Microbial determinants of obesity, a cross-sectional study

Peihan Song (Student ID:2703170S)

Supervisor: Dr Umer Zeeshan Ijaz

Co-Supervisor: Joe Gibbs

August 26, 2022

MSc Computer Systems Engineering

School of Engineering

Abstract

According to the World Health Organization, from 1975 to 2016, the number of overweight or obese people almost tripled. Nowadays, obesity has become a widespread health problem in the world. Based on the aetiology of obesity, obesity can be classified as simple obesity and hypothalamic obesity. Simple obesity can be attributed to multiple factors, including sedentary work and a diet high in sugars, fats, and salt.

There is more and more evidence proving that gut microbiota is important for human health in many ways, including immunity, metabolism, and nutrition.

The present research used multiple bioinformatic and statistical approaches, including alpha diversity analysis, beta diversity analysis, NRI/NTI analysis, NST, analysis, QPE analysis, and differential analysis. The aim of the present study is to investigate the relationship between gut microbiota, SCFAs, and obesity.

The results of the present study showed significant connections between gut microbiota, SCFAs and obesity. The gut microbiota is capable of producing short-chain fatty acids that can influence host metabolism and appetite, which in turn affects host nutrient and energy utilization and ultimately host body weight; in turn, the content of short-chain fatty acids affects the composition of the gut microbiota

*Keywords: Gut Microbiota; Obesity; Bioinformatics; SCFA*s; *NGS*

Table of contents

Abbreviations

ANOVA	Analysis of Variance
ASV	Amplicon Sequence Variant
BMI	Body Mass Index
βMNTD	Beta Mean Nearest Taxon Distance
βNTI	Beta Nearest Taxon Index
C2	Acetate
C3	Propionate
C4	Butyrate
C5	Valerate
C6	Caproate
C7	Enanthate
C8	Caprylate
ePCR	emulsion Polymerase Chain Reaction
FFAR 2 & 3	Free Fatty Acid Receptor 2 and 3
GLM	Generalised Linear Model
GPR 43 & 41	G Protein Coupled Receptor
IC4	Iso-Butyrate
IC5	Iso-Valerate
LDL	Low Density Lipoprotein
MNTD	Mean Nearest Taxon Distance
MPD	Mean Phylogenetic Distance
NB	Negative Binomial
NGS	Next-Generation Sequencing
NIDDM	Non-Insulin-Dependent Diabetes Mellitus
NRI	Net Relatedness Index
NSS	Normalised Selection Strength
NST	Normalised Stochasticity Ratio
NTI	Nearest Taxon Index
OTU	Operational Taxonomic Unit

PERMANOVA	Permutational Analysis of Variance
PWS	Prader-Willi Syndrome
QPE	Quantitative Process Estimates
RR	Relative Risk
SCFA	Short-Chain Fatty Acids
SD	Standard Deviation
SDS	Standard Deviation Score
SES	Standardised Effect Size
SS	Selection Strength
ST	Stochasticity Ratio
WHO	World Health Organization

Acknowledgements

Thank you to Dr Umer Zeeshan Ijaz for the help and guidance in bioinformatics. Thanks to Dr Uzma and Sara Bibi for helping me understand the topics covered in this study. Thank you so much to Dr Muhammad Jaffar Khan for providing the dataset used in this study. Thanks to the University of Glasgow for providing me with the chance to improve my knowledge and skills. Thank you to my teammates for supporting and helping me in my study.

Chapter 1

Introduction

1.1 Obesity

1.1.1 BMI and BMI SDS

Body Mass Index (BMI) is a commonly used weight-for-height metric for classifying overweight and obesity in adults, which is defined as

$$BMI = \frac{m_{kg}}{l_m^2}. \quad (1)$$

Where m_{kg} is the body weight in kilograms, l_m is height in metres.

However, the World Health Organization (WHO) uses BMI Standard Deviation Score (BMI SDS) to describe BMI for children. Also, the age of the child should be taken into consideration.

1.1.2 Obesity in adults

The WHO (2021) defined adult overweight as BMI equal to or greater than 25 and adult obesity as BMI equal to or greater than 30.

In 2016, according to the WHO (2021), more than 1.9 billion adults aged 18 or older were reported as overweight, among which more than 650 million adults were obese.

1.1.3 Obesity in children

For children under five years, the WHO (2021) defined overweight as BMI SDS greater than 2 SD above the WHO Child Growth Standards median and obesity as BMI SDS greater than 3 SD above the WHO Child Growth Standards median.

For children and adolescents between 5-19 years of age, the WHO (2021) defined overweight as BMI SDS greater than 1 SD above the WHO Child Growth Standards median and obesity as BMI SDS greater than 2 SD above the WHO Child Growth Standards median. The WHO estimated that in 2016 there were more than 340 million overweight or obese children and adolescents between 5-19 years of age, and around 38.2 million children were overweight or obese in 2019.

1.1.4 Risk factors contributing to childhood obesity

Obesity is a disease caused by known and unknown factors. Known causes of Obesity include genetic hormone deficiency or malfunctioning of the hypothalamic satiety centre. The hypothalamic satiety centre malfunctioning might be caused by chromosomal diseases like Prader-Willi Syndrome (PWS) or a tumour's (e, g. craniopharyngioma) erosion. In the present study, the obesity caused by hypothalamic satiety centre malfunctioning is called “hypothalamic obesity”. Unfortunately, a definitive risk factor cannot explain most of the global obesity epidemic. Those forms of obesity are called “simple obesity” in the present study.

1.1.4 Risk factors contributing to simple obesity

According to the WHO (2021), from 1975 to 2016, the prevalence of obesity worldwide almost tripled. The energy imbalance between consumed and expended calories; The increased ingestion of high fat and high sugar foods; and an increase in inactivity due to sedentary occupations, modern modes of transportation and urbanisation. All of those can lead to obesity.

1.1.5 Hypothalamic disorder and obesity

The Prader-Willi Syndrome (PWS) is the primary cause of syndromic obesity and a primary cause of metabolic problems in the hypothalamic obesity group; its most significant symptom is insatiable hunger, which causes the children’s change in behaviour, causing children to ingest food excessively.

According to Lustig and Mueller (2011), the pathology and symptomology can be described as "organic leptin resistance". This means a failure in leptin signalling in the afferent arm, caused by hypothalamic damage, thereby leading to the efferent arm's autonomic dysfunction, promoting inadequate energy expenditure and excessive energy storage. (Lustig and Mueller, 2011)

1.1.6 Complications of Obesity

According to Must and Strass (1999), the complications of childhood obesity are shown in Table 1.

Table1. complications of obesity

Short-term		
System	Disease/Symptoms	Risk
Orthopaedic	Slipped capital epiphysis	50% - 70%
	Blunt's disease, bowing of the legs and tibial torsion in response to unequal or early excess weight bearing	80%
Pulmonary	Asthma	30%
	Decrease of at least 15% in performance with exercise	More than 80%
	Sleep apnea with hypoventilation	Up to 94%
	Significant decrements in learning and memory function due to sleep apnea	
	Pickwickian syndrome, hypoventilation, somnolence, polycythemia and right ventricular hypertrophy and failure due to severe obesity	
Gastroenterological	Steatohepatitis due to insulin resistance	20% - 25%
	Steatohepatitis in severe obesity	40% - 50%
	Gallstones	8% - 33%
Endocrine	Reduction of insulin-stimulated glucose uptake	
	Higher levels of total cholesterol, Low Density Lipoprotein (LDL) cholesterol, and triglycerides	
	Non-Insulin-Dependent Diabetes Mellitus (NIDDM) in overweight adolescents	2.4%
	NIDDM in obesity	90%
	Menstrual abnormalities	
	Polycystic ovary Syndrome	
Social and economic	Poor emotional development	
	Dieting	50%

	Smoking to control weight	20%
	Bulimic (eating disorder) in adulthood	40%
	Mid-term	
System	Disease/Symptoms	Risk
Cardiovascular	Elevated systolic or diastolic blood pressure in obese children between the ages of 5-11y	20% - 30%
	High blood pressure in obese boys and girls	9 – 10-fold
	High blood pressure in overweight adolescents	8.5-fold
	Deleterious effects on total cholesterol and LDL-cholesterol in adulthood	
Persistence	Obese adolescence continues to be obese in adulthood	25% - 50%, vary by gender
	Long-term	
System	Disease/Symptoms	Risk
Adult morbidity	Risk of heart disease	1.5 Relative Risk (RR)
	Risk of atherosclerosis	
	Risk of colon cancer and gout in males	
	Risk of arthritis, hip fracture, and difficulty with activities of daily living in females	
Adult mortality	All-cause heart disease (independent of weight status or smoking)	2.0 RR
	Coronary heart disease (independent of weight status or smoking)	

1.1.7 Management of obesity

Obesity is a disorder caused by multiple known and unknown factors. Obesity interventions require not only individual efforts but also environmental and societal support. For individuals, the change in diet (e, g. reducing the ingestion of foods that are high in calories, fats and sugars, increasing the intake of fruits, vegetables, whole grains, nuts, and legumes) and lifestyles (e, g. doing sports for 60 minutes a day) is effective for preventing overweight and obesity. Nowadays, society also plays an

indispensable role in promoting healthy lifestyles. For instance, the food industry can reduce the sugars, fats, and salt in processed foods, limit the advertising of foods high in fats, sugars, and salt, especially to children and teenagers, and make healthy foods inexpensive and obtainable for every consumer.

1.2 Gut microbiota, short-chain fatty acids, and obesity

The gut microbiota is a dynamic ecosystem coevolving with its human host, which accounts for 1 kilogram of human body weight. (Gérard, 2016)

limit the advertising. According to Gérard (2016), the gut microbiota "ferments otherwise indigestible food components, synthesises vitamins and other essential micronutrients, metabolises dietary toxins and carcinogens, converts cholesterol and bile acids, assures the maturation of the immune system, affects the growth and differentiation of enterocytes, regulates intestinal angiogenesis, and protects against enteric pathogens".

