BMC Genomics

RESEARCH Open Access

A robust and transformation-free joint model with matching and regularization for metagenomic trajectory and disease onset

Qian Li[1*], Kendra Vehik[2], Cai Li[1], Eric Triplett[3], Luiz Roesch[3], Yi-Juan Hu[4] and Jeffrey Krischer[2]

Abstract

Background: To identify operational taxonomy units (OTUs) signaling disease onset in an observational study, a powerful strategy was selecting participants by matched sets and profiling temporal metagenomes, followed by trajectory analysis. Existing trajectory analyses modeled individual OTU or microbial community without adjusting for the within-community correlation and matched-set-specific latent factors.

Results: We proposed a joint model with matching and regularization (JMR) to detect OTU-specific trajectory predictive of host disease status. The between- and within-matched-sets heterogeneity in OTU relative abundance and disease risk were modeled by nested random effects. The inherent negative correlation in microbiota composition was adjusted by incorporating and regularizing the top-correlated taxa as longitudinal covariate, pre-selected by Bray-Curtis distance and elastic net regression. We designed a simulation pipeline to generate true biomarkers for disease onset and the pseudo biomarkers caused by compositionality. We demonstrated that JMR effectively controlled the false discovery and pseudo biomarkers in a simulation study generating temporal high-dimensional metagenomic counts with random intercept or slope. Application of the competing methods in the simulated data and the TEDDY cohort showed that JMR outperformed the other methods and identified important taxa in infants' fecal samples with dynamics preceding host disease status.

Conclusion: Our method JMR is a robust framework that models taxon-specific trajectory and host disease status for matched participants without transformation of relative abundance, improving the power of detecting disease-associated microbial features in certain scenarios. JMR is available in R package mtradeR at https://github.com/qianli10000/mtradeR.

Keywords: Metagenomic trajectory, Compositional data, Joint model, Zero-inflated, Pseudo biomarker

Background

Gut microbiota profiled by 16s rRNA gene sequencing or metagenomic (i.e., whole-genome shotgun) sequencing has been frequently used in observational studies of environmental exposures, immune biomarkers, and disease onset [1–5]. One of the challenges in analyzing microbiota in an observational study is to incorporate the matching between participants based on certain confounding risk factors (e.g. gender, clinical site, etc.) and/or disease status (case-control), such as the DIABIMMUNE and TEDDY cohorts [1, 2, 5]. A matching design effectively eliminates the noise effect of sample collection, storage, shipment, sequencing batch, and environmental exposures confounding with disease outcomes, as well as reduces the sequencing costs. Statistical analyses of microbiota in matched sets included, but are not limited

*Correspondence: qian.li@stjude.org

[1] Department of Biostatistics, St. Jude Children's Research Hospital, Memphis 38105, TN, USA
Full list of author information is available at the end of the article

to, conditional logistic regression [1], non-parametric comparison PERMANOVA [6] and LDM [7] with extension to compare cases and controls within a matched set [8], which aimed to model and analyze microbiome data at independent time points.

Longitudinal profiling is a powerful strategy for the microbiome studies that aim to identify differential microbial trajectories between exposure groups or phenotypes [9, 10] or detect the time intervals of differential abundance [11]. However, most of these studies failed to test if the compositional trajectory of an operational taxonomic unit (OTU) signaled host disease status. To detect microbial trajectories predictive of disease outcome in matched sets, an intuitive method is the generalized linear mixed effect model with or without the zero-inflation component [9, 10, 12], in which a taxon's abundance and/or presence is the outcome variable and the disease status is the covariate of interest. The Zero-Inflated Beta Regression (ZIBR) model [9] tests the association between OTU and a covariate factor using a two-part model for the non-zero relative abundance and presence of each OTU, assuming the non-zero relative abundance and presence being independent. A similar framework [10] was proposed to analyze the longitudinal zero-inflated counts per OTU using a Negative Binomial distribution, without converting the raw counts to relative abundance. A semi-parametric approach for longitudinal taxon-specific relative abundance is the linear mixed effect model (LMM) with asin-square-root transformation, which has been implemented in MaAsLin 2 [12].

One concern about using generalized linear mixed model to test the association between 16S rRNA or metagenomic trajectory and disease onset is that the covariates in this model may contribute to disease risk. For example, the HLA haplogenotypes and early use of probiotics may affect infants' gut microbiota and should be included as covariates. These factors were also found associated with islet autoimmunity among children enrolled in TEDDY [13]. One usually added interaction terms between each covariate and the disease outcome [3, 12] to adjust for the association. However, a linear model with many interaction terms may lead to overfitting and reduce the detection power [14]. A sensible choice is the joint modeling of longitudinal biomarker and survival outcomes [15, 16], but there are limitations in applying this model to microbiome data in observational studies. First, the cost of metagenomic sequencing and the availability of fecal samples in a multi-center study may restrict the metagenome profiling to a subgroup of participants selected by certain criteria [1–3], whose survival outcome may deviate from common statistical assumptions. Second, the classic joint modeling approach aims to address repeated measurements of biomarkers in a time-to-event analysis rather than test if a biomarker's intercept or slope is predictive of host health condition. Third, in an observational study that selects and matches participants by certain factors, their risk of developing disease is also matched. Thus, a survival submodel may not be capable of characterizing the disease risk between matched participants.

Many of the existing methods for microbiome data are built on the transformed relative abundance, such as centered log-ratio or inter-quartile log-ratio. In our new method, transformation of compositional data is not considered, since transformation strategy may have profound impact on analysis result and interpretation [17]. The compositional change in true biomarkers (e.g., causal OTUs contributing to disease onset) always leads to simultaneous change in some other OTUs' composition because of sum-to-one constraint. In an observational study with matching design, it is common to collect and profile microbiota at many time points. The sum-to-one constraint and latent noise effect may yield pseudo biomarkers with relative abundance associated with host disease status but not contributing to disease development. Hence, a taxon-level model is built for relative abundance trajectory that adjusts for the dynamic interdependence between taxa and reduces pseudo biomarker rate. In addition, we illustrate the performance of our method by a simulation pipeline that mimics the negative correlation in microbial community.

The latent technical noise in microbiome was removed by converting raw counts to relative abundance, and Zero-Inflated Beta density [9] was adopted to model an OTU's non-zero relative abundance and presence, respectively. We employed a subject-level random effect to link the logistic regression model of disease to a two-part longitudinal submodel. The latent effect of exposures related to matched set indicator was modeled by another random effect nested with the subject-level random effect. The OTU-disease association was assessed by jointly testing the scaling parameters for the subject-level random effect in the two-part submodel. We benchmarked the robustness and power of our method by a comprehensive simulation study and an application in the TEDDY cohort. The results illustrated that our method controlled the rates of false discovery and pseudo biomarkers, as well as improved the efficacy of detecting microbial trajectories signaling disease outcome.

Results

For simplicity, the aim of present research is to link the matched longitudinal microbiome samples to hosts' matched disease risk and incorporate the unknown dependence between taxa in an univariate trajectory framework, without modeling the compositionality. Briefly, we develop a Joint model with Matching and Regularization (JMR) to detect taxon-specific compositional trajectory associated with disease onset, adjusting for the linear correlation with other taxa and matched-set-specific latent noises. According to the characteristics of disease risk and infant-age gut microbiota in the TEDDY cohort, we designed a simulation pipeline similar to [8], generated the observed counts of temporal microbiota and compared our method to LMM and ZIBR using the simulated data. We also applied these methods to the shotgun metagenomic sequencing data profiled from the 4-9 months stool samples of infants enrolled in TEDDY cohort.

Overview of TEDDY microbiome study

TEDDY is an observational prospective study of children at increased genetic risk of type 1 diabetes (T1D) conducted in six clinical centers in the U.S. and Europe (Finland, Germany, and Sweden). A total of 8,676 children were enrolled from birth and followed every 3 months for blood sample collection and islet autoantibody measurement up to 4 years of age, then every 3-6 months based on autoantibody status until the age of 15 years or diabetes onset [18]. A primary disease endpoint in TEDDY is islet autoimmunity (IA), defined as persistently positive for insulin autoantibodies (IAA), glutamic acid decarboxylase autoantibodies (GADA), or insulinoma-associated-2 autoantibodies (IA-2A) at two consecutive visits confirmed by the two TEDDY laboratories [18]. The participants' monthly stool samples were collected from 3-month age until the onset of IA or censoring with random missing samples [1, 2]. Based on the sample availability and metagenomic sequencing cost, the microbiome study in TEDDY selected all the participants (cases) who developed IA by the design cutoff date May 31, 2012 and the controls at 1:1 case-control ratio matched by clinical center, gender, family history of T1D to profile the temporal gut microbiota, resulting in S = 418 matched sets (or pairs [19]). These matching factors are known risk factors for type 1 diabetes. Some of the matched sets are at higher risk of IA than the others due to higher risk human leukocyte antigen (HLA) genotypes, geography or having family history of T1D. Hence, the matched participants have comparable risk of IA, but heterogeneity still exists between them according to the case-control status by the design freeze date. The observed metagenomic counts table in TEDDY was generated by the standard procedure of DNA extraction, PCR amplification, shotgun metagenomic sequencing, assembly, annotation and quantification, as described in [1]. We visualized the top abundant species in the metagenomes of TEDDY participants who had matched IA endpoint no later than 2 years of age (Fig. 1).

