UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Modelling Longitudinal Effects of Intervention on Microbiome Experiments

Mary Louise Dela Cruz

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

MENTOR :

Mevrouw Thi Huyen NGUYEN

2023
2024

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Modelling Longitudinal Effects of Intervention on Microbiome Experiments

Mary Louise Dela Cruz

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

MENTOR :

Mevrouw Thi Huyen NGUYEN

Acknowledgements

I would like to express my deepest appreciation to my thesis supervisors Prof. Dr. Ziv Shkedy and Thi Huyen Nguyen for their invaluable support and patience in sharing their expertise and knowledge, which were essential for the completion of this paper.

I am also grateful for the faculty and staff of the Data Science Institute of Universiteit Hasselt. The knowledge that I have gained through various courses have greatly contributed to the success of this research.

Many thanks to the VLIR-UOS for providing me the opportunity to pursue the Masters of Statistics and Data Science program at Universiteit Hasselt.

Special thanks to the Filipino community, my Ates and Kuyas, my fellow VLIR-UOS scholars, classmates, Team Bioinformatics - Marina, Tom and Anthony, to my dear friends Jessa, Nong Roel and MJ for their unending moral and emotional support throughout this journey.

To my parents Luis and Ma. Joyce Dela Cruz, my family and friends, who have been my inspiration and source of strength in overcoming the challenges along the way, my deepest gratitude.

Lastly, to our dear Lord, who have showered His wisdom, guidance and blessings and for His continuous provisions that have carried me through.

Contents

Abstract

Introduction: The role of medical research in contributing to various health aspects through disease diagnosis, disease prevention and development of treatment and interventions, has addressed public health issues and improved health care quality. With the utilization of a wide range of medical biological data sets, medical experiments were conducted to assess the efficacy of developed treatments and interventions. It was found that the human microbiome, specifically the gut microbiome alongside wih the fecal and intestinal microbiome can be linked to both health and existence of diseases, and the administration of treatments can change the microbial composition of a patient or subject.

Objectives: The main objective of this project is to investigate the effects of the treatment on the microbiome data through examining the longitudinal behavior of the microbiome and to explore changes on the microbiome in response to the treatment.

Methodology: There were two datasets that used in this study. The data from the CERTIFI Study which consists of the fecal microbiome of patients with Crohn's Disease, and a set of data from the T1D Study which is composed of fecal and intestinal microbiomes of non-obese diabetic (NOD) mice. The analyses were performed kingdom, family and OTU levels. The alpha diversity measures used as the response were the Shannon index at kingdom and family levels and the relative abundance at the OTU level. To address the objective of the study, a linear mixed model was fitted at each taxonomic level and for each family and OTU.

Results: The findings revealed that for the CERTIFI study, at kingdom level, the effect of the treatment on the microbiome was significant at weeks 0 and 6. At the family level, *Ruminococcaceae* exhibited significant difference in the microbiome at weeks 0 and 4. At OTU level, only *Otu0002* showed significant result at week 0. Besides, for the T1D study, at kingdom level there exists a significant difference in the microbiome at weeks 3 and 5. Significant results were observed from family *Lachnospiraceae* at weeks 3 and 5. At the OTU level, only OTUs *d963b59f19db6517a9f26908f684545d* yielded significant results at weeks 5 and 7.

Discussion: The results of the analysis indicated that the treatment has altered the diversity and composition of the microbiome. Some phyla have decreased or increased after administering the treatment. It was also reflected in the results that the impact of treatment varies over time. At each timepoint, the differences in the microbiome between the treatment groups are not the same.

Keywords: Linear Mixed Model, Fecal microbiome, Intestinal microbiome, Crohn's Disease, Type 1 Diabetes

1 Introduction

Medical research has continuously driven developments that have significantly enhanced public health in various aspects. One of these contributions can be attributed to disease diagnosis, disease prevention and development of treatments and interventions for a wide range of diseases. This is not limited to developing treatments that work efficiently but also treatments that works for a particular patient's condition. With this, effectiveness of a treatment can be provided and adverse effects can be identified. Additionally, medical research has greatly contributed in addressing public health issues and improving health care quality through treatment developments. This, in turn, plays a significant role in improving the economy [3]. Alongside this development of research, an increasing availability of large medical biological data sets are being utilized to identify characteristics of patients that corresponds to healthy and pathological conditions [13].

Microbiota or often called as microbiome is a collection of microbes which also are ranges of bacteria, archaea, fungi, microbial eukaryotes and viruses. The human microbiota highly contributes in several functions in the body which includes aiding in food digestion, vitamin production, immune system regulation and protection against pathogenic microbes that cause diseases [7]. It is particularly important to note that microbiomes can be linked to both health and disease. Healthy and unhealthy microbiomes, distinguished by their microbial characteristics, are identified through various technologies and methodologies. Commonly, experiments and researches have utilized the gut microbiome, composed of bacteria and small portion of nonbacterial microbes. The microbiomes found in other parts of the body also play a crucial role for both disease detection and protection. Since it is known that not all people share the same microbiome ecology, extreme caution must be taken into account, considering the differences of the microbiome composition and the blood tests for different individuals. An additional consideration in microbiome testing is the correct method of sampling for analysis, in which in the present remains arguable. Also, in terms of sample, microbiomes are mostly accessed through stool samples as these samples are mostly beneficial and utilized in studying diseases in cross-sectional and also longitudinal studies [1]. Specifically, microbiomes can aid in determining the efficacy of a treatment for several diseases.

Among the several diseases is Type 1 Diabetes. This is a chronic condition that is characterized by a lifelong absolute insulin deficiency, due to T-cell mediated destruction of the pancreatic β-cells. The prevalence, incidence and the associated mortality with Type 1 Diabetes is a major concern globally, even in high-income countries, due to its association with several conditions, including reduced quality of life, long-term complications, shorter life expectancy, and the high costs of treatment and healthcare. It has also been found that in low-income countries, complications and early-age mortality are highly associated with Type 1 Diabetes. Over time, both the prevalence and incidence of the disease have increased significantly [6]. There are also scarcity of data including the incidence and prevalence of the disease in both younger children and adults. Additionally a lack of diagnosis of the disease at onset also led to underestimation of the incidence of the disease. Furthermore, the global prevalence of type 1 diabetes is increasing. Despite the implementation and development of necessary treatments and strategies, the decreasing age of onset remains a significant concern [5]. It has been observed that early-life intestinal microbiota play a crucial role in shaping the development of the immune system. Specifically, the lower intestinal tract contains most of the active microbes in the body. During early life,

exposure to antibiotics has been shown to alter intestinal microbiota, thereby influencing immune system functions and potentially contributing to an earlier onset of Type 1 Diabetes.

In addition to intestinal microbiota, fecal microbiota is also rich in biomarkers that help predict various disease states. As a result, investigating the association between fecal microbiota and patients' responses to therapy for Crohn's disease has become a global concern. Crohn's Disease, a type of inflammatory bowel disease, is an immunologically mediated inflammatory condition of the any part of the gastrointestinal tract characterized by inflammation that affects the entire thickness of the bowel wall [16]. The disease usually has a chronic and often progressive trajectory. Common symptoms are diarrhea, abdominal pain, nausea, and vomiting. Systemic manifestations include weight loss, fever, and fatigue. Occurring typically in patients aged 15-35 years, Crohn's Disease can extend beyond the lower part of the small intestine to affect the large intestine, stomach, esophagus, and the mouth. Moreover, it is highlighted that a higher prevalence of inflammatory bowel disease (IBD), particularly Crohn's disease (CD) can be found in regions traditionally known for lower incidence rates. While clinical symptoms in children and adults often overlap, issues such as delayed growth have also been documented. Additionally, while the disease may develop at an early age, diagnosis is typically feasible during adolescence [18].The rising prevalence and incidence of Crohn's Disease present significant economic and healthcare system challenges. Thus, identifying prognostic biomarkers to help clinicians determine which patients are responding to treatment would be highly beneficial [14]. Furthermore, early diagnosis and management of the disease can enhance patients' quality of life and treatment outcomes.

The evaluation of treatment impacts on the microbiome has shown that these treatments lead to changes in the microbial composition. For instance, the findings of the study on "Antiobiotic-induced Microbiome Depletion is Associated with Resilience in Mice after Chronic Social Defeat Stress", the treatment with antibiotic that was administered to the subjects have reduced the α-diversity of the gut microbiome as compared to the control group. Overall findings have shown that the treatment induced changes in both the diversity and composition of the gut microbiome. Specific phyla were also identified to have increased and decreased after administering the treatment [21]. It has also been emphasized that apart from several factors that affects the variation in human microbiome, namely age, geographical location, diet and hygiene, medications can also change the microbial composition. A systematic review on the effect of antibiotics on the human microbiome examined changes in the relative or differential abundance of the microbiome at various taxonomic levels across several studies. Additionally, diverse combinations of treatments, methodologies, and tools were employed, with numerous volunteers participating. The analyses were conducted over both short and long-term periods, focusing on the gut microbiota. The results showed that combinations of antibiotics have a significant impact on the human microbiota, particularly on alpha diversity. Furthermore, it was found that the impact of the treatment varies depending on the duration of the study [12].

1.1 Objectives

This research aims to investigate the effects of the treatment on the microbiome data. Specifically, the objective is to examine the longitudinal behavior of the microbiome over a period of time by using the alpha diversity measure, and to explore changes and trends on the microbiome

in response to the treatment. Additionally, the research will assess the influence of different intervention factors on the microbiome by comparing the microbiome profiles of different treatment groups at each of the time point.

2 Data Description

2.1 The CERTIFI Study

One of the datasets utilized in this research was from the Crohn's Evaluation of Response to Ustekinumab Anti-Interleukin-12/23 for Induction (CERTIFI) study [14], which is a Phase II dose finding clinical development program for Ustekinumab (UST) in Crohn's Diseases (CD) patients. Moreover, this study focuses on exploring the association between the fecal microbiota and the patients' response to the UST therapy for treating CD. Primarily, there were a total of 500 subjects involved in the study, for which has a moderate to severe CD condition. The stool samples of the subjects composed of the fecal communities were collected in a span of 22 weeks [5]. The composition was characterized by sequencing 16srRNA gene. The study was a randomized, double-blind placebo-controlled phase 2 clinical study [13]. In the induction phase, there were only 306 patients who were randomized to the treatment groups. The patients having complete sample and microbiome information were included for the analysis. Out of 130 patients, there were 29 belonging to the placebo group and 101 patients from the treatment group and their information were measured at three time points (week 0, 4 and 6), which comprise a total of 2353 OTUs.

2.2 The T1D Study

The second dataset was from the T1D study wherein the interest is to assess the effect of antibiotic pulse on the time to develop type 1 diabetes in non-obese diabetic (NOD) mice [13]. The main objective is to study the relationship between the perturbed microbiota and the development of the disease. The pregnant mice were monitored for diabetes weekly and their fecal and intestinal microbiota were measured. Specifically, the mice were randomized into treatment groups, wherein the treatment includes the 1PAT and 3PAT, corresponding to a single and three-courses of antibiotic, respectively. They were all measured at weeks 3, 5 and 7 with a total of 2720 OTUs. In this analysis, the subjects randomized to the 3PAT treatment groups were used, wherein 40 mice were from the treatment group and 41 were from the control group.

In this research, the response variable for both of the datasets are the alpha diversity measure of the microbiome, specifically the Shannon indices for the kingdom and family levels and the relative abundance for the OTU level. The Shannon index is mostly used as a diversity index in ecological and microbial studies. This measure reveals the richness of a community, that is, the higher the value of the Shannon index indicates a higher microbial community diversity [2]. For instance, the Shannon indices were obtained for each sample for each taxonomic level. These values reflect the richness of each sample. In research, relative abundance refers to the frequencies of taxa within a sample. Specifically, it is defined as the fraction of a taxon observed relative to the total sum of all taxa in the sample [11].

2.3 Data Filtering

Preceding the exploration of the longitudinal profile, the data were filtered based on some criteria. In particular, for kingdom level, all OTUs were included in exploring the longitudinal profile of the microbiome. As a necessary step before fitting the appropriate models for the analysis, in the family level, only the active families were identified and the inactive families were filtered out. An inactive family is a family that has zero measurements across all samples. The less active families were the families that contain zero counts in the control or the treatment group of at least 70% of the samples. These families were removed since these do not contain any information or is referred to as rare taxa which do not occur in most of the samples. Removing these families is helpful for the subsequent analysis. At the OTU level, most of the counts are zeros, a lower threshold for filtering was imposed. Less active OTUs has zero counts that are greater than or equal to 40% of the total counts for the CERTIFI data and greater than or equal to 80% for the T1D data, filtering out these OTUs narrows down the total number of OTUs that are considered to be active or abundant. This particular strategy can be classified as a prevalent filtering wherein the proportion of zeros prevalent in the treatment and control group was the criteria in filtering and selecting active OTUs. The OTUs removed are also termed as non-informative OTUs, these only are present in few samples, thus do not have a significant contribution on the analysis [22].

2.4 Data Transformation

The Shannon index was calculated at both the kingdom and family levels. For the family level, inactive families were initially removed. Subsequently, families with zero counts in either the control or treatment group in 70% or more of the samples were excluded. These steps were performed for each of the time point. At the OTU level, after filtering, the counts were subsequently transformed into relative abundances for the analysis by dividing each OTU count by the total count of all OTUs in the sample. This normalization process allows for the calculation of the proportion of each OTU relative to the total microbial community present in each sample.

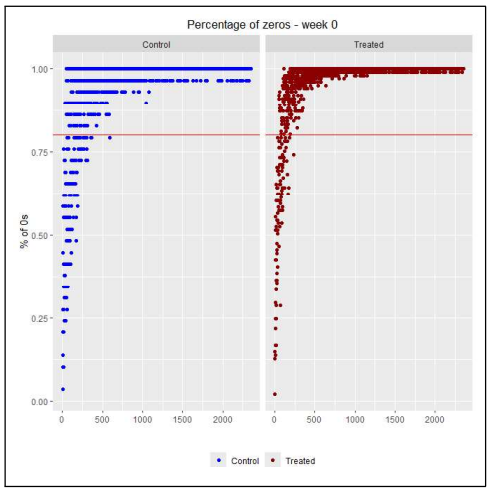
2.5 Exploratory Data Analysis

2.5.1 Percentage of Zeros

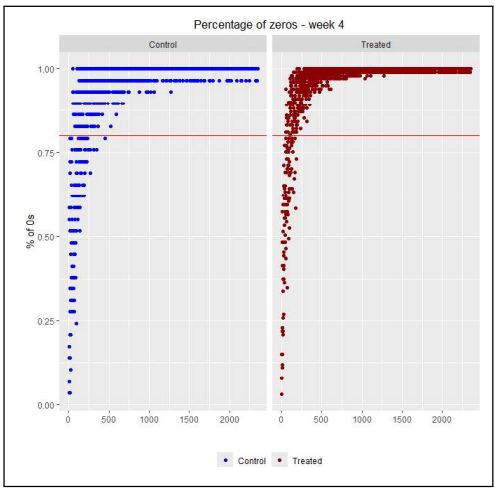
Microbiome data are characterized to have a large percentage of zeros. Several measures and strategies are applied to handle a high proportion of zeros that are present in microbiome data. These measures include adding a pseudo-count and probability modelling of excess zeros [8]. Moreover, the microbiome contains numerous operational taxonomic units (OTUs) in which majority are rare [4]. These rare OTUs typically have a majority of zero counts in most of the samples. Such difficulty is addressed by filtering out these rare OTUs for an efficient analysis. However, there is no general strategy to handle this characteristic of the data and taking into account this characteristic in processing the data for downstream analysis.