The gut microbiota can process dietary plant polysaccharides that are otherwise unreachable to humans into monosaccharides and Short-Chain Fatty Acids (SCFAs), principally acetate, propionate, and butyrate, which provide roughly 10% energy supply in omnivores but also act as signalling molecules, influencing energy ingestion and metabolism; moreover, SCFAs are ligands for the Free Fatty Acid Receptor 2 or the G-Protein Coupled Receptor 43 (FFAR2 or GPR43), and the Free Fatty Acid Receptor 3 or the G-Protein Coupled Receptor 41 (FFAR3 or GPR41); it is also suggested that the fibre administration results in increased production of SCFAs and thereby leads to increased satiety and reduced ingestion of foods. (Gérard, 2016)

1.3 Next-Generation Sequencing approach for analysing gut microbiota

Recent years have seen the shift of sequencing tools from traditional Sanger sequencing technology to Next-Generation Sequencing (NGS) technology. (Panek et al., 2018)

The Sanger sequencing technology, due to its low error rate, long read length, and large insert size, was considered the "gold standard", which would improve the outcomes of assembly for shotgun data; however, Sanger sequencing is labour-

intensive when it comes to its associated bias against the genes that are toxic for the cloning host. (Thomas et al., 2012)

The shotgun metagenomics technology provides complete information on the sample's gene pool; however, the amount of data generated from this technology is much too high and requires extensive effort in sequence analysis. (Panek et al., 2018)

16S rRNA amplicon sequencing provides cost-effectiveness, adequate resolution and sequencing depth and covers variable gene regions; it is the optimal method for bacterial community composition research on clinical and environmental samples. (Thomas et al., 2012; Panek et al., 2018)

The emulsion polymerase chain reaction (ePCR) is commonly used for amplifying DNA molecules in physically separated picolitre-volume water-in-oil droplets that act as reduced "reaction tubes"; the ePCR technology is used for determining the number of copies with digital droplet PCR, and preparative-scale applications, such as NGS for RNA profiling, molecular evolution, and genome-scale DNA and aptamer library construction; however the shortcoming of the conventional ePCR method for preparative-scale applications includes the hazardous organic solvents (like diethyl ether (DEE) and butanol) are used in breaking the emulsion, the silica-based columns are used in purifying PCR products, and those chemicals, even in trace amounts, will possibly interfere with the downstream applications, thus need to be removed by taking extra steps. (Verma et al., 2020)

Verma et al. (2020) developed two novel ePCR methods without hazardous organic solvents, which are proven to be equally effective for purifying PCR products after ePCR as the traditional ePCR method: a) the "spin + column" method avoids using DEE but involves centrifuging the emulsion to remove the oil and then using the QIA PB buffer to break the emulsion, and then the QIAquick spin columns are adopted to purify the PCR product; b) the "Quick ePCR extraction protocol" involves directly adding the QIA PB buffer to the PCR product for breaking the emulsion and then use the QIAquick spin columns to purify the PCR product, thus simplifying the process of

extracting DNA after ePCR and making it amenable with high-throughput applications.

1.4 Operational Taxonomic Units

The concept of the Operational Taxonomy Units (OTUs) was first coined by Sneath and Sokal (1963), meaning the "thing (s) being studied". The "thing(s)" concept is quite broad; it could be an organism, a named taxonomic group like species or genus, or even a group with undetermined evolutionary relationships that shares a set of traits. (Edgar, 2017)

Wayne et al. (1987) coined that, in general, the phylogenetic definition of a species would include about 70% or greater DNA-DNA relatedness between strains. Stackebrandt and Goebel (1994) found that 97% similarity of 16S rRNA sequences corresponds to around 70% of DNA reassociation in bacteriology.

However, the clustering threshold of 97% identity is not always accurate. Edgar (2018a) used a blinded test, and he found out 249490 identical sequences were annotated conflictedly in the SILVA release 128 (SILVA, 2018) and the Greengenes 13.5 (Second Genome, 2018) database at ranks up to phylum level, showing that the annotation error rate of those databases is around 17%. Edgar (2018b) used a large set of high-quality, full-length 16S rRNA sequences to assess the OTUs' correspondence to species; as a result, the conventional 97% threshold was proven too low. Edgar (2018b) also suggested that the optimal thresholds should be increased to at least 99% for both the V4 region and full-length sequences.

1.5 Aims and objectives

The present cross-sectional study aims to determine the relationship between gut microbiota and obesity. Specifically, the present study looks into the dataset previously obtained and investigates:

1. The difference in gut microbiota between different groups of people.
2. The relationship between gut microbiota, SCFAs, and obesity.
3. The difference in gut microbiota between the healthy and hypothalamic groups.

CHAPTER 2

Methodology

2.1 Dataset description

Data used in this study was collected and provided by Dr Muhammad Jaffar Khan. There were 151 faecal samples collected from people from the United Kingdom. Of those samples, 52 were from the healthy lean control (Healthy Lean Control) group, 29 were from the simple obese (Healthy Obese) group, 22 were from the hypothalamic lean (Hypoth. Lean) group, 19 were from the hypothalamic obese (Hypoth. Obese) group, and another 29 were from the parents of the groups. The samples were collected from October 2011 to January 2013. The sequencing method involved in the present study was 16S rRNA (ribosomal RNA) sequencing. The sequences were clustered into OTUs using the QIIME 2 pipeline (Bolyen et al., 2019) with the Silva Release 138 taxonomy database (SILVA, 2019). The BIOM file and the NEWICK format phylogenetic tree file were generated using the DADA2 pipeline (Callahan et al., 2016).

2.2 Statistical analysis

The statistical analysis involved in the present study was run in R 4.2.0 (RCoreTeam, 2022), using the BIOM file (feature_w_tax.biom) and NEWICK format phylogenetic tree file (tree.nwk), along with the related metadata (meta_data.csv). The metadata table includes the group where the sample was from and the contents of C2 (Acetate), C3 (Propionate), IC4 (Iso-Butyrate), C4 (Butyrate), IC5 (Iso-Valerate), C5 (Valerate), C6 (Caproate), C7 (Enanthate), and C8 (Caprylate) in the dried sample.

2.2.1 Alpha diversity

Alpha diversity was calculated for Healthy Lean Control, Healthy Obese, Hypoth. Lean, and Hypoth. Obese groups, respectively, with the vegan package (Oksanen et al., 2022). Since a single alpha diversity measure cannot characterise the diversity completely, a total of 5 measures were adopted:

1. Species richness (Whittaker, 1972) is a simple species count.

2. Shannon entropy (Shannon, 1948) measures the balance of a community. A higher Shannon's index means that the community is more balanced.

3. Pielou evenness (Pielou, 1966) measures the evenness of communities.

4. Fisher alpha (Fisher, 1972) compares the communities with various numbers of individuals.

5. Simpson's index (Simpson, 1949) also measures the evenness of communities but ranges from 0 to 1.

2.2.2 MPD and MNTD

The Mean Phylogenetic Distance (MPD) is the average phylogenetic distance between all pairs of OTUs in a sample. The MPD can be expressed as

$$MPD = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{i,j}p_i p_j}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_i p_j}. \quad (2)$$

Where $d_{i,j}$ is the phylogenetic distance between OTU i and OTU j, p_i is the relative abundance of the OTU i. (Stegen et al., 2012)

The Mean Nearest Taxon Distance (MNTD) is the average distance between each OTU in a sample and the OTU's closest relative in the phylogenetic tree.

The MNTD can be expressed as

$$MNTD = \sum_{i_k=1}^{n_k} f_{i_k} \min\left(\Delta_{i_k j_k}\right). \quad (3)$$

Where f_{i_k} is the relative abundance of OTU i in sample k, $\min\left(\Delta_{i_k j_k}\right)$ is the nearest distance between OTU i and other OTUs j in the sample k, n_k is the number of OTUs in sample k. (Stegen et al., 2012)

2.2.3 NRI and NTI

The present study examined environmental filtering for Healthy Lean, Healthy Obese, Hypoth. Lean, and Hypoth. Obese groups to recognise phylogenetic clustering or overdispersion. The phylogenetic distances were measured by calculating Net Relatedness Index (NRI) and Nearest Taxon Index (NTI). (Webb et al., 2000)

NRI is expressed as

$$NRI = -1 \times \frac{MPD_{observed} - MPD_{randomised}}{sdMPD_{randomised}}. \quad (4)$$

Where $MPD_{observed}$ is the observed MPD, $MPD_{randomised}$ is the expected MPD of the randomised community assemblages generated from null modelling ($n = 999$), $sdMPD_{randomised}$ is the standard deviation of the MPD of randomised community assemblages. (Stegen et al., 2012)

NTI is expressed as

$$NTI = -1 \times \frac{MNTD_{observed} - MNTD_{randomised}}{sdMNTD_{randomised}}. \quad (5)$$

Where $MNTD_{observed}$ is the observed MNTD, $MNTD_{randomised}$ is the expected MNTD of the randomised community assemblages generated from null modelling ($n = 999$), $sdMNTD_{randomised}$ is the standard deviation of the MNTD of randomised community assemblages. (Stegen et al., 2012)

The positive NRI or NTI values suggest phylogenetic clustering, while negative NRI or NTI values suggest phylogenetic overdispersion. (Kraft et al., 2007) Phylogenetic clustering suggests the community assembly results from the non-biotic assembly (i.e., environmental filtering) because closely related species tend to share the traits for surviving a specific set of environmental conditions; phylogenetic overdispersion means that the community assembly results from the biotic assembly (e.g., competition), as relatedness is believed to increase the similarity in traits and increase the possibility of competitive exclusion. (Gerhold et al., 2015)

In the present study, the picante package (Kembel et al., 2010) was used for calculating NRI and NTI with the phylogenetic distance matrix generated from the phylogenetic tree.

When used with a phylogenetic distance matrix, the ses.mpd() function returns the Standardised Effect Size (SES) of MPD in the communities (equal to NRI multiplied by -1) (DataCamp, 2022). And the ses.mnntd() function returns the SES of MNTD in the communities (equivalent to NTI multiplied by -1) (DataCamp, 2022). In the picante package, seven null models are implemented (DataCamp, 2022):

1. taxa.labels: Shuffle distance matrix labels (across all taxa included in the distance matrix).

2. richness: Randomize community data matrix abundances within samples (maintains sample species richness).

3. frequency: Randomise community data matrix abundances within species (maintains species occurrence frequency).

4. sample.pool: Randomise community data matrix by drawing species from a pool of species occurring in at least one community (sample pool) with equal probability.

5. phylogeny.pool: Randomise community data matrix by drawing species from a pool of species occurring in the distance matrix (phylogeny pool) with equal probability.

6. independentswap: Randomise community data matrix with the independent swap algorithm maintaining species occurrence frequency and sample species richness.

7. trialswap: Randomise community data matrix with the trial-swap algorithm maintaining species occurrence frequency and sample species richness.