Simulation

Disease outcome for the matched participants are simulated by the procedure below. The observed relative abundance per taxon were simulated by different scenarios. We first generated raw counts for a single OTU by Beta-Binomial distribution to assess the robustness and power of our method JMR without covariate taxa. We also designed a shifting procedure to mimic inherent negative correlation in the true composition of microbiota and generated the temporal high-dimensional raw counts table to evaluate the performance of compared methods.

Generate disease outcome in matched sets

We defined matched sets and subjects as 'high-risk' and 'low-risk' to generate the temporal OTU counts prior to disease onset. Subjects are matched at 1:1 ratio. For participant $j = 1, 2$ in matched set s ($s = 1, \ldots, S$), we first generated subject-level and set-level random effects from a standard Normal distribution $a_{s_j} \sim N(0, 1)$, $b_s \sim N(0, 1)$. Each random effect was converted to a binary variable by the median value. That is $A_{s_j} = \boldsymbol{I}(a_{s_j} > \text{median}(a_{s_j}))$, $B_s = \boldsymbol{I}(b_s > \text{median}(b_s))$, where $A_{s_j} = 1$ (or $B_s = 1$) represents a 'high-risk' subject (or set). Next, we simulated a host genotype G_{s_j} as disease risk factor, and the host disease status by a Bernoulli distribution $O_{s_j} \sim B(p_{s_j})$, where $\text{logit}(p_{s_j}) = \alpha_0 + \alpha_1 G_{s_j} + \alpha_2 A_{s_j} + \alpha_3 B_s$. We fixed $(\alpha_0, \alpha_1, \alpha_3) = (0.5, -2, 1)$, which is the JMR estimate from real data, and set $\alpha_2 \in \{0.5, 0.75, 1, 1.25, 1.5\}$ to generate different datasets.

Scenario A: single OTU counts.

We first simulated the observed counts of a single OTU by Beta-Binomial [14] distribution to compare the univariate trajectory methods without adjusting for covariate taxa. The true relative abundance of an OTU at the earliest time point $t = 1$ was drawn from a Beta distribution $\mu_1 \sim Beta(\mu_0, \phi_0)$, where parameters μ_0, ϕ_0 were estimated by applying Beta-Binomial MLE to the metagenomic raw counts of an OTU selected at a given relative abundance level in the TEDDY data. To simplify the age-dependent effect, the relative abundance of this OTU at later time points $t > 1$ was generated by linearly increasing μ_1 to μ_t. The baseline relative abundance at time t in a matched set s was generated by $\mu_{st} \sim Beta(\mu, \phi_t)$, and was increased or decreased by $\Delta\mu_{st}$ if the set was labeled as 'high-risk'. The true relative abundance of this OTU for subject j in set s at time point t was simulated by $\mu_{s_j t} \sim Beta(\mu_{st}, \phi_{st})$, and was increased or decreased by $\Delta\mu_{s_j t}$ if the subject was 'high-risk'. The total counts

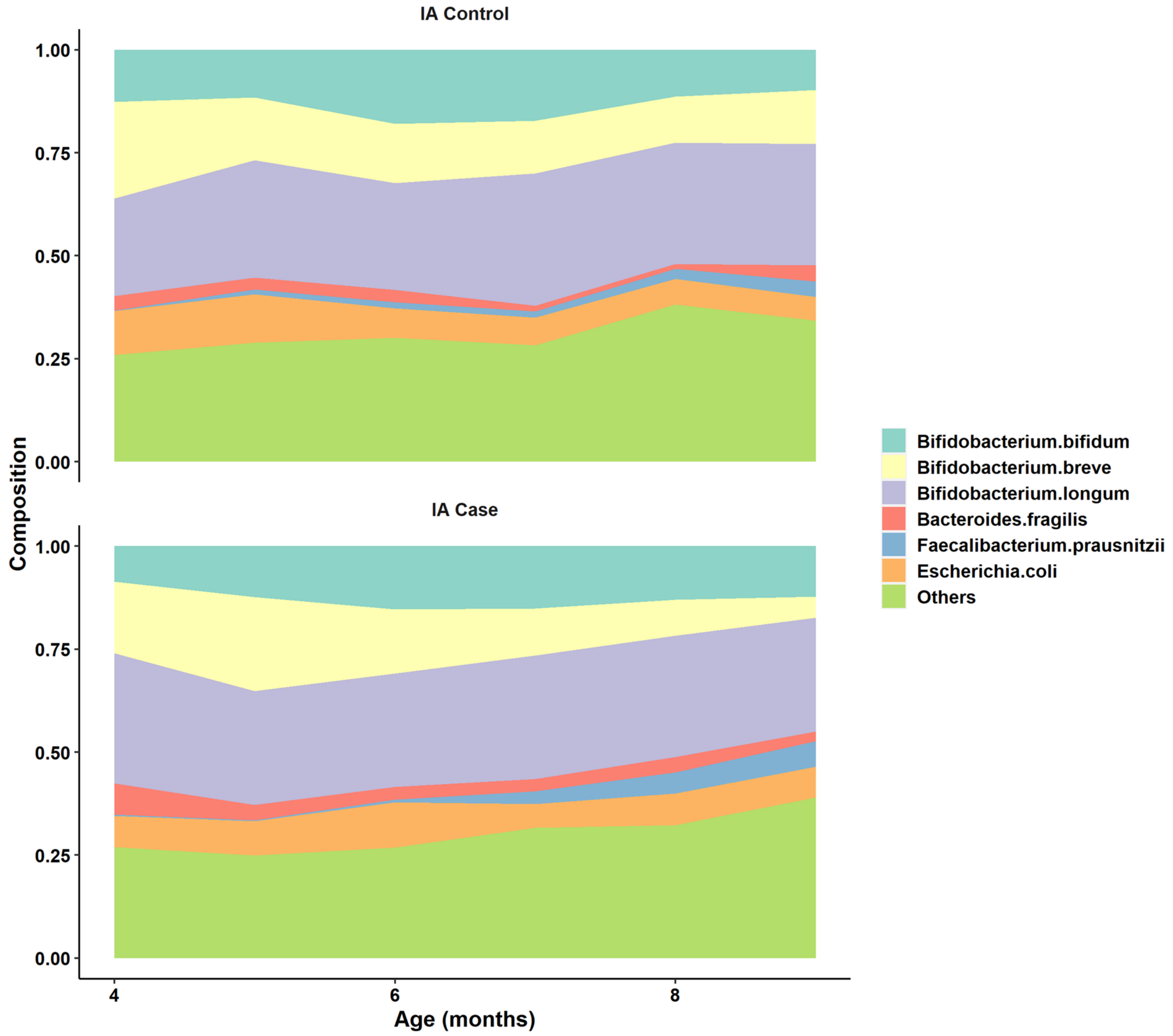

Fig. 1 Mean compositional trajectory of top abundant species in infant-age metagenomes grouped by host islet autoimmunity status at 2 years of age

per sample, i.e., library size was drawn from a Poisson distribution $N_{s_jt} \sim PS(100000)$, and the counts for this OTU is generated from a Binomial (BN) distribution $Y_{s_jt} \sim BN(N_{s_jt}, \mu_{s_jt})$.

Scenario B1: counts table with random intercept and pseudo biomarkers

We also generated a high-dimensional counts table with $P = 1030$ OTUs to demonstrate the performance of each method, so that the covariate taxa can be used in JMR. The true composition of each microbiome sample $\bar{\eta}_{s_jt}$ was simulated by a shifting procedure combined with Dirichlet distribution to account for the negative correlation within microbial community. The sample-wise library size was generated by a Poisson distribution, and the observed raw counts were sampled from a Multinomial distribution. Details of data generation process for this scenario is available in Methods, with a visualization for dimension of $P = 4$ in Fig. 2.

For a subject labeled as 'high-risk', we increased 15% OTUs (denoted by M^+) in $\bar{\eta}_{s_jt}$ by Δ_{s_jt}, and reduced another 15% OTUs (denoted by M^-) by $d\Delta_{s_jt}$ $(0 < d < 1)$. The subsets M^+ and M^- are the true biomarkers for disease status. We randomly selected a third subset (denoted by M^0) from the remaining 70% OTUs in $\bar{\eta}_{s_jt}$ and reduced the composition of M^0 by a total of $(1-d)\Delta_{s_jt}$. There may exist OTUs never selected in M^+, M^-, or M^0, which are the 'null' OTUs.