As illustrated in Figure 1, for the CERTIFI study, the percentage of zeros were examined for each week the fecal microbiome were measured for all the OTUs, across the treatment groups, that is, the group receiving the UST therapy and the control group. Among the 2353 OTUs, majority has zero counts. Similarly, this is observed in the T1D study as shown in Figure 2.

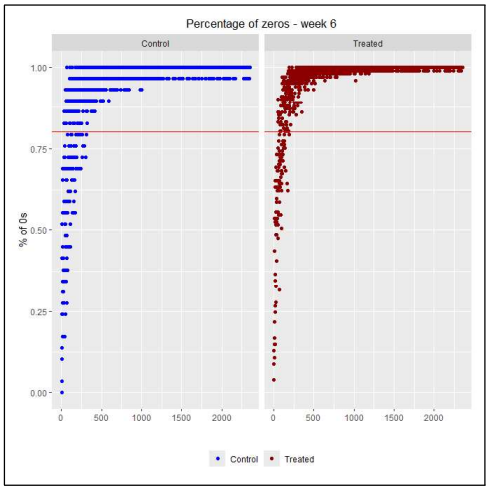
For weeks 3, 5 and 7, the fecal and intestinal microbiome mostly contains zero counts across the groups receiving the 3PAT treatment and the control group. It is noticeable that among the 2720 OTUs, only a small percentage of the counts were not zeros. Approximately, for both of the datasets, 80% of the counts are zeros, which is highlighted in the individual plots.



(a) Percentage of zeros - Week 0.

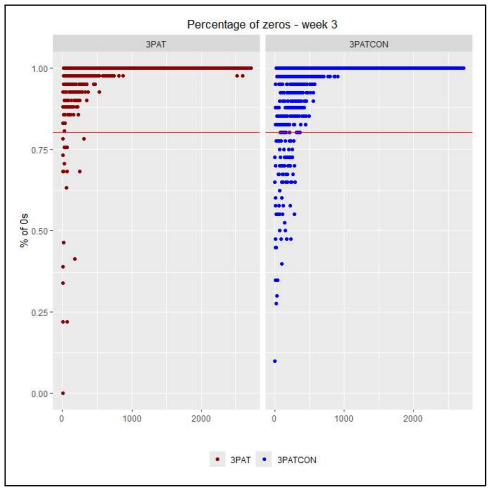


(b) Percentage of zeros - Week 4.

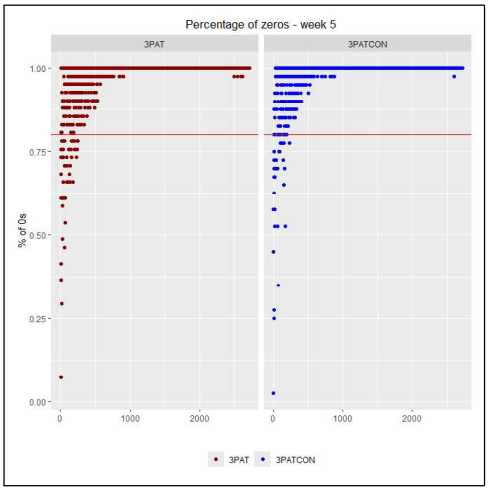


(c) Percentage of zeroes - Week 6

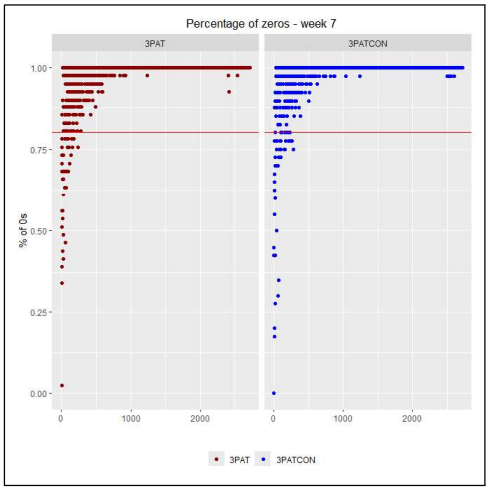
Figure 1: Percentage of zeros for the CERTIFI Study.



(a) Percentage of zeros - Week 3.



(b) Percentage of zeros - Week 5.

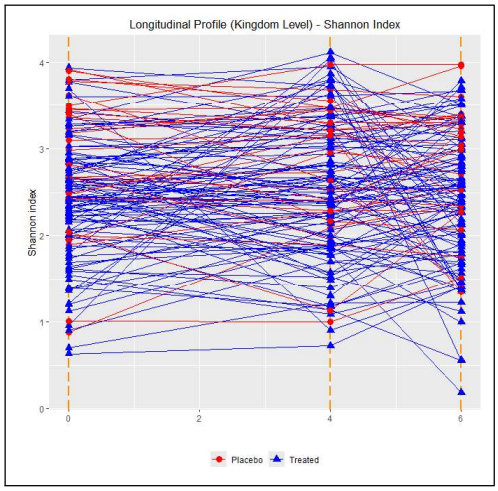


(c) Percentage of zeros - Week 7.

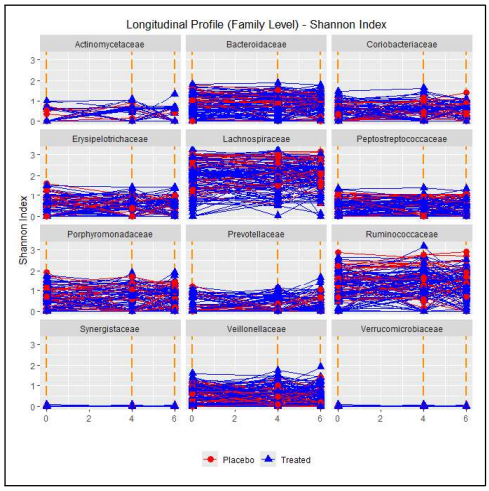
Figure 2: Percentage of zeros for the T1D Study.

2.5.2 Longitudinal Profiles

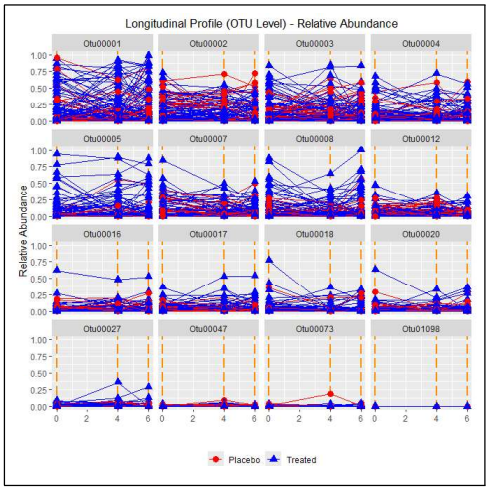
In exploring the longitudinal behavior of the Shannon index over time for the kingdom and family level and relative abundance over time for the OTU level, the measures of the microbiome, classified across treatment groups were plotted against the individual time points. Through this, trends and variations in microbiome and possible effects of treatment can be evaluated.



(a) Longitudinal Profile - Kingdom Level.



(b) Longitudinal Profile - Family Level.

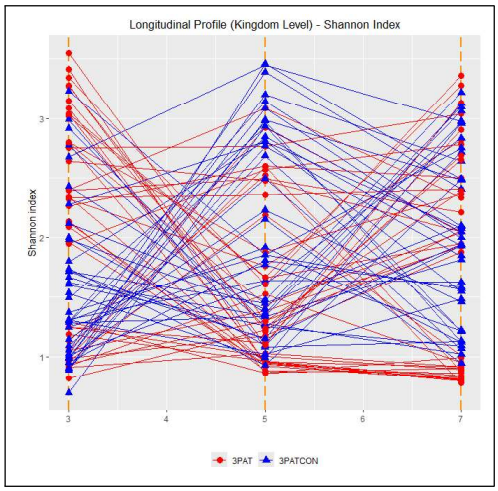


(c) Longitudinal Profile - OTU Level.

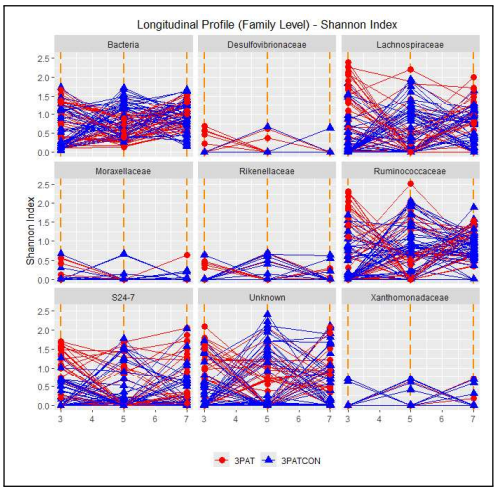
Figure 3: Longitudinal Profile for the CERTIFI Study.

As observed in the CERTIFI Study, at the kingdom level, in Figure 3a, the value for the Shannon index in each sample at Week 0 of visit are seen to be different. This observation suggests an inclusion of a random intercept. Furthermore, the longitudinal profile at family level was also plotted for both active and less active families. There were 11 active families for the CERTIFI study. Shown in Figure 3b are the active and less active families in the data. It is evident that among these active families, *Bacteroidaceae*, *Lachnospiraceae* and *Ruminococcaceae* have the highest Shannon indices. On the contrary, it is seen that families *Synergistaceae* and *Verrucomicrobiaceae* are less active since they exhibit lower Shannon indices over time. At the OTU level, with a total of 2353 OTUs, there were 15 OTUs that have high prevalence in all the samples from the study. Figure 3c shows examples of the longitudinal profiles of both the active and less active OTUs. Among these OTUs, it is noticeable that *Otu0001* has the highest relative abundance across the time points. Some of the less active OTUs from the data are *Otu00027*,

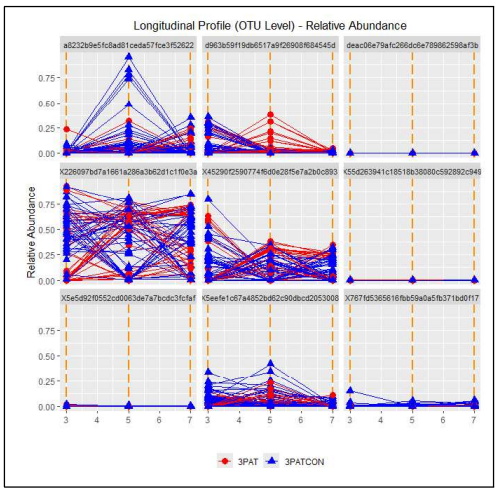
Otu00047, *Otu00073*, and *Otu01098*. These OTUs are seen to have lower relative abundances across all time points when compared to the active OTUs.



(a) Longitudinal Profile - Kingdom Level.



(b) Longitudinal Profile - Family Level.



(c) Longitudinal Profile - OTU Level.

Figure 4: Longitudinal Profile for the T1D Study.

Similarly, the same mechanism is observed for the T1D study wherein at the kingdom level, the Shannon indices were different at the starting point of the measurement, that is, at Week 3. Out of the 5 families that are active for the T1D study, illustrated in Figure 4b, the highest Shannon indices are observed in family *Lachnospiraceae* and *Ruminococcaeceae*. In addition to these, the plot also highlights less active families within the data. These families are, *Desulfovibrionaceae*, *Moraxellaceae*, *Rikenellaceae* and *Xanthomonadaceae*. At the OTU level, only 5 OTUs were highly prevalent among 2720 OTUs in total. In figure 4c, *X226097bd7a1661a286a3b62d1c1f0e3a* shows to be relatively abundant among the active OTUs. Also there are 4 less active OTUs shown in the plot. It can be observed that OTUs, *X5e5d92f0552cd0063de7a7bcdc3fcfaf*,

X55d263941c18518b38080c592892c949, *deac06e79afc266dc6e789862598af3b* show approximately zero values for the relative abundaces across the timepoints. Overall, these explorations do not indicate final conclusions.

3 Methodology

3.1 Model Formulation

In addressing the objectives of this research and drawing inferences, the appropriate statistical models were fitted. Initially the model that corresponds to the mean structure includes parameter estimates for the time, treatment indicator and the interaction effect of time and treatment indicator. Subsequently, the test for covariance structures for repeated measurement and test for inclusion of random effects were performed to formulate the final model.

3.1.1 The Linear Mixed Model

Longitudinal data, which are composed of repeated measurements available for all subjects are often characterized as highly unbalanced. Thus, it is essential to account for the subject-specific longitudinal profiles, wherein using multivariate regression techniques cannot adequately address. By fitting linear regression functions, both the within-subject and between-subject variability can be modelled through simultaneously estimating the fixed effects and the random [20]. This approach allows for a more accurate and comprehensive analysis of longitudinal data.

The linear mixed model is defined as follows:

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

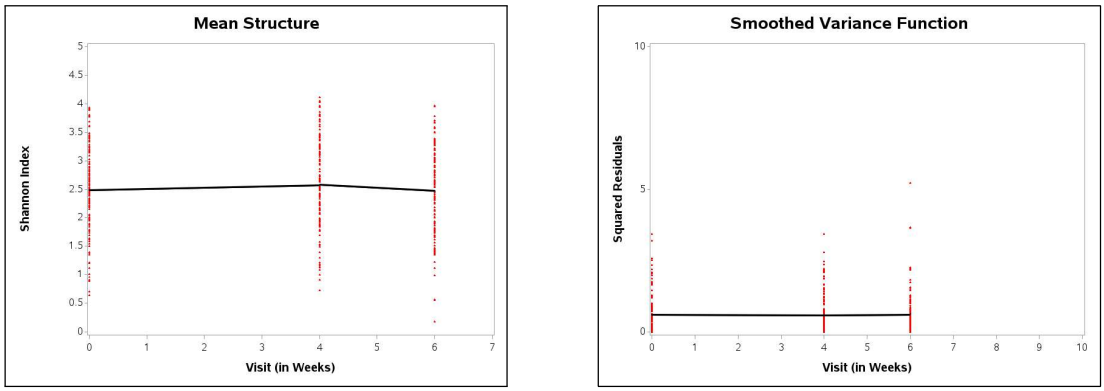
where X_i and Z_i are matrices of known covariates, β is a vector of fixed effects and b_i is a vector of the random effects. Additionally, $b_i \sim N(0, D)$ and $\varepsilon_i \sim N(0, \Sigma_i)$.

