

# Power Analysis of Microbiome Data

Michael Agronah

December 8, 2022

## Contents

<b>1</b>	<b>Investigating the Power of Differential Abundance Microbiome Studies</b>	<b>1</b>
1.1	Abstract . . . . .	1
1.2	Introduction . . . . .	2
1.3	Research goals . . . . .	2
1.4	Research Design and Methodology . . . . .	3
1.4.1	Data collection and processing . . . . .	3
1.5	Method . . . . .	3
1.5.1	Model description . . . . .	3
1.5.2	Investigating the relationship between control abundances and effect sizes . . . . .	3
1.6	Results . . . . .	5
1.7	Conclusion . . . . .	9
1.8	Future work . . . . .	9
	<b>References</b>	<b>10</b>

## 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 log<sub>2</sub> fold changes and control is used to refer to log<sub>2</sub> 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 ( $\alpha$ ): 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 Type 1 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 analysis 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** pipeline which involves filtering and trimming, error estimation, denoising, merging paired reads and chimeras removal (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

$$H_0 : \mu_C - \mu_T = 0 \quad (1)$$

$$H_a : \mu_C - \mu_T \neq 0, \quad (2)$$

where  $\mu_C$  and  $\mu_T$  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

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

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

$$E_{ij} = \sum_r x_{jr} \beta_{ir}, \quad (5)$$

where  $g$  is a link function,  $\beta_{ir}$  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  $\text{Var}K_{ij} = \mu_{ij} + \alpha\mu_{ij}^2$ . In this project, the estimating procedure implemented in the **DESeq2** package in R is used for estimating  $\beta_{ir}$ ,  $\mu_{ij}$  and  $\alpha_i$ . 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 assessment 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.

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

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.

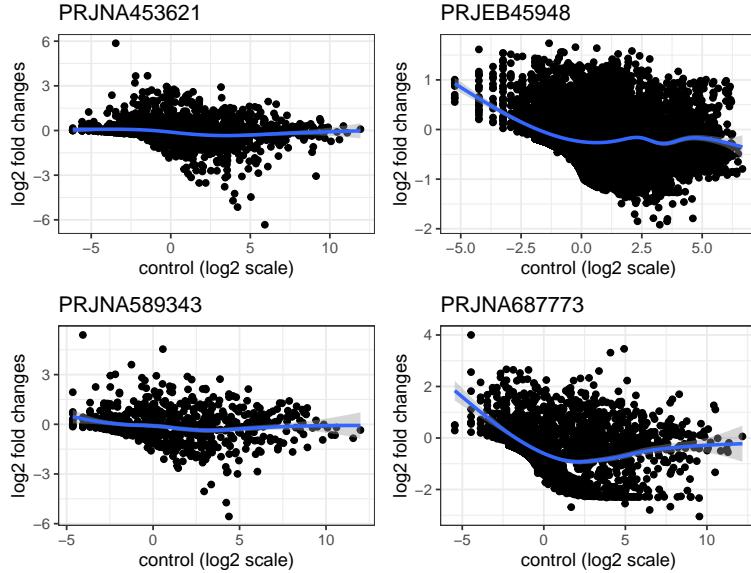


Figure 1: Plot of control against effect sizes for datasets with assessment numbers PRJNA453621, PRJEB45948, PRJNA589343 and PRJNA687773

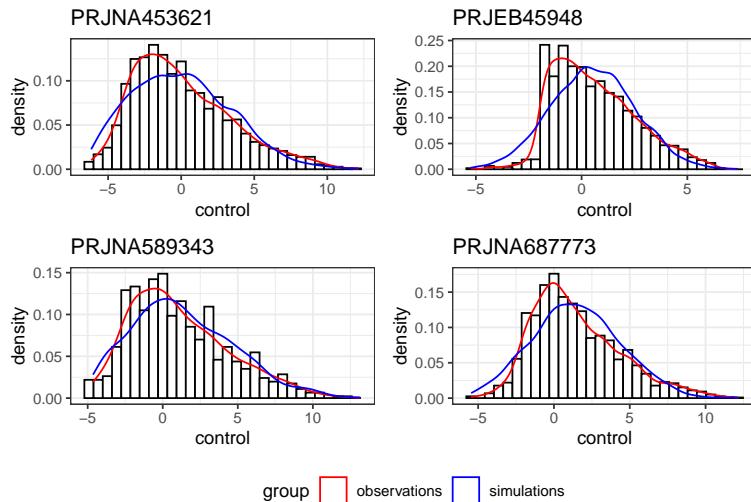


Figure 2: Histogram and density plots for control abundances and density from a simulated sample from fitting a truncated normal distribution

