Comprehensive Exam II: Report

Michael Agronah

December 8, 2022

Contents

Summary

There are two parts to this document. The first section, "Investigating the Power of Differential Abundance Microbiome Studies", aims to create a framework for comprehending how power, effect sizes, and abundances relate in differential microbiome abundance studies. A Multivariate Negative Binomial Mixed Model (MNBMM) is proposed in Part 2 as a model for examining trends in longitudinal microbiome data while accounting for the correlation structure between observations from individual subjects across time. MNBMM has the capability to handle overdispersion in microbiome data and can be used to model the relationship between longitudinal data and factors of interest such as age, disease status and groups.

1 Investigating the Power of Differential Abundance Microbiome Studies

1.1 Abstract

Microbiome studies are generally underpowered. Consequently, significance results reported in the literature can be misleading. Furthermore, the dynamics underlying the relationships among effect sizes, power and abundances in differential abundance studies is unclear. The ability to predict, even before the commencement of a research, how many species a researcher can reliably detect power for is an important question in differential abundance studies. This project seeks to develop a paradigm for understanding the relationships among effect sizes, power and abundances as well as the underpinning mechanism driving these relationships.

We investigate these relationships using ten (10) 16S microbiome datasets from children with autism spectrum disorder (ASD) (as treatment group) and neurotypical children (as control group). Effect size in this study is defined by log2 fold changes and control is used to refer to log2 control abundances. An R package will be developed from this project where new datasets can be fed into this package to explore the power, effect and abundance relationships.

1.2 Introduction

A key question in differential abundance studies is to determine differences in microbiome abundance between various groups of interest. For instance, what are the differences between microbiome abundances in control and treatment groups? This question can be translated into a hypothesis testing problem. The standard procedure in hypothesis testing is to first define a null hypothesis and then choose a significant level. The test rejects the null hypothesis whenever p-value is below the significant level and fails to reject the null hypothesis whenever p-value is above the significant level. This technique has two potential drawbacks. First, failing to reject the null hypothesis when the p-value is greater than the null hypothesis merely by accident. This results in what is known as the Type 1 error or the false positive error. The second drawback is rejecting the null hypothesis when the p-value is less than the significant level merely by chance, known as Type 2 error or false negative error.

Hypothesis testing depends on 4 parameters: (i) the effect size: a measure of the magnitude of differences between groups of interest; (ii) the sample size (n): the number of samples to be used in the studies; (iii) the power of a study ($1-\beta$): the probability of correctly rejecting the null hypothesis; and (iv) the confidence level (α): the probability of getting a false positive, that is, rejecting the null hypothesis when it is actually true. In general, power, sample sizes, significance level and effect size are positively related. For example, increasing the significance level increases power and a higher effect size will generally lead to higher power (Xia et al. 2018). Power studies are important to minimize Type 1 and Type 2 errors (Xia et al. 2018) and to determine the success of a study design even before commencing. In many funding applications, power analysis on a simulation study is required to evaluate how many sample sizes are needed to detect a given effect sizes and power of the studies.

The majority of microbiome studies, however, are underpowered (Kers and Saccenti 2021; Brüssow 2020). Consequently, significance results reported in the literature of differential abundance microbiome studies are influenced by Type 1 and Type 2 errors. Failure to correct for these errors has resulted in numerous instances

of conflicting conclusions reported in the literature of differential abundance microbiome studies (Brüssow 2020). Early research on the relationship between the gut microbiota of obese and non-obesity groups for instance, found significant differences in the diversity of the gut microbiota and significant differences in the Bacteroidetes/Firmicutes (B/F) ratio between obese and non-obese people (Ley et al. 2006). Later research conducted using larger samples found only slight differences in microbial diversity and no statistically significant change in the B/F ratio between non-obese and obese individuals (Sze and Schloss 2016). This problem makes it challenging to reproduce findings from microbiome studies reported in literature (Kers and Saccenti 2021). Due to the failure to correct Type1 and Type 2 errors, an average effect size and power of studies in the literature of differential abundance microbiome studies is unknown. Thus, it difficult to make informed decisions for conducting a power analyis of a design prior to commencing a study.

One potential remedy is to gain a better understanding of the relationships between power, effect size and abundances in differential abundance microbiome studies. This project seeks to develop a paradigm for understanding the relationships among effect sizes, power and abundances as well as the underpinning mechanism driving these relationships. Using this framework from this project, a new dataset can be explored to determine the relationships between power,effect size and abundance in the context of differential abundance studies.

1.3 Research goals

The goal of this project is to

1. investigate the relationships among power, effect sizes and abundances with the aim of understanding the mechanisms behind these relationships.
2. investigate the number of individual taxa in a differential abundance studies that power can be detected reliably, for a given effect and sample size.
3. investigate the number of sample sizes required to detect a given power and a given effect size.

1.4 Research Design and Methodology

1.4.1 Data collection and processing

Using the search terms *"autism[All Fields] AND 16S[All Fields]"*, *"autism[All Fields] AND 16S[All Fields] AND Fecal[All Fields]"*, raw sequence data from 10 projects that examined the microbiome of children with autism spectrum disorder were gathered from the European Nucleotide Archive (EBA) and the National Center for Biotechnology Information (NCBI). The following are the accession numbers for the 10 projects in the NCBI and ENA archives: PRJNA687773, PRJNA624252, PRJNA453621, PRJNA642975, PRJNA355023, PRJNA168470, PRJNA644763, PRJNA578223, PRJNA589343, and PRJEB45948. Children with autism spectrum disorders represent the treatment group in these datasets whereas neurotypical children are the control groups. Adaptor and primer sequences were removed using "**cutadapt**" function implemented in a bash script. The trimmed sequence from the "**cutadapt**" were then processed into Amplicon Sequence variants (ASVs) data using the **Dada2** pineline which involves filtering and trimming, error estimation, denoising, merging paired reads and chimeras removal (Y. Chen et al. 2020).

