Hierarchical Models for Longitudinal Microbiome Data

Hainan Xu

Department of Mathematics and Statistics
McMaster University

Supervisor
Dr. Pratheepa Jeganathan

Abstract

The microbiome is a complex and dynamic microbial community of various environments. High-throughput sequencing has revolutionized the classification and quantification of microbes, enabling researchers to collect longitudinal microbiome data and gain insights into the dynamics of microbial communities. First, this thesis outlines the components of microbiome data, the process of obtaining a count table from raw sequences, and differential abundance analysis on the cross-sectional microbiome data. Then, we evaluate the topic-based mixed-effect model, which can identify microbiome communities and quantify the effect of interventions in longitudinal experiments. The results show that the prediction error increases after intervention if we use random intercept models.

Acknowledgement

I would like to thank my supervisor Dr. Pratheepa Jeganathan, for her invaluable support and patience throughout the process. Her mentorship introduced me to the fascinating data science and omics data fields. I would also like to extend my thanks to my family and friends who have supported me during the past four years at McMaster.

Contents

1 Introduction

The contingency tables play a crucial role in understanding complex biological systems. For example, the microbial communities, which can be summarized into a count table in humans, animals, plants, and other natural environments, are associated with ecosystem functioning. For instance, cross-sectional soil microbiome experiments reveal the association between microbial diversity and ecosystem functions (Banerjee & van der Heijden, 2022). Further, the ocean microbiome supports the food map and Earth's biogeochemical cycles (Paoli et al., 2022). In addition, the temporal changes in the vaginal microbiome community are one of the important factors in predicting premature labor among pregnant women (DiGiulio et al., 2015).

Researchers process large amounts of raw sequence data to obtain count tables. High-throughput sequencing is a widely-used method for extracting microbial sequencing data from various environments. Standard methods used to quantify microbes in the environment include marker-gene sequencing and shotgun metagenomic sequencing. Particularly, 16s rRNA gene sequences amplify the fourth (V4) to the sixth hypervariable (V6) region of the gene, which is then used for microbial taxonomy classification.

The raw microbial sequences consist of biological and technical variations. Therefore, technical noise is removed before understanding microbiome distribution. `DADA2` is one of the tools to denoise the raw sequences (Callahan et al., 2016). Given the raw sequences, the `DADA2` pipeline goes through preprocessing, filtering, denoising, and taxonomy assignments.

There are four components of microbiome data. They are a count matrix, sample information table, taxonomy table, and phylogenetic tree. All these components are essential to understand the heterogeneity in the microbial distributions in the environments. Thus, all the components should be integrated into a container. `phyloseq` is one of the R packages for storing all the components of the data (McMurdie & Holmes, 2013). Moreover, `phyloseq` provides exploratory and visualization tools such as bar plots, heatmaps, ordination plots, and interfaces to generalized linear models.

One of the important questions in microbiome research is quantifying changes in microbial communities temporally or within a time interval or after various perturbations. In general, longitudinal data are collected repeatedly over time from a subject or environment. Compared with cross-sectional data, the major advantage of longitudinal data is its capacity to distinguish the changes or trends within a certain subject over time. Thus, longitudinal microbiome data allows us to observe the differential abundance of the microbiome over time and other factors of interest, such as sample groups, demographic factors, or clinical factors.

Despite a large amount of data, analyzing microbiome data can be challenging due to its unique characteristics. The characteristics of microbiome data, for example, present difficulties such as additive errors and multiplicative errors. The additive error refers to the possible DNA contaminations in the data. The multiplicative error refers to the library size variation, which means the total number of taxa counts in each sample (library size) differs due to sequencing technology. In addition, a change in the abundance of a single taxon could affect the microbial distribution in the entire sample (Greenacre et al., 2022), and the co-occurrence of taxa is important (Jeganathan & Holmes, 2021).

Some tools for removing microbiome contamination are based on regression (decontam) (Davis et al., 2018) and negative control samples (microDecon, BARBI) (McKnight et al., 2019; Cheng et al., 2019). One of the approaches to computing the library effect size is the median-of-ratio method, which uses the geometric mean of taxa to generate pseudo sample counts and calculates the library size scaling factor based on the pseudo sample (Anders & Huber, 2010).

In addition to the challenges posed by the nature of microbiome data, experimental design can further complicate the analysis process. For example, study designs may incorporate hierarchical, longitudinal, or spatial structures and thus bring dependencies to the data. Furthermore, the microbial variance might change over time for individuals or treatment groups. Therefore, statistical methods and tools that take account of correlations between samples are needed (Jeganathan & Holmes, 2021). Additionally, repeatedly collecting samples from a subject for every time point may not always be practical. Thus, many missing samples may also pose challenges for longitudinal data analyses. Also, imputing the missing points of observed individuals can be challenging due to heterogeneity (Kodikara et al., 2022).

Many microbiome data exploratory and visualization tools exist to understand microbiome diversity within and between subjects, outliers, the relation among microbial distribution, and experimental designs. For example, alpha and beta diversity measures the microbial variation within-sample and between-sample, respectively. The alpha-diversity is the mean number of unique taxa observed in a sample. The beta diversity measures the between-sample variation using different sample distance metrics.

The beta-diversity measure has also been used to test microbial distribution across environments. For example, the permutational multivariate analysis of variance (PERMANOVA) facilitates testing beta diversity under different environments. The PERMANOVA decomposes total dissimilarity into fixed and random effects (Anderson, 2017). However, because these measures focus on the overall diversity of the environments, they may not provide insight into the differences in taxa or bacterial communities or changes in taxa at or after interventions (Grantham et al., 2020).

For longitudinal data, mixed models are one of the approaches to detect the associations between the response and environments (Diggle et al., 2013). These methods have been extended to longitudinal microbiome data that can handle overdispersion and zero inflation in the linear models (Zhang et al., 2017; Kodikara et al., 2022). In addition, there are several R/Bioconductor packages available to conduct taxon-wise differential abundance analysis, such as `DESeq2`, `edgeR`, and `Voom/limma`. The most common statistic to test the taxa differences is the logarithmic fold change of abundance between factors. Moreover, `fido`, `mdsine`, and `SplinectomeR` have been used for microbiome dynamic analysis. The above packages differ in their model assumptions, handling library size differences, zero inflation, model estimation, and transformation.

In a microbiome ecosystem, bacteria with similar functionalities live together as a community and may change their dynamics at or after an intervention. They live in communities to protect against environmental stresses, communication, and adapt to changing conditions. These bacterial communities are latent variables that cannot be explicitly identified using exploratory data tools or taxon-wise analysis. Therefore, we propose to use Latent Dirichlet Allocation (LDA), a topic model, to identify and characterize these underlying microbiome communities and use a negative binomial mixed model on topics to quantify the intervention effect sizes.

Section 2 introduces the preliminaries, including the `DADA2` pipeline, which is used to denoise the raw sequences and cluster the raw sequences into ASVs. Container, notations, and taxon-wise differential abundance analysis are also presented. Then, we illustrate a workflow with examples of analyzing 16sRNA datasets to process the data, followed by differential abundance analyses using `DESeq2`. Section 3 describes the topic model, where the generative process is illustrated, and the analyses for the example longitudinal data are included. Section 4 consists of the negative binomial mixed models for topics and quantifies the intervention effect size using hold-out samples. Finally, in Section 5, we summarize the results, discuss alternative methods for hyperparameter search for the topic model, and other possible approaches for capturing the effect sizes in longitudinal microbiome data.

