UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

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

***Master's thesis***

***Comparative Analysis of the Female Microbiome Diversity and External Influences: Evidence from FLORA and Isala Projects***

**Mary Grace Lapurga**

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

**SUPERVISOR :**
Prof. dr. Olivier THAS
Mevrouw Kato MICHIELS

**SUPERVISOR :**
Prof. Dr. Thomas DEMUYSER
Prof. Dr. Shari MACKENS

**2024**
**2025**

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***

***Comparative Analysis of the Female Microbiome Diversity and External Influences: Evidence from FLORA and Isala Projects***

**Mary Grace Lapurga**

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

**SUPERVISOR :**
Prof. dr. Olivier THAS
Mevrouw Kato MICHIELS

**SUPERVISOR :**
Prof. Dr. Thomas DEMUYSER
Prof. Dr. Shari MACKENS

# Acknowledgements

The author extends her sincere gratitude to all who contributed to the formulation of this thesis and the completion of her master's degree.

Many thanks to her thesis supervisors, Prof. dr. Olivier Thas (UHasselt), Prof. dr. Shari Mackens (UZBrussel), Prof. dr. Thomas Demuyser (UAntwerp), and Ms. Kato Michiels (UHasselt) for entrusting her with the FLORA and Isala Projects, and for their invaluable guidance, support and feedback throughout the preparation of this report. The author also thanks Charlotte Van Roy (UZBrussel) and Tim Van Rillaer (UAntwerp) for their assistance in handling the survey and microbiome datasets.

Heartfelt thanks go to all the women who participated in the Isala and FLORA projects. Their willingness to voluntarily share personal and sensitive data has been indispensable to this research.

Deepest gratitude also goes to the Hasselt University and VLIR-UOS for the scholarship grant that made her graduate studies in Belgium possible.

Utmost appreciation to the Filipino community in Belgium, friends, mentors, and family, especially her parents, Mr. Marcelo Lapurga and Mrs. Regina Lapurga, for their unwavering support, encouragement, and belief in her.

Above all, praise and glory be to God for His grace, which has provided the author comfort and motivation to keep moving forward despite all the challenges.

# Abstract

Understanding the composition of the female microbiome and its link to external factors is important to promote reproductive and maternal health. While several works have linked the vaginal microbiome with infertility and external factors, direct comparative studies between subfertile and healthy women remain scarce. This limits understanding of the microbiome-driven mechanisms of fertility issues. In this study, a comparative analysis was done on the alpha diversity and taxon abundance in the vaginal microbiome of subfertile women who are currently undergoing fertility treatment ("subfertile group") with those who do not have reproductive and fertility issues ("healthy group") and with those who had once initiated a fertility program ("benchmark group"). It also assessed the link between external factors to the microbiome composition. Linear regression models of the Shannon index and Chao1 index were fitted to assess differences in alpha diversity between groups, adjusted for age, body mass index (BMI), smoking status, birth method, and hours of sleep. Analyses of Compositions of Microbiomes with Bias Correction (ANCOM-BC) were also performed to identify differentially abundant taxa and factors associated with the abundance of specific taxa. The results revealed that the three groups have a low Shannon index, which can be attributed to the dominance of the *Lactobacillus* species. The subfertile group has reduced microbial richness (Chao1 index) compared to the healthy and benchmark group. For women in the same group, the effect of age, BMI, and hours of sleep on the Shannon index and Chao1 index were small, which may not be indicative of a biologically relevant shift in alpha diversity. The study also found that smoking and the birth method have no significant effect on alpha diversity. In terms of microbiome composition, subfertile women exhibited lower abundance of *Lactobacillus jensenii* and higher abundance of *Lactobacillus gasseri* and *Lactobacillus crispatus* than the healthy group. In comparison with the benchmark group, subfertile women showed higher abundance of *Lactobacillus crispatus*, *Lactobacillus gasseri*, and *Lactobacillus iners*. These findings suggest that different *Lactobacillus* species may be differentially associated with reproductive status, highlighting the complexity of microbial influences on fertility. The results also suggest a significant association between the abundance of at least one *Lactobacillus* species and external factors, namely age, BMI, hours of sleep, and alcohol intake for subfertile women. These factors may be considered when designing targeted interventions or personalized fertility treatment. This study acknowledges certain limitations, including the absence of potentially important variables that could influence microbiome diversity and abundance. Moreover, the goal of the study was not to identify specific biomarkers of infertility, but to explore the associations between fertility status, external factors, and vaginal microbiome composition.



Keywords: *infertility, vaginal microbiome, alpha diversity, differential abundance*

# Contents

# List of Tables



# List of Figures

# 1 Introduction

## 1.1 Background

The human body is composed of microbiota that play a crucial role in maintaining normal body functions, the immune system, and overall health conditions. Microbiota refers to the collection of microorganisms, such as bacteria, fungi, archaea, viruses, and parasites, that live within the human body. Microbiota is used interchangeably with microbiome, which encompasses the entire ecosystem in which these organisms exist, including their genetic material and environmental conditions (Del Campo-Moreno et al., 2018). These microorganisms can be found in the oral cavity, skin, gastrointestinal tract, and reproductive tract, among others. Their function includes synthesizing essential vitamins, breaking down food to extract nutrients, improving the immune system function, and producing beneficial anti-inflammatory compounds that prevent diseases (Moreno et al., 2018).

In a woman's body, specifically, the genital tract is composed of microbiota that are essential for reproductive and maternal health (F. Liu et al., 2021, Chen et al., 2017). The reproductive tract of women is dominated by *Lactobacillus* species, which include *L. crispatus*, *L. jensenii*, *L. gasseri*, and *L. iners*. These species are considered pivotal in maintaining a healthy vaginal environment. *Lactobacillus* species secrete metabolic by-products in cervicovaginal fluid that help protect against pathogens and infections (Petrova et al., 2015). They help break down glycogen to lactic acid that reduces vaginal pH. This acidic environment makes it difficult for opportunistic pathogens to thrive, thereby preventing various vaginal infections (Pagar et al., 2024, Zaino et al., 1994). Reports show that a low amount of *L. crispatus* and colonization of *Streptococcus agalactiae* is linked to premature birth (Fettweis et al., 2019, Son et al., 2018, Haque et al., 2017). Reduced abundance of *Lactobacillus* species and presence of *Gardnerella vaginalis* and *Atopobium vaginae* characterizes bacterial vaginosis, a common vaginal infection in women of reproductive age caused by an imbalance of bacteria in the vagina (Shipitsyna et al., 2013). Bacterial vaginosis can increase the risk of pelvic inflammatory disease, miscarriage, and premature birth (Bradshaw and Sobel, 2016).

Recent studies have also linked the vaginal microbiome with infertility and external factors. Infertility, also termed subfertility, is characterized by the inability to conceive after at least 12 months of regular, unprotected sexual activity (World Health Organization, 2023). Among the possible causes of infertility are disorders of the reproductive health system (e.g., blocked fallopian tubes, endometriosis, polycystic ovarian syndrome or PCOS), hormonal imbalance, or lifestyle factors (e.g., smoking, obesity, alcohol intake). Various research has also described infertile women in terms of their microbiome composition. Women with fertility issues were characterized with reduced microbiome diversity, lower levels of *Lactobacillus*, and higher abundance of *Atopobium*, *Aerococcus*, and *Bifidobacterium* (Zhao

et al., 2020). Another study of women who underwent in vitro fertilization (IVF) treatment showed that a low level of *Lactobacillus*, a high level of *Proteobacteria*, and presence *Gardnerella vaginalis* are associated with a low pregnancy rate, while a high abundance of *Lactobacillus crispatus* is linked with a higher chance of getting pregnant (Koedooder et al., 2019). Differences in microbiome composition have been associated with age, estrogen level, personal hygiene practices, and antimicrobial medications (Lehtoranta et al., 2022). Body mass index, hours of sleep, and smoking were also found to have a significant association with the diversity of the vaginal microbiome (Lebeer et al., 2022). Furthermore, drinking alcohol, smoking, and being exposed to psychosocial stress were linked with an increased risk of having bacterial vaginosis (Morsli et al., 2024).

Despite existing research, direct comparative studies between subfertile and healthy women remain scarce. Most of them focused on the characterization of infertile women, while some have healthy controls with a few samples. This limits understanding of the microbiome-driven mechanisms of fertility issues. In this thesis, the primary interest is to compare the microbiome profiles of subfertile and healthy women, and to identify the external factors associated with the vaginal microbiome profiles of these two groups. Data from subfertile women came from the FLORA Project, a prospective clinical trial at the Brussels University Hospital (UZ Brussel) in Belgium, in collaboration with the University of Antwerp (UAntwerp). The FLORA project aims to investigate the role of the microbiota in fertility and its impact on IVF outcomes. It includes subfertile women currently undergoing IVF treatment at UZ Brussel. Meanwhile, data from a healthy cohort came from the Isala Project, a citizen science project by UAntwerp. The Isala Project also involved women who had once initiated the fertility program due to their or their partner's reduced fertility (with a known or unknown cause), among other reasons. This cohort was also used to benchmark the microbiome profile of subfertile women in the FLORA Project, which is the secondary objective of this thesis.

## 1.2 Research questions

Using the FLORA and Isala Project dataset, this study aims to answer the following research questions:

1. Are the microbiome profiles (i.e., alpha diversity and abundance of specific genera) of subfertile women different from those of healthy women?
2. Are the microbiome profiles of subfertile women different from those who had once initiated a fertility program?
3. What are the external factors associated with women's microbiome profiles?

## 1.3 Relevance, stakeholders, and ethics

Understanding the composition of the female microbiome and its link to external factors is essential in advancing reproductive and maternal health. By directly comparing the microbiome profiles of healthy and subfertile women, this thesis contributes to the growing body of research aimed at identifying patterns in microbiome composition and its association with subfertility. Although the causal link between specific vaginal microbiota and infertility remains unclear, this study provides evidence-based findings that may support existing studies or provide another perspective. By specifying the unique characteristics of subfertile women and how they are influenced by external factors, this research may provide insights on disease prevention and potential reproductive health intervention, among others.

The Isala and FLORA project are initiatives that were made possible through stakeholder collaboration. They involved women volunteers who generously provided their personal and sensitive data to serve the project goals, i.e., to better understand the vaginal microbiome and fertility. Behind these projects are experts and researchers in the field of health, microbiology, epidemiology, and gynecology, among others, who play a crucial role in study design and implementation. This thesis aims to benefit these groups by providing a meaningful statistical analysis of microbiome data and its link to reproductive health status and external influences. The findings of this study may also be used to guide the formulation of the study design for targeted-interventional studies or personalized fertility treatment. The output of this thesis contributes to existing studies that can be used by other students and data analysts doing similar research. Most importantly, this study may also benefit women of reproductive age and their family by knowing the factors that could help maintain balance in the genital flora and improve their reproductive health and fertility treatment outcomes.

In terms of ethical consideration, the Isala project was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of UZ Brussel/VUB (1432022000115, approved on 5 July 2022). The FLORA project also followed the same ethical principle. The Declaration of Helsinki, developed by the World Medical Association, outlines the ethical guidelines for conducting medical research involving human subjects. It emphasizes (i) protection of the health, safety, and well-being of human subjects; (ii) informed consent; (iii) privacy and confidentiality of information; (iv) special protection for vulnerable individuals; and (v) conformity of research to generally accepted scientific principles (World Medical Association, 2013). Both the Isala and FLORA projects collected data with the informed consent of the participants. The Isala project obtained a digital informed consent, while the FLORA project used Institutional Review Board-approved informed consent forms administered in a clinic. The Isala project provided volunteers with all the necessary support and materials to ensure informed and safe participation. This

includes sampling kits, brochures, and instructional videos to facilitate the self-sampling, storage, and transport of vaginal swabs. Meanwhile, the FLORA project followed standard clinical practice, avoiding discomfort to patients during vaginal swabs. It also ensured that the procedures do not require anaesthetic, interfere with the fertility treatment, or involve any additional costs to the patients. Both projects ensured strict data confidentiality. Only study staff, the Ethics Committee, and health authorities have access to the medical record and information of the participants. Both projects strictly implemented the non-disclosure of the name and any data related to the identity of participants. The project investigators used sample identification numbers or codes to replace the participant's name prior to transmitting the collected data to the database manager. The organizations behind the projects also recognized and followed the European General Data Protection Regulation (GDPR) in the collection, storage, processing, and protection of personal data of participants.

The datasets from the Isala and FLORA projects were shared with Hasselt University with patient anonymity. All data were handled and processed with strict confidentiality. The information gathered from the participants was used solely for the purposes of this thesis. Statistical analyses were carried out with integrity, ensuring accuracy and transparency in the reporting of findings to the best of the author's ability. Relevant previous studies were appropriately cited to acknowledge existing research and support the current work.

## 2 Data description

In March 2020, UAntwerp opened a call for women volunteers at least 18 years old and not pregnant to be part of the Isala Project, a citizen science project that aims to better understand vaginal microbiome and its association with external factors. After obtaining a digital informed consent, volunteers answered an extensive questionnaire about their background information. Sampling kits were sent to them and they self-collected vaginal swabs in a standardized way (Lebeer et al., 2022). These samples were sent to UAntwerp for laboratory analysis. The raw dataset of the Isala project consists of 3,349 respondents. To extract only healthy women from the dataset, exclusion criteria were applied (**Appendix A**). These criteria include having endometriosis or polycystic ovary syndrome (PCOS), initiated fertility program, and with diabetes or hematologic disorder. Respondents who are currently breastfeeding, had antibiotic treatment in the past three months, and currently using/used contraception were also excluded, as these could already have influenced their microbiome composition. The final dataset for the healthy group also does not include respondents who are older than 46 years to make it comparable to those of the FLORA Project. Of the original 3,349 respondents, the sample size was reduced to 390. This represents the "healthy group" in the subsequent analyses. A separate group (not included in the 390 samples) was also determined from the Isala project comprising the 146 women who reported that a fertility program was once initiated for them. This represents the "benchmark group" in this study and was also compared with women in the FLORA project.

Meanwhile, the FLORA project aims to include 1,000 patients undergoing diagnostic hysteroscopy prior to their IVF treatment in UZ Brussel. In February 2025, the project had gathered the data from 353 patients, which represent the "subfertile group" in this study. Their demographic profiles were collected through a questionnaire and registered in an electronic system by a study nurse. Table 1 summarized the variables from the questionnaire that were used in the analysis of alpha diversity and differential abundance in the subsequent sections. Vaginal swabs were also collected from these patients in UZ Brussel and analyzed in UAntwerp.

Both projects have separate datasets on patient's microbiome samples (i.e., Operational Taxonomic Units). This is based on 16S rRNA amplicon sequencing conducted by the UAntwerp. The raw count datasets of the healthy group and the benchmark group contain 285 and 260 unique taxa at the genus level, respectively, while the FLORA Project has 151 unique taxa. Prior to data analysis, data filtering was performed for the three groups. A prevalence cut-off of 1% was imposed for both the Isala and FLORA datasets to filter out possible contaminants and non-informative taxa. This retained only taxa that are present in at least 1% of the samples.

Table 1: Selected variables from the Isala and FLORA datasets

| Variable | Project | Description |
|---|---|---|
| age | Isala and FLORA | age in years |
| BMI | Isala and FLORA | body mass index categorized into 1-normal (ranging from 18.5 to 25)[1], 0-otherwise |
| smoking | Isala and FLORA | 1-yes, 0-no |
| sleep | Isala | Hours spent sleeping, 0 if from 3 to 6.5; 1 if from 6.5 to 9; and 2 if from 9 to 12 hours |
| | FLORA | Originally collected as actual number of hours, but recoded following the Isala dataset |
| born | Isala and FLORA | Method how the respondent was born, 1-natural means (i.e., vaginal birth), 0-caesarean section |
| fermented food<br>alcohol<br>sweet beverages | Isala | Frequency of consuming fermented food, alcohol, and sweet beverages in the past 3 months (0-never, 1-seldom, 2-monthly, 3-weekly, 4-more than three times a week, 5-daily, 6-multiple times a day) |
| bread | FLORA | 1-eating dark bread, white bread, or sourdough bread, 0-not eating bread |
| alcohol | FLORA | number of glasses consumed in a week |
| soft drinks | FLORA | number of glasses consumed in a week |

# 3 Methodology

## 3.1 Comparison between subfertile and healthy women

The primary objective of this study is to compare subfertile group (from the FLORA project) with healthy group (from the Isala project), and identify external factors associated with their microbiome profiles. The comparison was done through data visualization, estimation of alpha diversity, and analysis of differential abundance.

### 3.1.1 Data exploration

Exploratory data analysis was done to assess the distribution of patients according to characteristics, namely, age, body mass index, smoking habit, hours of sleep, and birth method. Data visualization was done to identify the most abundant taxa for the subfertile and healthy groups. This is based on the relative abundance, which is defined as the proportion of the observed count of a specific taxon and the total observed count in a sample.

[1]The WHO considers BMI less than 18.5 as underweight and above 25 as overweight among adults. (World Health Organization, 2025)

Alpha diversity, a measure of richness and evenness of species within a sample, was estimated using the Shannon diversity index.

$$H_i' = -\sum_{j=1}^{R} p_{ij} \ln p_{ij} \tag{1}$$

where $H_i'$ is the Shannon index for sample $i$, $R$ is the number of unique taxa in sample $i$, $p_{ij}$ is the proportion of the $j$th taxon to the total observed count for sample $i$.

Species richness was also measured using Chao1 index:

$$S_{Chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{2(n_2 + 1)} \tag{2}$$

where $S_{Chao1}$ is the estimated number of species, $S_{obs}$ is the observed total number of genera in a sample, $n_1$ is the number of singletons, and $n_2$ is the number of doubletons. Singletons are genera with counts equal to one in a sample, while doubletons have counts equal to two (Chao, 1987).

### 3.1.2 Modeling of the alpha diversity

The subfertile and healthy groups were compared based on their Shannon index and Chao1 index. One common statistical test to determine whether two groups are different is the t-test. However, in the presence of confounders, a linear regression model is deemed more appropriate. Confounders are variables that affect both the response variable and the covariate of interest, and they should be accounted for in the analysis. Confounding can distort the true or potential outcome-exposure relationship. Failing to account or recognize the confounding effect may result in invalid estimate of causal effect or false conclusion. Therefore, to test whether the alpha diversity of the subfertile group is different from that of the healthy group, the following linear regression model was fitted using the merged datasets of the two groups.

$$y_i = \beta_0 + \beta_1 \text{Group}_i + \beta_2 \text{Age}_i + \beta_3 \text{BMI}_i + \beta_4 \text{Smoking}_i + \beta_5 \text{Sleep1}_i + \beta_6 \text{Sleep2}_i + \beta_7 \text{Born}_i + \epsilon_i \tag{3}$$

where $y_i$ is the alpha diversity measure (Shannon index, Chao1 index) for sample $i$, $i = 1, 2, ..., 743$; $\beta$s are the regression coefficients; $\text{Group}_i$ is the dummy variable for the group to which sample $i$ belongs, i.e., 1 for healthy group and 0 for subfertile group; $\epsilon_i$ is the error term, assumed to be normally distributed with mean 0 and variance $\sigma^2$. While $\text{Group}_i$ is the main covariate of interest, age (in years), BMI (1-normal, 0-otherwise), and smoking (1-yes, 0-no) were also included in the model as confounders (Figure 1). Confounders in this study were identified based on existing literature. Studies have shown that age, smoking, and BMI influence both the alpha diversity and fertility (Bayoumi et al., 2024, Darıcı et al.,

2025, Lebeer et al., 2022). The effect of two other variables, namely, being born through natural means and number of hours spent sleeping in a day were likewise explored to make use of the variables that are common and measured similarly in both datasets. $\text{Sleep1}_i$ is 1 if the hours spent sleeping is between 6.5 and 9 hours, 0 otherwise. $\text{Sleep2}_i$ is 1 if the hours spent sleeping is between 9 and 12 hours, 0 otherwise. $\text{Born}_i$ is 1 if born through natural means (vaginal birth), 0 if through caesarean section. A model with all possible interaction terms between the group and each of the other variables was initially explored. The final model was determined after removing insignificant interaction terms and model diagnostics.

