Power Analysis of Microbiome Data

Michael Agronah

January 20, 2023

Contents

Summary

This document constitute 2 parts. Part 1 is a project the investigates the relationship between power, effect sizes and abundances in differential microbiome abundance studies. Part 2 focuses on investigating the modeling of the trends in longitudinal microbiome data and the relationship between longitudinal microbiome data with variables of interest such as the disease status of patients, and environmental factors.

1 Past 1 Power Analysis of Microbiome Data

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. Control is used to refer to the log2 control abundances.

Current results from the project show a positive relationship between standardized effect size and power. An R package will be developed from this project where a new dataset can be fed into this package to explore the power, effect and abundance relationships.

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 procedure. The standard process in hypothesis testing is to first define a null hypothesis, choose a significant level and reject the null hypothesis whenever p-value is below the significant level or fail 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 (resulting in what is known as the Type 1 error or the false positive error) and second, rejecting the null hypothesis when the p-value is less than the significant level by chance (known as Type 2 error false negative error).

Hypothesis testing depends on 4 parameters: (i) the effect size, which is 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 will when it is actually true. In general, power, sample sizes, significance level and effect size are all 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 what effect sizes and the power of the studies.

The majority of microbiome studies 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 in the literature of microbiome studies (Brüssow 2020). Early research on the relationship between the gut microbiota of obese and non-obesity groups, for example, 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, however, 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 generally 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. This makes is difficult to make informed decisions for conducting a power studies 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 new dataset can be to explored to determine the relationships between power,effect size and abundance in the context of differential abundance studies.

2.1 Research goals

2.1.0.1 Project Contributions: This project contributes to the current literature on power analysis in microbiome studies in the following ways

1. develop a framework for estimating the power and effect sizes at the taxon level and describing the relationship between effect sizes and power and abundances for differential abundance studies that use the Negative Binomial Model. Two commonly used packages that implement the Negative Binomial Model are Deseq2 and edgeR. For simplicity, this study focuses on the implementation in the Deseq2 package. However, the same analysis can be repeated using the edgeR package as well. This will help researchers to bla bla bla

2. develop and R package with visualization tools for power calculations and effect size at these. This will help researchers to bla bla bla...

The project will answer the following questions.

1. What is the relationship between effect size, power and abundances? For example, what is the average power and effect size for abundances in the range.

2.How many sample sizes ste needed to detect a given power and at a given effect size.

3. For a given effect size will determine the number of taxa with power greater than a desire value

2.1.0.2 Challenges: To do power analysis, one needs to know the effect size and the sample size. A typical approach is to get the effect sizes reported in the literature for the power. The accuracy of these effect sizes, however, is unclear. Often these effect sizes of microbiome studies are not reported [@]. ...showed that the effects from these studies are biased .

Mm reviewed... papers . In this studies, 100 publications reviewed did not report effect sizes from the studies

Therefore it is unclear what an average effect size for differential abundance analysis in microbiome research is.

Power is more likely overestimated in differential abundance studies in the literature and the relationship between power, effect size and abundances is unknown.

5o the nest of pur knowledge, no studies exist that explores the effect sizes and power for different ranges of abundances.

2.2 Research Design and Method

2.2.1 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 children with neurotypical or typical development are the control groups. Adaptor and primer sequences were removed using "cutadapt" function implemented in a bash script. The trimmed sequence from the "cutadpt" 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 (Chen et al. 2020).

2.2.2 Models description

The null hypothesis (H_0) and the alternative hypothesis (H_a) for testing differences between 2 groups (control and treatment) is formulated by $H_0 : \mu_{control} - \mu_{treatment} = 0$ and $H_a : \mu_{control} - \mu_{treatment} \neq$ respectively, where $\mu_{control}$ and $\mu_{treatment}$ denote the means counts in control and treatment groups respectively. Effect sizes,a measure of the tdifference in the mean can then be tested. The negative binomial distribution has often 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

$$K_{ij} \sim NB(mean = \mu_{ij}, dispersion = \alpha_i), \tag{1}$$

$$\mu_{ij} = g^{-1}(E_{ij}) \tag{2}$$

$$E_{ij} = \sum_r x_{jr}\beta_{ir}, \tag{3}$$

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 $VarK_{ij} = \mu + \alpha\mu^2$. 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 is stated in the paper by Love, Huber, and Anders (2014) and by Anders and Huber (2010).

2.2.3 Modelling the relationship between control abundances and effect sizes

Plots of the control abundances and effect sizes are shown in figure 2. The plots display an average effect size of zero. According to the plot, variations around the smooth curve are greatest in the middle and lowest at the ends of the smooth curves. A quadratic function can be used to describe the variance of the effect sizes as a function of control abundances. A scale location plot shown in figure 1 validates the quadratic relationship for the variance. A skewed normal distribution and a truncated Cauchy distribution, respectively, can be used to approximate the distribution of control abundances and effect sizes. Figure 3 and figure 4 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 defined with zero location and scale parameters defined by a quadratic function of control abundance, as shown in equation by 4.

$$y_i = Cauchy(location = 0, scale = \gamma_i), \gamma_i = \exp(k_0 + k_1 x_i + k_2 {x_i}^2), \tag{4}$$

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.