1.5 Method

1.5.1 Model description

The null hypothesis (H_0) and the alternative hypothesis (H_a) for testing differences between 2 groups (control and treatment) is formulated as

$$H_0 : \mu_C - \mu_T = 0 \tag{1}$$
$$H_a : \mu_C - \mu_T \neq 0, \tag{2}$$

where μ_C and μ_T denote the mean counts in control and treatment groups respectively. Effect sizes,a measure of the difference in the mean can then be tested. The negative binomial distribution has often been used to model microbiome count data due to the presence of overdispersion in microbiome count data.

Let K_{ij} denote the count data for the i^{th} taxa in the j^{th} sample. K_{ij} is modeled by negative binomial distribution defined by

$$K_{ij} \sim NB(mean = \mu_{ij}, dispersion = \alpha_i), \tag{3}$$

$$\mu_{ij} = g^{-1}(E_{ij}) \tag{4}$$

$$E_{ij} = \sum_r x_{jr}\beta_{ir}, \tag{5}$$

where g is a link function, β_{ir} are estimates of the effect sizes and x_{jr} are the covariates. The relationship between the variance of counts and the dispersion is given by $VarK_{ij} = \mu_{ij} + \alpha\mu_{ij}^2$. In this project, the estimating procedure implemented in the **DESeq2** package in R is used for estimating β_{ir}, μ_{ij} and α_i. Details concerning this procedure are explained in the paper by Love, Huber, and Anders (2014) and by Anders and Huber (2010).

1.5.2 Investigating the relationship between control abundances and effect sizes

The analysis presented in part 1 of this document is based on 4 of the 10 datasets. These 4 datasets are those with the assession numbers PRJNA453621, PRJEB45948,PRJNA589343 and PRJNA687773. Plots of the relationship between control abundances and effect sizes are shown in figure 1. The plots display an average effect size of zero. Variations around the smooth curve are greatest in the middle and lowest at the ends of the smooth curves. Thus, a quadratic function can be used to describe the variance of the effect sizes as a function of control abundances. A truncated normal distribution and a cauchy distribution, respectively, can be used to approximate the distribution of control abundances and effect sizes. Figure 2 and figure 3 show a histogram and density of the control abundance and effect sizes, as well as a plot of the density from simulation from fitting these distributions to the data. Effect sizes were fitted to Cauchy distributions with zero location and scale parameters defined by a quadratic function of control abundance, as shown in equation 6.

$$y_i = Cauchy(location = 0, scale = \gamma_i), \gamma_i = \exp(k_0 + k_1 x_i + k_2 x_i^2), \tag{6}$$

where y_i and x_i are the effect size and control abundance, respectively, for the i^{th} taxa and $k_j; j = 1, 2, 3$ are constants to be estimated.

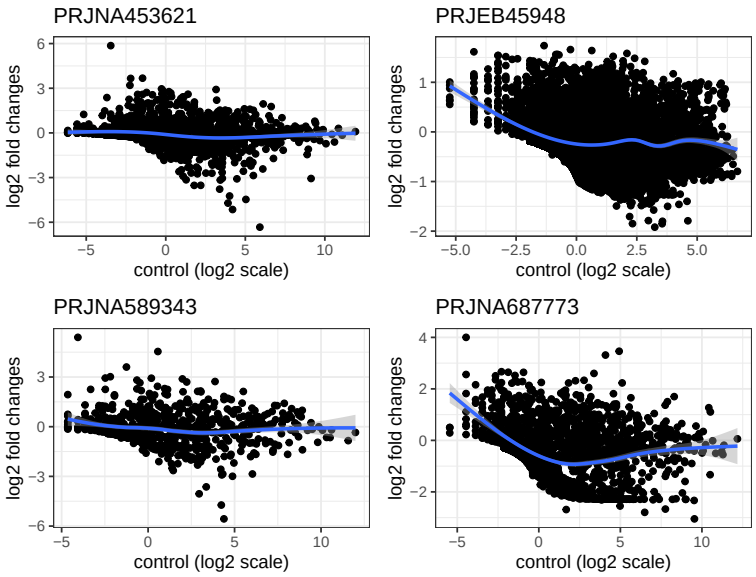

Figure 1: Plot of control against effect sizes for datasets with assession numbers PRJNA453621, PRJEB45948, PRJNA589343 and PRJNA687773

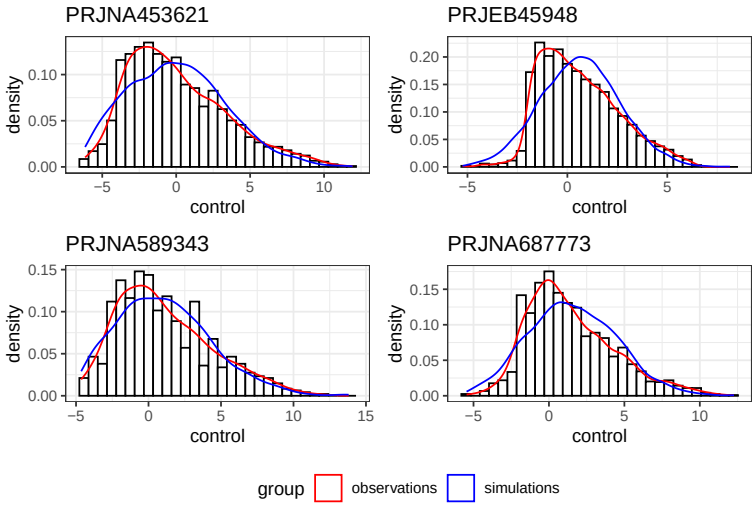

Figure 2: Histogram and density plots for control abundances and density from a simulated sample from fitting a truncated normal distribution