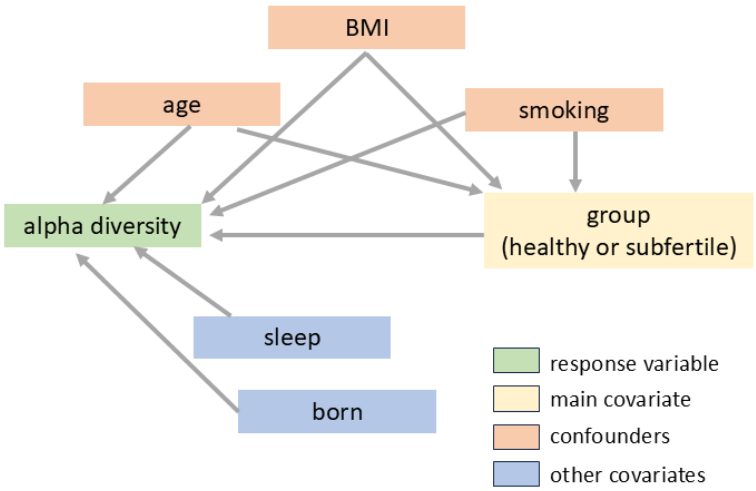



Figure 1: Directed acyclic graph of variables in alpha diversity model

### 3.1.3 Differential abundance analysis

In microbiome data analysis, the compositional effect is observed when changes in the absolute abundance of certain taxa in relation to covariates cause a shift in the relative abundance of all other taxa. As a result, the use of standard statistical methods that do not account for the compositionality in the data can lead to high false discovery rates. One of the recently developed approaches that accounts for this issue is the Analysis of Compositions of Microbiomes with Bias Correction or ANCOM-BC (Lin and Peddada, 2020). Unlike other differential abundance analysis methods, ANCOM-BC takes into account the sampling fraction in the normalization of the data, instead of relying solely on the library size. Failing to account for differences in the sampling fraction can lead to bias and a false conclusion that taxa are not differentially abundant (Lin and Peddada, 2020). ANCOM-BC results also include the analysis of sensitivity to pseudo-count addition. Previous studies have found that the choice of pseudo-counts could influence the results of differential abun-

dance analysis leading to an increased rate of false positives. To avoid this, ANCOM-BC performs pseudo-count sensitivity analysis. Prior to log transformation of observed counts, ANCOM-BC considers pseudo-count values of 0.1, 0.5, and 1 for zero counts. It computes the sensitivity score based on the proportion of times the adjusted p-value exceeds a specified significance level. A taxon is regarded as insensitive to the pseudo-count addition if the taxon's adjusted p-values consistently indicate either significance or non-significance across various pseudo-count adjustments and align with the results from the complete data (i.e., without pseudo-count addition). Lin and Peddada (2022) strongly recommended also taking into consideration the results of the sensitivity analysis in the final assessment of significance. ANCOM-BC uses a log-linear model formulated as follows:

$$y_{ijk} = d_{ik} + u_{jk} + \epsilon_{ijk} \tag{4}$$

where $y_{ijk}$ is the log of the read count of the $j$th taxon of the $i$th sample in the $k$th group;
$d_{ik}$ is the log of the sampling fraction in the $i$th sample from the $k$th group. Lin and Peddada (2020) defined sampling fraction as "the ratio of the expected absolute abundance to the corresponding absolute abundance in the ecosystem, which could be empirically estimated by the ratio of library size to the microbial load";
$u_{jk}$ is the log of the expected absolute abundance of the $j$th taxon in the $k$th group; and
$\epsilon_{ijk}$ is the error term, assumed to be independently distributed with mean 0 and heteroskedastic variance

In this thesis project, ANCOM-BC was utilized to identify differentially abundant taxa and factors associated with taxon abundance in the subfertile and healthy group. First, ANCOM-BC was performed using the merged dataset of the Isala and FLORA project, adjusting for the group variable (1=healthy, 0=subfertile) and confounders (i.e., age, BMI, smoking). This was done to identify which taxa are differentially abundant between groups. Second, ANCOM-BC was performed separately for the two groups, adjusting for eight covariates. Similar to the alpha diversity model, age, BMI, smoking, being born through natural means, and hours of sleep were considered as covariates. In addition, the effect of dietary habits, namely consumption of fermented food, alcoholic beverages, and soft drinks, was also of interest. However, information on these variables was measured differently for the subfertile and healthy group. Specifically, the subfertile group has a binary response (1-yes, 0-no) on eating bread and quantitative response (i.e., number of glasses per week) on drinking alcohol and soft drinks. Meanwhile, the healthy group used the six-level Likert scale (0-never, 1-seldom, 2-monthly, 3-weekly, 4-more than three times a week, 5-daily, 6-multiple times a day) on how frequent the respondents consume the food items in the past three months. Given these differences, the datasets of the two groups were analyzed separately to account for the effect of dietary habits. The separate model for the healthy group and subfertile group can be formulated as follows:

$$y_{ij} = d_i + b_j^T x_i + \epsilon_{ij} \quad (5)$$

where $y_{ij}$ is the log of the read count of the $j$th taxon of the $i$th sample;
$d_i$ is the log of the sampling fraction in the $i$th sample;
$x_i = (1, x_{i1}, x_{i2}, ..., x_{ip})^T$ are the covariates of interest for the $i$th sample;
$b_j = (b_{j0}, b_{j1}, ..., b_{jp})^T$ are the corresponding coefficients for $x_i$ for taxon $j$; and
$\epsilon_{ij}$ is the error term, assumed to be independently distributed with mean 0 and heteroskedastic variance (Lin and Peddada, 2024).

ANCOM-BC allows for the use of the raw count data and specification of the prevalence cut-off and library size cut-off. In this study, prevalence cut-off was set at 10%, which is one of the commonly used thresholds in microbiome analysis. This threshold means that a taxon with nonzero counts in less than 10% of the samples were excluded from the analysis. This narrows down the dimension of the analyses and removes non-informative or rare taxa. A library size cut-off was equal to 1000, which excludes samples with library sizes less than 1000. All tests were adjusted for multiple comparisons using the Benjamini–Hochberg procedure. The false discovery rate was set at 5%. Results of pseudo-count sensitivity analysis were also presented.

### 3.2 Benchmarking subfertile women against those who previously initiated a fertility program

The secondary objective of this study is to compare subfertile women who are currently undergoing fertility treatment (from the FLORA project, referred to in this report as "subfertile" group) with women who had once initiated a fertility program at any point in their lives (from the Isala project, referred to as "benchmark" group). Women in these groups both have fertility issues and are undergoing/underwent fertility treatment, which might have affected their microbiome composition. The comparison aims to assess whether the two groups have the same or different microbiome diversity and taxon abundance. This was done using data visualization, specifically using plots of the most abundant taxa and plots of the Shannon index and Chao1 index. To determine whether there is a difference in alpha diversity between the subfertile and benchmark group, the following linear regression model was fitted using the merged datasets of the two groups.

$$y_i = \beta_0 + \beta_1 \text{Group}_i + \beta_2 \text{Age}_i + \beta_3 \text{BMI}_i + \beta_4 \text{Smoking}_i + \beta_5 \text{Sleep1}_i + \beta_6 \text{Sleep2}_i + \beta_7 \text{Born}_i + \epsilon_i \quad (6)$$

where $y_i$ is the alpha diversity measure (Shannon index, Chao1 index) for sample $i$, $i = 1, 2, ..., 499$; $\beta$s are the regression coefficients; $\text{Group}_i$ is the dummy variable for the group to which sample $i$ belongs, i.e., 1 for benchmark group and 0 for subfertile group; $\epsilon_i$ is the error term, assumed to be normally distributed with mean 0 and variance $\sigma^2$. Age (in years),

BMI (1-normal, 0-otherwise), and smoking (1-yes, 0-no) were also included in the model as confounders. The effect of two other variables, namely, being born through natural means and number of hours spent sleeping in a day were likewise explored. $\text{Sleep1}_i$ is 1 if the hours spent sleeping is between 6.5 and 9 hours, 0 otherwise. $\text{Sleep2}_i$ is 1 if the hours spent sleeping is between 9 and 12 hours, 0 otherwise. $\text{Born}_i$ is 1 if born through natural means (vaginal birth), 0 if through caesarean section. A model with all possible interaction terms between the group and each of the other variables was initially explored. The final model was determined after removing insignificant interaction terms and model diagnostics.

To determine differentially abundant taxa and factors associated with the microbiome composition of the benchmark group, differential abundance analysis was also done using ANCOM-BC. First, ANCOM-BC was performed using the merged dataset of the Isala and FLORA project, adjusting only for the group variable (1=benchmark, 0=subfertile) and confounders (age, BMI, smoking). Second, ANCOM-BC was performed separately for the two groups, adjusting for eight covariates, namely age, BMI, smoking, being born through natural means, hours spent sleeping, consumption of fermented food, drinking alcohol, and drinking soft drinks. A prior assessment of the correlation matrix of these covariates was performed to ensure that none of these factors is strongly correlated with another factor (**Appendix B**).

### 3.3 Software

The analyses in this project were performed using the R version 4.4.3 software (R Core Team, 2025). Among the R libraries used were *ggplot*2 (Wickham, 2016), *phyloseq* (McMurdie and Holmes, 2013), *tidyverse* (Wickham et al., 2019), *tidytacos* (Wittouck et al., 2025), *vegan* (Oksanen et al., 2025), *psych* (William Revelle, 2025), *ANCOMBC* (Lin and Peddada, 2024), *fossil* (Vavrek, 2011), and *pheatmap* (Kolde, 2019).

# 4 Results

## 4.1 Comparison between subfertile and healthy women

The primary objective of this study is to compare subfertile women with healthy women. This section focuses on assessing the similarities or differences between these groups in terms of respondents' profiles, alpha diversity, and taxon abundance. Factors associated with the microbiome composition were also discussed.

### 4.1.1 Data exploration

Table 2 shows the summary statistics for the respondents' profiles. The age of subfertile women has a mean of 36 and ranges from 24 to 46, while the age of healthy women has a mean of 31 and ranges from 18 to 46. Almost 90% of the women in both groups were born through natural means (i.e., vaginal birth). More than half of the subfertile (54%) and healthy (64%) group has a normal BMI, while the rest are underweight, overweight, or obese. Only a few respondents (5%) reported that they were smoking, and the majority (78-80%) have 6.5 to 9.0 hours of sleep every night.

Table 2: Summary statistics of respondents' profiles

| Variable | Subfertile (n=353) | Healthy (n=390) |
|---|---|---|
| Age (in years) | | |
| *Mean* | 36 | 31 |
| *Std. dev.* | 4.6 | 6.4 |
| *Minimum* | 24 | 18 |
| *Maximum* | 46 | 46 |
| Born through natural means | 89.7% | 89.0% |
| BMI (normal) | 54.2% | 63.7% |
| Smoking (yes) | 4.8% | 4.6% |
| Hours of sleep | | |
| *From 3 to 6.5 hours* | 9.7% | 14.1% |
| *From 6.5 to 9 hours* | 77.7% | 80.0% |
| *From 9 to 12 hours* | 12.6% | 5.9% |

In terms of the microbiome data, the subfertile and healthy cohorts exhibit sparsity and skewness in library sizes. Figure 2 illustrates a highly skewed distribution of library size in both groups. The subfertile group has an average library size of 31,310 reads, ranging from 1,663 to 140,381, while the healthy group averages 26,183 reads, with a range of 2,457 to 160,768. Despite the skewness, the average library sizes of the two groups are relatively similar. Regarding sparsity, the data reveal that 89% of the read counts in the subfertile group and 85% in the healthy group are zeros, reflecting a high degree of sparsity in both datasets.

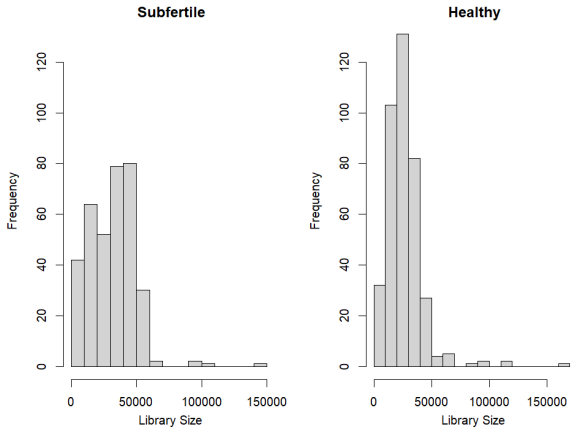



Figure 2: Histogram of library size of the subfertile and healthy group

Figure 3 presents the most abundant taxa in the two groups based on mean relative abundance. The subfertile group is dominated by *Lactobacillus crispatus* (37.9%), *Lactobacillus iners* (30.7%), *Bifidobacterium* (8.6%), *Lactobacillus jensenii* (5.8%), and *Lactobacillus gasseri* (4.2%). It is also composed of *Fannyhessea* (3.8%), *Prevotella* (2.9%), *Sneathia* (1.6%), *Anaerococcus* (0.6%), and other taxa (3.8%). For the healthy group, the most abundant taxa are *Lactobacillus crispatus* (36.9%), *Lactobacillus iners* (27.6%), *Bifidobacterium* (9.9%), *Lactobacillus jensenii* (5.5%), and *Prevotella* (4.2%). Also present in this group's microbiome are *Lactobacillus gasseri* (3.4%), *Fannyhessea* (1.9%), *Anaerococcus* (1.3%), *Streptococcus* (1.3%), and other taxa (8.1%).

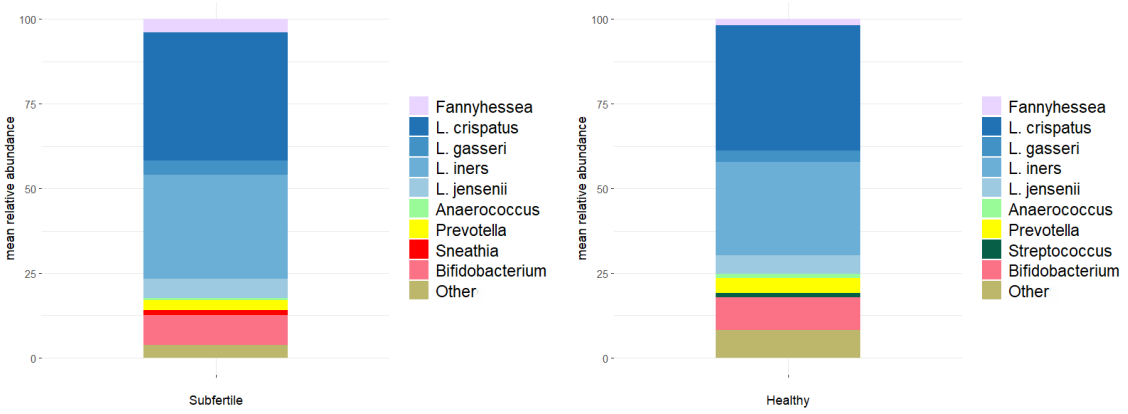



Figure 3: Mean relative abundance of top 10 taxa in subfertile and healthy group

Figure 4 shows the relative abundance of the top taxa by sample in the subfertile and healthy group. The microbiome composition of most women in both groups is dominated by *Lactobacillus crispatus* and *Lactobacillus iners*. Some subfertile women exhibit a high

relative abundance of *Sneathia*, which was not observed in any of the women in the healthy group. Meanwhile, some women in the healthy group show a high relative abundance of *Streptococcus*, which was not observed in any women in the subfertile group.

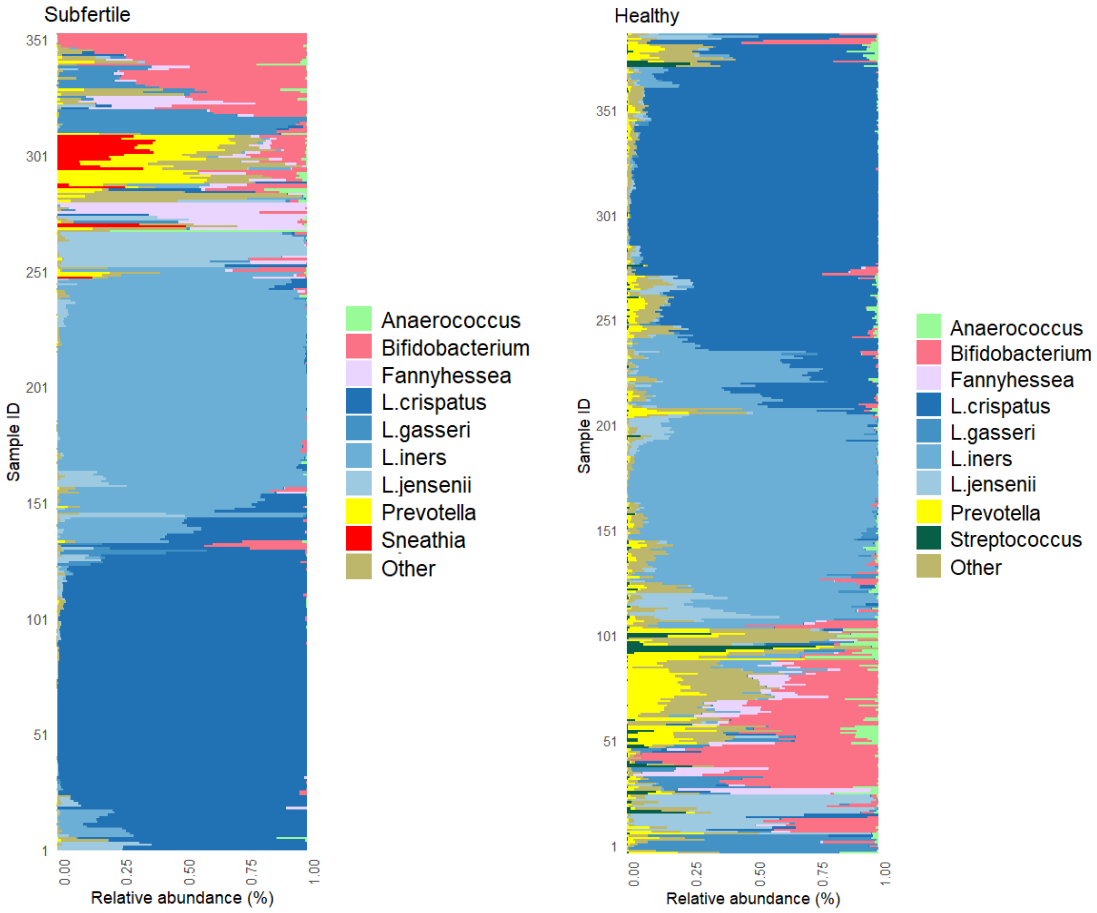



Figure 4: Relative abundance of top 10 taxa by sample and group

Using the respondents' profiles, differences in species richness were initially assessed using Chao1 index plots. Chao1 index is a common estimator of species richness that takes into account the observed number of species and the number of rare species (i.e., singletons and doubletons). Figure 5 reveals that the subfertile group has a lower median Chao1 index (4) than the healthy group (19). This indicates that compared to the healthy group, the subfertile group has lower microbial richness, even after accounting for rare taxa (singletons and doubletons). Slightly higher Chao1 index was also observed for women with less than 6.5 hours of sleep (16) than those with 6.5 to 9 hours of sleep (12) and more than 9 hours of sleep (5). Meanwhile, the plot of the Chao1 index against age does not exhibit an increasing or decreasing trend with respect to changes in age. Chao1 index was also observed to be similar between normal (11) and not normal BMI (12); smoker (13) and non-smoker (11);

and women born through natural means (12) and caesarean section (11).