2.2.4 Beta diversity

To examine the difference in gut microbiome composition between different groups, beta diversity was analysed using three distance metrics:

1. Bray-Curtis distance (Bray and Curtis, 1957), which tests whether the abundance counts of OTUs differ significantly between groups, can be expressed as

$$d_{ij} = \frac{\sum_{k=1}^{n}|x_{ik}-x_{jk}|}{\sum_{k=1}^{n}(x_{ik}+x_{jk})}. \quad (6)$$

2. UniFrac distance (Lozupone and Knight, 2005), which calculates between pairs of OTUs in the phylogeny to test whether the gut microbiome compositions of different groups are significantly different based on phylogeny. For two samples, if a branch leads to an OTU which exists in both of the two samples, the branch is marked as a "shared branch"; otherwise, if a branch leads to an OTU that only exists in one sample, the branch is marked as an "unshared branch". (Lozupone and Knight, 2005) The UniFrac distance can then be expressed as the fraction of total lengths of unshared branches, which is

$$\frac{sum\ of\ lengths\ of\ unshared\ branches}{sum\ of\ lengths\ of\ all\ branches}. \quad (7)$$

3. Weighted UniFrac distance (Lozupone et al., 2007) is a variant of UniFrac distance which takes into account the abundance of each OTU. The raw weighted UniFrac value can be expressed as

$$u = \sum_i^n b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|. \quad (8)$$

Where n is the total number of branches in the phylogenetic tree, $b_i$is the length of branch i; A_i is the abundance of the OTU that branch i leads to in sample A, B_i is the abundance of the OTU that branch i leads to in sample B, A_T is the total abundance of all OTUs in sample A, B_T is the total abundance of all OTUs in sample B. (Lozupone et al., 2007)

To normalise the weighted UniFrac value, the u value is then divided by a scaling factor, which can be expressed as

$$D = \sum_j^n d_j \times \left(\frac{A_j}{A_T} + \frac{B_j}{B_T} \right). \quad (9)$$

Where n is the total number of branches in the phylogenetic tree, d_j is the distance of the OUT that branch j leads to; A_j is the abundance of the OTU that branch i leads to in sample A, B_j is the abundance of the OTU that branch i leads to in sample B, A_T is the total abundance of all OTUs in sample A, B_T is the total abundance of all OTUs in sample B. (Lozupone et al., 2007)

For the normalised u value ($\frac{u}{D}$), the value of 0 indicates that the two samples are identical, while the value of 1 indicates that the two samples are non-overlapping.

In the present study, Bray-Curtis distance, UniFrac distance and weighted Unifrac distance are calculated with the phyloseq package (McMurdie and Holmes, 2013).

The Permutational Analysis of Variance (PERMANOVA) was also performed to identify the cause of the variation, using the adonis() function in the vegan package (Oksanen et al., 2022). If the PERMANOVA of a variable is reported to be significant, then the R^2 value indicates the proportion of variability that the variable can explain.

2.2.5 Observation of the top 25 most abundant taxa

The top 25 most abundant microbiome genera were analysed for Healthy Lean, Healthy Obese, Hypoth. Lean, and Hypoth. Obese groups to visualise the gut microbiome taxa abundance difference between the groups.

2.2.6 PCOA

Principal Coordinate Analysis (PCOA) was applied to the dataset to convert the beta diversity results to a two-dimensional plot. In the ordination plot, two samples are similar if they are close to each other.

2.2.7 Differential analysis: DESeq2 and MA plot

Differential analysis was applied to identify the significant differences in gut microbiome composition between different groups (Healthy Lean vs Healthy Obese, Healthy Lean vs Hypothalamic Lean, Healthy Obese vs Hypothalamic Obese). The present study used the DESeqDataSetFromMatrix() function in the DESeq2 package (Love et al., 2014) to convert the abundance table to the DESeqDataSet object. Then the DESeq() function in the DESeq2 package was used for performing differential analysis on the dataset based on the Negative Binomial (NB) Generalised Linear Model (GLM) fitting to calculate the maximum likelihood estimates of OTUs' logarithm to base 2-fold changes between two groups. (DataCamp, 2022) The DESeq() function returned a DESeqDataSet object. Then the results of the DESeq analysis were tested using the Wald significance test. The results table of the log 2-fold changes and p-values can be extracted by the results() function in the DESeq2 package, where the p-values were adjusted. (DataCamp, 2022) For the present study, the OTU is significant if its adjusted p-value is less than 0.05 and its absolute value of log 2-fold change is greater than 2.

The differential analysis generated an MA plot showing the mean abundance and the log 2-fold change of each OTU in the two groups. The significant OTUs were shown as red dots. As well as the MA-plot, a boxplot showing the normalised logarithmic relativeness of each significant OTU in the two groups was also generated. The CSV file generated shows the OTU up-regulated in the two groups. An OTU is up-

regulated in a group means that its presence in the group is identified as an increased presence compared to the other group.

2.2.8 Core gut microbiome analysis

The core microbiome analysis identifies the OTUs that are prevalent in more than a specific proportion of samples. Traditionally, the high prevalence threshold is 85%, which was defined in previous studies (Shetty et al., 2017). The microbiome package (Lahti et al., 2019) was used for performing the core microbiome analysis on every group. The results can be expressed in heat maps.

2.2.9 Beta MNTD and beta NTI

The Beta Mean Nearest Taxon Distance (βMNTD) and the Beta Nearest Taxon Index (βNTI) are the phylogenetic measures based on beta diversity.

βMNTD can be expressed as

$$\beta MNTD = 0.5\left[\sum_{i_k=1}^{n_k} f_{i_k} \min\left(\Delta_{i_k j_m}\right) + \sum_{i_k=1}^{n_k} f_{i_m} \min\left(\Delta_{i_m j_k}\right)\right]. \quad (10)$$

Where f_{i_k} is the relative abundance of OTU i in sample k, $\min\left(\Delta_{i_k j_k}\right)$ is the nearest phylogenetic distance between OTU i and other OTUs j in the sample k, n_k is the number of OTUs in sample k. (Stegen et al., 2012)

βNTI can be expressed as

$$\beta NTI = \frac{\beta MNTD_{observed} - \overline{\beta MNTD_{randomised}}}{sd(\beta MNTD_{randomised})}. \quad (11)$$

Where $\beta MNTD_{observed}$ is the observed βMNTD, $\overline{\beta MNTD_{randomised}}$ is the average βMNTD of the randomised community assemblages generated from null modelling ($n = 999$), $sd(\beta MNTD_{randomised})$ is the standard deviation of the βMNTD of randomised community assemblages. (Stegen et al., 2012)

βNTI is used to calculate the observed βMNTD's standard deviations from the average null distribution; the average null distribution is calculated by shuffling the OTUs in the phylogeny and recalculating the βMNTD 999 times. (Stegen et al., 2012)

2.2.10 Null modelling approaches for performing QPE analysis

The null modelling was used to analyse the different groups' gut microbial community assembly.

The Quantitative Process Estimates (QPE) were based on the two-step procedure by Stegen et al. (2015). Firstly, the pairwise βNTI values of all samples were calculated. If the βNTI value of a sample equals to or greater than 2, the community in the sample was assembled by variable selection; if the βNTI value of a sample equals to or less than -2, the community in the sample was assembled by homogeneous selection; if the absolute βNTI value of a sample is less than 2, the community assemble in the sample could be the result of homogenising dispersal, dispersal limitation, or undominated. (Vass et al., 2020) Secondly, the abundance-based (Raup-Crick) beta diversity was calculated using the pairwise Bray-Curtis distance measure ($\beta_{RC_{bray}}$). The communities not assembled by variable or homogeneous selection were involved in this step. If the $\beta_{RC_{bray}}$ value of a pair of samples is greater than 0.95, the two samples were assembled by dispersal limitation coupled with undominated; If the $\beta_{RC_{bray}}$ value of a pair of samples is less than -0.95, the two samples were assembled by homogenising dispersal; If the absolute $\beta_{RC_{bray}}$ value of a pair of samples is equal to or less than 0.95, the two samples were assembled by undominated. (Vass et al., 2020) Among the assembly processes, variable selection and homogeneous selection are deterministic, while dispersal limitation, homogenising dispersal and undominated are stochastic. (Stegen et al., 2015)

In the present study, the QPE process can provide the percentage of homogeneous selection, the percentage of variable selection, the percentage of dispersal limitation, the percentage of homogenising dispersal and the percentage of undominated in the gut microbial community assembly process.

The QPE and the $\beta_{RC_{bray}}$ were calculated with the picante package (Kembel et al., 2010), the ape package (Paradis and Schilap, 2019), and the ecodist package (Goslee and Urban, 2007).

2.2.11 NST

Based on the null modelling approach, the null expectation of the similarity and dissimilarity between two communities can be calculated by randomising the metacommunity, recalculating the similarity between the two communities 1000 times, and summing the average value. (Ning et al., 2019)

If communities are deterministically assembled, causing the communities to be more similar, the similarity between the two communities will be higher than the null expectation. (Ning et al., 2019) In this case, the Selection Strength (SS), which is called type-A selection strength, of the two communities can be expressed as

$$SS_{ij}^{A} = \frac{C_{ij} - \overline{E_{ij}}}{C_{ij}} \left(C_{ij} \geq \overline{E_{ij}}\right). \quad (12)$$

Where i, j are two communities, C_{ij} is the observed similarity between community i and community j, $\overline{E_{ij}}$ is the null expectation of the similarity between community i and community j. (Ning et al., 2019) Correspondingly, the type-A Stochasticity Ratio (ST) can be expressed as

If communities are deterministically assembled, causing the communities to be more dissimilar, the dissimilarity between the two communities will be higher than the null expectation. (Ning et al., 2019) In this case, the SS, which is called type-B SS, of the two communities can be expressed as

$$SS_{ij}^{B} = \frac{D_{ij} - \overline{G_{ij}}}{D_{ij}} = \frac{\overline{E_{ij}} - C_{ij}}{1 - C_{ij}} \left(C_{ij} < \overline{E_{ij}}\right). \quad (2\text{-}13)$$

Where i, j are two communities, D_{ij} $(= 1 - C_{ij})$ is the observed dissimilarity between community i and community j, G_{ij} $(= 1 - \overline{E_{ij}})$ is the null expectation of the dissimilarity between community i and community j. (Ning et al., 2019) Correspondingly, the type-B ST can be expressed as

$$ST_{ij}^{B} = 1 - SS_{ij}^{B} = \frac{\overline{G_{ij}}}{D_{ij}} = \frac{1 - \overline{E_{ij}}}{1 - C_{ij}} \left(C_{ij} < \overline{E_{ij}}\right). \quad (2\text{-}14)$$

The average pairwise SS of type-A, type-B, and total are expressed as

$$SS^{A} = \frac{\sum_{ij}^{n^A} SS_{ij}^{A}}{n^A}, \quad (2\text{-}15)$$

$$SS^B = \frac{\sum_{ij}^{n^B} SS_{ij}^B}{n^B}, \tag{2-16}$$

$$SS = \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} SS_{ij}}{n} = \frac{\sum_{ij}^{n^A} SS_{ij}^A + \sum_{ij}^{n^B} SS_{ij}^B}{n^A + n^B}. \tag{2-17}$$

Where n^A is the number of pairwise similarities higher than null expectations, n^B is the number of pairwise similarities lower than null expectations. (Ning et al., 2019)

The average pairwise stochasticity ratio of type-A, type-B, and total are expressed as

$$ST^A = \frac{\sum_{ij}^{n^A} ST_{ij}^A}{n^A}, \tag{2-18}$$

$$ST^B = \frac{\sum_{ij}^{n^B} ST_{ij}^B}{n^B}, \tag{2-19}$$

$$ST = \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} ST_{ij}}{n} = \frac{\sum_{ij}^{n^A} ST_{ij}^A + \sum_{ij}^{n^B} ST_{ij}^B}{n^A + n^B}. \tag{2-20}$$

Where n^A is the number of pairwise dissimilarities higher than null expectations, n^B is the number of pairwise dissimilarities lower than null expectations. (Ning et al., 2019)

Ideally, if the assembly process of a community is deterministic without stochasticity, the selection strength should be 100%, and the stochasticity ratio should be 0%; likewise, if the assembly process of a community is stochastic without determinism, the selection strength should be 0%, and the stochasticity should be 100%; however, since null model simulates stochastic assembly, the ST would always overestimate stochasticity. (Ning et al., 2019) Based on this, Ning et al. (2019) coined the Normalised Selection Strength (NSS) and the Normalised Stochasticity Ratio (NST).