2 Preliminaries

This section uses the 16S rRNA gene sequencing Zymo mock bacterial community dataset to demonstrate denoising sequences using `DADA2`. In addition, we reuse the soil microbiome dataset in Jeganathan & Holmes (2021) to find differentially abundant taxa between experimental factors using `DESeq2`.

2.1 DADA2

First, we investigate the quality profile plots of some samples to decide on trimming sequences. Figure 1 shows an example of a quality profile plot, where the x-axis is the quality score, and the y-axis is the cycle (length of the reads). The grey-scaled heat map shows the frequencies of sequences in different cycles with their corresponding quality scores. The green line indicates the mean score, the orange colour is the median, and the dashed orange lines are the 25th and 75th quantiles of the quality score.

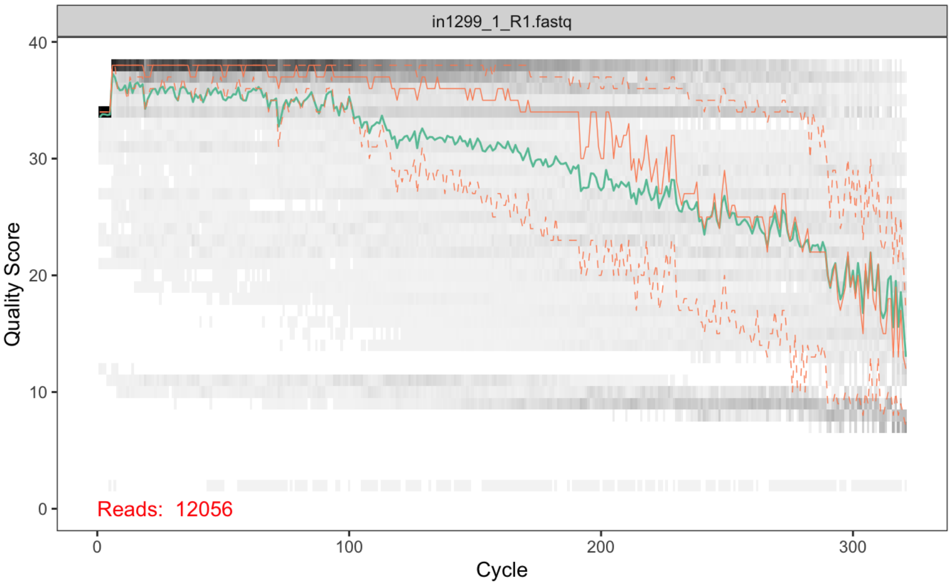


Figure 1: A quality profile plot of one sample. The x-axis is the quality score, and the y-axis is the cycle (length of the reads). The grey-scaled heat map shows the frequencies of sequences in different cycles with their corresponding quality scores.

Though the sequence with longer read lengths usually contains more information, there is a trade-off between quality score and length of reads. Combining quality score plots and bioinformatic experts' suggestions, we can trim sequence cycles that do not contain sufficient information, such as reads that are less than 320 nucleotides in length.

Next, `DADA2` uses a parametric error model to denoise the sequences. In this process, we estimate the error model using the whole or subset of samples, dereplicate the sequencing reads, and identify amplicon sequencing variants (ASV) (Callahan et al., 2016).

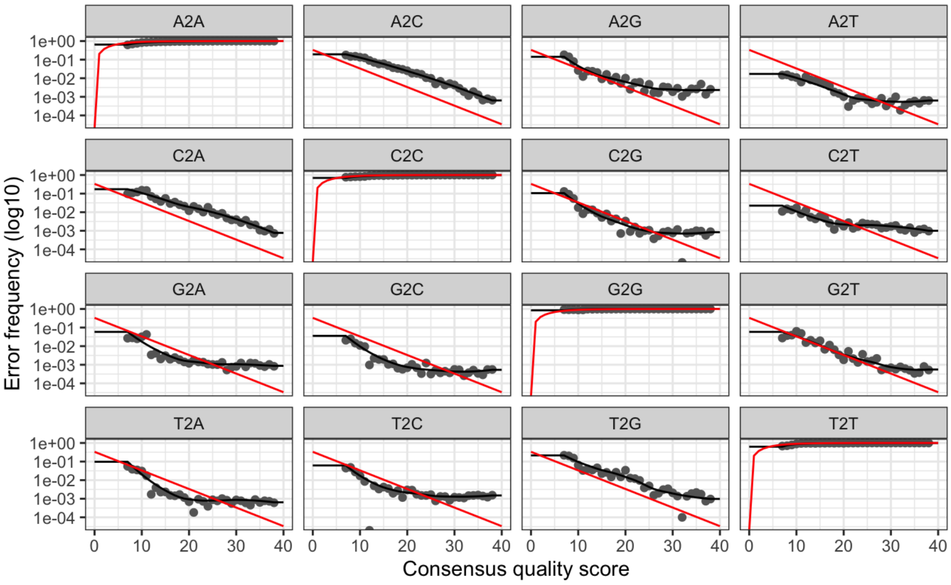


Figure 2: Learned error model for a set of samples. The black points are the observed error rate. The red line is the expected error rate under the quality score change.

Figure 2 shows the learned error model for the Zymo mock bacterial community dataset. The black points are the observed error rate. The red line is the expected error rate under the quality score change. If the error model is a good fit, we expect the black lines to coincide with the red lines, and the error rate will drop with an increased quality score.

We merge the paired reads using designated primers if we have the paired-end reads. Then, we construct a sequence table (contingency/count table). The sequence table contains the counts of different ASVs (taxa) in each sample. Finally, we use reference databases to construct a taxonomy table that provides taxonomic classifications of the ASVs. Table 1 shows the count table obtained after denoising the raw sequences containing V taxa and D samples. y_{vd} represents the number of times taxa$_v$ being observed in sample d. We assume $v = 1, \dots, V$ and $d = 1, \dots, D$. The library size for sample$_d$ is estimated by $S_d = \sum_{i=1}^{v} y_{id}$.

Table 1: Count table $\mathbf{Y} \in \mathbb{N}^{V \times D}$ of D samples and V taxa. y_{vd} denotes the number of times taxa$_v$ being observed in sample$_d$.

	Sample$_1$	...	Sample$_d$	...	Sample$_D$
Taxa$_1$	y_{11}		y_{1d}		y_{1D}
Taxa$_2$	y_{21}		y_{2d}		y_{2D}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Taxa$_v$	y_{v1}	...	y_{vd}	...	y_{vD}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Taxa$_V$	y_{V1}		y_{vd}		y_{VD}

Additionally, information regarding samples and experiment design are stored in sample data as shown in Table 2. The taxonomy table is as shown in Table 3.

Table 2: Sample data of D samples collected repeatedly from Q subjects over time.