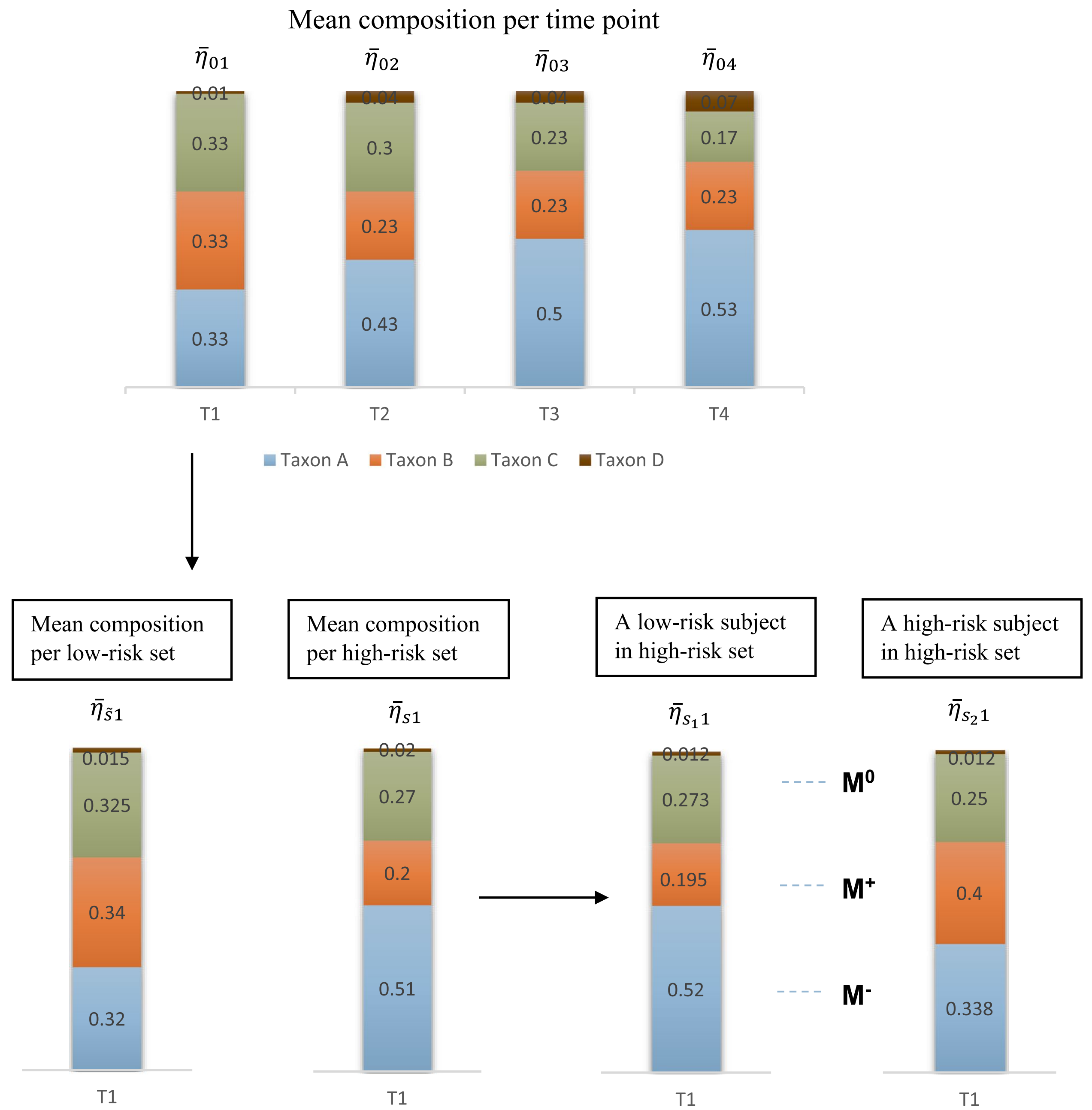

Fig. 2 Shifting procedure in simulation scenario B1. From time T1 to T4, taxa A-C decrease and taxon D increases in relative abundance. Compared to a low-risk set, taxa A, D are more abundant and taxa B, C are less abundant in a high-risk set. Taxon A is less abundant (i.e., M^{-}) and taxon B is more abundant (i.e., M^{+}) in a high-risk subject compared to the matched low-risk subject, both being true biomarkers at each time point. Taxon C is a pseudo biomarker (randomly selected as M^{0} at time point T1) with relative abundance automatically changed due to sum-to-one constraint, while taxon D is unchanged at T1

The OTUs selected in M^0 are the pseudo biomarkers due to random shift in frequency. We set the total shift $\Delta_{s_jt} = \lambda\Delta^0_{s_jt}$ at distinct effect size $\lambda \in \{0.5, 0.6, 0.7, 0.8\}$, where $\Delta^0_{s_jt}$ is the maximum shift restricted by sum-to-one.

Scenario B2: counts table with random slope and pseudo biomarkers

Data generation process for this scenario is similar to Scenario B1, except that the shift (Δ_{s_jt}) in microbiota true composition between 'low-risk' and 'high-risk' subjects varies by time points. It's worth to note that we cannot distinguish 'false positive' from 'pseudo positive' in scenarios B1 and B2. Hence, we use the sum of false positive rate and pseudo positive rate, i.e., false or pseudo positive rate (FPPR) as a performance metric for scenarios B1,B2. That is FPPR $= \frac{\#\text{ of positives in }(M^{+} \cup M^{-})^{c}}{\#\text{ of OTUs in }(M^{+} \cup M^{-})^{c}}$.

Scenario C: counts table without pseudo biomarkers

In this scenario we considered random intercept signaling the disease onset and fixed half OTUs in $\bar{\eta}_{s_j t}$ as 'null' in order to evaluate the FPR and FDR of each method, although this scenario is not applicable to real data. Among the other half OTUs, we selected 10% OTUs in $\bar{\eta}_{s_j t}$ as M^{+} and 40% OTUs as M^{-} without a subset of pseudo biomarkers (M^{0}).

Performance of competing methods

In scenario A, we compared JMR not adjusting for correlated taxa (JMR-NC) with the following methods: a) a joint model with regularization but without matching indicator and correlated taxa (JR-NC); b) the ZIBR model with a Wald statistic jointly testing OTU-specific abundance or presence using either a single random effect (ZIBR-S) or nested random effects (ZIBR-N); c) LMM with arcsin-square-root transformation using either a single random effect (LMM-S) or nested random effects (LMM-N). For LMM and ZIBR methods, we used R package gamlss and set the sample age, genotype, disease status, and genotype-disease interaction term as the fixed effect covariates. It's worth to note that the nested random effects used in LMM and ZIBR are independent of host disease risk, different from those in JMR.

We randomly selected 6 OTUs with different relative abundance from TEDDY data and estimated the baseline parameters for each. These OTUs are *Acinetobacter sp. NIPH 236, Brachyspira murdochii, Streptococcus phage YMC-2011, Erysipelatoclostridium ramosum, Ruminococcus gnavus*. Then we generated $n = 10000$ replicates for each OTU with $S \in \{50, 100\}$. The type I error rate and power of each method was calculated at statistical significance level $p < 0.05$, shown in Table 1. The results showed that JMR-NC persistently controlled the type I error and provided higher detection power at distinct abundance levels except for the OTUs with $-\log_{10}(y) \in (2, 3]$ and $(5, 6]$. Type I error of the reduced model JR-NC was severely inflated in some datasets and its power was lower than JMR-NC. LMM consistently controlled type I error, with power lower than JMR-NC in most simulated OTUs. The ZIBR method yielded inflated type I error rate and low efficacy regardless of sample size in this single-OTU scenario.

In scenarios B1, B2 and C, we generated 10 replicates for each OTU table to assess the performance of competing methods. We evaluated the performance of each method at different size of set-level random effect $\gamma \in \{0.6, 0.7, 0.8\}$. The taxa associated with disease onset in each OTU table are detected by FDR cutoff $q < 0.15$. The FPPR in scenario B1 (Fig. 3) showed that adjusting for the top-correlated taxa in JMR successfully controlled the rate of pseudo biomarkers across different scenarios and was more powerful than LMM at larger sample size ($N = 200$), although JMR showed lower detection power compared to JMR-NC. The results in Fig. 4 also demonstrated the outperformance of JMR-NC, JMR, and ZIBR in the sensitivity of detecting taxon-specific trajectory heralding disease outcome. The power of ZIBR in either intercept or slope analysis was higher than the competing methods regardless of sample size in the high-dimensional scenarios, but this model yielded inflated FPPR (Fig. 3). The LMM methods were powerful in the test of intercept, but this model occasionally produced inflated FPPR (Fig. 3) regardless of the set-level random effect size. Furthermore, the power of LMM was unstable in intercept analysis, while its power in slope analysis was nearly zero. To confirm the impact of γ on performance, we also compared the metrics in Figs. 3 and 5 between different values of γ, using Kruskal-Wallis test. A larger set-level random effect led to significant change in FPPR, FPR, FDR for LMM and ZIBR methods ($p < 10^{-5}$), but this impact was trivial in JMR or JMR-NC ($p > 0.1$). JMR showed the best performance in slope analysis, with higher sensitivity (Fig. 4) and the lowest FPPR (Fig. 3). Results of scenario C in Fig. 5 showed that JMR and LMM effectively controlled the FPR, while LMM produced higher FPR at larger matched-set-specific random effect (γ). The FDR of JMR at $\gamma = 0.6$ was relatively higher than that of LMM due to lower sensitivity. The inflated FPR and FDR of ZIBR in scenario C (Fig. 5) is consistent with the FPPR in scenario B (Fig. 3).