The NSS and the NST can be expressed as

$$NSS^A = \frac{SS^A - {}^T SS^A}{{}^D SS^A - {}^T SS^A} = \frac{\sum_{ij}^{n^A} SS_{ij}^A - \min_k\left\{\sum_{ij}^{n^A} \xi(E_{ij}^{(k)}, \overline{E_{IJ}})\right\}}{\sum_{ij}^{n^A}(1-\overline{E_{IJ}}) - \min_k\left\{\sum_{ij}^{n^A} \xi(E_{ij}^{(k)}, \overline{E_{IJ}})\right\}}, \tag{2-21}$$

$$NSS^B = \frac{SS^B - {}^T SS^B}{{}^D SS^B - {}^T SS^B} = \frac{\sum_{ij}^{n^B} SS_{ij}^B - \min_k\left\{\sum_{ij}^{n^B} \xi(E_{ij}^{(k)}, \overline{E_{IJ}})\right\}}{\sum_{ij}^{n^B} \overline{E_{IJ}} - \min_k\left\{\sum_{ij}^{n^B} \xi(E_{ij}^{(k)}, \overline{E_{IJ}})\right\}}, \tag{2-22}$$

$$NSS = \frac{SS - {}^T SS}{{}^D SS - {}^T SS} = \frac{\sum_{ij} \xi(C_{ij}, \overline{E_{IJ}}) - \min_k\left\{(E_{ij}^{(k)}, \overline{E_{IJ}})\right\}}{\sum_{ij} \xi({}^D C_{ij}, \overline{E_{IJ}}) - \min_k\left\{\sum_{ij} \xi(E_{ij}^{(k)}, \overline{E_{IJ}})\right\}}, \tag{2-23}$$

$${}^D C_{ij} = \begin{cases} 0 & C_{ij} \geq \overline{E_{IJ}} \\ 1 & C_{ij} < \overline{E_{ij}} \end{cases}, \tag{2-24}$$

$$\xi(x,y)=\frac{x-y}{x-\delta}\quad \delta=\begin{cases}0 & x\geq y\\ 1 & x<y\end{cases}, \tag{2-25}$$

$$NST=1-NSS. \tag{2-26}$$

Where ${}^{D}SS$ is the extreme value of SS under completely deterministic assembly, ${}^{D}SS$ is the extreme value of SS under completely stochastic assembly, the superscript A indicates type-A ($C_{ij}\geq\overline{E_{ij}}$) pairwise comparisons, the superscript B indicates type-B ($C_{ij}<\overline{E_{ij}}$) pairwise comparisons, ${}^{D}C_{ij}$ is the similarity between community i and community j under completely deterministic assembly, $E_{ij}^{(k)}$ is one of the null expected values of similarity between between community i and community j under stochastic assembly, ξ is a generalized function for SS_{ij} under observed, stochastic or completely deterministic assembly. (Ning et al., 2019)

the vegan package (Oksanen et al., 2022), the ape package (Paradis and Schilap, 2019), and the NST package (Ning et al., 2019) were used to calculate the NST values. There are nine null models available for NST analysis (Ning et al., 2019):

Table 2. null models for QPE

Abbreviation	Way to constrain taxa occurrence frequency	Ways to constrain species richness in each sample
EE	Equiprobable	Equiprobable
EP	Equiprobable	Proportional
EF	Equiprobable	Fixed
PE	Proportional	Equiprobable
PP	Proportional	Proportional
PF	Proportional	Fixed
FE	Fixed	Equiprobable
FP	Fixed	Proportional
FF	Fixed	Fixed

Note: For taxa occurrence frequency, “Equiprobable” means that all taxa have an equal probability of occurring, “Proportional” means that the occurrence probability of a taxon is proportional to its observed occurrence frequency, and “Fixed” means that the occurrence frequency of a taxon is fixed as observed; for species richness in each sample, “Equiprobable” means that all samples have an equal probability of containing a taxon, “Proportional” means the occurrence probability in a sample is proportional to the observed richness in this sample, and “Fixed” means the occurrence frequency of a taxon is fixed as observed. (Cheaib et al., 2021)

There are several phylogenetic distance measures, namely "manhattan", "mManhattan", "euclidean","mEuclidean", "canberra", "bray", "kulczynski", "jaccard", "gower", "altGower", "mGower", "morisita", "horn", "binomial", "chao", "cao". (Ning et al., 2019)

In the NST package, “ruzicka” is not the possible value of the dist.method parameter of the tNST() function, but the Ruzicka distance measure is selected by specifying dist.method=”jaccard” and abundance.weighted=TRUE.

2.2.12 Subset regression

The subset regression approach was used in the present study to infer the relationship between alpha diversity and the SCFAs.

The subset regression takes as many subsets of explanatory variables as possible to fit the models. For N independent explanatory variables, there will be $2^N - 1$ possible regression models. The generalised form of the models can be expressed as:

$$Y = \beta_0 + \sum_i^p \beta_i X_i. \quad (2\text{-}27)$$

Where Y is the dependent variable, β_0 is the intercept, β_i is the beta-coefficient of the ith explanatory variable X_i, p is the number of the explanatory variables in the model. However, the regression model can also be expressed in the form of $Y \sim X$. $Y = \beta_0 + \beta_i X_i$, for instance, can be expressed as

$$Y \sim X_i. \quad (2\text{-}28)$$

When fitting a model, the cross-validation technique should be adopted to make the model robust. Typically, two types of cross-validation approaches are available:

1. Leave-one-out cross-validation. Fit the model on the samples, leave one sample out in each fitting and calculate the errors. Then calculate the root mean square of all the errors. This can be expressed as

$$e_{cv} = \sqrt{\sum_{i=1}^n e_i^2}. \quad (2\text{-}29)$$

Where n is the number of samples, e_{cv} is the cross-validation error, and e_i is the error when leaving out the ith sample.

2. The m-fold cross-validation. Like the leave-one-out cross-validation (leave-one-out cross-validation is called 1-fold cross-validation), however, m samples are left out in each fitting.

The lower cross-validation error means that the regression model is better. Therefore, the best model would be the one with the lowest cross-validation error.

In the present study, subset regression was completed by using the regsubsets() function in the leaps package (Miller, 2020).

CHAPTER 3

Results

3.1 Diversity

3.1.1 Alpha Diversity

Alpha diversity analysis revealed the difference in gut microbiome composition between the Healthy Lean Control, Healthy Obese, Hypoth. Lean, and Hypoth. Obese groups (Figure 1).

The Analysis of Variance (ANOVA) reported that the richness and the Fisher alpha measures are significant.

Generally, the gut microbiome in the Healthy Obese group appeared to be significantly less diverse than in the Healthy Lean Control group.

3.1.2 NRI/NTI

In general, all the groups except the Healthy Obese group have positive NRI and NTI values, indicating that strong phylogenetic clustering caused by environmental filtering exists throughout the phylogeny of the gut microbiome. (Figure 1) Therefore, in Healthy Lean Control, Hypoth. Lean and Hypoth. Obese groups, the OTUs are more related to each other than expected. However, the Healthy Obese group's NRI was reported to be negative, showing that phylogenetic dispersion caused by competitive exclusion existed in the phylogeny.

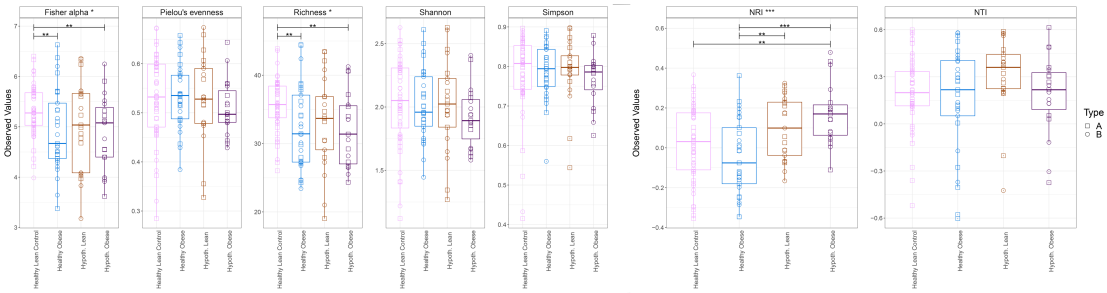

Figure1. The alpha diversity and NRI/NTI measures for each of the four groups. The "*" on the boxplot shows the significance (by ANOVA) of the results ($***: p < 0.001$, $**: p < 0.01$, $*: p < 0.05$)

3.1.3 Beta diversity

From the beta diversity analysis based on the Bray-Curtis distance measure, sequences from the Healthy Lean Control and the Hypoth. Lean groups are significantly close to each other, whereas the Healthy Obese sequences are far away from these (Figure 2). The PERMANOVA reported the groups as the predictor of the variation in abundance between the four groups ($p = 0.002, R^2 = 0.04856, df = 3$). Additionally, the beta diversity analysis based on Unweighted UniFrac distance indicates sequences from the Healthy Obese and the Hypoth. Obese groups are relatively close to each other. The PERMANOVA reported the groups as the predictor of the similar variation ($p = 0.002, R^2 = 0.04183, df = 3$).

3.1.4 Top 25 most abundant taxa for each group

The top 25 most abundant taxa are shown for each group (Figure 2). For the Healthy Lean Control and the Healthy Obese groups, the taxa analysis reported a significant difference in the abundance of *Bifidobacterium* and *Blautia*.

According to the taxa plot, the *Bifidobacterium* abundance and the *Blautia* abundance are relatively less abundant in lean groups than in obese groups.