	Sample ID	Subject ID	Time	Control	Treatment
Sample_1	Sample_1	Subject_1	t_1	TRUE	FALSE
Sample_2	Sample_2	Subject_1	t_2	TRUE	FALSE
⋮	⋮	⋮	⋮		
Sample_d	Sample_d	Subject_q	t_2	FASLE	TRUE
⋮	⋮	⋮	⋮		
Sample_D	Sample_D	Subject_Q	t_n	FALSE	TRUE

Table 3: Taxonomy table that contains the taxonomy information of each taxa in count matrix.

	Kingdom	Phylum	Class	Order	Family
Taxa_1	*Bacteria*	*Cyanobacteria*	*Chloroplast*	*Streptophyta*	
Taxa_2	*Bacteria*	*Chloroflexi*	*Anaerolineae*	*Anaerolineales*	*Anaerolinaceae*
⋮	⋮	⋮	⋮		
Taxa_v	*Bacteria*	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	S24-7
⋮	⋮	⋮	⋮		
Taxa_V	*Bacteria*	*Chloroflexi*	*Chloroflexia*	*Herpetosiphonales*	*Herpetosiphonaceae*

2.2 Container: phyloseq

We store the count table in Table 1, sample information in Table 2, and taxonomy table in Table 3 in a `phyloseq` object. The `phyloseq` R/Bioconductor package provides methods to preprocess, visualize, ordination, and exploratory data analysis (McMurdie & Holmes, 2013).

2.3 DESeq2

One of the goals of a microbiome study is to find differences in bacterial distribution across environments or experimental factors. Here we use `DESeq2` to test taxon differences across plant root microbiome environments. We consider a dataset with different treatments of peptide nucleic acids (pPNA) on the plants' root samples. In plant microbiome studies, pPNAs have been used to block host amplification. After removing host sequences, our goal is to test microbial distribution between two types of pPNA.

We preprocess the data to remove taxa that do not appear in any samples or less than 25 counts in at least five samples. Then, we convert the `phyloseq` object to `DESeq2` object. First, `DESeq2` estimates the library size scaling factor using the median-ratio method. Assuming that the sample count y_{dv} for each taxon follows a negative binomial distribution, and the dispersion does not vary across samples, DESeq2 computes the maximum likelihood estimators of mean μ_{dv} and dispersion parameters γ_v. Next, DESeq2 test each taxon differences between environments using Wald test statistic (standardized log fold change) or likelihood ratio statistic. Finally, DESeq2 shrinks the estimates using the mean-variance relationship. We can use either parametric or nonparametric models to do shrinkage. Figure 3 shows mean-variance estimation using local linear regression.

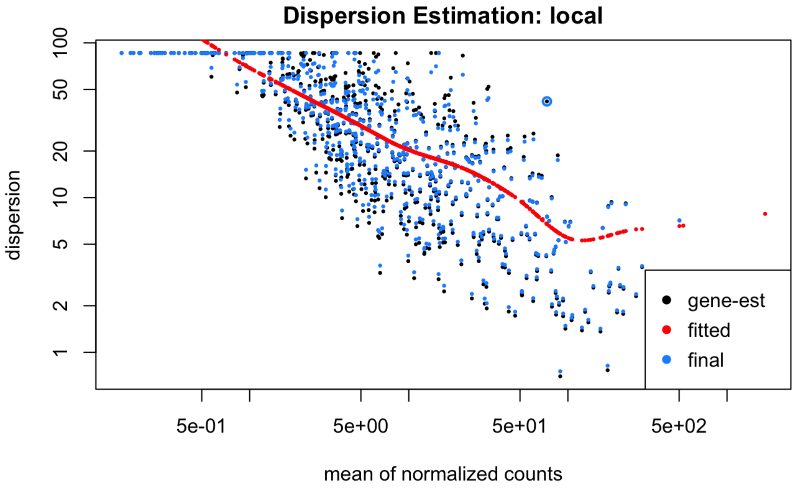


Figure 3: Mean-variance function estimation using local linear regression.

The negative binomial(NB) distribution is a way to model overdispersed count data. If the expected value of the negative binomial distribution is μ, then the variance is $\sigma^2 = \mu + \mu^2/\nu$ where ν is the dispersion parameter that controls the amount of overdispersion.

In DESeq2, we can choose the test statistic and the mean-variance estimation methods. Here we provide the result using the likelihood ratio test and local polynomial regression. This combination has the lowest residual for the fitted negative binomial model for each taxon. Moreover, DESeq2 shrinks the log2 fold changes and performs multiple hypothesis testing for differential abundances. In addition, it uses the Benjamini-Hochberg method to control the false discovery rate.

For our dataset, the hypothesis is that each taxon is not differentially abundant between two pPNA treatments. Figure 3 shows the DESeq2 results of using the likelihood ratio test and local polynomial regression to estimate the mean-variance relationship. This combination has the lowest residual for each taxon's fitted negative binomial model. DESeq2 shrinks the log2 fold changes and performs multiple hypothesis testing for differential abundance analysis. Moreover, it uses the Benjamini-Hochberg method to control the false discovery rate (Benjamini

& Hochberg, 1995).

Table 4: Paritial DESeq2 Output. The taxa are arranged in ascending order based on their p-values, from the lowest to the highest.

Taxa#	baseMean	log2FoldChange	lfcSE	stat	pvalue
136	16.966	21.680	2.155	10.050	8.3e-24
577	6.629	5.987	2.045	2.928	3.4e-03
67	5.288	5.660	2.341	2.418	1.5e-02
65	8.281	6.305	2.786	2.263	2.3e-02
524	6.474	-5.129	2.269	-2.261	2.4e-02
410	10.111	-5.134	2.324	-2.210	2.7e-02

In Table 4, we present taxon with adjusted p-values lower than 0.05 as differentially abundant. We found six taxa are differentially abundant between two pPNA treatments out of 5808 taxa.

Figure 4 visualizes the significant and non-significant taxa using the MA plot, a scatter plot of log2 fold changes (M) versus the average of normalized counts (A). The x-axis is the mean of normalized counts, and the y-axis is the log fold change. The triangle indicates the significant taxa, whereas the gray points have no sufficient evidence to conclude the significance. We observe again that most taxa were not significantly different between pPNA treatments.

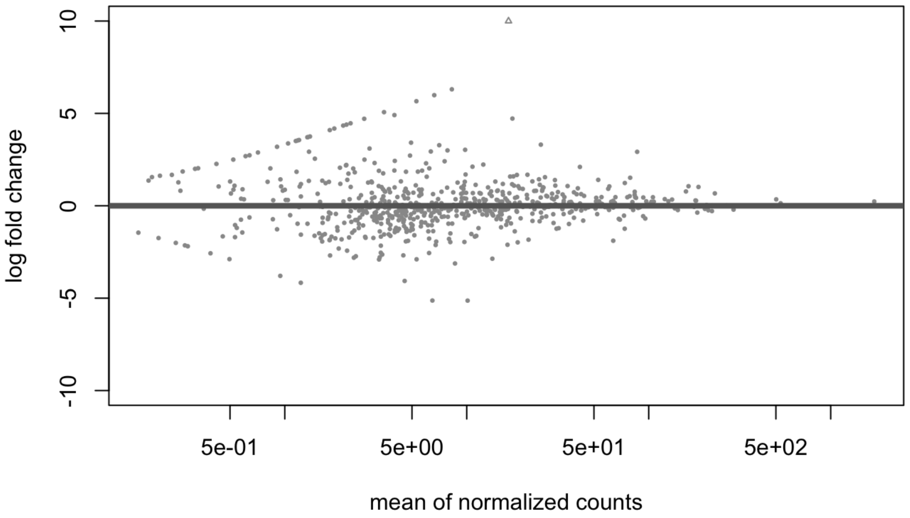


Figure 4: MA plot. The x-axis is the mean of normalized counts (A), and the y-axis is the log fold change (M).