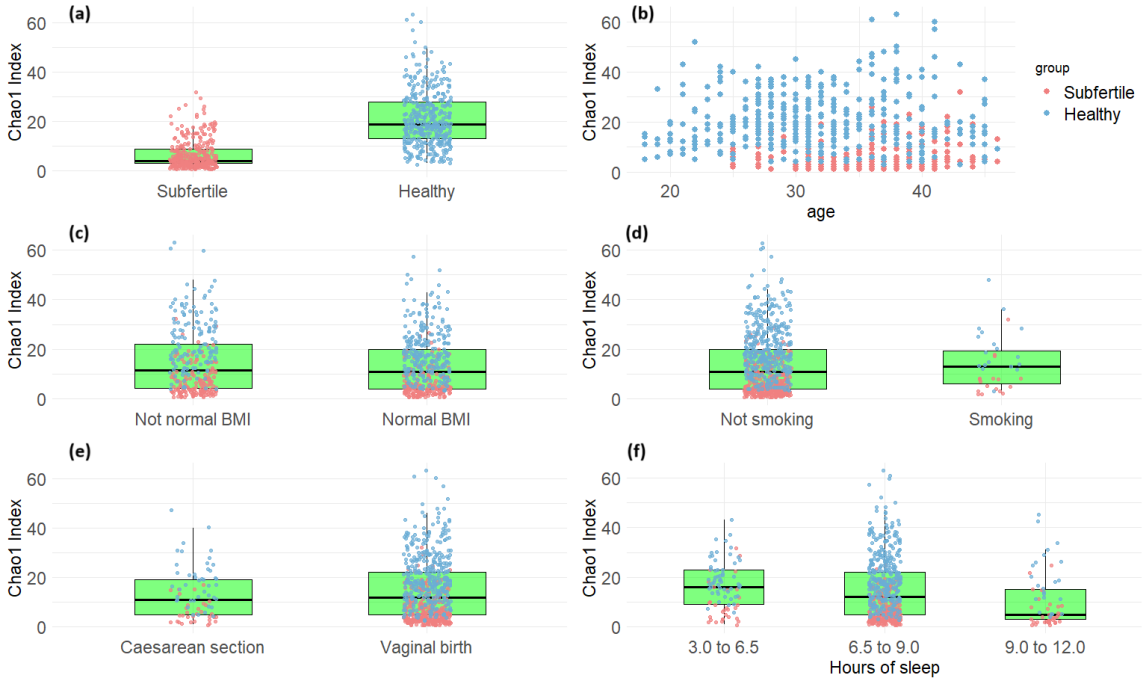



Figure 5: Plots of Chao1 index by respondents' profile

The five panels show the boxplot of Chao1 index (a) between subfertile and healthy group; (c) by BMI category; (d) by smoking status; (e) by method of birth; and (f) by number of hours spent sleeping. Panel (b) is a scatterplot of samples by Chao1 index and age. Dots in all panels represent the samples, color coded by group.

Similarly, the differences in Shannon index by respondent's profiles were also assessed. Shannon index is a measure of species richness and evenness in a sample. Figure 6 shows a low Shannon index for the two groups, with subfertile and healthy women having a median Shannon index of 0.19 and 0.73, respectively. The low Shannon index can be attributed to the dominance of *Lactobacillus* species in either group. Women who have less time spent sleeping (3 to 6.5 hours) exhibited a slightly higher alpha diversity (0.95) than those with 6.5 to 9 hours (0.49) or 9 to 12 hours of sleep (0.42). The statistical significance of these differences was discussed in Section 4.1.2, while biological relevance was discussed in Section 5.1. In contrast, the plot of the Shannon index against age does not exhibit an increasing or decreasing trend in alpha diversity with respect to changes in age. Shannon index was also observed to be similar between normal (0.49) and not normal BMI (0.56); smoker (0.61) and non-smoker (0.50); and women born through natural means (0.55) and caesarean section (0.43).

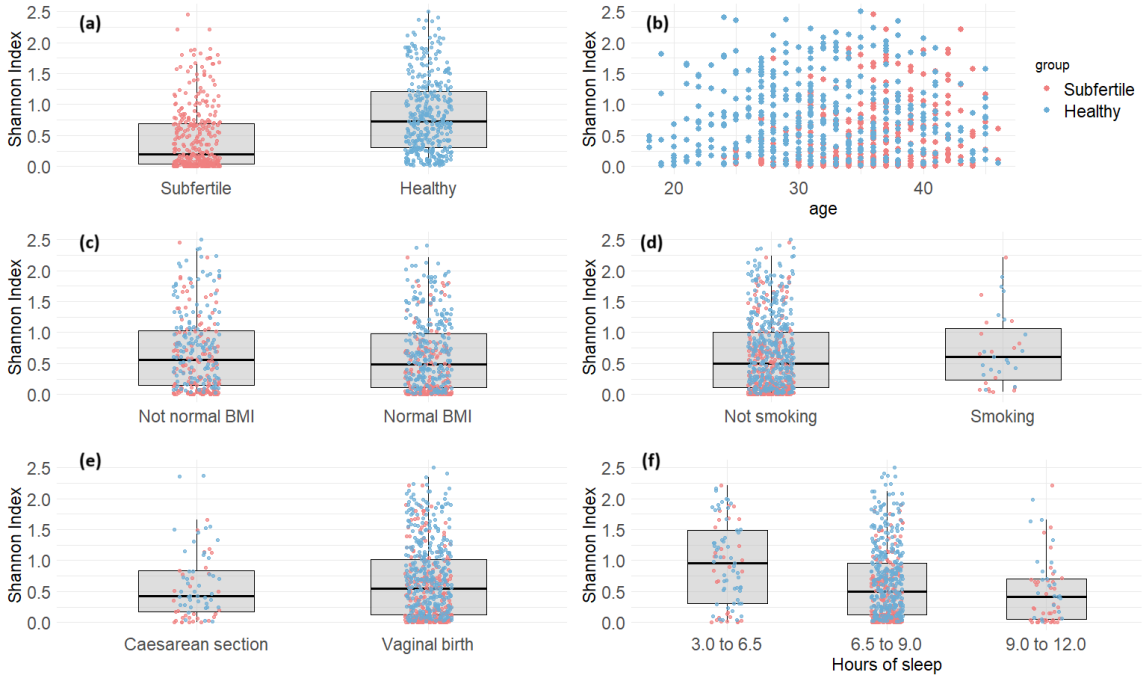



Figure 6: Plots of Shannon index by respondents' profile

The five panels show the boxplot of Shannon index (a) between subfertile and healthy group; (c) by BMI category; (d) by smoking status; (e) by method of birth; and (f) by number of hours spent sleeping. Panel (b) is a scatterplot of samples by Chao1 index and age. Dots in all panels represent the samples, color coded by group.

### 4.1.2 Model for alpha diversity

While Section 4.1.1 described the data through visual inspection, Section 4.1.2 was designed to statistically test the difference between two groups. A linear regression model, as formulated in Equation 3 of Section 3.1.2, was initially fitted. The first model used the estimated Shannon index as the response variable, six main effects (i.e., group, age, BMI, smoking, sleep, born) and the interaction terms between group and each of the five other covariates. This model was then reduced to remove insignificant interaction terms and retained the six main effects. The model diagnostics was performed prior to the interpretation of the regression estimates (**Appendix C**).

Table 3 presents the parameter estimates from the final model for Shannon index. The results showed that the Shannon index for healthy women is significantly higher than that of subfertile women by 0.40 on average, holding other factors constant. The true difference in Shannon index from that of subfertile women could lie somewhere between 0.31 and 0.49. This suggests that the microbiome composition of subfertile women is slightly lower in richness and evenness compared to that of healthy women. Interpretation or possible

biological relevance of this difference was discussed in Section 5.1.

Table 3: Regression estimates from the model for Shannon index

| Parameter | Estimate | Std. error | p-value | 95% confidence interval |
|---|---|---|---|---|
| Intercept | 0.5482 | 0.1628 | 0.0008 | 0.2285, 0.8678 |
| group (healthy) | 0.3972 | 0.0458 | <0.0001 | 0.3073, 0.4872 |
| age | 0.0039 | 0.0038 | 0.2992 | -0.0035, 0.0113 |
| BMI (normal) | -0.0911 | 0.0432 | 0.0354 | -0.1760, -0.0062 |
| smoking (yes) | -0.0167 | 0.1020 | 0.8693 | -0.2170, 0.1835 |
| sleep (6.5 to 9h) | -0.2832 | 0.0652 | <0.0001 | -0.4112, -0.1553 |
| sleep (9 to 12h) | -0.3377 | 0.0943 | 0.0005 | -0.5129, -0.1424 |
| born (natural means) | 0.0711 | 0.0681 | 0.2963 | -0.0625, 0.2048 |

Holding other factors constant (i.e., women are of the same group, age, BMI, smoking habit, and birth method), the Shannon index is significantly lower for women who spend more than 6.5 hours of sleep every night. Specifically, the Shannon index for women with 6.5 to 9 hours of sleep is estimated to be lower by 0.16 to 0.41 than for those with less than 6.5 hours of sleep. Similarly, a Shannon index lower by 0.14 to 0.51 is estimated for women with 9 to 12 hours of sleep than for those with less than 6.5 hours of sleep.

In addition, women with normal BMI were estimated to have a Shannon index lower by 0.006 to 0.176 than those who were underweight, overweight or obese. For each year of increase in age, women's Shannon index varies between a decrease of 0.004 and an increase of 0.011. The difference in Shannon index between smokers and non-smokers ranges from -0.22 to 0.18. Compared to women delivered by caesarean section, those born naturally could have Shannon index lower by 0.06 or higher by as much as 0.20. None of these factors - age, smoking status, or birth method - showed statistical significance.

Table 4 shows the parameter estimates from the model for Chao1 index. The results showed that the Chao1 index for healthy women is significantly higher than that of subfertile women. Holding other factors constant, the true difference in the Chao1 index from that of subfertile women could be somewhere between 14 and 17. This suggests that subfertile women could have fewer genera present in their microbiome composition than healthy women. Women with normal BMI were also found to have a significantly lower Chao1 index by 2 to 4 than women who are underweight, overweight or obese. In contrast, the effect of age, smoking, hours of sleep, and method of birth were found to be statistically insignificant.

Table 4: Regression estimates from the model for Chao1 index

| Parameter | Estimate | Std. error | p-value | 95% confidence interval |
|---|---|---|---|---|
| Intercept | 3.9706 | 2.6269 | 0.1310 | -1.1872, 9.1284 |
| group (healthy) | 15.3844 | 0.7392 | <0.0001 | 13.9331, 16.8356 |
| age | 0.0804 | 0.0608 | 0.1860 | -0.0388, 0.1997 |
| BMI (normal) | -2.9664 | 0.6976 | <0.0001 | -4.3361, -1.5966 |
| smoking (yes) | 0.1010 | 1.6456 | 0.9510 | -3.1300, 3.3319 |
| sleep (6.5 to 9h) | -0.1940 | 1.0513 | 0.8540 | -2.2582, 1.8702 |
| sleep (9 to 12h) | -1.3804 | 1.5225 | 0.3650 | -4.3697, 1.6088 |
| born (natural means) | 1.4728 | 1.0981 | 0.1800 | -0.6833, 3.6289 |

#### 4.1.3 Differential abundance analysis

As discussed in Section 3.1.3, ANCOM-BC was first implemented using the merged dataset of the Isala and FLORA projects, adjusting for the group variable (1=healthy, 0=subfertile) and confounders (age, BMI, smoking). This was done to identify differentially abundant taxa between groups. This was followed by ANCOM-BC performed separately for the subfertile and healthy group, adjusting for eight covariates.

Figure 7 illustrates the log-fold change of differentially abundant taxa, with only the group variable (and confounders) in the model. These 28 taxa were found to be significant at the 5% FDR level and passed the sensitivity analysis for the pseudo-count addition. Adjusted for the effect of age, BMI, and smoking habit, the abundance of *Lactobacillus jensenii*, *Sarcina*, *Prevotella*, *Corynebacterium* and 13 other taxa are expected to be higher for the healthy group than subfertile group. Specifically, *Lactobacillus jensenii* showed a log-fold change of 0.47 indicating that it is approximately 1.6 [exp(0.47)] times higher in the healthy group than in the subfertile group. The log-fold change of *Sarcina* (1.9), *Prevotella* (1.1), and *Corynebacterium* (1.1) suggests that these taxa are approximately 3 to 6 times higher in the healthy group. In contrast, the abundance of *Mobiluncus*, *Aerococcus*, *Ureaplasma*, *Bifidobacterium*, *Fannyhessea*, *Lactobacillus crispatus*, *Lactobacillus gasseri*, and four other taxa are lower in the healthy group than subfertile group. The log-fold change for these taxa ranging from -1.90 to -0.35 suggests that their abundance is approximately 30% to 85% lower in the healthy group than in the subfertile group.

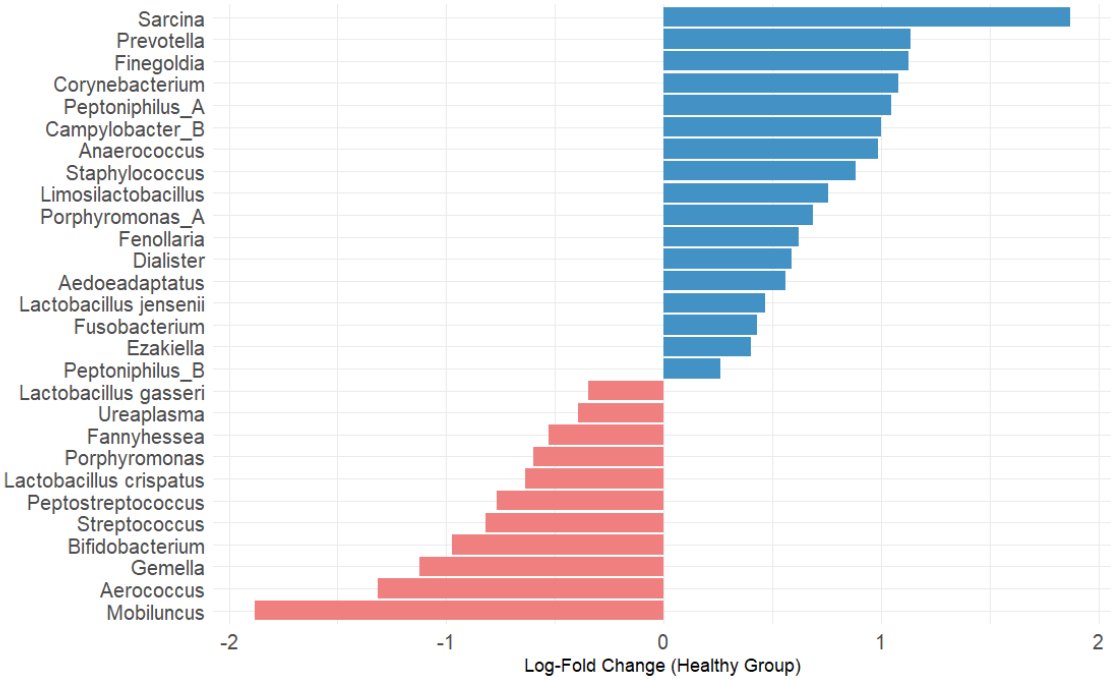



Figure 7: Log-Fold Change of Differentially Abundant Taxa, adjusted for group and confounders
This figure shows the log-fold change for each differentially abundant taxa, adjusted for group, age, BMI, and smoking habit. The 28 taxa are significant at the 5% FDR level and passed the ANCOM-BC sensitivity analysis for pseudo-count addition. Blue bars indicate higher abundance of a taxon in the healthy group, while red bars indicate higher abundance in the subfertile group.

## Healthy group

Figure 8 presents the log-fold change of the 39 differentially abundant taxa for the healthy group, adjusting for the eight covariates. These taxa were found to be significant at the 5% FDR level for at least one covariate (marked with black and green asterisks). Twelve taxa (marked with a green asterisk) passed the sensitivity analysis for a specific covariate, indicating that they are consistently significant across various pseudo-count adjustments. Taxa that did not pass the sensitivity analysis means that their statistical significance may be more sensitive to the choice of pseudo-counts used to address zero values in the data. However, this does not necessarily indicate that they are false positives, particularly given the known conservativeness of the ANCOM-BC2 sensitivity test, which may exclude biologically relevant signals.