3.1.2 Mean and Variance Structures

For further analysis, the mean and variance structures of the data were examined. This step serves as a tool that extremely aids in the selection of the appropriate models [19]. The mean structure illustrates the expected value of the response which is the Shannon index over time. This was produced from employing the Local Regression procedure, also referred to as the LOESS procedure. This is commonly used to capture existing trend or patterns in the data [17] and to show the relationship between the time and response variable. Also, the variance structure was explored for initial plausible variance function over time [20]. Assessing the change in variability over time aids in using the appropriate models or modeling techniques that would account for the variability. Furthermore, the measurement of the fixed-effects are corrected that would also improve the precision of the estimates. In this step, the LOESS procedure was also performed on the mean residuals obtained from the mean structure. Thereafter, the squared residuals were plotted against time.

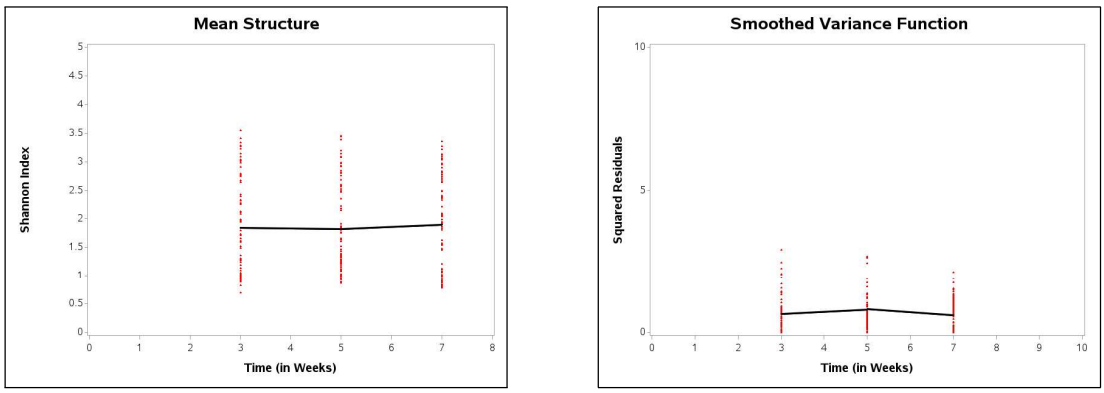
Evidently, Figure 5a shows that at week 4 of visit, the Shannon index has increased and is at its maximum value, as compared to the other week of visits. The plot also suggests an approximately

stable value of the response variable over time. Considering the plot for the mean structure of the data from the T1D study, shown in Figure 6a, similar to that of the previous plot, there is no clear trend that can be observed. The Shannon index, on average shows a constant behaviour, over time. In addition, the variance structures shown in Figures 5b and 6b show no clear pattern or trend. Dispersion of residuals around the mean does not change over time, implying that there is a stable variability in the data over time. Nonetheless, these illustrations only limit to preliminiary observations of the behavior, trend or patterns that exist in the data. Hence, inferences cannot be drawn at this stage. In the following sections, further exploration of the data in the three taxonomic levels were also shown.



(a) Mean Structure (b) Variance Structure

Figure 5: Mean and Variance Structure for the CERTIFI Study.



(a) Mean Structure (b) Variance Structure

Figure 6: Mean and Variance Structure for the T1D Study.

3.1.3 Test for Covariance Structure and Random Effects

Table 1: Comparison of covariance structures for CERTIFI Study.

Model	-2log-likelihood
(1) Unstructured	809.3
(2) Compound symmetry	811.0

Table 2: Comparison of covariance structures for T1D Study.

Model	-2log-likelihood
(1) Unstructured	559.9
(2) Compound symmetry	575.5

The covariance structures for the repeated measurements were examined through comparing between an unstructured covariance structure and compound symmetry. By assuming an unstructured covariance structure, this implies that there is no specific structure for the variance and covariances. On the other hand, using a compound symmetry as covariance structure implies that the variances are homogenous and the correlation between two separate measurements, regardless of the distance between these measurements, are assumed to be constant [9]. Tables 1 and 2 showed that based on the likelihood ratio test, as shown in the corresponding log likelihood values, an unstructured covariance structure is appropriate for the model as the value of its -2log-likelihood is smaller than that of the compound symmetry.

Furthermore, as seen in the exploratory analysis of the variance function of the data, the variances were constant. However, it is necessary to test whether an inclusion of only a random intercept or both random intercept and slopes of time should be included in the final model. Ideally, this step is performed through the use of the likelihood ratio test based on restricted maximum likelihood estimation and computation of the the p-values based on an asymptotic null distribution of two mixtures for the for the χ^2_1 and χ^2_2 distributions with equal weights of 0.5 [19]. However, in the case of the data used, convergence criteria were not met upon inclusion of a random slope in the model. Hence, only a random intercept is included in the final model.

3.1.4 Final Models

The final model for the CERTIFI study is defined as follows:

$$Y_{ij} = \beta_0 + \beta_1 Visit_{ij} + \beta_2 Treatment_i + \beta_3 Visit_{ij} * Treatment_i + b_{0i} + \varepsilon_{ij}.$$

where Y_{ij} corresponds to the Shannon indices for kingdom and family level, and corresponds to the relative abundances for the OTU level, $Visit$ is the time in weeks 0, 4 and 6, $Treatment$ as the treatment group indicator, $\beta_i s$ are the fixed effects, b_{0i} is the random intercept, ε_{ij} are the error terms, i represents individual microbiome and j represents the time point in weeks. Also, $b_{0i} \sim N(0, D)$ and $\varepsilon_{ij} \sim N(0, \Sigma_{ij})$.

Also the final model for the T1D study is written as follows:

$$Y_{ij} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Treatment_i + \beta_3 Time_{ij} * Treatment_i + b_{0i} + \varepsilon_{ij}$$

where Y_{ij} corresponds to the Shannon indices for kingdom and family level, and corresponds to the relative abundances for the OTU level, $Time$ is the time in weeks 3, 5 and 7, $Treatment$ as the treatment group indicator, $\beta_i s$ are the fixed effects, b_{0i} is the random intercept, ε_{ij} are the error terms, i represents individual microbiome and j represents the time point in weeks. Also, $b_{0i} \sim N(0, D)$ and $\varepsilon_{ij} \sim N(0, \Sigma_{ij})$.

3.2 Multiplicity Correction

A comparison of the effect of the treatment is assessed for each time point. This process is done for all identified active families and OTUs. With this, multiplicity correction was considered to control the false discovery rate while testing for multiple hypotheses [10]. The Benjamini-Hochberg false discovery rate (BH-FDR) procedure is employed in most of the recent researches specifically in large high dimensional data sets that involves numerous features, thus a multiplicity correction is necessary. This procedure have been widely used due to its simplicity wherein it arranges the p-values in an ascending order. Each p-value corresponds to a rank and is denoted by i where i ranges from 1 up to j. Then, the adjusted p-values are calculated with the following formula:

$$\tilde{p_i} = min(1, min_{j \geq i}(\frac{mp_j}{j})),$$

where $\tilde{p_i}$ is the adjusted p-value for the ith ranked p-value and m is the total number of hypothesis tests. The adjusted p-value $\tilde{p_i}$ is obtained through first taking the ratio $\frac{mp_j}{j}$ for each p-values where $j \geq i$. Taking the smallest value from the ratios and comparing it to 1, the minimum value is the adjusted p-value $\tilde{p_i}$ [15]. These adjusted p-values will then be compared to the 0.05, and from there the actual significant p-values will be used for making inferences.

3.3 Multiple Comparison Tests

Mainly, the analysis is focused in determining if differences in the means exist between the treatment groups. In this case, multiple comparison test is applied to compare the means of each treatment group at each time point. Below, the hypotheses are given by:

$$H_0 : \mu_C = \mu_T,$$
$$H_1 : \mu_C \neq \mu_T.$$

where μ_C is the mean for the control group and μ_T is the mean for the treatment group. This test is performed on each taxonomic level, that is for kingdom level, for each active families and active OTUs at each time point separately.

3.4 Softwares

The analyses performed in this research were facilitated by two software tools. In particular, RStudio version 4.4 was used for the pre-processing of the datasets and majority of the data exploration. Also, SAS on Demand, which is a cloud-based platform by Statistical Analysis System was mainly used for the model fitting and analyses including some of the data explorations.

4 Results

4.1 CERTIFI Study

4.1.1 Linear mixed model - Kingdom Level

To address the research objectives, a linear mixed model was fitted for each taxonomic level. In this section, examples of active families and OTUs are also illustrated. Firstly, examining the table for the results of fitting the model using the data from the CERTIFI study, at the kingdom level, Table 3 reveals that the effect of treatment at week 0 is significant at 5% level of significance. The estimate of 0.4590 (p-value = 0.0026) indicates that patients receiving the UST treatment has a higher Shannon index by 0.4590 when compared to patients in the control group. The linear effect of treatment on weeks 4 and 6 was found to be insignificant. This is also an indication that there is no difference in the Shannon indices of the patients across the treatment groups in these weeks.

Multiple comparison tests were conducted to further assess the effect of treatment on the microbiome for every visit, examining differences between groups at each time point. This process is necessary in order to know how the microbiome differs from one group to another. Table 4 reveals that the differences between the treatment groups were significant at Visits on week 0 and 6. For week 0, the Shannon index of the treatment group is found to be lower than the control group by 0.04590 (p-value = 0.0026) and by 0.3927 (p-value=0.0149) at week 6. The plot of the means per time point is shown in Figure 7. After doing the multiplicity correction, it revealed that at all time points the difference of microbiome between the groups is only significant at week 6.

Table 3: Parameter Estimates - Kingdom Level.

Effect	**Estimate**	**SE**	**DF**	**Test Statistic**	**p-value**
Intercept	2.3858	0.0713	128	33.46	< 0.0001
Visit (4)	0.1288	0.0745	256	1.73	0.0850
Visit (6)	-0.0024	0.0737	256	-0.03	0.9743
Treatment	0.4590	0.1510	256	3.04	0.0026
Visit (4): Treatment	-0.1819	0.1577	256	-1.15	0.2498
Visit (6): Treatment	-0.0663	0.1559	256	-0.43	0.6710

Table 4: Least Squares Estimates.

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Visit : Treatment	0	-0.4590	0.1510	256	-3.04	0.0026	0.0078
Visit : Treatment	4	-0.2771	0.1643	256	-1.69	0.0929	0.0929
Visit : Treatment	6	-0.3927	0.1603	256	-2.45	0.0149	0.0224

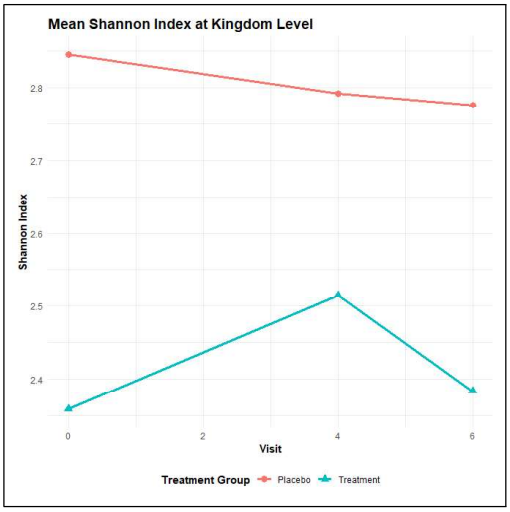


Figure 7: Mean Shannon Index - Kingdom Level (CERTIFI Study).

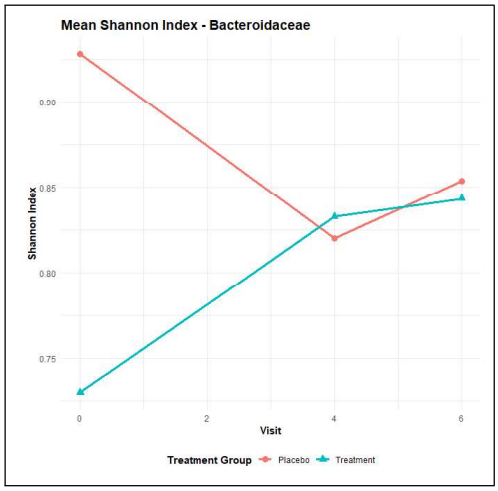
4.1.2 Linear Mixed Model - Family Level

Separate models were fitted for both active families and OTUs. In examining the effect of treatment at family level, multiplicity correction for multiple hypothesis tests was employed to control for false discoveries. For the CERTIFI Study, 11 active families were identified and are shown in Table 5 with the corresponding p-values for each time point. These p-values were obtained after employing the Benjamini-Hochberg correction for multiple testing. Among these families, it is evident that *Lachnospiraceae* shows significant difference in the Shannon indices of the treatment and control group at week 0. This case is the same with family *unclassified* and *Porphyromonadaceae* where the difference in the microbiome is only significant at week 0. The family *Ruminococcaceae* shows significant differences in the Shannon index for weeks 0 and 4.

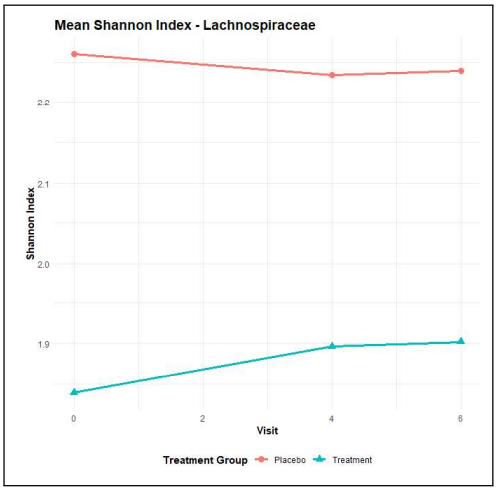
Table 5: The p-values of 11 active Families after Multiplicity Correction.

Family	**Visit (0)**	**Visit (4)**	**Visit (6)**
Bacteroidaceae	0.1379	0.9063	0.9243
Coriobacteriaceae	0.5831	0.4933	0.9243
Enterobacteriaceae	0.7039	0.1131	0.6119
Erysipelotrichaceae	0.1515	0.6247	0.9243
Lachnospiraceae	0.0330	0.0762	0.0766
Peptostreptococcaceae	0.0715	0.4933	0.0766
Porphyromonadaceae	0.0330	0.0762	0.3986
Ruminococcaceae	0.0143	0.0319	0.1015
Streptococcaceae	0.2703	0.4933	0.8128
unclassified	0.0349	0.0762	0.0766
Veillonellaceae	0.6596	0.3421	0.3986

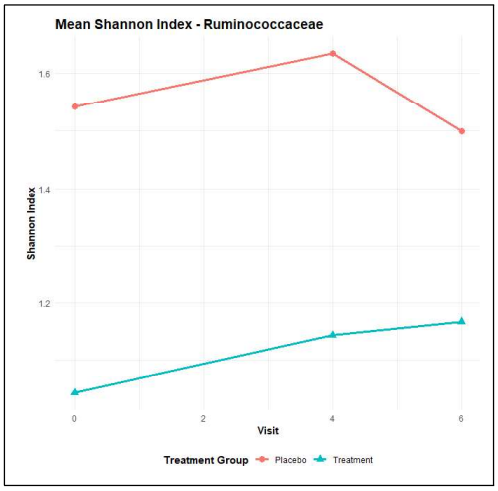
In the succeeding tables, three of the active families are presented with the estimates from fitting the model and the results from the comparison between groups. Table 6 shows the results for *Bacteroidaceae* wherein the effect of the treatment at week 4 is significant at 5% level with an estimate of -0.2110 (p-value = 0.0418). This means that the Shannon index of the treatment group is 0.2110 less than that of the control group, at week 4. It is also shown that when compared to week 0, the Shannon index is higher for week 4 and week 6 by 0.1033 and 0.1135, respectively. These results were also significant. Further comparison test was done for this family and it shows that the differences between the Shannon indices of the groups were not significant for all time points as seen in Table 7. Figure 8a shows the mean estimates for the Shannon indices for each group at each timepoint indicating that the Shannon index for the placebo group is higher than the treatment group at weeks 0 and 6.