Next, we will demonstrate the topic model to identify microbial communities in a longitudinal microbiome dataset.

3 Topic Modeling

Table 1 is a multivariate count data. Each cell indicates the number of taxa observed for the taxa within each sample. Various methods have been proposed to analyze the count data, out of which the topic model, which is frequently used to extract topics in text analyses, has been proposed to discover and extract the communities in microbiome data.

These latent variables can be extracted using latent Dirichlet allocation (LDA). In applying LDA to microbiome analyses, a document in text analysis corresponds to a sample, a term corresponds to a taxon, a topic corresponds to a community, a word corresponds to a sequence read, and the corpus corresponds to the environment. Based on this connection, we apply LDA to Table 1 to quantify topic proportion within each sample.

3.1 Latent Dirichlet Allocation

In this section, we describe the generative process related to LDA, which allows fractional membership across a set of topics (Sankaran & Holmes, 2019; Blei et al., 2003). Under LDA, samples can be represented by mixtures of topics, and topics can be represented by mixtures of taxa.

Let D be the number of samples, V be the number of taxa, T be the number of topics, and S_d be the total number of sequences in d^{th} sample. w_{dn} represents n^{th} sequencing reads in the d^{th} sample, with z_{dn} as the associated topic. θ_d is a T-dimensional vector that represents the d^{th} samples mixture over T topics (Table 5). β_k is a V-dimensional vector that represents the t^{th} topic mixture across different samples (Table 6). α and γ are the hyperparameters in the Dirichlet distribution to generate the distribution of β_t and θ_d, respectively. The generative process of LDA is as follows.

1. For each topic t,
   (a) draw $\beta_t \sim \text{Dir}(\gamma)$, where γ is a V-dimensional vector
2. For each sample d,
   (a) draw $\theta_d \sim Dir(\alpha)$, where α is a T-dimensional vector
   (b) For each n^{th} sequence read in sample d,
      i. Draw $z_{dn|\theta_d} \sim Mult(1, \theta_d)$
      ii. $w_{dn|\beta_k, z_{dn}} \sim Mult(1, \theta_d)$

The generative process above can be summarized as follows.

$$\begin{aligned}
w_{dn} \mid \beta_t, z_{dn} &\overset{iid}{\sim} \text{Mult}(1, \beta_{z_{dn}}) \quad \text{for } d = 1, \dots, D \text{ and } n = 1, \dots, S_d, \\
z_{dn} \mid \theta_d &\overset{iid}{\sim} \text{Mult}(1, \theta_d) \quad \text{for } d = 1, \dots, D \text{ and } n = 1, \dots, S_d, \\
\theta_d &\overset{iid}{\sim} \text{Dir}(\alpha) \quad \text{for } d = 1, \dots, D, \\
\beta_t &\overset{iid}{\sim} \text{Dir}(\gamma) \quad \text{for } t = 1, \dots, T.
\end{aligned}$$

Let's consider y_{dv} as the number of times the v^{th} taxon appears in the d^{th} sample and

$$y_{dv} = \sum_{n=1}^{S_d} I(w_{dn} = v),$$

where S_d is the number of sequencing reads in the d^{th} sample (library size), and $I(w_{dn} = v)$ is an indicator function that equals to 1 if the n^{th} read in the d^{th} sample is equal to the v^{th} taxon, and 0 otherwise.

Then we can write the model generative process as follows.

$$\begin{aligned}
y_{d\cdot} \mid (\beta_t)_{t=1}^{T} &\overset{\text{iid}}{\sim} \text{Mult}(S_d, B\theta_d) \text{ for } d = 1, \dots, D, \\
\theta_d &\overset{\text{iid}}{\sim} \text{Dir}(\alpha) \text{ for } d = 1, \dots, D, \\
\beta_t &\overset{\text{iid}}{\sim} \text{Dir}(\gamma) \text{ for } t = 1, \dots, T.
\end{aligned}$$

Joining all topics column-wise, we obtain a $V \times T$ matrix B, where $B = [\beta_1, ...\beta_T]$. Thus, the unknown parameters are θ_d and B (Tables 5 and 6).

Table 5: Topic proportions over samples $\theta_d, d = 1, \dots D$.

	Sample$_1$	$\dots$	Sample$_d$	$\dots$	Sample$_D$
Topic1	θ_{11}		θ_{1d}		θ_{1D}
Topic2	θ_{21}		θ_{2d}		θ_{2D}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Topic$_t$	θ_{t1}	$\dots$	θ_{td}	$\dots$	θ_{tD}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Topic$_T$	θ_{T1}		θ_{Td}		θ_{TD}

Table 6: Taxa proportions over topics (B).

	Topic$_1$	$\dots$	Topic$_t$	$\dots$	Topic$_T$
Taxa1	β_{11}		β_{1t}		β_{1T}
Taxa2	β_{21}		β_{2t}		β_{2T}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Taxa$_v$	β_{v1}	$\dots$	β_{vt}	$\dots$	β_{vT}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Taxa$_V$	β_{V1}		β_{Vt}		β_{VT}

3.2 Choosing Optimal Number of Topics

When applying LDA to microbiome data, the number of topics T needs to be prespecified, which defines the complexity of the model. In addition, changing the number of topics may change the interpretation of the result. Therefore, choosing an appropriate T is crucial to the downstream analysis.

To find the best T, we first fit LDA across various choices of T. However, due to the label-switching problem, it is difficult to decide the relationships between topics generated by different T. To tackle this problem, we align the topics based on optimal transport, which finds the path with minimum transport cost between the topics discovered for varied T (Fukuyama et al., 2021). The alignment diagnostics can be conducted by looking at the coherence and refinement scores.

3.3 Parameter Estimation

After specifying the number of topics T, we consider the hyperparameters α and γ to be less than one to generate sparse mixtures.

Using Bayes' theorem, the joint posterior distribution of θ and B is as follows:

$$P(\theta_d, B | y_{d.}, S_d, \alpha, \gamma) = \frac{P(y_{d.}|\theta_j, B, S_d) P(\theta_d, B|\alpha, \gamma)}{P(y_{d.})} \tag{1}$$

where $B = [\beta_1, \cdots, \beta_T]$. $P(y_{d.})$ is the marginal distribution integrating the numerator in equation 1, which is difficult to compute. Thus, the posterior distributions of parameters β and θ are estimated using variational expectation maximization (Blei et al., 2003).

3.4 Results of the Topic Model

We apply LDA to the Vancomycin-Resistant Enterococcus Faecium (VREfm) data, a longitudinal study investigating the gut microbiome under the VREfm colonization following antibiotic treatment. The VREfm dataset includes three components, count data, sample data, and taxonomy table. The count data includes the counts from longitudinal samples. Sample data includes information relating to the samples and experiment designs. In this experiment, samples are collected from nine mice over 14 days with two steps treatment, including antibiotic treatments (days 6 and 7) and VREfm colonization (day 9). The nine mice were raised in three cages labeled A, B, and C, respectively.