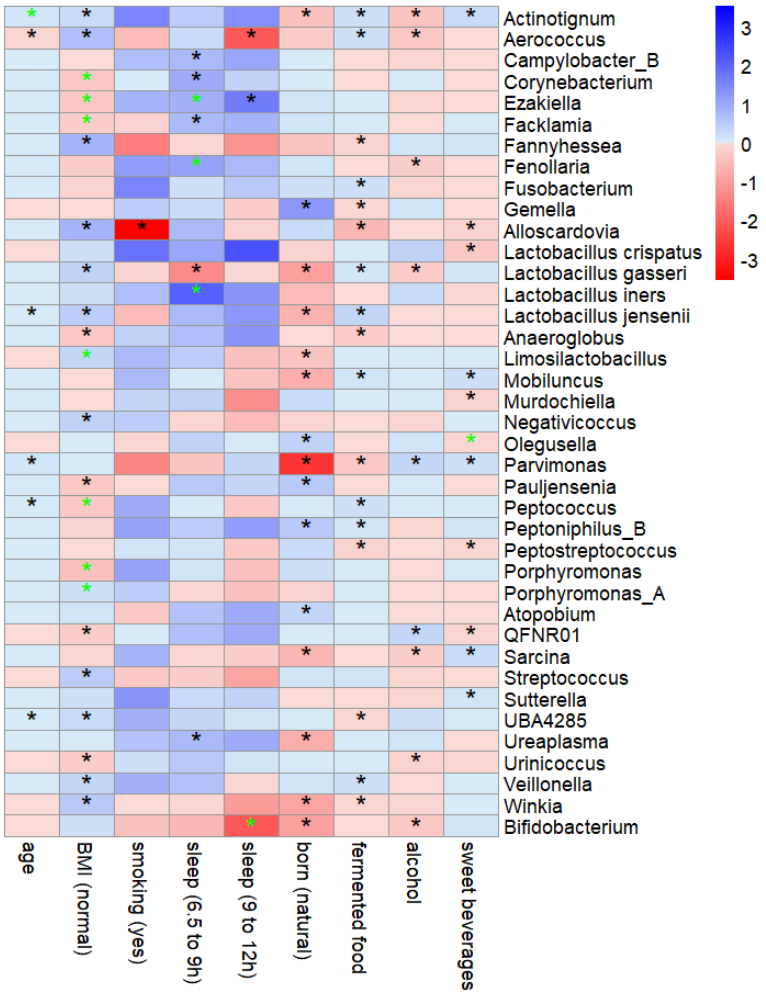



Figure 8: Log Fold Change of the Differentially Abundant Taxa for the Healthy Group

The figure presents the differentially abundant taxa, adjusted for age, BMI, smoking habit, hours of sleep, frequency of consuming fermented food, alcohol intake, and drinking sweet beverages. Red cells represent a log-fold change between 0 and -3, suggesting lower taxon abundance. Blue cells represent a log-fold change between 0 and 3, suggesting higher taxon abundance. Cells with asterisk indicate significance at 5% FDR level. Asterisks in green indicate that taxa have successfully passed the ANCOM-BC2 sensitivity analysis for pseudo-count addition. While this analysis is useful for assessing robustness, it is known to be conservative and may exclude biologically relevant signals. Therefore, all differentially abundant taxa, regardless of passing/failing the sensitivity analysis, were presented.

*Lactobacillus* species are among the 39 differentially abundant taxa found for the healthy group. *Lactobacillus* species are known to have an essential role in maintaining a healthy vaginal environment. The results revealed that the abundance of *Lactobacillus crispatus* is negatively associated with the frequency of drinking sweet beverages. The abundance of *Lactobacillus gasseri* is expected to be higher for healthy women with normal BMI than for those who are underweight, overweight, or obese. In contrast, it is lower for women with 6.5 to 9 hours of sleep than for those with less than 6.5 hours of sleep. It is also lower for women

who were born through vaginal birth than for those delivered through caesarean section. While the abundance of *Lactobacillus gasseri* is positively associated with the frequency of eating fermented food, it is negatively associated with drinking alcohol. The abundance of *Lactobacillus iners* is higher for women with 6.5 to 9 hours of sleep than for those with less than 6.5 hours. The abundance of *Lactobacillus jensenii* is positively associated with age and frequency of eating fermented food. It is expected to be higher for women with normal BMI and lower for women who were born through natural means.

The abundance of *Actinotignum* is positively associated with age and frequency of consuming fermented food and sweet beverages, but negatively associated with the frequency of alcohol intake. It is also expected to be higher for women with normal BMI, but lower for women who were born through vaginal birth. The abundance of *Corynebacterium* is expected to be lower for women with normal BMI than for those with non-normal BMI. It is higher for women with 6.5 to 9 hours of sleep than for those with less than 6.5 hours. The abundance of *Ezakiella* tends to be lower for women with normal BMI and higher for those with more than 6.5 hours of sleep. The abundance of *Facklamia* is estimated to be lower for women with normal BMI and higher for women with 6.5 to 9 hours of sleep. The abundance of *Fenollaria* is likewise higher for women who spent 6.5 to 9 hours sleeping. *Limosilactobacillus* is higher for women with normal BMI but lower for those born through vaginal birth. The abundance of *Olegusella* is higher for women born naturally than those born by caesarean section, but it is negatively associated with the frequency of drinking sweet beverages. The abundance of *Peptococcus* is higher for women with normal BMI and those who frequently eat fermented food. The abundance of *Porphyromonas* is lower for women with normal BMI than for those who are underweight, overweight or obese. The abundance of *Bifidobacterium* is higher for women with 9 to 12 hours of sleep than for those with less than 6.5 hours of sleep.

Results further revealed significant association of the abundance of *Aerococcus*, *Parvimonas*, *Peptococcus*, *UBA4285* with age. The abundance of *Aerococcus*, *Fannyhessea*, *Alloscardovia*, and five other taxa is lower for women with normal BMI. The abundance of *Alloscardovia* is expected to be lower for smokers. The abundance of *Campylobacter B*, *Corynebacterium*, *Facklamia*, and *Ureaplasma* is higher for women with more than 6.5 hours of sleep. The abundance of *Parvimonas*, *Mobiluncus*, *Sarcina*, *Ureaplasma* and *Winkia* is expected to be lower for women born through natural means. A positive association was found between frequency of eating fermented food and the abundance of *Aerococcus*, *Fusobacterium*, *Mobiluncus*, *Peptococcus*, *Peptoniphilus B*, and *Veillonella*. Meanwhile, the abundance of *Aerococcus*, *Fenollaria*, *Sarcina*, and *Urinicoccus* is negatively associated with alcohol intake. A positive association was found between frequency of drinking sweet beverages and the abundance of *Mobiluncus*, *Parvimonas*, *Sarcina*, and *Sutterella*; while it has a negative association with *Alloscardovia*, *Murdochiella*, *Peptostreptococcus*,

and *QFNR01*.

### Subfertile group

Figure 9 shows the log-fold change of the 13 differentially abundant taxa for the subfertile group, adjusting for eight covariates. The abundance of *Lactobacillus crispatus* is higher for women with normal BMI than for underweight, overweight, or obese. It is also expected to be higher for women with 9 to 12 hours of sleep than for those with less than 6.5 hours of sleep. The abundance of *Lactobacillus gasseri* is expected to be higher for women with normal BMI, but lower for women who were born through vaginal birth. The abundance of *Lactobacillus iners* is negatively associated with age. It is lower for women with normal BMI, but higher for women with more than 6.5 hours of sleep than for those with less than 6.5 hours of sleep.

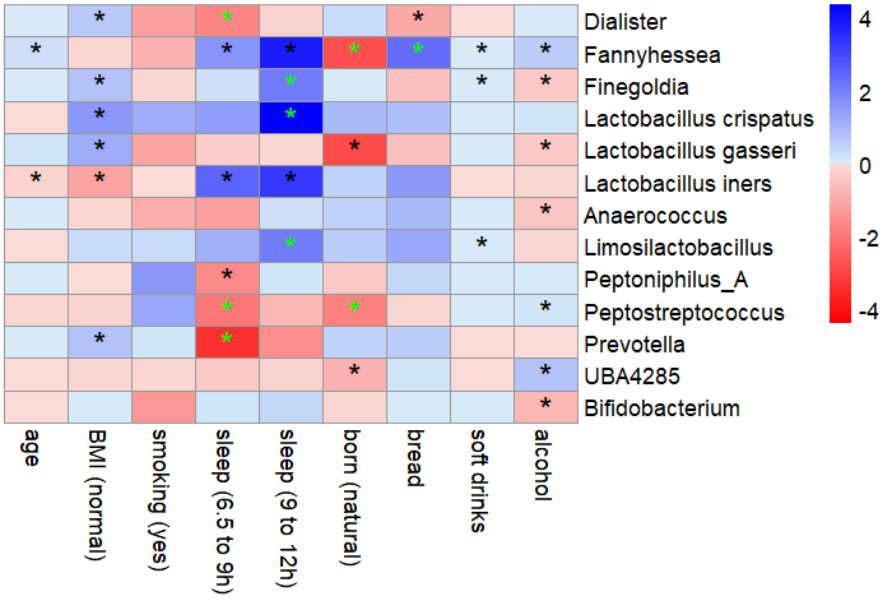



Figure 9: Log Fold Change of the Differentially Abundant Taxa for the Subfertile Group
The figure presents the differentially abundant taxa, adjusted for age, BMI, smoking habit, hours of sleep, frequency of consuming fermented food, alcohol intake, and drinking sweet beverages. Red cells represent a log-fold change between 0 and -4, suggesting lower taxon abundance. Blue cells represent a log-fold change between 0 and 4, suggesting higher taxon abundance. Cells with asterisk indicate significance at 5% FDR level. Asterisks in green indicate that taxa have successfully passed the ANCOM-BC2 sensitivity analysis for pseudo-count addition. While this analysis is useful for assessing robustness, it is known to be conservative and may exclude biologically relevant signals. Therefore, all differentially abundant taxa, regardless of passing/failing the sensitivity analysis, were presented.

Results further revealed that the abundance of *Dialister* is higher for subfertile women with normal BMI than those with non-normal BMI. In contrast, it is lower for women with 6.5 to 9 hours of sleep than for those with less than 6.5 hours. The abundance of *Fannyhessea* is positively associated with age, alcohol intake, and soft drink consumption. It is also higher for women with more than 6.5 hours of sleep and those who eat bread, but lower for women born through natural delivery. The abundance of *Finegoldia* is expected to be higher for women with normal BMI than non-normal BMI. It is also higher for those with 9 to 12 hours of sleep than those with less than 6.5 hours of sleep. The abundance of *Limosilactobacillus* is higher for women with 9 to 12 hours of sleep than for those with less than 6.5 hours of sleep, and for women who eat bread. The abundance of *Peptostreptococcus* is positively associated with alcohol intake. It is lower for women with 6.5 to 9 hours of sleep than for those with less than 6.5 hours of sleep. It is also lower for women born naturally. The abundance of *Prevotella* is expected to be lower for women with normal BMI. It is also lower for those with 6.5 to 9 hours of sleep than for those with less than 6.5 hours of sleep. The abundance of *Anaerococcus* is negatively associated with alcohol intake. The abundance of *Peptoniphilus A* is expected to be lower for women with 6.5 to 9 hours of sleep than for those with less than 6.5 hours. The abundance of $UBA4285$ is expected to be lower for women born through natural means. It is also positively associated with alcohol intake.

## 4.2 Benchmarking subfertile women against those who initiated fertility program

The secondary objective of this study is to compare subfertile women who are currently undergoing fertility treatment (from the FLORA project, referred to in this report as "subfertile" group) with women who had once initiated a fertility program at any point in their lives (from the Isala project, referred to as "benchmark" group). Women in these groups both have fertility issues and are undergoing/underwent fertility treatment, which might have affected their microbiome composition. This section focuses on assessing whether the two groups have the same or different microbiome diversity and taxon abundance. The comparison between subfertile and benchmark groups was done through data visualization, alpha diversity model, and differential abundance analysis.

### 4.2.1 Data exploration

Table 5 shows that the age of women in the benchmark group has a mean of 30 and ranges from 18 to 46. More than 90% of them were born through natural means (i.e., vaginal birth). Around 60% have a normal BMI, while the rest are underweight, overweight, or obese. Only a few (6%) reported that they were smoking, and the majority (77%) have 6.5 to 9.0 hours of sleep every night. These characteristics are almost similar to those of women in the subfertile group.

Table 5: Summary statistics of respondents' profiles

| Variable | Subfertile (n=353) | Benchmark (n=146) |
|---|---|---|
| Age (in years) | | |
| *Mean* | 36 | 30 |
| *Std. dev.* | 4.6 | 6.2 |
| *Minimum* | 24 | 18 |
| *Maximum* | 46 | 46 |
| Born through natural means | 89.7% | 90.4% |
| BMI (normal) | 54.2% | 59.6% |
| Smoking (yes) | 4.8% | 6.2% |
| Hours of sleep | | |
| *From 3 to 6.5 hours* | 9.7% | 16.4% |
| *From 6.5 to 9 hours* | 77.7% | 77.4% |
| *From 9 to 12 hours* | 12.6% | 6.2% |

Figure 10 presents the most abundant taxa in the two groups based on mean relative abundance. As discussed in Section 4.1.1, the subfertile group is dominated by *Lactobacillus crispatus* (37.9%), *Lactobacillus iners* (30.7%), *Bifidobacterium* (8.6%), *Lactobacillus jensenii* (5.8%), and *Lactobacillus gasseri* (4.2%). It is also composed of *Fannyhessea* (3.8%), *Prevotella* (2.9%), *Sneathia* (1.6%), *Anaerococcus* (0.6%), and other taxa (3.8%).

Meanwhile, for the benchmark group, the most abundant taxa are *Lactobacillus crispatus* (44.6%), *Lactobacillus iners* (20.4%), *Bifidobacterium* (8.2%), *Lactobacillus jensenii* (6.8%), and *Prevotella* (3.9%). Also present in this group's microbiome are *Lactobacillus gasseri* (2.5%), *Streptococcus* (2.1%) *Anaerococcus* (1.5%), *Finegoldia* (0.9%), and other taxa (9.1%).

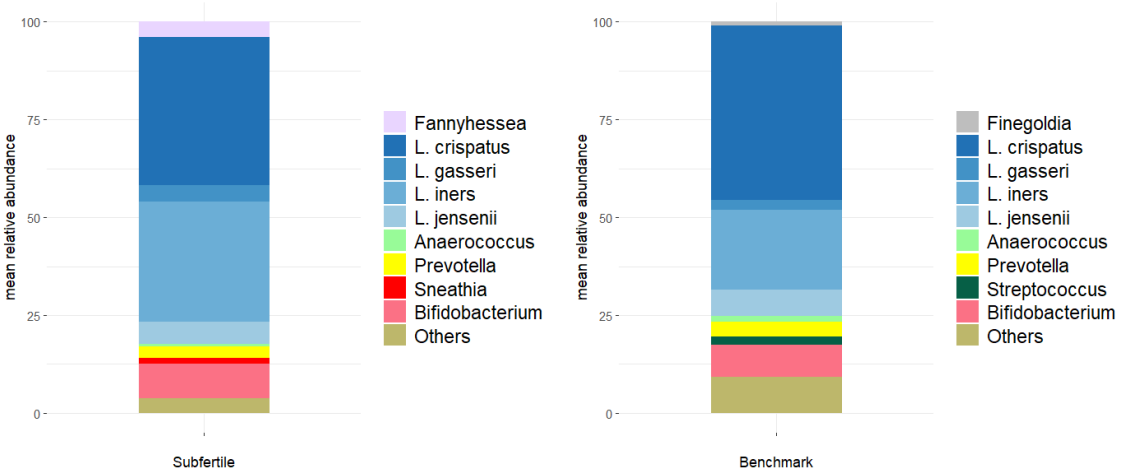



Figure 10: Mean relative abundance of top 10 taxa in subfertile and benchmark group

Figure 11 shows the relative abundance of the top taxa by sample in the subfertile and benchmark group. The microbiome composition of most women in both groups is dominated by *Lactobacillus crispatus* and *Lactobacillus iners*. Some subfertile women exhibit a high relative abundance of *Sneathia*, which was not observed in any of the women in the benchmark group. Meanwhile, some women in the benchmark group show a high relative abundance of *Streptococcus*, which was not observed in any women in the subfertile group.

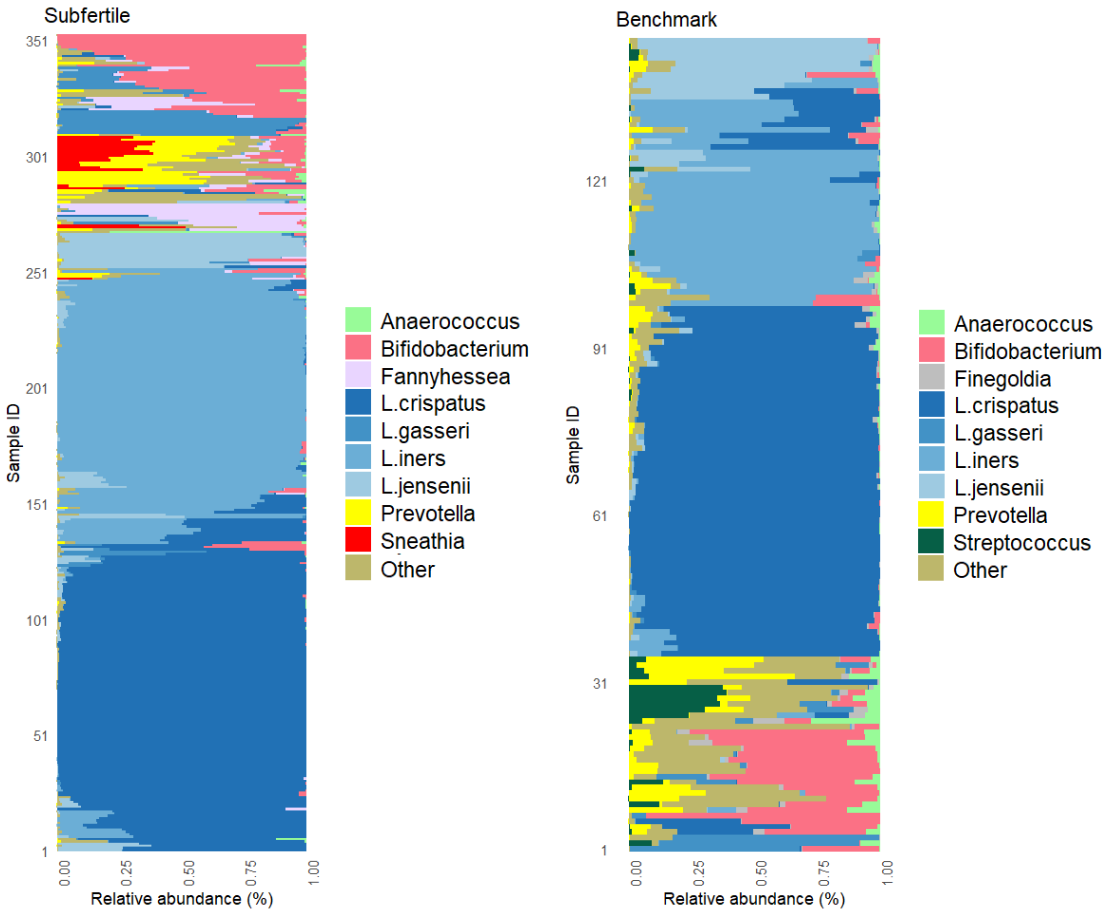



Figure 11: Relative abundance of top 10 taxa by sample and group

Figure 12 presents the plot of the Chao1 index by prespondent's profile. Subfertile group has a lower median Chao1 value (4) compared to the benchmark group (17). This suggests that subfertile women have reduced microbial richness, i.e., fewer genera, compared to the benchmark group. Slightly higher Chao1 index was also observed for women with less than 6.5 hours of sleep (12) than those with 6.5 to 9 hours of sleep (7) and more than 9 hours of sleep (4). Meanwhile, the plot of the Chao1 index against age does not exhibit an increasing or decreasing trend with respect to changes in age. Chao1 index was also observed to be similar between normal (6) and not normal BMI (8); smoker (8) and non-smoker (6); and women born through natural means (7) and caesarean section (8).