(a) Mean Shannon Index - Bacteroidaceae.



(b) Mean Shannon Index - Lachnospiraceae.



(c) Mean Shannon Index - Ruminococcaceae.

Figure 8: Mean Shannon Index - Family Level (CERTIFI Study).

Table 6: Parameter Estimates - Bacteroidaceae.

Effect	**Estimate**	**SE**	**DF**	**Test Statistic**	**p-value**
Intercept	0.7297	0.0524	128	13.92	< 0.0001
Visit (4)	0.1033	0.0487	256	2.12	0.0350
Visit (6)	0.1135	0.0505	256	2.25	0.0253
Treatment	0.1983	0.1110	256	1.79	0.0752
Visit (4): Treatment	-0.2110	0.1032	256	-2.05	0.0418
Visit (6): Treatment	-0.1878	0.1069	256	-1.76	0.0801

Table 7: Least Squares Estimates (Bacteroidaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Visit : Treatment	0	-0.1983	0.1110	256	-1.79	0.0752	0.1379
Visit : Treatment	4	0.0127	0.1081	256	0.12	0.9063	0.9063
Visit : Treatment	6	-0.0105	0.1052	256	-0.10	0.9206	0.9243

Table 8: Parameter Estimates - Lachnospiraceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	1.8389	0.0756	128	24.33	< 0.0001
Visit (4)	0.0574	0.0656	256	0.88	0.3820
Visit (6)	0.0632	0.0642	256	0.98	0.3257
Treatment	0.4211	0.1600	256	2.63	0.0090
Visit (4): Treatment	-0.0834	0.1389	256	-0.60	0.5485
Visit (6): Treatment	-0.0840	0.1359	256	-0.62	0.5368

Table 8 shows the results for the family *Lachnospiraceae*. As seen in the table, the effects of treatment on the microbiome for each timepoint were not significant at 5% level of significance. Further examining the differences of the effect of treatment for all time points, the results for the multiple comparison in Table 9 shows that for all time points, there is a significant difference in the effect of treatment between the control group and the treated group. At week 0, the treatment group has Shannon index lower than the control group by 0.4211 (p-value = 0.0090). Furthermore, the Shannon index at week 4 is 0.3377 (p-value = 0.0277) lower for the treatment group than the control and at week 6 it is 0.3371 (p-value = 0.0193) lower. Also, as shown in Figure 8b, the Shannon index is higher for the placebo group in all time points.

Table 9: Least Squares Estimates (Lachnospiraceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Visit : Treatment	0	-0.4211	0.1600	256	-2.63	0.0090	0.0330
Visit : Treatment	4	-0.3377	0.1525	256	-2.21	0.0277	0.0762
Visit : Treatment	6	-0.3371	0.1431	256	-2.35	0.0193	0.0766

Table 10: Parameter Estimates - Ruminococcaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	1.0437	0.0721	128	14.47	< 0.0001
Visit (4)	0.0999	0.0835	256	1.20	0.2323
Visit (6)	0.1236	0.0843	256	1.47	0.1438
Treatment	0.4983	0.1527	256	3.26	0.0013
Visit (4): Treatment	-0.0082	0.1767	256	-0.05	0.9630
Visit (6): Treatment	-0.1667	0.1785	256	-0.93	0.3511

Table 11: Least Squares Estimates (Ruminococcaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Visit : Treatment	0	-0.4983	0.1527	256	-3.26	0.0013	0.0143
Visit : Treatment	4	-0.4901	0.1632	256	-3.00	0.0029	0.0319
Visit : Treatment	6	-0.3316	0.1581	256	-2.10	0.0369	0.1015

Another active family observed from the data is *Ruminococcaceae*. The findings reveal that the effect of treatment at each week of visit is not significant. The effect of treatment is seen to be significant with an estimate of 0.4983 (p-value = 0.0013) suggesting that the Shannon index of the treatment group is higher by 0.4983 than the control group. By evaluating the difference in the Shannon index between the treatment and control group, Table 11 shows that there is a significant difference in the Shannon index of the groups at each time point. The results exhibit there is a difference of 0.4983(p-value = 0.0013) at week 0, 0.4901 (p-value = 0.0029) at week 4, and 0.3316 (p-value = 0.0369) at week 6, in the Shannon index between the two groups, highlighting that the Shannon index is higher for the control/placebo group. Figure 8c indicates that at each time point, the Shannon index is higher for the placebo group.

4.1.3 Linear Mixed Model - OTU Level

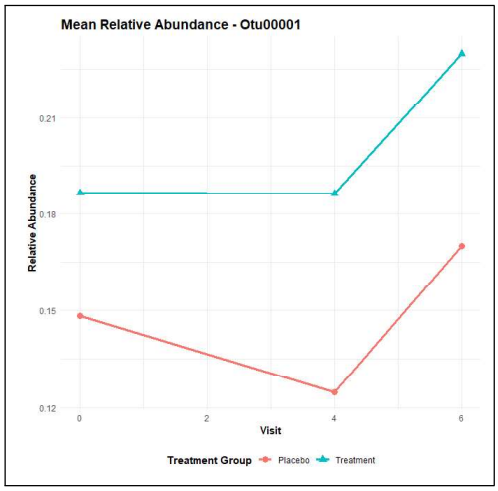
Analyzing the results at the OTU level revealed the presence of 15 active OTUs. Shown in Table 12 are the p-values for each OTU at each week of visit, after the multiplicity correction. It was observed that *Otu00005* and *Otu00067* did not yield any result, this is because the model fitted for these OTUs did not converge. Prior to fitting the models, it was evident that the counts for these OTUs were mostly zeros, and converting them to relative abundances resulted to smaller values. Accordingly, among the active OTUs, only *Otu00002* reveals that the relative abundances between groups are different at week 0. Moreover, a linear mixed model is fitted for each of these OTUs. The following tables show the results for three of the active OTUs.

Table 12: The P-values of 13 active OTUs after Multiplicity Correction.

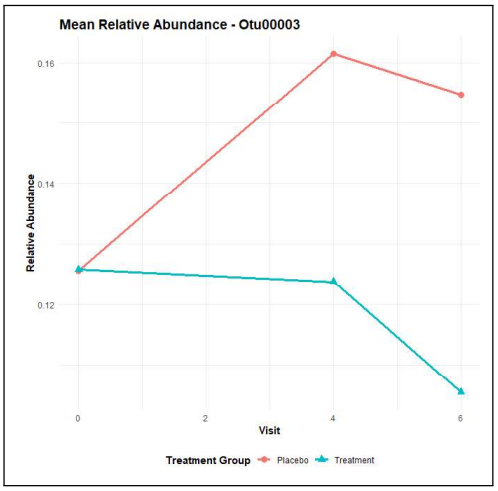
OTU	Visit (0)	Visit (4)	Visit (6)
Otu00001	0.7386	0.6427	0.9101
Otu00002	0.0351	0.6427	0.3081
Otu00003	0.9958	0.6639	0.7614
Otu00004	0.7386	0.8116	0.9101
Otu00007	0.7386	0.8116	0.9101
Otu00008	0.6483	0.6427	0.7614
Otu00012	0.6483	0.6427	0.9101
Otu00016	0.7386	0.8116	0.9101
Otu00017	0.9239	0.6639	0.9101
Otu00018	0.7386	0.7058	0.9101
Otu00020	0.7386	0.8687	0.9101
Otu00022	0.7386	0.6427	0.9101
Otu00025	0.7386	0.7058	0.9101

Table 13 shows the results for *Otu00001*. Following the multiple comparisons to assess the

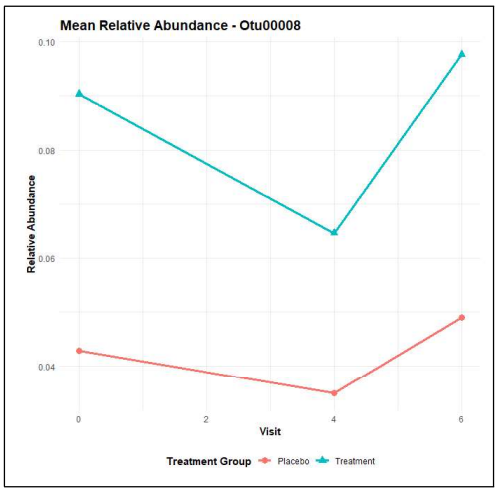
difference between the relative abundances of the treatment and control group, there were no significant results. Although not significant, it is evident that the treatment group has higher relative abundance for all the time points. The relative abundance of the treatment group is higher by 0.0381 at week 0, 0.0616 at week 4 and 0.0598 at week 6. For this OTU, the plot for the mean relative abundances is shown in Figure 9a, indicating that the relative abundances are higher for the treatment group than the placebo group.



(a) Mean Relative Abundance - Otu00001.



(b) Mean Relative Abundance - Otu00003.



(c) Mean Relative Abundance - Otu00008.

Figure 9: Mean Relative Abundance - OTU Level (CERTIFI Study).

Table 13: Parameter Estimates - Otu00001.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.1864	0.0245	128	7.62	< 0.0001
Visit (4)	-0.0002	0.0253	256	-0.01	0.9927
Visit (6)	0.0434	0.0249	256	1.75	0.0817
Treatment	-0.0381	0.0518	256	-0.74	0.4624
Visit (4): Treatment	-0.0235	0.0536	256	-0.44	0.6612
Visit (6): Treatment	-0.0217	0.0526	256	-0.41	0.6801

Table 14: Least Squares Estimates (Otu00001).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Visit : Treatment	0	0.0381	0.0518	256	0.74	0.4624	0.7386
Visit : Treatment	4	0.0616	0.0499	256	1.24	0.2176	0.6427
Visit : Treatment	6	0.0598	0.0582	256	1.03	0.3052	0.9101

Additionally, the results for *Otu00003* in Table 15 reveal that the relative abundance is higher for the treatment groups. However, looking at the multiple comparisons in Table 16, at week 0, the relative abundance is higher by 0.0002 (p-value = 0.9958) in the treatment group. Also, at week 4, the placebo group is relatively abundant than the treatment group by 0.0376 and by 0.0492 at week 6. These estimates were not statistically significant at 5% level. Figure 9b shows that the relative abundances are higher for the treatment group at weeks 4 and 6, and are not significantly different at week 0.

Table 15: Parameter Estimates - Otu00003.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.1258	0.0174	128	7.22	< 0.0001
Visit (4)	-0.0019	0.0201	256	-0.10	0.9232
Visit (6)	-0.0202	0.0156	256	-1.30	0.1962
Treatment	-0.0002	0.0369	256	-0.01	0.9958
Visit (4): Treatment	0.0378	0.0426	256	0.89	0.3760
Visit (6): Treatment	0.0494	0.0331	256	1.49	0.1362

Table 16: Least Squares Estimates (Otu00003).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Visit : Treatment	0	0.0002	0.0369	256	0.01	0.9958	0.9958
Visit : Treatment	4	-0.0376	0.0371	256	-1.01	0.3117	0.6639
Visit : Treatment	6	-0.0492	0.0335	256	-1.47	0.1428	0.7614

The results for *Otu00008* are shown in Tables 17 and 18. The results exhibit a higher relative abundance for the treatment group as indicated by the estimates. Specifically, the relative abundance is of the treatment group is higher than the control group by 0.0475, 0.0297 and

0.0488, at weeks 0, 4 and 6, respectively. The plot for the mean relative abundances is shown in Figure 9c where it can be seen that the mean relative abundances are higher for the treatment group for all time points.

Table 17: Parameter Estimates - Otu00008.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0903	0.0155	128	5.83	< 0.0001
Visit (4)	-0.0257	0.0143	256	-1.80	0.0725
Visit (6)	0.0073	0.0146	256	0.50	0.6179
Treatment	-0.0475	0.0328	256	-1.45	0.1496
Visit (4): Treatment	0.0177	0.0302	256	0.59	0.5576
Visit (6): Treatment	-0.0013	0.0310	256	-0.04	0.9660

Table 18: Least Squares Estimates (Otu00008).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Visit : Treatment	0	0.0475	0.0328	256	1.45	0.1496	0.6483
Visit : Treatment	4	0.0297	0.0214	256	1.39	0.1662	0.6427
Visit : Treatment	6	0.0488	0.0359	256	1.36	0.1757	0.7614

4.2 T1D Study

4.2.1 Linear mixed model - Kingdom Level

Table 19: Parameter Estimates - Kingdom Level.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	2.1261	0.1232	79	17.25	< 0.0001
Time (5)	-0.5404	0.1915	158	-2.82	0.0054
Time (7)	-0.3369	0.1462	158	-2.30	0.0225
Treatment	-0.5595	0.1732	158	-3.23	0.0015
Time (5): Treatment	1.0312	0.2692	158	3.83	0.0002
Time (7): Treatment	0.7795	0.2055	158	3.79	0.0002

Table 20: Least Squares Estimates.

Effect	Time	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Time : Treatment	3	0.5595	0.1732	158	3.23	0.0015	0.0045
Time : Treatment	5	-0.4718	0.1700	158	-2.78	0.0062	0.0093
Time : Treatment	7	-0.2201	0.1767	158	-1.25	0.2148	0.2148

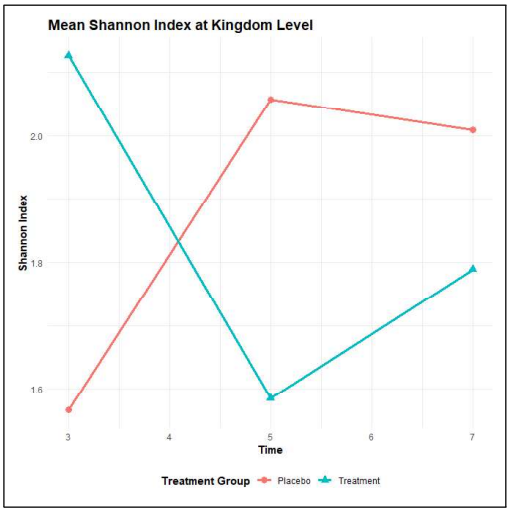


Figure 10: Mean Shannon Index - Kingdom Level.