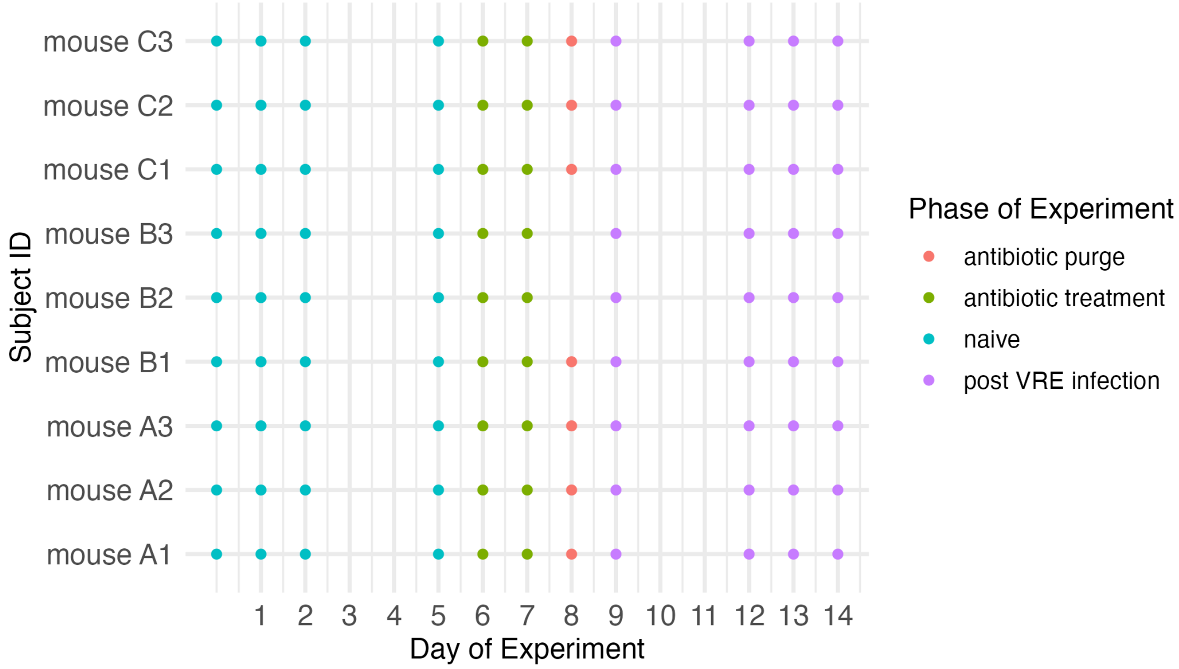


Figure 5: Sampling schedule plot. The mice are given two-step treatments: 1. antibiotic treatments on both day 6 and day 7; 2.VREfm colonization at day 9.

The sampling schedule is shown in Figure 5. Each mouse went through the four phases of the experiment. They are as follows.

1. Naive phase (day 1 - day 5): mice are not given any special treatment;

2. Antibiotic phase (day 6 - day 7): mice are given antibiotic treatment on days 6 and 7;

3. Antibiotic purge phase (day 8) The administration of antibiotics to all mice is discontinued to prevent the development of antibiotic resistance ;

4. Post VREfm (day 9 - day 14): mice are given one-time VREfm colonization on day 9.

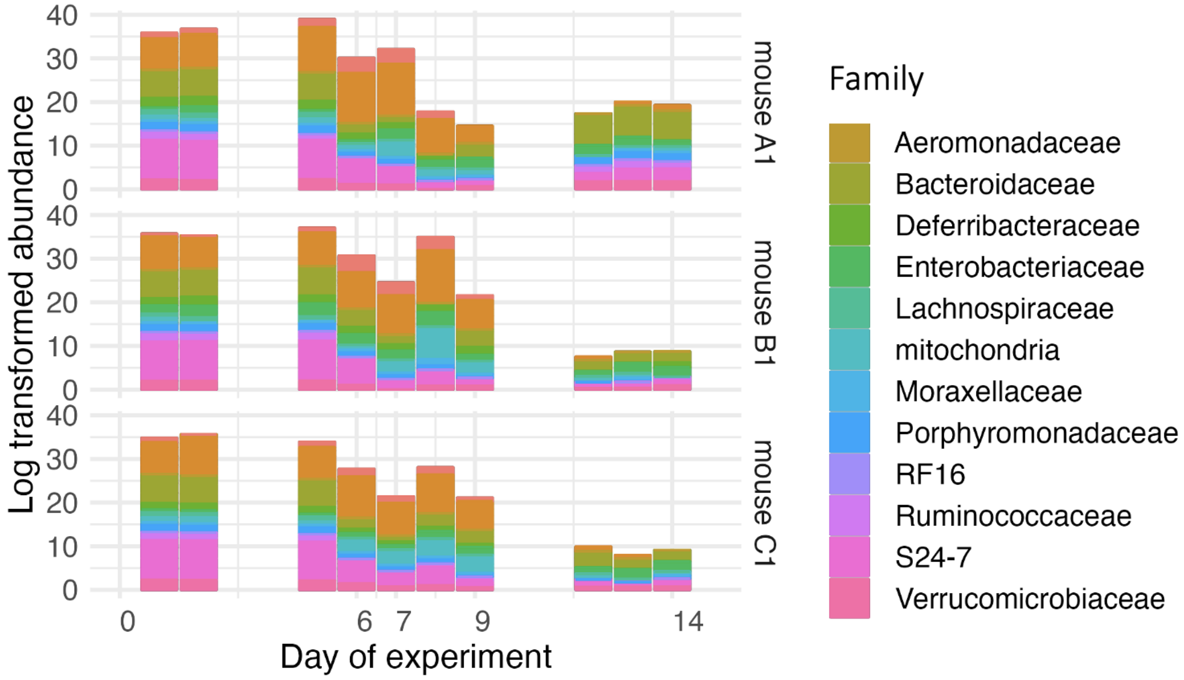


Figure 6: Temporal change of top 50 taxa.

The observed count table includes information on 3574 taxa across 97 samples. We filter taxa by removing low-abundance taxa. The reduced data contains 396 taxa across 97 samples after selecting taxa with at least ten reads in at least two samples. Figure 6 displays the temporal change of the taxa over three mice in different groups.

