(0.5 + \sum_{j:b_j^{(\ell)} \text{ in top quantile}} \boldsymbol{\mathcal{C}}_{ijt}\right) \Big/ \left(0.5 + \sum_{j:b_j^{(\ell)} \text{ in bottom quantile}} \boldsymbol{\mathcal{C}}_{ijt}\right)\right), \tag{3}$$

where $\mathcal{C}_{ijt}$ is the read count of Feature j in the tth sample of Subject i. Details of the pseudocount chosen in this log ratio transformation can be found in Methods.

Data-driven simulation

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 (Additional file 1: Fig. S6). This dataset was chosen due to previously observed differences in longitudinal trajectories between delivery modes [1, 2]. The second simulation utilizes the FARMM [16] 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 (Additional file 1: Fig. S7). 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, 16]. 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 (see 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 the performance of TEMPTED as compared to CTF, microTensor, FTSVD, 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 testing data (see Methods) and performing out-of-sample prediction, although TCAM cannot handle missingness in time points (Fig. 2a,c). 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. It is also important to note that among all the non-PERMANOVA methods, TEMPTED is the only method treating time as a continuous variable without discretization or imputation. While CTF and microTensor allow missingness in the time points, they require discretizing time into intervals. On the other hand, TCAM and FTSVD require temporal sampling to be identical across subjects without any missing time points. To use these methods for highly varying temporal sampling, as in our ECAM-based simulation, without sample imputation, we transformed time into the order of samples (Fig. 2) or monthly intervals (Additional file 1: Fig. S5)

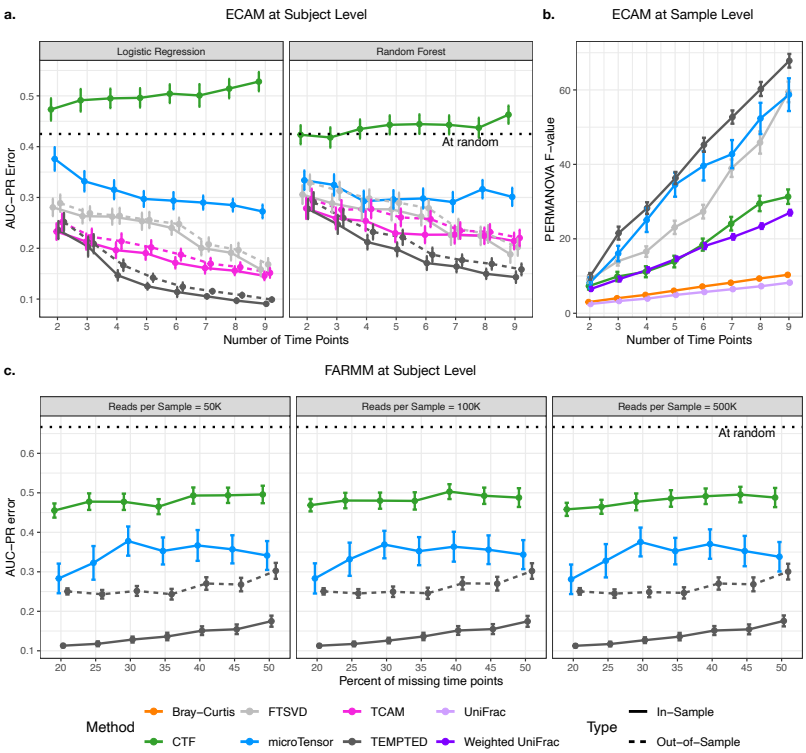

Fig. 2 TEMPTED outperforms CTF, microTensor, TCAM, FTSVD, and PCoA in identifying group structures. TEMPTED demonstrates superior performance in reducing host-level phenotype discriminatory error by more than 50% compared to methods capable of handling missing time points (a, c) and improves the sample-level group discriminatory power (b). Sample-level discriminatory power is quantified by PERMANOVA pseudo F-statistic based on the sample-level Euclidean distance constructed using the first two components from each method (b) (for TEMPTED see Eq. (2)). Host-level group classification error is quantified by AUC-PR (1 - area under the precision-recall curve) with the first two components of each method as predictors. 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 (See Methods) 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. For ECAM-based simulation, we randomly choose a given number of samples from each subject such that CTF, microTensor, FTSVD 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 and FTSVD cannot. EMBED was not included in the benchmarking because it does not provide host-level or sample-level beta diversity analysis. Different reads per sample are obtained by resampling reads in each sample.

Case studies

To further 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] (Additional file 1: Fig. S8). 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 (Additional file 1: Figs. S11-13, Tables S1-2). 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, Additional File 1: Fig. S14, Additional File 2). Additionally, TEMPTED identified ASVs that clearly separate wildtype and Pax5+/- mice (Fig. 3b-c, Additional file 1: Fig. S15). 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 [17–19].