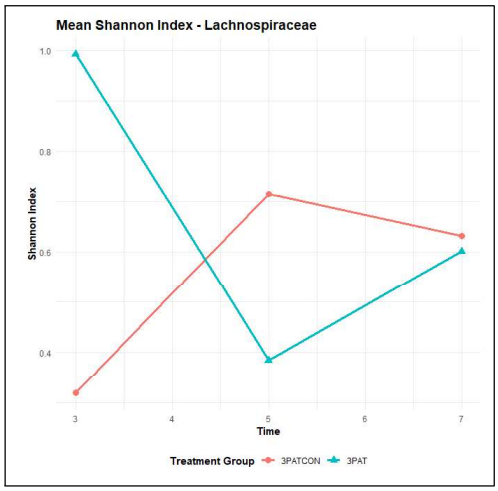
Similar to the previous analysis, to address the same objective of assessing the effect of treatment on the microbiome, the linear mixed model was also fitted for the data from the T1D study. The model was fitted using the data at different taxonomic levels which are the kingdom, family and OTU level. In this data, the samples were grouped into the treatment group (3PAT) and the control group (3PATCON). Firstly, for the kingdom level, the results are shown in Tables 19 and 20. Focusing on the estimates in Table 19, the results indicate a significant effect of treatment on the microbiome at 5% level of significance. Particularly, the effect of treatment has an estimate of -0.5595 (p-value =0.0015). This is an evidence that the Shannon index is higher for the control group by 0.5595. Additionally, a significant effect of treatment at week 5 and 7 was also observed. At week 5, the Shannon index is higher for the treatment group by 1.0312 (p-value = 0.0002). At week 7, the treatment group has higher Shannon index than the control group by 0.7795 (p-value = 0.0002). Furthermore, multiple comparisons were employed to evaluate differences in the Shannon indices of the treatment and control group at each week. In Table 20, the results for weeks 3 and 5 were significant at 5% level. The Shannon index of the treatment group is higher than the control group by 0.5595 (p-value = 0.0015) and lower by 0.4718 (p-value = 0.0062), at weeks 3 and 5, respectively. These values are indeed significant after multiplicity correction.

4.2.2 Linear Mixed Model - Family Level

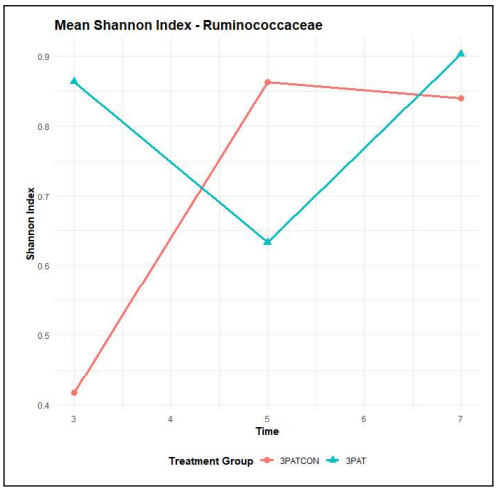
The Shannon index was also used as the microbiome at the family level. Fitting separate models for the families in the data requires a multiplicity correction to avoid false discoveries. Prior to model fitting, there were 5 active families from the T1D study. These families are shown below in Table 21 with their respective p-values after the multiplicity correction. Among these families, it is observed that at weeks 3 and 5, *Lachnospiraceae* has significant results. At week 5, the result was significant for *Bacteria* and at week 3, the results were significant for *S24-7* and *Unknown* family.

Table 21: The P-values of 5 active Families after Multiplicity Correction.

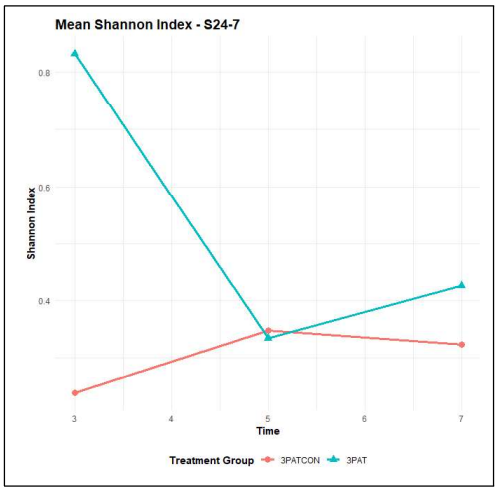
Family	Time (3)	Time (5)	Time (7)
Bacteria	0.1397	0.0005	0.7991
Lachnospiraceae	0.0002	0.0338	0.7991
Ruminococcaceae	0.0058	0.1960	0.7923
S24-7	0.0002	0.9160	0.7923
Unknown	0.0002	0.3665	0.7923



(a) Mean Shannon Index - Lachnospiraceae.



(b) Mean Shannon Index- Ruminococcaceae.



(c) Mean Shannon Index - S24-7.

Figure 11: Mean Shannon Index - Family Level (T1D Study).

Table 22 shows the results for *Lachnospiraceae*. It shows that the effect of treatment is statistically significant with an estimate of -0.6725 (p-value = < 0.0001), indicating that the Shannon index

of the treatment group is lower than the control group by 0.6725. The effect of treatment at weeks 5 and 7 were also significant at 5% level. In particular, to investigate the differences of the effect of the treatment between the two groups, multiple comparisons were employed. In Table 23 shows that the effect of treatment at weeks 3 and 5 were significant. The microbiome is higher for the treatment group by 0.6725 (p-value = < 0.0001) at week 3 and is lower by 0.3309 (p-value = 0.0135) at week 5. Figure 11a shows that the mean Shannon indices are higher for the control group in week 5 and 7.

Table 22: Parameter Estimates - Lachnospiraceae.

Effect	**Estimate**	**SE**	**DF**	**Test Statistic**	**p-value**
Intercept	0.9919	0.1071	79	9.26	< 0.0001
Time (5)	-0.6072	0.1491	158	-4.07	< 0.0001
Time (7)	-0.3906	0.1254	158	-3.12	0.0022
Treatment	-0.6725	0.1505	158	-4.47	< 0.0001
Time (5): Treatment	1.0035	0.2095	158	4.79	< 0.0001
Time (7): Treatment	0.7032	0.1762	158	3.99	0.0001

Table 23: Least Squares Estimates (Lachnospiraceae).

Effect	**Time**	**Estimate**	**SE**	**DF**	**Test Statistic**	**p-value**	**Adjusted p-value**
Time : Treatment	3	0.6725	0.1505	158	4.47	< 0.0001	0.0002
Time : Treatment	5	-0.3309	0.1324	158	-2.50	0.0135	0.0338
Time : Treatment	7	-0.0306	0.1201	158	-0.25	0.7991	0.7991

Table 24: Parameter Estimates - Ruminococcaceae.

Effect	**Estimate**	**SE**	**DF**	**Test Statistic**	**p-value**
Intercept	0.8636	0.1108	79	7.80	< 0.0001
Time (5)	-0.2305	0.1675	158	-1.38	0.1707
Time (7)	0.03938	0.1203	158	0.33	0.7439
Treatment	-0.4473	0.1557	158	-2.87	0.0046
Time (5): Treatment	0.6769	0.2354	158	2.88	0.0046
Time (7): Treatment	0.3843	0.1691	158	2.27	0.0244

For *Ruminococcaceae*, the effect of treatment is also significant with an estimate of -0.4473 (p-value = 0.0046) suggesting that the control group has Shannon index higher than the treatment group by 0.4473. Further analysis on the differences of the effect of treatment was done through multiple comparisons. Table 25 shows that the difference between the groups is only significant at week 3. The treatment group has higher Shannon index than the control by 0.4473 (p-value = 0.0046). The plot for the mean Shannon indices for the treatment groups are shown in Figure 11b and it exhibits a higher Shannon index for the treatment group at weeks 3 and 7.

Table 25: Least Squares Estimates (Ruminococcaceae).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Time : Treatment	3	0.4473	0.1557	158	2.87	0.0046	0.0058
Time : Treatment	5	-0.2296	0.1459	158	-1.57	0.1176	0.1960
Time : Treatment	7	0.0630	0.0880	158	0.72	0.4754	0.7923

Among the other families, *S24-7* exhibits significant effect of the treatment. The Shannon index is lower for the treatment group by 0.5935 (p-value = < 0.0001) than the control group. Also, the results reveal that the linear effect of treatment at each time point were statistically significant. Considering the results from the multiple comparisons, it is evident that only at week 3, the effect of treatment was significant. The Shannon index is higher for the treatment group than that of the control group by 0.5935 (p-value = < 0.0001). Figure 11c illustrates that the Shannon indices are higher for the treatment group at weeks 3 and 7 and slightly lower than the control group at week 5.

Table 26: Parameter Estimates - S24-7.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.8319	0.0772	79	10.78	< 0.0001
Time (5)	-0.4973	0.1153	158	-4.31	< 0.0001
Time (7)	-0.4056	0.0945	158	-4.29	< 0.0001
Treatment	-0.5935	0.1085	158	-5.47	< 0.0001
Time (5): Treatment	0.6064	0.1621	158	3.74	0.0003
Time (7): Treatment	0.4906	0.1329	158	3.69	0.0003

Table 27: Least Squares Estimates (S24-7).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Time : Treatment	3	0.5935	0.1085	158	5.477	< 0.0001	0.0002
Time : Treatment	5	-0.0129	0.1226	158	-0.11	0.9165	0.9160
Time : Treatment	7	0.1030	0.1258	158	0.82	0.4144	0.7923

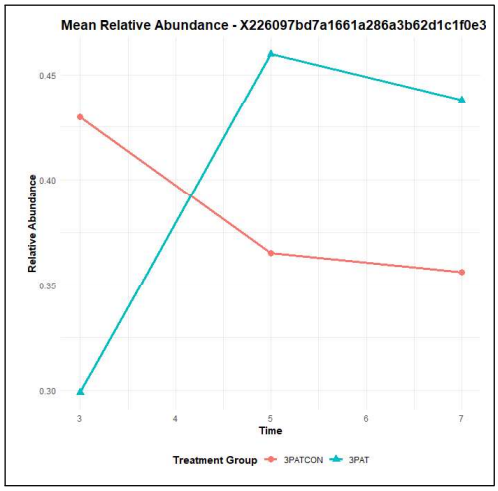
4.2.3 Linear Mixed Model - OTU Level

Subsequent to obtaining the relative abundances for all the OTUs, there were 5 active OTUs determined. By applying multiplicity correction, *d963b59f19db6517a9f26908f684545d* showed significant results for weeks 5 and 7, *a8232b9e5fc8ad81ceda57fce3f52622* exhibit significant result for week 5, and both *X226097bd7a1661a286a3b62d1c1f0e3a* and *X5eefe1c67a4852bd62c90dbcd2053008* have significant result for week 3. The results for each family are shown in the succeeding tables.

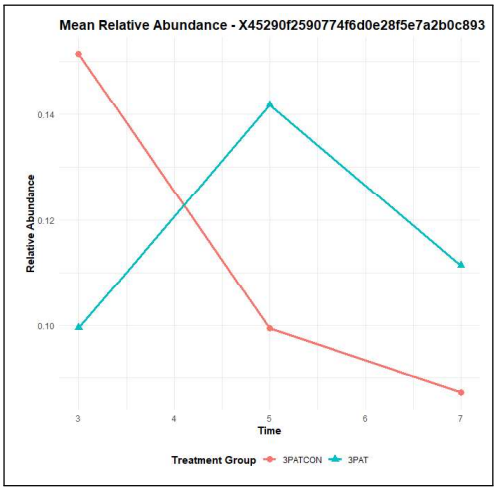
Table 28: The P-values of 5 active OTUs after Multiplicity Correction.

OTU	**Time (3)**	**Time (5)**	**Time (7)**
X226097bd7a1661a286a3b62d1c1f0e3a	0.0075	0.1464	0.4375
X45290f2590774f6d0e28f5e7a2b0c893	0.2871	0.1464	0.5470
d963b59f19db6517a9f26908f684545d	0.2003	0.0030	0.0005
X5eefe1c67a4852bd62c90dbcd2053008	0.0075	0.1464	0.6255
a8232b9e5fc8ad81ceda57fce3f52622	0.2889	0.0043	0.7594

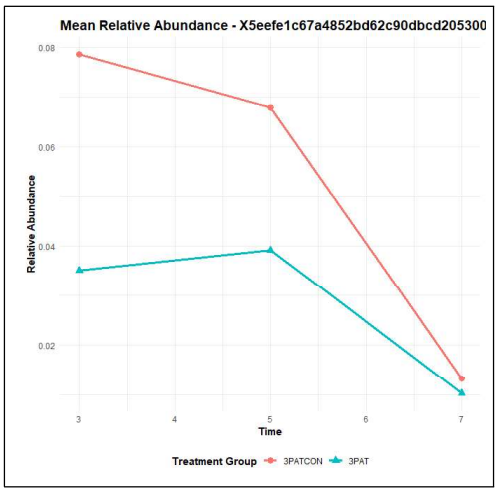
The treatment effect is seen to be significant for OTU *X226097bd7a1661a286a3b62d1c1f0e3a*. The relative abundance of the treatment group is higher by 0.1939 (0.0015) than the control group. In addition, the effect of treatment at each timepoint were also significant. In Table 30, examining the results from the multiple comparisons, at week 3, there is a significant difference in the relative abundances of the two groups, wherein the control group has higher relative abundance than the treatment group by 0.1939 (p-value = 0.0015). Figure 12a shows that the mean relative abundance are higher for the treatment group at weeks 5 and 7.



(a) Mean Relative Abundance - X226097bd7a1661a286a3b62d1c1f0e3a.



(b) Mean Relative Abundance - X45290f2590774f6d0e28f5e7a2b0c893.



(c) Mean Relative Abundance - X5eefe1c67a4852bd62c90dbcd2053008.

Figure 12: Mean Relative Abundance - OTU Level (T1D Study).

Table 29: Parameter Estimates - X226097bd7a1661a286a3b62d1c1f0e3a.

Effect	**Estimate**	**SE**	**DF**	**Test Statistic**	**p-value**
Intercept	0.2987	0.0426	79	7.01	< 0.0001
Time (5)	0.1611	0.0609	158	2.64	0.0090
Time (7)	0.1392	0.0531	158	2.62	0.0096
Treatment	0.1939	0.0599	158	3.24	0.0015
Time (5): Treatment	-0.2885	0.0856	158	-3.37	0.0009
Time (7): Treatment	-0.2756	0.0747	158	-3.69	0.0003

Table 30: Least Squares Estimates (X226097bd7a1661a286a3b62d1c1f0e3a).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Time : Treatment	3	-0.1939	0.0599	158	-3.24	0.0015	0.0075
Time : Treatment	5	0.0947	0.0636	158	1.49	0.1386	0.1464
Time : Treatment	7	0.0817	0.0600	158	1.36	0.1750	0.4375

The results show that for *X45290f2590774f6d0e28f5e7a2b0c893*, the effect of treatment was not significant. Also, in the multiple comparisons, the results reveal that there were no significant differences in relative abundances of the treatment and control groups in all time points. This merely suggests that there is insufficient evidence to conclude that the treatment has an effect on the relative abundance of the OTU. Figure 12b shows that the mean relative abundance is higher for the treatment group in weeks 5 and 7. In contrast, the effect of treatment is found to be significant for OTU *X5eefe1c67a4852bd62c90dbcd2053008* as shown in Table 33. The relative abundance of the treatment group is higher by 0.0434 (p-value = 0.0030) than the control group. Specifically, this effect is also statistically significant at week 3 as shown in the results of the multiple comparisons. The treatment group has lower relative abundance by 0.0434 (p-value = 0.0030). The mean relative abundance is observed to be higher for the control group for all time points as shown in Figure 12c, although only a small difference is seen at week 7.