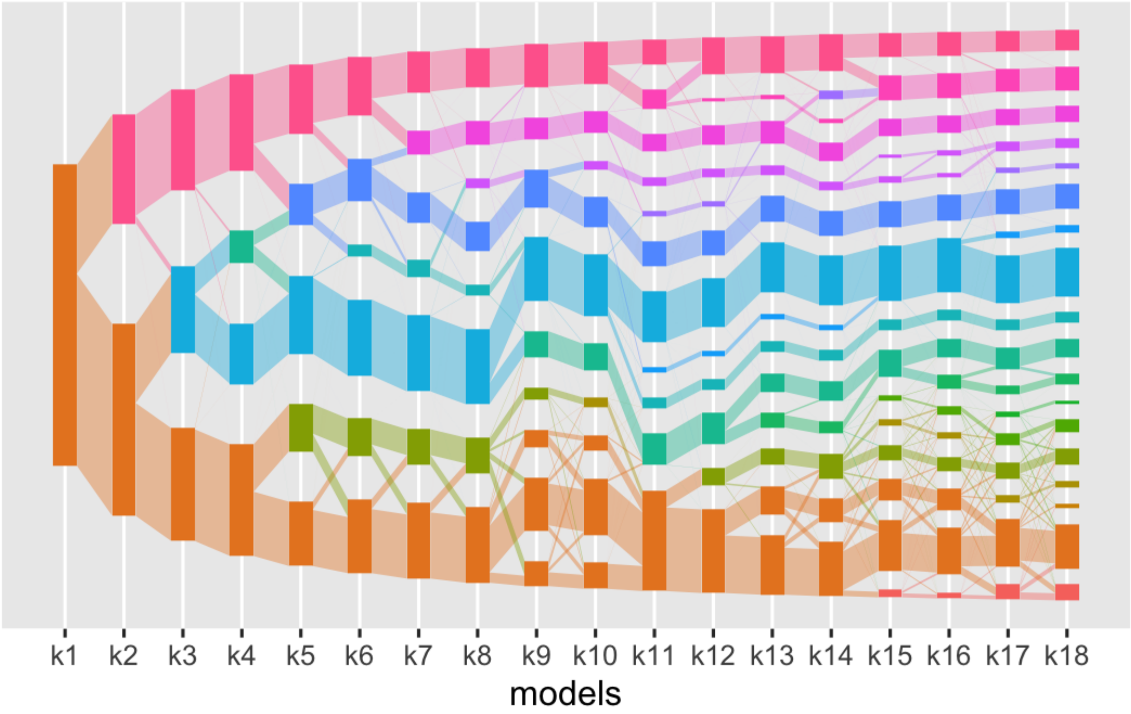
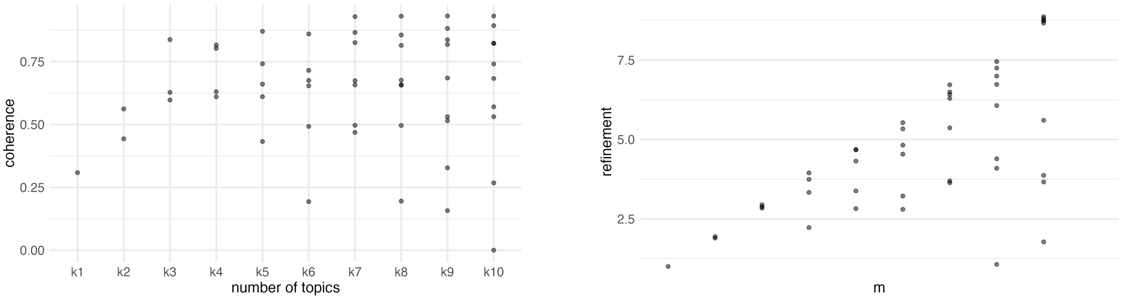


Figure 7: Topic alignment using optimal transport.



(a) Coherence Score

(b) Refinement scores

Figure 8: A high coherence and a high refinement score indicate a good choice of k. We also consider small dispersion in the scores.

We varied the number of topics from 1 to 10 to decide the number of topics. Based on Figures 8 and 7, we choose the number of topics as $T = 7$.

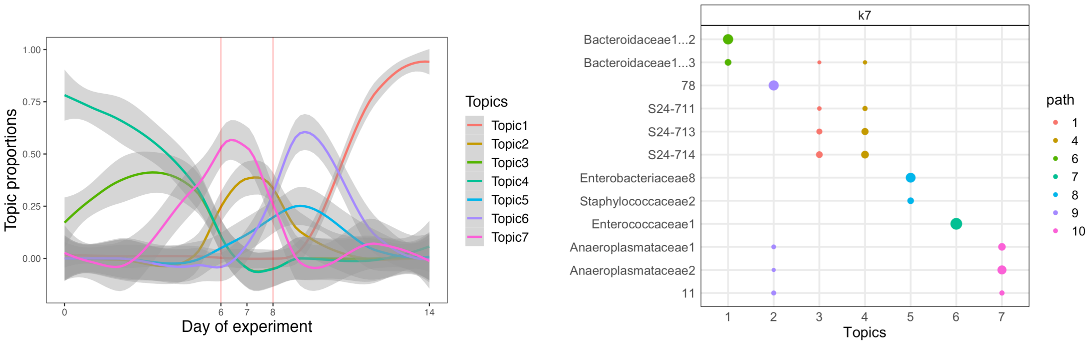


Figure 9: (a) Y-axis is the average across samples of the median proportion of topics. X-axis is the day of the experiment. (b) Y-axis is the selected taxa at the family level. X-axis is the topics. Size is the median taxa proportions within each topic.

Figures 9(a) and (b) show the temporal change of topic proportions and the taxa proportions in each topic, respectively. Moreover, Figure 9(a) indicates differentially abundant topics across time. For example, during the post-VREfm infection phase (after day 9), Topic 1 becomes differentially abundant, mainly constituted by taxa with the Family of *Bacteroidaceae*.

4 Mixed-effect Model

4.1 Negative Binomial Mixed Model for Topic Abundance

In longitudinal microbiome data, repeated measures are used to collect the samples. Suppose the optimal number of topics is T. We have T topics, Q subjects, and D samples. Let K denote the topic abundance. The topic abundance K_{td} of topic$_t$ in sample$_d$ can be calculated by $S_d \times \beta_{td}$. S_d is the library size of sample$_d$, where $S_d = \sum_{v=1}^{V} y_{dv}$ and β_{td} is median of posterior estimation of the topic proportion for taxa$_t$ in sample$_d$.

Table 7: Topic abundance K.

	Sample$_1$	$\dots$	Sample$_d$	$\dots$	Sample$_D$
Topic$_1$	K_{11}		K_{1d}		K_{1D}
Topic$_2$	K_{21}		K_{2d}		K_{2D}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Topic$_t$	K_{t1}	$\dots$	K_{td}	$\dots$	K_{tD}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
Topic$_T$	K_{T1}		K_{Td}		K_{TD}

A negative binomial linear mixed model for each topic abundance $K_{t.}$ is as follows.

$$(K_{td}|B = b) \sim NB(X\beta + Zb, \sigma^2 W^{-1}), d = 1, \dots, D, \tag{2}$$

where K_{td} is the t-th topic abundance in d-th sample, β is a (p+1) dimensional vector for fixed effect coefficients, B is the random effect coefficient, and Z is the random effect design matrix.

We can write β, X, Z and $B = b$ (random intercepts) are as follows.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{D1} & x_{D2} & \cdots & K_{Dp} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1Q} \\ z_{21} & z_{22} & \cdots & z_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ z_{D1} & z_{D2} & \cdots & z_{DQ} \end{bmatrix},$$

$$b = \begin{bmatrix} b_1, b_2, \dots, b_Q \end{bmatrix}^T.$$

The design matrix Z, which contains only zeros and ones, relates random effects b to the response variable K. Each column of Z represents a subject, and each row corresponds to a sample. Given sample$_d$, if it belongs to subject$_q$, then z_{dq} equals to one, and the rest of the row is equal to 0. b_d is a Q dimensional vector that contains the random effect coefficients.

