# Faculty of Sciences

## *School for Information Technology*

## Master of Statistics and Data Science

***Master's thesis***

***The effects of solid food introduction on the infant microbiome and metabolome***

**Frédérique Vilenne**

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

**SUPERVISOR :**

Prof. dr. Dirk VALKENBORG

**SUPERVISOR :**

Dr. John PENDERS

**2021**

**2022**

UHASSELT
KNOWLEDGE IN ACTION

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

# Faculty of Sciences

## *School for Information Technology*

Master of Statistics and Data Science

***Master's thesis***

***The effects of solid food introduction on the infant microbiome and metabolome***

**Frédérique Vilenne**

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

**SUPERVISOR :**

Prof. dr. Dirk VALKENBORG

**SUPERVISOR :**

Dr. John PENDERS

# Abstract

The introduction of solid foods play a major role in the maturation of the gut microbiome. Both the microbial composition and function changes due to the introduction of solid foods. This research will look into these changes of the gut microbiome in a short interval at the time of solid food introduction, together with changes in the metabolome. In addition, the association between the microbiome and metabolome during this time period is being examined. The microbiome data analysis started out by filtering the microbial taxa (amplicon sequence variants, ASVs) using a filter based on a prevalence of 5% and minimum relative abundance of 0.01%. The data was normalized using the centered logarithmic ratio, a compositional data analysis approach to account for the compositional nature of microbiome data. The data was analyzed using ANCOM-BC. A model containing covariates for the age of an infant and whether or not solid foods were given with a random effect per infant was used. Correction for multiple testing was done using the Benjamini-Hochberg False Discovery Rate. None of the microbial taxa were statistically significantly associated with solid food introduction at $\alpha = 0.05$. The metabolomics data was cleaned first and normalized using a natural logarithm and pareto-scaling. Differential abundance testing was done using a variety of methods to gain robustness in the results. The Wilcoxon rank-sum test was reported with a correction for multiplicity using Benjamini-Hochberg. None of the metabolites were found to be differentially abundant due to the introduction of solid foods at $\alpha = 0.05$. The rationale behind not finding any differentially abundant ASVs and metabolites due to the introduction of solid foods is most likely due to the fact that around the days of solid food introduction, these dietary changes might not be drastic enough to cause major changes or that the microbiome needs more time to adapt. Lastly, the association between the microbiome and metabolome was investigated using DIABLO. Hyperparameter tuning was done using PLS and LOOCV. A final model was fit using 3 principal components, a custom covariance matrix and only key contributors, ASVs and metabolites. A total of 119 high correlations (> 0.7 or < -0.7) were found between ASVs and metabolites and between metabolites. Amongst those correlations were 11 different bacteria. *Bacteroides* ovatus had a total of 21 associations between 3 ASVs and 7 metabolites, which strengthens the proof of association. All bacteria-metabolite associations could be related back to metabolites produced by humans, the bacteria itself and food sources. *Butyricicoccus pullicaecorum*, *Bacteroides caccae* and *Bacteroides ovatus* were associated with histidine, a precursor metabolite for histamine. This proves the presence of an early-life association between the microbiome and metabolome not only when looking towards the processing of food but also as a key-player in the immune system of infants.

# Table of Contents

# List of Equations

# List of Figures

# List of Tables

# Abbreviations

*Table 1 Abbreviations*

| Abbreviation | Full meaning |
|---|---|
| µM | Micromolar |
| Å | Angstrom |
| ALR | Additive Log-Ratio |
| ANCOM-BC | Analysis of Compositional Data with Bias Correction |
| ANOVA | Analysis of Variance |
| ASV | Amplicon Sequence Variant |
| ATP | Adenosine triphosphate |
| BER | Balanced Error Rate |
| BMI | Body Mass Index |
| Bp | Base pairs |
| CLR | Centre Log-Ratio |
| Cm | Centimetre |
| CoDa | Compositional Data analysis |
| CV | Cross-Validation |
| DIABLO | Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies |
| DIMS | Direct Infusion Mass Spectrometry |
| ER | Error Rate |
| FDR | False Discovery Rate |
| G | Grams |
| ILR | Isometric Log-Ratio |
| IV | Intravenous |
| LOD | Level Of Detection |
| LOOCV | Leave-One-Out Cross-Validation |
| MCAR | Missing Completely At Random |
| MRM | Multiple Reaction Monitoring |
| MS | Mass Spectrometry |
| N | Amount |
| NMR | Nuclear Magnetic Resonance |
| NOESY | Nuclear Overhauser Effect Spectroscopy |
| OUT | Operational Taxonomic Units |
| PCA | Principal Component Analysis |
| PCR | Polymerase Chain Reaction |
| PFG | Pulsed-Field Gradient |
| PLS | Partial Least Squares |
| Ppm | parts per million |
| RNA | Ribonucleic acid |
| RPLC | Reversed Phase Liquid Chromatography |
| rRNA | Ribosomal ribonucleic acid |
| SCFA | Short Chain Fatty Acids |
| sGCCA | sparse Generalized Canonical Correlation Analysis |
| sPLS-DA | Sparse Partial Least Squares Discriminant Analysis |
| Std. dev. | Standard deviation |
| TCA | Tricarboxylic acid |
| UPLC | Ultra-Performance Liquid Chromatography |
| α | Alpha |

# 1 Introduction

## 1.1 Research context

The human microbiome is composed of bacteria, archaea, viruses and eukaryotic microbes residing in and on our bodies. They interact with each other and with their host. It is often referred to as the forgotten organ or the second genome. The microbiome composition is unique for an individual and is established early in life [1, 2, 3, 4]. The microbiome composition is considerably site-specific [5, 6].

The human gut microbiome is described as all these microbes who reside in the human gut and is of interest given the important functionalities of the host and the high abundance of microbes in the human body. It has a symbiotic relationship with its human host and has key functions in the human body [4, 7, 8, 9]. Some of these functions are food degradation and metabolism [10]. Disturbances in the gut microbiome could lead to a vast number of pathologies which underlines the importance of a healthy microbiome [11, 12]. Early alterations in the infant gut microbiome have been linked to the development of chronic diseases [11, 13, 14, 15], weaker immune systems [16], vaccine responses [17] and drug metabolism [18]. The human gut microbiome is influenced by a variety of factors throughout life. Diet, antibiotics, probiotics, gender, age and disease can change the composition of the gut microbiota [19, 20]. The importance of the human gut microbiome is a well-established topic.

Therefore, gaining an understanding into the changes and maturation of the infant gut microbiome plays a crucial role in disease prevention and a healthy development of the infant. The infant gut microbiome develops from a relatively simple microbiome to a microbiome of adult state during the first three years of life [21]. There are a lot of different factors influencing this development. It starts out at birth, where babies born by caesarean section are missing key microbes [22]. After this, a variety of factors play a key role in the development of the infants' gut microbiome such as maternal milk versus infant formula feeding babies [23], probiotics [24], antibiotics [25] and the exposure to solid foods [26]. During this research, a focus is placed on the introduction of solid foods. This change of diet plays an important role in the maturation of the microbiome as different metabolites get introduced in the infants' body [27, 28].

Metabolites can be measured by metabolomics. Metabolomics is the study of the raw materials, for example food, and products of the body's biochemical reactions, molecules that are smaller than most proteins, DNA and other macromolecules. Similarly, to the microbiome, the metabolome is unique for each individual [29]. The human metabolome can be influenced by several factors such as age, disease, drugs, environment, genetic factors, lifestyle and nutrition [30].

There is an association between the metabolome and the microbiome in the gut [31]. These molecules can be nutrients that shape the composition of the microbiome [32] or important by-products of host-microbe nutrient co-metabolism [33, 34, 35]. Studies suggest that the faecal metabolome can be used to gain insight into the metabolic functions of the gut microbiome. A large cohort study in adults indicated that around 60% of the faecal metabolome is associated with the microbial composition and, on average, 67% of the variance in the metabolome can be explained by the microbiome [36, 37].

There is only a small amount of studies focussing on the correlation between the gut microbiome and metabolome. Most of these studies have more spread out sampling points. This study will allow for a focus on a short time interval where an important event, such as the introduction of solid foods, for the infant gut microbiome takes place.

## 1.2 Research questions

There is still a lot which is not yet understood regarding this topic. What drives the change in the human gut microbiome at the time of introduction of solid foods? Previous studies have shown a link between the changing ratios of fat, protein carbohydrates and fibre in the diet. Therefore, a more in-depth analysis of the infant gut microbiome and metabolome could provide essential information on the development of the microbiome at the time of introduction of solid foods. Gaining a further understanding of the development of the gut microbiome in infants could provide useful insights into the possible causation of health related problems in a later stadium. This research will focus on the following research questions:

- What is the impact of the introduction of solid foods on the gut microbiome?
- What is the impact of the introduction of solid foods on the metabolome?
- Is there an early-life association between the gut microbiome and metabolome?

## 1.3 Outline

The research starts with the elaborating upon the data and methods in Chapter 2. A first step is discussing the study design and data collection in Chapter 2.1. The study design covers the study used to address the research questions at hand. The data collection describes the methods used to collect data about the microbiome and metabolome. Chapter 2.2 discusses the methods used in depth. The pre-processing, exploratory tools and statistical methods applied for each of the data sets. It is ended by discussing the method to study the association between the microbiome and metabolome.

The results are shown in Chapter 3. The microbiome data is discussed first in Chapter 3.1, followed by the metabolomics data in Chapter 3.2 and the association between the microbiome and metabolome in Chapter 3.3. For the microbiome and metabolome data, the effects of pre-processing the data is illustrated, followed by exploratory tools and finished with a statistical analysis. For the association study, the hyper parameter tuning is addressed followed by the final results.

The results are discussed in Chapter 4 where a relationship is made between literature and the obtained results. A reason is sought behind the obtained results and related back to literature. The advantages and drawbacks of the research are discussed.

A final conclusion about the research is given in Chapter 5 with relation to the research questions. Potential future research is proposed based on the obtained results.

# 2 Data and methods

## 2.1 Data

### 2.1.1 Study design

The data used for investigating the effects of solid food introduction in infants originates from the LucKi Gut Study. This is a sub-study of the LucKi study, a longitudinal cohort study from Maastricht in the Netherlands [38]. The LucKi Gut Study uses questionnaires to gather information about the birth, diet, medication and other exposures. In addition to the questionnaire, eligible participants will be asked to provide data and collect stool samples over a 14-day period. During this period, solid foods are introduced. Consent was given by the caregivers of the infants and data was pseudonymized. The LucKi Gut Study was approved by the Medical Ethics Committee Maastricht University Medical Centre in the Netherlands. A summary of the study can be found in Table 2 [27, 39].

*Table 2 Descriptive summary of the study design of the LucKi-Gut Cohort study*

| | **LucKi Gut study** |
|---|---|
| **Sample size** | 9 infants |
| **Source population** | South-Limburg, Netherlands |
| **Inclusion criteria** | • Full-term (> 37 weeks)<br>• Singleton<br>• Low risk (defined as being followed through Baby Welfare Clinics) |
| **Exclusion criteria** | • Caesarean section birth<br>• Admission to the neonatal intensive care unit<br>• Full weaning prior to introduction of solid food<br>• Use of oral or IV antibiotics within 4 weeks of introduction of solid foods<br>• Parent or guardian unable to communicate in Dutch |
| **Follow-up time period** | A 14-day sampling period at the time of solid food introduction |

Fresh stool samples were frozen by the caregivers upon defecation. Research staff was informed by the caregivers to arrange sample pick-up [39]. The stool samples were used to investigate the microbiome and metabolome. Additionally, caregivers were asked to fill out a study diary every day. Caregivers were asked to report if a stool sample was collected, the consistency of the sample based on the Bristol Stool Chart, if the sample had contact with diaper cream, the type of diaper the sample was collected from, number of bower movements from the infant per day, medications the mother was on, time spent asleep by the infant, how many times the infant woke up during the night and the time spent awake during these periods, interaction with other children or animals, food consumption and medication [27, 39]. Many of these variables however, were not used during this research as this was not the scope of this topic.

The study consisted out of 5 male and 4 female infants. The infants had a mean age of 158 days (std. dev. 21.66) at the time of introduction into the study, a mean weight of 5658.88 grams (std. dev. 1623.96) and mean height of 61.22 cm (std. dev. 6.78). All 9 infants were given breastfeeding prior to the introduction of solid foods. One infant initially received breastfeeding but switched to formula feeding at a certain point in time prior to the introduction of solid foods. The delivery mode was vaginal for 8 out of 9 infants and unknown for 1 infant. One infant was subjected to antibiotics at the time of birth and antibiotic use was unknown for one of the infants. None of the infants were subjected to any probiotics (Table 3).

*Table 3 Summary of the participants in the LucKi Gut Study.*

| ***Variable*** | | **Summary statistic** |
|---|---|---|
| ***Gender*** | *Male* | n = 5 |
| | *Female* | n = 4 |
| ***Mean height (cm)*** | | 61.22 (std. dev. 6.78) |
| ***Mean weight (gram)*** | | 5658.88 (std. dev. 1623.96) |
| ***Mean age at the time of introduction (days)*** | | 158 (std. dev. 21.66) |
| ***Breastfeeding*** | | n = 9 |
| ***Delivery mode (vaginal)*** | | n = 8 |
| ***Antibiotics*** | | n = 1 |
| ***Probiotics*** | | None |

### 2.1.2 Data collection

All collected faecal samples from each infant were subjected to microbiota profiling using 16S ribosomal RNA (rRNA) V4 hypervariable gene region sequencing, while a selection of samples were subjected to metabolomics using nuclear magnetic resonance spectroscopy (NMR), direct infusion-mass spectrometry (DIMS) and ultra-performance liquid chromatography (UPLC). It was opted to have at least 1 and preferably 2 samples prior to the introduction of solid foods which included metabolomics data. After the introduction of solid foods, a maximum of 3 samples were used to acquire metabolomics data (Table 4).

*Table 4 Study design of the LucKi Gut Study (N = 9).* [1]

| **Infant** | **Day 1** | **Day 2** | **Day 3** | **Day 4** | **Day 5** | **Day 6** | **Day 7** | **Day 8** | **Day 9** | **Day 10** | **Day 11** | **Day 12** | **Day 13** | **Day 14** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **P** | 3 | 1 | | | 3 | 1 | | 3 | | | | | | |
| **Q** | 1 | 4 | 1 | 4 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 4 | | |
| **R** | | | 4 | 1 | 4 | 1 | 1 | | 4 | 1 | 1 | 1 | 1 | 4 |
| **S** | 4 | 1 | 4 | 1 | 4 | 1 | 1 | 4 | 1 | | | | | |
| **T** | 3 | 1[2] | 1 | | 4 | 1 | 1 | 1 | | 4 | 1 | | 4 | |
| **U** | 4 | 1 | 4 | | | 1 | 1 | 4 | 1 | | 1 | | | 4 |
| **V** | 4 | 1 | 4 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 1 | 1 | 4 |
| **W** | | 4 | | 4 | 1 | | 4 | | | 1 | 4 | | | |
| **X** | 4 | 4 | | 1 | 1 | 1 | 1 | 4 | 1 | | 1 | | 1 | 4 |

[1]: The red numbers denote stool samples taken before the introduction of solid foods. The green numbers denote stool samples taken after the introduction of solid foods. The numbers denote the following, 1 for samples of which only 16S rRNA sequencing has performed, 2 for samples for whom 16s sequencing and NMR was performed, 3 for samples for whom 16s sequencing, NMR and DIMS was performed and lastly 4 for stool samples for which 16S rRNA sequencing, NMR, DIMS and UPLC has been performed.

[2]: This sample was used for UPLC. There was no sample left of day 1 and it was opted to use the sample of day 2 instead.

A total of 87 samples were collected and analysed (Table 5). The majority of samples (65/87, 74,71%) were collected after the introduction of solid foods. Figure 1 shows a Venn-diagram of the methods used per sample. It is observed that only 35,63% (31/87) of the samples were measured using all 4 measurement methods and 58.62% (51/87) only using 16S rRNA sequencing.

*Table 5 Overview of measurement times for samples with relation to solid food introduction.*

| ***Summary*** | **N** | **Percentage** |
|---|---|---|
| *Pre-introduction of solid foods* | 22 | 25,29% |
| *Post-introduction of solid foods* | 65 | 74,71% |
| *Total amount of Samples* | 87 | 100,00% |

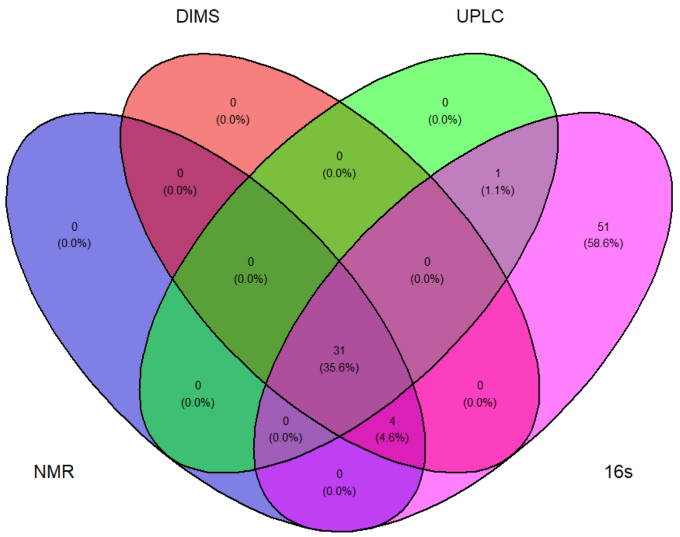

*Figure 1 Venn diagrams of the methods used per sample.*

A drawback of the study is the small sample size and unbalanced design due to the fact that there isn't a sample available on a daily basis of all infants, due to absence of bowel movements, and not all data collection methods were applied on all samples. This could cause a potential loss of power to get statistically significant results. The missing samples can be assumed to be missing completely at random (MCAR) and a complete case analysis is appropriate.

#### 2.1.2.1 Microbiome data

The microbiota composition was investigated through sequencing of the 16S rRNA hypervariable V3-4 gene region. The 16S rRNA gene sequence is used to study the bacterial phylogeny and taxonomy in the stool samples. This gene is the most commonly used genetic marker because it is present in almost all bacteria and is large enough for informatics purposes [40]. The 16s rRNA gene is a highly conserved region in the rRNA with variable and constant regions. The constant regions make amplification possible by using universal PCR primers. Meanwhile, sequencing of the variable regions allows for discrimination between different micro-organisms such as bacteria and archaea [41].

The microbial DNA from stool samples was extracted as previously described in the article by J.C. Stearns et al. [42]. Amplification of the 16S rRNA gene was performed as previously described by A.K. Bartram et al. [43]. Sequencing was performed in the McMaster Genomics Facility in Hamilton, Canada. Illumina libraries were paired-end sequenced with 250bp sequencing in the forward and reverse directions on the Illumina MiSeq instrument. Sequencing data was processed with standard software [44] and amplicon sequence variants (ASVs) were inferred using the DADA2 pipeline [45].

#### *2.1.2.2 Metabolome data*

The metabolome was investigated through three metabolic methods. Both untargeted, DIMS and targeted, NMR were used to assess short chain fatty acids and other organic acids and alcohols. Both methods measure different metabolites. The third method used was an additional targeted method, namely UPLC, to quantify bile-acid profiles. All metabolic methods were performed and analysed at The Metabolomics Innovation Centre located at the University of Alberta, Canada.

##### 2.1.2.2.1 Nuclear Magnetic resonance spectroscopy

Quantitative NMR spectroscopy was used for targeted metabolomic analysis of water-soluble metabolite classes including amino acids, saccharides, alcohols, organic acids, amines, tricarboxylic acid (TCA) cycle intermediates and short chain fatty acids (SCFAs). Total metabolites were measured with nuclear resonance spectrometry (NMR). All $^{1}$H-NMR spectra were collected on a 700 MHz Avance III (Bruker) spectrometer equipped with a 5 mm HCN Z-gradient pulsed-field gradient (PFG) cryoprobe. $^{1}$H-NMR spectra were acquired at 25°C using the first transient of the Nuclear Overhauser Effect Spectroscopy (NOESY) pre-saturation pulse sequence (noesy1dpr), chosen for its high degree of quantitative accuracy. All free induction decays were zero-filled to 250.000 data points. The singlet produced by the DSS methyl groups was used as an internal standard for chemical shift referencing (set to 0 ppm). All $^{1}$H-NMR spectra were processed and analysed using the Chenomx NMR Suite Professional software package version 8.1 (Chenomx Inc., Edmonton, AB). The concentration of the metabolites is expressed in µmol/g.

##### 2.1.2.2.2 Direct Flow Injection Mass Spectrometry

For targeted metabolomic analysis of biogenic amines, amino acids, acylcarnitines, phospholipids and sphingolipids, direct flow injection mass spectrometry (DIMS) was used. Untargeted metabolites were measured with direct flow injection mass spectrometry with an Agilent 1100 series HPLC system (Agilent, Palo Alto, CA) and an Agilent reversed-phase Zorbax Eclipse XDB C18 column (3.0 mm × 100 mm, 3.5 µm particle size, 80 Å pore size) with an AB SCIEX QTRAP® 4000 mass spectrometer (AB SCIEX, CA, U.S.A.). The controlling software was Analyst® 1.6.2. The mass spectrometer was set to positive electrospray ionization with multiple reaction monitoring (MRM) mode. The concentration is expressed in µM.

##### 2.1.2.2.3 Ultra-High Performance Liquid Chromatography

UPLC was used as a targeted metabolomic method to examine bile acids in the stool samples. Bile acids were measured with Ultra-Performance Liquid Chromatography-Tandem Mass Spectrometry (UPLC) on an Agilent 1290 system coupled to a 4000 QTRAP mass spectrometer. The MS instrument was operated in the multiple-reaction monitoring (MRM) mode with negative-ion (-) detection. A Waters BEH 15-cm long, 2.1-mm I.D. and C18 LC column was used, and the mobile phase was (A) 0.01% formic acid in water and (B) 0.01% formic acid in acetonitrile for binary-solvent gradient elution by RPLC. Linear regression calibration curves were constructed between analyte-to-internal standard peak area ratios (As/Ai) versus molar concentrations (nmol/mL). The final concentrations of the bile acids are expressed in nmol/g.

## 2.2 Methods

In the methods section, the complete analysis pipeline performed is explained. Each data set was treated independently prior to doing an association study between the microbiome and metabolome data. Data pre-processing steps, exploratory tools and statistical methods are explained and elaborated upon. All code written in the Master thesis was written using R in RStudio version 4.2.0 [46]. Packages and their versions are listed in the Appendix in Table 10.

In addition to the microbiome data and the metabolome data, metadata was also provided in a csv-format. The metadata file contains variables such as gender, height, weight, age and additional information collected by using the study diaries. A total of 188 variables were recorded. However, many of these variables carry little to no information essential to this research. Variables suspected to be essential based upon literature research were retained. Additionally, any identifying variables were further anonymized for visualizations.

### 2.2.1 Microbiome data and analysis

For the microbiome data, 87 measurements of the 9 infants were acquired by 16s rRNA sequencing. The data was processed using the DADA2 pipeline [45]. ASV read counts were obtained and formatted in a csv-file. ASVs are reads with identical sequences and an alternative to Operational Taxonomic Units (OTUs) which are bins of reads based upon a certain similarity threshold. ASVs have a higher sensitivity and specificity [47]. The raw data contains counts of 8787 different ASVs. These ASVs were identified up to different taxonomic ranks (Figure 2).

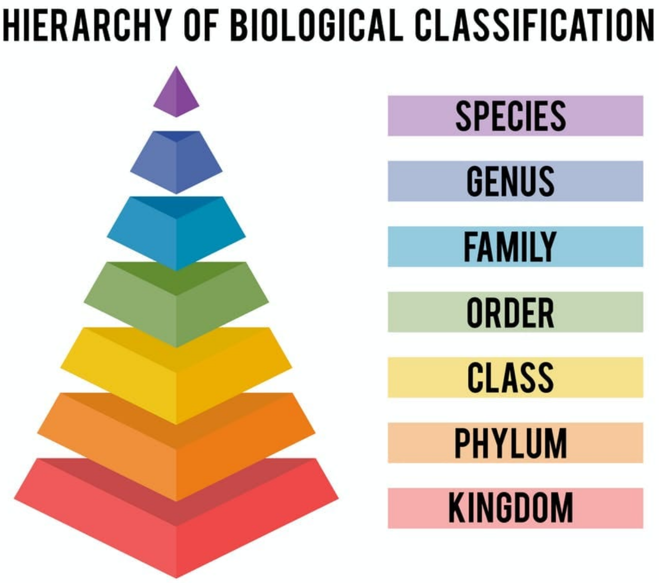

*Figure 2 The taxonomic ranks identified in the microbiome data [48].*

These taxonomic ranks are Kingdom, Phylum, Class, Order, Family, Genus and Species. Some ASVs were not able to be identified up to the lowest taxonomic rank and are therefore named by the lowest taxonomic rank to which they were able to be classified. For instance, order instead of species. The taxonomic data was provided in a separate csv-file.

#### *2.2.1.1 Phyloseq*

A first step in the analysis was the creation of a Phyloseq object for the microbiome data. Phyloseq is a package used to import, store, visualise and analyse complex microbiome data [49]. It combines the ASV counts, taxonomic data and metadata into a single Phyloseq object. The Phyloseq object contained data on 8787 taxa for 87 samples and 188 metadata variables.

#### 2.2.1.2 Filtering and cleaning of the microbiome data

Prior to doing any forms of analysis, the microbiome data was filtered. Filtering is required due to the sparseness of the microbiome data sets. This is due to the fact that it contains a large number of rare taxa observed in only a small number of samples [50]. Quality control studies indicate that rare taxa appear due to various reasons such as sequencing artefacts [51], contamination and/or sequencing errors [52, 53, 54, 55]. Filtering reduces the complexity of the microbiome data by removing rare taxa while retaining informative taxa. This leads to a reduction of technical variability allowing for more reproducible and comparable results in the data analysis [50]. An additional advantage of filtering the microbiome data is the dimension reduction, leading up to less hypotheses tests to be conducted and a higher statistical power.

Most filtering approaches are based on the rules of thumb, which vary from lab-to-lab. An important point to consider is that the filtering must be independent of the test statistic evaluated. This means that the filtering must be done across all samples and not within one group compared with another. Different approaches are cut-offs for the prevalence or the abundance of taxa across samples [56].

The first filtering method used in this article was based on a hard cut-off prevalence threshold of 5%. Rationale behind the threshold was that the infant with the lowest number of measurements, infant "P", only had 5 measurements. A total of 87 samples were present. So having 5 measurements amounted to 5.75% of the sample size, rounded down to 5%. This allowed for a taxon to be specific for all samples of a single infant.

A second filter was applied to filter out rare taxa further. The filter applied was based on the relative abundances of the taxa. The relative abundances are calculated for each taxon by summing up the counts for each taxon and dividing by the total count of all taxa, shown in Equation 1. A hard cut-off value of at least a relative abundance of 0.01% was applied based on domain knowledge.

*Equation 1 Taxa specific relative abundances.*

$$Relative\ abundance_{taxa} = \frac{Count_{taxa}}{Total\ Count} \times 100$$

A visualization using a density plot was shown to see the distribution of the ASV counts prior and posterior to filtering the microbiome data.

Additionally, upon data exploration of the microbiome data, it was observed that not all taxa were identified up till species rank. These were giving the lowest identified possible taxonomic rank making use of the MicroViz package [57].

#### 2.2.1.3 Normalisation of the microbiome data

A next step performed was the normalization of the microbiome data. Differential abundance testing will be performed to see which taxa change due to the introduction of solid foods. To do so, normalization is required because due to the varying library sizes of each sample being an obstacle for differential abundance testing. Library sizes are the sum of all taxa counts in a sample. These vary between samples which is regarded as a technical artefact. Therefore, the counts can only be compared using relative abundances. Failing to normalize the data will results into a systematic bias that increases the false discovery rate [58].

This introduces a challenge, as data naturally described by proportions, such as the relative abundances for the microbiome data, are referred to as compositional data. Proportions have a sum constraint of 1, shown in Equation 2. This is defined in mathematics as the Aitchison Simplex [59].

*Equation 2 Compositional data sum constraint. For count N of  taxon j = 1 till k in sample i divided by the library size in sample i.*

$$\sum_{j=1}^{k} \frac{N_{ik}}{L_i} = 1$$

There are two major problems with the compositionality of data:

- An increase or decrease of abundance, may be the consequence of the true decrease or increase of the abundance of one or more other taxa. This is due to compensation of the sum constraint shown in Equation 2. A graphical depiction is shown in Figure 3 [60].
- Removing taxa may result in changes of commonly used distance measures such as the Euclidian distance or Bray-Curtis distance. This indicates that these distance measures are not sub-compositional incoherent [60].