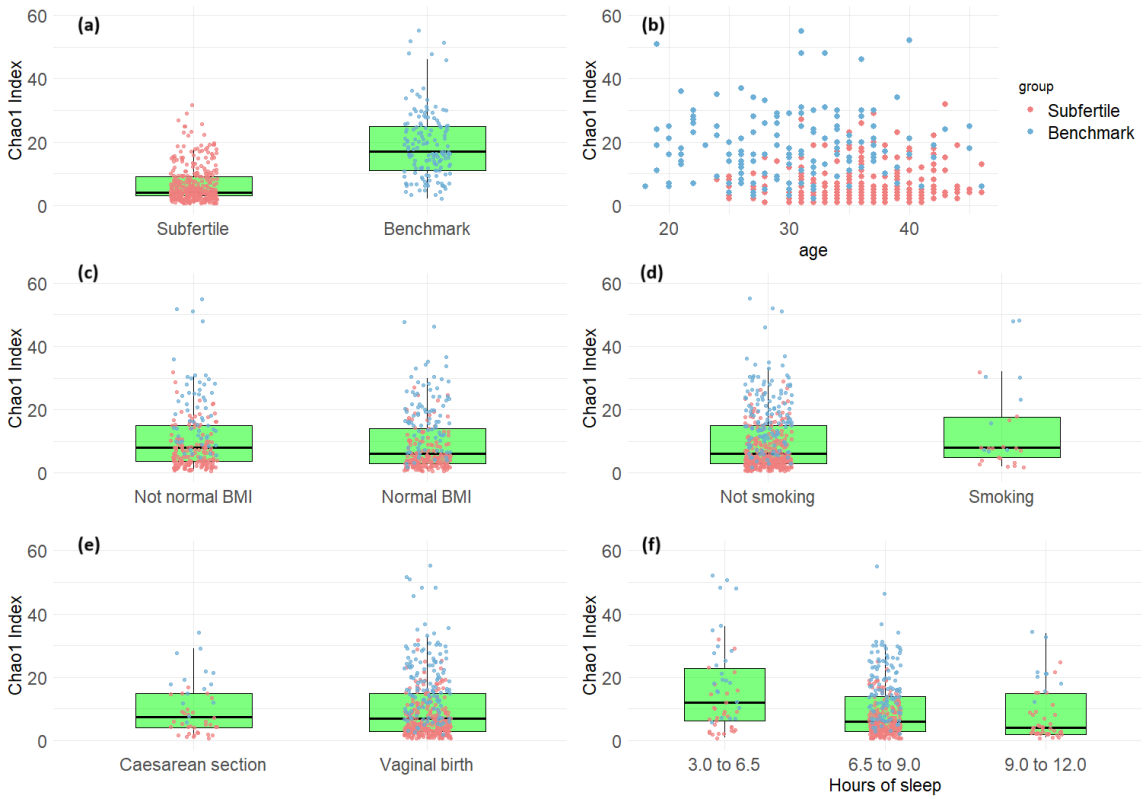



Figure 12: Plots of Chao1 index by respondents' profile

The five panels show the boxplot of Chao1 index (a) between subfertile and benchmark group; (c) by BMI category; (d) by smoking status; (e) by method of birth; and (f) by number of hours spent sleeping. Panel (b) is a scatterplot of samples by Chao1 index and age. Dots in all panels represent the samples, color coded by group.

Figure 13 shows the plot of Shannon index by respondents' characteristics. The benchmark group (0.59) has slightly higher median Shannon index than the subfertile group (0.19). The low Shannon index in both groups can be attributed to the dominance of *Lactobacillus* species. Women who have less time spent sleeping (3 to 6.5 hours) exhibited a higher Shannon index (0.80) than those with 6.5 to 9 hours (0.35) or 9 to 12 hours of sleep (0.30). The statistical significance of these differences was discussed in Section 4.2.2, while the biological relevance was discussed in Section 5.2.

In contrast, the plots of the Shannon index against age do not exhibit an increasing or decreasing trend with respect to changes in age. Shannon index was also observed to be similar between normal (0.28) and not normal BMI (0.41); smoker (0.67) and non-smoker (0.33); and women born through natural means (0.36) and caesarean section (0.28).

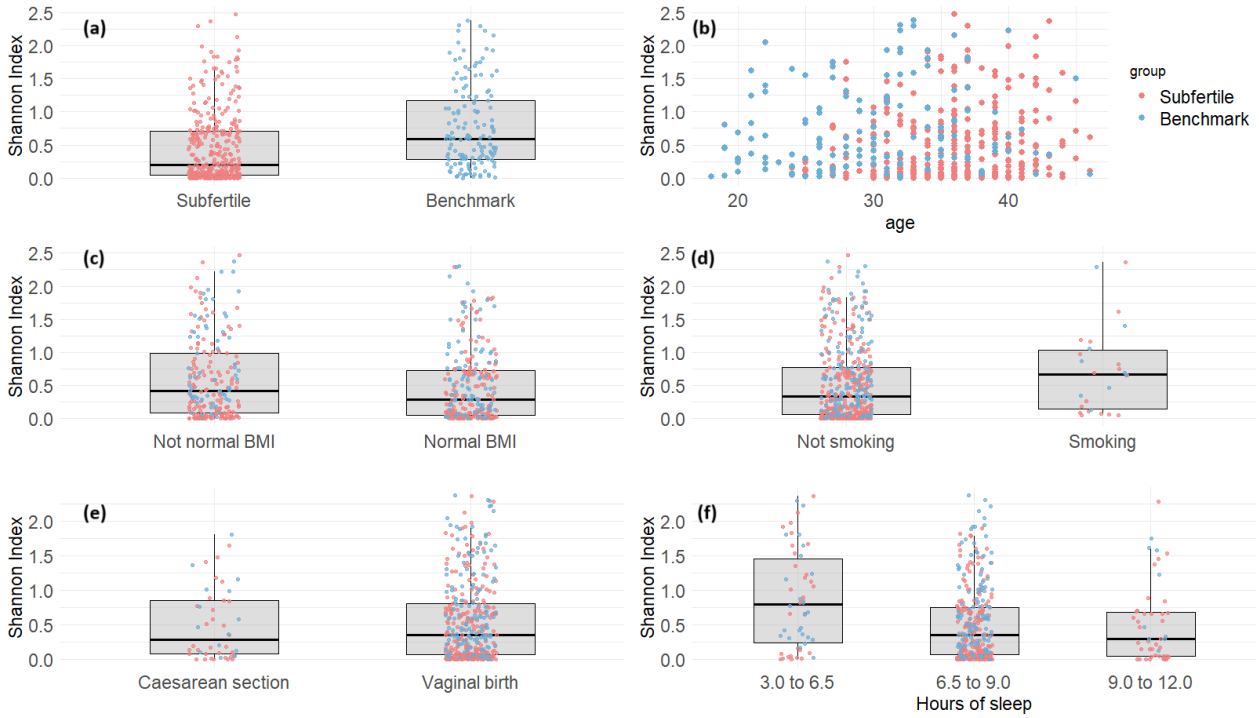



Figure 13: Plots of Shannon index by respondents' profile

The five panels show the boxplot of Shannon index (a) between subfertile and benchmark group; (c) by BMI category; (d) by smoking status; (e) by method of birth; and (f) by number of hours spent sleeping. Panel (b) is a scatterplot of samples by Chao1 index and age. Dots in all panels represent the samples, color coded by group.

### 4.2.2 Model for alpha diversity

To statistically test the difference in alpha diversity between two groups, a linear regression model was initially fitted, as formulated in Equation 6 in Section 3.2 (with interaction). The model diagnostics was performed prior to the interpretation of the regression estimates for the final model (**Appendix D**).

Table 6 presents the parameter estimates from the final model for the Shannon index that include a significant interaction term (i.e., group:sleep) and six main effects. The results showed that there is no significant difference between the Shannon index of the subfertile and benchmark group. The true difference in the Shannon index between these groups could be somewhere between -0.19 and 0.40.

Table 6: Regression estimates from the model for Shannon index

| Parameter | Estimate | Std. error | p-value | 95% confidence interval |
|---|---|---|---|---|
| Intercept | 0.2584 | 0.2145 | 0.2289 | -0.1632, 0.6801 |
| group (benchmark) | 0.1051 | 0.1493 | 0.4819 | -0.1885, 0.3987 |
| age | 0.0149 | 0.0049 | 0.0027 | 0.0052, 0.0247 |
| BMI (normal) | -0.1098 | 0.0513 | 0.0329 | -0.2107, -0.0089 |
| smoking (yes) | 0.0624 | 0.1183 | 0.5981 | -0.1700, 0.2948 |
| sleep (6.5 to 9h) | -0.4313 | 0.1057 | <0.0001 | -0.6391, -0.2236 |
| sleep (9 to 12h) | -0.4176 | 0.1319 | 0.0017 | -0.6767, -0.1584 |
| born (natural means) | 0.1038 | 0.0842 | 0.2184 | -0.0617, 0.2694 |
| group (benchmark):sleep (6.5 to 9h) | 0.3143 | 0.1583 | 0.0477 | 0.0032, 0.6254 |
| group (benchmark):sleep (9 to 12h) | 0.5116 | 0.2462 | 0.0383 | 0.0278, 0.9954 |

The results revealed that sleep duration affects each group differently. In the benchmark group, compared to women with less than 6.5 hours of sleep, those sleeping 6.5 to 9 hours showed a Shannon index 0.12 lower on average, while those sleeping 9 to 12 hours had an index 0.09 higher.[2] For the subfertile group, women with 6.5 to 9 hours of sleep have Shannon index lower by 0.43 than those with less than 6.5 hours of sleep, while women with 9 to 12 hours of sleep have an index lower by 0.42.

Age showed a significant positive effect on the Shannon index, with the true effect on alpha diversity ranging between 0.005 and 0.025 for each year of increase in age. BMI also demonstrated a statistically significant effect on the Shannon index, with women of normal BMI showing a lower Shannon index by 0.01 to 0.21. The results also revealed two statistically insignificant factors: smoking and birth method. The effect of smoking on the Shannon index ranged between -0.17 and 0.29, while women born through natural delivery could have a Shannon index lower by 0.06 or higher by 0.27 than those born through caesarean section.

Table 7 presents the parameter estimates from the Chao1 index model. The results showed that the Chao1 index of the benchmark group is significantly higher than that of the subfertile group. The true difference could be somewhere between 12 and 16. This indicates that women who had once initiated a fertility program show higher microbial richness than women who are currently undergoing fertility treatment. Holding other factors constant (i.e., women are of the same group, age, BMI, smoking status, and birth method), women who have 6.5 to 9 hours of sleep have a significantly lower Chao1 index by around 4 than those with less than 6.5 hours of sleep. The results further revealed that the effect of age, BMI, smoking habit, and birth method on Chao1 index are not statistically significant.

[2]The effect of 6.5 to 9 hours of sleep compared to the reference category was estimated as 0.1051+(-0.4313)+0.3143-0.1015 = -0.1170. Meanwhile, the effect of 9 to 12 hours sleep was estimated as 0.1015+(-0.4176)+0.5116-0.1015 = 0.0940

Table 7: Regression estimates from the model for Chao1 index

| Parameter | Estimate | Std. error | p-value | 95% confidence interval |
|---|---|---|---|---|
| Intercept | 5.1640 | 3.3275 | 0.1214 | -1.3756, 11.7036 |
| group (benchmark) | 13.6890 | 0.9869 | <0.0001 | 11.7494, 15.6286 |
| age | 0.1173 | 0.0801 | 0.1442 | -0.0402, 0.2747 |
| BMI (normal) | -1.3456 | 0.8337 | 0.1073 | -2.9840, 0.2929 |
| smoking (yes) | 1.8051 | 1.9216 | 0.3481 | -1.9716, 5.5817 |
| sleep (6.5 to 9h) | -3.8196 | 1.3189 | 0.0039 | -6.4117, -1.2275 |
| sleep (9 to 12h) | -3.3041 | 1.7793 | 0.0640 | -6.8009, 0.1928 |
| born (natural means) | 1.0527 | 1.3663 | 0.4414 | -1.6325, 3.7379 |

#### 4.2.3 Differential abundance analysis

As discussed in Section 3.2, ANCOM-BC was first implemented using the merged dataset of the Isala and FLORA projects, adjusting for the group variable (1=benchmark, 0=subfertile) and confounders (age, BMI, smoking). This was followed by ANCOM-BC performed separately for the subfertile and benchmark group, adjusting for eight covariates.

Figure 14 illustrates the log-fold change of differentially abundant taxa, with only the group variable and confounders in the model. The results revealed that 11 taxa are more abundant in the benchmark group than in the subfertile group. These taxa include *Sarcina*, *Finegoldia*, and *Anaerococcus*, among others. In contrast, 12 taxa, which include *Lactobacillus gasseri*, *Lactobacillus iners*, *Lactobacillus crispatus*, *Fannyhessea*, and *Ureaplasma*, have lower abundance in the benchmark group than in the subfertile group.

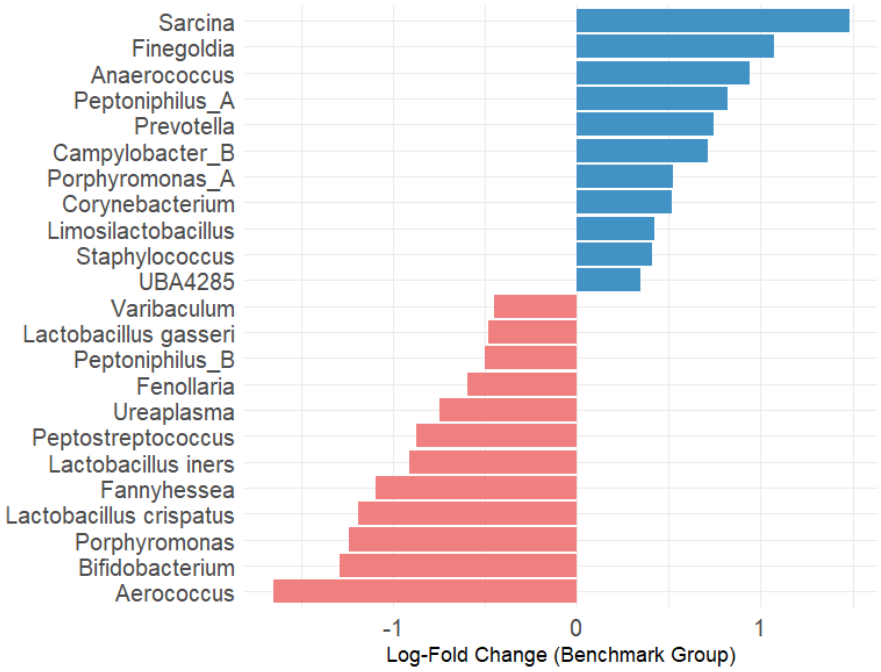



Figure 14: Log-Fold Change of Differentially Abundant Taxa, adjusting for group and confounders
This figure shows the log-fold change for each differentially abundant taxa, adjusted for group, age, BMI, and smoking habit. The 23 taxa are significant at the 5% FDR level and passed the ANCOM-BC sensitivity analysis for pseudo-count addition. Blue bars indicate higher abundance of a taxon in the healthy group, while red bars indicate higher abundance in the subfertile group.

Figure 15 presents the log-fold change of the 30 differentially abundant taxa for the benchmark group, adjusting for the eight covariates. These taxa include *Lactobacillus* species that are known to be pivotal in maintaining a healthy vaginal environment. The results revealed that the abundance of *Lactobacillus crispatus* is negatively associated with age. The abundance of *Lactobacillus gasseri* is expected to be higher for women with normal BMI than for those who are underweight, overweight or obese. It is also positively associated with the frequency of drinking sweet beverages. In contrast, it is lower for women who were born through vaginal birth than those born through caesarean section. The abundance of *Lactobacillus iners* is higher for women born through vaginal birth than those born through caesarean section. The abundance of *Lactobacillus jensenii* is positively associated with the consumption of fermented food, but negatively associated with alcohol intake. It is expected to be lower for women with normal BMI and those born through natural means. In contrast, it is higher for women with 9 to 12 hours of sleep than for those with less than 6.5 hours of sleep.

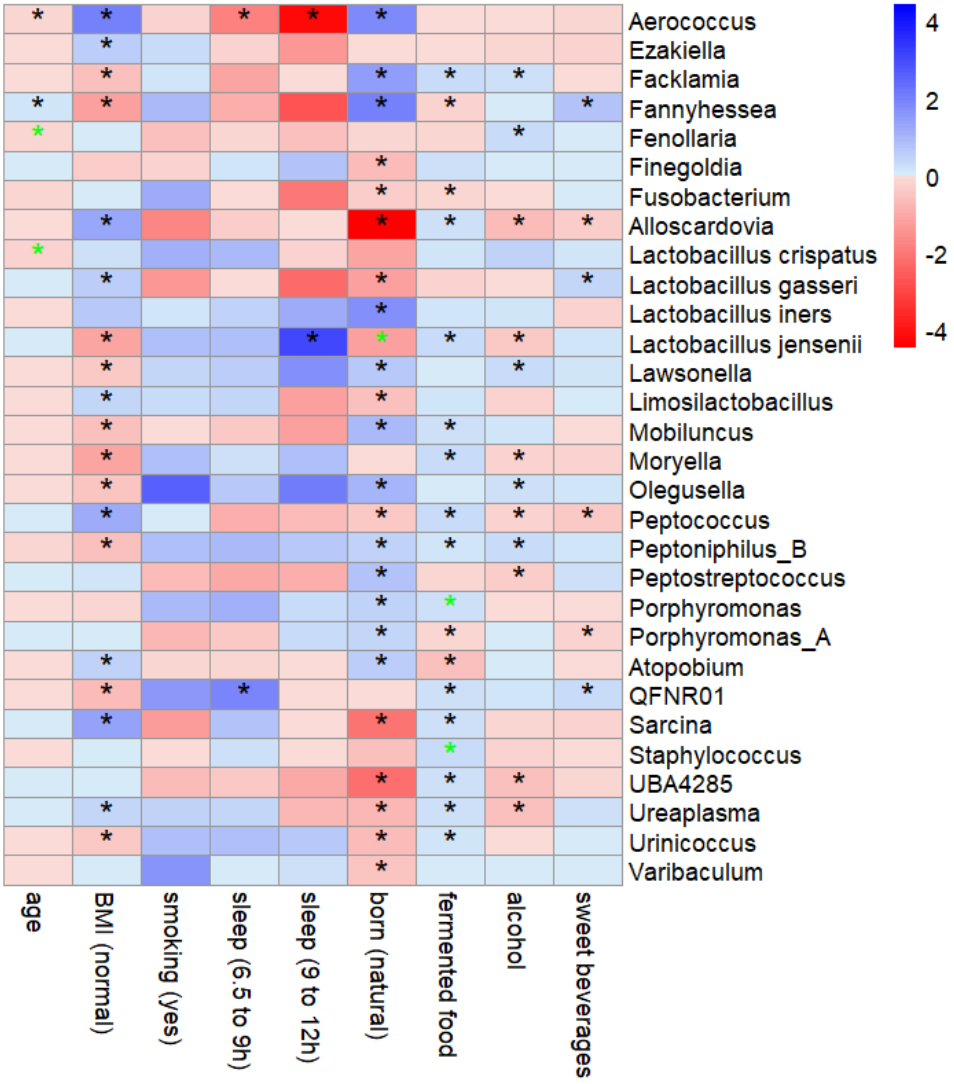



Figure 15: Log Fold Change of the Differentially Abundant Taxa for the Benchmark Group
The figure presents the differentially abundant taxa, adjusted for age, BMI, smoking habit, hours of sleep, frequency of consuming fermented food, alcohol intake, and drinking sweet beverages. Red cells represent a log-fold change between 0 and -4, suggesting lower taxon abundance. Blue cells represent a log-fold change between 0 and 4, suggesting higher taxon abundance. Cells with asterisk indicate significance at 5% FDR level. Asterisks in green indicate that taxa have successfully passed the ANCOM-BC2 sensitivity analysis for pseudo-count addition. While this analysis is useful for assessing robustness, it is known to be conservative and may exclude biologically relevant signals. Therefore, all differentially abundant taxa, regardless of passing/failing the sensitivity analysis, were presented.