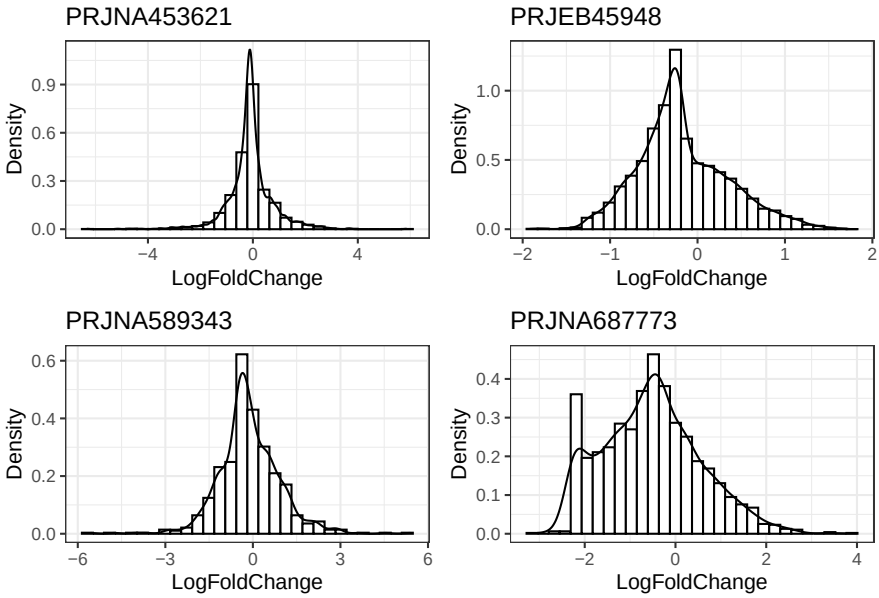

Figure 3: Histogram and density plots for effect sizes

1.6 Results

Figure 4 displays heatmaps of the power and total control abundances (on the non-log scale) for different ranges of control abundances and effect sizes. In each plot, regions with total control abundances less than 100 have been filtered out.

Figure 5 shows a contour plot illustrating how power changes with control abundances and effect sizes. Apart from the plot on the bottom left hand corner where power increases systematically with increasing effects in the region of the data points, all other plots show an initial increase or decrease in power, a peak around the center of the plot and then decrease or increase thereafter. These patterns may be better understood by standardizing the effect sizes to account for errors in the fold change estimates.

The delta method with a first order taylor series approximation was used to estimate the standard errors of the effect size estimates. The contour plot with standardized effect size is shown in figure 6. The plot at the top left corner of figure 6 shows increasing power with increasing standardized effect sizes, whereas all other plots show an initial increase in power as standardized effect increases and a later decrease in power. A higher order approximation of the delta method for calculating the standard error may result in uniformity in these patterns. This will be the next focus of this project.

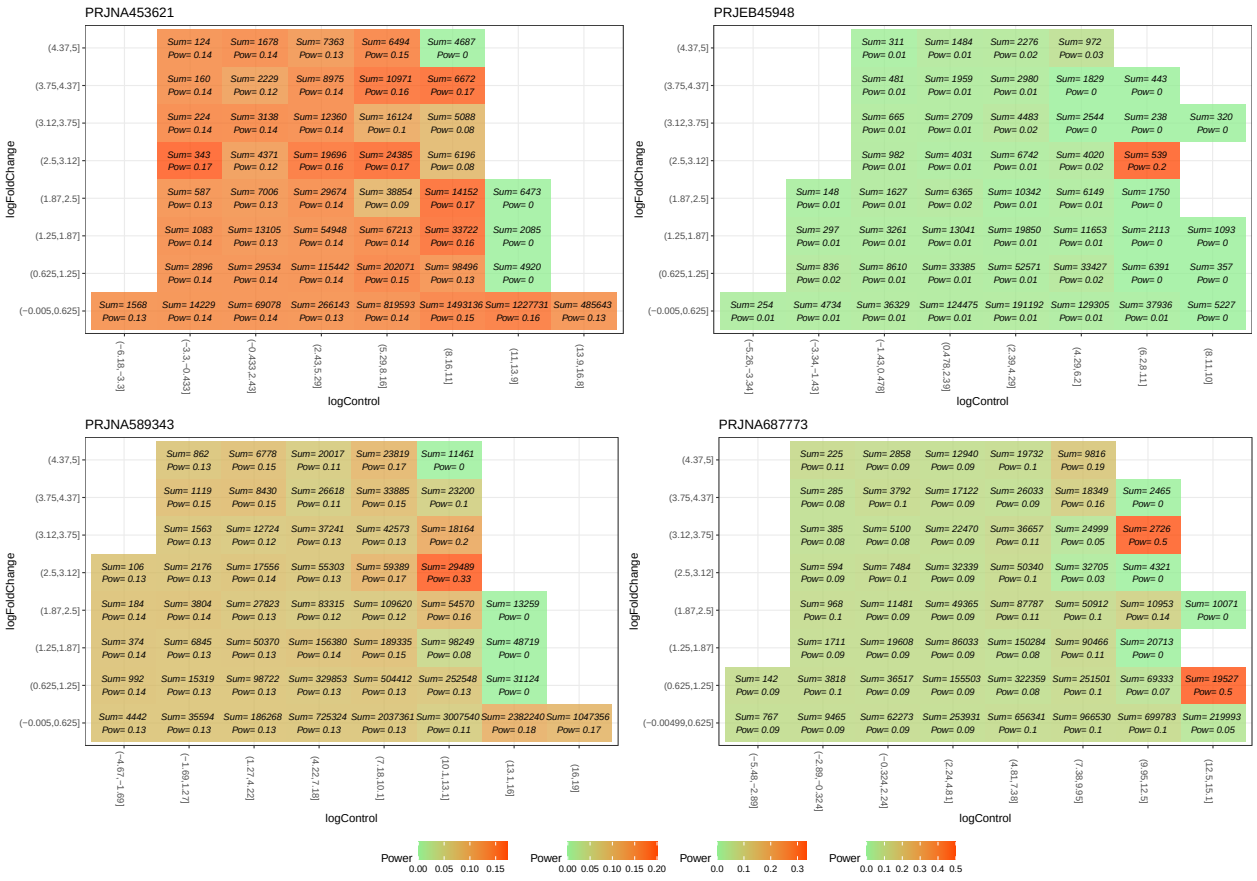

Figure 4: Heatmap showing the power and total control abundances for various ranges of control and effect sizes