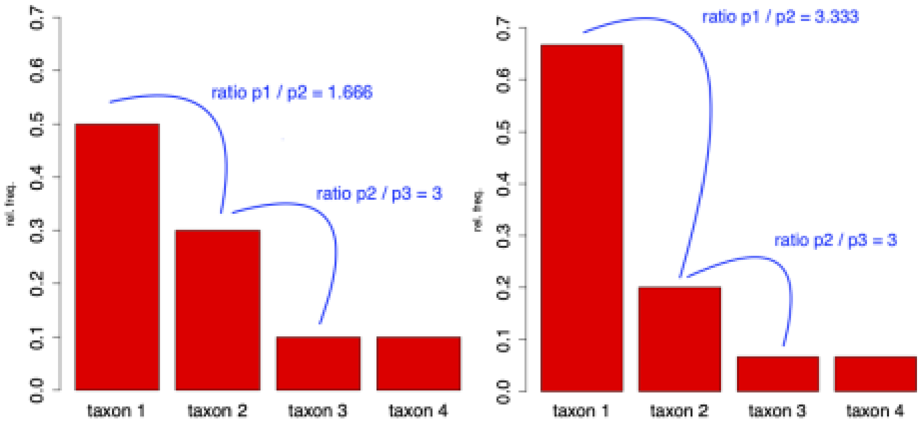

*Figure 3 Graphical depiction of the first problem with compositionality. The absolute abundance of taxon 1 is doubled. This influences the relative abundances of the other taxa too due to the sum constraint.*

Due to these problems with compositionality, standard statistical methods are not appropriate for analysing compositional data. If the compositional feature of the microbiome data is not taken into consideration during differential abundance analysis, the false discovery rates are inflated [58]. The principle of compositional data analysis or CoDa methods is circumventing the problems of compositionality by working with ratios instead of read counts. There are different ways of transforming the read counts to ratios such as additive log-ratio (ALR), centre log-ratio (CLR) or isometric log-ratio (ILR) [56, 60].

During the exploratory data analysis of the microbiome data, it was opted to choose for the centred log ratio approach. The centred log-ratio transformation is a CoDa approach using the geometric mean of the read counts of all taxa within a sample as the denominator for that sample [56, 60]. The formula is shown in Equation 3. It was opted to choose for the CLR due to disadvantages of the other methods. The ALR transformation requires a reference taxon as denominator. However, this choice is arbitrary as the results are dependent on the choice of reference taxon [61, 62]. The ILR transformation has the disadvantage that there is no one-to-one relationship between the original components and the transformed variables [60].

*Equation 3 Centred Log-Ratio.*

$$clr(N_{ij}) = \log \frac{N_{ij}}{g(N_i)} \quad where\ g(N_i) = (\prod_{j=1}^{p} N_{ij})^{1/p}$$

During the hypotheses testing for differentially abundant taxa, a different normalisation method was used. This is elaborated upon in Chapter 2.2.1.6.

#### 2.2.1.4 Variable selection

The selection of variables to investigate was based on literature and by consulting experts with domain knowledge. These variables were shortly discussed and elaborated upon on why they were in- or excluded during the statistical analysis.

Gender is known to play a significant role in affecting the gut microbiome composition. This is due to several of reasons. A first major factor is gender specific hormones affecting the microbiome. The $\alpha$-diversity becomes significantly different between males and females after puberty [63]. Gender however was not included in to statistical analysis. The rationale being that the infants included in the study don't produce any gender-specific hormones yet. Therefore, their microbiomes were not yet altered by it. Gender was investigated during the exploratory data analysis.

Another major factor influencing the gut microbiome is the delivery mode of the infant. The microbiota differ between caesarean born and vaginally delivered infant over the first year of life. Previous studies show that vaginally born children show an enrichment of *Bifidobacterium* spp. and reduction of *Enterococcus* and *Klebsiella* spp. [64]. As the study design excluded any infants born through a caesarean section, it is not possible to include this during the analysis.

The same conclusion was given for breastfeeding. Breastfeeding shapes the gut microbiota in early life, both by directly exposing the infant to milk microbiota and indirectly through maternal milk factors that affect bacterial growth and metabolism [23, 65]. As all infants in the study were breastfed prior to the introduction of solid food, no data was available on formula-fed babies and was not further investigated.

Other variables such as antibiotics and probiotics were not used prior or during the study and not included in the models. Yet they play a major role. Antibiotics cause a decrease in abundance in the gut microbiome, probiotics cause an increase [19]. However, there was some form of data available on the disease status of the infants. This data was free text describing the general state of the infant such as a weeping nose, fever, coughs and snotty eyes. Due to the sparse data, all disease states were generalized into a binary covariate. It has to be taken into consideration that neither of these symptoms might describe an actual infection. This was further investigated during the exploratory data analysis.

A final variable considered was the age of the infant. The gut microbiome changes rapidly during the first three years of infancy [19, 20]. A glance was given during the exploratory data analysis at simple demographics such as weight and length. However, it is unknown if these truly affect the gut microbiome. For instance, the BMI is known to affect the gut microbiome [66]. However, the measuring of the length of infants is inaccurate and was therefore only considered during the exploratory data analysis.

### 2.2.1.5 Exploratory data analysis of the microbiome data

A first step during the exploratory data analysis was performing a principal component analysis (PCA) on the filtered and normalised microbiome data. PCA is a technique which reduces the dimensionality of datasets, increases the interpretability, and minimizes the loss of information. This is done by creating principal components which are uncorrelated and maximize the variance present in the data [67]. PCA is particularly useful to gain insight into data by introducing different colours and shapes for variables of interest in order to reveal potential clusters happening within the data. Another way of observing clustering is by constructing a heat map. In the present study, the heat map for the microbiome data was based on the Aitchison distance, which is simply the Euclidean distance between the CLR-normalised data. The linkage to calculate the grouping was based on Ward's minimum variance method [68]. In Ward's minimum variance method, the distance between two clusters is the Analysis of Variance (ANOVA, [69]) sum of squares between the two clusters added up over all the variables.

A final step in this part of the exploratory data analysis was the creation of bar plots based on their relative abundance. This was done at family rank to make the visualizations more interpretable. In the filtered dataset, only 19 families are present. The disadvantage of visualizing at family rank is that increasing or decreasing abundances for certain species is unobserved. A general example may be that *Enterobacter cloacae* might increase but *Enterobacter aerogenes* might decrease. This causes the entire family of *Enterobacter* spp. to remain steady. The bar plots will give a first indication of potentially differentially abundant taxa.

#### 2.2.1.5.1 Alpha diversity measures

A set of $\alpha$-diversity measures were calculated from the filtered microbiome data to visualise different variables of interest. The $\alpha$-diversity summarizes the structure of the gut microbiota with respect to its richness and evenness. The richness is the number of taxonomic groups, the evenness describes the deviation of the abundances from the uniform distribution [70]. The principle is shown in Figure 4.

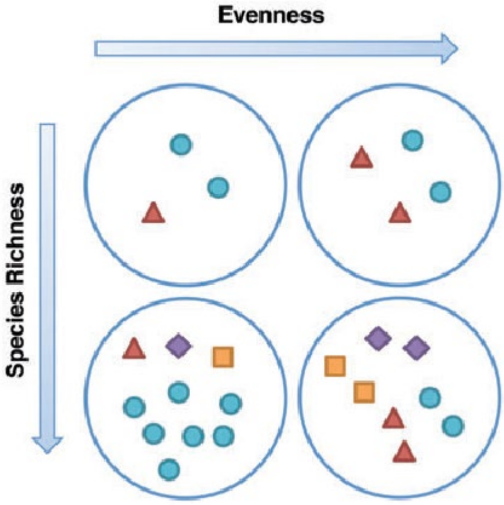

*Figure 4 Alpha diversity conceptualized. As the amount of different species in the sample increases, so does the richness. The evenness increases as the distribution of the abundances of species becomes more uniformly distributed. Reference [71].*

Different alpha-diversity measures are considered:

- The observed richness is a count of the different ASVs occurring at least once in a sample. There is no correction for taxa not observed in the sample but present in the microbiome. The formula is shown in Equation 4.

- The inverse Simpson index is a measure for the evenness of the microbiome and the inverse of the classical Simpson diversity estimator. The classical Simpson diversity estimator is the probability that two randomly selected micro-organisms belong to the same taxon. It is a measure of un-evenness. By taking the inverse, it becomes a measure for evenness [72]. The formula is shown in Equation 5.
- The Shannon-index or Shannon entropy is the uncertainty or entropy associated with the prediction of a randomly sampled taxon. High values for the Shannon-index indicate a diverse ecosystem [73]. The formula is shown in Equation 6.

*Equation 4 Observed richness. Where R denotes the richness and $S_0$ denotes the number of taxa observed at least once in a sample.*

$$R = S_0$$

*Equation 5 Inverse Simpson. Where $\lambda_i$ denotes the Simpson index for sample i and $P_{ij}$ the proportion of taxon j in sample i computed by dividing the abundance of taxon j in sample i by the library size of sample i.*

$$\frac{1}{\lambda_i} = \frac{1}{\sum_{j=1}^{R} P_{ij}^2}$$

*Equation 6 Shannon-index. Where $H_i$ denotes the Shannon-index in sample i and $P_{ij}$ the proportion of taxon j in sample i computed by dividing the abundance of taxon j in sample i by the library size of sample i.*

$$H_i = -\sum_{j=1}^{R} P_{ij} ln(P_{ij})$$

The alpha diversity indices were used during the exploratory data analysis for a variety of purposes. They were used to visualize whether an increase of alpha diversity had taken place after the introduction of solid food, longitudinal trends and as a summary statistic for other potential covariates described in Chapter 2.2.1.4.

#### 2.2.1.6 Statistical analysis for differentially abundant ASVs

The first main scope of this research was finding if the introduction of solid foods affected the microbiome. The interest was finding if there were any ASVs differentially abundant after the introduction of solid food. To test for differentially abundant ASVs, analysis of compositions of microbiomes with bias correction or ANCOM-BC was performed [74].

ANCOM-BC is based on Aitchison’s methodology which considers the compositional nature of microbiome data. The relative abundances are used to do inference about the absolute abundances. Based on simulation studies, ANCOM-BC performs well in controlling the False Discovery Rate (FDR) while maintaining a high power. However, it requires a sample size of at least 5 per group, which is satisfied for the LucKi gut Cohort [75, 74].

To perform ANCOM-BC, the filtered, non-normalized microbiome data was used. This is due to the fact of ANCOM-BC having a built-in normalization step. ANCOM-BC assumes that the sample is an unknown fraction of the entire system. It accounts for a sampling fraction by introducing an offset term in a linear regression framework. This offset term is estimated from the observed data and also serves as the bias correction. The linear regression framework in the logarithmic scale is equivalent to the log-ratio transformation to account for the compositional nature of the microbiome data. The models identify taxa that are differentially expressed [74].

ANCOM-BC allows for the specification of several settings. The formula used to model the data included covariates for the age of the infants in days and whether or not solid foods were given. Due to the clustering within infants, it was opted to allow for grouping of the data within infants. This can be regarded as a random effect. Filters provided by ANCOM-BC based on prevalence and library sizes were specified to avoid any further exclusion of the taxa still present in the filtered data set. It was opted not to search for structural zeros in the data as this could also lead to exclusion of taxa. As ANCOM-BC uses the Expectation-Maximization algorithm, convergence was set at $1 \times 10^{-5}$ with a maximum number of iterations at 100. Lastly, ANCOM-BC allows for global tests of significance for the parameters. This was disabled since the absence of taxa in some samples didn't allow for this global test. The correction for multiplicity was done using the Benjamini-Hochberg procedure. This is less conservative in comparison to other methods to correct for multiplicity and control the FDR at a significance level $\alpha = 0.05$ [76]. The log-linear modelling framework makes several assumptions such as a linear relationship between the outcome and the covariates, normality of the error term, homoscedasticity, or constant variance and little to no multicollinearity between covariates. The hypothesis of interest is shown in Equation 7. The results are shown using volcano plots.

*Equation 7 Hypothesis test for differentially abundant ASVs for ASV i.*

$$H_o: Effect\ of\ introduction\ of\ solid\ foods_i\ = insignificant$$
$$H_a: Effect\ of\ introduction\ of\ solid\ foods_i\ = significant$$

### 2.2.2 Metabolomics data

Metabolomics data was gathered through three different methods. Two methods, NMR and DIMS, are untargeted methods to identify metabolites such as fatty acids and other organic acids and alcohols. UPLC was specifically used to identify bile-acids. The number of metabolites are shown in the Venn-diagram in Figure 5. Each method identified a unique set of metabolites, no metabolite was identified by more than 1 method. NMR identified 41 metabolites in 35 samples, DIMS identified 116 metabolites in 35 samples and UPLC identified 75 metabolites in 32 samples.

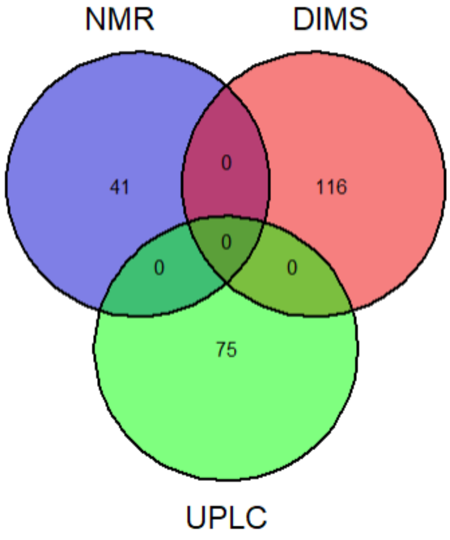

*Figure 5 Venn diagram of the metabolites identified per method.*

#### 2.2.2.1 Data cleaning

Prior to starting any analysis, the data sets were screened to see whether or not all data was valid and if any data cleaning was needed. Several cleaning steps were performed.

All methods expressed the metabolite concentrations in different units. These were recalculated to have the same units, namely µmol/gram. Next, was the removal of several observations. All three data sets contained data from time points measured outside of the LucKi Gut cohort intensive sampling study. These data points were removed as they are not eligible for the association study performed.

Also, in the DIMS data, 49 data points were below the level of detection of the method (LOD). There are different approaches to deal with these observations. Literature suggests different approaches based on the LOD, such as replacing it with a value equal to half of the LOD. However, the level of detection is unknown for the used method and devices. A different approach suggests replacing these values with 0. This approach was also used during this research.

Lastly, in the UPLC data set, 5 metabolites were removed due to being completely absent. The following bile-acids were removed:

- 6,7-Diketolithocholic acid: A bile-acid derived from lithocholic acid. It plays a role as a bacterial metabolite produced by *Bacillus* species [77]. *Bacillus* species was absent in the faecal samples (Figure 9).
- Glycodehydrocholic acid: A bile-acid glycine conjugate [77].
- Lithocholic acid-3-Glucuronide: A bile-acid found in human urine samples [77].
- Lithocholic acid-24-Glucuronide: A bile-acid found in human urine samples [77].
- Taurodeoxycholic acid-3-sulfate: A bile-acid taurine conjugate [77].

After the data cleaning procedures, the metabolite data sets are ready for data normalization.

#### 2.2.2.2 Data normalization

Prior to conducting any exploratory data analysis or statistical analysis, the metabolomics data was normalized. The benefits and purpose of normalizing microbiome data have been discussed during Chapter 2.2.1.3. For metabolome data, other factors play a role into normalising data [78]:

- There are differences in orders of magnitudes between measured metabolites. Highly abundant metabolites like ATP for example are not necessarily more important than those present at low concentrations. Normalizing will rescale all metabolites to the same order of magnitude.
- There are differences in fold changes in metabolite concentrations due to induced variation. Metabolites from the central metabolism are generally relatively constant. Metabolites from secondary metabolism tend to show larger fluctuations in concentrations depending on the environmental conditions.
- Sometimes metabolites show large fluctuations under identical experimental conditions, this phenomenon is called uninduced biological variation.
- Technical variation due to sampling, isolation techniques, measurement errors.
- Heteroscedasticity.

There are several normalization methods of dealing with this unwanted variation in the field of metabolomics [78].

- Centring: Centring converts all the concentrations to fluctuate around zero instead of around the mean of the metabolite concentrations. This way, it removes the offset in the data. However, it does not remove any heteroscedasticity in the data. It is quite often combined with data scaling and transformations.
- Scaling: Scaling methods normalize the data by dividing each variable by a factor, a scaling factor. Which differs for each variable. The main aim is to adjust for the differences in fold changes between the different metabolites. The undesirable effect however is an inflation of small values. There are two subclasses of scaling, measures of data dispersion and size measures.

- Transformations: Transformations are non-linear transformations of the data like a logarithmic or power transformations. They are generally applied to correct for heteroscedasticity and to make skewed distributions more symmetric [79].

It was opted to use two normalization steps. A first normalization step was performing a natural logarithmic transformation on the abundances. The logarithmic transformation corrects for heteroscedasticity present in the data. However, it reduces large values in the data set relatively more than small values. The transformation has a pseudo scaling effect as differences between large and small values in the data are reduced. Due to the pseudo scaling effect of the logarithmic transformation, it is rarely sufficient to fully adjust for the magnitude differences. So applying a scaling method together with a logarithmic transformation can be beneficial. One problem with the logarithmic scaling is the inability to deal with zeros as they are transformed to minus infinity. This is solved by adding 1 to the abundances. The rationale of adding 1 is that metabolites that had an abundance of 0 in an infant will become zero again after the logarithmic transformation. The formula is shown in Equation 8 [78].

*Equation 8 Logarithmic transformation. Where $\tilde{x}_{ij}$ is the natural log-transformed metabolite abundance of metabolite i in infant j and $x_{ij}$ is the non-normalized metabolite abundance of metabolite i in infant j.*

$$\tilde{x}_{ij,logaritmic} = \log(x_{ij} + 1)$$

A second step was performing Pareto scaling [80]. Pareto scaling is a form of scaling based on the data dispersion. It uses the square root of the standard deviation as the scaling factor. The formula is shown in Equation 9 and also includes centring the abundances around 0 by subtracting the mean per metabolite. It reduces the relative importance of large values but keeps the data structure partially intact. Large fold changes are decreased more than small fold changes, which makes them less dominant [78].

*Equation 9 Pareto scaling. Where $\tilde{x}_{ij,log-pareto}$ is the normalized metabolite abundance of metabolite i in infant j, $\tilde{x}_{ij,logarithmic}$ is the logarithmic transformed metabolite abundance of metabolite i in infant j, $\bar{x}_i$ the mean abundance of metabolite i and $s_i$ the standard deviation of metabolite i. .*

$$\tilde{x}_{ij,log-pareto} = \frac{\tilde{x}_{\ ij,logarithmic} - \bar{x}_i}{\sqrt{s_i}}$$

The effects for the normalization were shown during the exploratory data analysis.

#### 2.2.2.3 Exploratory data analysis

Prior to performing any statistical analysis, an exploratory data analysis was conducted. This was done to gain an insight into the data and decide upon the statistical analysis used to find differentially expressed metabolites. A set of visualization techniques were used.

The first visualizations made were the effects of the normalizations using boxplots per sample for each of the data sets prior and after normalization.

Next, a PCA was conducted [67]. If clustering within children takes place, this has to be accounted for during the statistical analysis. Additionally, a biplot was constructed for the UPLC data to see which metabolites carry a lot of information. A biplot overlays the PCA plot where the vectors are the projected variables.

This clustering was further investigated by constructing a heat map with dendrograms. The distances were calculated using the Euclidean distance and linkage was based on Ward's minimum variance method [68]. The dendrograms resemble the clustering of metabolites and samples.

#### 2.2.2.4 Statistical analysis for differentially abundant metabolites

Literature proposes a wide variety of tests to find differentially abundant metabolites ranging from simple tests to more complex models such as parametric, semi-parametric and non-parametric approaches. The parametric methods make distributional assumptions whom have to be met in order to be valid. The semi-parametric and non-parametric methods are robust to these distributional assumptions but require a larger amount of samples per group of 10 to 15 samples per group. Literature suggests that the best approach is using a wider variety of methods and compare if the results are similar to make the inference more robust [81]. A selection of techniques were applied such as a Wilcoxon rank-sum test [82], two-sample t-test, ANOVA [69] and a linear mixed model. Similar results were obtained and one method was reported.

The Wilcoxon rank-sum test was elaborated upon during the report due to the fact that during the exploratory data analysis, highly skewed metabolite distributions were observed. Unlike the microbiome data, no clustering of infants was present. The Wilcoxon rank-sum test is a non-parametric test equivalent to a t-test. Due to the Wilcoxon rank-sum test being a non-parametric test, it makes no distributional assumptions. This solves the problem of the skewed distributions and sparseness in the metabolomics data. It is used to compare two independent groups. In this case, the abundance of metabolites before the introduction of solid food and after the introduction of solid. The test is based solely on the order in which the observations from the two samples fall. Each observation is ordered and ranked from smallest to largest. The ranks for each sample is summed and an exact p-value is calculated. The hypothesis is given in Equation 10.

*Equation 10 Hypothesis test for differentially abundant metabolites. Where before means the sum of ranks before the introduction of solid food and after means the sum of ranks after the introduction of solid food. Done for each metabolite j.*

$$H_o: Before_j = After_j$$
$$H_a: Before_j \neq After_j$$

A disadvantage of using a non-parametric test such as the Wilcoxon rank-sum test is that they are usually less powerful compared to their parametric counterparts. However, as mentioned before, the parametric assumptions must hold for those tests to be used. But due to the lower power, it is less likely to reject hypotheses. In order for the non-parametric test to be valid, a sample size of 10 to 15 per group is required. This requirement was satisfied for each of the metabolomics data sets.

The change in abundance of metabolites was deemed to be significant based on a significance level $\alpha = 0.05$. Correction for multiplicity and to control the false discovery rate was done using the Benjamini-Hochberg procedure [76]. Results were visualized using volcano plots.

### 2.2.3 Association study between the microbiome and metabolome

#### 2.2.3.1 Principle

A correlation analysis was performed in order to investigate the association between the microbiome and metabolome. This was done using Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO). DIABLO is a multi-omics methods that identifies key omics variables during the integration process and discriminate phenotypic groups [83].

DIABLO is an extension of the sparse Generalized Canonical Correlation Analysis (sGCCA) to a classification or supervised framework. sGCCA is a multivariate dimension reduction technique using the singular value decomposition, selecting correlated variables from several omics datasets. sGCCA maximizes the covariance between linear combinations of variables and projects it into a smaller dimensional space, spanned by the components. The selection of correlated variables is done using a L1-penalty. A L1-penalty will minimize the residual sum of squares while setting many of the parameters equal to zero, shown in  Equation 11 [83, 84].

*Equation 11 L1-penalty.*

$$minimize \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_j \beta_j x_{ij}\right)^2 subject\ to \sum_j |\beta_j| \leq t\ with\ t \geq 0$$

DIABLO is an extension of sGCCA to the classification framework and can also be referred to as multiblock sparse Partial Least Squares Discriminant Analysis (sPLS-DA). sPLS-DA performs a variable selection and classification in a one-step procedure using the L1-penalty from Equation 11 [83, 84]. Partial Least Squares (PLS) is a supervised alternative for dimension reduction. It identifies a new set of features $Z_1, \dots, Z_M$ that are linear combinations of the original ASVs or metabolites. A linear model is fit via least squares using these M new features. It is a supervised approach as it identifies the ASVs or metabolites correlated the most to the response, the introduction of solid foods. The equation to compute these new features is given in Equation 12. PLS computes the first direction $Z_1$ by setting each $\phi_{j1}$ in Equation 12 equal to the coefficient from the simple linear regression of Y onto $X_j$, which is proportional to the correlation between Y and $X_j$ [85]. DIABLO requires hyperparameter tuning such as a covariance matrix, the amount of principal components and the optimal amount of features, namely ASVs and metabolites.

*Equation 12 Partial Least Squares feature equation.*

$$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$$

Even though DIABLO doesn’t make any assumptions about the distributions of the microbiome and metabolome data, it still requires the data to be normalized and pre-processed prior to performing DIABLO. The appropriate field-specific methods can be used. Hence, the microbiome data was filtered and CLR-transformed. The metabolome data was transformed using the natural logarithm transformation and pareto scaling. Additionally, each ASV and metabolite was centred and scaled internally [83]. The outcome is denoted by Y, the introduction of solid foods being either before or after. DIABLO can only work with samples where all methods were applied. This means that only 31 samples were used as shown in Figure 1. Infant “P” was excluded from the association study and infant “T” had no measurement before the introduction of solid food.

#### *2.2.3.2 Hyperparameter tuning*

##### 2.2.3.2.1 Covariance matrix

A covariance matrix C had to be specified with the dimensions of the amount of data sets used in the analysis. In the case of the current research, this was a 4 x 4 matrix. The values of the covariance matrix C range from 0 to 1 indicating the association between the data sets. A null and full design of the covariance matrix is shown in Equation 13. The covariance matrix C can be determined using either prior knowledge or a data-driven approach. A correlation between the microbiome-metabolome has been investigated earlier and described in literature [36, 37]. However, neither of these studies were performed during infancy at the time of introduction of solid foods with intensive sampling such as the LucKi-Gut study. This was the motivation to choose the data-driven approach. The data driven approach is done using PLS that models the pairwise association between each of the omics datasets and looking at the correlation between the first component of each of the omics data sets [86]. The diagonal was set equal to 0 in order to make sure the algorithm doesn't compute a relationship for a data set to itself.

*Equation 13 Examples of covariance matrices for DIABLO.*

$$C_{null} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \; and \; C_{full} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Based on the results of the correlations between the first components, of each omics dataset, an appropriate covariance matrix was determined.

#### 2.2.3.2.2 Principal component tuning

An initial blocked sparse PLS-DA was fitted afterwards using an arbitrary amount of principal components, 25. To evaluate the most optimal amount of principal components, cross-validation (CV) techniques can be used. DIABLO offers two CV options. The first option would be to use Leave One Out Cross-Validation (LOOCV). LOOCV is a cross-validation approach where just a single sample is held out for validation while all other samples are used for training. This is done by iterating over each sample. The final result is then calculated by taking the mean of all the individual evaluations. LOOCV tends to have a higher bias towards the dataset used to build the model. This can be relativised by the fact that the model will not be used for prediction and the goal is to find the best possible relationships between the ASVs and metabolites present in the data set. The second option would be to use repeated K-fold CV. During repeated K-fold CV, a different split of the data set into K-folds can be implemented and repeated over a set amount of times. A large amount of repeats is required because the results vary largely based on the splits. With a large amount of repeats, this allows for a better generalization of the model. Since only a small data set is available (n = 31), the repeated K-fold CV requires a very large amount of repeats in order to give consistent results. DIABLO requires a large amount of computational power and using a large amount of repeats ends up costing a considerable amount of time in order to evaluate the model. It was chosen to use the LOOCV to evaluate the performance of the model using different amounts of principal components. This was also advised by the creators of the package when using small data sets.

Different distance measures were used in order to assign the classes, either before or after the introduction of solid foods, of the test observations during the LOOCV.

- The maximum distance: The maximum distance assigns the class with the largest predicted score. It performs quite well when using only a single data set. This is an important point to consider since 4 data sets are used in the current setting.
- The centroid distance: The centroid distance computes the centroid ($G_k$) per component for each class K using the training samples belonging to that class. Afterwards, for each prediction the Euclidean distance to $G_k$ is calculated. The centroid of the class that minimizes this Euclidean distance is assigned to that sample. It is more robust compared to the maximum distance and less susceptible towards outliers in the training data.
- The Mahalanobis distance: The Mahalanobis distance is similar to the centroids distance. Instead of using the Euclidean distance, the Mahalanobis distance is used. This distance measure takes into account the correlation between the components [87].

The performance was measured via the overall misclassification error rate (ER) and balanced error rate (BER). BER is more appropriate in case of an unbalanced number of samples per class. It calculates the average proportion of wrongly classified samples in each class. This is weighted by the number of samples in each class. In the current research setting, these classes will be either before or after the introduction of solid foods. The BER is less biased towards the class containing most samples [88]. The performance of the model is visualized in order to properly evaluate the most optimal amount of principal components. All distance measures and both error rate measures were included.

###### 2.2.3.2.3 Feature selection

A final hyperparameter being tuned is the amount of features, ASVs or metabolites being kept in the model. This was done by performing a grid search over all microbiome and metabolome datasets selecting the key contributors. The process was done for each of the principal components selected during the principal component tuning. This yields a significant amount of models to be fit. To speed up the process, the grid search started with taking larger steps to filter out most of the insignificant ASVs and metabolites contributing less information. Afterwards, the grid search was done in smaller steps to select the final important features. To evaluate the different combinations of features, LOOCV or repeated K-fold CV can be used again. Similarly to the principal component tuning, LOOCV was chosen.

The algorithm will identify the key contributor ASVs and metabolites per principal component. A final DIABLO model was ran and results were visualized.

#### *2.2.3.3 Results*

A first visualization made was at a sample level using a correlation plot. The n$^{th}$ principal component of each data set was plotted against each other containing the samples. The samples were coloured based on the introduction of solid foods, either before or after the introduction of solid foods. Additionally, 95% confidence ellipses were added to visualize if the principal components are able to discriminate between both categories of solid food introduction. This was done for all principal components.