In (2), W is a diagonal matrix with known prior weights. We assume the random effect variable B is multivariate normal with variance-covariance matrix, Σ,

$$B \sim N(0, \Sigma). \tag{3}$$

We assume each subject is independent, so the variance-covariance matrix Σ is $\sigma^2 I$.

For the VREfm dataset, we explain the variability in the response K_{td} using a given day of the experiment (x_1), phase of the experiment (x_2), and subject indicator Z.

$$log(E(K_{td}|x_{1d}, x_{2d}, B = b)) = \beta_0 + \beta_1 x_{1d} + \beta_2 x_{2d} + Zb. \tag{4}$$

We use pseudo-likelihood estimation to fit the model in (4) implemented in `lmer` R package. Moreover, we access the goodness-of-fit using Akaike Information Criterion (AIC).

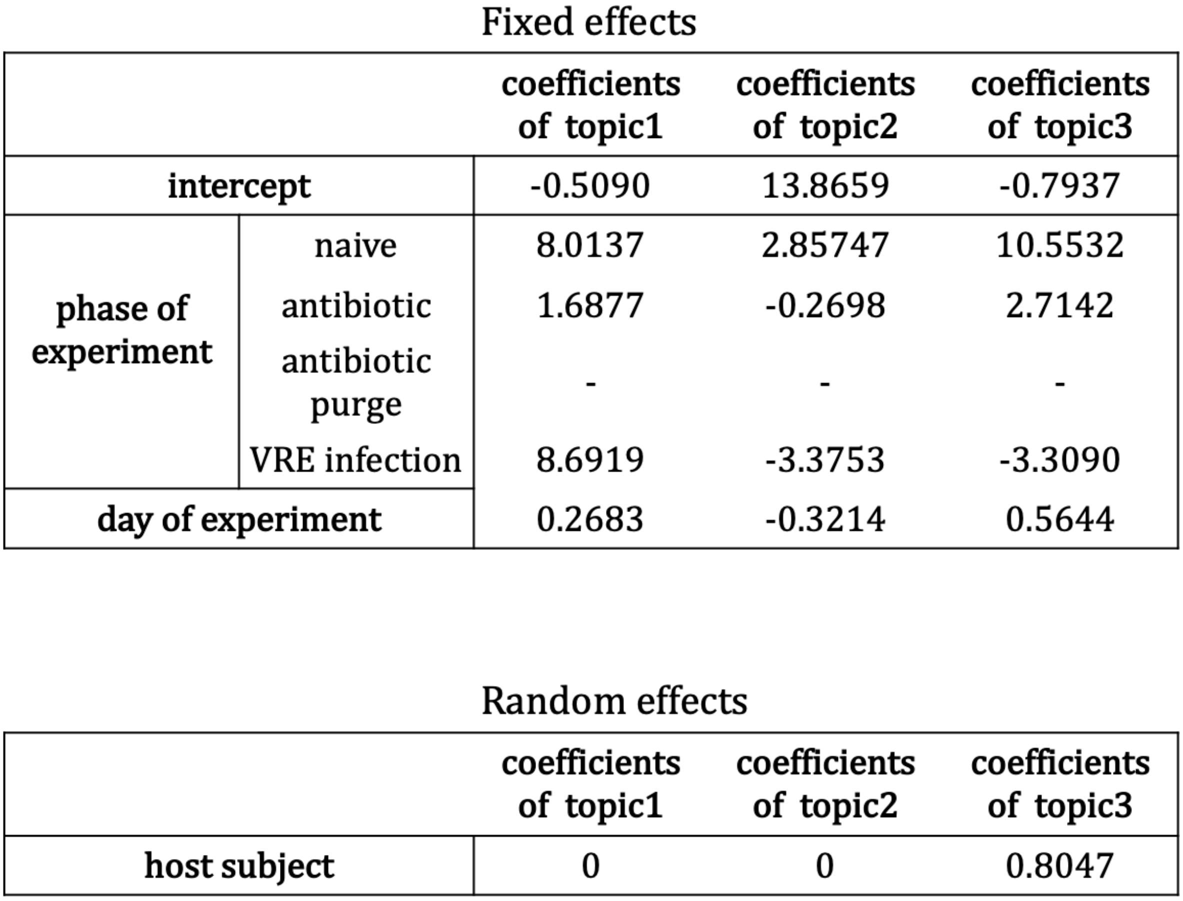
4.2 Predicting Effect Size

We use hold-out samples to evaluate the linear mixed model in estimating intervention effect sizes. First, we randomly hold out for samples from different subjects to quantify the effect size (two samples from the naive phase and two samples after the VREfm. Then, we fit the model using the rest of the data. Next, we predict the topic abundance of the four samples. Finally, we use the mean squared error (MSE) to check the performance of the model.

4.3 Results

We apply the mixed-effect model on the topics found in Section 3.4. Figures 10 and 11 display the results of the model fit and MSE, respectively.

Figure 10: Mixed-effect model results on three topics.



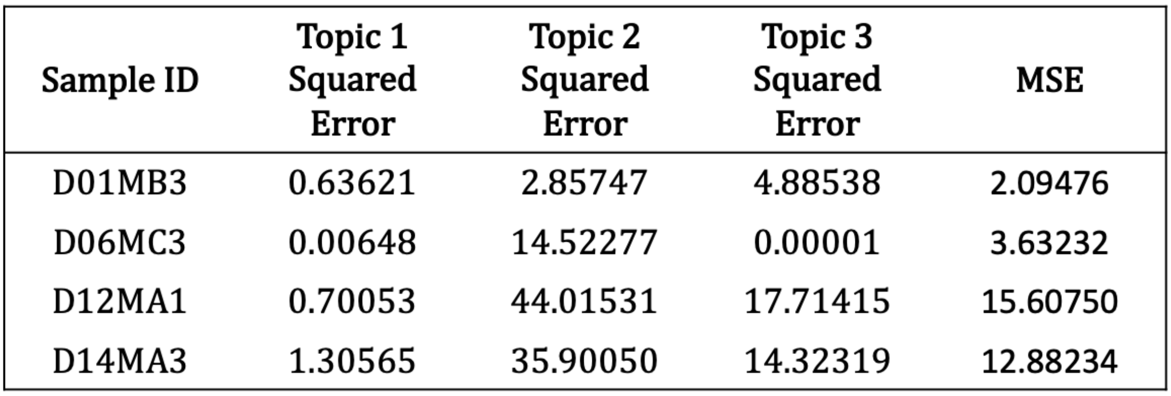
Fixed effects

		coefficients of topic1	coefficients of topic2	coefficients of topic3
intercept		-0.5090	13.8659	-0.7937
phase of experiment	naive	8.0137	2.85747	10.5532
	antibiotic	1.6877	-0.2698	2.7142
	antibiotic purge	-	-	-
	VRE infection	8.6919	-3.3753	-3.3090
day of experiment		0.2683	-0.3214	0.5644

Random effects

	coefficients of topic1	coefficients of topic2	coefficients of topic3
host subject	0	0	0.8047

Figure 11: MSE on holdout samples.