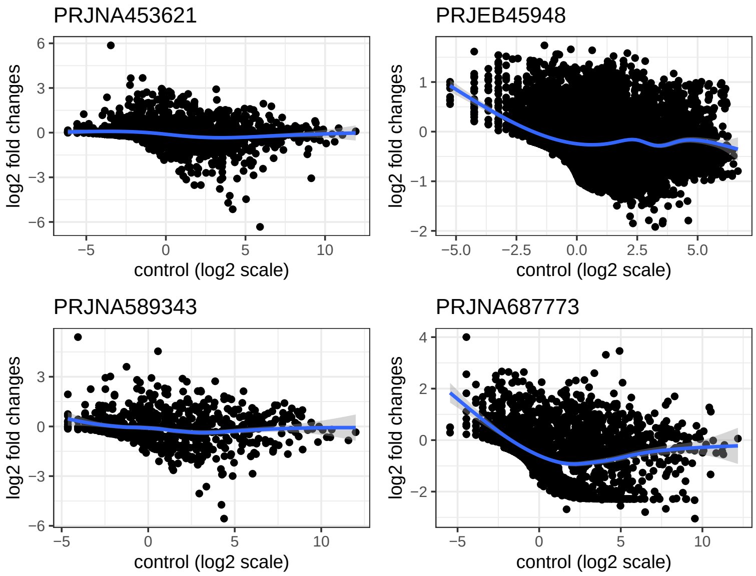

Figure 1: Plot of control against effect sizes for datasets with assession numbers PRJNA453621, PRJEB45948, PRJNA589343 and PRJNA687773

2.2.4 Estimation of standard errors

2.2.5 The delta method

Write up on the delta method. The delta method will be used to estimate the standardised error for the control abundance and the effect sizes defined

The delta method is a method for claucluationg the assymptotic bal bal bale. For a random variable X and parameter $\theta, \hat{\theta}$ and any smooth function f with the property that $f'(\theta)$ exists and is non-zero valued, the

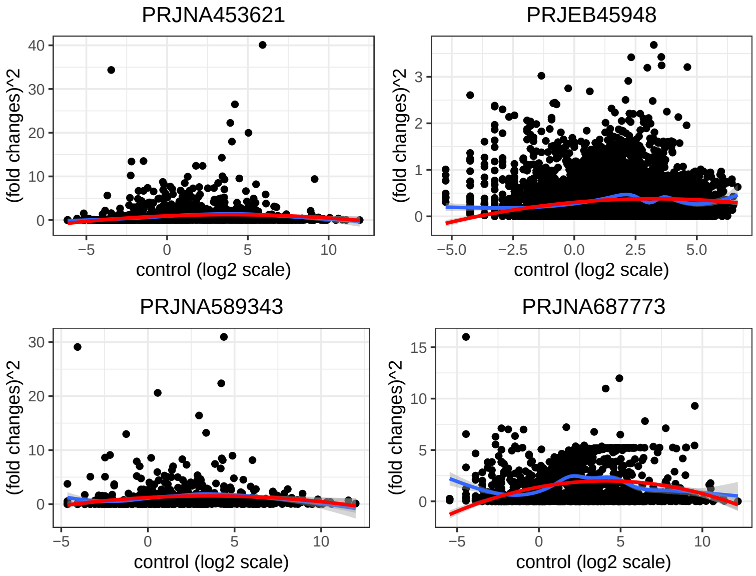

Figure 2: Scale-location plot of control against squared fold changes for datasets with assession numbers PRJNA453621, PRJEB45948, PRJNA589343 and PRJNA687773