Next, visualizations were made at feature level. A first visualization was made by creating a circos plot. A circos plot can be used to gain an idea on how the selected features from each data set relate to each other. This was done for all principal components simultaneously. Each data set has its own colour. A threshold for the correlation between features was set and are located within the circular plot. On the outside, lines were added, showing the expression of that feature for both before and after the introduction of solid foods [89]. A second and final visualization was a heatmap of the correlations between the different metabolites and ASVs selected by the final DIABLO model. The Euclidean distance was used to compute the distances.

# 3 Results

The results are discussed methodically for each of the data sets separately followed up by the results of the correlation analysis between the microbiome and metabolome.

## 3.1 Microbiome data

### 3.1.1 Exploratory data analysis

A first step in the exploratory data analysis was oriented towards the effects of filtering. A first filter being applied was based on the prevalence of the taxa. A taxon had to be present in at least 5% of the samples in order to not be filtered out. This narrowed down the original 8787 taxa to 168 taxa, which corresponded to 1.91% of the original amount of taxa. A second filter being applied was a filter based on the relative abundances of the taxa. A taxon had to have a relative abundance of 0.01% in order to not be filtered out. This resulted in a final number of 121 taxa that were retained for analysis, equal to

1.38% of the original amount of taxa. The effects of filtering were visualized using density plots displaying the distributions of the library sizes (Figure 6). The distribution of the filtered library sizes are shown in blue, the results of the unfiltered library sizes are shown in red. An unwanted effect of filtering would be a major shift of the filtered library sizes towards the left of the plot or towards the smaller library sizes. However, results show that the distribution of the filtered library sizes remained almost identical to the distribution of the unfiltered library sizes. This indicated that while more than 98% of the taxa were removed, only a very limited amount of data were removed (i.e., only very sparse taxa were removed).

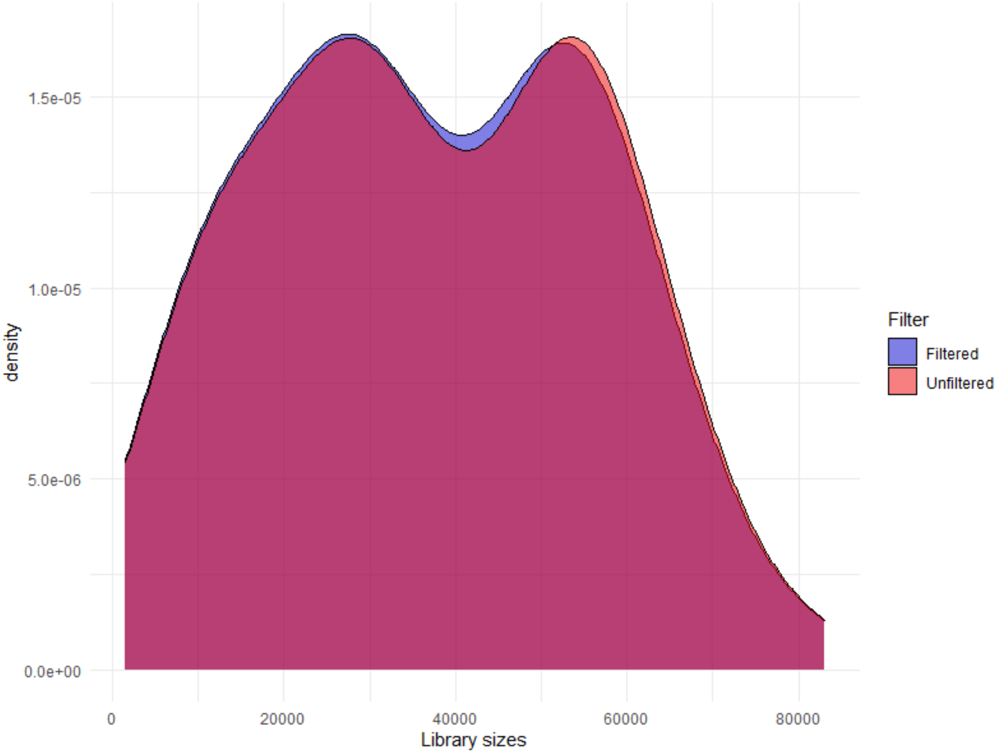

*Figure 6 Effects of filtering displayed using density plots. Red shows the distribution of the unfiltered library sizes and blue the distribution of the filtered library sizes.*

A first start in exploring the filtered, normalized microbiome data was performing a principal component analysis. The principal component analysis uses the Aitchison distance as distance measure. A scree plot was provided in the appendix (Figure 23) to illustrate the variance explained by each principal component. The first principal component explained 22.4% of the variance present in the entire normalized and filtered microbiome data set. The second principal component explained 15.1% of the variance, adding up to a total of 37.5% (Figure 7). Colours denote the different infants in the Lucki Gut study and shapes the introduction of solid foods. It is observed that samples from the same infants cluster together strongly and are different from each other. A second observation made is that sample before and after the introduction of solid food were mixed within an infant. This indicated that, using a PCA, samples before and after the introduction of solid foods were not able to be separated. A last observation made was that the ellipses from the PCA from before and after the introduction of solid food almost overlapped. Making differentiation between both food types not possible.

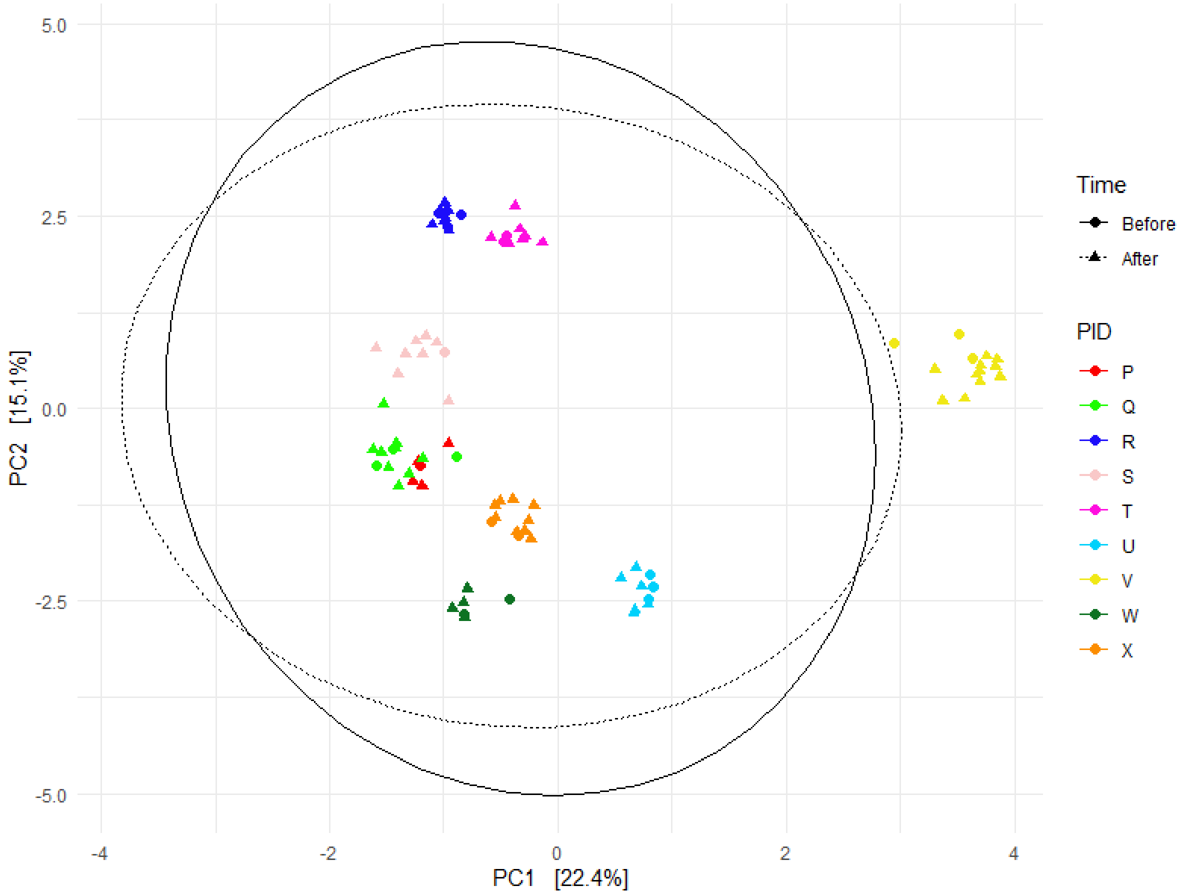

*Figure 7 Principal component analysis of the microbiome data. The distance measure used is the Aitchison distance. Colours denote the different infants present in the Lucki Gut study. Shapes denote before and after the introduction of solid food.*

Identically to the PCA, a heat map showed that samples from infants cluster together (Figure 8). This was illustrated by the dendrograms which identified 9 clusters that were linked to the 9 individual children. The heat map added an additional layer of information. A block-like structure was observed within infants, showing ASVs specific to each infant. There were also a set of taxa present in all infants. The bar plot indicating the sampling times with relation to the introduction of solid foods showed the same mixed pattern as observed in the PC, i.e., no clear separation was visible based on the introduction of solid foods within an infant and/or between all infants.

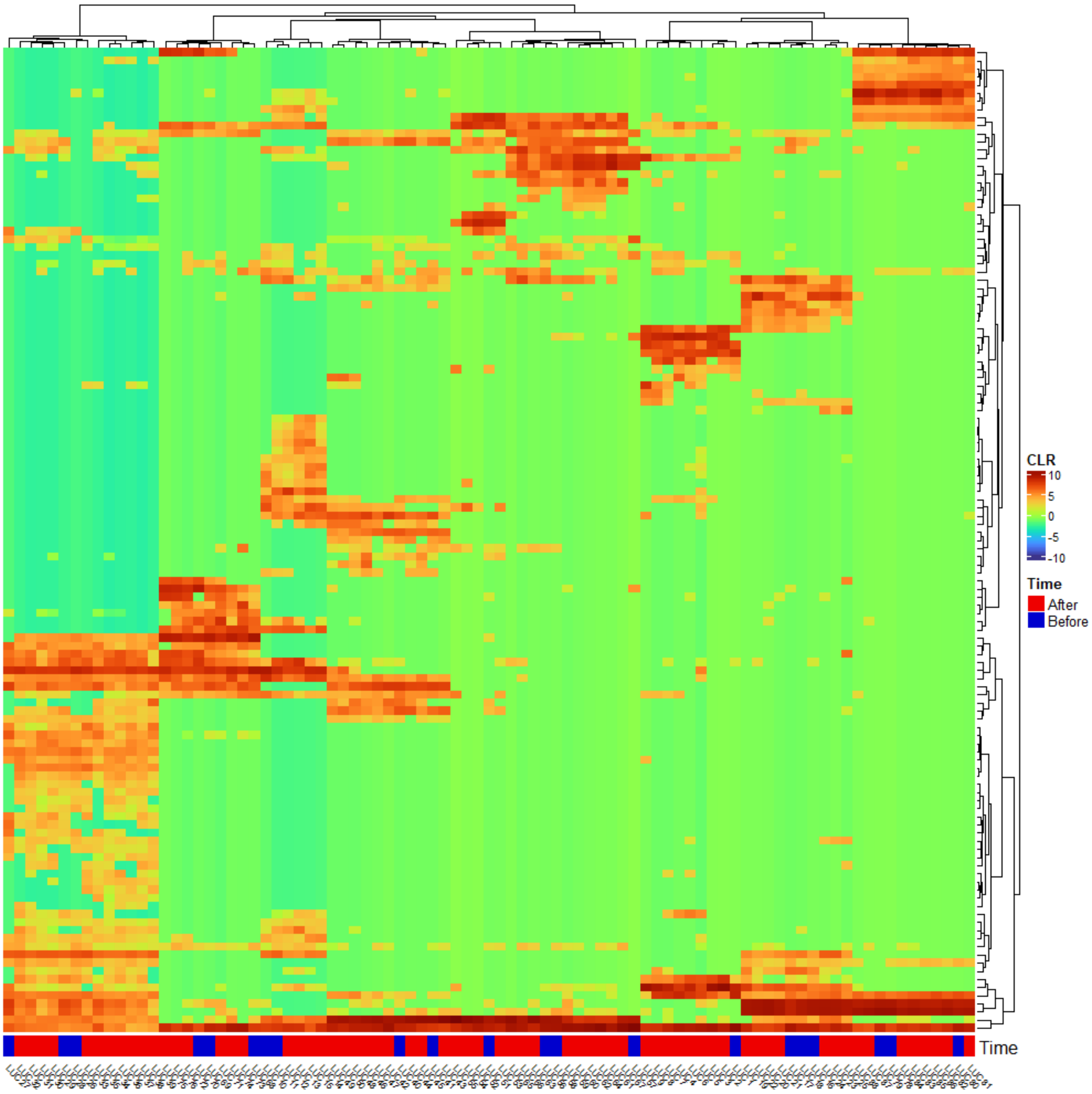

*Figure 8 A heatmap of the normalized microbiome data. Rows indicate the different taxa, columns indicate the different samples. Clustering with the dendrograms is done using the Aitchison distance and Ward's minimum variance method. The bar plot at the bottom indicates the sampling times, either before (blue) or after (red) the introduction of solid foods.*

When examining the relative abundance at the family level (Figure 9, Figure 24 - Figure 32), several observations were made. There were a set of dominant families which were commonly present both before and after the introduction of solids foods. Among these families were *Bacteroidaceae*, *Bifidobacteriaceae, Enterobacteriaceae*. *Veillonellaceae* was a family commonly present at lower abundances in a large proportion of the samples and seemed to increase slightly after the introduction of solid foods. It was also observed that in some infants the gut microbiome composition changed after the introduction of solid foods. In infant "S", an increase of *Lachnospiraceae* was observed. In infant "R", an increase of *Porphyromonadaceae* was also apparent. Lastly, in infant "V", a slight increase of *Clostridiaceae* was seen. Most other families remained somewhat constant within an infant before and after the introduction of solid foods. No systematic trends of increasing or decreasing families was observed.

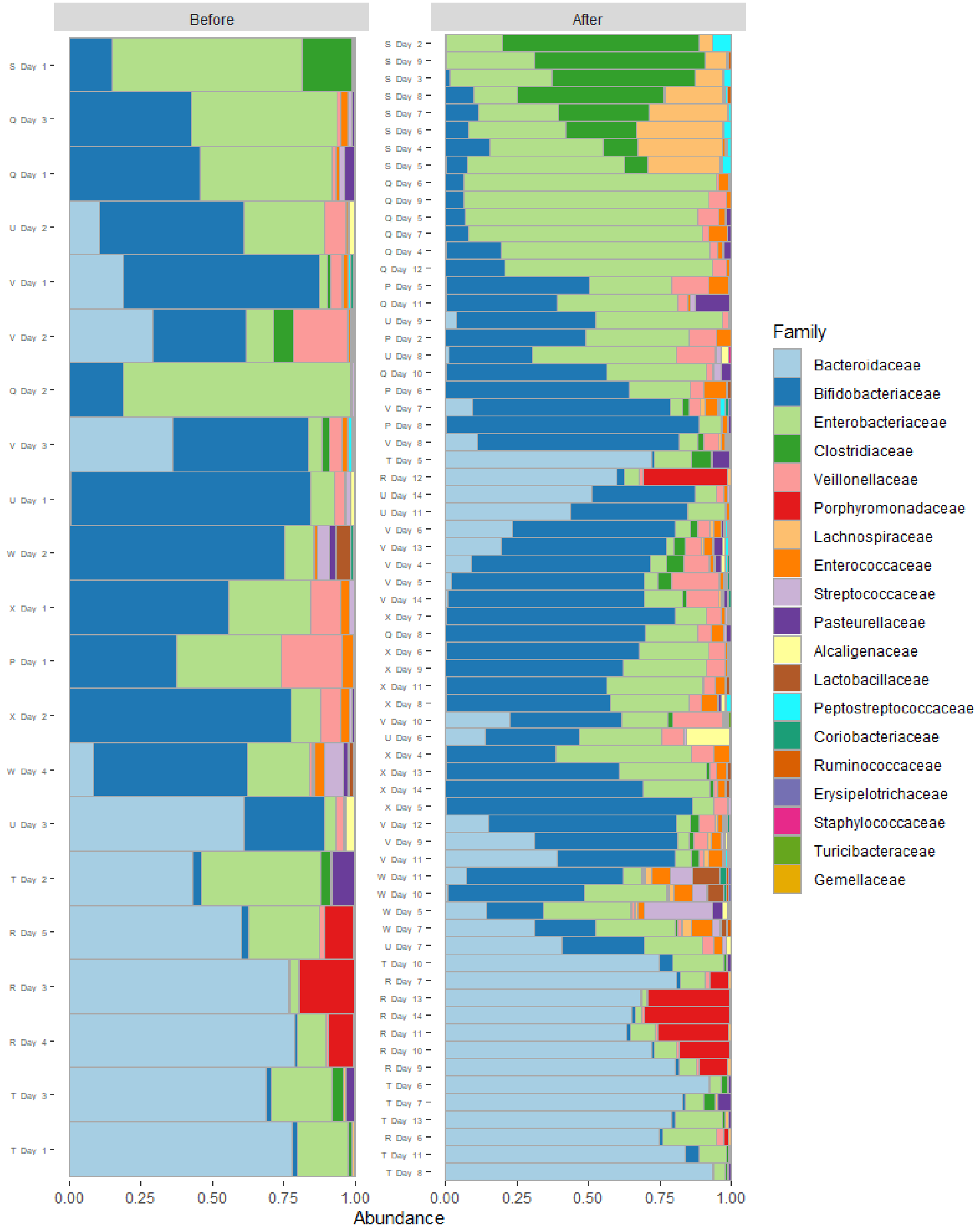

*Figure 9 Bar plot of the relative abundances in samples at family rank. Stratified over the introduction of solid foods.*

Lastly, alpha diversity measures were calculated in order to visualize other covariates of interest. The observed richness, inverse Simpson index and Shannon index were used. No increasing trends were witnessed for any of the alpha diversity indices (Figure 10). However, for some individual infants increasing trends could be witnessed for the alpha diversity indices with increasing age (Figure 11). This could be observed for infants “R”, “S”, “U” and “W”. However, decreasing trends were also observed such as in infant “P”. Other children oscillated around the same values of the alpha diversity indices. Other covariates such as gender (Figure 33), length (Figure 34), weight (Figure 35) and disease status (Figure 36) were also explored. Gender showed that female infants had higher alpha diversities compared to males. The length, weight and disease status showed no increasing or decreasing trends.

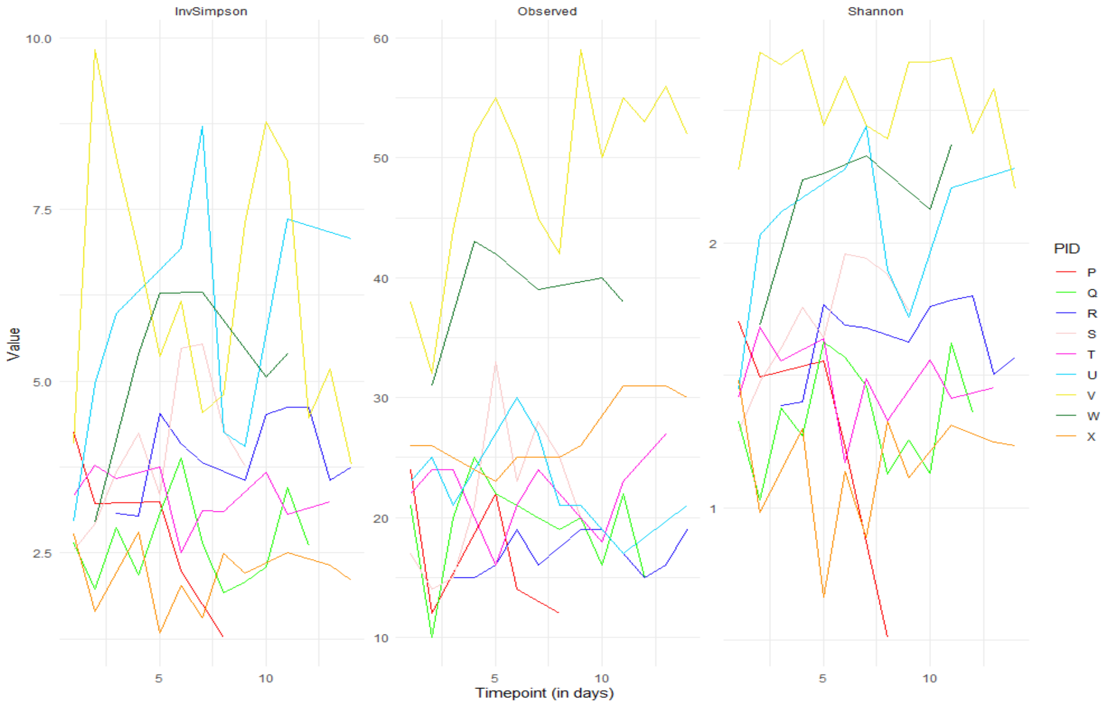

*Figure 10 Longitudinal trends of alpha diversities of time per infant. The different colours resemble different infants.*

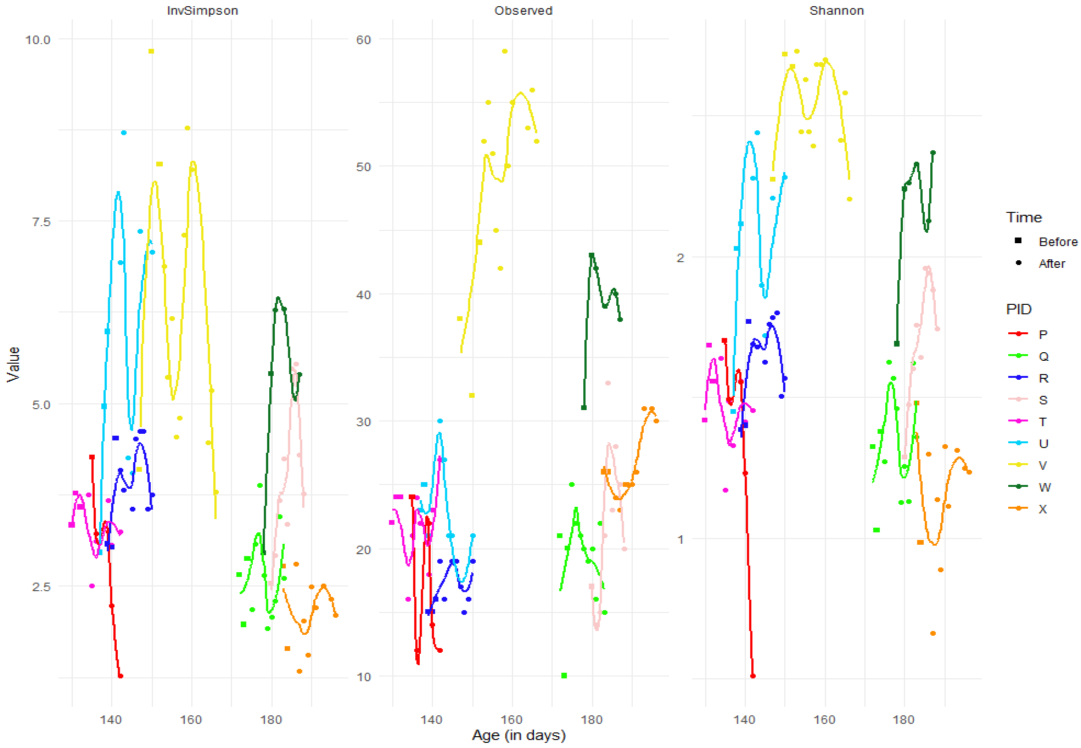

*Figure 11 Alpha diversity indices in function of age per infant. The different colours resemble the different infants. The different shapes denote the sampling times.*

### 3.1.2 Statistical analysis for differentially abundant ASVs

The statistical analysis of the differentially abundant ASVs for solid food introduction resulted in 7 (5.79%) statistically significant ASVs prior to any correction for multiple hypothesis testing. ASVs for *Haemophilus* parainfluenza, *Streptococcus luteciae*, *SMB53* sp., *Streptococcus sp.*, *Enterobacter* cloacae, *Staphylococcus* aureus and *Sutterella* sp. were found to be differentially abundant prior to correction for multiple testing. After correction for multiple hypothesis testing, no statistically significant results were found (Figure 12). For the age of infants, 69 differentially abundant ASVs were found prior to correction for multiple hypothesis testing. After correction for multiplicity, 60 ASVs (49.59%) were found to be differentially abundant (Figure 37). The full summary of the results can be found in the Appendix (Table 11).

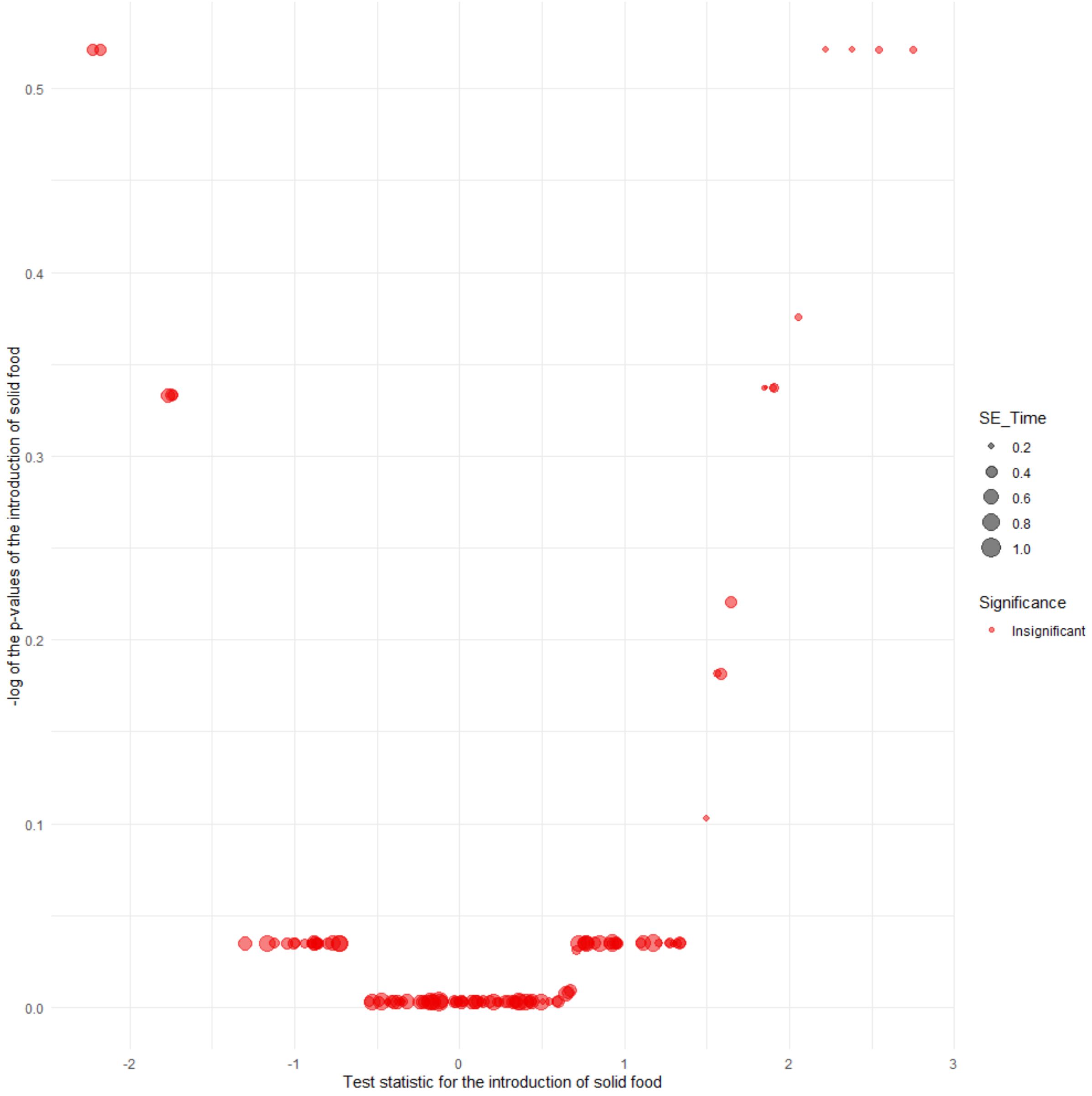

*Figure 12 Results for the introduction of solid foods of the microbiome data after adjustment of multiplicity. The x-axis denotes the test statistic for the introduction of solid food and the y-axis the negative logarithmic transformed p-values after adjustment of multiplicity. Colours denote the significance of the results, red being insignificant. The size of the dots denote the standard error of the test statistic.*

## 3.2 Metabolomics data

### 3.2.1 Exploratory data analysis

The metabolite distributions per sample for the NMR metabolites prior to any normalization steps were highly skewed (Figure 13). Metabolites with a different order of magnitude were observed. Potential technical and experimental variation might be present between the different samples. Similar figures can be found for the DIMS (Figure 38) and UPLC metabolites (Figure 39) in the Appendix. For each of the metabolite detection methods, the same distributions were observed.