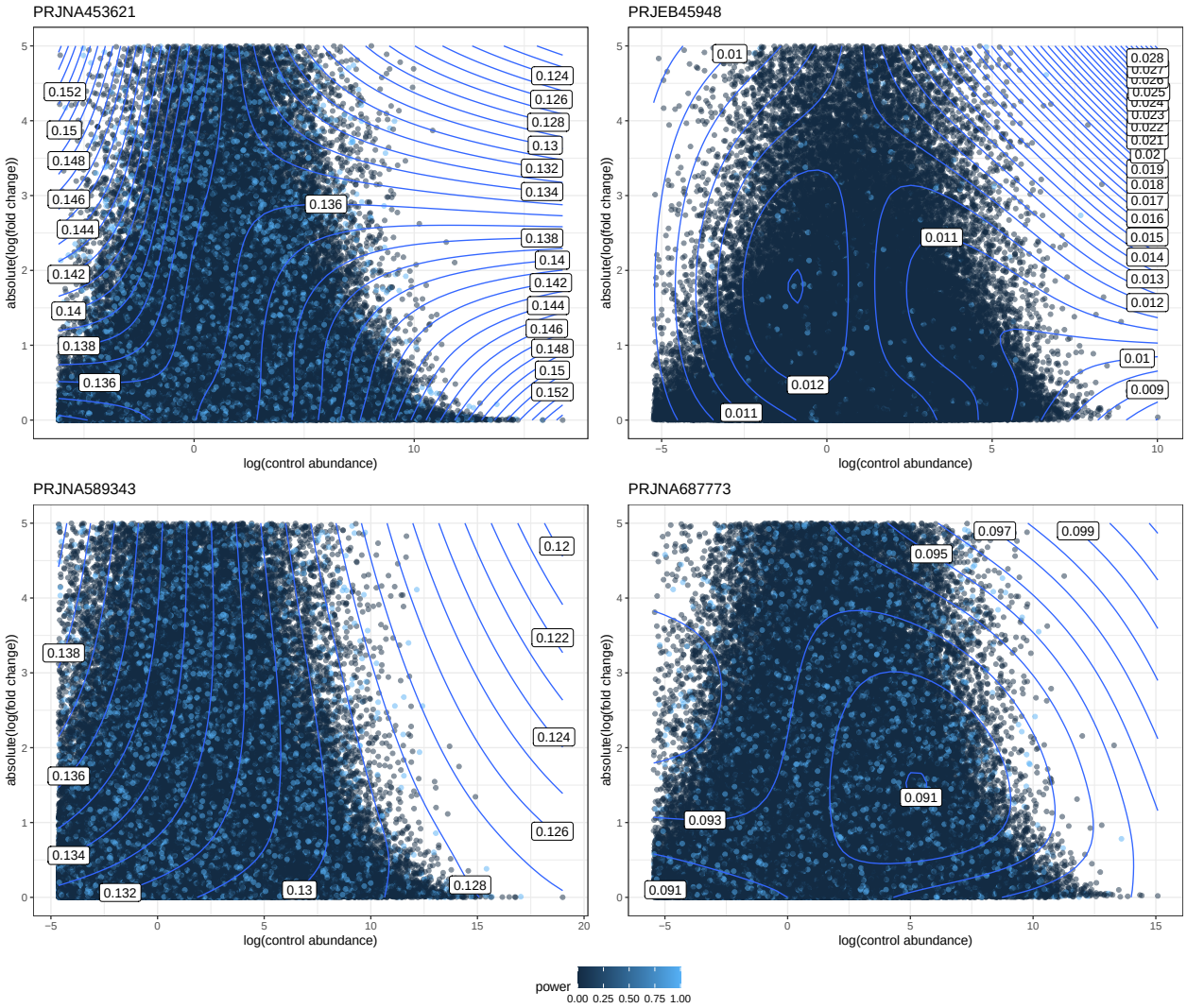

Figure 5: Contour plot showing changes in power in relation to control abundances and effect sizes