Next, we applied TEMPTED to two independent studies that longitudinally sampled the vaginal microbiome during pregnancy: one study sequenced shotgun metagenomics and used metagenome-assembled genomes (MAGs) [4, 20] and while the second sequenced 16S rRNA genes and used ASVs [4, 5] (Additional file 1: Fig. S9-10). Remarkably, TEMPTED consistently captured differences in the vaginal microbiome trajectories of pregnancies that resulted in term versus 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, h), which can potentially serve as a biomarker to detect preterm birth early in pregnancy (Fig. 3f, i). Of note, the dynamics of these features are required to differentiate between term and preterm birth as this separation was not found when performing differential abundance analysis using ALDEx2 [21] at any single time point (Additional Files 3-4; 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, h), consistent with existing literature [4, 20, 22]. Overall, our results highlight the robustness of TEMPTED in multiple studies and sequencing technologies in uncovering microbial dynamics underlying host phenotypes.

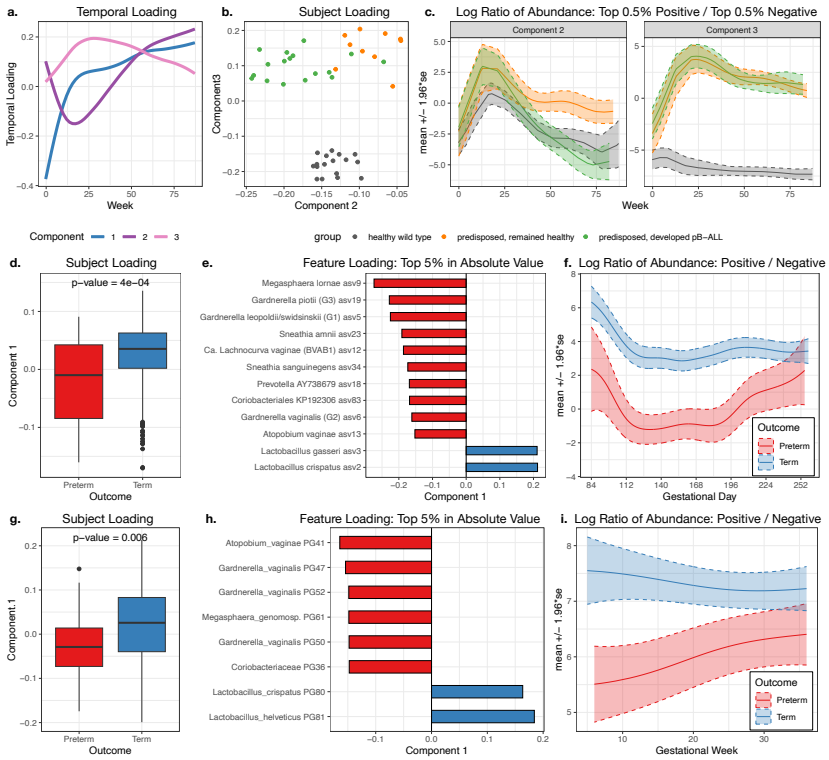

Fig. 3 TEMPTED reveals latent trajectories distinguishing health phenotypes. (a)-(c) show the results of applying TEMPTED to a mouse study on leukemia (17 wild-type mice, 27 Pax5+/- mice of which 17 developed leukemia)3. (a) Temporal loadings capture major temporal patterns in the central-log-ratio-transformed 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, 20]. (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.

Discussion

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, FTSVD, and TCAM, TEMPTED extracts prominent structures from the data, but does not guarantee the capture of the phenotype of interest in

its leading components or a single component or look for differential temporal trends between host groups. TEMPTED's smoothness assumption (see Methods) on the temporal loading allows it to extract key temporal trends beneath the noisy observed data, but such an assumption also makes it unsuitable for change point detection. Currently, TEMPTED cannot handle individual missing entries in the data tensor, which could be addressed in future work. In addition, TEMPTED does not guarantee its rank-r decomposition to have the best reconstruction accuracy because it obtains its low-rank structure for each component sequentially to achieve the uniqueness of the decomposition instead of looking for the best rank-r approximation for a given rank. 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.

Conclusions

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 patterns shared across hosts and features, TEMPTED efficiently extracts important low-dimensional structures from high-dimensional longitudinal data, identifies major temporal dynamics and key contributing features, 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 extract low-dimensional data structures associated with 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 to transfer the learned low-dimensional representation from training to testing data, make it a highly practical tool for microbiome research, promoting the research reproducibility. The flexible model setting of TEMPTED also makes it applicable to a wide range of high-dimensional temporal data, potentially benefiting more research beyond longitudinal microbiome studies.