Table 31: Parameter Estimates - X45290f2590774f6d0e28f5e7a2b0c893.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0995	0.0306	79	3.25	0.0017
Time (5)	0.0422	0.0413	158	1.02	0.3084
Time (7)	0.0119	0.0340	158	0.35	0.7264
Treatment	0.0519	0.0431	158	1.21	0.2297
Time (5): Treatment	-0.0943	0.0581	158	-1.62	0.1066
Time (7): Treatment	-0.0762	0.0478	158	-1.59	0.1130

Table 32: Least Squares Estimates - Multiple comparisons (X45290f2590774f6d0e28f5e7a2b0c893).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Time : Treatment	3	-0.0519	0.0431	158	-1.21	0.2297	0.2871
Time : Treatment	5	0.0424	0.0290	158	1.46	0.1464	0.1464
Time : Treatment	7	0.0242	0.0247	158	0.98	0.3282	0.5470

Table 33: Parameter Estimates - X5eefe1c67a4852bd62c90dbcd2053008.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0350	0.0102	79	3.42	0.0010
Time (5)	0.0041	0.0158	158	0.26	0.7965
Time (7)	-0.0249	0.0114	158	-2.19	0.0298
Treatment	0.0434	0.0144	158	3.02	0.0030
Time (5): Treatment	-0.0147	0.0223	158	-0.66	0.5094
Time (7): Treatment	-0.0404	0.0160	158	-2.53	0.0123

Table 34: Least Squares Estimates - Multiple comparisons (X5eefe1c67a4852bd62c90dbcd2053008).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value	Adjusted p-value
Time : Treatment	3	-0.0434	0.0144	158	-3.02	0.0030	0.0075
Time : Treatment	5	-0.0287	0.0178	158	-1.61	0.1088	0.1464
Time : Treatment	7	-0.0030	0.0045	158	-0.68	0.5004	0.6255

5 Software

The model fitting was mainly performed through the SAS software. The analyses were done at kingdom level, at family level where the linear mixed model was fitted for each active family, and at OTU level where the model was also fitted for each active OTU. Generally, the linear mixed model is written as:

$$Y_{ij} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Treatment_i + \beta_3 Time_{ij} * Treatment_i + b_{0i} + \varepsilon_{ij}$$

where Y_{ij} corresponds to the Shannon indices for kingdom and family level, and corresponds to the relative abundances for the OTU level, $\beta_i s$ are the fixed effects, b_{0i} is the random intercept, and ε_{ij} are the error terms. The analyses were performed by initially fitting the models and obtaining the estimates, followed by examining the differences between the microbiome of the two groups and applying multiplicity correction for multiple testing.

6 Ethics, Societal Relevance, Stakeholder Awareness

Ethical considerations were important in this research, especially regarding confidentiality. Measures were implemented to ensure that participants' personal information were securely protected and not disclosed or used for any purposes other than those related to the study. Additionally, the findings were reported honestly, ensuring that plagiarism was avoided and results were not duplicated from similar studies. In terms of societal relevance, conducting this study and analyzing the data can significantly contribute to addressing various community health issues, specifically in the areas of disease management, drug development, and medical interventions. Particularly, this study aimed to contribute in addressing the increasing incidence and prevalence of Type 1 diabetes and Crohn's Disease. This enables clinicians to identify patients who are more likely to respond to the treatments. Lastly, the findings can potentially benefit medical institutions, health personnel, patients and researchers through enhanced medical treatments, disease diagnosis and prevention.

7 Discussion

This project investigated the effects of treatment on microbiome data. Specifically, the microbiome data utilized in this study is a longitudinal data from two studies which are the CERTIFI study that focuses on the effect of the Ustekinumab Therapy in Crohn's Disease (CD) patients. The data used in this project focused on the measurements that were taken from three time points (week 0, 4 and 6). For the T1D study, the dataset obtained was from an experiment in non-obese diabetic mice that were given the three-course antibiotic pulse treatment. The measurements were taken from the subjects at three time points (week 3, 5 and 7). The analyses were done at three taxonomic levels, i.e., kingdom, family and OTU level. The response variable for the datasets were the Shannon indices at kingdom and family level and the relative abundance at the OTU level. Preliminary to fitting the models, an exploratory analysis was done for both of the datasets. The datasets were clearly highly sparse, so data filtering was considered necessary. Thus, reducing the size of the dataset for the subsequent analyses. The excluded taxa are non-informative taxa, which all has zero counts and some less active taxa. Moreover, the Shannon index and the relative abundances were obtained from the counts, and these measures were used in fitting the models. To draw inferences and answer the objective of this project, linear mixed models were fitted at kingdom level, each for all active families and active OTUs. The model specification was also assessed and the appropriate mean and variance structures were applied to the final model. At family and OTU level, a multiplicity correction was considered to control for the false discoveries since in these taxonomic levels, multiple hypotheses were tested.

The data from the CERTIFI study revealed 11 active families and 15 active OTUs. These were identified by determining which family or OTU had the majority of nonzero counts across all samples. Initially, in the exploratory data analysis, among the 11 active families, there were three families seen to have higher Shannon indices compared to the other families. These are *Bacteroidaceae*, *Lachnospiraceae* and *Ruminococcaceae*. At the OTU level, among the 15 active OTUs, *Otu00001* revealed the highest relative abundance across the time points. These findings are solely for exploratory analysis. In drawing inferences, the conclusions are based on the results from fitting the model to the data. At the kingdom level, the effect of treatment on the microbiome was significant at 5% level. Specifically, after multiplicity correction, there were significant differences in the microbiome between the treatment and control group at weeks 0 and 6. Furthermore, the analysis at family level exhibited that after multiplicity correction, *Lachnospiraceae*, *unclassified* and *Porphyromonadaceae* showed significant differences in the Shannon indices between the treatment and control group at weeks 0, while *Ruminococcaceae* has shown significant differences in microbiome between groups at weeks 0 and 4. In analysing the data at OTU level, the findings revealed that only *Otu00002* has significant results at week 0.

At kingdom level, the results from the T1D study dataset showed that there exists a significant difference in the Shannon indices between the treatment and control group, specifically at weeks 3 and 5. Moreover, after employing the multiplicity correction, significant results were observed for families *Lachnospiraceae* at weeks 3 and 5, *Bacteria* at week 5 and *S24-7* and *Unknown* family at week 3. At OTU level, *X226097bd7a1661a286a3b62d1c1f0e3a* and *X5eefe1c67a4852bd62c90dbcd2053008* showed significant results at week 3, *a8232b9e5fc8ad81ceda57fce3f52622* yielded significant result at week 5 and *d963b59f19db6517a9f26908f684545d* showed significant results at weeks 5 and 7. These findings were obtained after the multiplicity correction.

Generally, the results obtained from the data analysis align with the implications of the previously mentioned study on microbiome depletion [21],which suggests that the treatment has altered the diversity and composition of the microbiome. Additionally, the study highlighted changes in specific taxa, with some increasing and others decreasing after the treatment. These changes are also observed in the results of the analyses. In particular, from both datasets, the family *Lachnospiraceae* has only indicated significant differences in its microbial diversity, the Shannon index, only in specific time points. Also, in the analysis at the OTU level, the OTUs have exhibited significant at specific time points. Nevertheless, these results imply that changes in the alpha diversity can be brought by administering the treatment or intervention.

Moreover, a review on the effect of antibiotics on the human microbiome [12] suggested that the impact of treatment varies with the study duration. This is also reflected in the obtained results. These are also shown in the results from the multiple comparisons where the differences in the microbiome between groups at each time point are not the same. The difference in Shannon indices and relative abundances varies, it can be increasing or decreasing at each time point.

While the findings helped understand the effect of treatment or medical intervention on the microbial composition of the microbiome, there are some limitations to consider. Given the substantial datasets that were utilized in the analysis, a more comprehensive conclusion can be drawn to provide a clearer understanding in the changes in the microbiome at several time points. Likewise, this study does not aim to address biomarker detection but rather focused on assessing the impact of the treatment on the microbiome. Therefore, future research should incorporate a broader range of data to identify potential biomarkers associated with the diseases and to better understand the effects of the treatment, and hence provide biological interpretations.

8 References

[1] Celeste Allaband et al. "Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians". eng ; jpn. In: *Clinical Gastroenterology and Hepatology* 17.2 (2019), pp. 218–230. ISSN: 1542-3565.

[2] Yu Chen et al. "Association Between Gut Dysbiosis and Sepsis-Induced Myocardial Dysfunction in Patients With Sepsis or Septic Shock". In: *Frontiers in Cellular and Infection Microbiology* 12 (2022). ISSN: 2235-2988. DOI: 10.3389/fcimb.2022.857035. URL: https://www.frontiersin.org/articles/10.3389/fcimb.2022.857035.

[3] Dr. Peace Chikezie. *The Importance of Medical Research*. https://infiuss.com/blog/the-importance-of-medical-research-zRce. June 2022.

[4] Arnaud Cougoul et al. "Rarity of microbial species: In search of reliable associations". In: *PLoS ONE* 14 (2018). URL: https://api.semanticscholar.org/CorpusID:80627792.

[5] Matthew K. Doherty et al. "Fecal Microbiota Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease Patients". In: *mBio* 9.2 (2018), 10.1128/mbio.02120–17. DOI: 10.1128/mbio.02120-17. eprint: https://journals.asm.org/doi/pdf/10.1128/mbio.02120-17. URL: https://journals.asm.org/doi/abs/10.1128/mbio.02120-17.

[6] Gabriel A Gregory et al. "Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study". In: *The Lancet Diabetes Endocrinology* 10.10 (2022), pp. 741–760. ISSN: 2213-8587. DOI: https://doi.org/10.1016/S2213-8587(22)00218-2. URL: https://www.sciencedirect.com/science/article/pii/S2213858722002182.

[7] Rima Hajjo, Dima A. Sabbah, and Abdel Qader Al Bawab. "Unlocking the Potential of the Human Microbiome for Identifying Disease Diagnostic Biomarkers". In: *Diagnostics* 12.7 (2022). ISSN: 2075-4418. DOI: 10.3390/diagnostics12071742. URL: https://www.mdpi.com/2075-4418/12/7/1742.

[8] Abhishek Kaul et al. English. In: *Frontiers in Microbiology* 8.NOV (Nov. 2017). Publisher Copyright: © 2017 Kaul, Mandal, Davidov and Peddada. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02114.

[9] Chuck Kincaid. *Guidelines for Selecting the Covariance Structure in Mixed Model Analysis*. Director, SAS Center of Excellence, COMSYS Information Technology Services, Inc. 5278 Lovers Lane, Portage, MI 49002. URL: http://www.comsyssas.com.

[10] Sangseok Lee and Dong Kyu Lee. "What is the proper way to apply the multiple comparison test?" In: *Korean Journal of Anesthesiology* 71 (2018), pp. 353–360. URL: https://api.semanticscholar.org/CorpusID:52129871.

[11] H. Lin and S.D. Peddada. "Analysis of microbial compositions: a review of normalization and differential abundance analysis". In: *NPJ Biofilms Microbiomes* 6.1 (Dec. 2020), p. 60. DOI: 10.1038/s41522-020-00160-w.

[12] Kristien Nel Van Zyl et al. "Effect of antibiotics on the human microbiome: a systematic review". In: *International journal of antimicrobial agents* 59.2 (Feb. 2022), p. 106502. ISSN: 0924-8579. DOI: 10.1016/j.ijantimicag.2021.106502. URL: https://doi.org/10.1016/j.ijantimicag.2021.106502.

[13] Olajumoke Evangelina Owokotomo, Rudradev Sengupta, and Ziv Shkedy. "Development of Microbiome Biomarkers in Intervention Studies". In: *Journal of Applied Microbiology* 134.3 (2023), lxac052.

[14] Abdulelah Almutairdi Paulo Gustavo Kotze Christopher Ma and Remo Panaccione. "Clinical utility of ustekinumab in Crohn's disease". In: *Journal of Inflammation Research* 11 (2018). PMID: 29445293, pp. 35–47. DOI: 10.2147/JIR.S157358. eprint: https://www.tandfonline.com/doi/pdf/10.2147/JIR.S157358. URL: https://www.tandfonline.com/doi/abs/10.2147/JIR.S157358.

[15] George Pipis. *The Benjamini-Hochberg procedure (FDR) and P-Value Adjusted Explained*. https://www.r-bloggers.com/2023/07/the-benjamini-hochberg-procedure-fdr-and-p-value-adjusted-explained/. July 2023.

[16] IR Ranasinghe, C Tian, and R Hsu. *Crohn Disease*. Updated 2024 Feb 24. Available from: https://www.ncbi.nlm.nih.gov/books/NBK436021/. Treasure Island (FL): StatPearls Publishing, Jan. 2024.

[17] Steven P. MPH Sanderson II. *Unveiling the Magic of LOESS Regression in R: A Step-by-Step Guide with mtcars*. 2023. URL: https://example.com/loess_guide.

[18] Seyed Saeid Seyedian, Forogh Nokhostin, and Mehrdad Dargahi Malamir. "A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease". jpn ; eng. In: *Journal of Medicine and Life* 12.2 (2019), pp. 113–122. ISSN: 1844-122X.

[19] G. Verbeke and Molenberghs G. *Longitudinal Data Analysis (Course notes)*.

[20] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer, 1997.

[21] Siming Wang et al. "Antibiotic-Induced Microbiome Depletion Is Associated with Resilience in Mice after Chronic Social Defeat Stress". In: *SSRN Electronic Journal* (Jan. 2019). DOI: 10.2139/ssrn.3424192.

[22] Ruwen Zhou et al. "Data Pre-processing for Analyzing Microbiome Data – A Mini Review". In: *Computational and Structural Biotechnology Journal* 21 (Oct. 2023). DOI: 10.1016/j.csbj.2023.10.001.

9 Appendices

9.1 Appendix A. Tables and Figures

9.1.1 CERTIFI Study

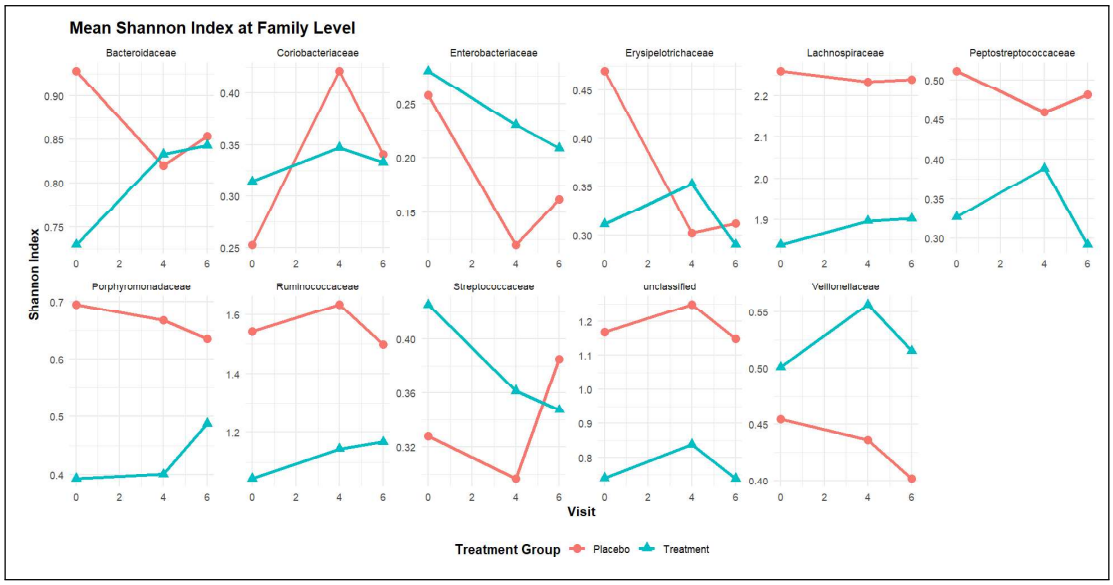


Figure 13: Mean Shannon Index - Family Level.