Sample ID	Topic 1 Squared Error	Topic 2 Squared Error	Topic 3 Squared Error	MSE
D01MB3	0.63621	2.85747	4.88538	2.09476
D06MC3	0.00648	14.52277	0.00001	3.63232
D12MA1	0.70053	44.01531	17.71415	15.60750
D14MA3	1.30565	35.90050	14.32319	12.88234

Figure 10 shows that the Topic 1 and 2 effect sizes are insignificant. For Topic 3, subject differences impact the result significantly. Subject differences have different effects on each topic. In addition, Figure 11 depicts that the MSE increases considerably after the intervention in samples D12MA1 and D14MA3. Thus, we need to incorporate a structured covariance matrix or intervention as a covariate to improve prediction after the intervention.

5 Discussion

We described a complete workflow for preprocessing and analyzing the abundance of taxa and microbiome communities over time. First, `DADA2` is implemented to denoise the raw sequences and cluster the sequences into ASVs. After that, the negative binomial generalized linear model is used to perform differential abundance analyses of taxa. Next, we apply latent Dirichlet allocation (LDA) to extract the microbial communities. We also decided on the optimal number of communities using topic alignment, coherence score, and refinement score. Finally, we used topic-based mixed-effect models to identify differentially abundant microbial communities and their effect sizes after the intervention.

Instead of taxon-wise analysis, we applied LDA to identify bacterial communities over longitudinal microbiome data. LDA can summarize high-dimensional multivariate microbiome data into a few key topics. This is particularly useful in the context of longitudinal microbiome data, as it helps to uncover the temporal variation in microbial communities and to characterize the changes in microbial communities at or after the intervention. However, selecting appropriate hyperparameters for the Dirichlet distribution in LDA can be challenging. The hyperparameters control the sparsity and density of the Dirichlet distribution, which impacts the interpretability and computational efficiency of the model.

Prior sensitivity analysis for LDA can ensure that the topics generated by the LDA model accurately reflect the underlying structure of the microbiome data. For example, we can find appropriate hyperparameters for Dirichlet distributions in LDA by fitting multiple models with different hyperparameters. In this project, with the goal of sparse topic proportions, we set the hyperparameters to be less than one. Despite this, other tools like `Rstan`, which uses the Hamiltonian Monte Carlo- No-U -Turn (NUTS) method, could be considered in future studies. High-performance computing can be employed to improve the efficiency of the computation, especially when using computationally intensive tools such as `Rstan`.

We used a mixed-effect model with random intercepts for subjects to estimate the temporal changes in topic abundance and intervention effect sizes. This method accounts for some of the heterogeneity in the data, allowing us to capture subject-specific differences and the time effect on the topic abundance. However, there are challenges when applying the mixed-effect model to topic abundances, such as not accounting for the time sequence (direction), underlying correlations between samples from the same subject, and the intervention pattern. For example, the intervention may cause a sudden change or delayed response in the topics' behavior at or after the intervention time point.

In the mixed-effect model, we can incorporate a covariance structure, such as AR(1), into the model to better account for time sequences in longitudinal data analyses. Furthermore, considering a more complex structure for the mixed-effect model would help to address the potential impact of intervention time on topics' behaviour.

In addition to the mixed effects models with random intercept, several alternative methods exist to estimate the temporal changes in microbiome data. One example is the mixed effects model with random intercepts and slopes, allowing individual variation. Besides, one can use spline models to capture the non-linear relationships between response, fixed, and random effects and model the dependence within the residuals.

In summary, studying longitudinal data can help understand the association between the microbiome ecosystem and health. The Bayesian hierarchical model poses a great potential to analyze these rich-structured, high-dimensional microbiome data. With the several hierarchies in the model, we can capture more complex structures and interactions between taxa under or after perturbation, which could aid the development of preventive or predictive medicine for challenging diseases.

References

Anders, S., & Huber, W. (2010, October). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. doi: 10.1186/gb-2010-11-10-r106

Anderson, M. J. (2017). Permutational multivariate analysis of variance (PERMANOVA). In *Wiley StatsRef: Statistics reference online* (pp. 1–15). John Wiley & Sons, Ltd. doi: 10.1002/9781118445112.stat07841

Banerjee, S., & van der Heijden, M. G. A. (2022). Soil microbiomes and one health. , 1–15. (Publisher: Nature Publishing Group) doi: 10.1038/s41579-022-00779-w

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. , *13*(7), 581–583. doi: 10.1038/nmeth.3869

Cheng, H., Tan, S., Sweeney, T., Jeganathan, P., Briese, T., Khadka, V., ... Relman, D. (2019). *Combined use of metagenomic sequencing and host response profiling for the diagnosis of suspected sepsi* (Tech. Rep.). (Type: article) doi: 10.1101/854182

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018, December). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 226. doi: 10.1186/s40168-018-0605-2

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, a. S. (2013). *Analysis of Longitudinal Data* (Second Edition, Second Edition ed.). Oxford, New York: Oxford University Press.

DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., ... Relman, D. A. (2015). Temporal and spatial variation of the human microbiota during pregnancy. , *112*(35), 11060–11065. doi: 10.1073/pnas.1502875112

Fukuyama, J., Sankaran, K., & Symul, L. (2021). Multiscale analysis of count data through topic alignment. , kxac018. doi: 10.1093/biostatistics/kxac018

Grantham, N. S., Guan, Y., Reich, B. J., Borer, E. T., & Gross, K. (2020). MIMIX: A bayesian mixed-effects model for microbiome data from designed experiments. , *115*(530), 599–609. (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2019.1626242) doi: 10.1080/01621459.2019.1626242

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., & Quinn, T. (2022). *Aitchison's compositional data analysis 40 years on: A reappraisal* (No. arXiv:2201.05197). arXiv. doi: 10.48550/arXiv.2201.05197

Jeganathan, P., & Holmes, S. P. (2021). A statistical perspective on the challenges in molecular microbial biology. , *26*(2), 131–160. doi: 10.1007/s13253-021-00447-1

Kodikara, S., Ellul, S., & Lê Cao, K.-A. (2022). Statistical challenges in longitudinal microbiome data analysis. , *23*(4), bbac273. doi: 10.1093/bib/bbac273

McKnight, D., Huerlimann, R., Bower, D., Schwarzkopf, L., Alford, R., & Zenger, K. (2019, May). microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environmental DNA*, *1*. doi: 10.1002/edn3.11

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. , *8*(4), e61217. doi: 10.1371/journal.pone.0061217

Paoli, L., Ruscheweyh, H.-J., Forneris, C. C., Hubrich, F., Kautsar, S., Bhushan, A., ... Sunagawa, S. (2022). Biosynthetic potential of the global ocean microbiome. , *607*(7917), 111–118. (Number: 7917 Publisher: Nature Publishing Group) doi: 10.1038/s41586-022-04862-3

Sankaran, K., & Holmes, S. P. (2019). Latent variable modeling for the microbiome. , *20*(4), 599–614. doi: 10.1093/biostatistics/kxy018

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. , *18*(1), 4. doi: 10.1186/s12859-016-1441-7

Supplementary Materials

All the works are reproducible at https://github.com/PratheepaJ/HMixM_microbiome.