Methods

In this section, we provide details of our methods, including data preprocessing, the TEMPTED algorithm, tensor reconstruction, and how the dimension reduction

learned from training data can be transferred to new testing data. Additionally, we present details of case studies (Pax5 Mice Leukemia data, shotgun metagenomic vaginal data, 16S Vaginal data), and data-driven simulation studies.

Microbiome Data Preprocessing

Here, we present the data transformation we adopted for microbiome sequencing data specifically to address their compositionality and highly skewed distribution. For other sources of data, users can choose their desired transformation and normalization before using TEMPTED. Let $i = 1, \ldots, n$ denote subjects, $j = 1, \ldots, p$ denote features, and $\boldsymbol{\mathcal{C}}_{ijt}$ denote the observed read count of feature j from subject i at time points $t \in T_i = \{t_{i1}, \ldots, t_{im_i}\}$. We apply the centered-log-ratio (CLR) transformation to read counts added by .5 [23, 24]

$$\boldsymbol{\mathcal{Y}}_{ijt} = \log\left(\frac{\boldsymbol{\mathcal{C}}_{ijt} + .5}{\left(\prod_{s=1}^{p}(\boldsymbol{\mathcal{C}}_{ist} + .5)\right)^{1/p}}\right). \tag{4}$$

Adding the pseudocount .5 instead of other values is theoretically justified by [24]. [24] showed that $\log(\boldsymbol{\mathcal{C}}_{ijt} + \alpha_i/2)$ has the smallest bias in the estimation of log(mean of $\boldsymbol{\mathcal{C}}_{ijt}$) when $(\boldsymbol{\mathcal{C}}_{i1t}, \ldots, \boldsymbol{\mathcal{C}}_{ipt})$ follow Dirichlet-Multinomial distribution with α_i being the overdispersion parameter. Since microbiome sequencing count data are generally equally or more dispersed than multinomial distribution [25], when the estimation of α_i is difficult, we opt to estimate α_i with 1, leading to the pseudocount of .5. In Additional file 1: Fig. S1, we also compare the pseudocount of 0.5 with 0.1 and 1. While using 0.1 yields slightly worse performance in subject-level and sample-level dimension reduction, choosing between 0.5 and 1 has little difference. It is also worth noting that among the methods we benchmarked against in our simulation analyses, TCAM, CTF, and FTSVD all require log transformation of the count data. microTensor models count data directly without adding pseudocounts and log transformation, yet it is still outperformed by our method.

TEMPTED algorithm

The goal of TEMPTED is to obtain estimates of λ_ℓ, $a^{(\ell)}$, $b^{(\ell)}$ and $\xi^{(\ell)}(t)$ in the approximately CP low-rank structure (1). To overcome the issue of scaling identifiability (e.g., $(\lambda_l, a^{(l)}, b^{(l)}, \xi^{(l)}(t))$ and $(x_1\lambda_l, x_2a^{(l)}, x_3b^{(l)}, x_4\xi^{(l)}(t))$ essentially represent the same component whenever $x_1x_2x_3x_4 = 1$), we opt to estimate each component ℓ sequentially for $\ell = 1, \ldots, r$ by minimizing the following objective function (5).

$$\min_{a_i, b_j, \xi} \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{t\in T_i}\left(\boldsymbol{\mathcal{Y}}_{ijt} - a_i^{(\ell)}b_j^{(\ell)}\zeta^{(\ell)}(t)\right)^2 + C_K||\zeta^{(\ell)}(t)||_{\mathcal{H}}^2 \tag{5}$$

$$\text{subject to } \sum_{i=1}^{n}(a_i^{(\ell)})^2 = \sum_{j=1}^{p}(b_j^{(\ell)})^2 = 1. \tag{6}$$

While the sequential estimation of components does not guarantee the smallest reconstruction error, it preserves the uniqueness of each component regardless of the chosen rank r and offers the significant advantage of allowing users to explore additional components without impacting those previously obtained.

In (5), $\|\zeta\|_{\mathcal{H}}^2$ is the *reproducing kernel Hilbert space (RKHS) norm* of function $\zeta(t)$ with the rescaled Bernoulli polynomial as the reproducing kernel,

$$\mathbb{K}(s,t) = 1 + k_1(s)k_2(t) + k_2(s)k_2(t) - k_4(|s-t|), \tag{7}$$

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 $\zeta(t)$s to be absolutely continuous and squared-integrable in its second-order derivative, and C_K is a tuning parameter controlling the smoothness of $\zeta(t)$s. Such smoothness assumption does not require the temporal trends themselves to be polynomials. Commonly seen temporal trends such as monotone, unimodal, bimodal, seasonal or circadian trends, can all satisfy this smoothness assumption. It is worth noting that the smoothness assumption is on the underlying structure of the tensor, not the observed tensor itself. Thus, variation of microbiome data due to noises does not necessarily violate our smoothness assumptions.