For each raw counts table in scenarios B and C, more than half of simulated OTUs have the observed zero-inflation probability (i.e., 1−prevalence) between 2% and 90%, although there are a few OTUs with the observed prevalence at 100%. The overall prevalence of each OTU table in scenarios B and C is similar between different datasets, which cannot be specified in the Dirichlet-Multinomial distribution or the shifting procedure. Hence, we assess the impact of zero-inflation on performance only in scenario A, whereas the prevalence is related to taxon-specific relative abundance. We also visualized the prevalence of six OTUs generated in scenario A (Fig. 6). The power of JMR-NC (Table 1) was better for the prevalence at a medium level, i.e., replicates with $-\log_{10}(y)$

Table 1 The type I error and power based on 10000 simulated replicates for a taxon at different levels of mean relative abundance ($\boldsymbol{y}$)

$-\log_{10}(\boldsymbol{y})$		**(0, 1]**	**(1, 2]**	**(2, 3]**	**(3, 4]**	**(4, 5]**	**(5, 6]**
Type I error							
	JMR-NC	0.01	0.003	0.008	0.0004	0	0
	JR-NC	0.061	0.03	0.001	0.002	0.0003	0.0005
$N = 100$	LMM-N	0.045	0.031	0.026	0.033	0.068	0.052
$(S = 50)$	LMM-S	0.041	0.036	0.023	0.036	0.066	0.049
	ZIBR-N	0.057	0.066	0.080	0.053	0.164	0.369
	ZIBR-S	0.049	0.079	0.076	0.063	0.163	0.366
	JMR-NC	0.018	0.004	0.022	0.051	0.001	0.0003
	JR-NC	0.162	0.091	0.049	0.066	0.035	0.002
$N = 200$	LMM-N	0.043	0.039	0.068	0.057	0.079	0.058
$(S = 100)$	LMM-S	0.046	0.051	0.066	0.060	0.079	0.056
	ZIBR-N	0.056	0.067	0.099	0.078	0.170	0.324
	ZIBR-S	0.059	0.088	0.099	0.088	0.174	0.324
Power							
	JMR-NC	0.399	0.5	0.372	0.833	0.798	0.136
	JR-NC	0.32	0.65	0.051	0.057	0.548	0.040
$N = 100$	LMM-N	0.040	0.084	0.474	0.107	0.552	0.512
$(S = 50)$	LMM-S	0.043	0.090	0.468	0.108	0.347	0.468
	ZIBR-N	0.06	0.088	0.412	0.123	0.743	0.666
	ZIBR-S	0.057	0.095	0.405	0.123	0.686	0.664
	JMR-NC	0.452	0.723	0.619	0.944	0.917	0.388
	JR-NC	0.677	0.699	0.194	0.836	0.737	0.324
$N = 200$	LMM-N	0.06	0.115	0.429	0.478	0.287	0.253
$(S = 100)$	LMM-S	0.068	0.134	0.423	0.321	0.273	0.206
	ZIBR-N	0.057	0.332	0.433	0.667	0.280	0.379
	ZIBR-S	0.063	0.344	0.430	0.648	0.265	0.377

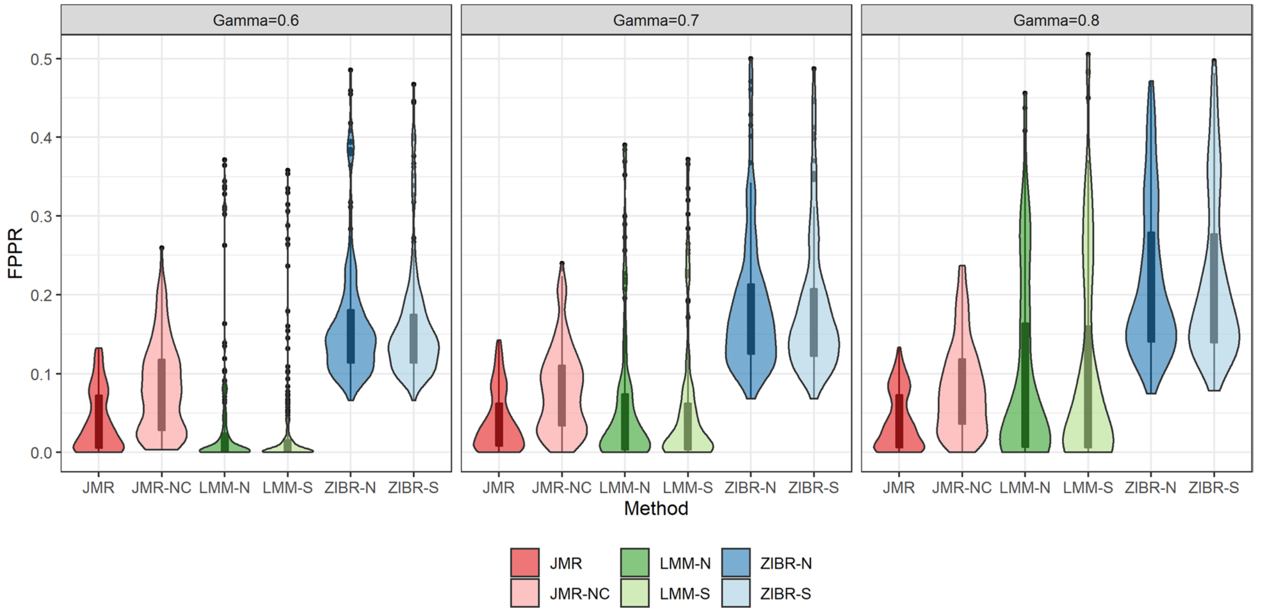

Fig. 3 FPPR of each method for scenario B1 at different size of matched-set-specific random effect

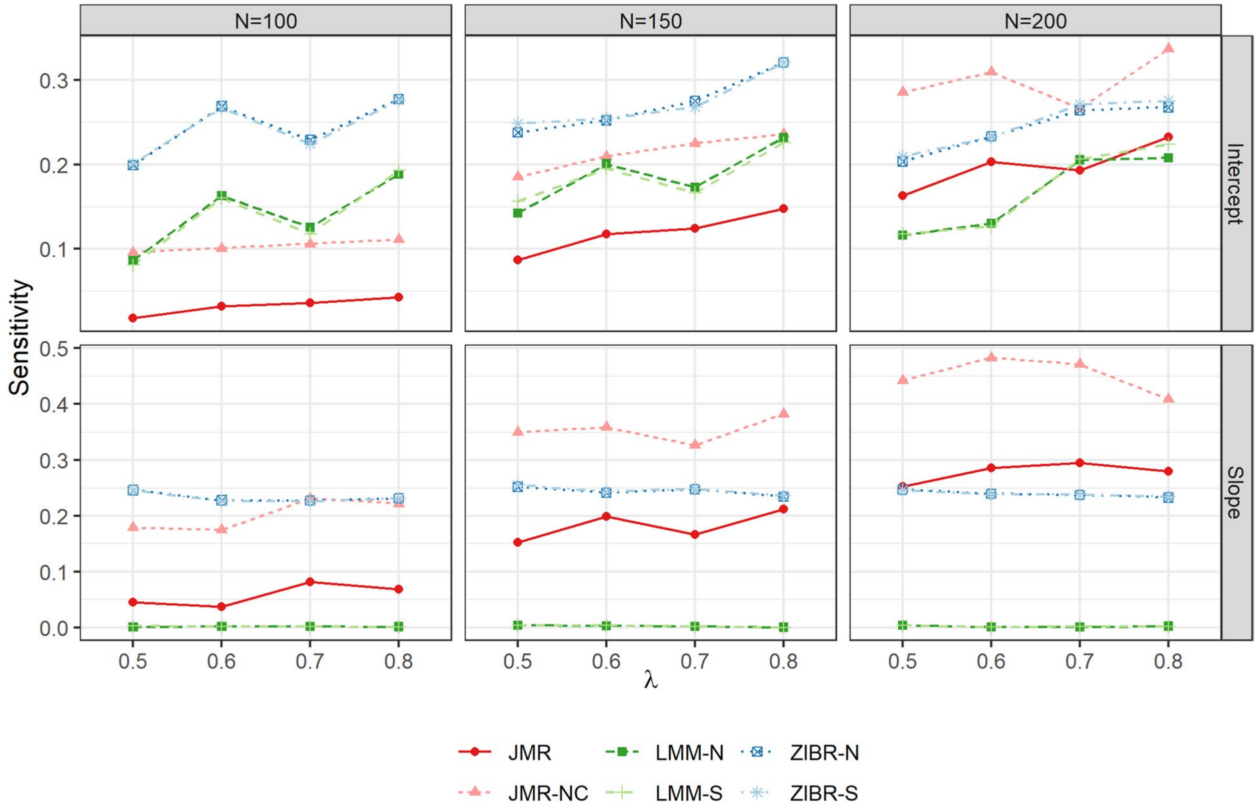

Fig. 4 Sensitivity of each method for scenarios B1 and B2 at different sample size (*N*) and different effect size (*λ*)