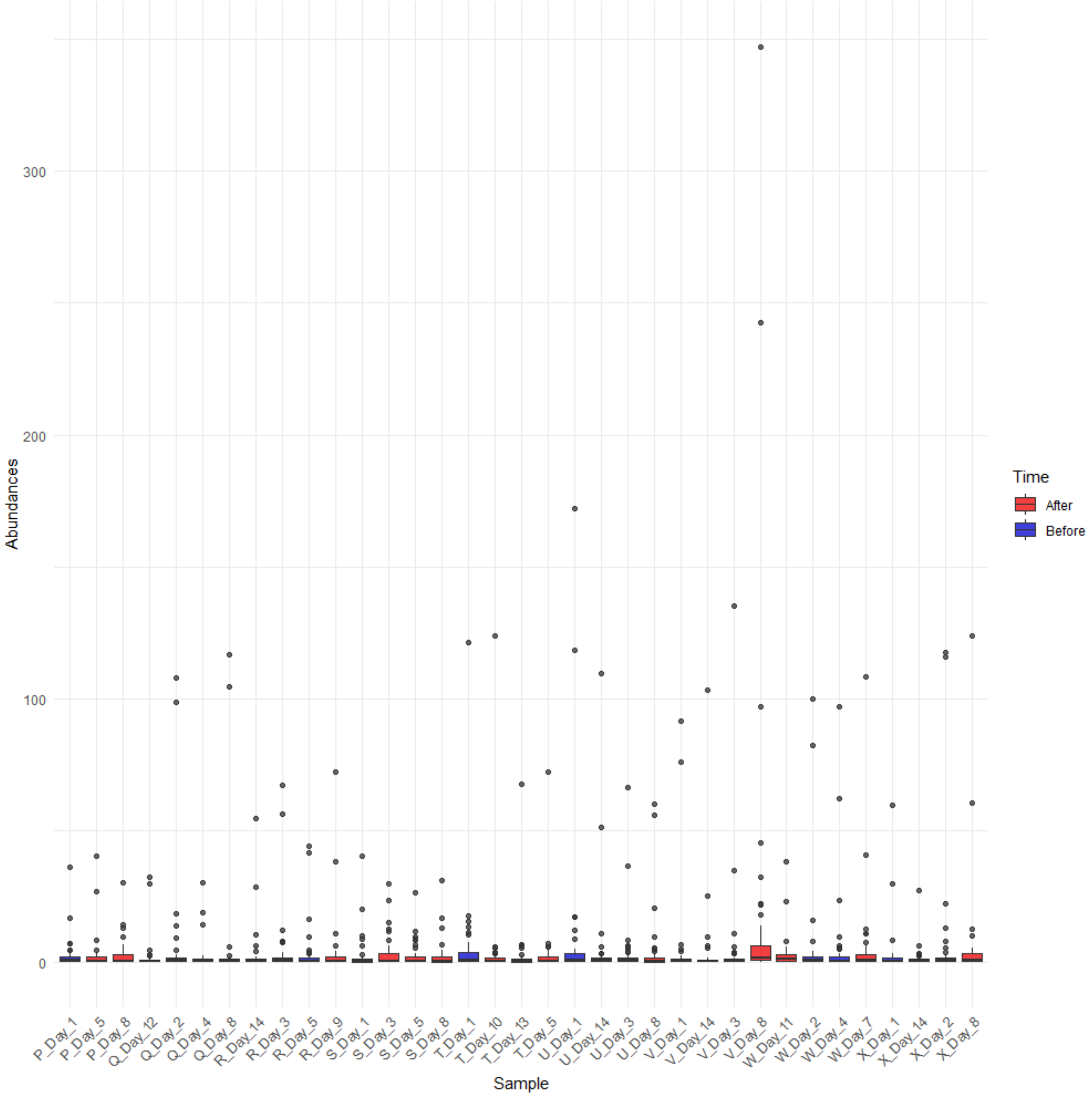

*Figure 13 Non-normalized metabolite distributions per sample for NMR metabolites. The colours indicate the sampling times, either before (blue) or after (red) the introduction of solid foods.*

Upon a natural logarithm and pareto scaling, the NMR metabolite distributions per sample no longer showed any large outliers and were approximately centred around 0 (Figure 14). The normalization accounted for the metabolites with different orders of magnitude as no large abundances were observed. Potential technical and experimental variation was removed. The data was properly normalized for the exploratory and statistical data analysis. The results for the normalization of the DIMS (Figure 40) and UPLC (Figure 41) metabolites are shown in the Appendix and were similar to the results of the NMR metabolites.

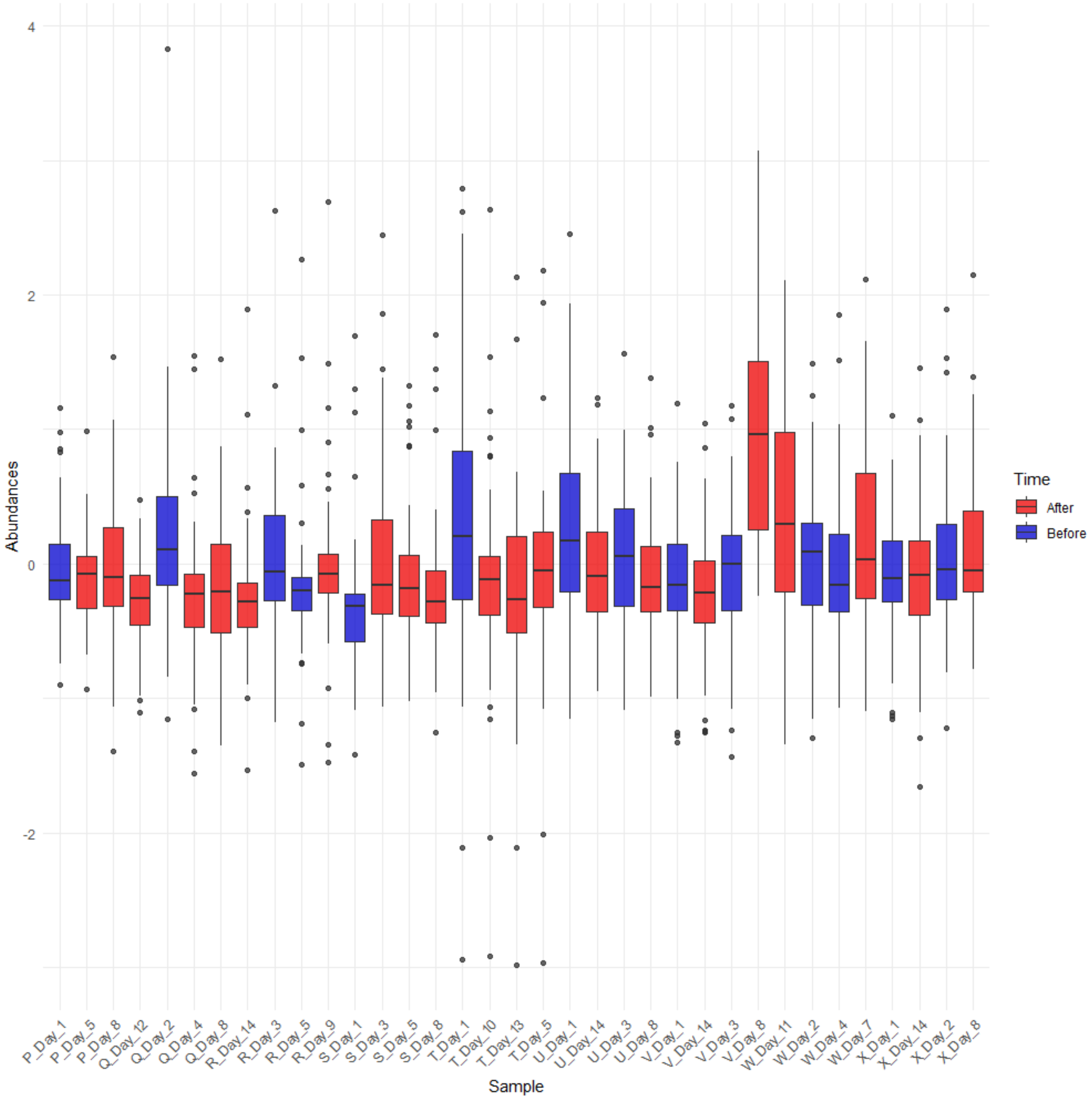

*Figure 14 Normalized metabolite distributions for NMR metabolites after a logarithmic-pareto normalization. The colours indicate the sampling times, either before (blue) or after (red) the introduction of solid foods.*

A first step in the exploratory data analysis was conducting a PCA for each of the metabolome data sets. For the NMR data, the first two principal components described 34.7% of the total variability in the data set (Figure 15). Based on the variability, there was no real distinction between samples before or after the introduction of solid foods. A second observation was that samples from the same infant do not cluster together unlike in the microbiome data. The results for the PCA of the DIMS (Figure 42) and UPLC (Figure 43) metabolites are shown in the Appendix. For the DIMS data, the first two principal components explained 47.2% of the total variability in the data. It could be observed that based on the variability, samples after the introduction of solid foods seemed to cluster together, while samples before the introduction of solid foods seemed to have a higher variability. Similar to the NMR data, the samples from the same infants did not cluster together. For the UPLC data, 37.3% of the total variability was explained by the first two principal components. Here, the opposite was seen. The samples before the introduction of solid foods seemed to cluster together while the samples after the introduction of solid foods had a higher variability. Again, samples from the same infant did not cluster together. The PCA on the UPLC data showed 4 samples with a lot of variability in the first principal component towards the left. This was to rationale to construct a biplot of the UPLC data to know which metabolites drive this variability.

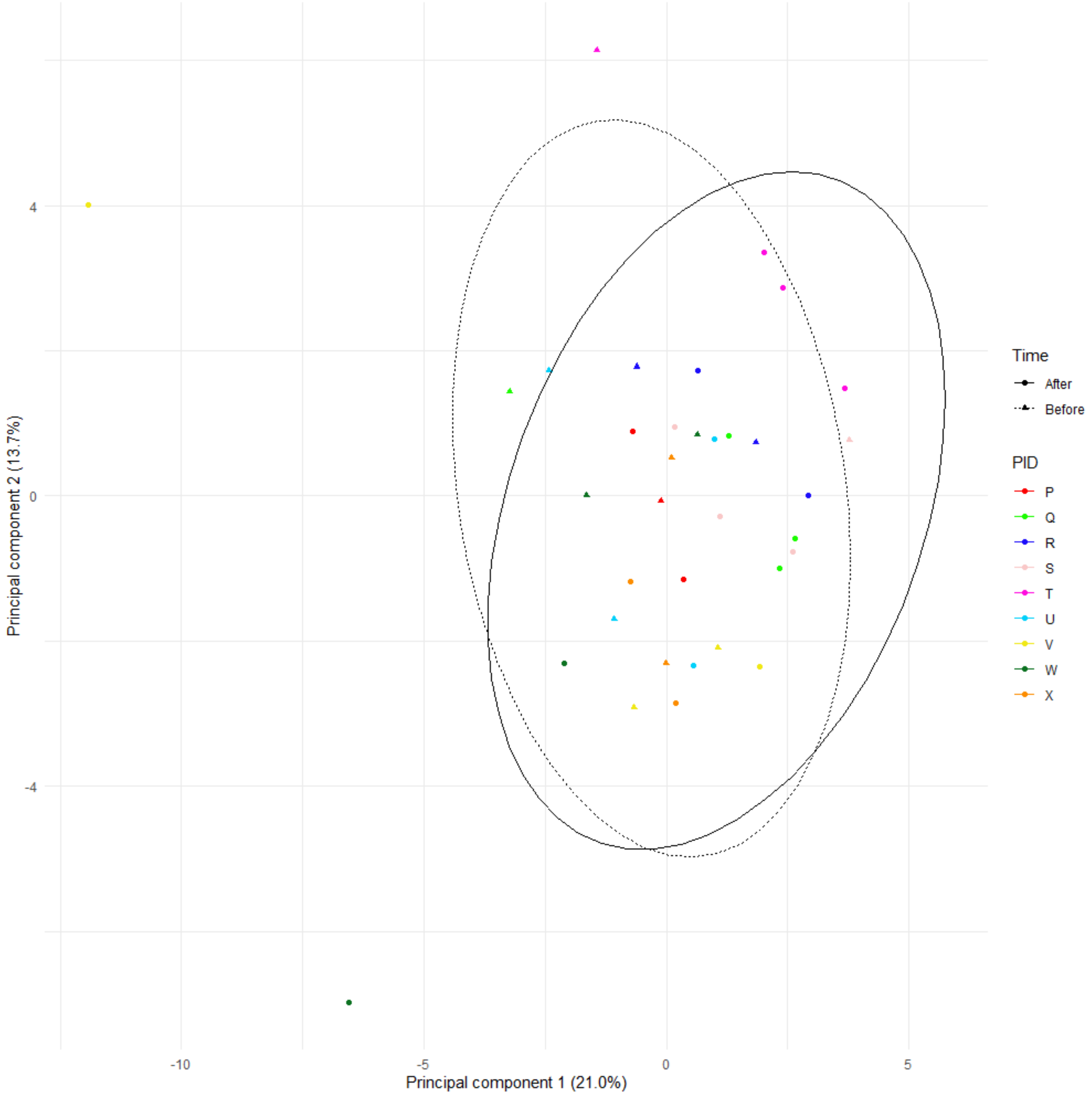

*Figure 15 PCA of the NMR metabolome data set. Colours denote samples from the same patient. The ellipses and shape of the dots describe the data from either before or after the introduction of solid foods.*

A biplot was constructed in order to find the metabolites causing a set of observations to be shifted to a large negative value of the first principal component of the UPLC data. The biplot shows the top 5 metabolites (Figure 16). The metabolites identified were:

- Taurocholic acid: A bile-acid taurine conjugate of cholic acid that usually occurs as the sodium salt of bile. It is involved in the emulsification of fats [77].
- Glycoursodeoxycholic acid-3-sulfate: A bile-acid glycine conjugate derived from ursoodeoxycholic acid. It is associated with neuroprotective properties [77].
- Tauroallocholic acid: A bile-acid taurine conjugate [77].
- Glycodeoxycholic acid-3-sulfate: A bile-acid glycine conjugate [77].
- Glycoallocholic acid

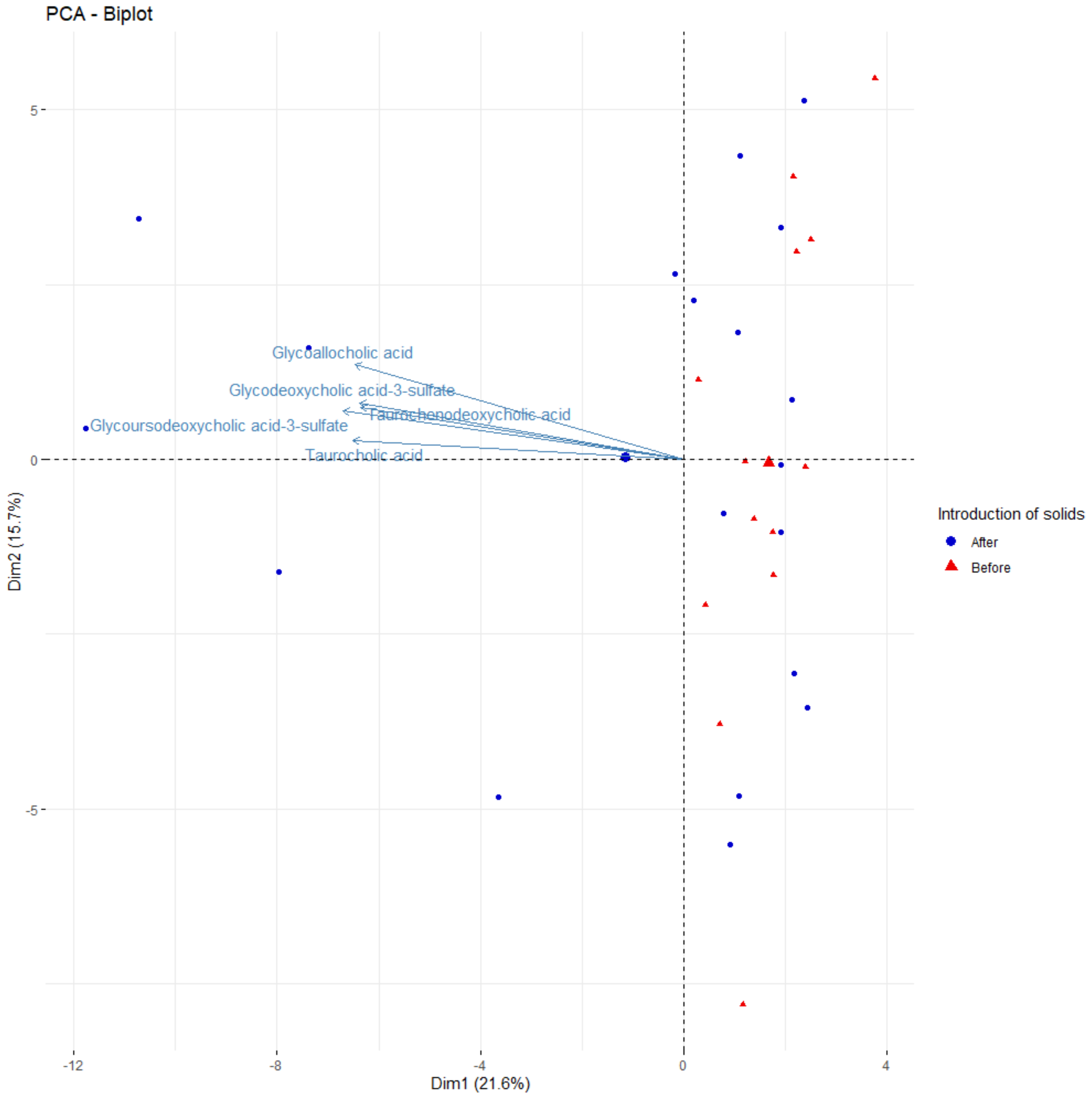

*Figure 16 Biplot of the UPLC data with the top 5 metabolites. The colours denote the sampling times, before the introduction of solid foods being red and after the introduction of solid foods being blue.*

A final exploratory tool used for each of the metabolite data sets were the construction of heatmaps. The results are shown for the NMR data set (Figure 17). Similar heatmaps were constructed for the DIMS data set (Figure 44) and UPLC data set (Figure 45) in the Appendix. Several conclusions could be drawn for each of the data sets. Samples from within an infant did not cluster together, they were mixed with each other. Similar conclusions were drawn from the PCA. The observations from before the introduction of solid foods and after the introduction of solid foods were mixed. Hence, there were no large changes in metabolite concentrations observed due to the introduction of solid foods. And particularly for the DIMS and UPLC data sets, only a key set of metabolites were present in the infants. Finally, a peculiar observation was made for the NMR data set. For infant “T”, the metabolites lactate, acetate and succinate were almost completely absent while this infant did receive breast feeding prior to the introduction of solid foods. Infant “T” did however have a higher abundance of pyruvate in comparison to the other infants. A similar pattern was observed for infant “R”, who had a lower abundance for lactate and a higher abundance for succinate.

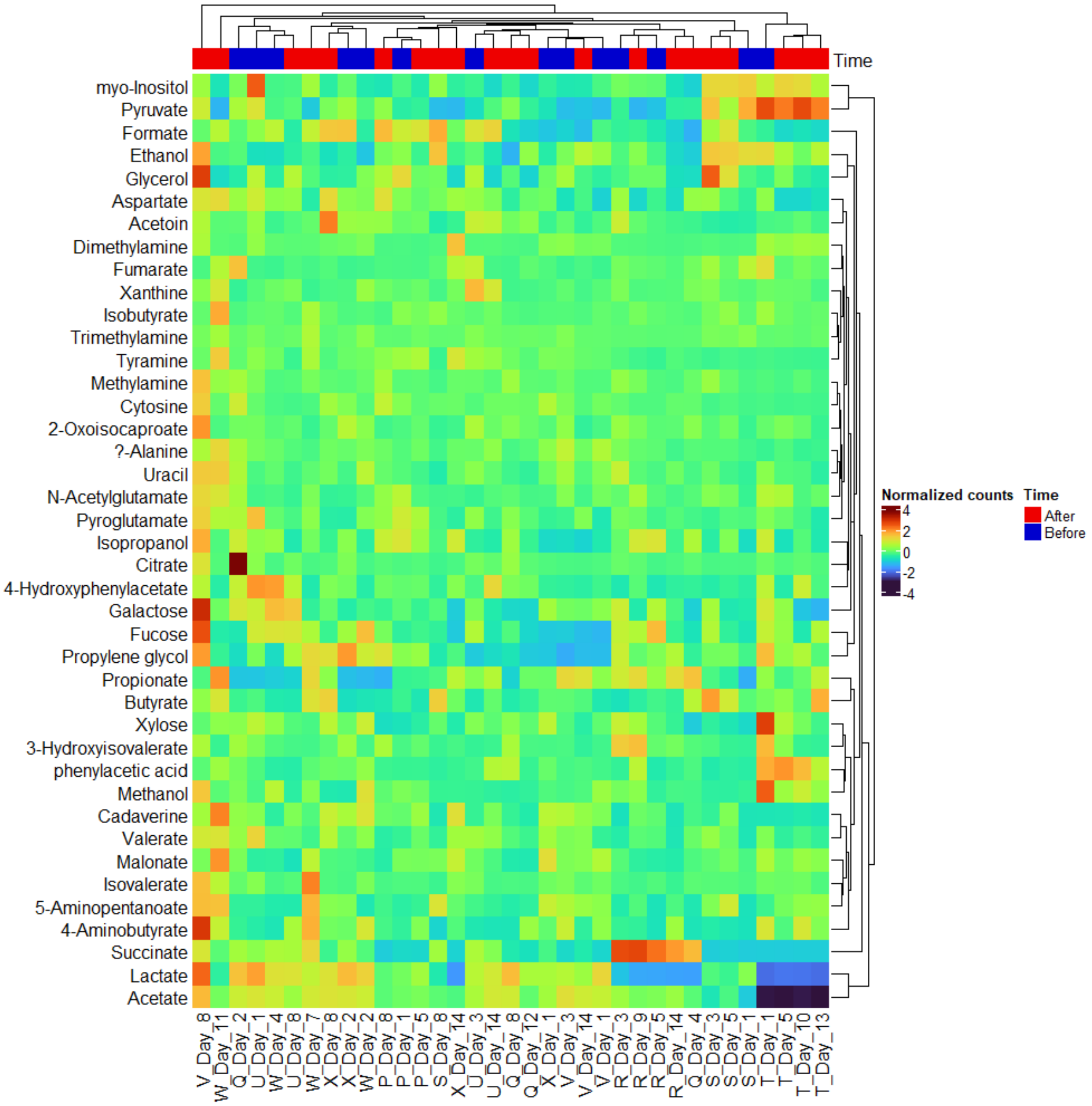

*Figure 17 Heatmap of the NMR data set. The normalized counts are shown. Dendrograms show the clustering of metabolites in the rows and samples in the columns. A barplot is constructed to indicate the sample period, before (blue) or after (red) the introduction solid foods.*

### 3.2.2 Statistical analysis for differentially abundant metabolites

The Wilcoxon rank-sum test was used to test for differentially abundant metabolites. Prior to adjustment for multiplicity, several metabolites in each of the data sets were found to be differentially abundant (Table 6). There were 4 differentially abundant metabolites in the NMR data (9.76%), 12 differentially abundant metabolites in the DIMS data (10.34%) and 13 differentially abundant metabolites in the UPLC data (18.57%). After adjustment for multiple hypothesis testing, no metabolites were found to be differentially abundant in either of the data sets based on the introduction of solid foods. The raw and adjusted p-values for each metabolite are provided in the Appendix (Table 12 - Table 14). The volcano plot for the NMR data showed no differentially abundant metabolites (Figure 18). The volcano plots for the DIMS (Figure 46) and UPLC (Figure 47) data sets are shown in the Appendix. Neither of the plots showed any significant results.

*Table 6 Wilcoxon rank-sum test results for differentially abundant metabolites after solid food introduction.*

| | NMR | DIMS | UPLC |
|---|---|---|---|
| ***Amount of metabolites*** | 41 | 116 | 70 |
| ***Amount of significant metabolites before adjustment for multiplicity*** | 4 (9.76%) | 12 (10.34%) | 13 (18.57%) |
| ***Amount of significant metabolites after adjustment for multiplicity*** | 0 (0.00%) | 0 (0.00%) | 0 (0.00%) |

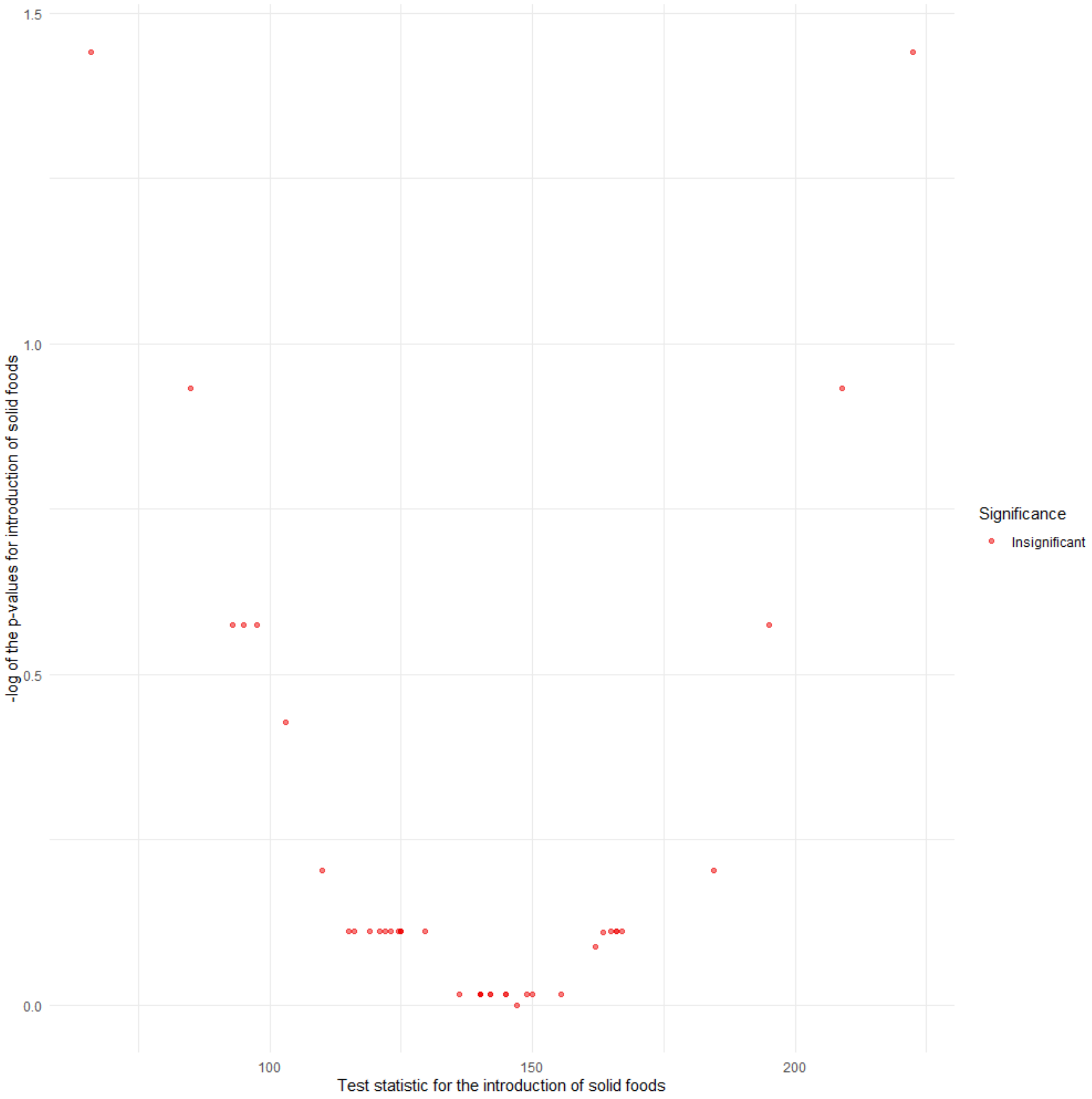

*Figure 18 Results for the differentially abundant metabolites after the introduction of solid foods in the NMR data. Results are shown after adjustment for multiplicity. The x-axis denotes the test statistic for the introduction of solid food and the y-axis the negative logarithmic transformed p-values after adjustment of multiplicity. Colours denote the significance of the results, red being insignificant.*

## 3.3 Association study between the microbiome and metabolome

### 3.3.1 Hyperparameter tuning

#### *3.3.1.1 Covariance matrix*

A covariance matrix was constructed by fitting a PLS model between each of the data sets and computing the correlation between the first principal components. The covariance matrix shown in Table 7 was constructed using this technique. The highest correlations were observed between the Microbiome and UPLC data sets (0.866391) and between the NMR and DIMS data sets (0.8399946). The lowest correlations were observed between the Microbiome and DIMS data sets (0.6923896) and between the NMR and UPLC data sets (0.6672782).