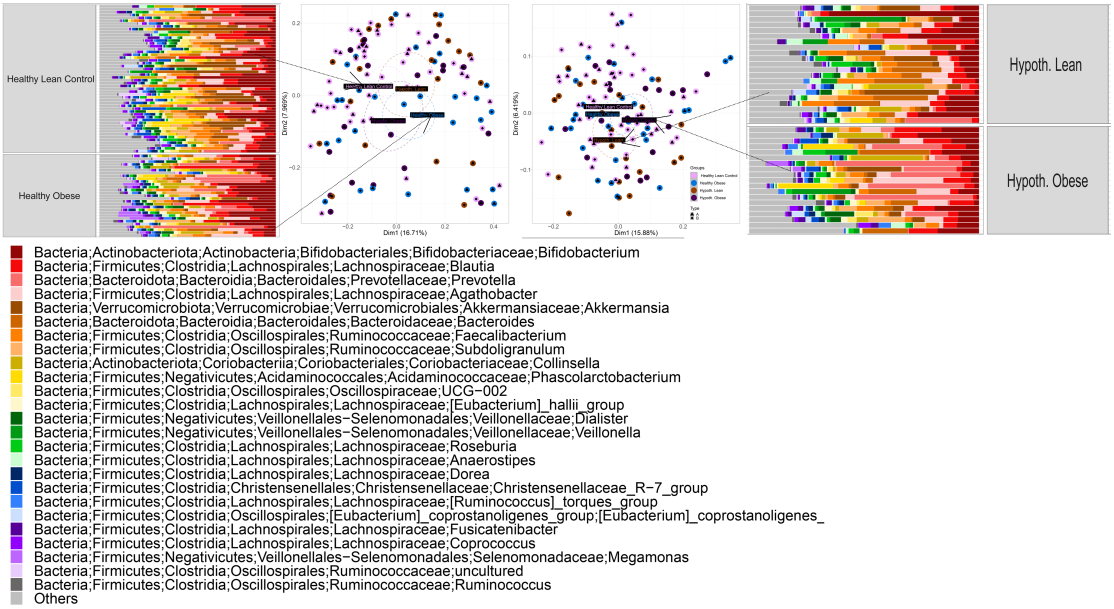

Figure 2. Beta diversity calculated with Bray-Curtis (left) and Unweighted UniFrac Distance (right) distance measures, shown with PCOA plots. The dashed ellipses are the standard errors of the four groups. Each group's top 25 most abundant taxa (at genus level) are shown around the PCOA plots, with the taxa keys on the bottom left side.

3.2 Core microbiome and differential analysis

3.2.1 Core microbiome analysis

The core microbiome of the Healthy Lean Control, Healthy Obese, Hypoth. Lean, and Hypoth. Obese groups are indicated with heat maps (Figure 3). The core microbiome is detected based on the 85% prevalence threshold.

The genera on the top of the heat map are the genera of low prevalence and low abundance, while the genera on the bottom of the heat map are the genera of high prevalence and high prevalence.

For the Health Lean Control group, the top 5 most persistent genera are *Bifidobacterium*, *Blautia*, *Agathobacter*, *Faecalibacterium*, and *Subdoligranulum*. For the Health Obese group, the top 5 most persistent genera are *Bifidobacterium*, *Blautia*, *Collinsella*, *Agathobacter*, and *Dorea*. For the Hypoth. Lean group, the top 5 most persistent genera are *Blautia*, *Bifidobacterium*, *Bacteroides*, *Faecalibacterium*, and

Agathobacter. For the Hypoth. Lean group, the top 5 most persistent genera are *Blautia*, *Bifidobacterium*, *Agathobacter*, *Faecalibacterium*, and *Subdoligranulum*.

As the detection threshold increases, the prevalence of each genus goes down, showing the difference in abundance between genera in each group.

The *Bifidobacterium* and the *Blautia* are reported as highly prevalent in all of groups, but they are significantly less abundant in the groups other than the Healthy Lean Control group.

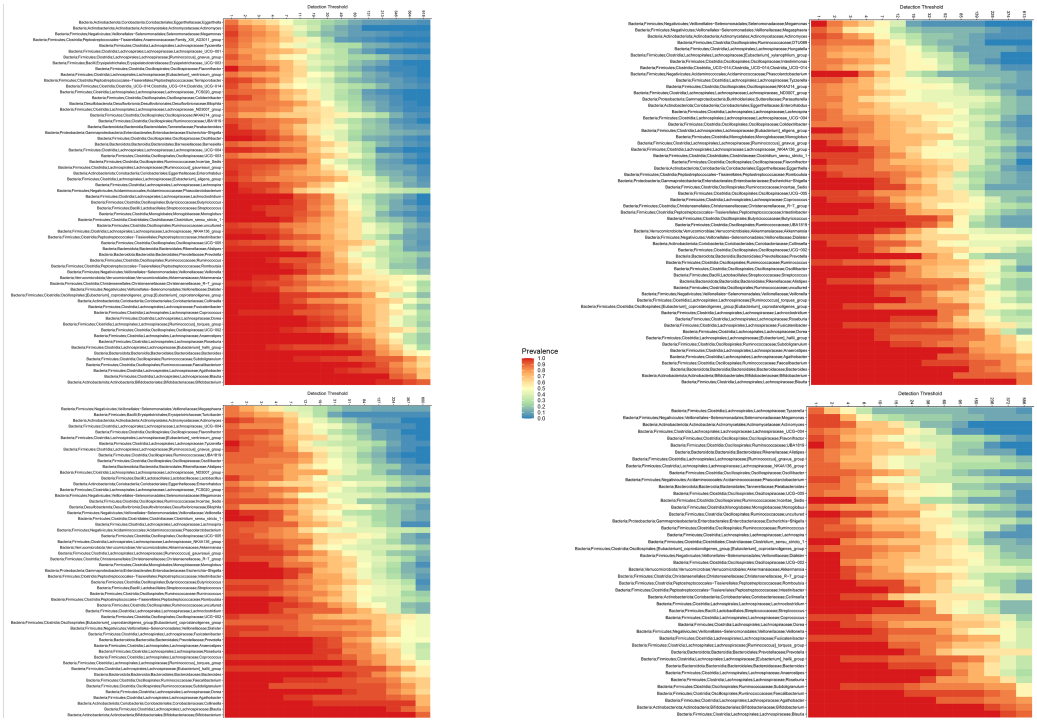

Figure 3. Core microbiome (at genus level) of the Healthy Lean Control(top-left), Healthy Obese (bottom-left), Hypoth. Lean (top-right), and Hypoth. Obese (bottom-right) groups. The detection threshold indicates the lowest level of abundance in a sample for a genus to be detected; the prevalence indicates the percentage of samples the genus detected with the threshold. (Note: please zoom in to look at the plots in detail)

3.2.2 Differential analysis

The results of differential analysis (at genus level) identified the genera significantly different between the groups. The results of comparisons are shown as MA plots (Figure 4), and a table showing the up-regulated group of the genus (Table 3).

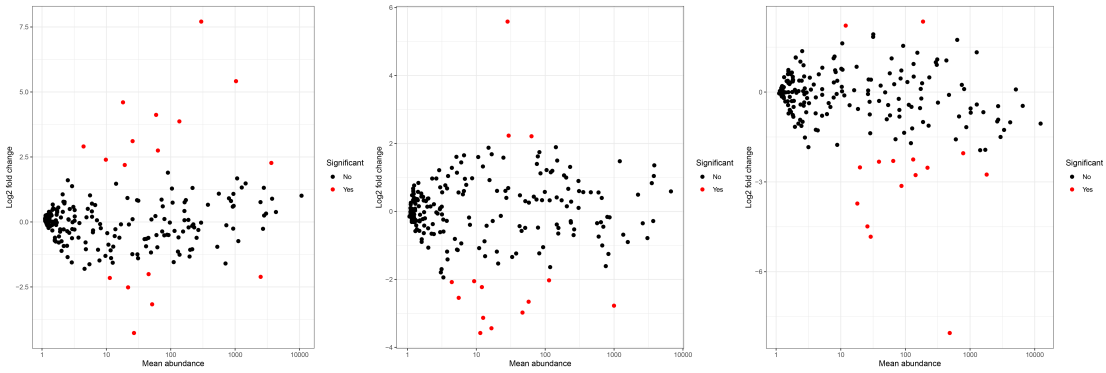

Figure 4. The MA plots showing the significant genera that are different between: the Healthy Lean Control and the Healthy Obese groups (left), the Healthy Lean Control and the Hypoth. Lean groups (middle), the Healthy Obese and Hypoth. Obese groups (right).

Table 3. The up-regulated genera in each group based on differential analysis

Healthy Lean Control vs Healthy Obese		Healthy Lean Control vs Hypoth. Lean		Healthy Obese vs Hypoth. Obese	
genus	Upregulated	genus	Upregulated	genus	Upregulated
Acidaminococcus	Healthy Obese	Catenibacterium	Healthy Lean Control	[Eubacterium]_coprostanoligenes_group	Healthy Obese
Akkermansia	Healthy Lean Control	Desulfovibrio	Healthy Lean Control	[Ruminococcus]_gnavus_group	Hypoth. Obese
Allisonella	Healthy Obese	Eisenbergiella	Hypoth. Lean	Acidaminococcus	Healthy Obese
Alloprevotella	Healthy Obese	Enorma	Healthy Lean Control	Allisonella	Healthy Obese
Catenibacterium	Healthy Obese	Holdemanella	Healthy Lean Control	Alloprevotella	Healthy Obese
Eisenbergiella	Healthy Lean Control	Hungatella	Hypoth. Lean	Catenibacterium	Healthy Obese
Enorma	Healthy Lean Control	Marvinbryantia	Healthy Lean Control	Clostridia_UCG-014	Healthy Obese
Howardella	Healthy Obese	Muribaculaceae	Healthy Lean Control	Enterorhabdus	Healthy Obese
Lactobacillus	Healthy Obese	Olsenella	Healthy Lean Control	Erysipelatoclostridium	Hypoth. Obese
Megamonas	Healthy Obese	Phascolarctobacterium	Healthy Lean Control	Lachnospiraceae_UCG-010	Healthy Obese
Megasphaera	Healthy Obese	RF39	Healthy Lean Control	Lactobacillus	Healthy Obese
Olsenella	Healthy Lean Control	Slackia	Healthy Lean Control	Paraprevotella	Healthy Obese
Paraprevotella	Healthy Obese	Succiniclasticum	Hypoth. Lean	Megamonas	Healthy Obese
Prevotella	Healthy Obese	UCG-003	Healthy Lean Control	Senegalimassilia	Healthy Obese
Sellimonas	Healthy Lean Control			Succiniclasticum	Healthy Obese
Slackia	Healthy Lean Control				
Succiniclasticum	Healthy Obese				

3.3 QPE and NST analysis

The result of QPE analysis of all the groups reported more stochastic process than deterministic process. (Figure 5) Among the processes, homogeneous selection and variable selection are deterministic, while dispersal limitation, homogenising dispersal and undominated are stochastic.