Due to the substantial differences in abundance among bacterial taxa across most time points, we offer an optional step we referred to as "mean subtraction", to extract and remove this time-invariant structure from $\boldsymbol{\mathcal{Y}}$ to improve the efficiency of our algorithm when this structure is not of interest. Specifically, we calculate the average of $\boldsymbol{\mathcal{Y}}_{ijt}$ across time points t, resulting in $\boldsymbol{\mathcal{Y}}_{ij\cdot}$, which forms the matrix $\bar{Y} = (\boldsymbol{\mathcal{Y}}_{ij\cdot})$. Then, we calculate the first singular component of $\bar{Y}$ as $\bar{Y}^{(0)} = \lambda_0 a^{(0)} b^{(0)\top}$, and subtract $\bar{Y}^{(0)}$ from $\boldsymbol{\mathcal{Y}}_t$ for each t. This optional mean subtraction step reduces the effect of highly abundant bacteria taxa and other time-invariant factors that may confound downstream analysis. We adopted this mean subtraction in all our simulation analyses in the main text and real data analyses. In Additional file 1: Section S3, we provide more insights into the effect of mean subtraction through simulation based on the ECAM data and FARMM data. In summary, we found that the reconstruction accuracy of TEMPTED with mean subtraction at rank r is comparable to but slightly better than that of TEMPTED without mean subtraction at rank $r+1$ (Additional file: Fig. S3).The ability of TEMPTED to capture group differences is greatly reduced without mean subtraction, but this reduction can be compensated by adding another component to the classification and is highly dependent on the dataset (Additional file 1: Fig. S4). These results suggest the possibility of TEMPTED capturing the mean structure in some of its components depending on its significance in the dataset. Depending on whether the mean structure contains useful information for downstream analyses, users can decide if mean subtraction is needed before applying TEMPTED.

Set $\widetilde{\boldsymbol{\mathcal{Y}}} = \boldsymbol{\mathcal{Y}}$ or $\boldsymbol{\mathcal{Y}}$ after mean subtraction. For each component $\ell = 1, \ldots, r$ sequentially, we perform the following Steps 1 to 3 to estimate a, b, and $\zeta(t)$.

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 $\widetilde{\boldsymbol{\mathcal{Y}}}$: i.e., the p-by-$\left(\sum_{i=1}^n |T_i|\right)$ matrix with $\{\widetilde{\boldsymbol{\mathcal{Y}}}_{i\cdot t} \in \mathbb{R}^p\}_{i=1,\ldots,n, t\in T_i}$ as its columns.

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

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{t\in T_i}\left(\widetilde{\boldsymbol{\mathcal{Y}}}_{ijt} - a_i b_j \zeta(t)\right)^2 + C_K\|\zeta\|_{\mathcal{H}}^2. \tag{8}$$

(a) Update $\hat{\zeta}$ by applying kernel ridge regression to solve

$$\hat{\zeta} = \arg\min_{\zeta\in\mathcal{H}} \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{t\in T_i}\left(\widetilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \hat{a}_i\hat{b}_j \zeta(t)\right)^2 + C_K\|\zeta\|_{\mathcal{H}}^2. \tag{9}$$

The details of this update are described in the next section.

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

$$\hat{a}_i = \frac{\sum_{j=1}^{p}\sum_{t\in T_i}\widetilde{\boldsymbol{\mathcal{Y}}}_{ijt}\hat{b}_j\hat{\zeta}(t)}{\sum_{t\in T_i}\hat{\zeta}(t)^2}, \quad \text{and } \hat{a} = \hat{a}/\|\hat{a}\|_2. \tag{10}$$

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

$$\hat{b}_j = \frac{\sum_{i=1}^{n}\sum_{t\in T_i}\widetilde{\boldsymbol{\mathcal{Y}}}_{ijt}\hat{a}_i\hat{\zeta}(t)}{\sum_{i=1}^{n}\sum_{t\in T_i}(\hat{a}_i\hat{\zeta}(t))^2}, \quad \text{and } \hat{b} = \hat{b}/\|\hat{b}\|_2. \tag{11}$$

Step 3: (Subtracting previous components) Normalize $\hat{\zeta}$ to $\hat{\xi} = \hat{\zeta}/\|\hat{\zeta}\|_2$. Update $\widetilde{\boldsymbol{\mathcal{Y}}}$ by

$$\widetilde{\boldsymbol{\mathcal{Y}}}_{ijt} = \widetilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \hat{\eta}\hat{a}_i\hat{b}_j\hat{\xi}(t), \tag{12}$$