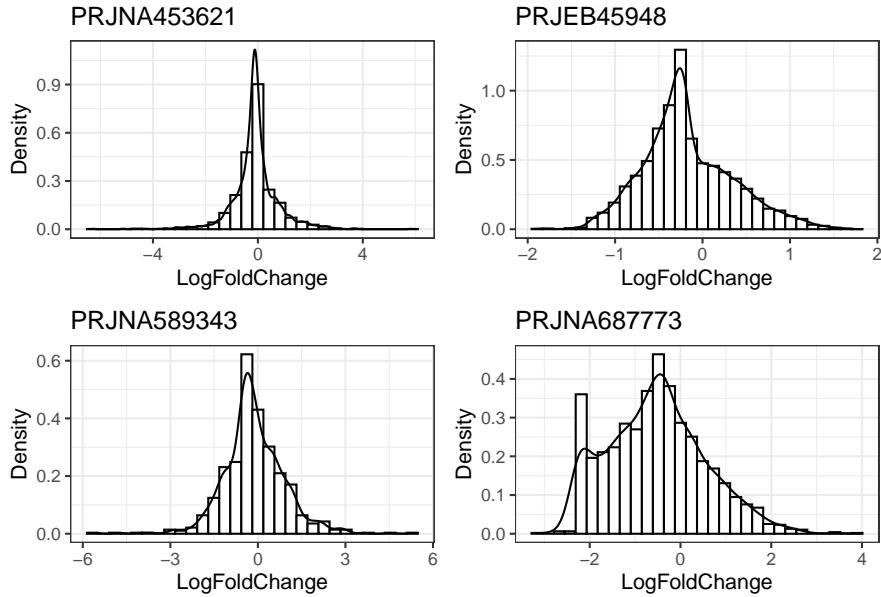


Figure 3: Histogram and density plots for effect sizes

## 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.

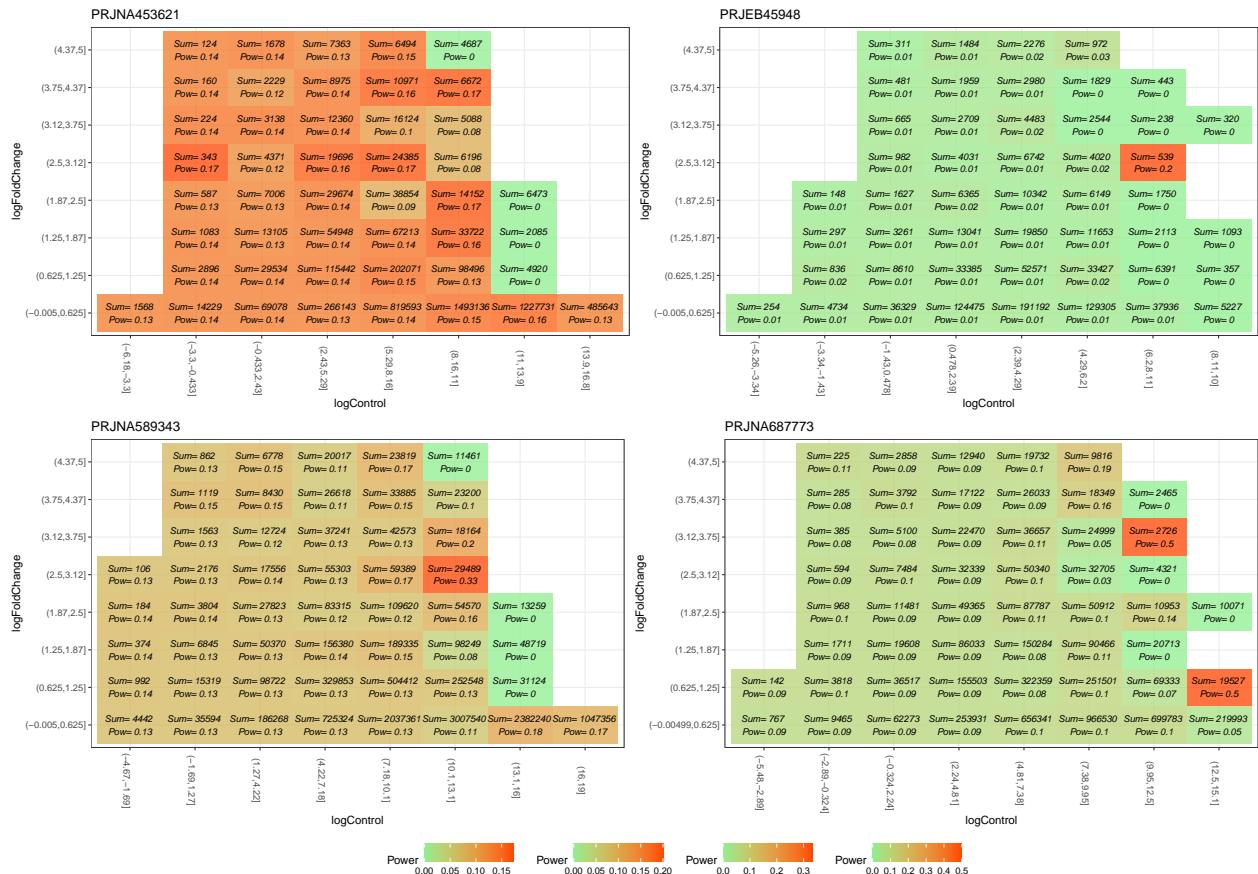