Table 35: Parameter Estimates - Coriobacteriaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.3144	0.04076	128	7.71	$< .0001$
Visit (4)	0.0329	0.0433	256	0.76	0.4480
Visit (6)	0.0185	0.0453	256	0.41	0.6823
Treatment	-0.0614	0.0863	256	-0.71	0.4771
Visit (4): Treatment	0.1355	0.0917	256	1.48	0.1408
Visit (6): Treatment	0.0693	0.0960	256	0.72	0.4713

Table 36: Least Squares Estimates (Coriobacteriaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	0.0615	0.0863	256	0.71	0.4771
Visit : Treatment	4	-0.0741	0.0885	256	-0.84	0.4036
Visit : Treatment	6	-0.0078	0.0822	256	-0.10	0.9243

Table 37: Parameter Estimates - Enterobacteriaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.2798	0.0277	128	10.12	< 0.0001
Visit (4)	-0.0497	0.0331	256	-1.50	0.1343
Visit (6)	-0.0709	0.0328	256	-2.16	0.0315
Treatment	-0.0223	0.0586	256	-0.38	0.7039
Visit (4): Treatment	-0.0877	0.0701	256	-1.25	0.2117
Visit (6): Treatment	-0.0246	0.0694	256	-0.35	0.7230

Table 38: Least Squares Estimates (Enterobacteriaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	0.0223	0.0586	256	0.38	0.7039
Visit : Treatment	4	0.1100	0.0562	256	1.96	0.0514
Visit : Treatment	6	0.0469	0.0544	256	0.86	0.3894

Table 39: Parameter Estimates - Erysipelotrichaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.3118	0.0447	128	6.97	< 0.0001
Visit (4)	0.0415	0.0439	256	0.95	0.3455
Visit (6)	-0.0213	0.0477	256	-0.45	0.6553
Treatment	0.1580	0.0947	256	1.67	0.0964
Visit (4): Treatment	-0.2088	0.0930	256	-2.25	0.0256
Visit (6): Treatment	-0.1359	0.1010	256	-1.35	0.1796

Table 40: Least Squares Estimates - Multiple comparisons (Erysipelotrichaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.1580	0.0947	256	-1.67	0.0964
Visit : Treatment	4	0.0507	0.0887	256	0.57	0.5679
Visit : Treatment	6	-0.0221	0.0866	256	-0.26	0.7987

Table 41: Parameter Estimates - Peptostreptococcaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.3271	0.0405	128	8.08	< 0.0001
Visit (4)	0.0613	0.0434	256	1.41	0.1586
Visit (6)	-0.0351	0.0362	256	-0.97	0.3320
Treatment	0.1843	0.0857	256	2.15	0.0325
Visit (4): Treatment	-0.1136	0.0918	256	-1.24	0.2174
Visit (6): Treatment	0.0057	0.0766	256	0.07	0.9412

Table 42: Least Squares Estimates - Multiple comparisons (Peptostreptococcaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.1843	0.0857	256	-2.15	0.0325
Visit : Treatment	4	-0.0708	0.0817	256	-0.87	0.3874
Visit : Treatment	6	-0.1900	0.0818	256	-2.32	0.0209

Table 43: Parameter Estimates - Porphyromonadaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.3929	0.0527	128	7.46	< 0.0001
Visit (4)	0.0075	0.0572	256	0.13	0.8964
Visit (6)	0.0950	0.0616	256	1.54	0.1240
Treatment	0.3014	0.1115	256	2.70	0.0073
Visit (4): Treatment	-0.0345	0.1211	256	-0.28	0.7761
Visit (6): Treatment	-0.1545	0.1303	256	-1.19	0.2368

Table 44: Least Squares Estimates - Multiple comparisons (Porphyromonadaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.3014	0.1115	256	-2.70	0.0073
Visit : Treatment	4	-0.2670	0.1151	256	-2.32	0.0212
Visit : Treatment	6	-0.1469	0.1121	256	-1.31	0.1913

Table 45: Parameter Estimates - Streptococcaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.4251	0.0354	128	11.99	< 0.0001
Visit (4)	-0.0638	0.0399	256	-1.60	0.1106
Visit (6)	-0.0777	0.0437	256	-1.78	0.0763
Treatment	-0.0972	0.0750	256	-1.29	0.1966
Visit (4): Treatment	0.0326	0.0844	256	0.39	0.6995
Visit (6): Treatment	0.1348	0.0925	256	1.46	0.1460

Table 46: Least Squares Estimates - Multiple comparisons (Streptococcaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	0.0972	0.0750	256	1.29	0.1966
Visit : Treatment	4	0.0645	0.0676	256	0.95	0.3406
Visit : Treatment	6	-0.0377	0.0701	256	-0.54	0.5911

Table 47: Parameter Estimates - Unclassified Family.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.7410	0.08026	128	9.23	< 0.0001
Visit (4)	0.0973	0.0858	256	1.13	0.2577
Visit (6)	-0.0019	0.0804	256	-0.02	0.9814
Treatment	0.4267	0.1699	256	2.51	0.0127
Visit (4): Treatment	-0.0168	0.1816	256	-0.09	0.9263
Visit (6): Treatment	-0.0168	0.1702	256	-0.10	0.9213

Table 48: Least Squares Estimates - Multiple comparisons (Unclassified Family).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.4267	0.1699	256	-2.51	0.0127
Visit : Treatment	4	-0.4099	0.1758	256	-2.33	0.0205
Visit : Treatment	6	-0.4098	0.1592	256	-2.57	0.0106

Table 49: Parameter Estimates - Veillonellaceae.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.5010	0.0416	128	12.06	< 0.0001
Visit (4)	0.0555	0.0508	256	1.09	0.2761
Visit (6)	0.0147	0.0543	256	0.27	0.7865
Treatment	-0.0462	0.0880	256	-0.53	0.5996
Visit (4): Treatment	-0.0739	0.1076	256	-0.69	0.4931
Visit (6): Treatment	-0.0672	0.1150	256	-0.58	0.5593

Table 50: Least Squares Estimates - Multiple comparisons (Veillonellaceae).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	0.0462	0.0880	256	0.53	0.5996
Visit : Treatment	4	0.1201	0.0907	256	1.32	0.1866
Visit : Treatment	6	0.1135	0.0918	256	1.24	0.2174

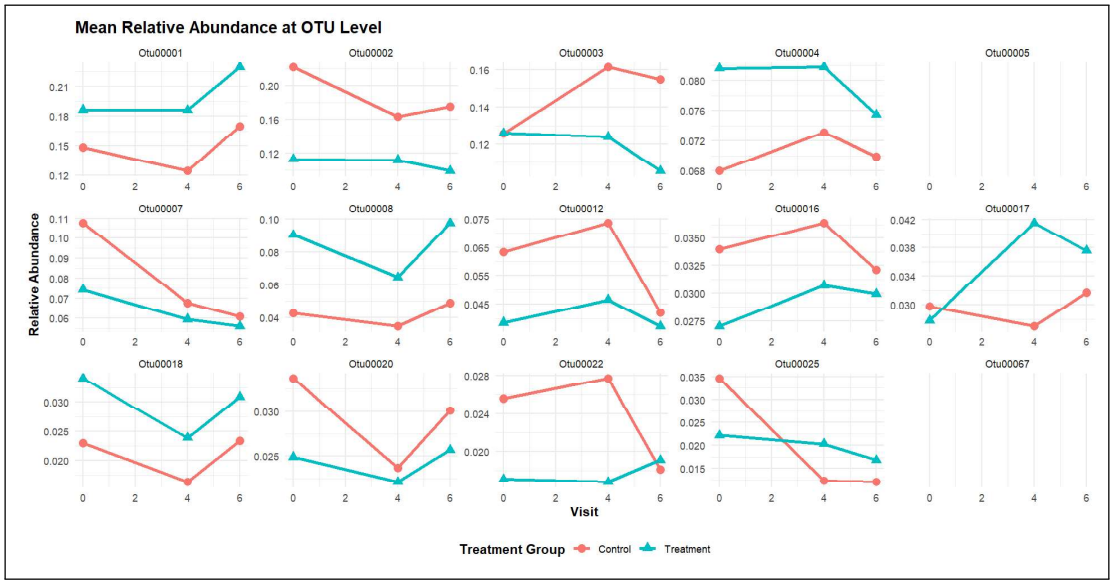


Figure 14: Mean Relative Abundance - OTU Level.

Table 51: Parameter Estimates - Otu00002.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.1127	0.0171	128	6.60	< 0.0001
Visit (4)	-0.0009	0.0153	256	-0.06	0.9542
Visit (6)	-0.0139	0.0186	256	-0.75	0.4557
Treatment	0.1094	0.0361	256	3.03	0.0027
Visit (4): Treatment	-0.0577	0.0323	256	-1.78	0.0755
Visit (6): Treatment	-0.0326	0.0395	256	-0.83	0.4089

Table 52: Least Squares Estimates - Multiple comparisons (Otu00002).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.1094	0.0361	256	-3.03	0.0027
Visit : Treatment	4	-0.0517	0.0337	256	-1.53	0.1265
Visit : Treatment	6	-0.0768	0.0337	256	-2.28	0.0237

Table 53: Parameter Estimates - Otu00004.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0816	0.0130	128	6.29	< 0.0001
Visit (4)	0.0002	0.0130	256	0.01	0.9887
Visit (6)	-0.0061	0.0156	256	-0.39	0.6981
Treatment	-0.0135	0.0275	256	-0.49	0.6225
Visit (4): Treatment	0.0049	0.0274	256	0.18	0.8593
Visit (6): Treatment	0.0079	0.0331	256	0.24	0.8107

Table 54: Least Squares Estimates - Multiple comparisons (Otu00004).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	0.0135	0.0276	256	0.49	0.6225
Visit : Treatment	4	0.0087	0.0271	256	0.32	0.7492
Visit : Treatment	6	0.0057	0.0250	256	0.22	0.8230

Table 55: Parameter Estimates - Otu00007.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0742	0.0139	128	5.34	< 0.0001
Visit (4)	-0.0146	0.0111	256	-1.32	0.1877
Visit (6)	-0.0182	0.0106	256	-1.72	0.0861
Treatment	0.0334	0.0295	256	1.14	0.2574
Visit (4): Treatment	-0.0255	0.0234	256	-1.09	0.2764
Visit (6): Treatment	-0.0286	0.0223	256	-1.28	0.2024

Table 56: Least Squares Estimates - Multiple comparisons (Otu00007).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.0334	0.0295	256	-1.14	0.2574
Visit : Treatment	4	-0.0079	0.0214	256	-0.37	0.7116
Visit : Treatment	6	-0.0049	0.0212	256	-0.23	0.8180

Table 57: Parameter Estimates - Otu00012.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0386	0.0071	128	5.45	< 0.0001
Visit (4)	0.0078	0.0081	256	0.97	0.3330
Visit (6)	-0.0013	0.0054	256	-0.25	0.8062
Treatment	0.0247	0.0150	256	1.65	0.1008
Visit (4): Treatment	0.0024	0.0171	256	0.14	0.8901
Visit (6): Treatment	-0.0199	0.0115	256	-1.74	0.0837

Table 58: Least Squares Estimates - Multiple comparisons (Otu00012).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.0247	0.0150	256	-1.65	0.1008
Visit : Treatment	4	-0.0271	0.0151	256	-1.79	0.0748
Visit : Treatment	6	-0.0048	0.0125	256	-0.39	0.7002

Table 59: Parameter Estimates - Otu00016.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0270	0.0067	128	4.05	< 0.0001
Visit (4)	0.0037	0.0048	256	0.77	0.4414
Visit (6)	0.0029	0.0050	256	0.57	0.5668
Treatment	0.0069	0.0141	256	0.49	0.6250
Visit (4): Treatment	-0.0012	0.0101	256	-0.12	0.9041
Visit (6): Treatment	-0.0047	0.0107	256	-0.44	0.6586

Table 60: Least Squares Estimates - Multiple comparisons (Otu00016).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.0069	0.0141	256	-0.49	0.6250
Visit : Treatment	4	-0.0057	0.0129	256	-0.44	0.6593
Visit : Treatment	6	-0.0022	0.0141	256	-0.16	0.8756

Table 61: Parameter Estimates - Otu00017.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0278	0.0052	128	5.31	< 0.0001
Visit (4)	0.0138	0.0074	256	1.86	0.0641
Visit (6)	0.0098	0.0067	256	1.47	0.1419
Treatment	0.0021	0.0111	256	0.19	0.8528
Visit (4): Treatment	-0.0166	0.0157	256	-1.06	0.2896
Visit (6): Treatment	-0.0079	0.0141	256	-0.56	0.5766

Table 62: Least Squares Estimates - Multiple comparisons (Otu00017).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.0021	0.0111	256	-0.19	0.8528
Visit : Treatment	4	0.0146	0.0158	256	0.92	0.3575
Visit : Treatment	6	0.0058	0.0156	256	0.37	0.7089

Table 63: Parameter Estimates - Otu00018.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0341	0.0090	128	3.77	0.0002
Visit (4)	-0.0102	0.0089	256	-1.15	0.2520
Visit (6)	-0.0033	0.0094	256	-0.35	0.7286
Treatment	-0.0111	0.0192	256	-0.58	0.5624
Visit (4): Treatment	0.0035	0.0188	256	0.19	0.8522
Visit (6): Treatment	0.0037	0.0199	256	0.19	0.8533

Table 64: Least Squares Estimates - Multiple comparisons (Otu00018).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	0.0111	0.0192	256	0.58	0.5624
Visit : Treatment	4	0.0076	0.0120	256	0.69	0.4886
Visit : Treatment	6	0.0074	0.0138	256	0.54	0.5915

Table 65: Parameter Estimates - Otu00020.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0249	0.0068	128	3.65	0.0004
Visit (4)	-0.0028	0.0064	256	-0.43	0.6685
Visit (6)	0.0008	0.0052	256	0.15	0.8809
Treatment	0.0088	0.0144	256	0.61	0.5433
Visit (4): Treatment	-0.0072	0.0136	256	-0.53	0.5972
Visit (6): Treatment	-0.0044	0.0110	256	-0.40	0.6931

Table 66: Least Squares Estimates - Multiple comparisons (Otu00020).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.0088	0.0144	256	-0.61	0.5433
Visit : Treatment	4	-0.0016	0.0094	256	-0.17	0.8687
Visit : Treatment	6	-0.0044	0.0122	256	-0.36	0.7173

Table 67: Parameter Estimates - Otu00022.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0171	0.0041	128	4.20	< 0.0001
Visit (4)	-0.0003	0.0056	256	-0.05	0.9579
Visit (6)	0.0019	0.0056	256	0.34	0.7319
Treatment	0.0085	0.0086	256	0.98	0.3288
Visit (4): Treatment	0.0024	0.0119	256	0.20	0.8388
Visit (6): Treatment	-0.0095	0.0118	256	-0.80	0.4242

Table 68: Least Squares Estimates - Multiple comparisons (Otu00022).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.0085	0.0086	256	-0.98	0.3288
Visit : Treatment	4	-0.0109	0.0094	256	-1.16	0.2472
Visit : Treatment	6	0.0010	0.0089	256	0.11	0.9101

Table 69: Parameter Estimates - Otu00025.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0222	0.0067	128	3.31	0.0012
Visit (4)	-0.0020	0.0054	256	-0.37	0.7153
Visit (6)	-0.0054	0.0062	256	-0.87	0.3826
Treatment	0.0123	0.0142	256	0.87	0.3868
Visit (4): Treatment	-0.0203	0.0115	256	-1.76	0.0802
Visit (6): Treatment	-0.0171	0.0131	256	-1.30	0.1931

Table 70: Least Squares Estimates - Multiple comparisons (Otu00025).