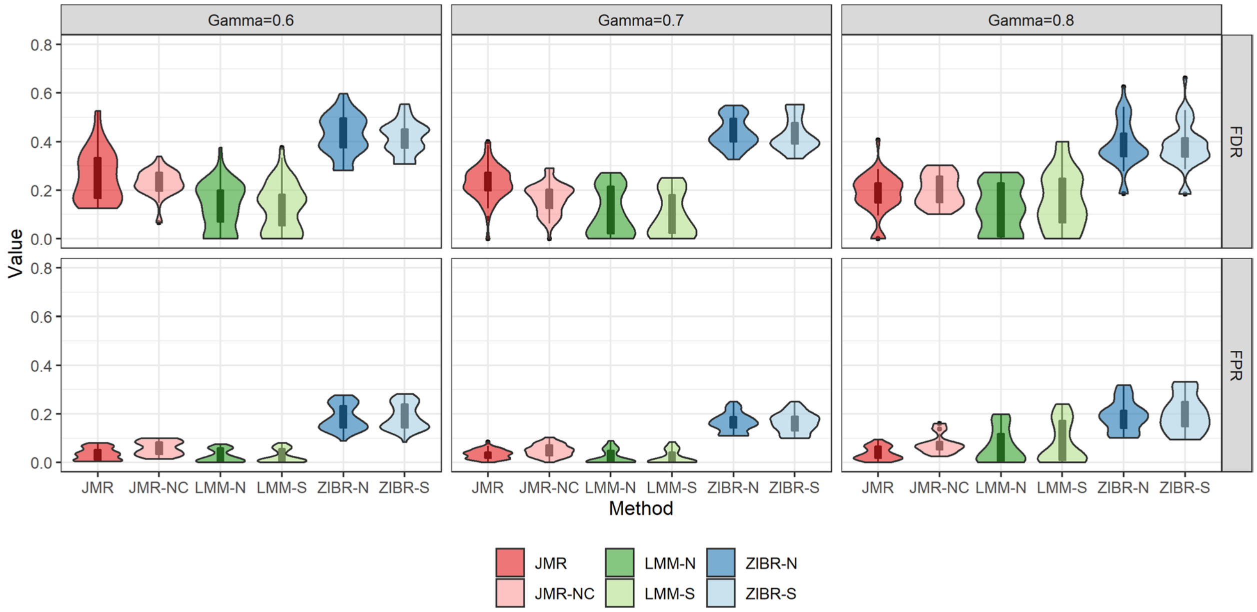

Fig. 5 FDR and FPR for each method in scenario C at different size of matched-set-specific random effect

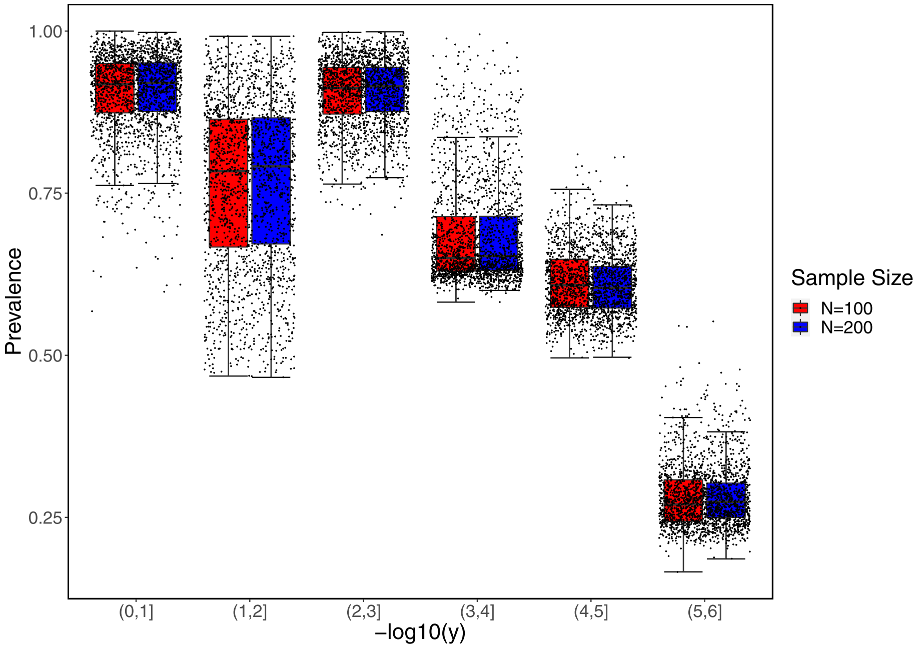

Fig. 6 Prevalence distribution for each OTU in scenario A

between (3, 4] or (4, 5] (Fig. 6). The OTUs with higher abundance and relatively lower prevalence (i.e., replicates with $-\log_{10}(y) \in (1, 2]$ in Fig. 6) showed better efficacy in Table 1. In general, OTU-specific prevalence being too high or too low may reduce the power of JMR.

Application in TEDDY

We applied the competing methods to the longitudinal metagenomes profiled from TEDDY children's monthly stool samples collected at the age of 4-9 months [1]. We included the cases developing IA between 9-month and 24-month age and their matched controls who remained IA-negative by the cases' diagnosis age. For each matched pair included in the present analysis, one participant was IA positive and the other one was negative at the age of 24 months. We excluded the participant(s) matched to multiple pairs, yielding $N = 152$ subjects ($S = 76$ pairs) and $n = 672$ metagenome samples. The cases who experienced IA onset after 24 months and their matched controls were not included in this analysis.

We first filtered OTUs at genus and species level by relative abundance $> 10^{-6}$ and prevalence $> 5\%$, selecting 125 out of 265 genera and 365 out of 750 species in downstream analysis. It's worth to note that there are 1797 species in total profiled and quantified in TEDDY cohort, with 750 species detected between 3- and 9-month age. The sample age and the hosts' breastfeeding status per time point were used as longitudinal covariates, while HLA DR3 &4 haplotype was included as time-invariant covariates. For the LMM and ZIBR methods, we used the interaction term between IA status and the binary HLA category (DR3 &4 vs. others) as a covariate to adjust for the association. We tested each OTU's association with IA by FDR cutoff $q < 0.05$ or $q < 0.1$, individually. The HLA DR3 &4 genotype was confirmed positively and significantly ($p < 0.05$ by Wald test) associated with IA in JMR. The results in Table 2 showed that JMR identified more OTUs than LMM in both intercept and slope analysis. The LMM methods only found a small subgroup of taxa associated with IA at either genus or species level. We also visualized the overlap and difference between JMR, JMR-NC, LMM-N selected by $q < 0.1$ in Fig. 7 with OTU names listed in Supplementary Table S1, and then compared Akaike Information Criterion (AIC) of JMR and JMR-NC for the 76 species detected by both methods. Adjusting for the correlated taxa in JMR did improve model fitting with lower mean AIC (-2631.847) compared to JMR-NC (-2615.571). LMM-N is not comparable to JMR or JMR-NC in terms of information criteria, since the taxon-specific relative abundance was transformed by asin-square-root.

The taxa with mean abundance (intercept) associated with IA onset exclusively detected by both JMR and JMR-NC at $q < 0.1$ include *Bifidobacterium breve*, *Bacteroides fragilis*, *Lactobacillus ruminis*, *Veillonella ratti*. *B.breve*, as one of the three species dominating infant-age gut microbiota in TEDDY, was less abundant in intercept (i.e. at 4- and 9-month) during infancy among

Table 2 The number of genera and species associated with IA detected by each method in a subgroup of TEDDY participants

		Intercept		Slope	
FDR		$q < 0.05$	$q < 0.1$	$q < 0.05$	$q < 0.1$
Genera	JMR	27	31	27	34
	JMR-NC	44	52	36	44
	LMM-N	11	16	3	4
	LMM-S	49	82	10	19
	ZIBR-N	26	30	31	36
	ZIBR-S	26	32	31	36
Species	JMR	75	94	37	46
	JMR-NC	147	166	112	125
	LMM-N	43	60	13	21
	LMM-S	40	63	7	16
	ZIBR-N	89	106	119	138
	ZIBR-S	83	105	120	140

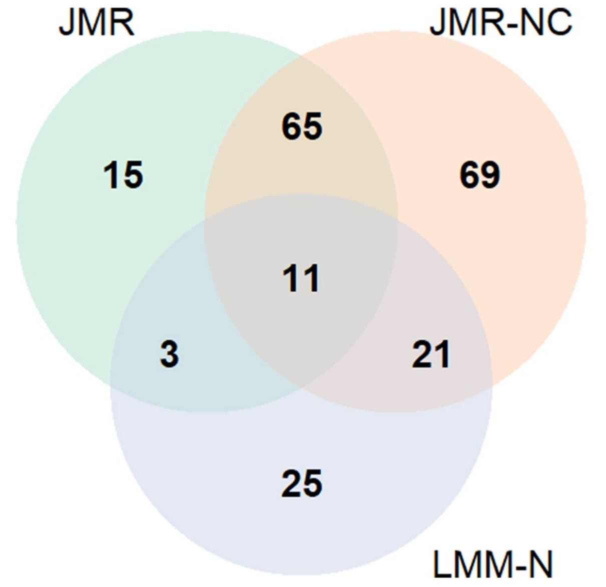

Fig. 7 Venn diagram for the intercept analysis in TEDDY data by JMR, JMR-NC, LMM-N

IA cases, with density shown in Fig. 8. The species *B.fragilis* as part of the normal microbiota in human colon was found more abundant among IA cases compared to their matched controls (Fig. 1). This *Bacteroides* species was also found differential between T1D cases and controls at only one time point in a small-size Finnish cohort [20].