According to Ning et al. (2019), when using the Ruzicka distance measure and PP or PF null model, the accuracy and the correctness are the highest for NST and ST metrics. Therefore, the present study used the PF-Ruzicka and PP-Ruzicka measures to calculate the values of Normalised Stochasticity ratio (NST) in each group. The

results are shown in the plots below (Figure 6). The results showed that the assembly process of the gut microbiota in the Healthy Obese and the Hypoth. Obese groups is more stochastic than the assembly process of the gut microbiota in the Healthy Lean control and the Hypoth. Lean groups.

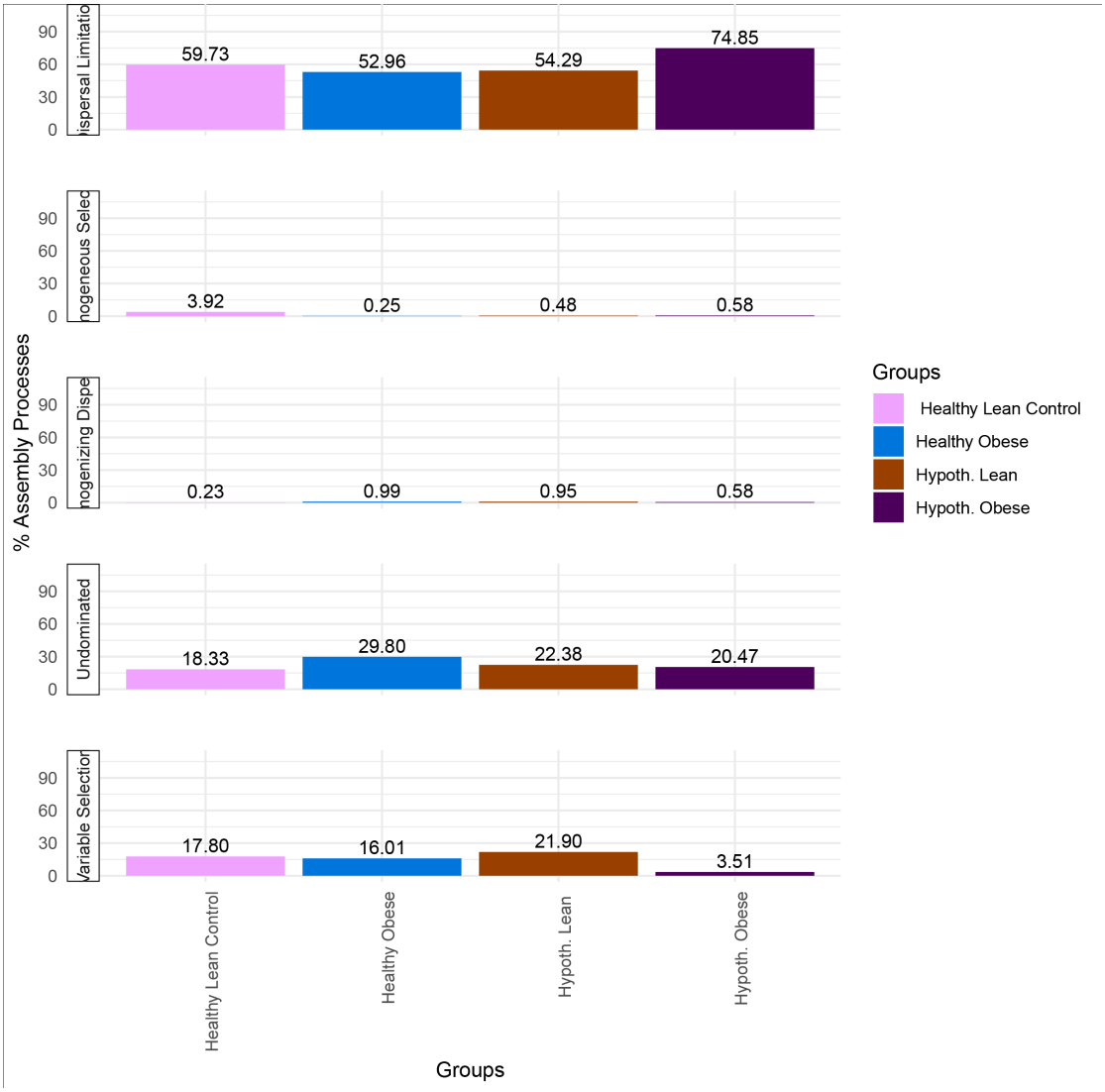

Figure 5. The proportion of community assembly process in all the groups. From top to bottom: Dispersal Limitation, Homogeneous Selection, Homogenising Dispersal, Undominated, Variable Selection.

Figure 6. The results of NST analysis using PF-Ruzicka (left) and PP-Ruzicka (right) measres.

3.4 Regression analysis

The regression models estimating the relationship between alpha diversity and SCFAs are fitted. The tables and the plots below show the optimal models for each distance measure, which are FisherAlpha ~ C2_dry + C3_dry + C5_dry + C6_dry (Table 4 and Figure 7), PielouEvenness ~ C2_dry + IC4_dry + C5_dry + C6_dry (Table 5 and Figure 8), Richness ~ C2_dry + IC4_dry + C6_dry + C8_dry (Table 6 and Figure 9), Shannon ~ C2_dry + IC4_dry (Table 7 and Figure 10), Simpson ~ C3_dry + IC4_dry (Table 8 and Figure 11). (Model parameters given with significant positive influencers highlighted in orange and negative in blue; the SCFAs are: C2=Acetate, C3=Propionate, IC4=Iso-Butyrate, C4=Butyrate, IC5=Iso-Valerate, C5=Valerate, C6=Caproate, C7=Enanthate, and C8=Caprylate) For each distance measure, only the best model will be shown here, the top 9 models and their cross-validation errors (e_{CV}) can be found in the Appendix I.

Table 4. FisherAlpha ~ C2_dry + C3_dry + C5_dry + C6_dry ($e_{CV} = 0.64961$)

FisherAlpha									
Predictors	*Estimates*	*std. Error*	*std. Beta*	*standardized std. Error*	*CI*	*standardized CI*	*Statistic*	*p*	*df*
(Intercept)	5.14969 ***	0.1888	0	0.0822	4.77558 – 5.52380	-0.16289 – 0.16289	27.27643	**5.00E-51**	111
C2 dry	-0.00097	0.00052	-0.1934	0.10361	-0.00199 – 0.00006	-0.39872 – 0.01192	-1.86651	6.46E-02	111
C3 dry	-0.00379 *	0.0019	-0.23215	0.11606	-0.00755 – -0.00004	-0.46214 – -0.00216	-2.00015	**4.79E-02**	111
C5 dry	0.03169 **	0.01103	0.30088	0.10477	0.00982 – 0.05355	0.09327 – 0.50849	2.87186	**4.89E-03**	111
C6 dry	0.03657 *	0.01445	0.22617	0.08937	0.00794 – 0.06520	0.04909 – 0.40326	2.53086	**1.28E-02**	111
Observations	116								
R^2 / R^2 adjusted	0.243 / 0.216								
									* p<0.05 ** p<0.01 *** p<0.001

Figure 7. The estimated distribution of the beta coefficients of FisherAlpha ~ C2_dry + C3_dry + C5_dry + C6_dry

Table 5. PielouEvenness ~ C2_dry + IC4_dry + C5_dry + C6_dry ($e_{CV} = 0.06999$)

	PielouEvenness								
Predictors	*Estimates*	*std. Error*	*std. Beta*	*standardized std. Error*	*CI*	*standardized CI*	*Statistic*	*p*	*df*
(Intercept)	0.49183 ***	0.02775	0	0.08411	0.43684 – 0.54682	-0.16666 – 0.16666	17.72278	**3.78E-34**	111
C2 dry	-0.00010 *	0.00005	-0.17964	0.08879	-0.00019 – -0.00000	-0.35559 – -0.00369	-2.02308	**4.55E-02**	111
IC4 dry	0.00706 ***	0.00178	0.43219	0.10908	0.00353 – 0.01059	0.21604 – 0.64835	3.9621	**1.32E-04**	111
C5 dry	-0.00277 *	0.00126	-0.24491	0.11145	-0.00527 – -0.00027	-0.46575 – -0.02407	-2.19751	**3.01E-02**	111
C6 dry	0.00269	0.00158	0.15509	0.09103	-0.00044 – 0.00583	-0.02529 – 0.33548	1.70373	9.12E-02	111
Observations	116								
R^2 / R^2 adjusted	0.208 / 0.179								

** p<0.05 ** p<0.01 *** p<0.001*

Figure 8. The estimated distribution of the beta coefficients of PielouEvenness ~ C2_dry + IC4_dry + C5_dry + C6_dry

Table 6. Richness ~ C2_dry + IC4_dry + C6_dry + C8_dry ($e_{CV} = 4.75411$)

Richness									
Predictors	*Estimates*	*std. Error*	*std. Beta*	*standardized std. Error*	*CI*	*standardized CI*	*Statistic*	*p*	*df*
(Intercept)	31.39122 ***	1.91731	0	0.08107	27.59194 – 35.19051	-0.16065 – 0.16065	16.3725	**2.23E-31**	111
C2 dry	-0.00966 **	0.0032	-0.25533	0.08461	-0.01600 – -0.00332	-0.42299 – -0.08767	-3.01767	**3.16E-03**	111
IC4 dry	0.28297 **	0.09894	0.24566	0.08589	0.08692 – 0.47902	0.07546 – 0.41586	2.86016	**5.06E-03**	111
C6 dry	0.29342 **	0.11065	0.2395	0.09031	0.07417 – 0.51267	0.06054 – 0.41847	2.6519	**9.18E-03**	111
C8 dry	0.2076	0.16521	0.11237	0.08943	-0.11976 – 0.53497	-0.06483 – 0.28958	1.25663	2.12E-01	111
Observations	116								
R^2 / R^2 adjusted	0.264 / 0.238								

** p<0.05 ** p<0.01 *** p<0.001*

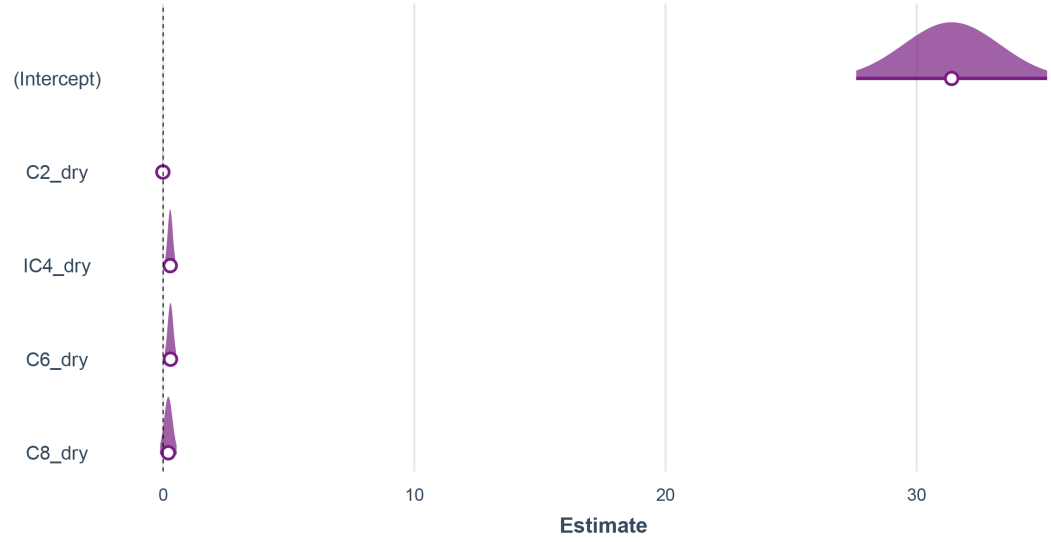

Figure 9. The estimated distribution of the beta coefficients of Richness ~ C2_dry + IC4_dry + C6_dry + C8_dry