*Table 7 Covariance matrix for the final DIABLO model.*

| | Microbiome | NMR | DIMS | UPLC | Outcome |
|---|---|---|---|---|---|
| **Microbiome** | 0 | 0.8241704 | 0.6923896 | 0.866391 | 1 |
| **NMR** | 0.8241704 | 0 | 0.8399946 | 0.6672782 | 1 |
| **DIMS** | 0.6923896 | 0.8399946 | 0 | 0.7526392 | 1 |
| **UPLC** | 0.866391 | 0.6672782 | 0.7526392 | 0 | 1 |
| **Outcome** | 1 | 1 | 1 | 1 | 0 |

#### *3.3.1.2 Principal component tuning*

LOOCV was used in order to find the most optimal amount of principal components to use in the DIABLO model. The results were visualized in Figure 19. The centroids distance achieved the lowest classification error rates for both the normal error rate and the balanced error rate using 7 principal components. A larger amount of principal components could also be used as the same error rate was achieved. However, for computational purposes it was better to choose the lowest amount of principal components. In the essence of those computational purposes, it was chosen to select 3 principal components as the difference in error rates between 3 and 7 principal components was only minimal.

#### *3.3.1.3 Feature selection tuning*

Feature selection was used in order to select the key contributors, ASVs and metabolites, of each data set. The covariance matrix in Table 7 was used with 3 principal components. LOOCV was used in order to evaluate model. After several grid searches, the key contributors in each data set were narrowed down. Table 8 summarizes the amount of key contributors or features per data set and principal component. Noteworthy was the large amount of metabolites in the second principal component measured through DIMS (60) selected. All other data sets required a relatively small amount of features per principal component.

*Table 8 Key contributors or features per data set and principal component.*

| Data set | Principal component 1 | Principal component 2 | Principal component 3 |
|---|---|---|---|
| **Microbiome** | 10 | 8 | 5 |
| **NMR** | 19 | 21 | 5 |
| **DIMS** | 13 | 60 | 13 |
| **UPLC** | 7 | 8 | 6 |

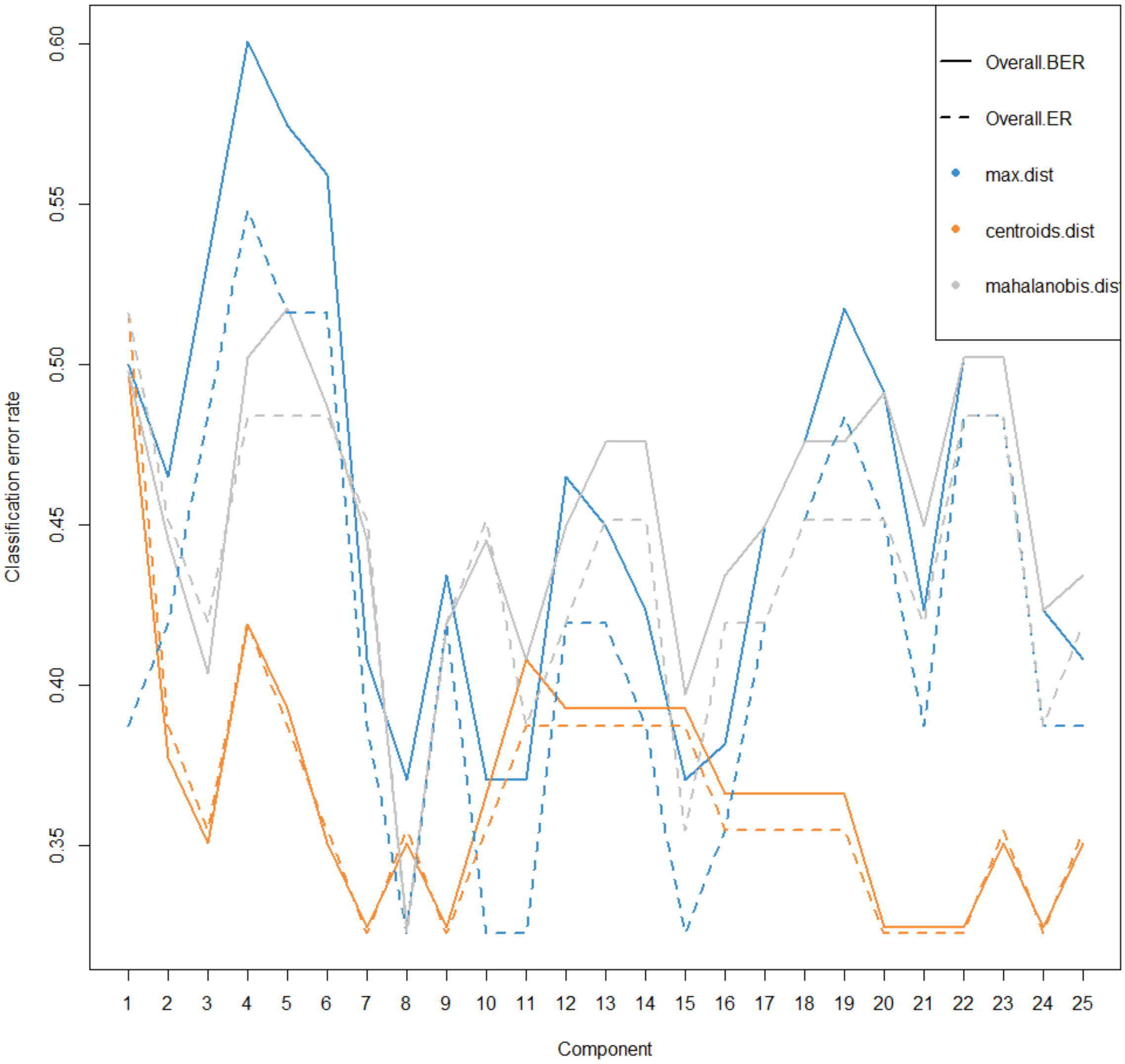

*Figure 19 Hyperparameter tuning of the principal components for the DIABLO model. The 3 distance measures are visualized with the maximum distance in blue, centroids distance in orange and Mahalanobis distance is grey. The solid lines is the classification error rate and the dotted line the balanced error rate.*

### 3.3.2 Results

A first visualization made was the correlation plot at sample level where the principal components for each data set were plotted against each other. The samples are coloured based on the introduction of solid foods, either before (red) or after (blue) the introduction of solid foods. Figure 20 shows the first principal component of each data set. Two combinations of data sets reached a correlation higher 0.8. The highest correlation of 0.82 was reached between the microbiome and NMR data. Between the NMR and DIMS data, a correlation of 0.8 was achieved. The lowest correlation of 0.67 was found between the first principal components of the DIMS and UPLC data. Based on the ellipses, it could be observed that a distinction is present between the samples from both groups, with an overlap always being present. For the first principal components of the microbiome and DIMS data, the ellipse of samples after the introduction of solid foods were located completely within the ellipse of samples before the introduction of solid foods. A correlation plot of the second principal component is shown in the Appendix (Figure 48). Only the correlation of the second principal component between the microbiome and DIMS data achieved 0.80. While all correlations with the UPLC data achieved a correlation lower than 0.70. It could also be observed that the ellipses were mostly overlapping between both groups of the introduction of solid foods. Lastly, the correlation plot between the third principal components was constructed and found in the Appendix (Figure 49). A large correlation was found between the microbiome and DIMS data (0.89) and between the DIMS and UPLC data (0.86). Only between the NMR and UPLC data, a correlation lower than 0.7 was observed (0.69). The ellipses of both groups always completely overlapped. However, it could be noted that in the third principal component, the ellipses of the group after the introduction of solid foods were larger compared to before the introduction of solid foods indicating a larger variability for the samples after the introduction of solid foods.

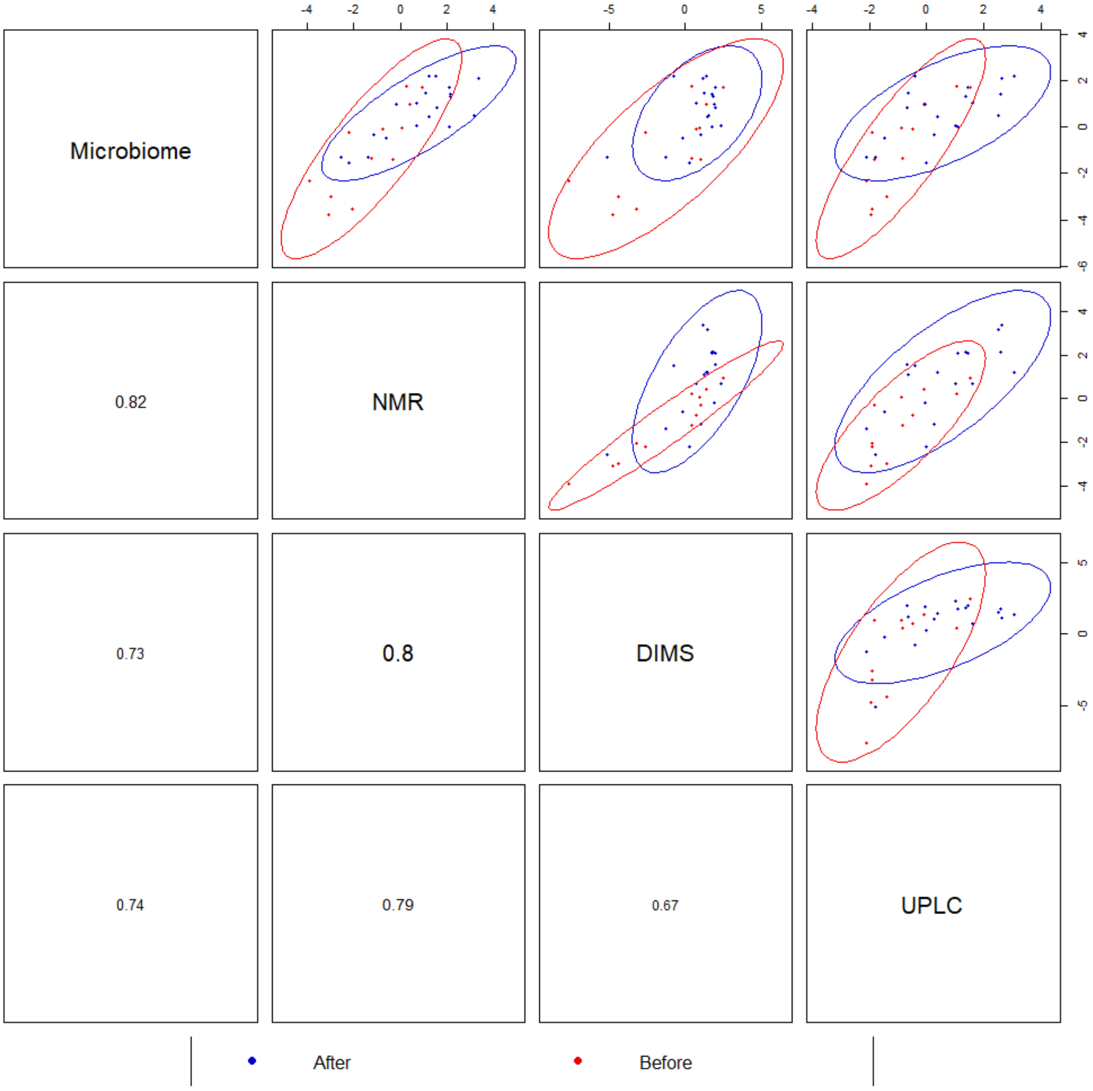

*Figure 20 Correlation plot of the correlation between the first principal component of each data set. Dots denote samples and the colour shows the sampling time, either before the introduction of solid foods (red) or after (blue). 95% confidence ellipses were added.*

Next, a circos plot was created (Figure 21). The circos plot shows all key contributors selected from each data set for all 3 principal components. Each data set is shown in a different colour. Inside the circle are lines present. These lines show the correlations larger than 0.7 or smaller than -0.70. On the outside of the circle, there are also two lines present. These lines show the expression level of that metabolite or ASV for both groups of the introduction of solid foods. It could be observed that there was a difference in expression levels for a portion of the selected features for both groups of the solid food introduction. Due to the large amount of key contributors selected by the DIABLO algorithm, the labels of the key contributors are hard to read. Therefore, all key contributors with correlations > 0.7 or < -0.7 are summarized in Table 15, amounting to a total of 119 strong associations between ASVs and metabolites or between metabolites. It could be seen that a set of ASVs were positively correlated with metabolites. One bacteria may be present multiple times in Table 15. This is due to the fact that different ASVs of that bacteria were associated with the same metabolite. As the main focus was the association of ASVs with metabolites, the associations between different metabolites were currently not looked into. *Bacteroides ovatus* had the largest amount of associations (21) with metabolites (Table 9, Table 15). Other associations between bacteria and metabolites were also investigated. Noteworthy associations were between:

- *Streptococcus infantis* and Beta-muricholic acid
- *Butyricicoccus pullicaecorum*, *Bacteroides caccae* and *Bacteroides ovatus* and histidine.

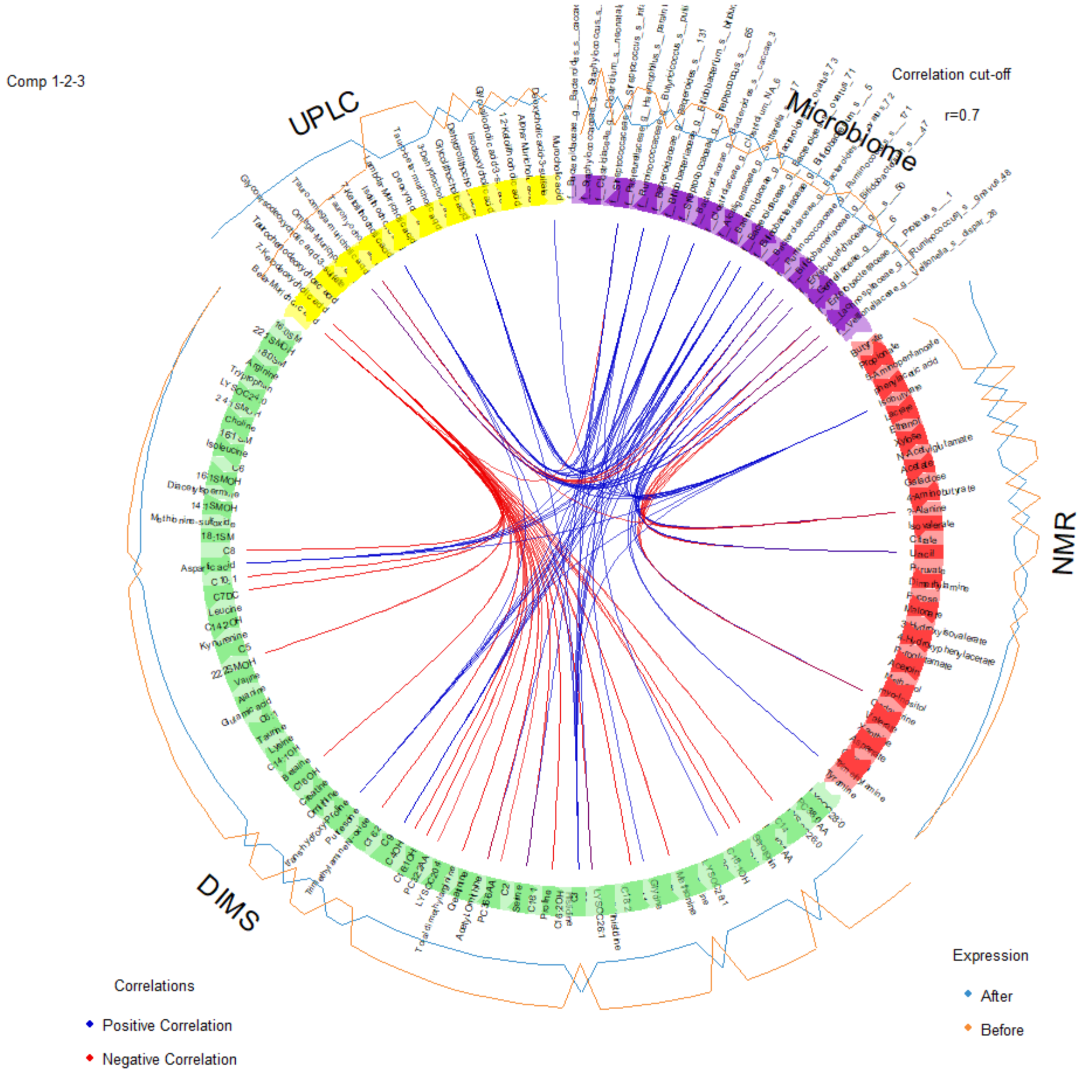

*Figure 21 A circos plot of the final DIABLO model for the LucKi Gut Study. Each data set is shown in a different colour with the microbiome data in purple, the NMR data in red, the DIMS data in green and UPLC data in yellow. Inside the circle, the blue lines show the positive correlations > 0.7 and the red lines the negative correlations < -0.7. The lines outside the circle show the expression of that metabolite or ASV for both groups of solid food introduction. Before in orange and after in blue.*

*Table 9 Amount of associations between ASVs and metabolites per bacteria.*

| Bacteria | Associations |
|---|---|
| ***Bacteroidaceae Bacteroides ovatus*** | 21 |
| ***Ruminococcaceae Butyricicoccus pullicaecorum*** | 7 |
| ***Bacteroidaceae Bacteroides caccae*** | 7 |
| ***Ruminococcaceae Ruminococcus* sp.** | 6 |
| ***Erysipelotrichaceae*** | 6 |
| ***Veillonellaceae Veillonella dispar*** | 5 |
| ***Lachnospiraceae [Ruminococcus] gnavus*** | 5 |
| ***Enterobacteriaceae Proteus*** | 5 |
| ***Streptococcaceae Streptococcus infantis*** | 4 |
| ***Bifidobacteriaceae Bifidobacterium* sp.** | 3 |
| ***Bifidobacteriaceae Bifidobacterium bifidum*** | 2 |

Figure 22 is a heat map of the correlations between all 159 metabolites and ASVs selected by the final DIABLO model. The Euclidean distance was used as a distance measure. A blocky structure could be observed between the different metabolites and ASVs used in the DIABLO model which indicated the grouping of metabolites and ASVs.

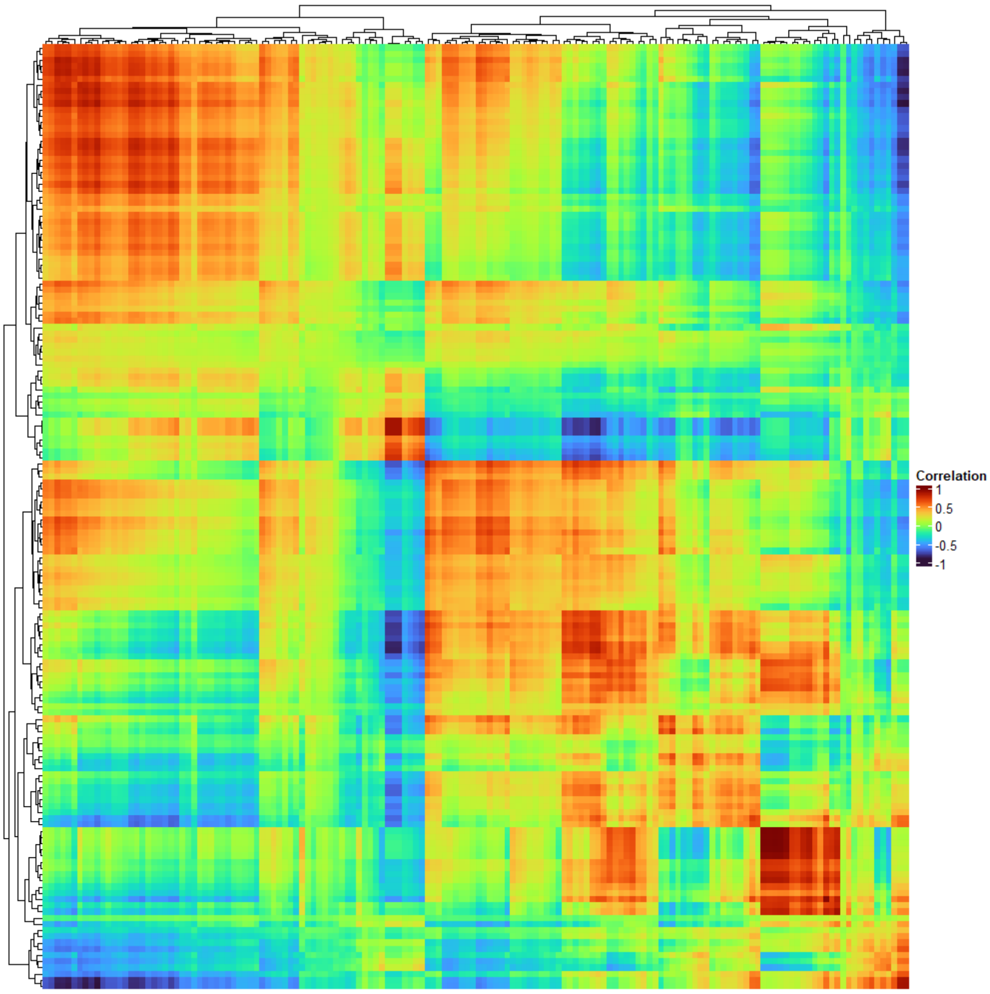

*Figure 22 Heat map of the correlations from the circos plot.*

# 4 Discussion

Within the present project, the effects of the introduction of solid foods on the gut microbiome and metabolome were investigated. A second scope of the study was to examine the association between the gut microbiome and metabolome. To this end, a selection of 9 infants from the LucKi Gut study who were intensively sampled during a 14-day period at the time of introduction of solid foods were used [39].

The impact of solid food introduction on the gut microbiome was investigated first. The microbiome data was filtered, removing 98% of the taxa. Afterwards, the counts were normalized using a CLR-transformation to account for the compositionality of the microbiome data [59]. The effects of solid food introduction and other potential covariates were explored using multiple exploratory tools. The introduction of solid foods showed no influence on the microbial composition or on individual microbial taxa. It was observed that the microbiome could be regarded as a unique fingerprint for each individual, which corresponds to literature [1, 2, 3, 4, 90]. The age of the infants was shown to influence the alpha diversity. Formal testing for differentially abundant ASVs due to solid food introduction was done using ANCOM-BC and the Benjamini-Hochberg procedure to account for multiple hypotheses testing [58, 74, 76]. Prior to adjustment for multiple hypotheses testing, 7 significant ASVs (5.79%) were found to be differentially abundant prior to the introduction of solid foods. These ASVs were all known gut commensals [91]. After adjusting for the multiple hypotheses testing, no significant results were found for the introduction of solid foods. This indicated that the abundances of the ASVs didn't change significantly under the introduction of solid foods. Literature suggests that the gut microbiome is able to respond rapidly to dietary changes, even within the order of hours and days [92]. However, at the time of introduction of solid foods, these dietary changes are not as drastic as changing a complete diet. The infants were still breastfed after the introduction of solid foods most of the time and received only small amounts of simple foods like fruit and porridges once or several times a day to substitute the breast milk. Consequently, the microbiome is likely still mainly adapted to the degradation of human milk oligosaccharides present in breast milk. For age however, the abundances of ASVs did show statistically significant results after adjustment for multiple hypotheses testing in 60 out of 121 ASVs (49.59%). This corresponds nicely with previous findings showing that age is an important factor in the maturation of the gut microbiome [19, 20]. This provides an interesting insight into the rapid diet-independent development of the gut microbiome with age.

Next, the metabolome was investigated. Prior to performing any analysis, the data was cleaned and normalized using a natural logarithm and pareto scaling. An exploratory data analysis was conducted afterwards. Both the PCA and heatmaps yielded similar conclusions. Visually, there was no distinct separation of the samples that were collected before and after the introduction of solid foods. In contrary to the microbiome data, samples from the same infant did not cluster together. This might be explained by the fact that, in contrast to the microbiota composition, the metabolome is more directly influenced by the food that was recently consumed. Since all infants followed a highly similar diet consisting mostly out of breastfeeding, a clear distinction between infants may not yet be present. An additional biplot was constructed for the UPLC data. The top 5 metabolites were visualized and showed a large amount of variability for 4 samples after the introduction of solid foods. The metabolites were all bile-acid taurine and glycine conjugates present in larger abundances compared to other samples. These 4 samples were also clustered together on the heatmap for the UPLC data. However, no patterns related to the microbiome or foods digested were discovered. The heatmap for the NMR data showed almost complete absence of lactate for infant "T" while receiving breastfeeding. A potential reason for this could be the high abundance of *Bacteroidaceae*. Previous studies have proven a higher abundance of *Bacteroidaceae* causes a lower abundance of lactate [93]. In infant "R", lactate was also only present in low abundances while succinate had a higher abundance compared to other infants. The lower lactate levels may have a similar reason as observed in infant "T", which was due to the high abundance of *Bacteroidaceae* [93]. A potential reason for the high succinate abundance could be due to the high abundance of *Porphyromonadaceae* in infant "R". This family is involved in the succinate pathway which yields succinate as the end-product [32]. Several methods such as a Wilcoxon rank-sum test, two sample t-test, ANOVA and linear mixed model were used to test for differentially abundant metabolites. Each of the methods yielded the same conclusions and the Wilcoxon rank-sum test was reported [82]. The Benjamini-Hochberg procedure was used to correct for multiplicity [76]. Prior to any adjustment for multiple hypotheses testing, several metabolites were found to be differentially abundant. The 4 differentially abundant metabolites in the NMR data prior to accounting for multiple testing were Galactose, Butyrate, 5-Aminopentanoate and Xylose. The 12 differentially abundant metabolites in the DIMS data prior to adjustment were part of the fatty-acid metabolism, central carbon metabolism and choline metabolism. For the UPLC data, these 13 differentially abundant bile-acids were taurine and glycine conjugates. In particular, the galactose is interesting as galactose is a part of lactose. Galactose had an average normalized abundance of 0.345 before the introduction of solid foods and -0.230 after the introduction of solid foods. This potentially indicates the reduction of milk in the diet of infants after solid food introduction. After correction for multiple hypothesis testing, no differentially abundant metabolites were found. A same reasoning can be given for the metabolomics results as for the microbiome results. The introduction of solid foods is a slow process, replacing breast milk once or several times a day by simple foods such as fruit porridge while still receiving breast milk most of the day. Which causes only a small shift of metabolites in the beginning. An alternative explanation could be due to the fact that a low sample size, large amount of metabolites and large variation between individuals leads up to very little power to retain significant associations after correcting for multiple hypotheses testing.

The final analysis was conducting an association study between the microbiome and metabolome data at the time of introduction of solid foods. DIABLO was used to perform this association study [83]. DIABLO required the tuning of several hyperparameters such as the covariance matrix, the number of principal components used and the amount of ASVs and metabolites used in the final model. The results for the final DIABLO model were first shown at sample level using a correlations plot between each of the n$^{th}$ principal component of each data set. In the correlation plots of the first principal components, the 95% confidence ellipses of both groups of solid food introduction were only partially overlapping. This indicated that separation between both groups was possible up to a certain height. For the second and third principal component correlation plots, this separation was no longer as straightforward. This was most likely due to the fact that the first principal components explained most of the variability in the data. The fact that separation between both groups of solid food introduction might not be entirely possible could be due to the fact that no differentially abundant ASVs or metabolites were found during the data analysis. Yet when comparing these results to the PCA performed for the microbiome or metabolome separately, the separation between both groups had significantly increased. This could indicate that using both microbiome and metabolome data combined, a distinguishment between both groups could be easier. A next step in the association study was looking into the results at feature level, either ASV or metabolite, using a circos plot. The circos plot had a filter where only correlations > 0.70 or < -0.70 were shown. A total of 119 bacteria-metabolite and metabolite-metabolite associations were found. There were 11 different bacteria present with metabolite associations. All bacteria-metabolite associations could be related back to metabolites produced by humans, the bacteria itself and from food sources [77, 94]. All of these bacteria were commensals of the gut microbiome. *Bacteroides ovatus* stood out with 3 ASVs associating with 7 different metabolites amounting to a total of 21 associations. This was due to the fact that multiple ASVs of *Bacteroides ovatus* were positively associated with the same set of metabolites. This strengthens the proof of an association between *Bacteroides ovatus* and these metabolites. *Bacteroides* sp. is a beneficial bacterium responsible for the metabolization of polysaccharides, oligosaccharides, providing nutrition and vitamins to the host and other intestinal microbial residents. *Bacteroides ovatus* is also associated with breaking down insulin [95]. An interesting finding was the bacteria-metabolite association between *Streptococcus infantis* and Beta-muricholic acid. Beta-muricholic acid is a bile-acid synthesized by mice while humans synthesize cholic acid [96]. A direct rationale could not be found behind the presence of this metabolite and its association the with *Streptococcus infantis*. Another interesting finding was an association between *Butyricicoccus pullicaecorum*, *Bacteroides caccae* and *Bacteroides ovatus* and histidine. The associations between both *Bacteroides* sp. and histidine have been described in literature, as both bacteria are involved in the histidine metabolism [97]. A link between *Butyricicoccus pullicaecorum* and histidine has not yet been established. Histidine is an essential amino-acid and a precursor metabolite to histamine. Histamine is a compound involved in multiple physiological processes such as gastric acid secretion and as a vital inflammatory agent in immune responses and allergic reactions [77]. A final visualization provided was a heat map of the correlations between the different metabolites and ASVs present in the final DIABLO model. A blocky structure was observed indicating higher correlations between different metabolites and ASVs. This showed that different metabolites and ASVs clustered together and potentially were related with each other. The associations found using DIABLO showed an early-life association between the gut microbiome and metabolome was present not only concerning the digestion of food but also as a key player in the immune system of infants.