Figure 4: Heatmap showing the power and total control abundances for various ranges of control and effect sizes

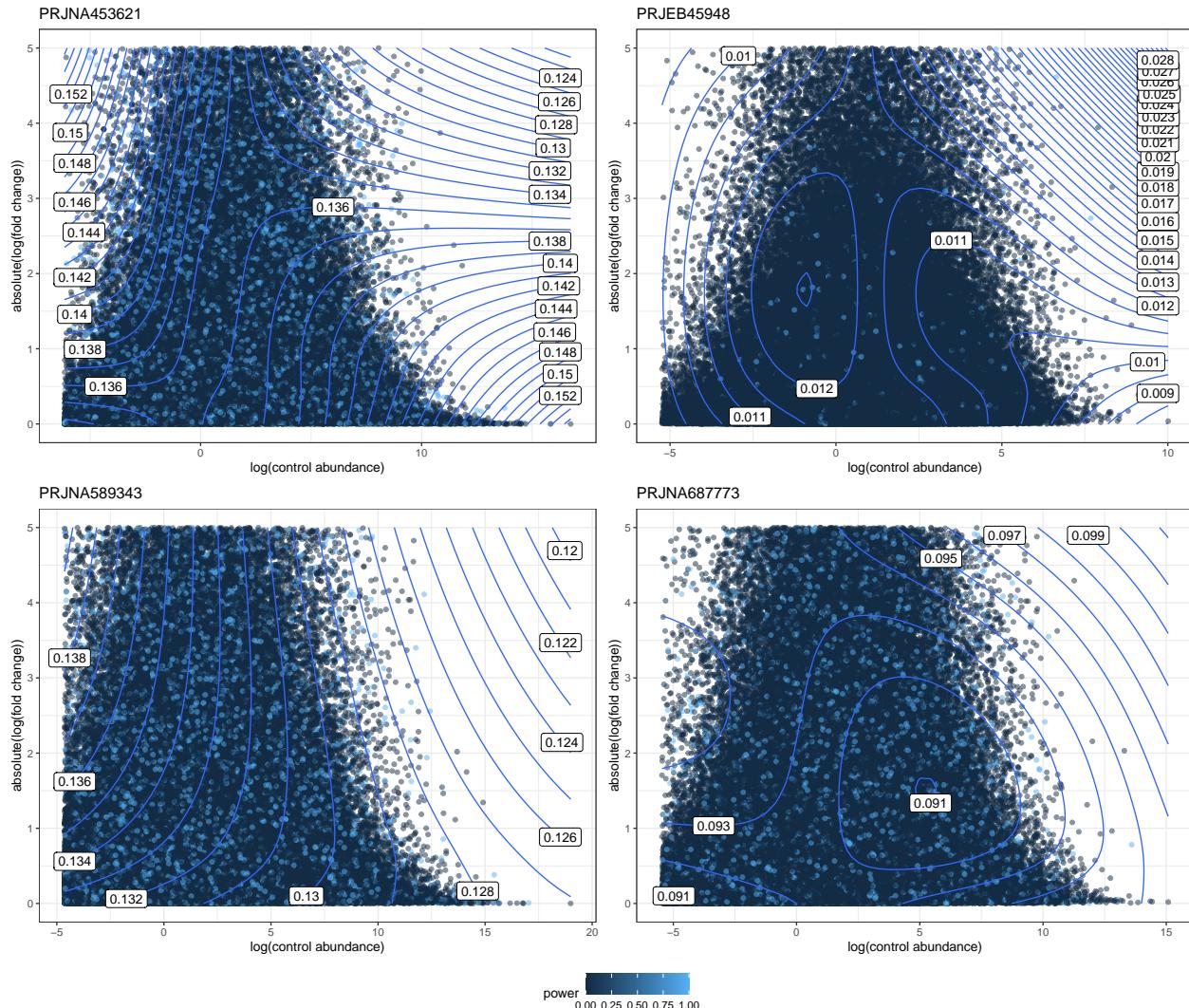


Figure 5: Contour plot showing changes in power in relation to control abundances and effect sizes