Table 7. Shannon ~ C2_dry + IC4_dry ($e_{CV} = 0.28664$)

Shannon									
Predictors	*Estimates*	*std. Error*	*std. Beta*	*standardized std. Error*	*CI*	*standardized CI*	*Statistic*	*p*	*df*
(Intercept)	1.88044 ***	0.11379	0	0.08408	1.65500 – 2.10587	-0.16657 – 0.16657	16.52602	**6.28E-32**	113
C2 dry	-0.00049 *	0.00019	-0.22074	0.087	-0.00087 – -0.00011	-0.39309 – -0.04838	-2.53732	**1.25E-02**	113
IC4 dry	0.02230 ***	0.00584	0.33209	0.087	0.01072 – 0.03387	0.15974 – 0.50444	3.81732	**2.21E-04**	113
Observations	116								
R^2 / R^2 adjusted	0.194 / 0.180								
									* p<0.05 ** p<0.01 *** p<0.001

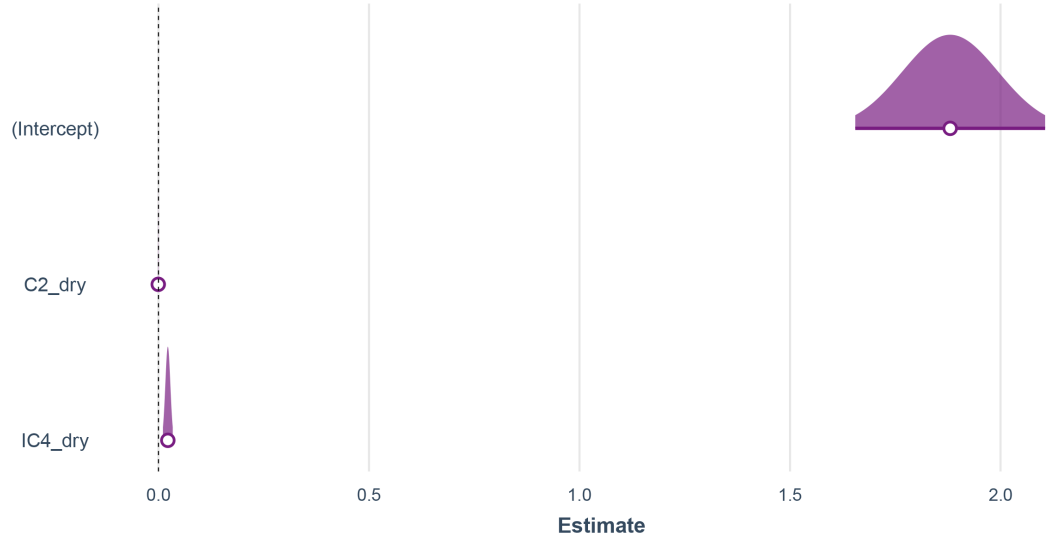

Figure 10. The estimated distribution of the beta coefficients of Shannon ~ C2_dry + IC4_dry

Table 8. Simpson ~ C3_dry + IC4_dry ($e_{CV} = 0.0806$)

	Simpson								
Predictors	*Estimates*	*std. Error*	*std. Beta*	*standardized std. Error*	*CI*	*standardized CI*	*Statistic*	*p*	*df*
(Intercept)	0.73949 ***	0.02687	0	0.08809	0.68626 – 0.79272	-0.17452 – 0.17452	27.52379	**6.37E-52**	113
C3 dry	-0.00036 *	0.00018	-0.17886	0.08865	-0.00070 – -0.00001	-0.35450 – -0.00323	-2.01757	**4.60E-02**	113
IC4 dry	0.00514 **	0.00164	0.27784	0.08865	0.00189 – 0.00838	0.10220 – 0.45347	3.13404	**2.20E-03**	113
Observations	116								
R^2 / R^2 adjusted	0.116 / 0.100								
									* p<0.05 ** p<0.01 *** p<0.001

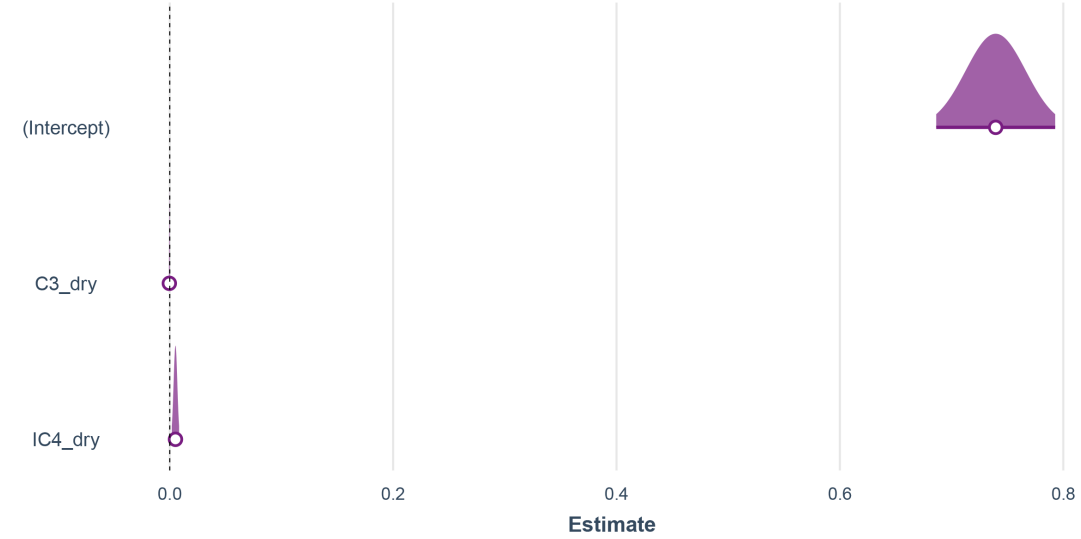

Figure 11. The estimated distribution of the beta coefficients of Simpson ~ C3_dry + IC4_dry

Chapter 4

Discussion

The genera *Bifidobacterium* and *Blautia* were found to be highly prevalent in all of the groups. However, the significant difference in the abundance of the two genera between the Healthy Lean Control and the Healthy Obese groups may explain the cause of simple obesity.

The *Bifidobacterium* is responsible for the metabolisation of oligosaccharides, including plant source fructo oligosaccharides and dairy source galacto oligosaccharides. *Bifidobacterium* also produces lactic acid and acetic acid. The *Blautia* produces butyric acid and acetic acid. (Ozato et al., 2019) Acetic acid, lactic acid, and butyric acid regulate the GPR41 and the GPR43, and reduce obesity. (Ozato et al., 2019) The *Blautia* is also responsible for reducing visceral fat. (Ozato et al., 2019)

The significant differences in gut microbiota could also account for the cause of obesity.

The effect of the genus *Lactobacillus* on body weight is multifaceted. On the one hand, the species *L. rhamnosus* and *L. acidophilus* are proven to be associated with weight gain, while the species *L. plantarum* and *L. curvatus* are proven to have beneficial effects on weight.

A significant limitation of the present study is that the Operational Taxonomic Units (OTUs), instead of the Amplicon Sequence Variants (ASVs), are used to represent the species of the gut microbiome. An OTU is constructed by clustering the sequences based on a fixed similarity threshold (usually 97%). Since the individuals in a species do not evolve or mutate at the same rate, the cluster of the similar sequences may not be able to accurately represent a species. However, an ASV is constructed based on a specific distribution model, thus the assignment of similarity is not needed.

Chapter 5

Conclusion

Here, with the modern bioinformatic and statistical techniques, I found out the relationship between gut microbiota, SCFAs and obesity. The gut microbiota is capable of producing short-chain fatty acids that can influence host metabolism and appetite, which in turn affects host nutrient and energy utilization and ultimately host body weight; in turn, the content of short-chain fatty acids affects the composition of the gut microbiota

Unfortunately, due to the absence of the original sequence data and a lack of time, in the present study, the in-depth investigation the specific species of gut microbiome was not able to be completed appropriately.

Therefore, the direction of the future studies should be conducted in the following directions.

1. Use ASV (Amplicon Sequence Variant) instead of OTUs to represent the species of the gut microbiome.
2. Perform a thorough investigation to figure out the functions and the properties of the core gut microbiome genera, especially those genera that are significantly different in abundance between different groups.
3. Investigate the relationship between hypothalamic disorders, metabolism, and gut microbiota.

References

Abenavoli, L., Scarpellini, E., Colica, C., Boccuto, L., Salehi, B., Sharifi-Rad, J., Aiello, V., Romano, B., de Lorenzo, A., Izzo, A. A., & Capasso, R. (2019). Gut Microbiota and Obesity: A Role for Probiotics. *Nutrients*, *11*(11), 2690.