There are two more abundant species *Faecalibacterium prausnitzii* and *Escherichia coli* visualized in Fig. 1 associated with IA in slope and exclusively detected by JMR. *F.prausnitzii*, as one of the most abundant and important commensal bacteria of human gut microbiota that produces butyrate and short-chain fatty acids from the fermentation of dietary fiber, increased faster in IA cases after 6-month of age. This rapid change and abnormally higher level of *F.prausnitzii* prior to IA seroconversion may be a result of the sudden change of dietary pattern during infancy.

Our method successfully detected the case-control difference in the slope of *E.coli*, which was found as an amyloid-producing bacteria with temperal dynamics heralding IA onset in a subset analysis in DIABIMMUNE cohort [21]. The relative abundance of *E.coli* in TEDDY smoothly decreased from 4-month to 9-month for both cases and controls (Fig. 1), and it was relatively more abundant in controls between 7- and 9-month with stratified densities shown in Fig. 8. The temporal change of *E.coli* prior to IA seroconversion in TEDDY detected by JMR was consistent with the decrease of *E.coli* reported in DIABIMMUNE cohort [21], which was possibly due to prophage activation according to the *E.coli* phage/*E.coli* ratio prior to *E.coli* depletion in that research.

Discussion

We developed a joint model with nested random effects to test the association between taxa and disease risk, and adjusted for the correlated taxa screened by a preselection procedure in abundance and prevalence, individually. We implemented our method in an R package mtradeR (metagenomic trajectory analysis with disease endpoint and risk factors) with illustration examples at https://github.com/qianli10000/mtradeR. The JMR function implemented the framework in equation (1) by parallel computing. We also provided simulation functions StatSim and TaxaSim to generate (binary) disease status and temporal high-dimensional metagenomic counts of matched sets. The runtime of each method for different sample size and different number of OTUs were compared on an 8-core computer, with mean and standard deviation shown in Table 3. The nested random effects were utilized in each method. For the univariate models without covariate taxa, LMM-N is the fastest algorithm and ZIBR-N is the slowest, both implemented in gamlss R package. Although the adjustment of correlated taxa in JMR requires additional computation, the runtime of JMR is still shorter than ZIBR-N in gamlss.

The simulation of single OTU demonstrated the performance of each method at different relative abundance levels, implying that LMM with either single or nested random effect is still a robust method. The simulation of high-dimensional OTU tables also illustrated LMM's overall performance in the test of intercept, but the unstableness of LMM is a concern in real data analysis.

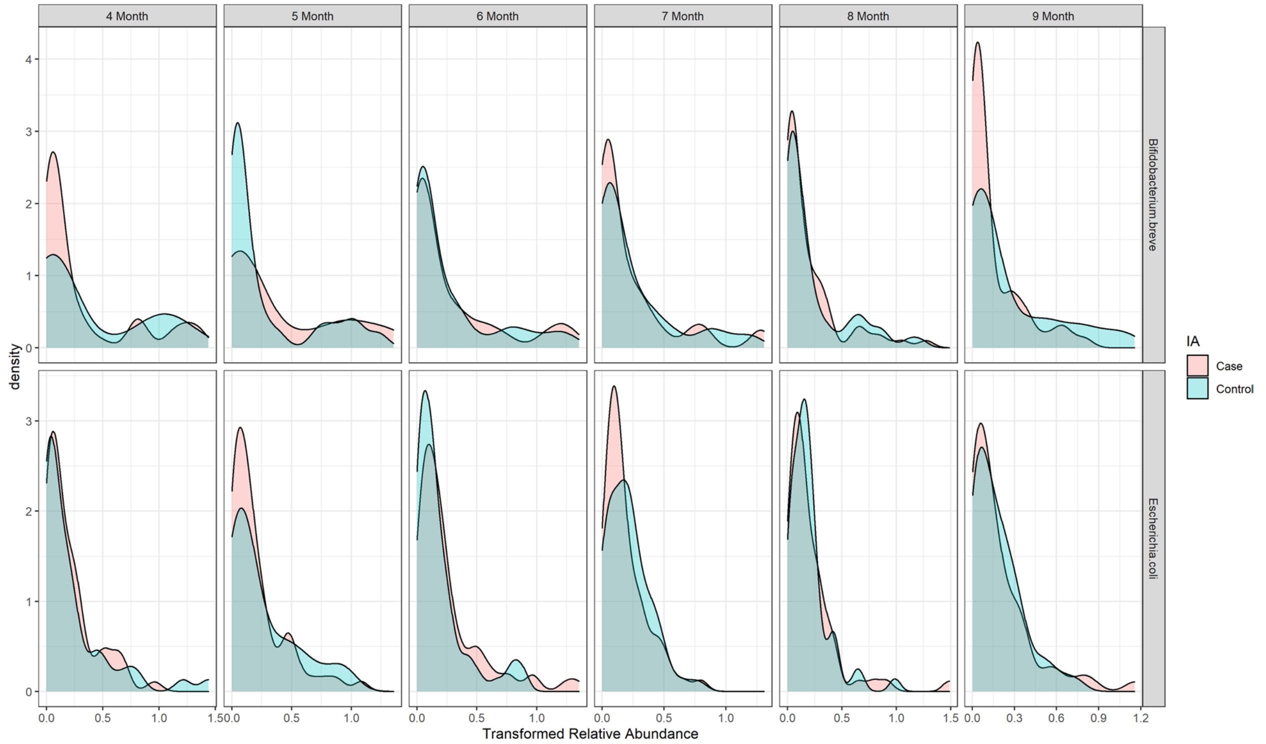

Fig. 8 Distribution of relative abundance for *B.breve* and *E.coli* per time point grouped by 2-year IA status

Table 3 The mean and standard deviation (SD) of runtime in minutes for 30 repeated runs by each method at different number of longitudinal samples (n) and filtered OTUs ($\tilde{P}$) in TEDDY data. The OTUs in each dataset are filtered by either relative abundance $> 10^{-6}$, prevalence $> 5\%$ or relative abundance $> 10^{-5}$, prevalence $> 10\%$

		Mean Runtime (SD)			
Dataset scale		JMR-NC	ZIBR-N	LMM-N	JMR
$n = 153$	$\tilde{P} = 359$	11.62 (0.95)	44.11 (4.72)	3.01 (0.36)	35.95 (3.6)
$n = 153$	$\tilde{P} = 234$	8.02 (0.26)	20.79 (0.31)	1.93 (0.04)	24.58 (0.52)
$n = 307$	$\tilde{P} = 370$	14.16 (0.65)	56.01 (4.78)	5 (0.45)	43.67 (3.09)
$n = 307$	$\tilde{P} = 247$	9.69 (0.63)	37.89 (2.31)	3.9 (0.24)	26.97 (0.68)

JMR yielded lower false or pseudo positive rate in the simulated datasets and higher detection power in slope analysis by adjusting for the top-correlated taxa. The preselection of top-correlated taxa in JMR was performed in relative abundance and presence, individually, being consistent with the two-part model strategy. According to the simulation study, a disadvantage of JMR is the limited power at small sample size and the dependence on tuning parameter. The prescreening procedure in JMR may occasionally select a true biomarker as covariate taxon, which is possibly confounding with the subject-level random effect. Hence, the adjustment of related taxa in JMR reduced the detection power compared to JMR-NC, although this strategy controlled the pseudo biomarker rate. Adding nodes in the GH approximation may improve the power of JMR, but more nodes will also lead to additional computation costs. Hence, future work should focus on improvement of JMR in both detection power and computation efficiency. Furthermore, the simulation results in Fig. 3 also suggested the minimum number of participants or matched pairs required based on set-level or subject-level random effect size. In an observational study with strong set-level noises in the microbiota (e.g., multi-center effect), a minimum sample size of $N = 200$ participants (i.e., $S = 100$ pairs) coupled with JMR can improve the detection power and control FPPR at each level of disease-associated random effect.

Another limitation of our method is the potential bias in scaling parameter (λ_r, λ_p) estimation, possibly caused by the L_2 regularization. Our current work only focused on the unsigned association between a taxon and host disease status by using a Wald statistic. An improvement in the estimate of scaling parameter and statistical inference should be considered in future work, such as the algorithm in ZINQ [14]. We did not use quantile regression in current research, since the performance of ZINQ required tuning of grid. But ZINQ provided an alternative approach for modeling zero-inflation in microbiota composition with fewer statistical assumptions.

The right-censoring of longitudinal biomarker measurements or a binary disease outcome always occurs in observational studies. Our model allows random missingness or censoring of microbiome samples at any time point. In an observational study like TEDDY, the controls' disease outcome was censored at or later than the matched cases' endpoint, because the case-control matching was based on the participants' disease status. Thus, right-censoring is not applicable to the disease status at matched endpoint. For a study matching participants solely based on confounding risk factors (e.g., DIABIMMUNE), the right-censoring of disease outcome should be addressed prior to the usage of JMR, such as multiple imputation. There are other important topics to be considered in the modeling of longitudinal microbiome data. One potential direction is high dimensional modeling framework, such as tensor singular value decomposition [22]. A promising extension of the current work in JMR is to exploit functional data analysis for multiple microbial trajectories. By employing a non-parametric joint modeling, we may be able to capture nonlinear trends and heterogeneous patterns of longitudinal biomarkers in microbiota, as well as negative correlations among taxa [23].