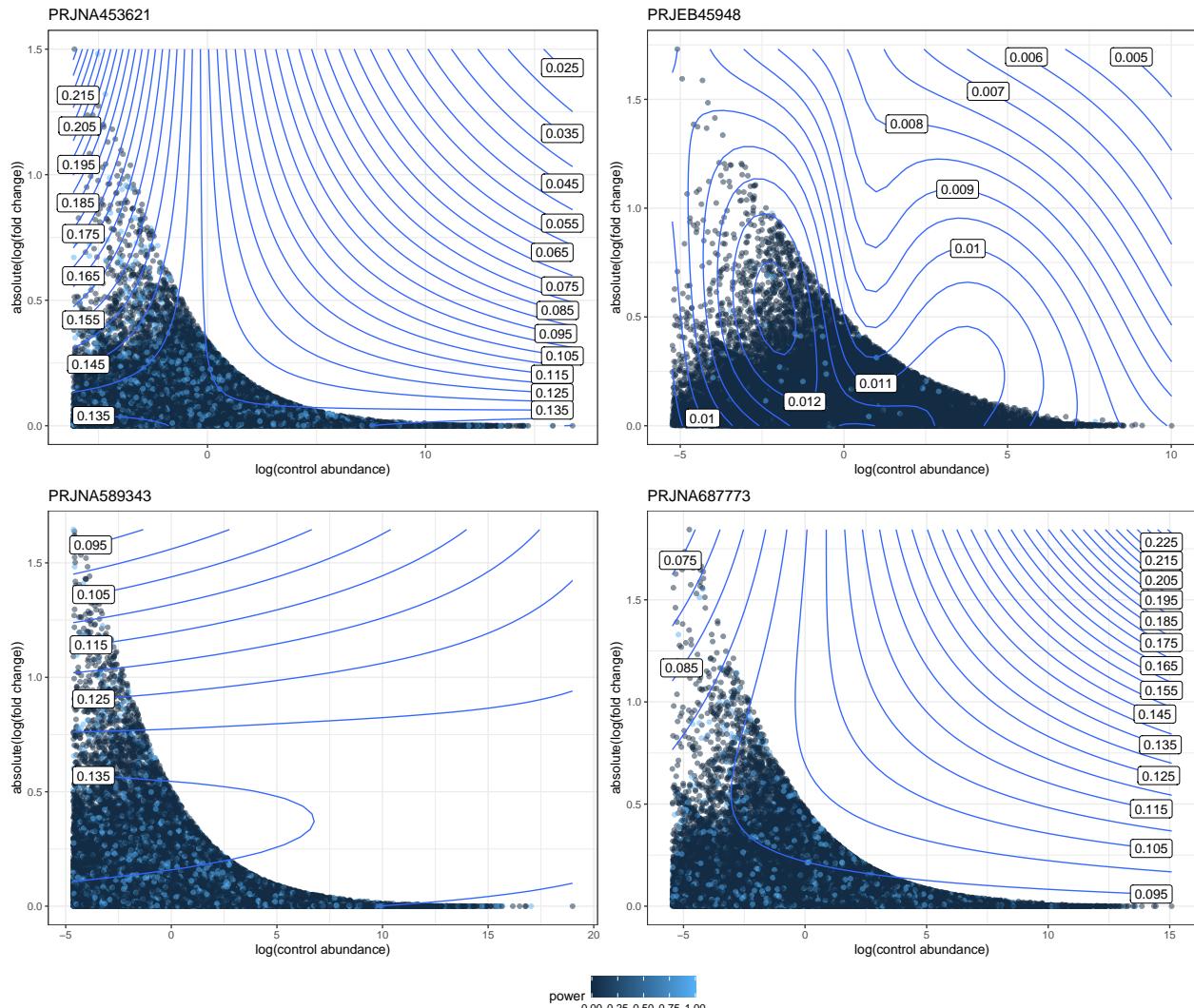


Figure 6: Contour plot showing changes in power in relation to control abundances and standardized effect sizes

## 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. 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.
- 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.