Gérard, P. (2016). Gut microbiota and obesity. *Cellular and Molecular Life Sciences*, *73*(1), 147–162. https://doi.org/10.1007/s00018-015-2061-5

Kelly, T., Yang, W., Chen, C. S., Reynolds, K., & He, J. (2008). Global burden of obesity in 2005 and projections to 2030. *International Journal of Obesity*, *32*(9), 1431–1437. https://doi.org/10.1038/ijo.2008.102

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

McKenna, A., Ijaz, U. Z., Kelly, C., Linton, M., Sloan, W. T., Green, B. D., Lavery, U., Dorrell, N., Wren, B. W., Richmond, A., Corcionivoschi, N., & Gundogdu, O. (2020). Impact of industrial production system parameters on chicken microbiomes: mechanisms to improve performance and reduce Campylobacter. *Microbiome*, *8*(1), 128. https://doi.org/10.1186/s40168-020-00908-8

Nikolova, C., Ijaz, U. Z., & Gutierrez, T. (2021). Exploration of marine bacterioplankton community assembly mechanisms during chemical dispersant and surfactant-assisted oil biodegradation. *Ecology and Evolution*, *11*(20), 13862–13874. https://doi.org/10.1002/ece3.8091

Ning, D., Deng, Y., Tiedje, J. M., & Zhou, J. (2019). A general framework for quantitatively assessing ecological stochasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(34), 16892–16898. https://doi.org/10.1073/pnas.1904623116

Stanislawski, M. A., Dabelea, D., Lange, L. A., Wagner, B. D., & Lozupone, C. A. (2019). Gut microbiota phenotypes of obesity. *Npj Biofilms and Microbiomes*, *5*(1), 18. https://doi.org/10.1038/s41522-019-0091-8

Stegen, J. C., Lin, X., Fredrickson, J. K., & Konopka, A. E. (2015). Estimating and mapping ecological processes influencing microbial community assembly. *Frontiers in Microbiology*, *6*(MAY). https://doi.org/10.3389/fmicb.2015.00370

Stegen, J. C., Lin, X., Konopka, A. E., & Fredrickson, J. K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME Journal*, *6*(9), 1653–1664. https://doi.org/10.1038/ismej.2012.22

Vass, M., Székely, A. J., Lindström, E. S., & Langenheder, S. (2020). Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-59182-1

Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *In Annual Review of Ecology and Systematics*, *33*, 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448

World Health Organization. (2021/06/09). *Obesity and overweight.* https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Lustig, R. H. (2011). Hypothalamic obesity after craniopharyngioma: mechanisms, diagnosis, and treatment. *Frontiers in endocrinology, 2,* 60.

Must, A. & Strauss, R. S. (1999). Risks and consequences of childhood and adolescent obesity. *International journal of obesity, 23*(2), S2-S11.

R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SILVA (2019). *Release 138.* URL https://www.arb-silva.de/documentation/release-138/

Westcott and Schloss (2015), De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ 3*, e1487; DOI 10.7717/peerj.1487

Panek, M., Čipčić Paljetak, H., Barešić, A., Perić, M., Matijašić, M., Lojkić, I., ... & Verbanac, D. (2018). Methodology challenges in studying human gut microbiota–

effects of collection, storage, DNA extraction and next generation sequencing technologies. *Scientific reports*, *8*(1), 1-13.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics-a guide from sampling to data analysis. *Microbial informatics and experimentation*, *2*(1), 1-12.

Verma, V., Gupta, A., & Chaudhary, V. K. (2020). Emulsion PCR made easy. *BioTechniques*, *69*(1), 64-69.

Sneath, P. H., & Sokal, R. R. (1973). Numerical taxonomy. *The principles and practice of numerical classification*.

Edgar, R. *Operational Taxonomic Units (OTUs).* drive5. URL https://www.drive5.com/usearch/manual7/otu_definition.html

Edgar, R. *Defining and interpreting OTUs*. URL https://drive5.com/usearch/manual/otus.html

SILVA (2016). *Release 128.* URL https://www.arb-silva.de/documentation/release-128/

Second Genome (2018). *Greengenes database 13.5.* URL https://greengenes.secondgenome.com/?prefix=downloads/greengenes_database/gg_13_5/

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., ... & Truper, H. G. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic and Evolutionary Microbiology*, *37*(4), 463-464.

Edgar, R. (2018). Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ*, *6*, e5030.

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371-2375.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*, 581-583. https://doi.org/10.1038/nmeth.3869.

Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P,

Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2022). *vegan: Community Ecology Package*. R package version 2.6-2, URL https://CRAN.R-project.org/package=vegan

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, *21*(2-3), 213-251.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *Journal of theoretical biology*, *13*, 131-144.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 42-58.

Simpson, E. H. (1949). *Measurement of diversity. Nature*, *163*(4148), 688-688.

S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics 26*:1463-1464.

DataCamp (2022). *ses.mpd: Standardized effect size of MPD.* RDocumetation. URL https://www.rdocumentation.org/packages/picante/versions/1.8.2/topics/ses.mpd

DataCamp (2022). *ses.mntd: Standardized effect size of MNTD.* RDocumetation. URL https://www.rdocumentation.org/packages/picante/versions/1.8.2/topics/ses.mntd

Webb, C. O. (2000). Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, *156*(2), 145-155.

Kraft, N. J., Cornwell, W. K., Webb, C. O., & Ackerly, D. D. (2007). Trait evolution, community assembly, and the phylogenetic structure of ecological communities. *The American Naturalist*, *170*(2), 271-283.

Gerhold, P., Cahill Jr, J. F., Winter, M., Bartish, I. V., & Prinzing, A. (2015). Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). *Functional Ecology*, *29*(5), 600-614.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs*, *27*(4), 326-349.

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, *71*(12), 8228-8235.

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, *73*(5), 1576-1585.

Paul J. McMurdie and Susan Holmes (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4): e61217.

Love, M.I., Huber, W., Anders, S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12):550

DataCamp (2022). *DESeq: Differential expression analysis based on the Negative Binomial (a.k.a. Gamma-Poisson) distribution*. RDocumetation. URL https://www.rdocumentation.org/packages/DESeq2/versions/1.12.3/topics/DESeq

DataCamp (2022). *results: Extract results from a DESeq analysis*. RDocumetation. URL https://www.rdocumentation.org/packages/DESeq2/versions/1.12.3/topics/results

Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H., & de Vos, W. M. (2017). Intestinal *microbiome* landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS microbiology reviews*, *41*(2), 182-199.

Leo Lahti et al. (2019) *microbiome R package*. URL: http://microbiome.github.io

Raup, D. M., & Crick, R. E. (1979). Measurement of faunal similarity in paleontology. *Journal of Paleontology*, 1213-1227.

Paradis E. & Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526-528.

Goslee, S.C. and Urban, D.L. 2007. The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, *22*(7), 1-19. DOI:10.18637/jss.v022.i07

Cheaib, B., Seghouani, H., Llewellyn, M., Vandal-Lenghan, K., Mercier, P. L., & Derome, N. (2021). The yellow perch (Perca flavescens) microbiome revealed resistance to colonisation mostly associated with neutralism driven by rare taxa under cadmium disturbance. *Animal microbiome*, *3*(1), 1-19.

Alan Miller (2020). *leaps: Regression Subset Selection.* R package version 3.1. URL <https://CRAN.R-project.org/package=leaps>.

Ozato, N., Saito, S., Yamaguchi, T., Katashima, M., Tokuda, I., Sawada, K., ... & Nakaji, S. (2019). Blautia genus associated with visceral fat accumulation in adults 20–76 years of age. *NPJ biofilms and microbiomes*, *5*(1), 1-9.

Appendix

Appendix I the cross-validation errors of the fitted models

	Fisher Alpha and SCFAs	
	Model	Cross-validation Errors
4	FisherAlpha ~ C2_dry + C3_dry + C5_dry + C6_dry	0.64961
5	FisherAlpha ~ C2_dry + C3_dry + IC5_dry + C5_dry + C6_dry	0.65295
2	FisherAlpha ~ C2_dry + C6_dry	0.65418
3	FisherAlpha ~ C3_dry + C5_dry + C6_dry	0.65725
6	FisherAlpha ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry	0.66919
7	FisherAlpha ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry + C8_dry	0.6755
8	FisherAlpha ~ C2_dry + C3_dry + IC4_dry + C4_dry + IC5_dry + C5_dry + C6_dry + C8_dry	0.68828
1	FisherAlpha ~ IC5_dry	0.6903
9	FisherAlpha ~ C2_dry + C3_dry + IC4_dry + C4_dry + IC5_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.69527
	Pielou Evenness and SCFAs	
	Model	Cross-validation Errors
4	PielouEvenness ~ C2_dry + IC4_dry + C5_dry + C6_dry	0.06999
5	PielouEvenness ~ C2_dry + C3_dry + IC4_dry + C5_dry + C6_dry	0.07032
3	PielouEvenness ~ C2_dry + IC4_dry + C5_dry	0.07052
2	PielouEvenness ~ C2_dry + IC4_dry	0.07096
6	PielouEvenness ~ C2_dry + C3_dry + IC4_dry + C5_dry + C6_dry + C7_dry	0.071
7	PielouEvenness ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry + C7_dry	0.0713
8	PielouEvenness ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.0716
1	PielouEvenness ~ IC4_dry	0.07225
9	PielouEvenness ~ C2_dry + C3_dry + IC4_dry + C4_dry + IC5_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.07229
	Richness and SCFAs	
	Model	Cross-validation Errors
4	Richness ~ C2_dry + IC4_dry + C6_dry + C8_dry	4.75411
3	Richness ~ C2_dry + IC4_dry + C6_dry	4.79648
5	Richness ~ C2_dry + C3_dry + IC4_dry + C6_dry + C8_dry	4.79727
2	Richness ~ C2_dry + C6_dry	4.95577
6	Richness ~ C2_dry + C3_dry + IC4_dry + C5_dry + C6_dry + C8_dry	4.96595
7	Richness ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry + C8_dry	5.03256
1	Richness ~ IC4_dry	5.11507
8	Richness ~ C2_dry + C3_dry + IC4_dry + C4_dry + IC5_dry + C5_dry + C6_dry + C8_dry	5.13719
9	Richness ~ C2_dry + C3_dry + IC4_dry + C4_dry + IC5_dry + C5_dry + C6_dry + C7_dry + C8_dry	5.23092
	Shannon Entropy and SCFAs	
	Model	Cross-validation Errors
2	Shannon ~ C2_dry + IC4_dry	0.28664
3	Shannon ~ C2_dry + IC4_dry + C6_dry	0.28696
4	Shannon ~ C2_dry + IC4_dry + C5_dry + C6_dry	0.29072
1	Shannon ~ IC4_dry	0.29402
5	Shannon ~ C2_dry + C3_dry + IC4_dry + C5_dry + C6_dry	0.29843
6	Shannon ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry	0.30143
7	Shannon ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry + C8_dry	0.30259
8	Shannon ~ C2_dry + C3_dry + IC4_dry + IC5_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.30443
9	Shannon ~ C2_dry + C3_dry + IC4_dry + C4_dry + IC5_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.30979
	Simpson Index and SCFAs	
	Model	Cross-validation Errors
2	Simpson ~ C3_dry + IC4_dry	0.0806
3	Simpson ~ C3_dry + IC4_dry + C6_dry	0.08064
4	Simpson ~ C2_dry + C3_dry + IC4_dry + C6_dry	0.08104
5	Simpson ~ C2_dry + C3_dry + IC4_dry + C6_dry + C8_dry	0.08132
6	Simpson ~ C2_dry + C3_dry + IC4_dry + C5_dry + C6_dry + C8_dry	0.08178
1	Simpson ~ IC4_dry	0.08218
7	Simpson ~ C2_dry + C3_dry + IC4_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.08219
8	Simpson ~ C2_dry + C3_dry + IC4_dry + C4_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.08569
9	Simpson ~ C2_dry + C3_dry + IC4_dry + C4_dry + IC5_dry + C5_dry + C6_dry + C7_dry + C8_dry	0.08701