Conclusions

The proposed framework JMR successfully controlled the false or pseudo biomarkers in taxon-specific trajectory analysis with improved detection power by incorporating the matching of participants and adjusting for the dependence between taxa.

Methods

Joint model with matching and regularization

The probability for participant j ($j = 1, \ldots, J$) in matched set s ($s = 1, \ldots, S$) developing the disease of interest is $p_{s_j} = P(O_{s_j} = 1)$, where O_{s_j} is the binary disease status. There are J participants in each matched set. Let y_{s_jt} be the relative abundance of an OTU for participant j in matched set s at time point t ($t = 1, \ldots, T_{s_j}$). We denote the expected non-zero abundance by $\mu_{s_jt} = E(y_{s_jt}|y_{s_jt} > 0)$, and the probability of presence (or zero-inflation) by $\pi_{s_jt} = P(y_{s_jt} > 0)$, similar to [9]. For a microbiome study matching participants by the disease-associated factors and/or disease status (e.g., DIABIMMUNE, TEDDY), the matched participants are assumed to have comparable but distinct disease risk. Hence, we model the disease status by a logistic mixed effect model with nested random effects. A joint model for the host disease status and microbial trajectory in matched set is

$$\begin{aligned} \operatorname{logit}(p_{s_j}) &= \boldsymbol{u}_{s_j}\boldsymbol{\alpha} + a_{s_j} + b_s \\ \operatorname{logit}(\mu_{s_jt}) &= \boldsymbol{x}^{(1)}_{s_jt}\boldsymbol{\beta}_{11} + \boldsymbol{z}_{s_jt}\boldsymbol{\beta}_{12} + \tilde{z}_{s_jt}(\lambda_r a_{s_j} + \gamma_r b_s) \\ \operatorname{logit}(\pi_{s_jt}) &= \boldsymbol{x}^{(2)}_{s_jt}\boldsymbol{\beta}_{21} + \boldsymbol{z}_{s_jt}\boldsymbol{\beta}_{22} + \tilde{z}_{s_jt}(\lambda_p a_{s_j} + \gamma_p b_s) \end{aligned} \tag{1}$$

The host disease status is determined by a vector of fixed effect covariates $\boldsymbol{u}_{s_j}$ and the independent nested random effects a_{s_j}, b_s. The non-zero relative abundance μ_{s_jt} and presence π_{s_jt} per OTU are predicted by the same random effects rescaled by parameters λ_r, λ_p and a vector of clinical or bioinformatics technical covariates $\boldsymbol{z}_{s_jt}$. To model the unknown correlation between taxa, this OTU's non-zero abundance and presence per time point also depend on the other taxa with relative abundance $\boldsymbol{x}^{(1)}_{s_jt}$ and presence-absence $\boldsymbol{x}^{(2)}_{s_jt}$ measured at the same time point, preselected by a procedure described below. The two-part submodel of y_{s_jt} characterizes how the trajectory is affected by subject- and set-level latent factors contributing to disease risk, and how the OTU trajectory interacts with correlated taxa over time. If an OTU is a pseudo biomarker, then its relative abundance (y_{s_jt}) should be driven by the top-correlated taxa per time point instead of the disease-associated random effect a_{s_j}. On the other hand, the abundance of a true biomarker OTU at each time point is mainly determined by the latent risk of disease onset (a_{s_j}, b_s) and possibly associated with the top-correlated taxa.

We set $\tilde{z}_{s_jt} = 1$ in equation (1) to test intercept, and $\tilde{z}_{s_jt} = \text{age}$ to test slope. The nested random effects and parameters λ_r, λ_p provide flexibility in the modeling of between-subjects and between-sets heterogeneity, as well as model the abundance-presence correlation in each taxon by shared nested random effects instead of assuming independence between the two processes as in [9].

Parameter estimation and hypothesis testing

To account for the sum-to-one restriction on non-zero relative abundance ($0 < \mu_{s_jt} < 1$) and the binarized measurement $\boldsymbol{I}(y_{s_jt} > 0)$ of an OTU, we intuitively employ the Zero-Inflated Beta (ZIB) density function [9] to define the match-set-specific marginal likelihood for parameter estimation. That is $L(\theta; \boldsymbol{y}, \boldsymbol{O}) = \prod_{s=1}^{S} L_s$, where

$$L_s = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g_b(b_s) \prod_{j=1}^{J} l^{(1)}_{s_j}(a_{s_j}, b_s) l^{(2)}_{s_j}(a_{s_j}, b_s) g_a(a_{s_j}) \boldsymbol{da}_{\boldsymbol{s}_1} \ldots \boldsymbol{da}_{\boldsymbol{s}_J} \boldsymbol{db}_{\boldsymbol{s}} \tag{2}$$

$$\begin{aligned} l^{(1)}_{s_j}(a_{s_j}, b_s) &= \prod_{t=1}^{T_{s_j}} [(1 - \pi_{s_jt})\boldsymbol{I}(y_{s_jt} = 0) + \pi_{s_jt}\boldsymbol{I}(y_{s_jt} > 0) f(y_{s_jt}|y_{s_jt} > 0)] \\ l^{(2)}_{s_j}(a_{s_j}, b_s) &= p_{s_j}\boldsymbol{I}(O_{s_j} = 1) + (1 - p_{s_j})\boldsymbol{I}(O_{s_j} = 0) \end{aligned} \tag{3}$$

$g_a(a_{s_j}), g_b(b_s)$ are the Gaussian density functions with mean 0 and variance σ_a^2, σ_b^2, individually, and $f(y_{s_jt}|y_{s_jt} > 0)$ is the Beta density function with mean μ_{s_jt} and overdispersion ϕ. In the simulation study, we demonstrated that the robustness and performance of this model does not require the observed relative abundance being generated from ZIB distribution.

The estimate of overdispersion $\hat{\phi}$ without regularization is severely inflated and also leads to bias in the estimate of other parameters. Hence, we use L_2 (ridge) regularization to control the overdispersion and type I error in hypothesis testing. All the parameters θ are estimated by maximizing a penalized marginal likelihood function $\hat{\theta} = \arg max \tilde{L}(\theta; \boldsymbol{y}, \boldsymbol{O})$, where

$$\tilde{L}(\theta; \boldsymbol{y}, \boldsymbol{O}) = \ln L(\theta; \boldsymbol{y}, \boldsymbol{O}) - \rho ||\theta||_2^2 \tag{4}$$

and ρ is selected by a cross-validation described below.

There is no closed form of the multivariate integral L_s in equation (2) because of the Beta density in $l_{s_j}^{(1)}(a_{s_j}, b_s)$. Hence, L_s can be approximated by Gauss-Hermite (GH) quadrature, with details explained in Appendix. We test the association between OTU trajectory and host disease status with null hypothesis $H_0 : \lambda_r = \lambda_p = 0$ and a Wald statistic $W = \frac{\hat{\lambda}_r^2}{SE_{\lambda_r}^2} + \frac{\hat{\lambda}_p^2}{SE_{\lambda_p}^2}$, which follows a Chi-Square distribution $W \sim \chi^2(2)$. The false discovery rate (FDR) for multiple testing is corrected by the Benjamini-Hochberg (BH) procedure.

Pre-selection of correlated taxa and tuning parameter selection

For each OTU (y_{s_jt}) in equation (1), using all the other taxa as covariates is computationally inefficient. Hence, we use a data-driven procedure to pre-select $\boldsymbol{x}_{s_jt}^{(1)}$ and $\boldsymbol{x}_{s_jt}^{(2)}$, and then perform a post-selection hypothesis testing. The first step screens the taxa correlated with y_{s_jt} in abundance and presence, individually, using the Bray-Curtis distance less than 0.1 quantile. Our current method uses relative abundance in both pre-selection and modeling, since this method is developed for large-scale microbiome studies and the multi-center technical batch effect can be simply normalized by relative abundance. According to the comparison of dissimilarity metrics on microbiome compositional data [24], we choose Bray-Curtis dissimilarity to pre-select the related taxa. This step may still result in many covariate taxa at species level in metageonmic data due to high dimensionality. Thus, we employ elastic net regression to further select the taxa with relative abundance $\boldsymbol{x}_{s_jt}^{(1)}$ associated with y_{s_jt} or the taxa with presence $\boldsymbol{x}_{s_jt}^{(2)}$ associated with $\boldsymbol{I}(y_{s_jt} > 0)$, individually. In this pre-selection procedure, we model all the longitudinal metagenomes as independent samples regardless of time points (or age). One may restrict this procedure to a sub-community such as the species or subspecies of certain genera.

To reduce the computational burden of cross-validation for a high-dimensional OTU table, we randomly select P_0 OTUs from distinct relative abundance levels to represent the complexity of microbiota composition. The matched sets are divided into 5 folds, each being a validation fold for the model built on the other four (training) folds. The penalized log likelihood in equation (4) is the negative objective function in cross-validation. For each validation fold f and the selected OTU i, the loss function is $S_{fi} = -\tilde{L}(\hat{\theta}_{-f}^{i}; \boldsymbol{y}_f^i, \boldsymbol{O}_f)$, where $\hat{\theta}_{-f}^{i} = \arg\max \tilde{L}(\theta^i; \boldsymbol{y}_{-f}^i, \boldsymbol{O}_{-f})$. The optimal ρ is selected by the 'elbow point' minimizing $S = \sum_{f=1}^{5} \sum_{i=1}^{P_0} S_{fi}/(5P_0)$.