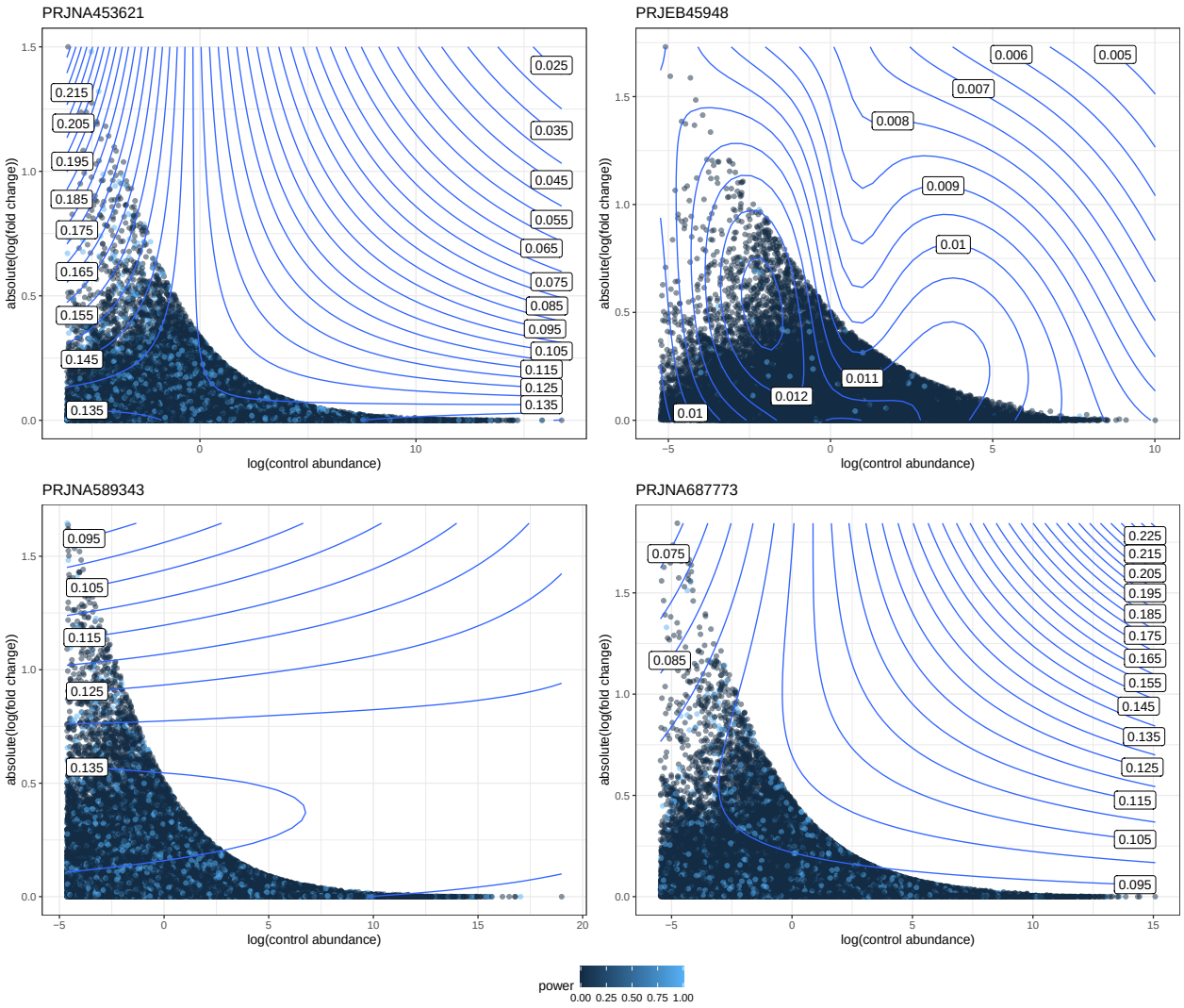

Figure 6: Contour plot showing changes in power in relation to control abundances and standardized effect sizes

1.7 Conclusion

In conclusion, this project seeks to understand the relationships among power, effect size and control abundances. Investigation of the relationship between control abundances and effect sizes show that there is an average effect size of zero across control abundances with variations in effect sizes being highest about midway of control abundances and lower variations at the high and low regions of control abundances (see figure 1). The variation in the effect sizes can be modeled as a quadratic function of control abundances. An While patterns between power and control and effects are unclear, a more precision in the estimation of the standard error may be desirable in understanding the current patterns seen in the datasets.

1.8 Future work

Future work on this project will focus on the following

1. Estimation of higher precision of the standard error and
2. Exploration of how the standard error influences power

References

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Nature Precedings*, 1–1.

Bäckhed, Fredrik, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, et al. 2015. "Dynamics and Stabilization of the Human Gut Microbiome During the First Year of Life." *Cell Host & Microbe* 17 (5): 690–703.

Brooks, Mollie E, Kasper Kristensen, Koen J Van Benthem, Arni Magnusson, Casper W Berg, Anders Nielsen, Hans J Skaug, Martin Machler, and Benjamin M Bolker. 2017. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling." *The R Journal* 9 (2): 378–400.

Brüssow, Harald. 2020. "Problems with the Concept of Gut Microbiota Dysbiosis." *Microbial Biotechnology* 13 (2): 423–34.

Chen, Eric Z, and Hongzhe Li. 2016. "A Two-Part Mixed-Effects Model for Analyzing Longitudinal Microbiome Compositional Data." *Bioinformatics* 32 (17): 2611–17.

Chen, Yu, Hui Fang, Chunyan Li, Guojun Wu, Ting Xu, Xin Yang, Liping Zhao, Xiaoyan Ke, and Chenhong Zhang. 2020. "Gut Bacteria Shared by Children and Their Mothers Associate with Developmental Level and Social Deficits in Autism Spectrum Disorder." *Msphere* 5 (6): e01044–20.

Kers, Jannigje Gerdien, and Edoardo Saccenti. 2021. "The Power of Microbiome Studies: Some Considerations on Which Alpha and Beta Metrics to Use and How to Report Results." *Frontiers in Microbiology* 12.

Kodikara, Saritha, Susan Ellul, and Kim-Anh Lê Cao. 2022. "Statistical Challenges in Longitudinal Microbiome Data Analysis." *Briefings in Bioinformatics* 23 (4): bbac273.

Lewis, James D, Eric Z Chen, Robert N Baldassano, Anthony R Otley, Anne M Griffiths, Dale Lee, Kyle Bittinger, et al. 2015. "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease." *Cell Host & Microbe* 18 (4): 489–500.

Ley, Ruth E, Peter J Turnbaugh, Samuel Klein, and Jeffrey I Gordon. 2006. "Human Gut Microbes Associated with Obesity." *Nature* 444 (7122): 1022–23.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 1–21.

Sze, Marc A, and Patrick D Schloss. 2016. "Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome." *MBio* 7 (4): e01018–16.

Xia, Yinglin, Jun Sun, Ding-Geng Chen, et al. 2018. *Statistical Analysis of Microbiome Data with r*. Vol. 847. Springer.

Zhang, Xinyan, Boyi Guo, and Nengjun Yi. 2020. "Zero-Inflated Gaussian Mixed Models for Analyzing Longitudinal Microbiome Data." *PloS One* 15 (11): e0242073.

Zhang, Xinyan, and Nengjun Yi. 2020. "NBZIMM: Negative Binomial and Zero-Inflated Mixed Models, with Application to Microbiome/Metagenomics Data Analysis." *BMC Bioinformatics* 21 (1): 1–19.

2 Multivariate Analysis of Longitudinal Microbiome data using Generalised Linear Mixed Negative Binomial Model

2.1 Abstract

Human microbiome is dynamic in nature. Understanding the dynamics in longitudinal microbiome data can help explain the mechanisms that underpin human health and diseases. However, longitudinal microbiome data analysis, whether 16S rRNA or metagenome shotgun, is difficult. Along with the difficulties presented by the characteristics of microbiome data such as sparsity, overdispersion, and compositionality, repeated measures introduce correlation between observations from the same individual at different time points. In an attempt to handle the problem of correlation, the majority of methods used in the literature to analyze longitudinal microbiome data only model one taxon at a single time point and do not account for the correlation structure between individual taxa across time. In this project, a Multivariate Negative Binomial Mixed Model (MNBMM) is proposed for analyzing longitudinal microbiome data. A special feature of

MNBMM is that it can handle the correlation structure of individual taxa over time. Fitting MNBMM to data, however, is very challenging because the number of parameters estimates for the unstructured covariance matrix grows quadratically with the dimensionality of the data. For atypical microbiome data with thousands of taxa, fitting this model is computationally expensive and almost impossible. A possible solution to implement a latent variable model for fitting a covariance matrix with fewer parameter estimates. In this project, a reduced rank method is used for fitting the data reduced rank method. The **glmmTMB** package in R provided a great functionality for fitting the reduced rank model. We demonstrate using from a simulation studies how implementation of the MNBMM works in **glmmTMB**.