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

$$\arg\min_{\eta} \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{t\in T_i}\left(\widetilde{\boldsymbol{\mathcal{Y}}}_{ijt} - \eta\hat{a}_i\hat{b}_j\hat{\xi}(t)\right)^2. \tag{13}$$

Step 4: After obtaining $\{\hat{a}^{(l)}, \hat{b}^{(l)}, \hat{\xi}^{(l)}(t)\}_{l=1}^{r}$ by sequentially running Steps 1-3, we estimate $\lambda = (\lambda_1, \ldots, \lambda_r)$ is estimated via the following least squares problem:

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda} \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{t\in T_i}\left(\boldsymbol{\mathcal{Y}}_{ijt} - \sum_{\ell=1}^{r}\lambda_\ell\hat{a}_i^{(\ell)}\hat{b}_j^{(\ell)}\hat{\xi}^{(\ell)}(t)\right)^2. \tag{14}$$

Kernel Ridge Regression

By the presenter theorem [26], the solution to (9) 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:

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

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

$$D = \begin{pmatrix} \overbrace{b_1^{(t)} a_1^{(t)} \ddots b_1^{(t)} a_1^{(t)}}^{|T_1| \text{ copies}} & & & \\ & \ddots & & \\ & & b_1^{(t)} a_n^{(t)} \overbrace{\ddots b_1^{(t)} a_n^{(t)}}^{|T_n| \text{ copies}} & \\ & & & \vdots \\ \vdots & & & \\ \overbrace{b_p^{(t)} a_1^{(t)} \ddots b_p^{(t)} a_1^{(t)}}^{|T_1| \text{ copies}} & & & \\ & \ddots & & \\ & & b_p^{(t)} a_n^{(t)} \overbrace{\ddots b_p^{(t)} a_n^{(t)}}^{|T_n| \text{ copies}} & \end{pmatrix}$$

The solution to formula (9) is equivalent to

$$\min_{\alpha \in \mathbb{R}^{|T_1|+\cdots+|T_n|}} (z - DK\alpha)^\top (z - DK\alpha) + C_K \alpha^\top K \alpha, \tag{15}$$

which can be solved by

$$\hat{\alpha} = (D^\top DK + C_K I)^{-1} D^\top z, \quad \hat{\zeta}(s) = \alpha^\top \left(\mathbb{K}(T_1, s), \ldots, \mathbb{K}(T_n, s)\right)^\top. \tag{16}$$

Here, I is the identity matrix, and $b^{(t)}$ will cancel out in $D^\top D$ because $||b^{(t)}||_2 = 1$. C_K is a tuning parameter that makes $\zeta(t)$ smoother when C_K is larger. The implementation of TEMPTED allows users to choose the value of C_K. We used $C_K = 0.0001$ for all our case studies and found it to work well for most datasets we analyzed with more than four time points. This is validated by a sensitivity analysis of C_K using the ECAM-based simulated data (Additional file 1: Fig. S2). Our results also indicate that while sample-level beta diversity derived from feature loadings is very insensitive to the choice of C_K, a larger C_K is needed for a smaller number of time points to ensure good performance in the subject-level beta-diversity analysis.

Tensor Reconstruction

In addition to estimating the low-rank components, we can also construct $\widehat{\boldsymbol{\mathcal{Y}}}$, a low-rank approximation of the target tensor $\boldsymbol{\mathcal{Y}}$, using the following method:

$$\widehat{\boldsymbol{\mathcal{Y}}}_{ijt} = \sum_{\ell=1}^{r} \hat{\lambda}_\ell \hat{a}_i^{(\ell)} \hat{b}_j^{(\ell)} \hat{\xi}^{(\ell)}(t). \tag{17}$$

When mean subtraction is applied, the subtracted mean $\bar{Y}^{(0)}$ needs to be added back to $\widehat{\boldsymbol{\mathcal{Y}}}_t$. We evaluate the reconstruction accuracy of the decomposition in terms of normalized Frobenius norm:

$$\frac{\|\widehat{\boldsymbol{\mathcal{Y}}} - \boldsymbol{\mathcal{Y}}\|_F^2}{\|\boldsymbol{\mathcal{Y}} - \text{mean}(\boldsymbol{\mathcal{Y}})\|_F^2} \tag{18}$$

This reconstruction error ranges from 0 to 1 and decreases as the rank r increases for TEMPTED and FTSVD. It can guide the selection of r in the same fashion as (1 - percent of variance explained) in principal component analysis.