The major strength of this study was the focus on looking into the influences of solid food introduction on a short, intensively sampled time-interval. Previous studies only looked into this over a larger period of time which didn't allow for a focus on the rapidly occurring changes in the microbiome and metabolome during infancy. The association study showed associations present between the gut microbiome and metabolome during this short time-interval. It opens new questions for future research and reach more knowledge about the development of the gut microbiome during infancy.

The major shortcoming of this study was the low amount of samples available. The low amount of samples caused a lower statistical power. This caused a low sensitivity to find differentially abundant ASVs and metabolites. This was also the case for the association study, which was done with even less samples. A larger sample size would yield more statistically robust results. A second shortcoming was the limited sampling time. Some infants only had samples of a few days, which could potentially be too short for the microbiome and metabolome to change significantly under the effects of solid food introduction.

# 5 Conclusion

This thesis focused at the changes in the microbiome and metabolome taking placing at the time of solid food introduction using an intensively sampling study. An additional topic was the association between the microbiome and metabolome at the time of solid food introduction. ANCOM-BC was used to test for differential abundance of ASVs. No differentially abundant ASVs were found due to the solid food introduction after correction for multiplicity. The age of infants was a second covariate in the model. There were 60 differentially abundant ASVs based on the age of infants after correction of multiplicity. A Wilcoxon rank-sum test, two sample t-test, ANOVA and linear mixed model was used to test for differentially abundant metabolites and the Wilcoxon rank-sum test was reported. No differentially abundant metabolites were identified after correction for multiplicity. A potential reason for not finding any differentially abundant ASVs and metabolites is due to the fact that the dietary changes at the time of solid food introduction may not be as drastic to cause any large changes in gut microbiome and metabolome compositions. The differentially abundant ASVs for the age of an infant corresponds with previous findings where the microbiome is affected by age. The final focus was the studying the association between the microbiome and metabolome. This was done using DIABLO. A set of associations were identified between different bacteria and metabolites. These metabolites could all be related back to metabolites produced by the infants, by the bacteria itself and from food sources. A set of bacteria were also associated to histidine, a precursor metabolite for histamine which is responsible for gastric acid secretion and immune responses. This research confirms an association between the gut microbiome and metabolome in the digestion of food but also as a key player for the immune response of infants at an early-life. This research provided a valuable insight into the gut microbiome during infancy and its functions. This study could lead up to a further knowledge about the development of the gut microbiome and potential health-related issues whom are related to the gut microbiome.

As future research, a proposal would be looking into the metabolites found to be associated with bacteria. One way of doing this would be by cultivating the bacteria on agars and adding the metabolite to the agar. If it affects the growth of the bacteria, this confirms a relationship between the metabolites and bacteria. The correlations were found on a data driven method and this would confirm their relationship in a laboratory setting. A second proposal would be increasing the sample size of the study as a larger sample size would potentially reveal biologically relevant changes in the gut microbiome and metabolome. A third proposal would be to collect samples from these infants after the breast feeding has completely stopped. This might provide more insight into the effects of solid food introduction and breast feeding.

## Acknowledgments

I would like to start out by thanking my internal supervisor, Prof. Dr. Dirk Valkenborg, and external supervisors, Dr. John Penders and Dr. Niels van Best. Each of you have provided me excellent guidance, counselling and valuable ideas during my master thesis. The feedback helped in my development and pushed the thesis to a higher level.

I would also like to thank the entire microbiome research group of Dr. John Penders. The weekly meetings, presentations and paper discussions have broadened my knowledge about the microbiome and drug resistance. It definitely sparked a big interest in me to continue with research.

A thank you to all the Professors and Doctors of uHasselt for providing the Master in Statistics and Data Science. I look forward to developing a career in this field and look up to every single one of you!

A big special thanks to my partner Sofie. Not everyone has the patience to stay with someone who works full-time and studies full-time. But you didn’t only manage to do that, you supported me every single moment. I would never have been able to complete this program without you. And finally, thanks to my family! For all the support, council and motivation to help me out.

# Appendices

## R packages

*Table 10 R packages used with their corresponding versions.*

| *R package* | Version |
|---|---|
| ANCOMBC | 1.6.0 |
| Biobase | 2.56.0 |
| BiocGenerics | 0.42.0 |
| BiocParallel | 1.30.0 |
| Biostrings | 2.64.0 |
| ComplexHeatmap | 2.12.0 |
| data.table | 1.14.2 |
| dplyr | 1.0.9 |
| DT | 0.23 |
| factoextra | 1.0.7 |
| forcats | 0.5.1 |
| GenomeInfoDb | 1.32.2 |
| ggplot2 | 3.3.6 |
| ggridges | 0.5.3 |
| ggtext | 0.1.1 |
| ggven | 0.1.9 |
| IMIFA | 2.1.8 |
| IRanges | 2.30.0 |
| lattice | 0.20-45 |
| lme4 | 1.1-29 |
| MASS | 7.3-57 |
| Matrix | 1.4-1 |
| matrixStats | 0.62.0 |
| MESS | 0.5.7 |
| metagMisc | 0.0.4 |
| microbiome | 1.18.0 |
| microViz | 0.9.0 |
| mixOmics | 6.20.0 |
| nlme | 3.1-157 |
| oligo | 1.60.0 |
| oligoclases | 1.58.0 |
| phyloseq | 1.40.0 |
| purrr | 0.3.4 |
| readr | 2.1.2 |
| S4Vectors | 0.34.0 |
| skimr | 2.1.4 |
| stringr | 1.4.0 |
| tibble | 3.1.7 |
| tidyr | 1.2.0 |
| tidyverse | 1.3.1 |
| Viridis | 0.6.2 |
| viridisLite | 0.4.0 |
| XVector | 0.36.0 |

## Addendum

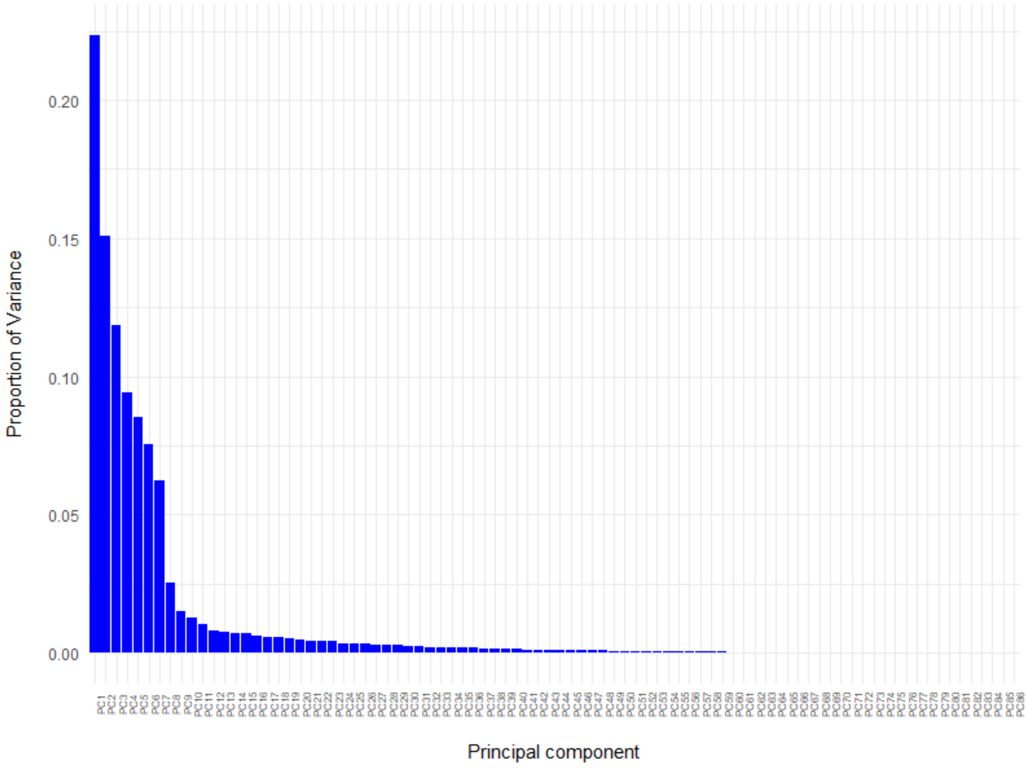

*Figure 23 Scree plot of the microbiome PCA shown in Figure 7.*

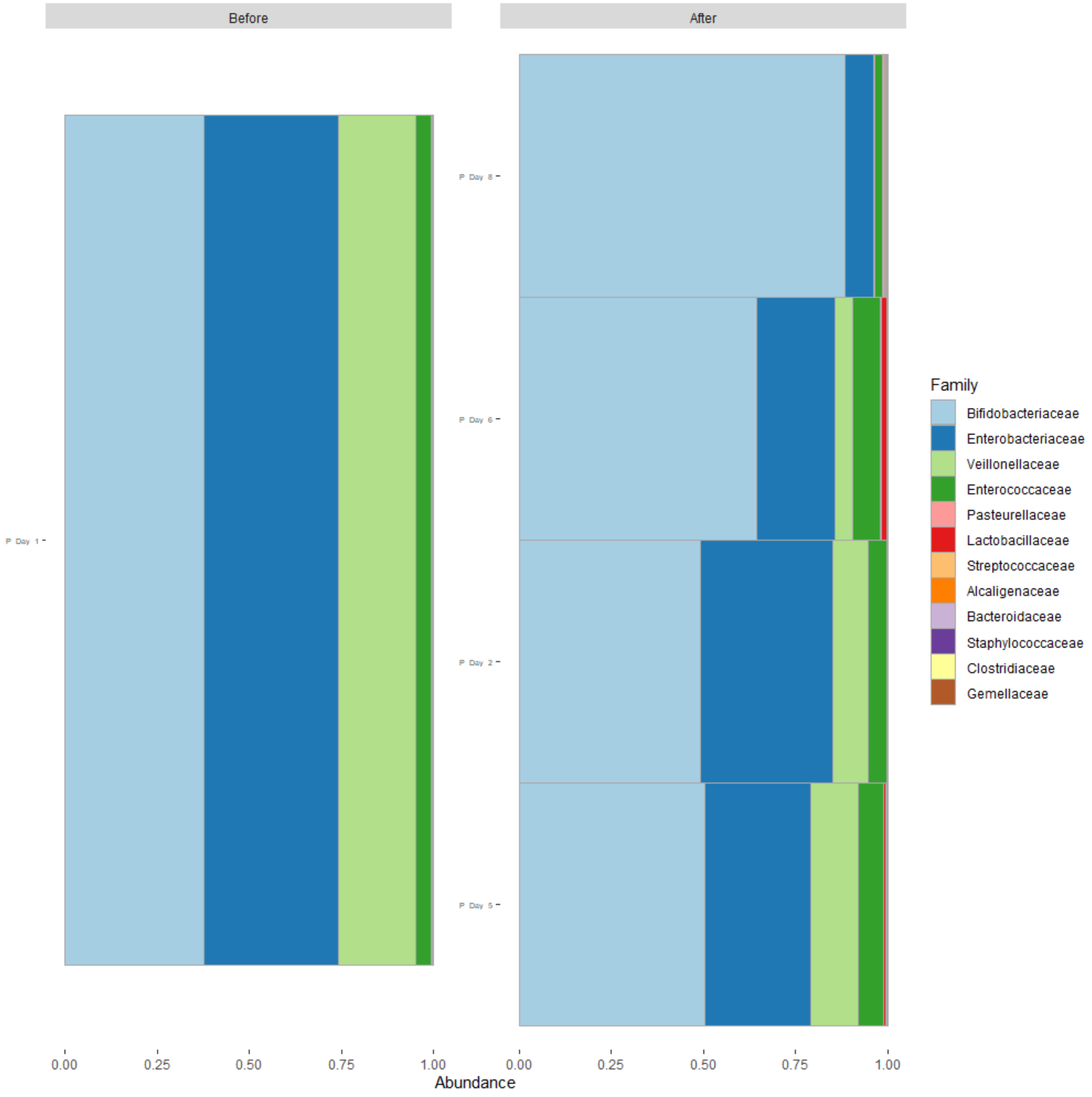

*Figure 24 Bar plot of the relative abundances in samples of infant "P" at family rank. Stratified over the introduction of solid foods.*

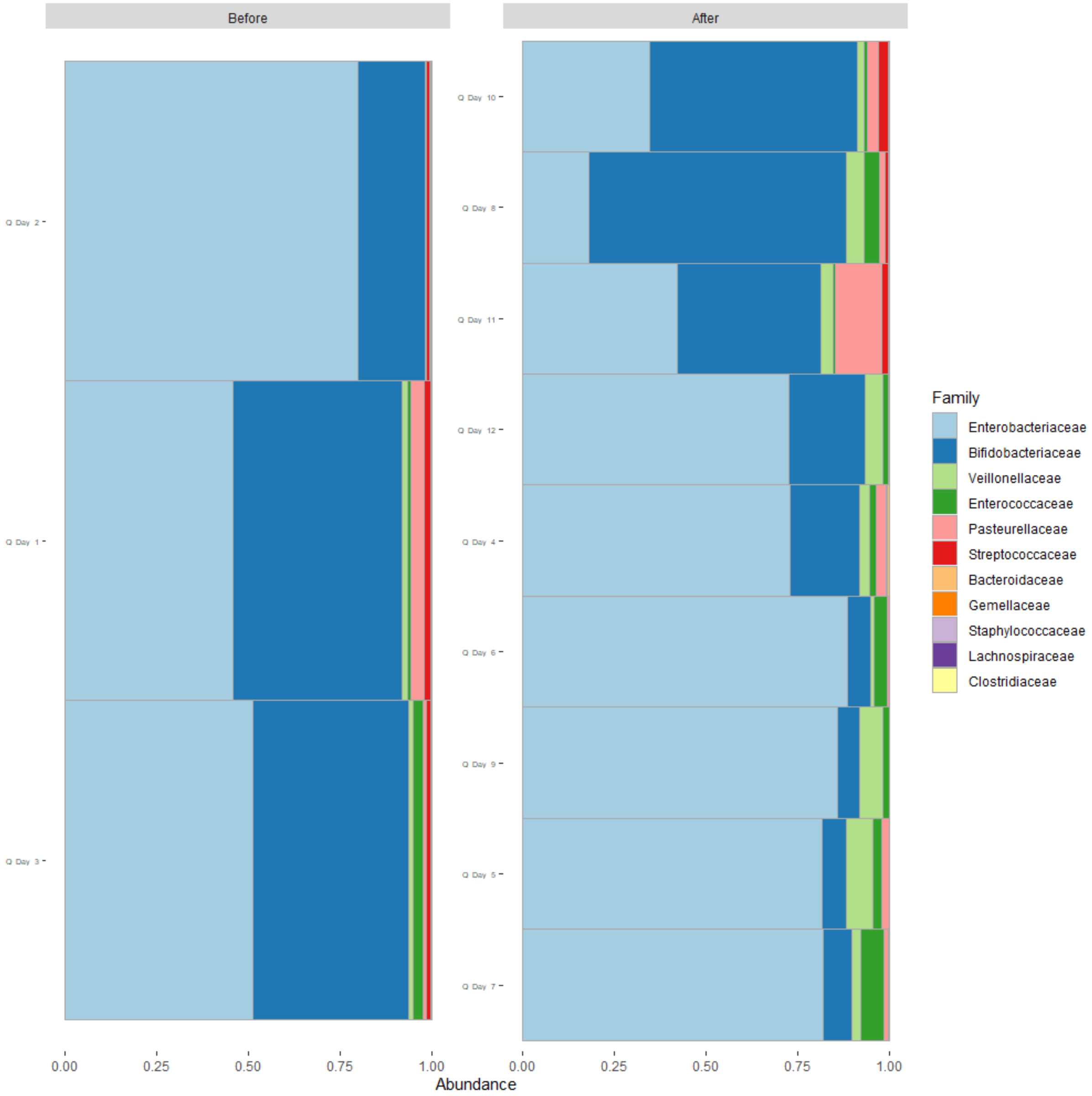

*Figure 25 Bar plot of the relative abundances in samples of infant "Q" at family rank. Stratified over the introduction of solid foods.*

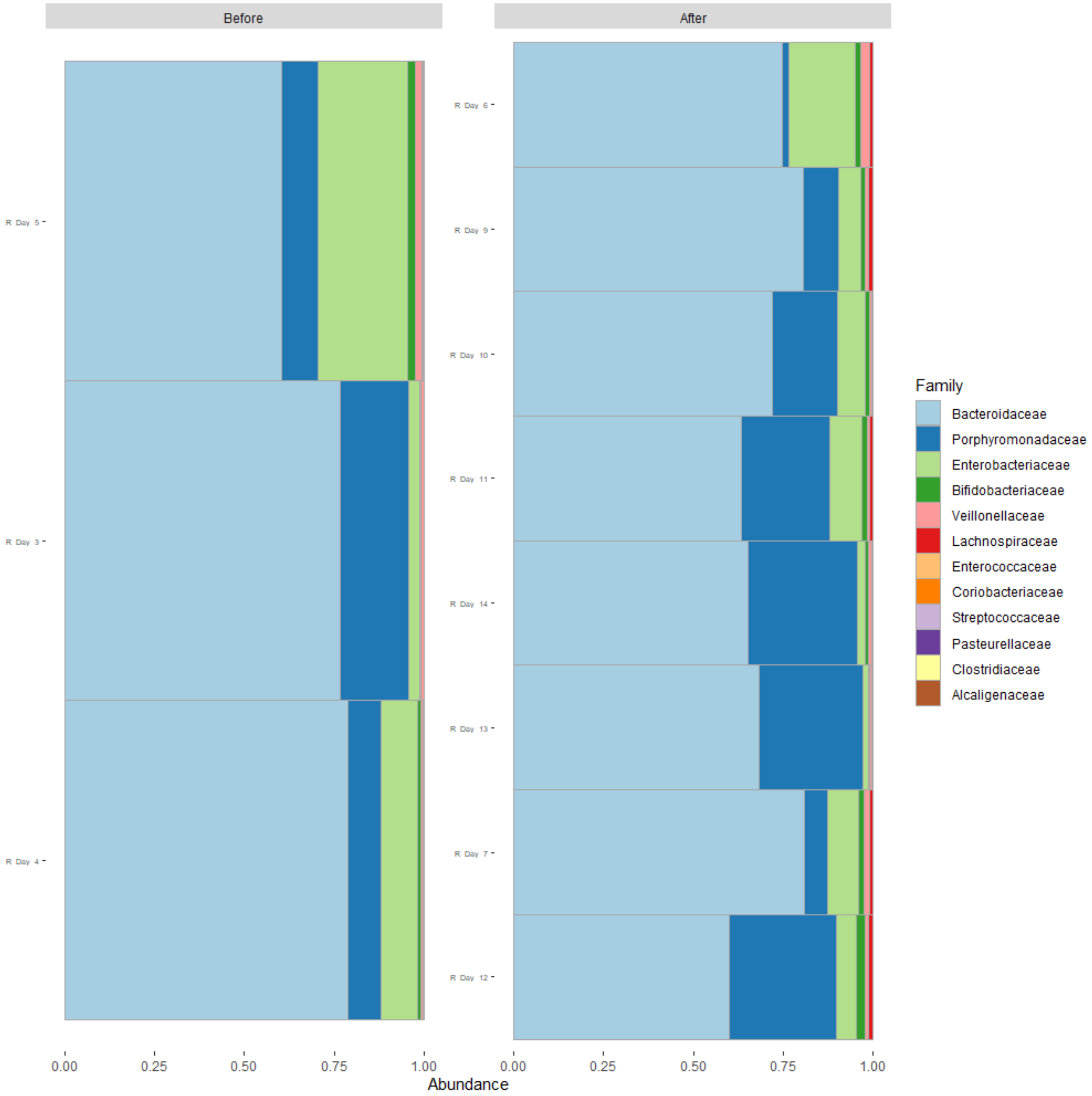

*Figure 26 Bar plot of the relative abundances in samples of infant "R" at family rank. Stratified over the introduction of solid foods.*

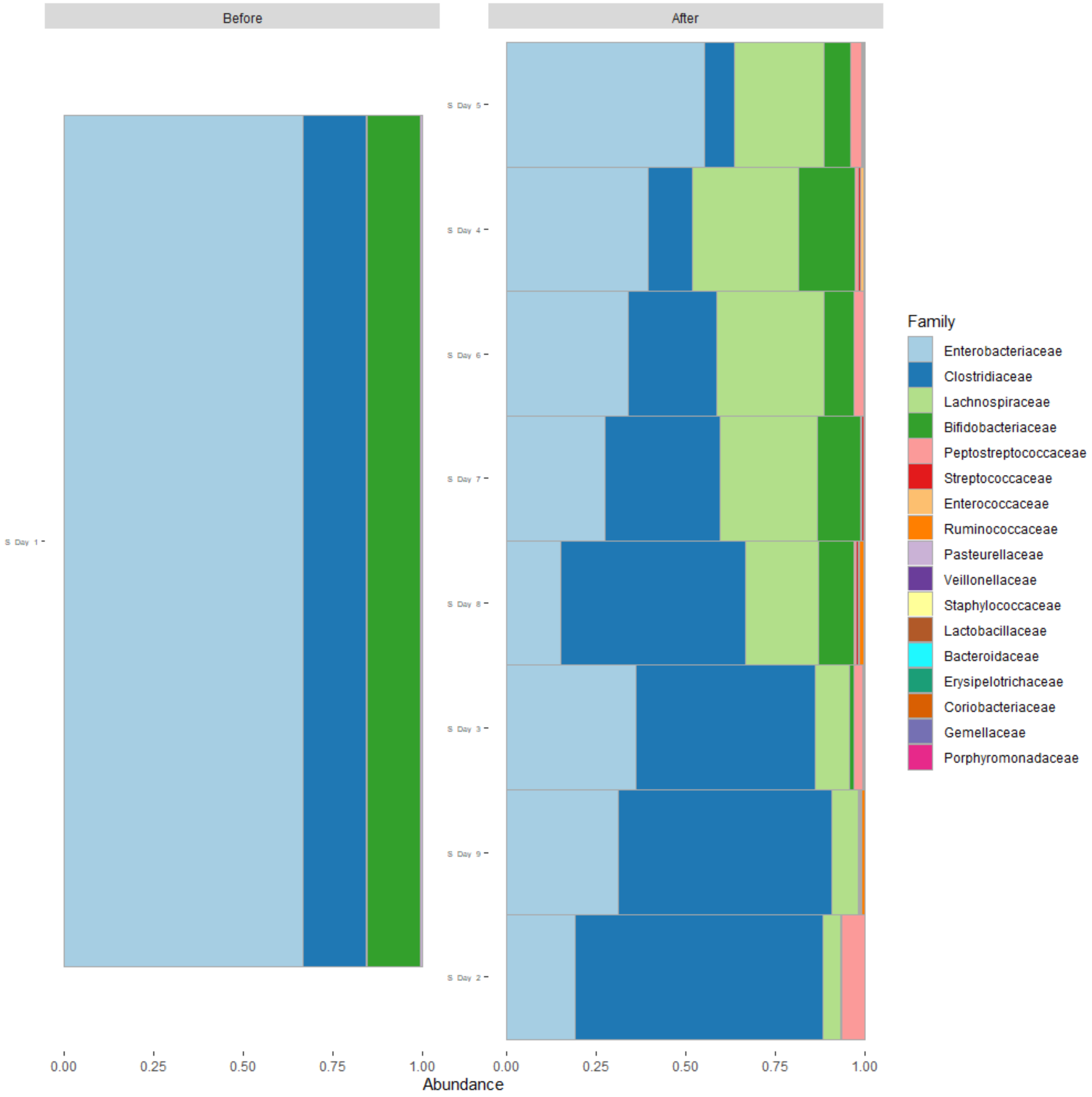

*Figure 27 Bar plot of the relative abundances in samples of infant "S" at family rank. Stratified over the introduction of solid foods.*

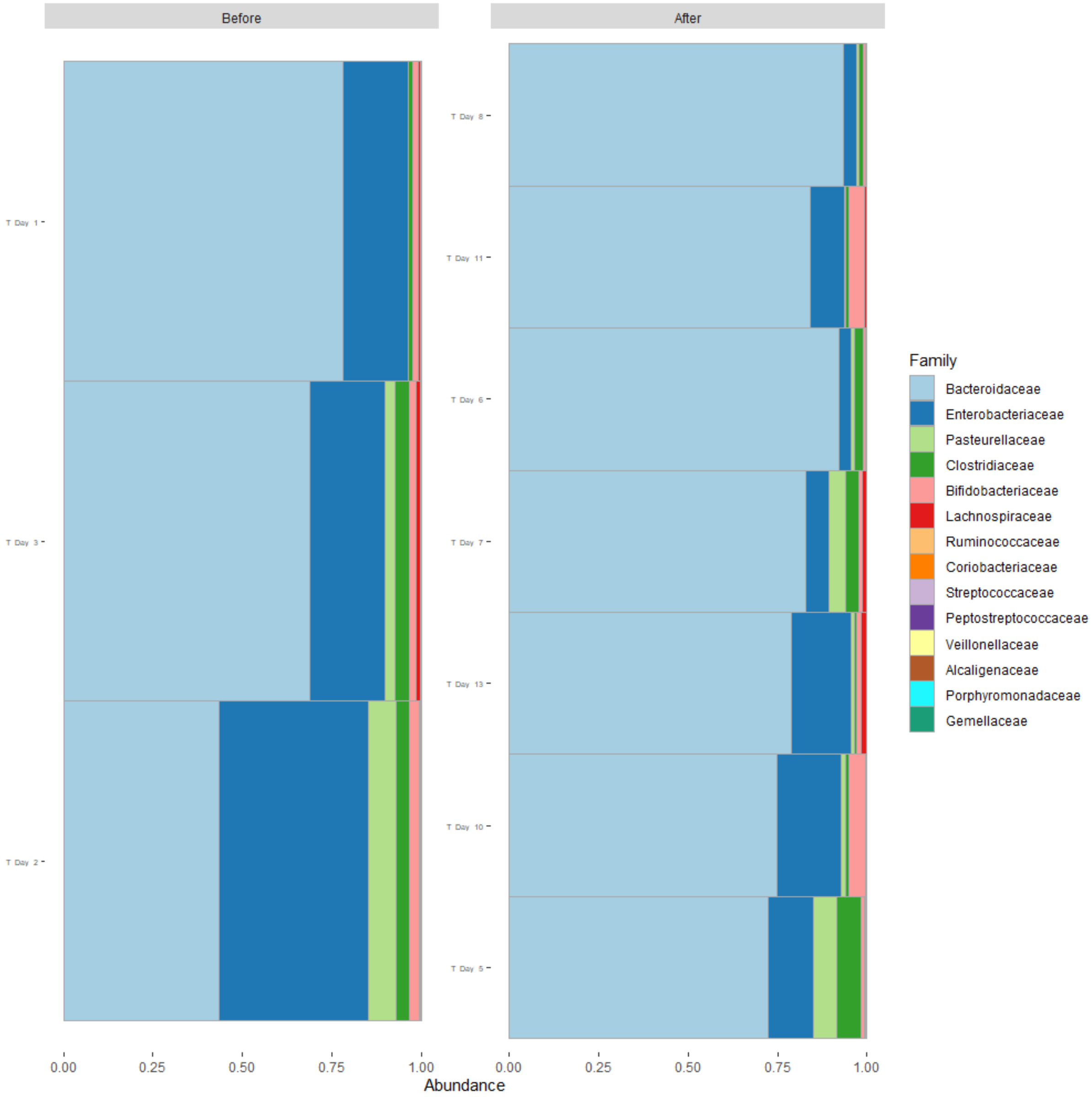

*Figure 28 Bar plot of the relative abundances in samples of infant "T" at family rank. Stratified over the introduction of solid foods.*