2.2 Introduction

Longitudinal data arises when repeated measures are collected for the same subject. For example; DNA taking the microbiome of pregnant women repeatedly over multiple time points. The literature of longitudinal microbiome investigations has generally responded mainly to two queries: (1). to determine how microbiome abundances change over time between groups (for example cases versus controls) (Lewis et al. 2015; Bäckhed et al. 2015), and 2. to investigate the relationship between microbiome abundances and other factors, for instance, environmental factors, clinical outcomes, etc) (Kodikara, Ellul, and Lê Cao 2022; E. Z. Chen and Li 2016).
Special features of microbiome data such as variation in sequencing depth, high dimensionality, overdispersion, excess zeros and compositionality, poses challenges when analyzing microbiome data(Xia et al. 2018). Variations in sequencing depth makes comparison of reads across samples difficult. The problem of overdispersion renders statistical methods that assume means to be less or equal to variance ineffective. For example, the Poisson model which assumes mean to be equal to variance will underestimate the variance in microbiome data. The high dimensionality of microbiome data leads to numerical instabilities and convergence problems. Due to compositionality of microbiome data, statistical methods with normal distributional assumption cannot be used (Xia et al. 2018).
Longitudinal microbiome data analysis poses additional challenges. First, repeated measurements exhibit correlation between observations taken from the same subject at various points. Longitudinal data also often tend to have more missing data from a person or data not taken from everyone at all the time period studied (Kodikara, Ellul, and Lê Cao 2022). For instance, a longitudinal study collecting fecal specimens from patients may have missing data due to patients dropping out of the study during the period of the specimen collection.
Several models such as the Zero-inflated beta regression Model, Negative binomial mixed model, Block Bootstrap Method, Zero-inflated Gaussian mixed models, Fast zero-inflated negative binomial mixed model and Bayesian semi-parametric generalized linear model have been proposed for the studies of longitudinal microbiome data (Kodikara, Ellul, and Lê Cao 2022). Many researchers have proposed a linear mixed model as a means to handle the correlation exhibited by taxa from repeated measures (Zhang, Guo, and Yi 2020). In order to overcome the problem of correlations between observations from individuals across time, the majority of longitudinal microbiome research analyzes only one taxa at a time (E. Z. Chen and Li 2016). This approach, however, ignores the correlation structure between taxa from the individual subject. In this project, a Multivariate Negative Binomial Mixed Model (MNBMM) is proposed for the studying longitudinal microbiome data. MNBMM has the capacity to account for correlation instruction between individual taxa over time.

2.3 Research Goals

The goal of this project is to

1. model the relationship between taxa abundances over time and covariate variables of interest (for instance, clinical and environmental factors such as age, groups or disease status of patients, among others), while accounting for the correlation structure between individual taxa within individual subjects. The unstructured correlation matrix, however, requires too many parameter estimates and is computationally infeasible for a typical ASV or OTU tables with thousands of taxa. Thus, a reduced rank latent variable model's estimation of the structured correlation matrix will provide an understanding of the correlation between taxa on the community level.

2. model how microbiome abundances change over time between groups and determine taxa with differential abundances over time between groups.

2.4 Methods

2.4.1 Model Description

The count data denoted by Y_{ijt} is modeled by a negative binomial model defined as

$$\begin{aligned} Y_{ijt} &\sim NB(mean = \mu_{ijt}, dispersion = \alpha_{it}), \\ \boldsymbol{\mu} &= g^{-1}(E) \\ E &= X\beta + Zb,\ b \sim MVN(0, \Sigma) \end{aligned} \tag{7}$$

where i, j and t denote the i^{th} taxa, j^{th} sample and the t^{th} time, respectively, and $Y_{i,j,t}$ denotes the count data for the i^{th} taxa in the j^{th} sample at time t. The function g is the link function and $\boldsymbol{\mu}$ is a vector of means with entries μ_{ijt}. X and Z are the fixed and random effect model matrices respectively and β is the fixed effect parameter. The vector b is a multivariate deviate from a gaussian distribution with a variance covariance matrix Σ.
Given an OTU or ASV table of dimension n, the parameter estimates required for Σ are given by $n(n+1)/2$. Thus, the number of parameter estimates grows quadratically with n. For an OTU table with 10 taxa for instance, $55(= 10 * 11/2)$ parameter estimates will be required for Σ. Consequently, the typical microbiome data with thousands of taxa will require millions of parameter estimates, which is computationally infeasible. Additionally, the unstructured covariance matrix Σ is often singular due to linear independence of the columns. This results in numerical instability and convergence problems when fitting the model. A possible remedy is to use a latent variable model, to specify a low dimensional latent variable that describes the full correlation structure. This project applies a reduced rank latent variable model for estimating the structured covariance matrix. The reduced rank model expresses the columns of Σ by a set of $r < n$ latent variables. The number of parameters required to be estimated for the reduced rank model with r latent variables is given by $nr - \binom{r}{2}$, hence, reducing the number of parameters to be estimated significantly to a linear function of n. The **glmmTMB** package (Brooks et al. 2017) in R provides a flexible framework for fitting mixed models with a range of response distributions and also has the capability to fit reduced rank variance structure. The **glmmTMB** package is used in this project for the estimation of the reduced rank model.