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 to $a_{\text{test}}^{(\ell)}$ trained by $a_{\text{train}}^{(\ell)}$. The prediction of phenotype for the testing data by such classifiers is purely out-of-sample since the dimension reduction is performed without any knowledge of $\boldsymbol{\mathcal{Y}}_{\text{test}}$, and classifiers trained by $a_{\text{train}}^{(\ell)}$ have no information from the testing data. The accuracy of such out-of-sample prediction is evaluated through data-driven simulation (Fig. 2 dashed lines).

Case Study: Pax5 Mice Leukemia Data

The Pax5 dataset was published by [3] and deposited at https://qiita.ucsd.edu/study/description/11953, artifact ID 75878 [27]. 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 [28]. Please refer to [3] 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. We used the top 3 components for all dimension reduction methods when analyzing this dataset. The smoothness parameter C_K was set to 10^{-4} for TEMPTED.

Case Study: Shotgun Metagenomic Vaginal Data

The dataset used in this study was collected and published by [4] and deposited under dbGaP (study no. 20280; accession ID phs001523.v1.p1) [29]. 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. We conducted a rank-2 decomposition of this dataset using TEMPTED, focusing specifically on interpreting the first component. The smoothness parameter C_K was set to 10^{-4}.

Case Study: 16S Vaginal Data

The 16S data used in this study were published by [4, 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 [30]. 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. We performed rank-2 decomposition of this dataset using TEMPTED, focusing specifically on interpreting the first component. The smoothness parameter C_K was set to 10^{-4}.

Data Driven Simulation

The ECAM dataset was published by [2] (Qiita ID 10249). Preprocessed data were obtained from [31]. 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 was used as the time variable. Since CTF and microTensor can handle some missingness, we also ran them with time points rounded to month. The results are summarized in Additional file 1: Fig. S5, which shows worse performance for microTensor and slightly improved performance for CTF, but the superiority of TEMPTED remained obvious. For TEMPTED, the smoothness parameter C_K was set to 1, 0.1, 0.01, 0.005, for $m = 2, 3, 4, 5$ respectively, and 10^{-4} for $m = 6, \ldots, 9$. These values of C_K match with the optimal C_K indicated by the sensitivity analysis in Additional file 1: Fig. S2.

The FARMM dataset was published by [16] (BioProject ID PRJNA675301). Preprocessed relative abundance data were obtained from [32]. 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 randomly dropped to achieve different percentages of missingness. Time points were used as is for TEMPTED, 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. For TEMPTED, the smoothness parameter C_K was set to 10^{-4} in all settings.

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. The number of ranks r is set to 2 for all methods in the simulation analysis in Fig. 2.

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).

Declarations

Availability of data and materials. The TEMPTED method is built into the R package "tempted" on CRAN (source codes and tutorial on GitHub

https://github.com/pixushi/TEMPTED and Zenodo [33]) and the Python package "gemelli" (source codes and tutorial on GitHub https://github.com/biocore/gemelli and Zenodo [34]). All data analysis and simulation are performed using R and Python scripts through the R package reticulate. All processed data and codes used for this paper are available on GitHub https://github.com/pixushi/TEMPTED_paper and Zenodo [35].

The Pax5 mice dataset [3] was deposited at Qiita with artifact ID 75878 [27] and in the ENA under accession PRJEB34720 [28].

The shotgun vaginal microbiome dataset [4, 20] was deposited under dbGaP (study no. 20280; accession ID phs001523.v1.p1) [29].

The 16S vaginal microbiome dataset [3, 5] was deposited under BioProjects PRJNA393472 (subjects enrolled at the University of Alabama, Birmingham) and PRJNA821262 (subjects enrolled at Stanford University). Preprocessed data were obtained from [30].

The ECAM dataset [2] was published under Qiita ID 10249. Preprocessed data were obtained from [31].

The FARMM dataset [16] was deposited under BioProject ID PRJNA675301. Preprocessed relative abundance data were obtained from [32].

Funding. This research was supported by NIH Director's New Innovator Award (DP2AI185753), NICHD R01HD092415, NIH Director's Pioneer Award (DP1AT010885), CDC BAA Award (75D301-22-C-14717), Emerald Foundation Distinguished Investigator Award, NCI R01CA241728, NCI U24CA248454, NSF CAREER-2203741, NHLBI R01HL169347, NHLBI R01HL168940, and Duke Microbiome Center.