The results further revealed that the abundance of *Fenollaria* is negatively associated with age, but positively linked to alcohol intake. The abundance of *Porphyromonas* is positively associated with frequency of eating fermented food. It is also expected to be higher for women born through natural means. The abundance of *Staphylococcus* is positively associated with the frequency of eating fermented food. The abundance of *Aerococcus* is nega-

tively associated with age. It is expected to be higher for women with normal BMI and for those born through natural means, but lower for women with more than 6.5 hours of sleep. The abundance of *Ezakiella*, *Alloscardovia*, Limosilactobacillus, *Peptococcus*, *Atopobium*, *Sarcina*, and *Ureaplasma* is expected to be higher for women with normal BMI. In contrast, the abundance of *Facklamia*, *Fannyhessea*, *Lawsonella*, *Mobiluncus*, *Moryella*, *Olegusella*, *Peptoniphilus B*, *QFNR*01, and *Urinicoccus* is lower for women with normal BMI. The abundance of *Facklamia*, *Fannyhessea*, and eight other taxa is higher for women born through natural means. In contrast, the abundance of *Alloscardovia*, *Sarcina*, *UBA*4285, and seven other taxa is lower for women born through natural means. The abundance of *Facklamia*, *Alloscardovia*, and nine other taxa is positively associated with the frequency of eating fermented food, while this factor is negatively associated with the abundance of *Fannyhessea*, *Fusobacterium*, *Porphyromonas A*, and *Atopobium*. The abundance of *Facklamia*, *Lawsonella*, *Olegusella*, and *Peptoniphilus B* is positively associated with alcohol intake. In contrast, the abundance of *Alloscardovia*, *Moryella*, *Peptococcus*, *Peptostreptococcus*, *UBA*4285, and *Ureaplasma* is negatively associated with alcohol intake. The abundance of *Fannyhessea* and *QFNR*01 has a positive association with the consumption of sweet beverages, but the abundance of *Allocardovia*, *Peptococcus*, and *Porphyromonas A* is negatively associated with this habit.

The comparison of the effect of external factors between the subfertile and the benchmark group was further discussed in Section 5.2.

# 5 Discussion

## 5.1 Comparison between subfertile and healthy group

The primary objective of this study is to compare the alpha diversity and taxon abundance between healthy and subfertile women. This comparison aimed to identify microbial patterns that may be associated with reproductive health and infertility.

The results revealed a low Shannon index for both groups. This can be attributed to the dominance of *Lactobacillus* species. The Shannon index is estimated to be slightly higher by 0.30 to 0.48 for healthy women than subfertile women. While statistically significant, this may not indicate a clinically relevant or abnormal shift in alpha diversity, specifically in terms of species evenness. In fact, previous studies reported that the average Shannon index they observed for healthy women (i.e., with no fertility issues or bacterial vaginosis) is 0.67 $\pm$0.59 standard deviation (Gottschick et al., 2017) and 0.80 $\pm$0.40 standard deviation (Ichiyama et al., 2021). While Shannon index slightly varies, Chao1 index revealed a more apparent difference between the two groups. Taking into account rare taxa, the Chao1 index is significantly higher by 14 to 17 for the healthy group than for the subfertile group. This suggests that the healthy group has a higher microbial richness, i.e., a greater number of genera present in their vaginal tract, compared to the subfertile group. However, it should be noted that a higher number of genera does not necessarily imply better vaginal health and fertility status, especially if beneficial microbiota are less abundant than other, potentially less favorable, taxa.

Holding other factors constant (i.e., women are of the same group, age, BMI, smoking habit, and birth method), Shannon index is lower for women with more than 6.5 hours of sleep. The difference in the Shannon index from those who sleep less than 6.5 hours is approximately between 0.14 and 0.51. While this is statistically significant, the estimated difference may also not be indicative of a biologically relevant change in alpha diversity level, specifically species evenness. The effect of sleep was found to be statistically insignificant for the Chao1 index. Previous studies have found that gut microbiome diversity is positively associated with hours of sleep (Smith et al., 2019, Li et al., 2020). However, there have been limited or perhaps no existing reports establishing a clear link between vaginal microbiome diversity and hours of sleep — an area that can be further studied.

Women with normal BMI in both groups were found to have a significantly lower Shannon index and Chao1 index. The true difference in Shannon index from those with non-normal BMI could range from 0.01 to 0.18. The true difference in Chao1 index could lie somewhere between 2 to 4 genera. While this is statistically significant, the estimated difference is small to suggest biological relevance or a significant shift in alpha diversity. Meanwhile, the birth method does not have a significant effect on the Shannon index and the Chao1

index. Similarly, the effect of age and smoking on Shannon index was found to be insignificant. This finding is not the same as the existing study (Lebeer et al., 2022), which can be attributed to the presence of a significant grouping variable (1=healthy, 0=subfertile) in the model and a small proportion of smokers in the dataset.

Previous studies emphasized that *Lactobacillus* species play an important role in maintaining a healthy vaginal environment. However, variations in the abundance of these species may not necessarily imply a direct link to subfertility. In this study, *Lactobacillus jensenii* was found to be more abundant in the healthy group, while *Lactobacillus crispatus* and *Lactobacillus gasseri* were more abundant in the subfertile group. These findings suggest that different *Lactobacillus* species may be differentially associated with reproductive status, highlighting the complexity of microbial influences on fertility. Among other taxa, $Fannyhessea$, $Mobiluncus$, and $Ureaplasma$ were found to be more abundant in the subfertile group. Previous studies have associated these genera with bacterial vaginosis, which is a condition marked by unusually high bacterial diversity and a reduction in typical *Lactobacillus* species (Margolis and Fredricks, 2015, P. Liu et al., 2023, Mendling et al., 2019). Similarly, *Aerococcus* and *Bifidobacterium* were also higher in subfertile women, which is consistent with the findings of Zhao et al. (2020). The role of other species in vaginal health and fertility remains unclear, and it is difficult to conclude that the abundance of specific species is the cause of infertility.

Results further revealed that the association of specific taxa with age may vary for the subfertile group and for the healthy group. Adjusted for BMI, smoking, birth method, and dietary habits, the abundance of *Lactobacillus iners* decreases with age among subfertile women, but not among healthy women. In contrast, the abundance of *Lactobacillus jensenii Actinotignum*, $Parvimonas$, $Peptococcus$, and $UBA4285$ increases with age among healthy women, but not among subfertile women. The abundance of $Fannyhessea$, which has been linked to bacterial vaginosis, increases with age among subfertile women, but not among healthy women. While age has been known to affect fertility among women (e.g., due to reduced ovarian reserve and hormonal shifts), the findings of this study suggest that aging in subfertile women may be accompanied by microbiome alterations that further compromise reproductive potential.

Some taxa showed similar associations with lifestyle and dietary habits in both groups, while others differed. The abundance of *Lactobacillus gasseri* and *Lactobacillus jensenii* was found to be higher for women with normal BMI in the healthy group than for those who are underweight, overweight or obese. The abundance of *Lactobacillus crispatus* and *Lactobacillus gasseri* was also higher for women with normal BMI in the subfertile group. The higher abundance of these *Lactobacillus* species in women with normal BMI may suggest that healthy body weight supports a more favorable vaginal microbiome. In the

subfertile group, abundance of *Lactobacillus crispatus* and *Lactobacillus iners* is expected to be higher for women with more than 6.5 hours of sleep than for those with less than 6.5 hours of sleep. In the healthy group, higher abundance of *Lactobacillus iners* and lower abundance of *Lactobacillus jensenii* is expected for women with more than 6.5 hours of sleep. These findings suggest that while the association of sleep duration with microbial composition differs between groups, lack of sleep could be a risk factor that can be linked to vaginal health and fertility. In terms of dietary habits, consumption of fermented food is negatively associated with *Fannyhessea* among the healthy group, but positive association among the subfertile group. Consumption of sweet beverages has a negative association with *Lactobacillus crispatus* among healthy women, but not among subfertile women. A consistent negative association was found between alcohol consumption and levels of *Bifidobacterium* and *Lactobacillus gasseri* for both groups. This suggests that alcohol intake could also alter the vaginal environment, noting also how detrimental it is to the gut microbiome (Lee and Lee, 2021). None of the taxa were found to have a significant association with smoking, except for *Alloscardovia* in the healthy group. The statistical insignificance of this factor may be attributed to the small proportion of smokers (5%) in the dataset, which could limit the ability to detect meaningful associations.

Consistent in both groups, *Lactobacillus gasseri* is expected to be less abundant in women born through natural means than in women delivered through caesarean section. The abundance of three other taxa in the subfertile group and nine other taxa in the healthy group is also lower for these women. This result should be interpreted with caution as caesarean section at birth has been associated with bacterial vaginosis in adulthood (Stennett et al., 2020).

## 5.2 Comparison between subfertile and benchmark group

The secondary objective of this study was to compare subfertile women who are currently undergoing fertility treatment (from the FLORA project, referred to in this report as "subfertile" group) with women who had once initiated a fertility program at any point in their lives (from the Isala project, referred to as "benchmark" group). Women in these groups both have fertility issues and are undergoing/underwent fertility treatment, which might have affected their microbiome composition.

The results showed that the two groups do not differ in Shannon index, but the Chao1 index for the benchmark group is significantly higher than the subfertile group. This difference cannot be further explained due to the absence of other pertinent health information about the benchmark group, such as the time since they received the treatment or complexity of their health condition. There could be an alteration in their vaginal microbiome composition over time, which cannot be attributed to fertility treatment. To reiterate, a higher

number of genera does not necessarily imply better vaginal health and fertility status, especially if beneficial microbiota are less abundant than other, potentially less favorable, taxa. A significant interaction effect on Shannon index was found between the group and hours of sleep, indicating that sleep duration affects each group differently. However, the true mean difference in Shannon index between women with less than 6.5 hours of sleep and women with more than 6.5 hours of sleep is estimated to be less than 0.5, which may not indicate relevant shift in microbial evenness. Holding other factors constant (i.e., women are of same group, age, BMI, smoking status, and birth method), women who have 6.5 to 9 hours of sleep also have a lower Chao1 index, i.e., by around four genera, than those with less than 6.5 hours of sleep. Age and BMI were also found to have a significant effect on the Shannon index. However, an increase of 0.01 for each year of increase in age and a reduction of 0.11 for normal BMI may not suggest a relevant shift in alpha diversity, particularly in terms of microbial evenness. Meanwhile, the effect of age, BMI, smoking, and birth method on Chao1 index was found to be insignificant.

Compared to the benchmark group, the subfertile group shows higher abundance of *Lactobacillus gasseri*, *Lactobacillus crispatus*, *Lactobacillus iners*. A clear link of these findings with fertility treatment is difficult to establish due to the lack of detailed health profiles of women in the benchmark group. It is possible that women in the benchmark group represent more complex or advanced cases of infertility, which could impact the composition of beneficial vaginal microbiota. However, this cannot be confirmed within the scope of this study.

The results also revealed that the relationship between beneficial vaginal microbiota and external factors is not consistent across groups. In the subfertile group, the abundance of *Lactobacillus crispatus* is associated with BMI and hours of sleep, while in the benchmark group, it is linked only to age. For *Lactobacillus gasseri*, a higher abundance is consistently observed in women with normal BMI across groups. In contrast, *Lactobacillus iners* shows lower abundance in women with normal BMI and higher abundance in those with more than 6.5 hours of sleep in the subfertile group; but these associations are not observed in the benchmark group. Additionally, *Lactobacillus iners* is more abundant in benchmark women born via vaginal delivery, an association not seen in the subfertile group. *Lactobacillus jensenii* demonstrates significant associations with BMI, hours of sleep, birth method, alcohol intake, and fermented food consumption in the benchmark group, but not in the subfertile group. These differences suggest that fertility treatment history may influence how external factors affect the microbiome composition. These findings raise the possibility that the vaginal microbiome's responsiveness to lifestyle or physiological influences changes along the fertility care continuum. Understanding these group-specific patterns could help identify critical windows for microbiome-targeted interventions and inform more personalized strategies to support reproductive health.

## 5.3 Possible drawbacks

A key limitation of this study is the restricted availability of variables measured consistently in both the Isala and FLORA projects. While other external factors might have been relevant to include in the model, the analysis could only use variables that were comparable between the two datasets. There is also a limitation in the comparison between the subfertile and benchmark group due to the unavailability of pertinent health information about women who had once initiated a fertility program. This information includes the specific reproductive health condition, kind of fertility treatment, and the duration of fertility program, among others.

Another possible drawback is the effectiveness of the exclusion criteria applied for the Isala dataset. As discussed in Section 2, the healthy group was identified by excluding individuals who reported reproductive health conditions (such as PCOS and endometriosis), diabetes, and hematologic disorders, among others. However, unreported fertility-related conditions may result in some overlap between the healthy and subfertile cohorts, meaning the groups are not necessarily mutually exclusive or perfectly defined.

As also discussed in Section 2, vaginal swabs in the Isala project were obtained through home self-sampling, while swabs in the FLORA project were collected in a clinical setting. This difference in collection method may introduce potential technical biases (e.g., contamination) and biological biases (e.g., unreported reproductive conditions or behavioral differences) that could affect estimates of microbiome diversity and abundance. This study acknowledges these limitations and addressed them through standard data filtering and further adjustments in the modeling of the diversity and abundance to account for the group effect and other potential confounders.

Despite these limitations, it is important to note that both the Isala and FLORA projects contributed valuable data on the vaginal microbiome and demographic profiles of women. This enabled meaningful comparative analyses of microbiome composition and its associations with external factors.

## 5.4 Future research

Researchers who are interested in characterizing healthy and subfertile women can consider formulating a study design in which microbiome data and patient's profiles are collected and measured in a similar manner for both groups. This will allow them to explore relevant external factors influencing the microbiome composition of both groups. Possible area for future study include (i) the link between hours of sleep and vaginal microbiome diversity, mentioned in Section 5.1; (ii) characterization of the role of specific taxa in vaginal health and fertility; and (iii) impact of microbiome-targeted interventions on the abundance of

*Lactobacillus* species and other beneficial microbiota.

# 6 Conclusion

This study characterized healthy women as having significantly higher microbial richness than subfertile women. However, this finding should be interpreted with caution, as a higher number of genera does not necessarily imply better vaginal health and fertility status, especially if beneficial microbiota are less abundant than other, potentially less favorable, taxa. In terms of microbiome composition, the findings suggest that different *Lactobacillus* species may be differentially associated with reproductive status, highlighting the complexity of microbial influences on fertility. The results also suggest a significant association between the abundance of at least one *Lactobacillus* species and external factors, namely age, BMI, hours of sleep, and alcohol intake, in subfertile women. These factors may be considered when designing targeted interventions or personalized fertility treatments. There may also be other important factors influencing the microbiome composition that were beyond the scope of this study. In addition to *Lactobacillus* species, several other taxa were found to be differentially abundant between the subfertile and healthy groups, though their roles in vaginal health and fertility are not yet clearly established.

This study acknowledges certain limitations, including the absence of potentially important variables that could influence microbiome diversity and taxon abundance. Moreover, the primary aim was not to identify specific biomarkers of infertility, but to explore the associations between fertility status, external factors, and vaginal microbiome composition. Future research should address these limitations and aim to identify potential microbiome-based biomarkers of infertility through more comprehensive and targeted analyses.

## References


Bayoumi, R. R., Hurt, L., Zhang, N., Law, Y. J., Venetis, C., Fatem, H. M., Serour, G. I., van der Poel, S., & Boivin, J. (2024). A critical systematic review and meta-analyses of risk factors for fertility problems in a globalized world. *Reproductive BioMedicine Online*, *48*(3), 103217. https://doi.org/https://doi.org/10.1016/j.rbmo.2023.04.008

Bradshaw, C. S., & Sobel, J. D. (2016). Current treatment of bacterial vaginosis—limitations and need for innovation. *The Journal of Infectious Diseases*, *214*(suppl$_1$), S14–S20.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 783–791.

Chen, C., Song, X., Wei, W., Zhong, H., Dai, J., Lan, Z., Li, F., Yu, X., Feng, Q., Wang, Z., et al. (2017). The microbiota continuum along the female reproductive tract and its relation to uterine-related diseases. *Nature communications*, *8*(1), 875.

Darıcı, E., Pais, F., Leemans, L., Strypstein, L., Tournaye, H., De Vos, M., De Waele, E., & Blockeel, C. (2025). Body composition in female infertility from body mass index to body composition in female infertility. *Reproductive BioMedicine Online*, 104941. https://doi.org/https://doi.org/10.1016/j.rbmo.2025.104941

Del Campo-Moreno, R., Alarcón-Cavero, T., D'Auria, G., Delgado-Palacio, S., & Ferrer-Martínez, M. (2018). Microbiota and human health: Characterization techniques and transference. *Enfermedades Infecciosas y Microbiología Clínica*, *36*, 241–245. https://doi.org/10.1016/j.eimce.2018.02.016

Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., Huang, B., Arodz, T. J., Edupuganti, L., Glascock, A. L., et al. (2019). The vaginal microbiome and preterm birth. *Nature medicine*, *25*(6), 1012–1021.

Gottschick, C., Deng, Z.-L., Vital, M., Masur, C., Abels, C., Pieper, D., Rohde, M., Mendling, W., & Wagner-Döbler, I. (2017). Treatment of biofilms in bacterial vaginosis by an amphoteric tenside pessary-clinical study and microbiota analysis. *Microbiome*, *5*, 119. https://doi.org/10.1186/s40168-017-0326-y

Haque, M. M., Merchant, M., Kumar, P. N., Dutta, A., & Mande, S. S. (2017). First-trimester vaginal microbiome diversity: A potential indicator of preterm delivery risk. *Scientific reports*, *7*(1), 16145.