2.5 Application

2.5.1 Data Simulation

Using the simulate function in **lme4**, we generated count data from a generalized linear mixed model with a negative binomial distribution. The simulation's parameters are displayed in Table 1. The plot of the trajectory for three taxa from the simulated data is shown in Figure 7. A rank reduced rank random slope model with subjects as the grouping variable, taxa as the random effect variable, and groups and the interaction between group and time as the fixed effect variable was fitted to the data. The optimal dimension of 3 for the reduced rank was determined by comparing the AICs of the model fitted to dimensions ranging from 2 to 10.

```
fit_list <- lapply(2:10,
                   function(d) {
                     fit.rr <-
                     glmmTMB(count ~ group*time+ rr(-1 + taxa|subject,d = d),
                                                data = dd, family = nbinom2)
                       })

# compare fits via AIC
aic_vec <- sapply(fit_list, AIC)
aic_vec - min(aic_vec, na.rm = TRUE)
```

Table 1: Parameter values used in data simulation

Effects	**Other parameters**
Intercept = 1	Number of subjects = 20
Group = 0.2	Number of Taxa = 100
Time =0.1	Length of Time = 10
Group*Time = 0.2	Number of Groups = 2
Dispersion parameter = 2	Random effect parameter (θ) = Identity

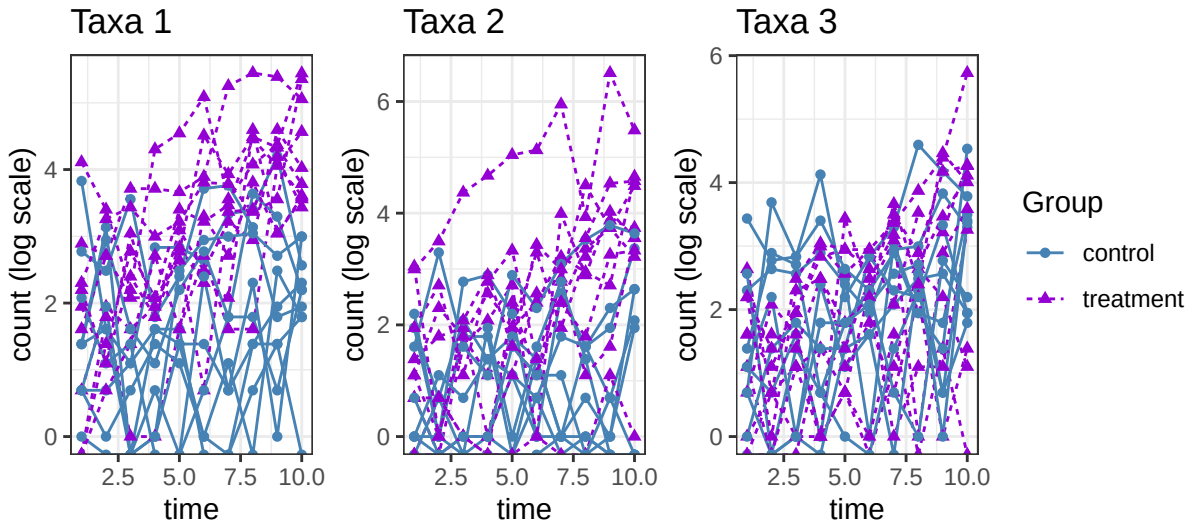

Figure 7: Some simulated taxa profiles with time, group and group * time effects and 0.2 dispersion

2.6 Results

2.6.1 Comparison with the Univariate Negative Binomial Mixed Model

The Univariate Negative Binomial Mixed Model (UNBMM) has often been used in the literature of longitudinal microbiome studies (Kodikara, Ellul, and Lê Cao 2022) . In this section we compare the effectiveness of UNBMM and MNBMM in predicting the fixed effect parameter. The UNBMM was fitted using the **glmm.nb** package developed by which is commonly used for longitudinal microbiome studies in the literature (Zhang and Yi 2020). Figure 8 shows the mean square errors of the fixed effect parameter estimated from both models for 50 simulations. For simplicity, 10 taxa are used for this comparison study. The plot shows lower errors from the MNBMM model. Consequently, accounting for correlations between the individual taxa provides better estimates of the fixed effect parameters.

2.6.2 Model Prediction

This section shows the mean squared error of our model prediction for individual taxa for 50 simulations. The points on the plot in figure 9 represent that average mean squared errors for individual taxa across time. Here, we consider 10 taxa. On average the mean squared error for these over taxa is approximately 4.96.