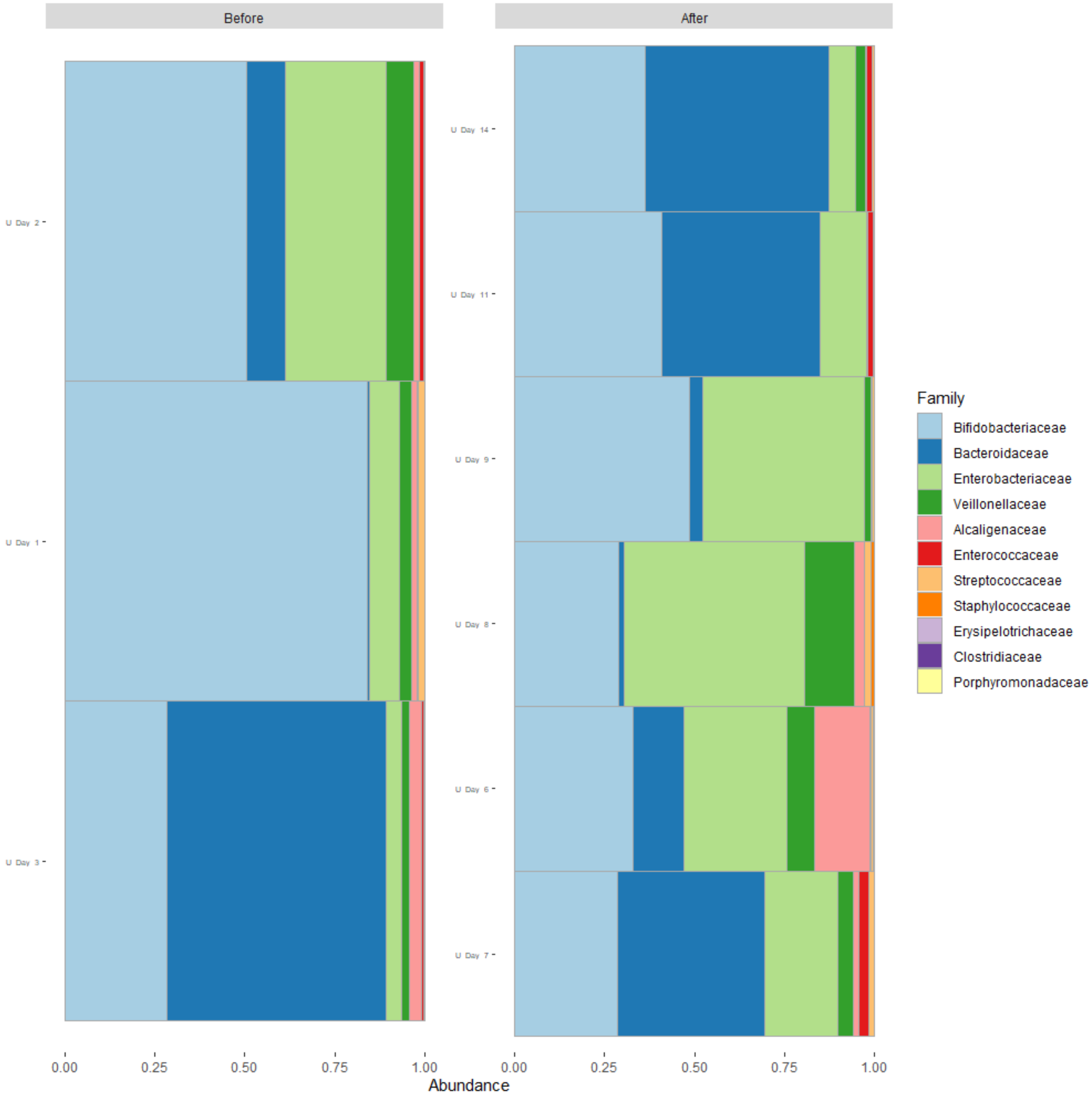

*Figure 29 Bar plot of the relative abundances in samples of infant "U" at family rank. Stratified over the introduction of solid foods.*

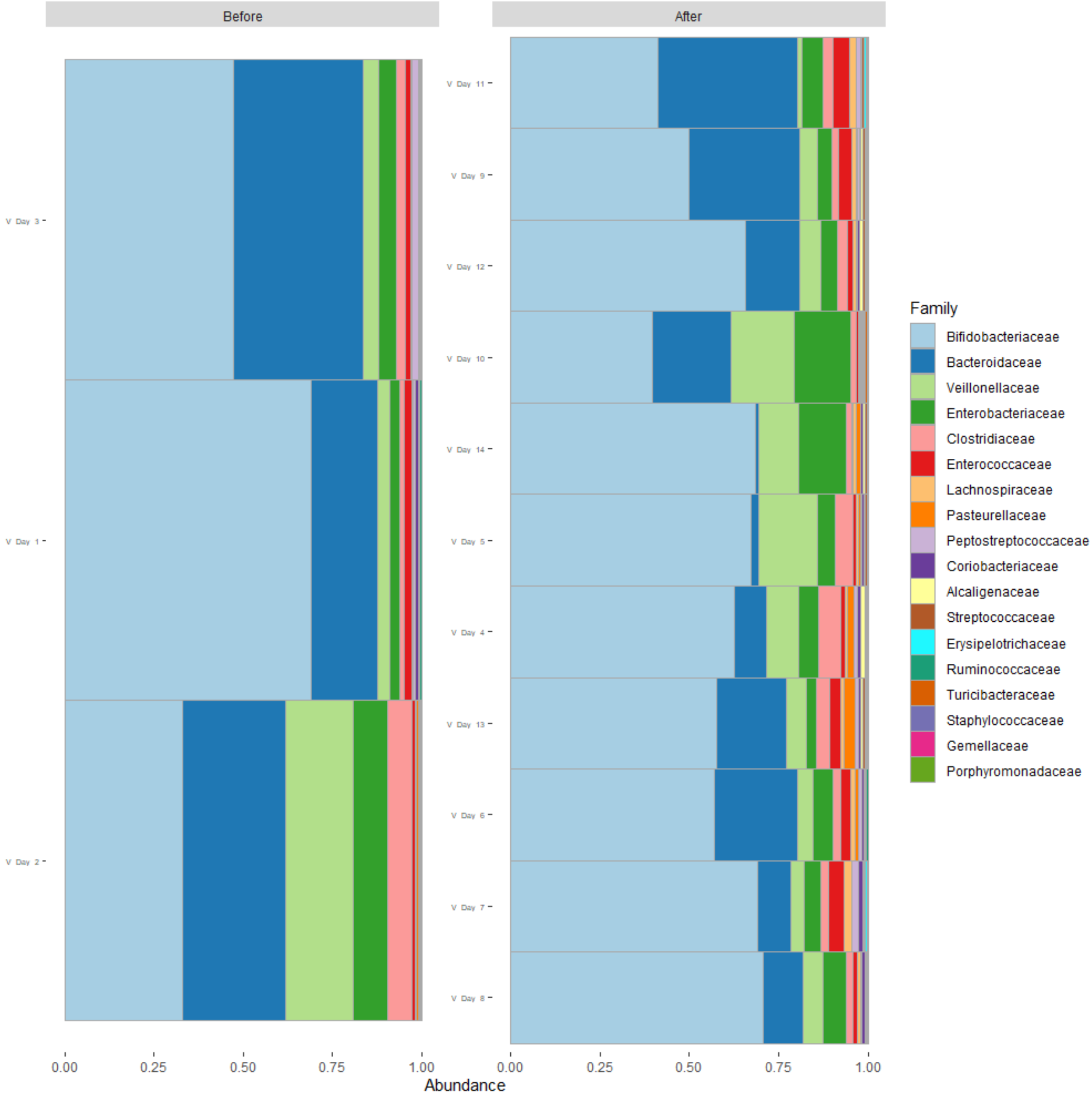

*Figure 30 Bar plot of the relative abundances in samples of infant "V" at family rank. Stratified over the introduction of solid foods.*

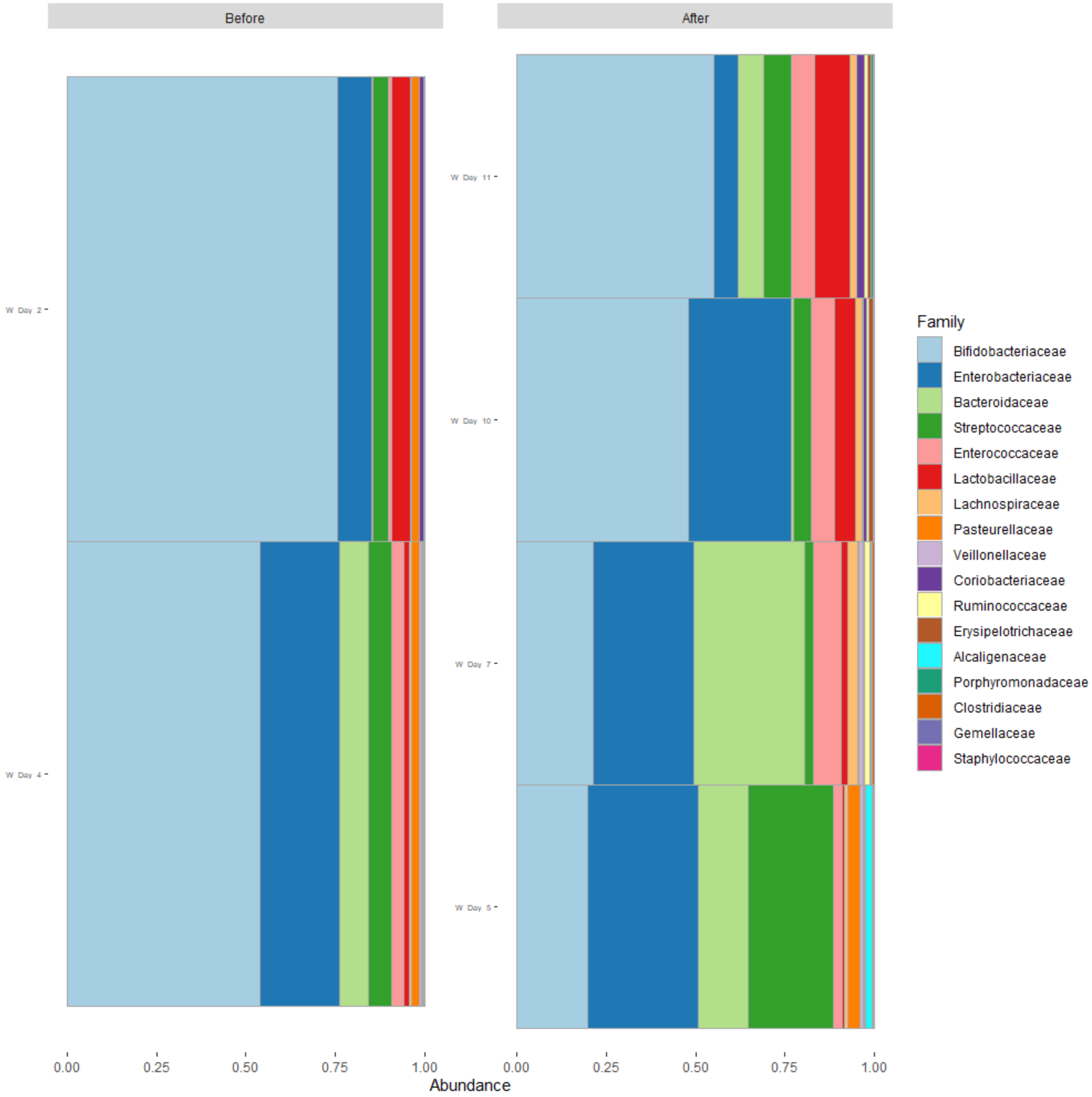

*Figure 31 Bar plot of the relative abundances in samples of infant "W" at family rank. Stratified over the introduction of solid foods.*

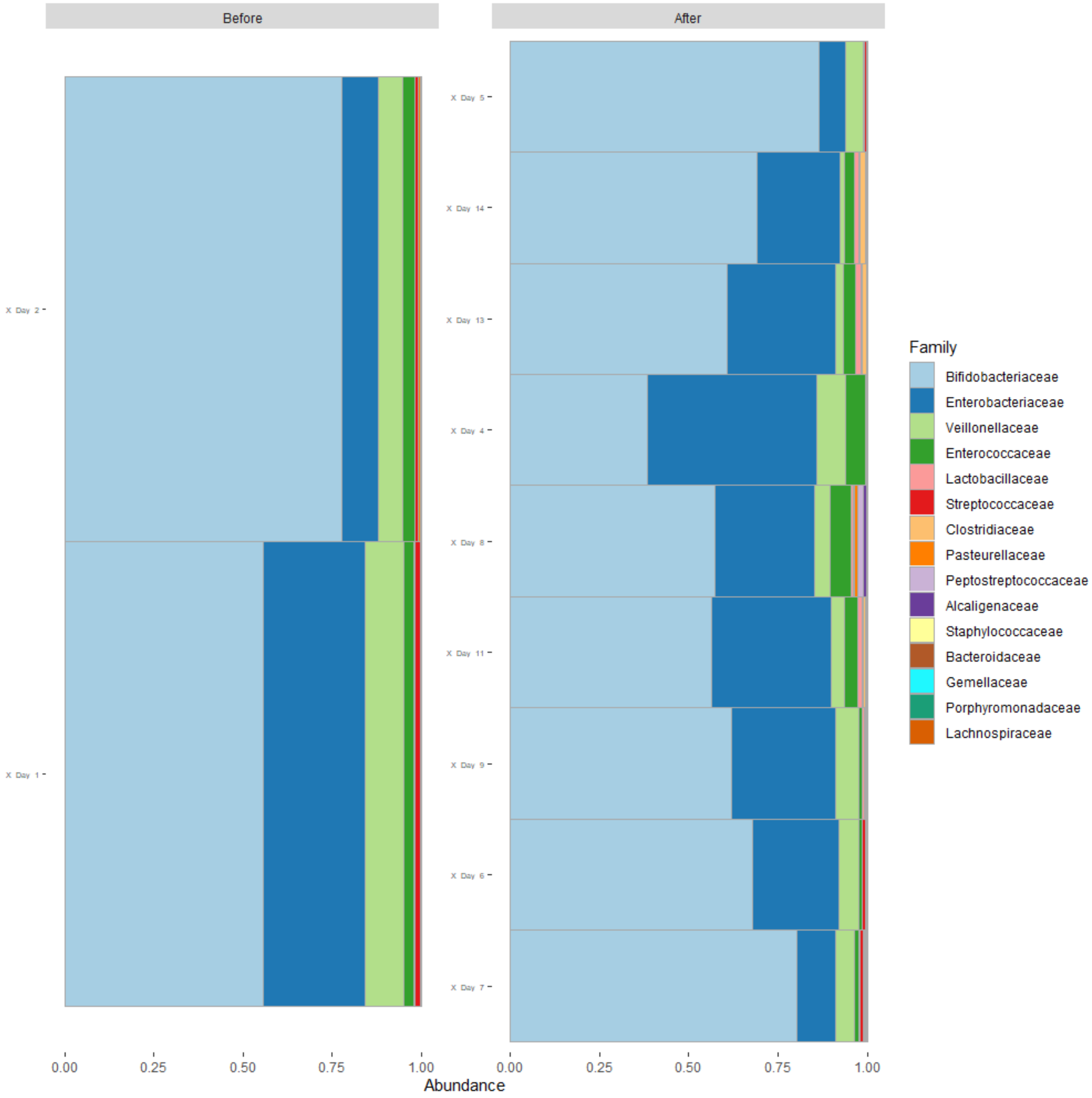

*Figure 32 Bar plot of the relative abundances in samples of infant "X" at family rank. Stratified over the introduction of solid foods.*

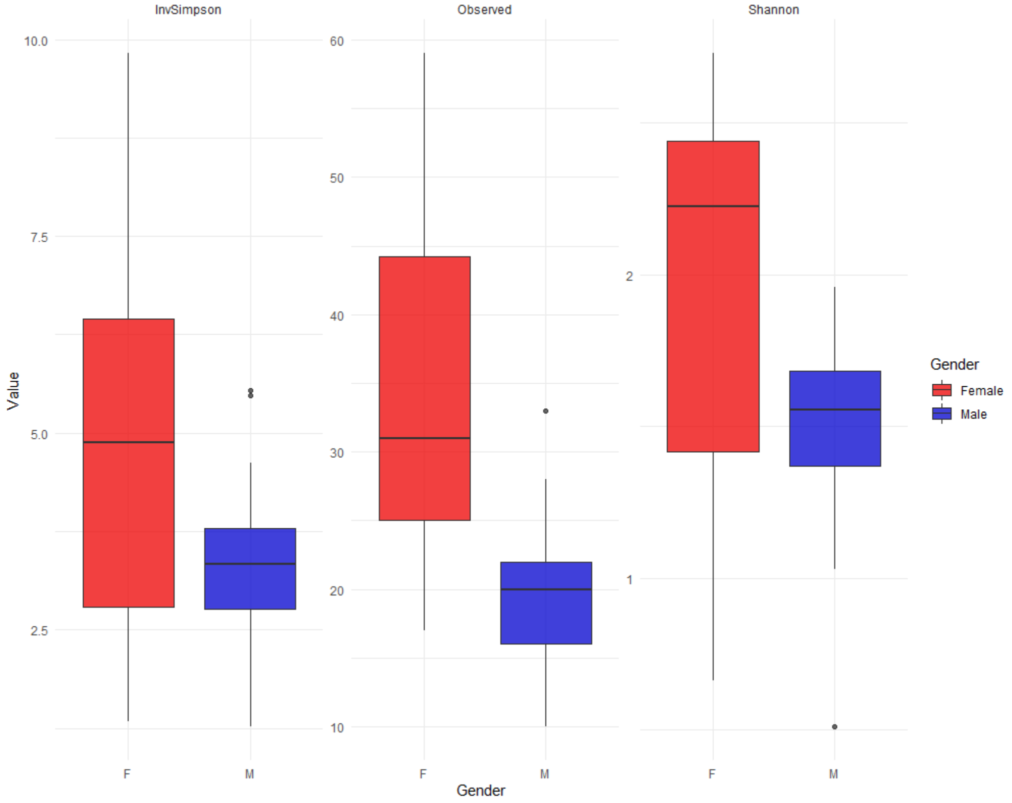

*Figure 33 Alpha diversity indices in function of gender. The colours denote the different genders with female in red and male in blue. Summarized for different alpha diversity statistics.*

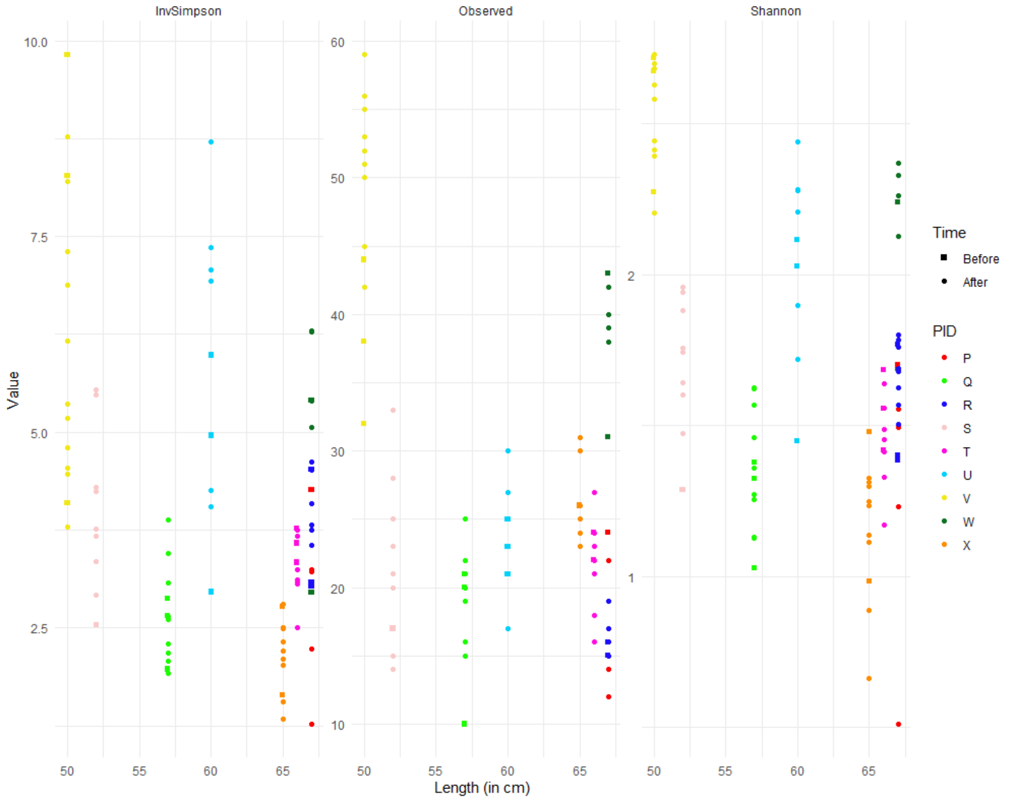

*Figure 34 Alpha diversity indices in function of length per infant. Coloured by infant and shapes denote the sampling times, either before (cubes) or after (circles) the introduction of solid foods. Summarized for different alpha diversity statistics.*

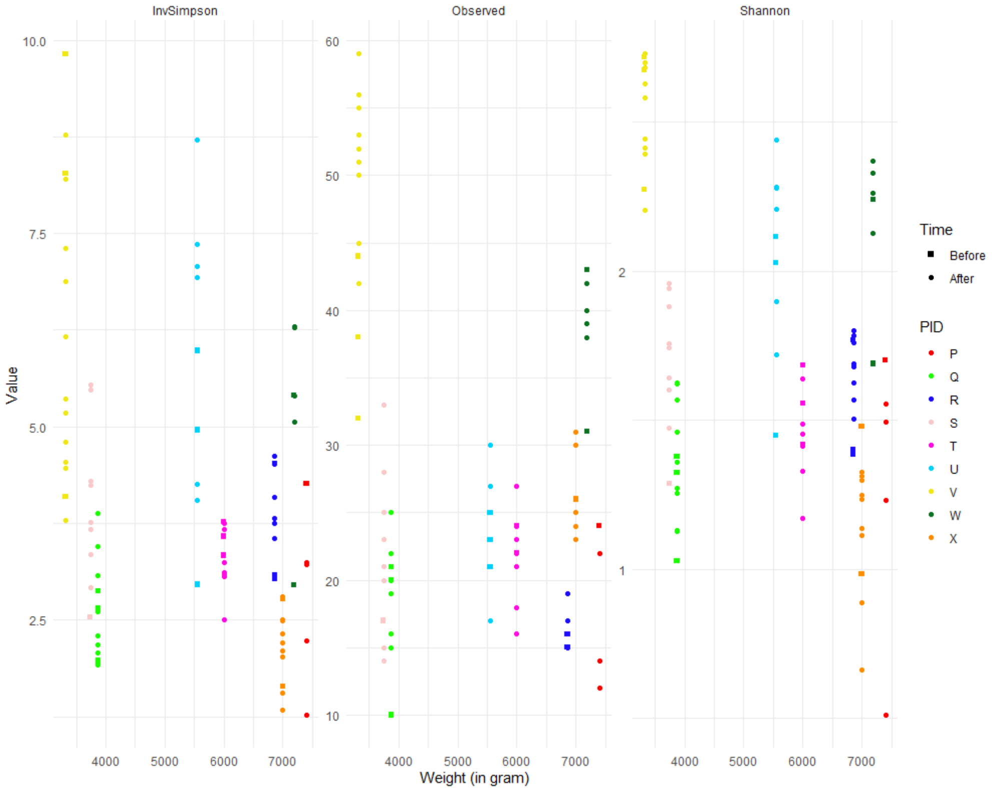

*Figure 35 Alpha diversity indices in function of weight per infant. Coloured by infant and shapes denote the sampling times, either before (cubes) or after (circles) the introduction of solid foods. Summarized for different alpha diversity statistics.*

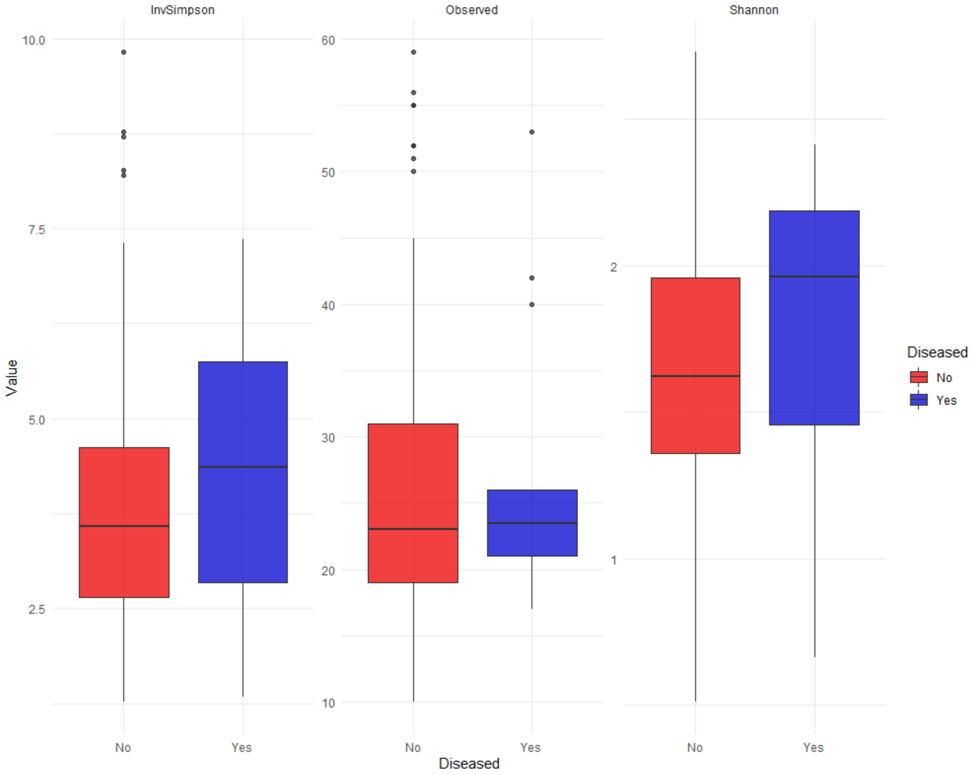

*Figure 36 Alpha diversity indices in function of disease status. The colours denote the disease status with “No” in red and “Yes”  in blue. Summarized for different alpha diversity statistics.*

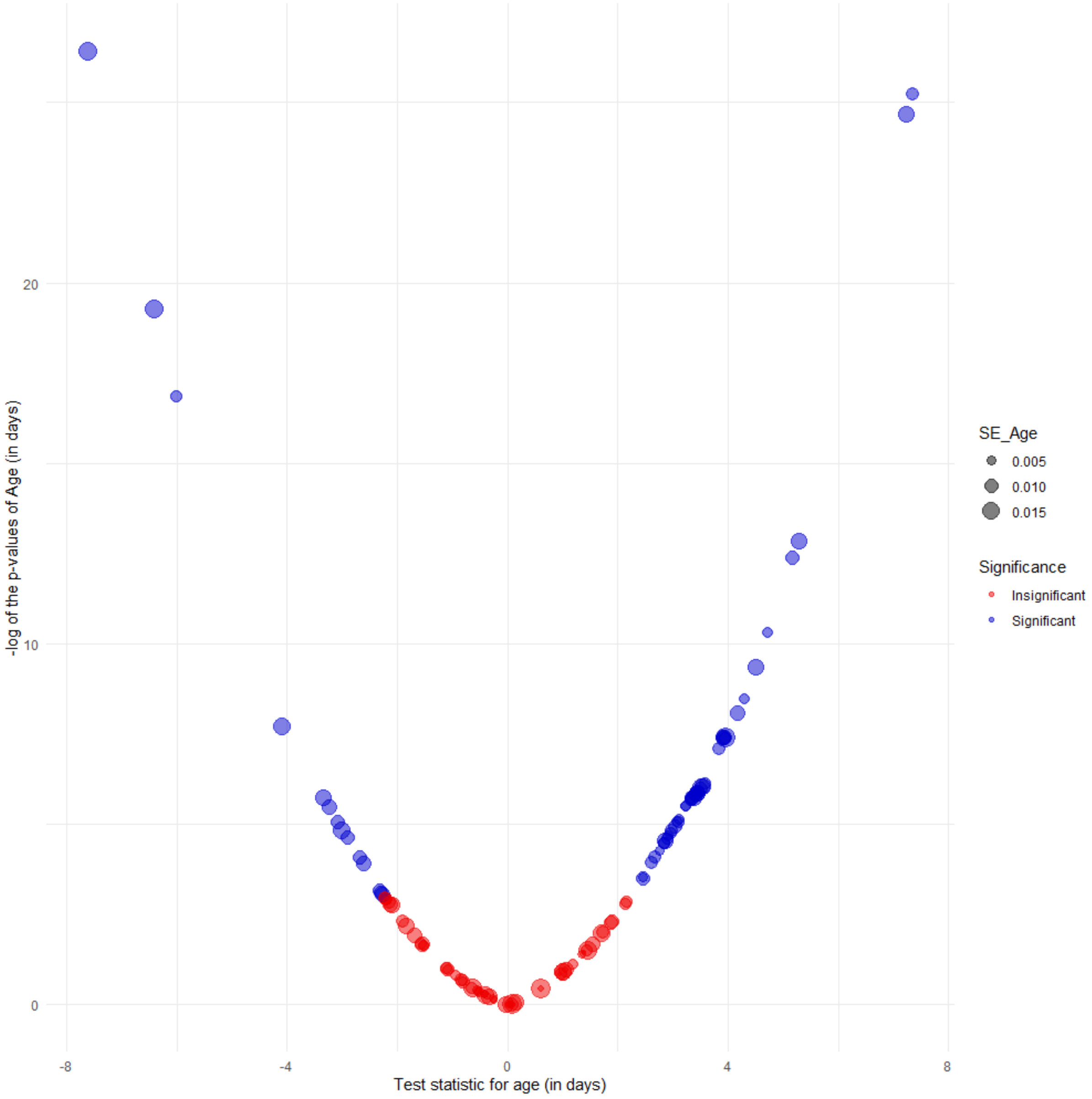

*Figure 37 Results for the age (in days) of the microbiome data after adjustment for multiplicity. The x-axis denotes the test statistic for the age in days and the y-axis the negative logarithmic transformed p-values after adjustment of multiplicity. Colours denote the significance of the results, red being insignificant and blue significant. The size of the dots denote the standard error of the test statistic.*