Ichiyama, T., Kuroda, K., Nagai, Y., Urushiyama, D., Ohno, M., Yamaguchi, T., Nagayoshi, M., Sakuraba, Y., Yamasaki, F., Hata, K., Miyamoto, S., Itakura, A., Takeda, S., & Tanaka, A. (2021). Analysis of vaginal and endometrial microbiota communities in infertile women with a history of repeated implantation failure [Published under CC BY 4.0 license]. *Reproductive Medicine and Biology*, *20*(1), 51–58. https://doi.org/10.1002/rmb2.12389

Koedooder, R., Singer, M., Schoenmakers, S., Savelkoul, P. H. M., Morré, S. A., de Jonge, J. D., Poort, L., Cuypers, W. J. S. S., Beckers, N. G. M., Broekmans, F. J. M., Cohlen, B. J., den Hartog, J. E., Fleischer, K., Lambalk, C. B., Smeenk, J. M. J. S.,

Budding, A. E., & Laven, J. S. E. (2019). The vaginal microbiome as a predictor for outcome of in vitro fertilization with or without intracytoplasmic sperm injection: A prospective study. *Human Reproduction*, *34*(6), 1042–1054. https://doi.org/10.1093/humrep/dez065

Kolde, R. (2019). *Pheatmap: Pretty heatmaps* [R package version 1.0.12]. https://CRAN.R-project.org/package=pheatmap

Lebeer, S., Ahannach, S., Wittouck, S., Gehrmann, T., Eilers, T., Oerlemans, E., Condori, S., Dillen, J., Spacova, I., Vander Donck, L., Masquiller, C., Bron, P., Van Beeck, W., Backer, C., Donders, G., & Verhoeven, V. (2022). Citizen-science map of the vaginal microbiome. https://doi.org/10.21203/rs.3.rs-1350465/v1

Lee, E., & Lee, J.-E. (2021). Impact of drinking alcohol on gut microbiota: Recent perspectives on ethanol and alcoholic beverage. *Current Opinion in Food Science*, *37*, 91–97. https://doi.org/https://doi.org/10.1016/j.cofs.2020.10.001

Lehtoranta, L., Ala-Jaakkola, R., Laitila, A., & Maukonen, J. (2022). Healthy vaginal microbiota and influence of probiotics across the female life span. *Frontiers in Microbiology*, *13*, 819958. https://doi.org/10.3389/fmicb.2022.819958

Li, Y., Zhang, B., Zhou, Y., Wang, D., Liu, X., Li, L., Wang, T., Zhang, Y., Jiang, M., Tang, H., Amsel, L. V., Fan, F., & and, C. W. H. (2020). Gut microbiota changes and their relationship with inflammation in patients with acute and chronic insomnia. *Nature and Science of Sleep*, *12*, 895–905. https://doi.org/10.2147/NSS.S271927

Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature Communications*, *11*(1), 3514. https://doi.org/10.1038/s41467-020-17041-7

Lin, H., & Peddada, S. D. (2024). Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. *Nature Methods*, *21*(1), 83–91. https://www.nature.com/articles/s41592-023-02092-7

Liu, F., Zhou, Y., Zhu, L., Wang, Z., Ma, L., He, Y., & Fu, P. (2021). Comparative metagenomic analysis of the vaginal microbiome in healthy women. *Synthetic and Systems Biotechnology*, *6*(2), 77–84. https://doi.org/https://doi.org/10.1016/j.synbio.2021.04.002

Liu, P., Wang, L., Li, R., & Chen, X. (2023). A rare bacteremia caused by fannyhessea vaginae in a pregnant woman: Case report and literature review. *Frontiers in Cellular and Infection Microbiology*, *13*. https://doi.org/10.3389/fcimb.2023.1278921

Margolis, E., & Fredricks, D. N. (2015). Chapter 83 - bacterial vaginosis-associated bacteria. In Y.-W. Tang, M. Sussman, D. Liu, I. Poxton, & J. Schwartzman (Eds.), *Molecular medical microbiology (second edition)* (Second Edition, pp. 1487–1496). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-397169-2.00083-4

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4), e61217. http://dx.plos.org/10.1371/journal.pone.0061217

Mendling, W., Palmeira-de-Oliveira, A., Biber, S., & Prasauskas, V. (2019). An update on the role of atopobium vaginae in bacterial vaginosis: What to consider when choosing a treatment? a mini review. *Archives of Gynecology and Obstetrics*, *300*(1), 1–6. https://doi.org/10.1007/s00404-019-05142-8

Moreno, I., Garcia-Grau, I., & Simón, C. (2018). Microbiota and pathogen screening in the female reproductive tract. In M. K. Skinner (Ed.), *Encyclopedia of reproduction (second edition)* (Second Edition, pp. 36–44). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-801238-3.64730-X

Morsli, M., Gimenez, E., Magnan, C., Salipante, F., Huberlant, S., Letouzey, V., & Lavigne, J.-P. (2024). The association between lifestyle factors and the composition of the vaginal microbiota: A review. *European Journal of Clinical Microbiology & Infectious Diseases*, *43*(10), 1869–1881. https://doi.org/10.1007/s10096-024-04915-7

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., ... Borman, T. (2025). *Vegan: Community ecology package* [R package version 2.6-10]. https://CRAN.R-project.org/package=vegan

Pagar, R., Deshkar, S., Mahore, J., Patole, V., Deshpande, H., Gandham, N., Mirza, S., Junnarkar, M., & Nawani, N. (2024). The microbial revolution: Unveiling the benefits of vaginal probiotics and prebiotics. *Microbiological Research*, *286*, 127787. https://doi.org/https://doi.org/10.1016/j.micres.2024.127787

Petrova, M. I., Lievens, E., Malik, S., Imholz, N., & Lebeer, S. (2015). Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. *Frontiers in physiology*, *6*, 81.

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Shipitsyna, E., Roos, S., Datcu, R., Hallen, A., Fredlund, H., Jensen, J. S., Engstrand, L., & Unemo, M. (2013). Composition of the vaginal microbiota in women of reproductive age–sensitive and specific molecular diagnosis of bacterial vaginosis is possible? *PloS One*, *8*(4), e60670. https://doi.org/10.1371/journal.pone.0060670

Smith, R. P., Easson, C., Lyle, S. M., Kapoor, R., Donnelly, C. P., Davidson, E. J., Parikh, E., Lopez, J. V., & Tartar, J. L. (2019). Gut microbiome diversity is associated with sleep physiology in humans. *PLoS One*, *14*(10), e0222394. https://doi.org/10.1371/journal.pone.0222394

Son, K.-A., Kim, M., Kim, Y. M., Kim, S. H., Choi, S.-J., Oh, S.-y., Roh, C.-R., & Kim, J.-H. (2018). Prevalence of vaginal microorganisms among pregnant women according to trimester and association with preterm birth. *Obstetrics and Gynecology Science*, *61*(1), 38–47.

Stennett, C. A., Dyer, T. V., He, X., Robinson, C. K., Ravel, J., Ghanem, K. G., & Brotman, R. M. (2020). A cross-sectional pilot study of birth mode and vaginal microbiota

in reproductive-age women. *PLoS One*, *15*(4), e0228574. https://doi.org/10.1371/journal.pone.0228574

Vavrek, M. J. (2011). Fossil: Palaeoecological and palaeogeographical analysis tools [R package version 0.4.0]. *Palaeontologia Electronica*, *14*(1), 1T.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

William Revelle. (2025). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.5.3]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Wittouck, S., Van Rillaer, T., & Smets, W. (2025). *Tidytacos: Manipulate taxonomic composition data of microbial communities* [R package version 1.0.6]. https://github.com/LebeerLab/tidytacos

World Health Organization. (2023). Infertility [Accessed: 2025-06-11]. https://www.who.int/news-room/fact-sheets/detail/infertility

World Health Organization. (2025). Global health observatory (gho) data – indicator metadata registry [Accessed: 2025-05-16]. https://www.who.int/data/gho/data/indicators

World Medical Association. (2013, October). Wma declaration of helsinki – ethical principles for medical research involving human subjects [Accessed: 2025-06-16]. https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/

Zaino, R. J., Robboy, S. J., Bentley, R., & Kurman, R. J. (1994). Diseases of the vagina. In *Blaustein's pathology of the female genital tract* (pp. 131–183). Springer.

Zhao, C., Wei, Z., Yang, J., Zhang, J., Yu, C., Yang, A., Zhang, M., Zhang, L., Wang, Y., Mu, X., Heng, X., Yang, H., Gai, Z., Wang, X., & Zhang, L. (2020). Characterization of the vaginal microbiome in women with infertility and its potential correlation with hormone stimulation during in vitro fertilization surgery. *mSystems*, *5*(4), e00450–20. https://doi.org/10.1128/mSystems.00450-20

# Appendices

## Appendix A List of exclusion criteria for the Isala dataset

| **Description of respondents to be excluded** | **From n=3349, sample size was reduced to** | **No. of excluded respondents** |
|---|---|---|
| 1. Greater than 46 years old | 3111 | 238 |
| 2. Initiated fertility program | 2965 | 146 |
| 3. Suffered from endometriosis or PCOS (polycystic ovary syndrome) | 1770 | 1195[a] |
| 4. Currently breastfeeding | 1707 | 63 |
| 5. Had antibiotic or antimycotic treatment in the past three months | 1347 | 360 |
| 6. With diabetes or hematologic disorder | 1219 | 128 |
| 7. Currently using/used contraception in the last three months to avoid getting pregnant | 390 | 829 |

a - Out of 1195, only 91 and 69 women reported that they had suffered from PCOS and endometriosis, respectively. Women with missing response (n=1042) were not assumed to be free of these conditions and were also excluded. Hence, the remaining samples of size 1770 were the women who directly reported that they had not suffered from PCOS or endometriosis.

# Appendix B Correlation matrix for Isala and FLORA covariates

**FLORA Correlation Matrix**

```
Call:corr.test(x = Flora.8vars[, c("age", "BMI.cat", "smoking", "sleep.cat",
    "born", "bread", "soft_drinks", "spirits")])
Correlation matrix
             age BMI.cat smoking sleep.cat  born bread soft_drinks spirits
age          1.00    0.00    0.02     -0.05 -0.04 -0.05        0.06   -0.08
BMI.cat      0.00    1.00   -0.03      0.11  0.02 -0.07        0.06   -0.05
smoking      0.02   -0.03    1.00     -0.12 -0.03 -0.07        0.02    0.04
sleep.cat   -0.05    0.11   -0.12      1.00  0.09 -0.01       -0.04    0.02
born        -0.04    0.02   -0.03      0.09  1.00 -0.05        0.07    0.06
bread       -0.05   -0.07   -0.07     -0.01 -0.05  1.00        0.12   -0.08
soft_drinks  0.06    0.06    0.02     -0.04  0.07  0.12        1.00   -0.02
spirits     -0.08   -0.05    0.04      0.02  0.06 -0.08       -0.02    1.00

Probability values (Entries above the diagonal are adjusted for multiple tests.)
             age BMI.cat smoking sleep.cat born bread soft_drinks spirits
age         0.00    1.00    1.00      1.00 1.00  1.00        1.00        1
BMI.cat     0.95    0.00    1.00      1.00 1.00  1.00        1.00        1
smoking     0.65    0.55    0.00      1.00 1.00  1.00        1.00        1
sleep.cat   0.41    0.06    0.04      0.00 1.00  1.00        1.00        1
born        0.53    0.77    0.54      0.12 0.00  1.00        1.00        1
bread       0.38    0.20    0.20      0.85 0.39  0.00        0.90        1
soft_drinks 0.28    0.28    0.68      0.52 0.22  0.03        0.00        1
spirits     0.17    0.41    0.51      0.67 0.31  0.15        0.71        0
```

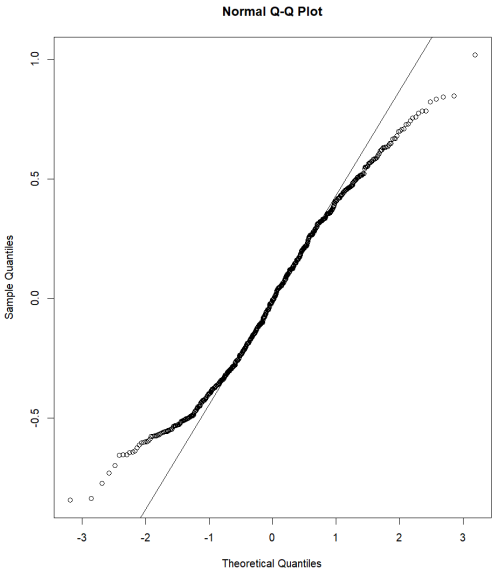
**Isala Correlation Matrix**

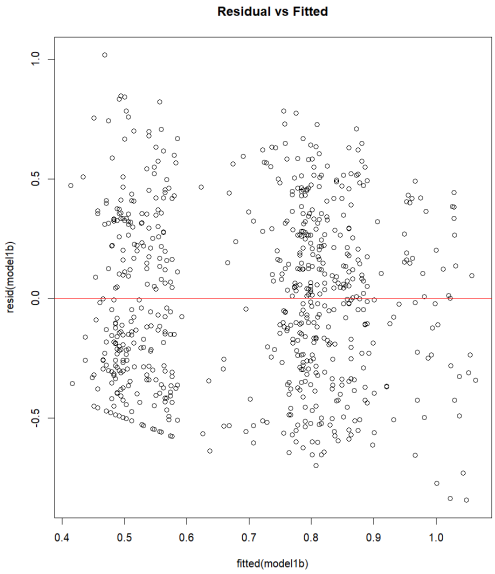
```
Call:corr.test(x = Isala.8vars[, c("age", "BMI.cat", "smoking", "sleep.cat",
    "born", "fermentedfd", "alcohol", "sugarbev")])
Correlation matrix
             age BMI.cat smoking sleep.cat  born fermentedfd alcohol sugarbev
age          1.00   -0.09    0.00     -0.15  0.14       -0.06    0.03    -0.22
BMI.cat     -0.09    1.00    0.01      0.02 -0.08        0.02    0.04     0.02
smoking      0.00    0.01    1.00     -0.04  0.04       -0.02    0.19    -0.06
sleep.cat   -0.15    0.02   -0.04      1.00 -0.10        0.03    0.05     0.00
born         0.14   -0.08    0.04     -0.10  1.00        0.02    0.02    -0.02
fermentedfd -0.06    0.02   -0.02      0.03  0.02        1.00    0.17     0.16
alcohol      0.03    0.04    0.19      0.05  0.02        0.17    1.00     0.04
sugarbev    -0.22    0.02   -0.06      0.00 -0.02        0.16    0.04     1.00

Probability values (Entries above the diagonal are adjusted for multiple tests.)
             age BMI.cat smoking sleep.cat born fermentedfd alcohol sugarbev
age         0.00    1.00    1.00      0.10 0.17           1    1.00     0.00
BMI.cat     0.07    0.00    1.00      1.00 1.00           1    1.00     1.00
smoking     0.98    0.82    0.00      1.00 1.00           1    0.01     1.00
sleep.cat   0.00    0.72    0.40      0.00 0.93           1    1.00     1.00
born        0.01    0.14    0.45      0.04 0.00           1    1.00     1.00
fermentedfd 0.24    0.73    0.73      0.52 0.75           0    0.02     0.04
alcohol     0.56    0.48    0.00      0.29 0.75           0    0.00     1.00
sugarbev    0.00    0.76    0.24      0.95 0.64           0    0.42     0.00
```

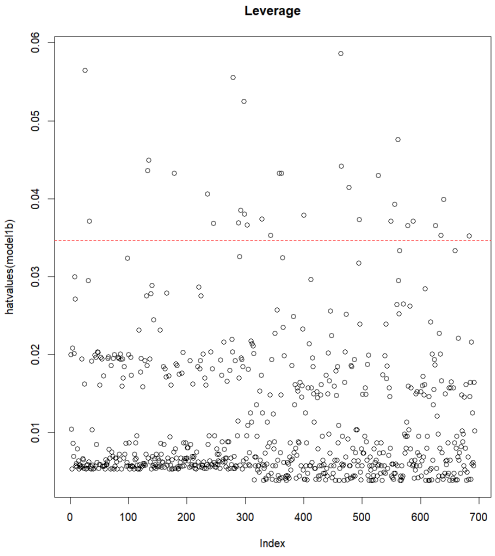
# Appendix C Diagnostics of diversity model in Section 4.1.2

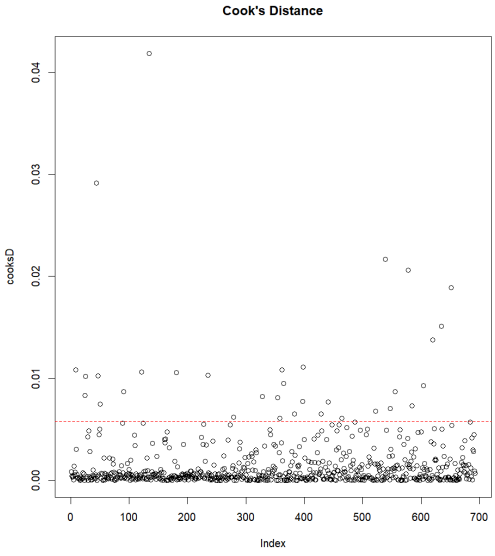
## C.1 Normality and Homoscedasticity



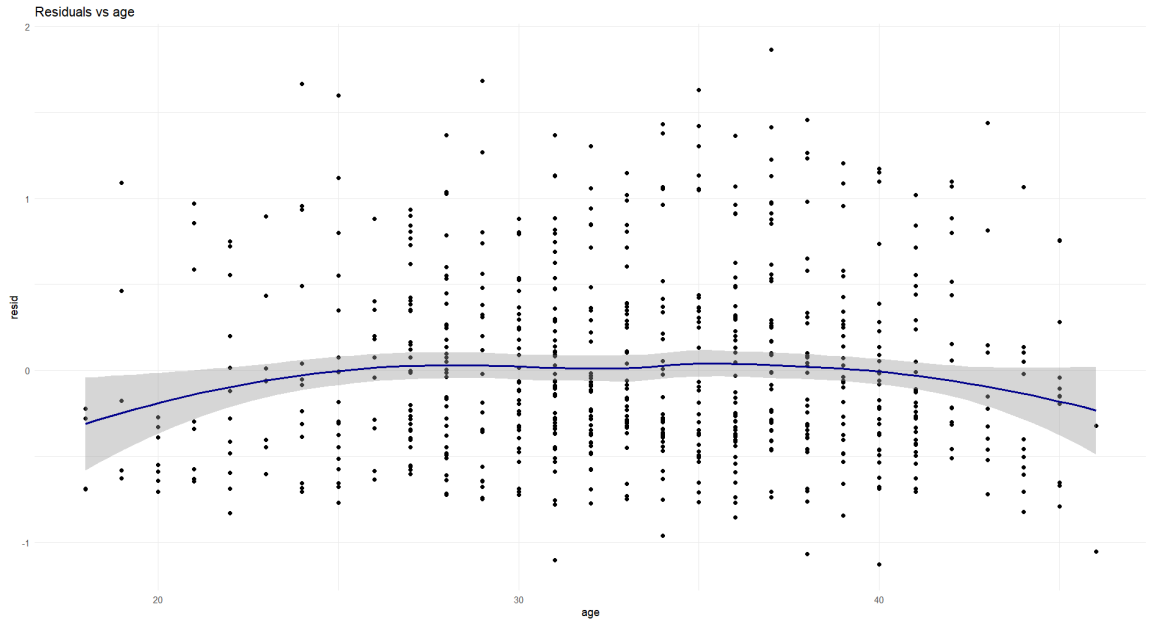




## C.2 Leverage/influential observations



## C.3 Linearity




## C.4 Independence

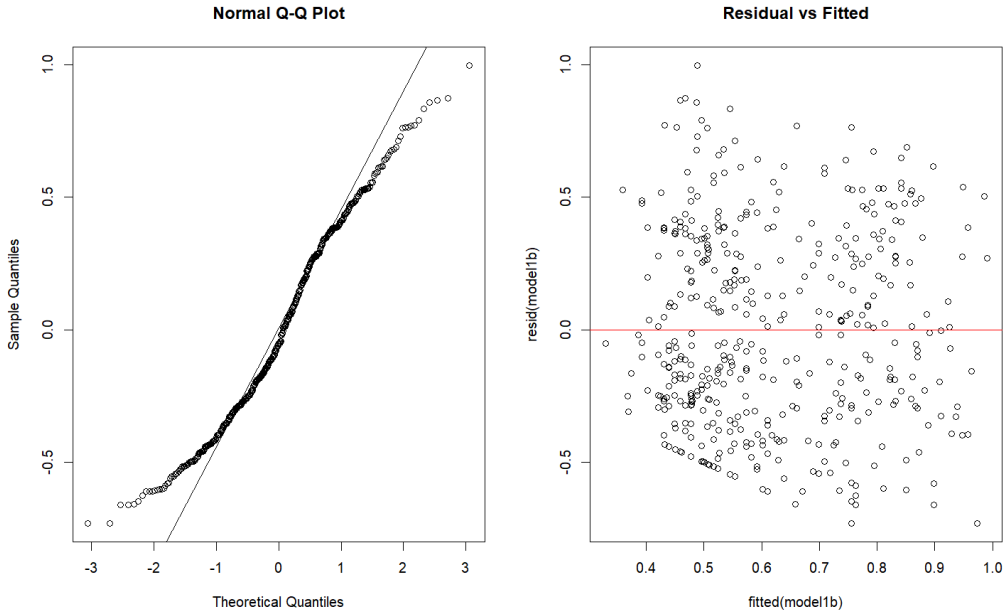
| Test | test statistic | p-value |
|---|---|---|
| Durbin-Watson test | 1.9578 | 0.2750 |