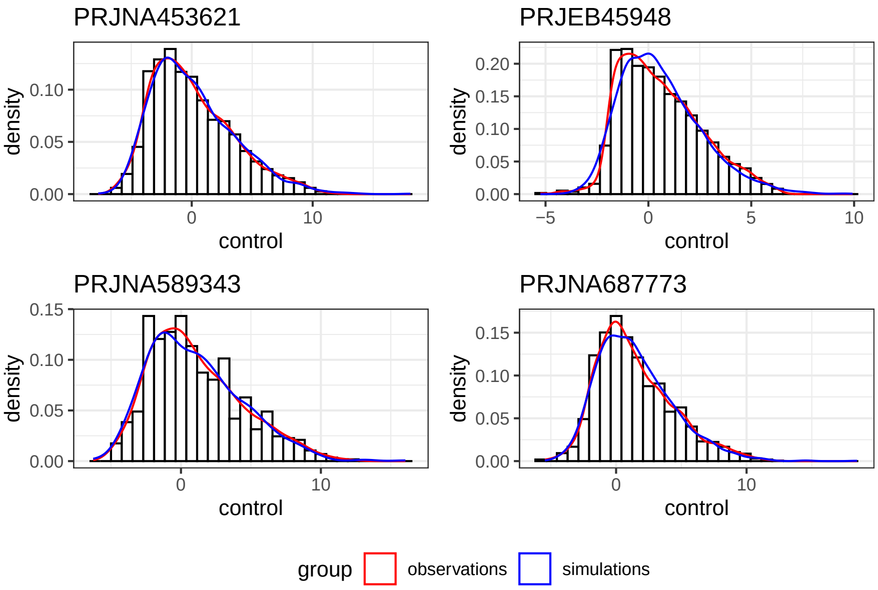

Figure 3: Histogram and density plots for control abundances and density from a simulated sample from fitting a truncated normal distribution

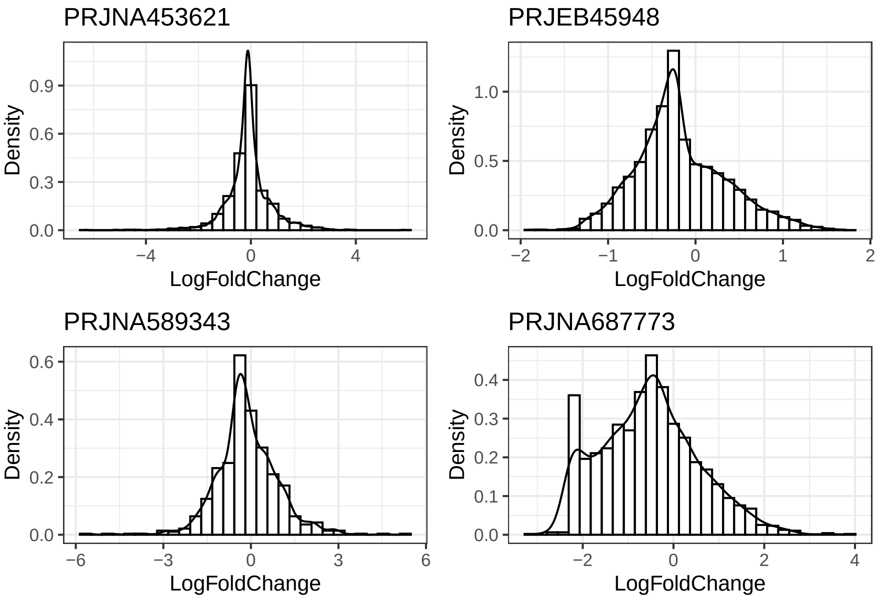

Figure 4: Histogram and density plots for control abundances and density from a simulated sample from fitting a Cauchy distribution to the data

delta method states that

$$\frac{\hat{\theta}-\theta}{S\hat{E}_\theta} \to N(0,1),\ then\frac{f(\hat{\theta})-f(\theta)}{f'(\theta)S\hat{E}_\theta} \to N(0,1)$$

and tha asymptotic mean and asymptotic standard error of $f(\hat{\theta})$ is $f(\theta)$ and $f('\hat{\theta})S\hat{E}(\theta)$ respectively.

Let C and T denote the control and treatment groups respectively. The standard error of fold changes is define be estimated as follows:

$$\mathbf{var}[\log_2\left(\frac{T}{C}\right)] = \mathbf{var}[\log_2(T)] + \mathbf{var}[\log_2(C)]$$

From Taylor series expansion $\mathbf{var}[f(X)] = \mathbf{var}[f(\mu_X + X - \mu_X)] \approx \mathbf{var}[f(\mu_X) + f'(\mu_X)(X-\mu_X) + \frac{f''(\mu_X)(X-\mu_X)}{2} + .. = f'(\mathbb{E}[X])^2\mathbf{var}[X] + \frac{f'(\mathbb{E}[X])^2\mathbf{var}[X]}{4}$,

$$\begin{aligned}
for f(X) = \log_2(X), f' &= \frac{1}{x\ln 2}\\
\mathbf{var}[\log_2\left(\frac{T}{C}\right)] &= \mathbf{var}[\log_2(T)] + \mathbf{var}[\log_2(C)]\\
&\approx \frac{1}{T\ln 2}\mathbf{var}[T] + \frac{1}{C\ln 2}\mathbf{var}[C]\\
\mathbb{SE}[\log_2\left(\frac{T}{C}\right)] &= \sqrt{\frac{1}{T\ln 2}\mathbf{var}[T] + \frac{1}{C\ln 2}\mathbf{var}[C]}
\end{aligned}$$

–>

–>

–> –> –> –> –> –>

3 Conclusion

1.Fit control abundace, no show the plot for the power. That is bettwer 2. Show the plots for the power

4 References

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

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

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.

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.