Authors' contributions. PS, RH, and AZ conceived TEMPTED. PS designed and implemented TEMPTED in R. CM implemented TEMPTED in Python. PS, CM, and LS conceived the data-driven simulation scheme. PS performed the data-driven simulation and analyzed the 16S vaginal microbiome study. SJ and PS analyzed the Pax5 mice leukemia study. LS analyzed the shotgun vaginal microbiome study. PS, CM, and LS wrote the manuscript, with input from RK, SJ, KO, GB, and MS. RK supervised CM and LS. All authors reviewed and approved the final manuscript.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Competing interests. Rob Knight is a scientific advisory board member, and consultant for BiomeSense, Inc., has equity and receives income. He is a scientific advisory board member and has equity in GenCirq. He is a consultant for DayTwo, and receives income. He has equity in and acts as a consultant for Cybele. He is a co-founder

of Biota, Inc., and has equity. He is a cofounder of Micronoma, and has equity and is a scientific advisory board member. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. Cameron Martino is the founder of Leaven Foods, inc. and has equity. Gregory Buck is on the Scientific Advisory Board for Juno LTC.

Additional Files

Additional file 1: Supplementary Materials. This file provides additional details and results from numerical studies, including Figs. S1-S15 and Tables S1-S2.

Additional file 2: Taxonomy of top ASVs. This table contains the detailed taxonomic assignment of top ASVs identified by TEMPTED in the Pax5 Mice Leukemia Data and plotted in Figs. S14 and S15.

Additional file 3: ALDEx2 on 16S Vaginal Data. This table contains the result of ALDEx2 applied to the 16S vaginal data.

Additional file 4: ALDEx2 on Shotgun Metagenomic Vaginal Data. This table contains the result of ALDEx2 applied to the shotgun metagenomic vaginal data.

References

[1] Martino, C., Shenhav, L., Marotz, C.A., Armstrong, G., McDonald, D., Vázquez-Baeza, Y., Morton, J.T., Jiang, L., Dominguez-Bello, M.G., Swafford, A.D., *et al.*: Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. Nature Biotechnology **39**(2), 165–168 (2021)

[2] Bokulich, N.A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., D. Lieber, A., Wu, F., Perez-Perez, G.I., Chen, Y., *et al.*: Antibiotics, birth mode, and diet shape microbiome maturation during early life. Science translational medicine **8**(343), 343–8234382 (2016)

[3] Vicente-Dueñas, C., Janssen, S., Oldenburg, M., Auer, F., González-Herrero, I., Casado-García, A., Isidro-Hernández, M., Raboso-Gallego, J., Westhoff, P., Pandyra, A.A., *et al.*: An intact gut microbiome protects genetically predisposed mice against leukemia. Blood, The Journal of the American Society of Hematology **136**(18), 2003–2017 (2020)

[4] Fettweis, J.M., Serrano, M.G., Brooks, J.P., Edwards, D.J., Girerd, P.H., Parikh, H.I., Huang, B., Arodz, T.J., Edupuganti, L., Glascock, A.L., *et al.*: The vaginal microbiome and preterm birth. Nature Medicine **25**(6), 1012–1021 (2019)

[5] Costello, E.K., DiGiulio, D.B., Robaczewska, A., Symul, L., Wong, R.J., Shaw, G.M., Stevenson, D.K., Holmes, S.P., Kwon, D.S., Relman, D.A.: Abrupt perturbation and delayed recovery of the vaginal ecosystem following childbirth. Nature Communications **14**(1), 4141 (2023)

[6] Ma, S., Li, H.: A tensor decomposition model for longitudinal microbiome studies. The Annals of Applied Statistics **17**(2), 1105–1126 (2023)

[7] Mor, U., Cohen, Y., Valdés-Mas, R., Kviatcovsky, D., Elinav, E., Avron, H.: Dimensionality reduction of longitudinal'omics data using modern tensor factorizations. PLoS Computational Biology **18**(7), 1010212 (2022)

[8] Han, R., Shi, P., Zhang, A.R.: Guaranteed functional tensor singular value decomposition. Journal of the American Statistical Association, 1–13 (2023)

[9] 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), 26 (2023)

[10] Happ, C., Greven, S.: Multivariate functional principal component analysis for data observed on different (dimensional) domains. Journal of the American Statistical Association **113**(522), 649–659 (2018)

[11] Gibson, T., Gerber, G.: Robust and scalable models of microbiome dynamics. In: International Conference on Machine Learning, pp. 1763–1772 (2018). PMLR

[12] Shenhav, L., Furman, O., Briscoe, L., Thompson, M., Silverman, J.D., Mizrahi, I., Halperin, E.: Modeling the temporal dynamics of the gut microbial community in adults and infants. PLoS Computational Biology **15**(6), 1006960 (2019)

[13] Äijö, T., Müller, C.L., Bonneau, R.: Temporal probabilistic modeling of bacterial compositions derived from 16s rRNA sequencing. Bioinformatics **34**(3), 372–380 (2018)

[14] 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 **6**, 1–20 (2018)