## C.5 Multicollinearity

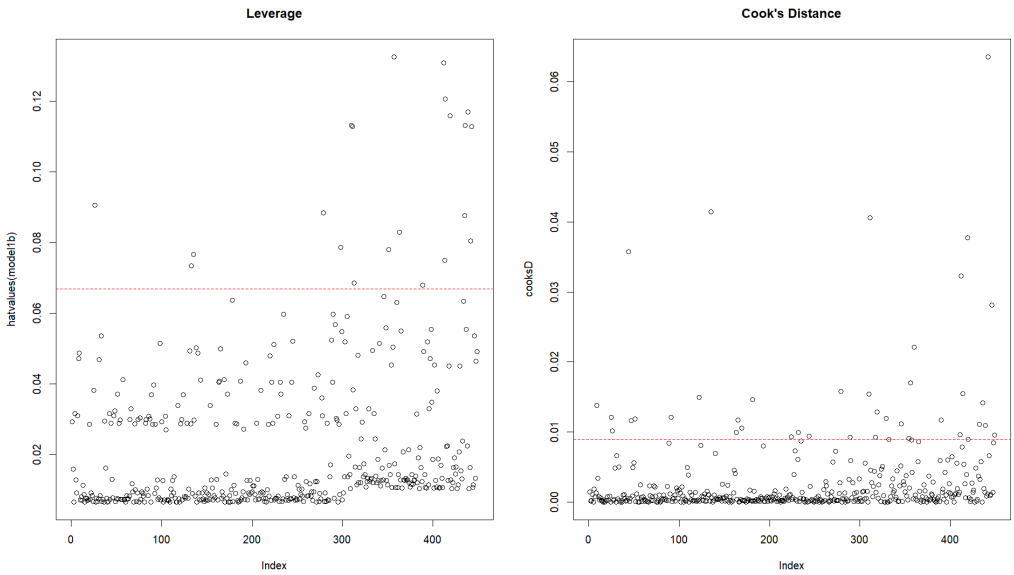
| | GVIF | Df | GVIF(1/(2*Df)) |
|---|---|---|---|
| group | 1.1715 | 1 | 1.0823 |
| age | 1.1569 | 1 | 1.0756 |
| BMI | 1.0258 | 1 | 1.0128 |
| smoking | 1.0083 | 1 | 1.0042 |
| sleep | 1.0518 | 2 | 1.0127 |
| born | 1.0142 | 1 | 1.0071 |

# Appendix D Diagnostics of diversity model in Section 4.2.2

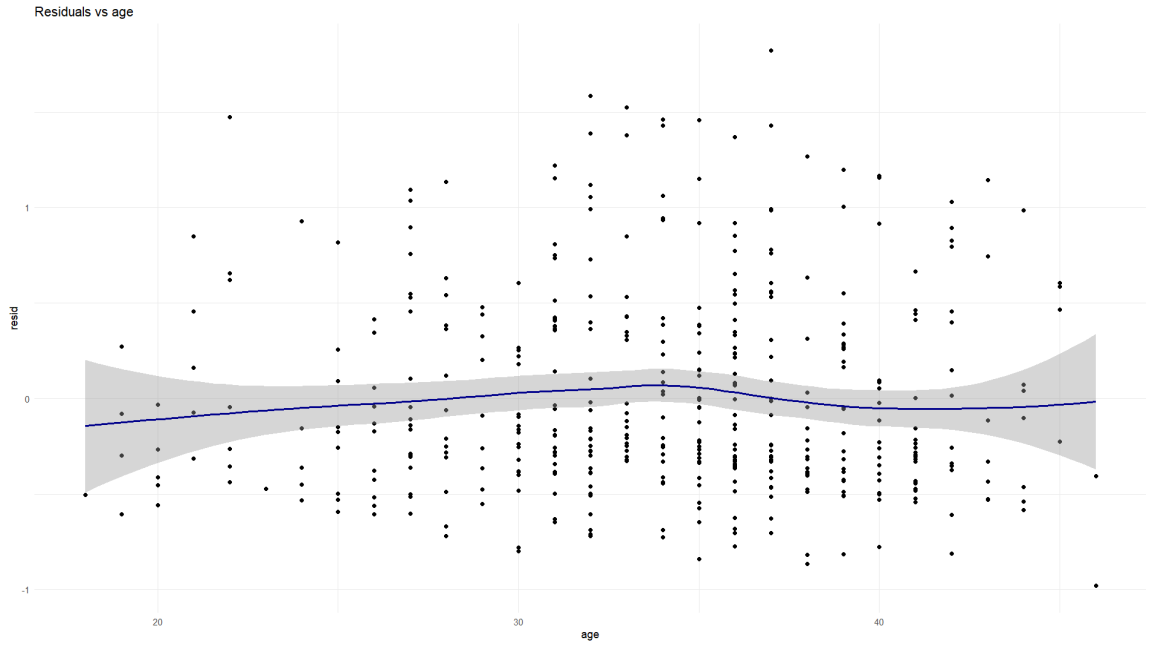
## D.1 Normality and Homoscedasticity




## D.2 Leverage/influential observations



## D.3 Linearity




## D.4 Independence

| Test | test statistic | p-value |
|---|---|---|
| Durbin-Watson test | 1.93 | 0.2111 |

## D.5 Multicollinearity

| | GVIF | Df | GVIF(1/(2*Df)) |
|---|---|---|---|
| group | 7.7639 | 1 | 2.78638 |
| age | 1.2663 | 1 | 1.1253 |
| BMI | 1.0340 | 1 | 1.0169 |
| smoking | 1.0333 | 1 | 1.0165 |
| sleep | 2.3428 | 2 | 1.2372 |
| born | 1.0344 | 1 | 1.0170 |
| group:sleep | 9.8168 | 2 | 1.7701 |

# Appendix E R codes

**For research question 1**

```
## Data filtering
Isala.OTU <- counts_matrix(Isala.count, sample_name = sample_id, taxon_name = taxon_id,
value = count)
prevalence1 <- colSums(Isala.OTU > 0) / nrow(Isala.OTU)
Isala.OTU <- Isala.OTU[, prevalence1 >= 0.01]
Flora.OTU <- counts_matrix(Flora.count, sample_name = sample_id, taxon_name = taxon_id,
value = count)
prevalence2 <- colSums(Flora.OTU > 0) / nrow(Flora.OTU)
Flora.OTU <- Flora.OTU[, prevalence2 >= 0.01]

## Shannon & Chao1 index for Isala
shannon <- diversity(Isala.OTU, index = "shannon")
shannon <- as.data.frame(shannon)
shannon$sample_id <- rownames(shannon)
chao1df <- as.data.frame(apply(Isala.OTU, 1, chao1))
chao1df$sample_id <- rownames(chao1df)
colnames(chao1df)[colnames(chao1df) == "apply(Isala.OTU, 1, chao1)"] <- "chao1"
IsalaDF.with.shannon <- merge(IsalaDF, shannon, by = "sample_id")
IsalaDF.for.model1 <- IsalaDF.with.shannon[,c("sample_id", "age", "BMI.cat", "smoking",
"sleep.cat", "born", "shannon")]
IsalaDF.for.model1 <- merge(IsalaDF.for.model1, chao1df, by = "sample_id")

## Shannon & Chao1 index for Flora
shannon <- diversity(Flora.OTU, index = "shannon")
shannon <- as.data.frame(shannon)
shannon$sample_id <- rownames(shannon)
chao1df <- as.data.frame(apply(Flora.OTU, 1, chao1))
chao1df$sample_id <- rownames(chao1df)
colnames(chao1df)[colnames(chao1df) == "apply(Flora.OTU, 1, chao1)"] <- "chao1"
sampledf.with.Shannon <- merge(Flora.samples, shannon, by = "sample_id")
FloraDF.for.model1 <- sampledf.with.Shannon[,c("sample_id", "age", "BMI.cat", "smoking",
"sleep.cat", "born", "shannon")]
FloraDF.for.model1 <- merge(FloraDF.for.model1, chao1df, by = "sample_id")

## Merge Isala & Flora datasets
MergedDF.for.model1 <- rbind(FloraDF.for.model1, IsalaDF.for.model1)
MergedDF.for.model1$group <- c(rep("0", 353), rep("1", 390))

## Fit linear models for alpha diversity
model1b <- lm(shannon ~ group + age + BMI.cat + smoking + sleep.cat + born,
data=MergedDF.for.model1)
model1b <- lm(chao1 ~ group + age + BMI.cat + smoking + sleep.cat + born,
```

```
data=MergedDF.for.model1)

## Model diagnostics
# QQ plot for normality
qqnorm(resid(model1b))
qqline(resid(model1b))
# Homoscedasticity
plot(fitted(model1b), resid(model1b), main="Residual vs Fitted")
abline(h = 0, col = "red")
# Leverage/influence
hatvals <- hatvalues(model1b)
p <- length(coefficients(model1b))
n <- nobs(model1b)
avg_leverage <- p / n
plot(hatvalues(model1b), main = "Leverage")
abline(h = 3*avg_leverage, lty = 2, col = "red")
cooksD <- cooks.distance(model1b)
n <- nobs(model1b)
plot(cooksD, main = "Cook's Distance")
abline(h = 4/n, lty = 2, col = "red")
outliers <- as.numeric(names(cooksD)[(cooksD > (4/n))])
# Linearity
modeldata <- model.frame(model1b)
modeldata$pred <- predict(model1b)
modeldata$resid <- residuals(model1b)
ggplot(modeldata, aes(x = age, y = resid)) +
  geom_point() + geom_smooth(method = "loess") + labs(title = "Residuals vs age") +
  theme_minimal()
# Independence
lmtest::dwtest(model1b)
# Multicollinearity
car::vif(model1b)
```

**For research question 2**

```
## Creating phyloseq object for merged Flora & Isala
all_taxa <- union(colnames(Isala.OTU), colnames(Flora.OTU))
for (col in setdiff(all_taxa, colnames(Isala.OTU))) {Isala.OTU[[col]]<- NA}
for (col in setdiff(all_taxa, colnames(Flora.OTU))) {Flora.OTU[[col]] <- NA}
Isala.OTU.all <- Isala.OTU[, all_taxa]
Flora.OTU.all <- Flora.OTU[, all_taxa]
Combined.otu <- rbind(Flora.OTU.all, Isala.OTU.all)
Combined.otu[is.na(Combined.otu)] <- 0
otu.tab.merged <- as.matrix(Combined.otu)
```

```
sample.data <- MergedDF.for.model1 %>% filter(!is.na(BMI.cat) & !is.na(born)
& !is.na(sleep.cat))
rownames(sample.data) <- sample.data$sample_id
metadata.phyloseq <- sample_data(sample.data)
merged.physeq <- phyloseq(otu_table(otu.tab.merged, taxa_are_rows = FALSE),
metadata.phyloseq)

## ANCOM-BC for merged dataset
Merged1.ancombc <-ancombc2(data = merged.physeq,
fix_formula = "group + age + BMI.cat + smoking", p_adj_method = "BH", group = "group",
prv_cut = 0.10, lib_cut = 1000, struc_zero = FALSE, iter_control = list(tol = 1e-5,
max_iter = 100, verbose=FALSE), alpha = 0.05, global = FALSE)

## Creating phyloseq object for Flora
rownames(Flora.samples) <- Flora.samples$sample_id
Flora.samples.phylo <- sample_data(Flora.samples)
Flora.phylo <- phyloseq(otu_table(Flora.OTU, taxa_are_rows = FALSE), Flora.samples.phylo)

## ANCOM-BC for Flora
Flora.ancombc <-ancombc2(data = Flora.phylo,
fix_formula = "age + BMI.cat + smoking + sleep.cat + born + bread + soft_drinks +
spirits", p_adj_method = "BH", group = NULL, prv_cut = 0.1, lib_cut = 1000,
struc_zero = FALSE, iter_control = list(tol = 1e-5,max_iter = 100,
verbose=FALSE), alpha = 0.05, global = FALSE)

## Creating phyloseq object for Isala
rownames(IsalaDF) <- IsalaDF$sample_id
IsalaDF.phylo <- sample_data(IsalaDF)
Isala.phylo <- phyloseq(otu_table(Isala.OTU, taxa_are_rows = FALSE), IsalaDF.phylo)

## ANCOM-BC for Isala
Isala.ancombc <-ancombc2(data = Isala.phylo,
fix_formula = "age + BMI.cat + smoking + sleep.cat + born + fermentedfd +
alcohol + sugarbev", p_adj_method = "BH", group = NULL, prv_cut = 0.10, lib_cut = 1000,
struc_zero = FALSE, iter_control = list(tol = 1e-5,max_iter = 100,verbose=FALSE),
alpha = 0.05, global = FALSE)
```

**For research question 3**

```
## Shannon & Chao1 index for Isala
shannon <- diversity(Isala.OTU2, index = "shannon")
shannon <- as.data.frame(shannon)
shannon$sample_id <- rownames(shannon)
chao1df <- as.data.frame(apply(Isala.OTU2, 1, chao1))
chao1df$sample_id <- rownames(chao1df)
```

```
colnames(chao1df)[colnames(chao1df) == "apply(Isala.OTU2, 1, chao1)"] <- "chao1"
IsalaDF2.with.shannon <- merge(IsalaDF2, shannon, by = "sample_id")
IsalaDF2.for.model1 <- IsalaDF2.with.shannon[,c("sample_id", "age", "BMI.cat",
"smoking", "sleep.cat", "born", "shannon")]
IsalaDF2.for.model1 <- merge(IsalaDF2.for.model1, chao1df, by = "sample_id")

## Merge Isala & Flora datasets
MergedDF2.for.model1 <- rbind(FloraDF.for.model1, IsalaDF2.for.model1)
MergedDF2.for.model1$group <- c(rep("0", 353), rep("1", 146))

## Fit linear model for alpha-diversity
model1b <- lm(shannon ~ group + age + BMI.cat + smoking + sleep.cat + born
+ group*sleep.cat, data=MergedDF2.for.model1)
summary(model1b)
model1b <- lm(chao1 ~ group + age + BMI.cat + smoking + sleep.cat + born,
data=MergedDF2.for.model1)
summary(model1b)

## Creating phyloseq object for merged Flora & Isala
all_taxa <- union(colnames(Isala.OTU2), colnames(Flora.OTU))
for (col in setdiff(all_taxa, colnames(Isala.OTU2))) {Isala.OTU2[[col]] <- NA}
for (col in setdiff(all_taxa, colnames(Flora.OTU))) {Flora.OTU[[col]]<- NA}
Isala.OTU2.all <- Isala.OTU2[, all_taxa]
Flora.OTU.all <- Flora.OTU[, all_taxa]
Combined.otu <- rbind(Flora.OTU.all, Isala.OTU2.all)
Combined.otu[is.na(Combined.otu)] <- 0
prevalence1 <- colSums(Combined.otu > 0) / nrow(Combined.otu)
Combined.otu <- Combined.otu[, prevalence1 >= 0.01]
otu.tab.merged <- as.matrix(Combined.otu)

sample.data <- MergedDF2.for.model1 %>% filter(!is.na(BMI.cat) & !is.na(born)
& !is.na(sleep.cat))
rownames(sample.data) <- sample.data$sample_id
IsalaDF2.phylo <- sample_data(sample.data)
merged.physeq2 <- phyloseq(otu_table(otu.tab.merged, taxa_are_rows = FALSE),
IsalaDF2.phylo)

## ANCOM-BC for merged dataset
Merged2.ancombc <-ancombc2(
  data = merged.physeq2,
  fix_formula = "group + age + BMI.cat + smoking",
  p_adj_method = "BH",
  group = "group",
  prv_cut = 0.10,
  lib_cut = 1000,
```

```
  struc_zero = TRUE,
  iter_control = list(tol = 1e-5,max_iter = 100,verbose=FALSE),
  alpha = 0.05, global = FALSE)

## Creating phyloseq object for Isala
rownames(IsalaDF2) <- IsalaDF2$sample_id
IsalaDF2.phylo <- sample_data(IsalaDF2)
Isala.phylo2 <- phyloseq(otu_table(Isala.OTU2, taxa_are_rows = FALSE), IsalaDF2.phylo)

## ANCOM-BC for IsalaDF2
Isala.ancombc2 <-ancombc2(
  data = Isala.phylo2,
  fix_formula = "age + BMI.cat + smoking + sleep.cat + born + fermentedfd +
  alcohol + sugarbev",
  p_adj_method = "BH",
  group = NULL,
  prv_cut = 0.1,
  lib_cut = 1000,
  struc_zero = FALSE,
  iter_control = list(tol = 1e-5,max_iter = 100,verbose=FALSE),
  alpha = 0.05, global = FALSE)
```