Effect	Visit	Estimate	SE	DF	Test Statistic	p-value
Visit : Treatment	0	-0.0123	0.0142	256	-0.87	0.3868
Visit : Treatment	4	0.0079	0.0108	256	0.74	0.4627
Visit : Treatment	6	0.0048	0.0069	256	0.70	0.4862

9.1.2 T1D Study

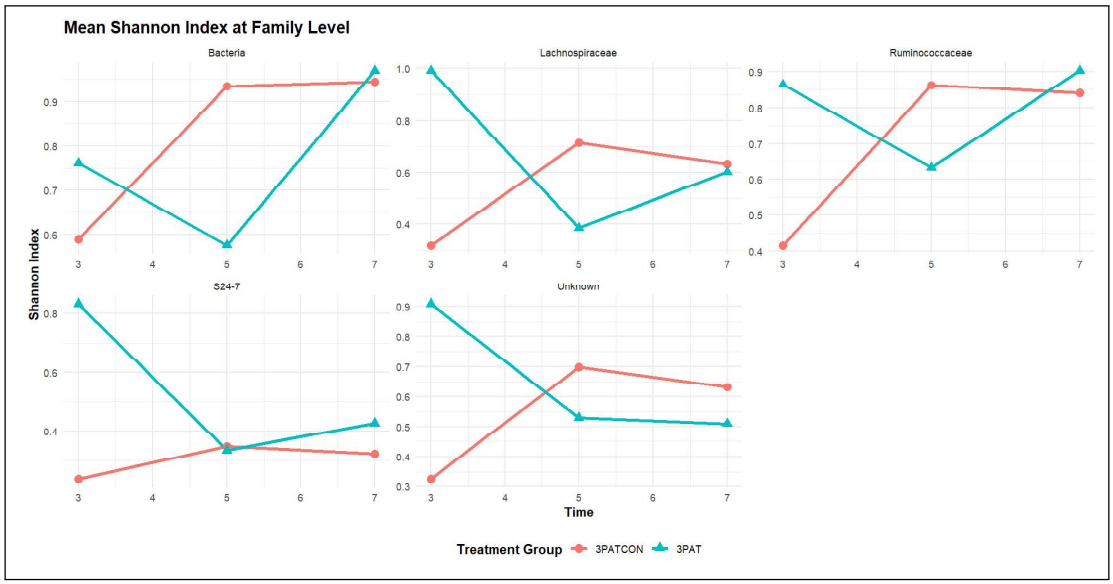


Figure 15: Mean Shannon Index - Family Level.

Table 71: Parameter Estimates - Bacteria.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.7616	0.0826	79	9.22	< 0.0001
Time (5)	-0.1852	0.1079	158	-1.72	0.0880
Time (7)	0.2077	0.0812	158	2.56	0.0115
Treatment	-0.1722	0.1160	158	-1.48	0.1397
Time (5): Treatment	0.5294	0.1516	158	3.49	0.0006
Time (7): Treatment	0.1468	0.1141	158	1.29	0.2002

Table 72: Least Squares Estimates - Multiple comparisons (Bacteria).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value
Time : Treatment	3	0.1722	0.1160	158	1.48	0.1397
Time : Treatment	5	-0.3572	0.0836	158	-4.27	< 0.0001
Time : Treatment	7	0.0255	0.0834	158	0.31	0.7604

Table 73: Parameter Estimates - Unknown Family.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.9078	0.0954	79	9.52	< 0.0001
Time (5)	-0.3784	0.1646	158	-2.30	0.0228
Time (7)	-0.3985	0.1164	158	-3.42	0.0008
Treatment	-0.5809	0.1340	158	-4.33	< 0.0001
Time (5): Treatment	0.7500	0.2314	158	3.24	0.0014
Time (7): Treatment	0.7032	0.1636	158	4.30	< 0.0001

Table 74: Least Squares Estimates - Multiple comparisons (Unknown Family).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value
Time : Treatment	3	0.5809	0.1340	158	4.33	< 0.0001
Time : Treatment	5	-0.1691	0.1603	158	-1.05	0.2932
Time : Treatment	7	-0.1222	0.1407	158	-0.87	0.3863

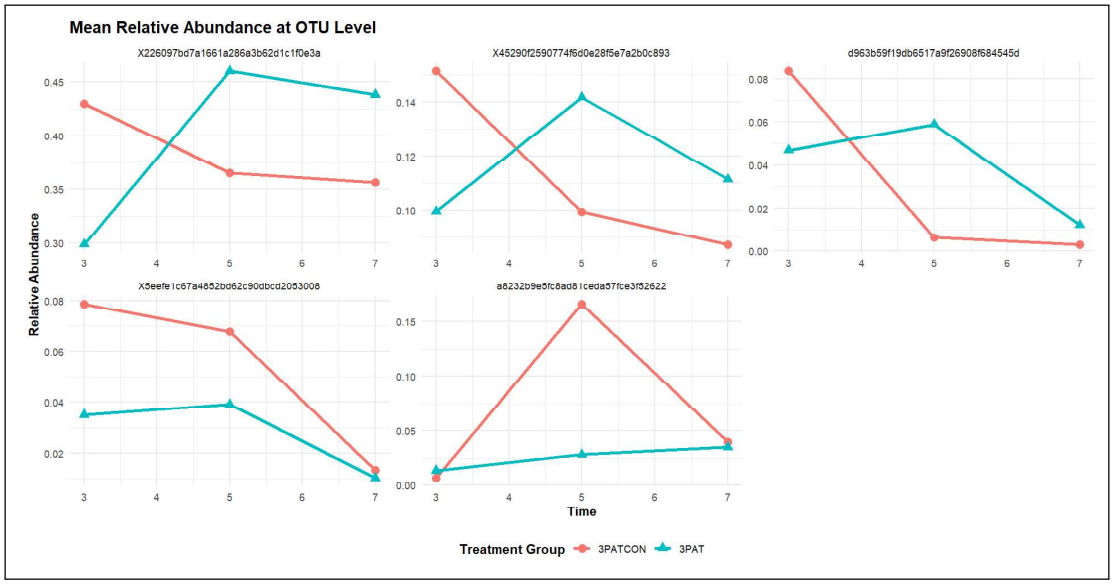


Figure 16: Mean Count - OTU Level.

Table 75: Parameter Estimates - d963b59f19db6517a9f26908f684545d.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0469	0.0167	79	2.80	0.0064
Time (5)	0.0120	0.0209	158	0.57	0.5671
Time (7)	-0.0345	0.0168	158	-2.06	0.0414
Treatment	0.0367	0.0235	158	1.56	0.1202
Time (5): Treatment	-0.0893	0.0294	158	-3.04	0.0028
Time (7): Treatment	-0.0460	0.0236	158	-1.95	0.0532

Table 76: Least Squares Estimates - Multiple comparisons (d963b59f19db6517a9f26908f684545d).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value
Time : Treatment	3	-0.0367	0.0235	158	-1.56	0.1202
Time : Treatment	5	0.0526	0.0150	158	3.52	0.0006
Time : Treatment	7	0.0093	0.0021	158	4.34	< 0.0001

Table 77: Parameter Estimates - a8232b9e5fc8ad81ceda57fce3f52622.

Effect	Estimate	SE	DF	Test Statistic	p-value
Intercept	0.0136	0.0046	79	2.98	0.0039
Time (5)	0.0146	0.0313	158	0.47	0.6416
Time (7)	0.0215	0.0117	158	1.83	0.0690
Treatment	-0.0069	0.0064	158	-1.06	0.2889
Time (5): Treatment	0.1445	0.0439	158	3.29	0.0012
Time (7): Treatment	0.0118	0.0165	158	0.72	0.4751

Table 78: Least Squares Estimates - Multiple comparisons (a8232b9e5fc8ad81ceda57fce3f52622).

Effect	Time	Estimate	SE	DF	Test Statistic	p-value
Time : Treatment	3	0.0069	0.0064	158	1.06	0.2889
Time : Treatment	5	-0.1376	0.0431	158	-3.19	0.0017
Time : Treatment	7	-0.0049	0.0161	158	-0.31	0.7594

Appendix B. R/SAS Codes

```
########### MODEL FITTING KINGDOM LEVEL ##############
########### SHANNON IS USED ON FILTERED DATA #########################
## First filter: remove OTUs with all 0s, from 2353 OTUs we now have
# otus by sample
otu_w0 = create_OTU(week0_otu2_t) #939
otu_w4 = create_OTU(week4_otu2_t) #1044
otu_w6 = create_OTU(week6_otu2_t) #932

# transpose the above matrix #
# sample by otus
otu_w0_t = t(otu_w0)
otu_w4_t = t(otu_w4)
otu_w6_t = t(otu_w6)

# computes all measures at once, this time filtered data
w0 = estimate_richness(otu_table(otu_w0_t,taxa_are_rows = FALSE))
w4 = estimate_richness(otu_table(otu_w4_t,taxa_are_rows = FALSE))
w6 = estimate_richness(otu_table(otu_w6_t,taxa_are_rows = FALSE))

# extracts only the Shannon index #
w0_shan = w0[,6]
w4_shan = w4[,6]
w6_shan = w6[,6]

# combines the sample info and shannon index #
w0_king = cbind(sample_w0, Shannon = w0_shan)
w4_king = cbind(sample_w4, Shannon = w4_shan)
```

```
w6_king = cbind(sample_w6, Shannon = w6_shan)

# longitudinal format of the data #
long_data_king = rbind(w0_king, w4_king, w6_king)
long_data_king <- long_data_king %>% arrange(USUBJID)
long_data_king <- long_data_king %>% mutate(id = group_indices(., USUBJID))
long_data_king$visitclass <- long_data_king$visit


############## MODEL FITTING FAMILY LEVEL ####################
# create a function that creates long format of family data then fits an LMM #
# needs fam_shan reduced per week and sample per week
long_data_family <- function(family, fam_shan_reduced_w0, fam_shan_reduced_w4,
                             fam_shan_reduced_w6, sample_w0,
                             sample_w4, sample_w6)
{
family_w0_fit = t(fam_shan_reduced_w0[family,])
family_w4_fit = t(fam_shan_reduced_w4[family,])
family_w6_fit = t(fam_shan_reduced_w6[family,])


family_w0_samp = cbind(sample_w0, family_w0_fit)
family_w4_samp = cbind(sample_w4, family_w4_fit)
family_w6_samp = cbind(sample_w6, family_w6_fit)

long_format = rbind(family_w0_samp, family_w4_samp, family_w6_samp)
long_format = long_format %>% arrange(USUBJID) %>% mutate(id = group_indices(., USUBJID),
                         visitclass = visit)

output_file <- paste0(family, ".csv")
write.csv(long_format, output_file, row.names = FALSE)
return(long_format)}

family_longdata <- lapply(common_families, function(family) {
  long_data_family(family, fam_shan_reduced_w0, fam_shan_reduced_w4, fam_shan_reduced_w6,
  sample_w0, sample_w4, sample_w6)})


############## MODEL FITTING OTU LEVEL ####################
# create a function that creates long format of OTU data then fits an LMM #
# needs otu_RA  per week and sample per week
long_data_otu <- function(otu, otu_RA_w0, otu_RA_w4,
                             otu_RA_w6, sample_w0, sample_w4, sample_w6)
{
  otu_w0_fit = t(otu_RA_w0[otu,])
  otu_w4_fit = t(otu_RA_w4[otu,])
  otu_w6_fit = t(otu_RA_w6[otu,])


  otu_w0_samp = cbind(sample_w0, otu_w0_fit)
```

```
  otu_w4_samp = cbind(sample_w4, otu_w4_fit)
  otu_w6_samp = cbind(sample_w6, otu_w6_fit)

  long_format = rbind(otu_w0_samp, otu_w4_samp, otu_w6_samp)
  long_format = long_format %>% arrange(USUBJID) %>% mutate(id = group_indices(., USUBJID),
                                                            visitclass = visit)
  output_file <- paste0(otu, ".csv")
  write.csv(long_format, output_file, row.names = FALSE)
  return(long_format)}

# using the function #
otu_longdata <- lapply(common_otu, function(otu) {
  long_data_otu(otu, otu_RA_w0, otu_RA_w4, otu_RA_w6, sample_w0, sample_w4, sample_w6)
})


###### MULTIPLICITY CORRECTION ####
# same code for family and OTU levels #
kingdom_mean <- read.csv("doherty_kingdom_means.csv")
kingdom_mean <- kingdom_mean %>%
mutate(Treatment = factor(Treatment, levels = c(0, 1), labels = c("Placebo", "Treatment")))

kingdom_pval <- read.csv("doherty_kingdom_pval.csv")
adjusted_p_values <- lapply(kingdom_pval[2:4], p.adjust, method = "BH")
names(adjusted_p_values) <- paste0(names(kingdom_pval)[2:4], "_adj")
fdr_table_kingdom <- cbind(kingdom_pval, adjusted_p_values)


##### MODEL FITTING IN SAS #####
/* final model and we do contrast using lm estimate */
proc mixed data=doherty_long method=reml;
class visitclass visit TRTGR_new;
model Shannon = visit TRTGR_new visit*TRTGR_new / solution;
random intercept / subject=id g gcorr v vcorr;
repeated visitclass/ type=un subject=id;

/* Compute least squares means for visit*TRTGR_new */
lsmeans visit*TRTGR_new / cl;

/* differences among means */
lsmestimate visit*TRTGR_new 'Group0 vs Group1 at Visit 0' -1 1  0 0   0 0 /e;
lsmestimate visit*TRTGR_new 'Group0 vs Group1 at Visit 4' 0  0  -1 1  0  0 /e;
lsmestimate visit*TRTGR_new 'Group0 vs Group1 at Visit 6' 0  0  0  0  -1 1/e;run;
```