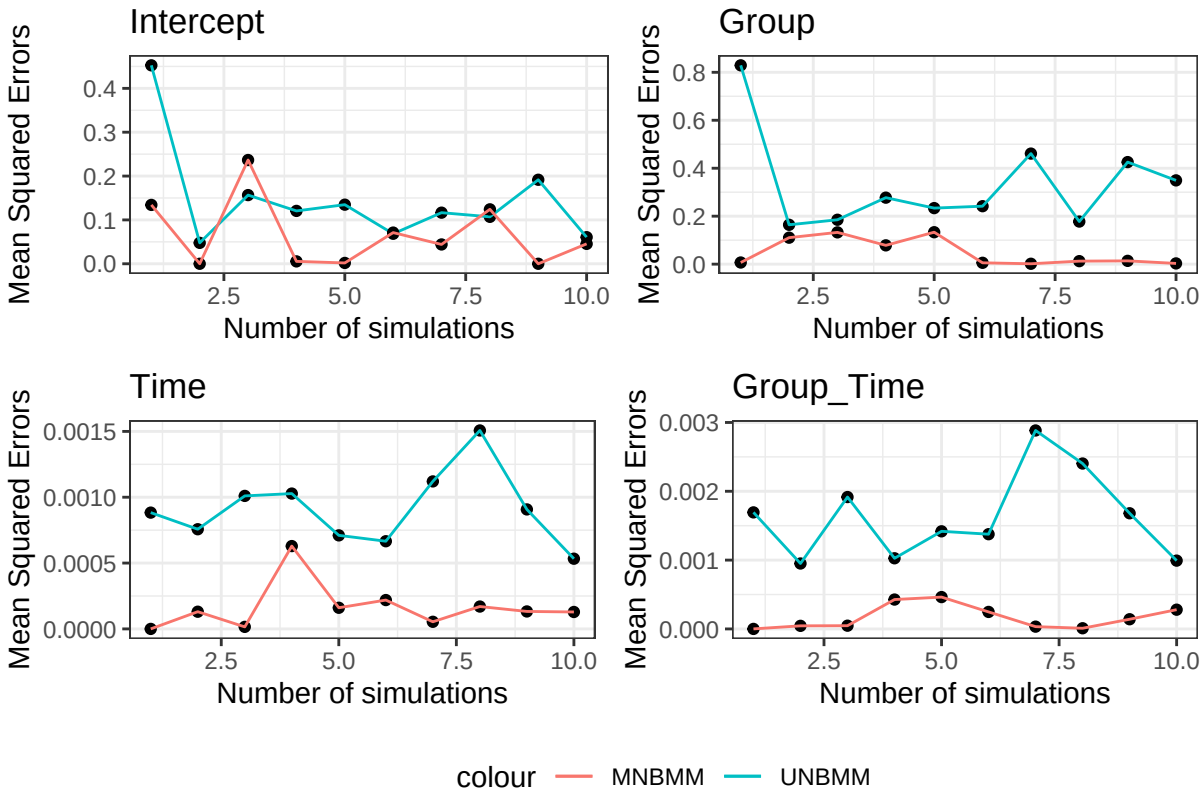

Figure 8: Comparison of the mean squared error for the fixed effect estimates from the Univariate and Multivariate Negative Binomial Mixed Models

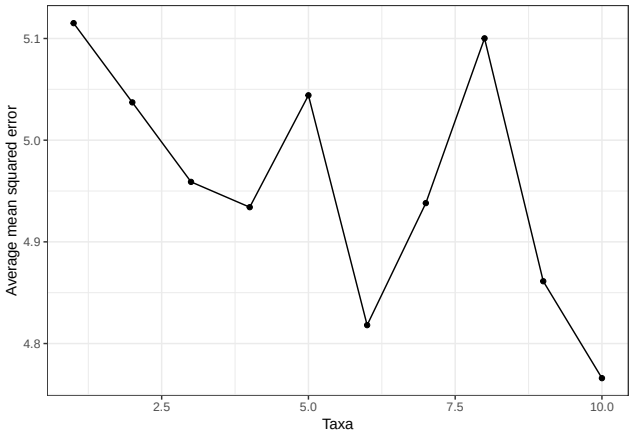

Figure 9: Average mean squared error for prediction of mean abundances of individual taxa across time

2.7 Conclusion

In this project, Multivariate Negative Binomial Mixed Model (MNBMM) was proposed for analyzing longitudinal microbiome data. The model is capable of handling correlations between taxa from individuals over time. The reduced rank component of MNBMM allows fitting of this model to microbiome data of large dimensions. MNBMM also has the capability to handle overdispersion in microbiome data. This project demonstrates how to fit the model in the **glmmTMB** package in R using simulation studies. We show using the simulation studies that the models provide better estimates of the fixed effect parameters compared to the traditional univariate Negative Binomial Mixed Model applied in the literature of longitudinal microbiome studies.

2.8 Further development

A limitation of MNBMM is that it does not handle zero inflation found in microbiome data. Additionally, MNBMM cannot handle microbiome data presented in propositions and does not handle compositionality in microbiome data. Further work on this project will extend this model to include zero inflation component so as to account for the zero inflation seen in microbiome data. A new model with a framework similar to MNBMM, for example a Multivariate Gaussian Mixed Model will be developed to handle microbiome data presented as proportions.

References

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Nature Precedings*, 1–1.

Bäckhed, Fredrik, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, et al. 2015. "Dynamics and Stabilization of the Human Gut Microbiome During the First Year of Life." *Cell Host & Microbe* 17 (5): 690–703.

Brooks, Mollie E, Kasper Kristensen, Koen J Van Benthem, Arni Magnusson, Casper W Berg, Anders Nielsen, Hans J Skaug, Martin Machler, and Benjamin M Bolker. 2017. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling." *The R Journal* 9 (2): 378–400.

Brüssow, Harald. 2020. "Problems with the Concept of Gut Microbiota Dysbiosis." *Microbial Biotechnology* 13 (2): 423–34.

Chen, Eric Z, and Hongzhe Li. 2016. "A Two-Part Mixed-Effects Model for Analyzing Longitudinal Microbiome Compositional Data." *Bioinformatics* 32 (17): 2611–17.

Chen, Yu, Hui Fang, Chunyan Li, Guojun Wu, Ting Xu, Xin Yang, Liping Zhao, Xiaoyan Ke, and Chenhong Zhang. 2020. "Gut Bacteria Shared by Children and Their Mothers Associate with Developmental Level and Social Deficits in Autism Spectrum Disorder." *Msphere* 5 (6): e01044–20.

Kers, Jannigje Gerdien, and Edoardo Saccenti. 2021. "The Power of Microbiome Studies: Some Considerations on Which Alpha and Beta Metrics to Use and How to Report Results." *Frontiers in Microbiology* 12.

Kodikara, Saritha, Susan Ellul, and Kim-Anh Lê Cao. 2022. "Statistical Challenges in Longitudinal Microbiome Data Analysis." *Briefings in Bioinformatics* 23 (4): bbac273.

Lewis, James D, Eric Z Chen, Robert N Baldassano, Anthony R Otley, Anne M Griffiths, Dale Lee, Kyle Bittinger, et al. 2015. "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease." *Cell Host & Microbe* 18 (4): 489–500.

Ley, Ruth E, Peter J Turnbaugh, Samuel Klein, and Jeffrey I Gordon. 2006. "Human Gut Microbes Associated with Obesity." *Nature* 444 (7122): 1022–23.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 1–21.

Sze, Marc A, and Patrick D Schloss. 2016. "Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome." *MBio* 7 (4): e01018–16.

Xia, Yinglin, Jun Sun, Ding-Geng Chen, et al. 2018. *Statistical Analysis of Microbiome Data with r*. Vol. 847. Springer.

Zhang, Xinyan, Boyi Guo, and Nengjun Yi. 2020. "Zero-Inflated Gaussian Mixed Models for Analyzing Longitudinal Microbiome Data." *PloS One* 15 (11): e0242073.

Zhang, Xinyan, and Nengjun Yi. 2020. "NBZIMM: Negative Binomial and Zero-Inflated Mixed Models, with Application to Microbiome/Metagenomics Data Analysis." *BMC Bioinformatics* 21 (1): 1–19.