Data generation process for simulation scenario B1

Step 1: Estimate the baseline mean composition (or frequency) of microbiota ($\bar{\eta}_0$) and the overdispersion ($\xi_0 = 0.04$) at the starting time point $t = 1$ in TEDDY data by Dirichlet-Multinomial (DM) maximum likelihood estimate (MLE) of the observed counts. Generate the mean frequency of microbiota at the first time point by Dirichlet (DL) distribution: $\bar{\eta}_{01} \sim DL(\bar{\eta}_0, \xi_0)$.

Step 2: The mean frequency $\bar{\eta}_{0t}$ at a later time point $t > 1$ is generated by the following shifting procedure: increase the frequency of some OTUs in $\bar{\eta}_{01}$ (denoted by M_{base}^{+}) with a sum of Δ_t and simultaneously reduce that of other OTUs in $\bar{\eta}_{01}$ (denoted by M_{base}^{-}) by Δ_t. The absolute shift size Δ_t represented the age effect on microbiota. This shifting strategy characterized the inherent correlation between M_{base}^{+} and M_{base}^{-} because of the simultaneous compositional change in these OTUs. All the OTUs in $\bar{\eta}_{0t}$ are assigned to either M_{base}^{+} or M_{base}^{-} to account for the impact of latent exposures across time points.

Step 3: At each time point, the heterogeneity between matched sets is the overdispersion estimated by DM MLE based on the samples per time point in TEDDY, denoted by ξ_t. The overdispersion at the first time point is $\xi_1 = 0.05$ and linearly decreases over time, which mimics the time-dependent overdispersion observed in the infant-age metagenome in TEDDY. We generated a mean frequency for each matched set s at time point t by $\bar{\eta}_{st} \sim DL(\bar{\eta}_{0t}, \xi_t)$. If a set is labeled as 'high-risk', we shifted all the OTUs in $\bar{\eta}_{st}$ using the procedure in Step 2 with shift size Δ_{st}, which is a proportion of the maximum shift size, i.e., $\Delta_{st} = \gamma \Delta_{st}^{0}$.

Step 4: The between-subject heterogeneity within each matched set was the median DM MLE of overdispersion per matched set based on the real data, that is $\xi^* = 0.03$. Hence, we generated the true microbiota composition for a sample collected from a 'low-risk' subject j in set s at time t by $\bar{\eta}_{s_jt} \sim DL(\bar{\eta}_{st}, \xi^*)$. The shift in $\bar{\eta}_{s_jt}$ between 'low-risk' and 'high-risk' subjects were described in Results.

Step 5: The library size for each sample is simulated by a Poisson distribution $N_{s_jt} \sim PS(100000)$, truncated by a minimum of 10000. The raw counts per sample is generated by Multinomial (MN) distribution $C_{s_jt} \sim MN(N_{s_jt}, \bar{\eta}_{s_jt})$.

Abbreviations

AIC: Akaike information criterion; FDR: False discovery rate; FPR: False positie rate; FPPR: False or pseudo positive rate; GH: Gauss-hermite; JMR: Joint model with matching and regularization; LMM: Linear mixed-effect model; OTU: Operational taxonomic unit; TEDDY: The environmental determinants of diabetes in the young; ZIBR: Zero-inflated beta regression.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08890-1.

Additional file 1: Appendix. Gauss-Hermite quadrature approximation for marginal likelihood.

Additional file 2: Supplementary Table S1. The list of OTUs detected by JMR, JMR-NC, LMM-N, individually, as shown in Figure 7.

Acknowledgements

The TEDDY study is funded by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Allergy and Infectious Diseases, National Institute of Child Health and Human Development, National Institute of Environmental Health Sciences, Centers for Disease Control and Prevention, and JDRF. We thank the TEDDY study data coordinating center at Health Informatics Institute, University of South Florida for data processing and sharing. We thank Suraj Sarvode Mothi for data cleaning support.

Authors' contributions

QL proposed and implemented the model and algorithm, performed the experiments, analyzed the data, and drafted the manuscript. KV conceived the real data analysis, interpreted results, and contributed to manuscript writing. QL and YH designed the experiments. CL contributed to methodology discussion and manuscript writing. ET and LR contributed to manuscript writing. JK supervised the study design, sample collection, data generation and analysis in the TEDDY cohort. The author(s) read and approved the final manuscript.

Funding

The work of QL was in part supported by U24DK097771 from the National Institute of Diabetes, Digestive and Kidney Diseases via the NIDDK Information Network's (dkNET) New Investigator Pilot Program in Bioinformatics. QL and CL are in part supported by P30 Cancer Center Support Grant (CA21765) funded by National Cancer Institute; and the American Lebanese Syrian Associated Charities (ALSAC).

Availability of data and materials

The TEDDY Microbiome WGS data that supports the findings of this study have been deposited in NCBI's database of Genotypes and Phenotypes (dbGaP) with the primary accession code phs001443.v1.p1. The R package mtradeR that implements JMR and the simulation pipeline is available at https://github.com/qianli10000/mtradeR.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis 38105, TN, USA. [2]Health Informatics Institute, University of South Florida, Tampa 33620, FL, USA. [3]Department of Microbiology and Cell Science, University of Florida, Gainesville 32611, FL, USA. [4]Department of Biostatistics and Bioinformatics, Emory University, Atlanta 30322, GA, USA.

Received: 19 June 2022 Accepted: 14 September 2022
Published online: 19 September 2022

References

1. Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature. 2018;562(7728):583–8.
2. Vatanen T, Franzosa EA, Schwager R, Tripathi S, Arthur TD, Vehik K, et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. Nature. 2018;562(7728):589–94.
3. Wang DD, Nguyen LH, Li Y, Yan Y, Ma W, Rinott E, et al. The gut microbiome modulates the protective association between a Mediterranean diet and cardiometabolic disease risk. Nat Med. 2021;27(2):333–43.
4. Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, et al. Linking the human gut microbiome to inflammatory cytokine production capacity. Cell. 2016;167(4):1125–36.
5. Vatanen T, Kostic AD, d'Hennezel E, Siljander H, Franzosa EA, Yassour M, et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. Cell. 2016;165(4):842–53.
6. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26(1):32–46.
7. Hu YJ, Satten GA. Testing hypotheses about the microbiome using the linear decomposition model (LDM). Bioinformatics. 2020;36(14):4106–15.
8. Zhu Z, Satten GA, Mitchell C, Hu YJ. Constraining PERMANOVA and LDM to within-set comparisons by projection improves the efficiency of analyses of matched sets of microbiome data. Microbiome. 2021;9(1):1–19.
9. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016;32(17):2611–7.
10. Zhang X, Yi N. NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. BMC Bioinformatics. 2020;21(1):1–19.
11. Metwally AA, Yang J, Ascoli C, Dai Y, Finn PW, Perkins DL. MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. Microbiome. 2018;6(1):1–12.
12. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, et al. Multivariable association discovery in population-scale meta-omics studies. PLoS Comput Biol. 2021;17(11): e1009442.
13. Uusitalo U, Liu X, Yang J, Aronsson CA, Hummel S, Butterworth M, et al. Association of early exposure of probiotics and islet autoimmunity in the TEDDY study. JAMA Pediatr. 2016;170(1):20–8.
14. Ling W, Zhao N, Plantinga AM, Launer LJ, Fodor AA, Meyer KA, et al. Powerful and robust non-parametric association testing for microbiome data via a zero-inflated quantile approach (ZINQ). Microbiome. 2021;9(1):1–19.
15. Luna PN, Mansbach JM, Shaw CA. A joint modeling approach for longitudinal microbiome data improves ability to detect microbiome associations with disease. PLoS Comput Biol. 2020;16(12): e1008473.

16. Hu J, Wang C, Blaser MJ, Li H. Joint modeling of zero-inflated longitudinal proportions and time-to-event data with application to a gut microbiome study. Biometrics. 2021. https://doi.org/10.1111/biom.13515.
17. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. GigaScience. 2019;8(9):giz107.
18. Group TS. The environmental determinants of diabetes in the young (TEDDY) study. Ann N Y Acad Sci. 2008;1150(1):1–13.
19. Lee HS, Burkhardt BR, McLeod W, Smith S, Eberhard C, Lynch K, et al. Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. Diabetes Metab Res Rev. 2014;30(5):424–34.
20. Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G, et al. Toward defining the autoimmune microbiome for type 1 diabetes. ISME J. 2011;5(1):82–91.
21. Tetz G, Brown SM, Hao Y, Tetz V. Type 1 diabetes: an association between autoimmunity, the dynamics of gut amyloid-producing E. coli and their phages. Sci Rep. 2019;9(1):1–11.
22. Han R, Shi P, Zhang AR. Guaranteed Functional Tensor Singular Value Decomposition. arXiv preprint arXiv:2108.04201. 2021.
23. Li C, Xiao L, Luo S. Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's Disease. Biometrics. 2021;78(2):435–47.
24. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016;10(7):1669–81.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.