[15] Fitzmaurice, G.M., Laird, N.M., Ware, J.H.: Applied Longitudinal Analysis, (2012)

[16] Tanes, C., Bittinger, K., Gao, Y., Friedman, E.S., Nessel, L., Paladhi, U.R., Chau, L., Panfen, E., Fischbach, M.A., Braun, J., *et al.*: Role of dietary fiber in the recovery of the human gut microbiome and its metabolome. Cell host & microbe **29**(3), 394–407 (2021)

[17] Buteyn, N.J., Santhanam, R., Merchand-Reyes, G., Murugesan, R.A., Dettorre, G.M., Byrd, J.C., Sarkar, A., Vasu, S., Mundy-Bosse, B.L., Butchar, J.P., *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**(7), 1988–1997 (2020)

[18] Lee, C., Hong, S.N., Paik, N.Y., Kim, T.J., Kim, E.R., Chang, D.K., Kim, Y.-H.: Cd1d modulates colonic inflammation in nod2-/- mice by altering the intestinal microbial composition comprising acetatifactor muris. Journal of Crohn's and Colitis **13**(8), 1081–1091 (2019)

[19] Chen, L., Zhai, Y., Wang, Y., Fearon, E.R., Núñez, G., Inohara, N., Cho, K.R.: Altering the microbiome inhibits tumorigenesis in a mouse model of oviductal high-grade serous carcinoma. Cancer research **81**(12), 3309–3318 (2021)

[20] Liao, J., Shenhav, L., Urban, J.A., Serrano, M., Zhu, B., Buck, G.A., Korem, T.: Microdiversity of the vaginal microbiome is associated with preterm birth. Nature Communications **14**(1), 4997 (2023)

[21] 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**(7), 67019 (2013)

[22] Sun, S., Serrano, M.G., Fettweis, J.M., Basta, P., Rosen, E., Ludwig, K., Sorgen, A.A., Blakley, I.C., Wu, M.C., Dole, N., *et al.*: Race, the vaginal microbiome, and spontaneous preterm birth. mSystems **7**(3), 00017–22 (2022)

[23] Aitchison, J., Aitchison, J.W.: The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability, (1986)

[24] Shi, P., Zhou, Y., Zhang, A.R.: High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis. Biometrika **109**(2), 405–420 (2022)

[25] McMurdie, P.J., Holmes, S.: Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Computational Biology **10**(4), 1003531 (2014)

[26] Kimeldorf, G., Wahba, G.: Some results on tchebycheffian spline functions. Journal of mathematical analysis and applications **33**(1), 82–95 (1971)

[27] Vicente-Dueñas, C., Janssen, S., Oldenburg, M., Auer, F., González-Herrero, I., Casado-García, A., Isidro-Hernández, M., Raboso-Gallego, J., Westhoff, P., Pandyra, A.A., et al.: Pax5 dataset. Qiita, University of California San Diego. artifact ID: 75878. https://qiita.ucsd.edu/study/description/11953

[28] PRJEB34720: Multi-Omic Microbiome Study Dataset. European Nucleotide Archive (ENA). Accession ID: PRJEB34720 (2019). https://www.ebi.ac.uk/ena/browser/view/PRJEB34720

[29] Multi-Omic Microbiome Study: Pregnancy Initiative (MOMS-PI). dbGaP. Accession ID: phs001523.v1.p1 (2017). https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001523.v1.p1

[30] Costello, E., DiGiulio, D., Robaczewska, A., Symul, L., Wong, R., Shaw, G., Stevenson, D., Holmes, S., Kwon, D., Relman, D.: Vaginal microbiome before and after childbirth. Stanford Digital Repository (2022). https://purl.stanford.edu/pz745bc9128

[31] Martino, C., Shenhav, L., Marotz, C., Armstrong, G., McDonald, D., Vazquez-Baeza, Y., Morton, J.T., Jiang, L., Dominguez-Bello, M.G., Swafford, A.D., Halperin, E., Knight, R.: Dataset and Code Capsule for "Context-aware dimensionality reduction deconvolutes gut microbial community dynamics". Code Ocean (2020). https://codeocean.com/capsule/6494482/tree/v1

[32] Ma, S., Li, H.: FARMM Dataset: Microbial Tensor Data for Analysis. GitHub (2023). https://github.com/syma-research/microTensor/tree/main/data/FARMM

[33] Shi, P.: TEMPTED. Zenodo. https://doi.org/10.5281/zenodo.14188190 (2024)

[34] Martino, C.: Gemelli. Zenodo. https://doi.org/10.5281/zenodo.14165979 (2024)

[35] Shi, P.: TEMPTED_paper. Zenodo. https://doi.org/10.5281/zenodo.14188193 (2024)