*Table 11 P-values for age and solid food introduction using ANCOM-BC.*

| *ASV* | **Age (in days) p-value** | **Solid foods p-value** | **Age (in days) adjusted p-value** | **Solid foods adjusted p-value** |
|---|---|---|---|---|
| ***f__Pasteurellaceae_g__Haemophilus_s__parainfluenzae_22*** | 0,611 | 0,006 | 0,698 | 0,594 |
| ***f__Streptococcaceae_g__Streptococcus_s__luteciae_1*** | 0,542 | 0,011 | 0,637 | 0,594 |
| ***f__Clostridiaceae_g__SMB53_s___12*** | 0,175 | 0,017 | 0,249 | 0,594 |
| ***f__Streptococcaceae_g__Streptococcus_s___67*** | 0,000 | 0,026 | 0,001 | 0,594 |

| | | | | |
|---|---|---|---|---|
| f__Enterobacteriaceae_g__Enterobacter_s__cloacae_9 | 0,290 | 0,027 | 0,385 | 0,594 |
| f__Staphylococcaceae_g__Staphylococcus_s__aureus_3 | 0,005 | 0,029 | 0,011 | 0,594 |
| f__Alcaligenaceae_g__Sutterella_s___75 | 0,937 | 0,040 | 0,976 | 0,687 |
| f__Lachnospiraceae_g__Blautia_s___49 | 0,130 | 0,057 | 0,197 | 0,714 |
| f__Enterobacteriaceae_g__Klebsiella_s___17 | 0,001 | 0,058 | 0,003 | 0,714 |
| f__Lachnospiraceae_g__[Ruminococcus]_s__gnavus_48 | 0,339 | 0,063 | 0,432 | 0,714 |
| f__Turicibacteraceae_g__Turicibacter_s___3 | 0,165 | 0,065 | 0,237 | 0,714 |
| f__Streptococcaceae_g__Streptococcus_s__infantis_2 | 0,014 | 0,076 | 0,030 | 0,717 |
| f__Streptococcaceae_g__Streptococcus_s___66 | 0,000 | 0,081 | 0,000 | 0,717 |
| f__Pasteurellaceae_g__Haemophilus_s__parainfluenzae_26 | 0,000 | 0,083 | 0,001 | 0,717 |
| f__Coriobacteriaceae_g__Collinsella_s__aerofaciens_3 | 0,000 | 0,099 | 0,000 | 0,802 |
| f__Lachnospiraceae_g__Lachnospira_s___23 | 0,028 | 0,113 | 0,054 | 0,834 |
| f__Alcaligenaceae_g__Sutterella_s___66 | 0,794 | 0,117 | 0,858 | 0,834 |
| f__Veillonellaceae_g__Veillonella_s__dispar_52 | 0,003 | 0,134 | 0,009 | 0,902 |
| f__Enterococcaceae_g__Enterococcus_s___60 | 0,420 | 0,179 | 0,513 | 0,966 |
| f__Bifidobacteriaceae_g__Bifidobacterium_s__longum_63 | 0,286 | 0,182 | 0,385 | 0,966 |
| f__Enterococcaceae_g__Enterococcus_s___69 | 0,678 | 0,188 | 0,759 | 0,966 |
| f__Enterococcaceae_g__Enterococcus_s___27 | 0,000 | 0,194 | 0,000 | 0,966 |
| f__Lachnospiraceae_g__[Ruminococcus]_s___4 | 0,809 | 0,196 | 0,866 | 0,966 |
| f__Lachnospiraceae_g__[Ruminococcus]_s__gnavus_22 | 0,033 | 0,202 | 0,060 | 0,966 |
| f__Lachnospiraceae_g___s___184 | 0,510 | 0,203 | 0,617 | 0,966 |
| f__Clostridiaceae_g__Sarcina_s___1 | 0,984 | 0,226 | 0,984 | 0,966 |
| f__Bacteroidaceae_g__Bacteroides_s__fragilis_14 | 0,000 | 0,239 | 0,000 | 0,966 |
| f__Pasteurellaceae_g__Haemophilus_s__parainfluenzae_21 | 0,087 | 0,244 | 0,141 | 0,966 |
| f__Coriobacteriaceae_g___s___34 | 0,001 | 0,263 | 0,003 | 0,966 |
| f__Bacteroidaceae_g__Bacteroides_s__caccae_4 | 0,003 | 0,264 | 0,008 | 0,966 |

| | | | | |
|---|---|---|---|---|
| f__Lachnospiraceae_g___s___290 | 0,000 | 0,270 | 0,002 | 0,966 |
| f__Enterobacteriaceae_g__Trabulsiella_s___14 | 0,000 | 0,295 | 0,001 | 0,966 |
| f__Lactobacillaceae_g__Lactobacillus_s__zeae_19 | 0,004 | 0,312 | 0,010 | 0,966 |
| f__Gemellaceae_g___s___6 | 0,000 | 0,319 | 0,000 | 0,966 |
| f__Enterococcaceae_g__Enterococcus_s___61 | 0,604 | 0,341 | 0,696 | 0,966 |
| f__Clostridiaceae_g__Clostridium_s__celatum_18 | 0,674 | 0,341 | 0,759 | 0,966 |
| f__Peptostreptococcaceae_NA_NA_1 | 0,283 | 0,343 | 0,385 | 0,966 |
| f__Lachnospiraceae_g__Dorea_s__formicigenerans_8 | 0,000 | 0,346 | 0,002 | 0,966 |
| f__Enterobacteriaceae_g__Enterobacter_s__cloacae_11 | 0,575 | 0,352 | 0,669 | 0,966 |
| f__Bifidobacteriaceae_g__Bifidobacterium_s___5 | 0,526 | 0,352 | 0,630 | 0,966 |
| f__Clostridiaceae_g__SMB53_s___47 | 0,351 | 0,364 | 0,443 | 0,966 |
| f__Lactobacillaceae_g__Lactobacillus_s__zeae_10 | 0,000 | 0,367 | 0,000 | 0,966 |
| f__Porphyromonadaceae_g__Parabacteroides_s___41 | 0,001 | 0,371 | 0,003 | 0,966 |
| f__Pasteurellaceae_g__Haemophilus_s__parainfluenzae_20 | 0,121 | 0,379 | 0,188 | 0,966 |
| f__Veillonellaceae_g__Veillonella_s___29 | 0,000 | 0,380 | 0,000 | 0,966 |
| f__Enterobacteriaceae_g__Citrobacter_s___17 | 0,117 | 0,387 | 0,184 | 0,966 |
| f__Alcaligenaceae_g__Sutterella_s___43 | 0,009 | 0,390 | 0,020 | 0,966 |
| f__Staphylococcaceae_g__Staphylococcus_s__epidermidis_4 | 0,007 | 0,392 | 0,017 | 0,966 |
| f__Bacteroidaceae_g__Bacteroides_s__fragilis_15 | 0,000 | 0,398 | 0,000 | 0,966 |
| f__Bacteroidaceae_g__Bacteroides_s___151 | 0,020 | 0,412 | 0,042 | 0,966 |
| f__Streptococcaceae_g__Streptococcus_s__luteciae_3 | 0,001 | 0,423 | 0,003 | 0,966 |
| f__Bifidobacteriaceae_g__Bifidobacterium_s___95 | 0,262 | 0,439 | 0,364 | 0,966 |
| f__Bacteroidaceae_g__Bacteroides_s__ovatus_10 | 0,001 | 0,442 | 0,004 | 0,966 |
| f__Alcaligenaceae_g__Sutterella_s___17 | 0,416 | 0,443 | 0,513 | 0,966 |
| f__Bacteroidaceae_g__Bacteroides_s___149 | 0,000 | 0,444 | 0,000 | 0,966 |
| f__Veillonellaceae_g__Veillonella_s___22 | 0,132 | 0,451 | 0,197 | 0,966 |

| | | | | |
|---|---|---|---|---|
| ***f__Enterobacteriaceae_g__Escherichia_s__coli_26*** | 0,876 | 0,463 | 0,929 | 0,966 |
| ***f__Enterobacteriaceae_NA_NA_39*** | 0,000 | 0,467 | 0,002 | 0,966 |
| ***f__Clostridiaceae_g__Clostridium_s__neonatale_5*** | 0,541 | 0,471 | 0,637 | 0,966 |
| ***f__Clostridiaceae_g__Clostridium_s__perfringens_19*** | 0,003 | 0,481 | 0,009 | 0,969 |
| ***f__Veillonellaceae_g__Veillonella_s__dispar_49*** | 0,061 | 0,500 | 0,102 | 0,991 |
| ***f__Enterobacteriaceae_NA_NA_66*** | 0,001 | 0,509 | 0,003 | 0,993 |
| ***f__Bacteroidaceae_g__Bacteroides_s___150*** | 0,009 | 0,517 | 0,020 | 0,993 |
| ***f__Lachnospiraceae_g__Blautia_s___109*** | 0,028 | 0,548 | 0,054 | 0,997 |
| ***f__Bifidobacteriaceae_g__Bifidobacterium_s___10*** | 0,005 | 0,552 | 0,011 | 0,997 |
| ***f__Ruminococcaceae_g__Ruminococcus_s___171*** | 0,014 | 0,585 | 0,029 | 0,997 |
| ***f__Pasteurellaceae_g__Haemophilus_s__parainfluenzae_23*** | 0,031 | 0,586 | 0,057 | 0,997 |
| ***f__Bifidobacteriaceae_g__Bifidobacterium_s__bifidum_3*** | 0,004 | 0,596 | 0,011 | 0,997 |
| ***f__Peptostreptococcaceae_g__Peptostreptococcus_s__anaerobius_1*** | 0,003 | 0,614 | 0,008 | 0,997 |
| ***f__Veillonellaceae_g__Veillonella_s__dispar_94*** | 0,693 | 0,621 | 0,769 | 0,997 |
| ***f__Enterobacteriaceae_g__Klebsiella_s___15*** | 0,705 | 0,626 | 0,775 | 0,997 |
| ***f__Bifidobacteriaceae_g__Bifidobacterium_s__longum_5*** | 0,000 | 0,638 | 0,001 | 0,997 |
| ***f__Ruminococcaceae_g__Oscillospira_s___272*** | 0,980 | 0,657 | 0,984 | 0,997 |
| ***f__Porphyromonadaceae_g__Parabacteroides_s__distasonis_11*** | 0,094 | 0,661 | 0,149 | 0,997 |
| ***f__Bacteroidaceae_g__Bacteroides_s__ovatus_72*** | 0,001 | 0,662 | 0,004 | 0,997 |
| ***f__Veillonellaceae_g__Veillonella_s__dispar_11*** | 0,026 | 0,667 | 0,052 | 0,997 |
| ***f__Bifidobacteriaceae_g__Bifidobacterium_s___11*** | 0,326 | 0,677 | 0,420 | 0,997 |
| ***f__Veillonellaceae_g__Veillonella_s___21*** | 0,032 | 0,680 | 0,058 | 0,997 |
| ***f__Enterococcaceae_g__Enterococcus_s___26*** | 0,314 | 0,690 | 0,408 | 0,997 |
| ***f__Streptococcaceae_g__Streptococcus_s___65*** | 0,000 | 0,692 | 0,001 | 0,997 |
| ***f__Lactobacillaceae_g__Lactobacillus_s__zeae_3*** | 0,000 | 0,707 | 0,000 | 0,997 |
| ***f__Bacteroidaceae_g__Bacteroides_s__ovatus_71*** | 0,002 | 0,713 | 0,006 | 0,997 |

| | | | | |
|---|---|---|---|---|
| f__Porphyromonadaceae_g__Parabacteroides_s__distasonis_10 | 0,035 | 0,716 | 0,062 | 0,997 |
| f__Bifidobacteriaceae_g__Bifidobacterium_s___7 | 0,001 | 0,717 | 0,003 | 0,997 |
| f__Veillonellaceae_g__Megamonas_s___2 | 0,029 | 0,728 | 0,054 | 0,997 |
| f__Enterobacteriaceae_g__Enterobacter_s__cloacae_8 | 0,006 | 0,730 | 0,014 | 0,997 |
| f__Bacteroidaceae_g__Bacteroides_s__ovatus_73 | 0,001 | 0,731 | 0,004 | 0,997 |
| f__Erysipelotrichaceae_g___s___50 | 0,082 | 0,738 | 0,134 | 0,997 |
| f__Veillonellaceae_g__Veillonella_s___26 | 0,004 | 0,746 | 0,010 | 0,997 |
| f__Bacteroidaceae_g__Bacteroides_s___152 | 0,035 | 0,750 | 0,062 | 0,997 |
| f__Enterobacteriaceae_g__Proteus_s___1 | 0,122 | 0,766 | 0,188 | 0,997 |
| f__Veillonellaceae_g__Veillonella_s__dispar_26 | 0,272 | 0,783 | 0,375 | 0,997 |
| f__Streptococcaceae_g__Streptococcus_s___20 | 0,057 | 0,807 | 0,099 | 0,997 |
| f__Enterobacteriaceae_NA_NA_61 | 0,002 | 0,811 | 0,006 | 0,997 |
| f__Bifidobacteriaceae_g__Bifidobacterium_s___6 | 0,981 | 0,821 | 0,984 | 0,997 |
| f__Veillonellaceae_g__Veillonella_s__dispar_50 | 0,232 | 0,824 | 0,326 | 0,997 |
| f__Enterobacteriaceae_g__Escherichia_s__coli_25 | 0,000 | 0,829 | 0,000 | 0,997 |
| f__Bifidobacteriaceae_g__Bifidobacterium_s__longum_6 | 0,067 | 0,837 | 0,111 | 0,997 |
| f__Bacteroidaceae_g__Bacteroides_s__caccae_3 | 0,002 | 0,838 | 0,006 | 0,997 |
| f__Pasteurellaceae_g__Haemophilus_s__parainfluenzae_24 | 0,024 | 0,848 | 0,048 | 0,997 |
| f__Clostridiaceae_g__Clostridium_s__neonatale_22 | 0,001 | 0,853 | 0,003 | 0,997 |
| f__Veillonellaceae_g__Veillonella_s___25 | 0,149 | 0,859 | 0,219 | 0,997 |
| f__Bacteroidaceae_g__Bacteroides_s___131 | 0,738 | 0,869 | 0,804 | 0,997 |
| f__Veillonellaceae_g__Veillonella_s__dispar_87 | 0,950 | 0,872 | 0,976 | 0,997 |
| f__Clostridiaceae_g__Clostridium_NA_6 | 0,022 | 0,883 | 0,044 | 0,997 |
| f__Clostridiaceae_g__Clostridium_s__perfringens_32 | 0,001 | 0,893 | 0,003 | 0,997 |
| f__Bifidobacteriaceae_g__Bifidobacterium_s___47 | 0,929 | 0,900 | 0,976 | 0,997 |
| f__Enterobacteriaceae_g__Klebsiella_s___16 | 0,000 | 0,905 | 0,001 | 0,997 |

| | | | | |
|---|---|---|---|---|
| ***f__Bacteroidaceae_g__Bacteroides_s__ovatus_29*** | 0,058 | 0,909 | 0,099 | 0,997 |
| ***f__Veillonellaceae_g__Veillonella_s__dispar_23*** | 0,000 | 0,920 | 0,003 | 0,997 |
| ***f__Clostridiaceae_g__Clostridium_s__neonatale_23*** | 0,307 | 0,924 | 0,404 | 0,997 |
| ***f__Enterobacteriaceae_g__Trabulsiella_s___3*** | 0,001 | 0,925 | 0,003 | 0,997 |
| ***f__Enterobacteriaceae_g__Klebsiella_s___14*** | 0,000 | 0,942 | 0,000 | 0,997 |
| ***f__Lachnospiraceae_g___s___353*** | 0,002 | 0,978 | 0,006 | 0,997 |
| ***f__Ruminococcaceae_g__Oscillospira_s___268*** | 0,153 | 0,979 | 0,222 | 0,997 |
| ***f__Clostridiaceae_g__Clostridium_s__neonatale_7*** | 0,007 | 0,980 | 0,017 | 0,997 |
| ***f__Ruminococcaceae_g__Oscillospira_s___269*** | 0,409 | 0,983 | 0,510 | 0,997 |
| ***f__Lachnospiraceae_g__Epulopiscium_s___2*** | 0,002 | 0,984 | 0,007 | 0,997 |
| ***f__Enterobacteriaceae_g__Escherichia_s__coli_20*** | 0,001 | 0,993 | 0,003 | 0,997 |
| ***f__Clostridiaceae_g__Clostridium_s__perfringens_13*** | 0,951 | 0,997 | 0,976 | 0,997 |
| ***f__Ruminococcaceae_g__Butyricicoccus_s__pullicaecorum_36*** | 0,001 | 0,997 | 0,003 | 0,997 |

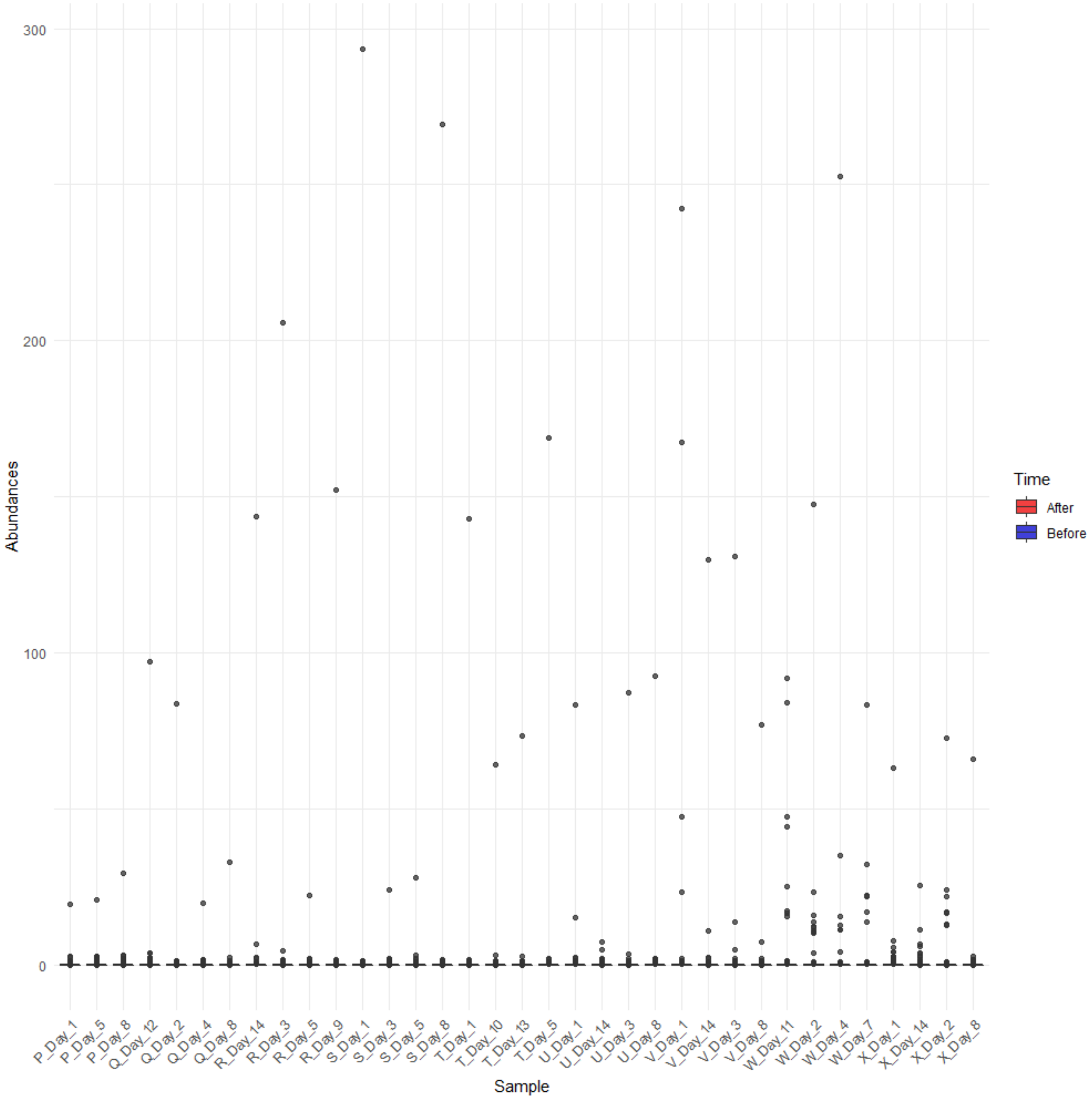

*Figure 38 Non-normalized metabolite distributions per sample for DIMS metabolites. The colours indicate the sampling times, either before (blue) or after (red) the introduction of solid foods.*

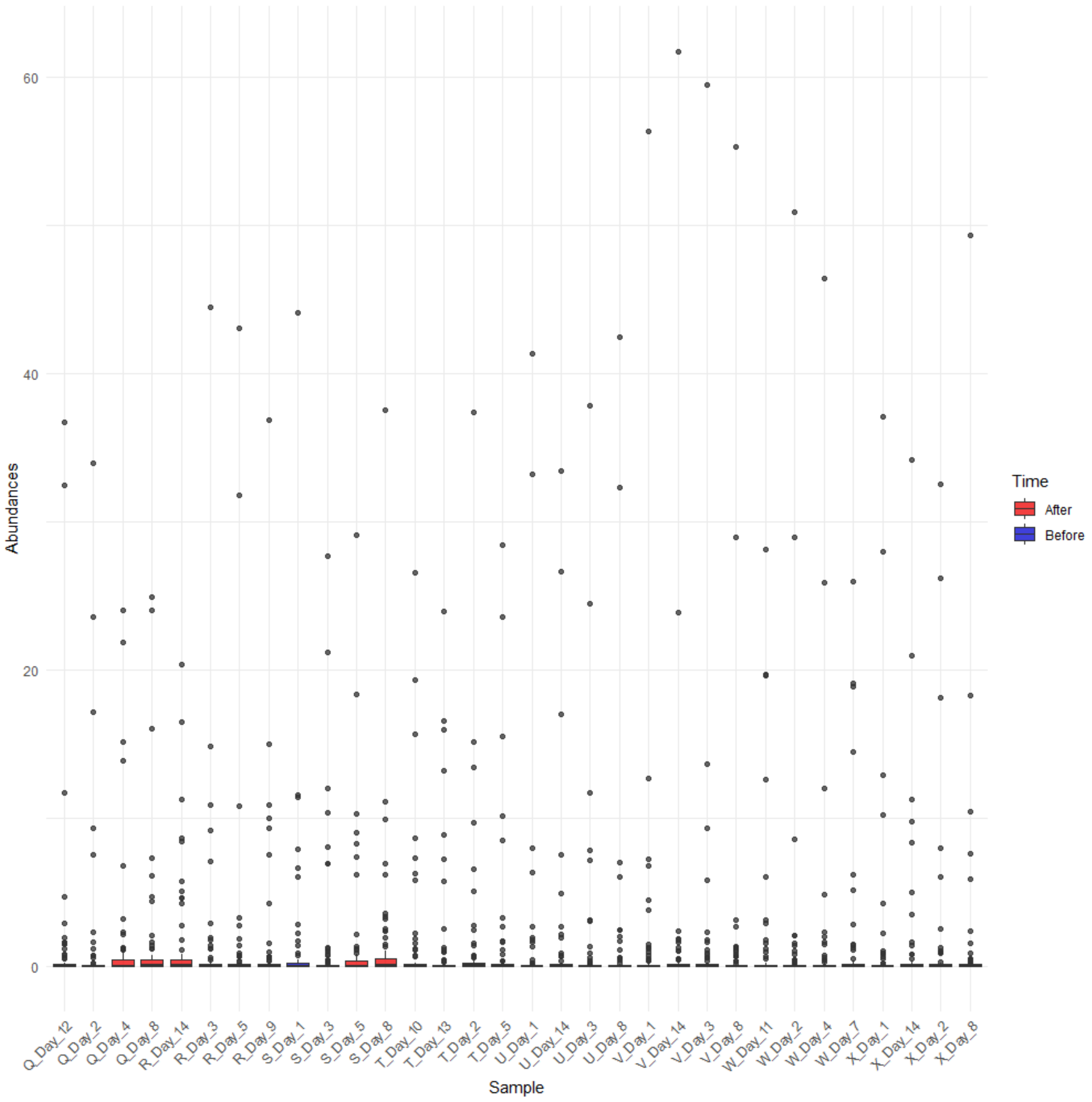

*Figure 39 Non-normalized metabolite distributions per sample for UPLC metabolites. The colours indicate the sampling times, either before (blue) or after (red) the introduction of solid foods.*

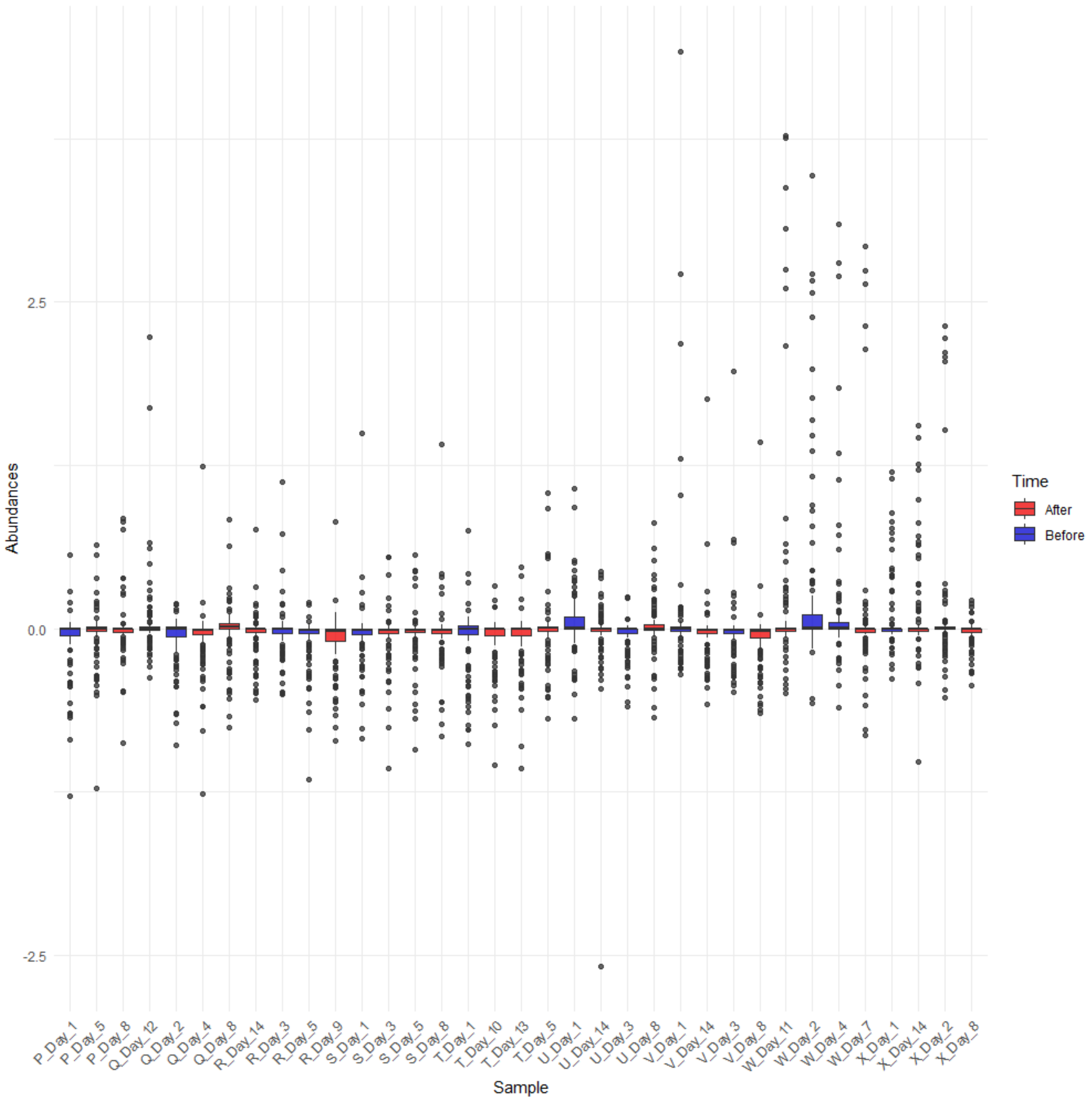

*Figure 40 Normalized metabolite distributions for DIMS metabolites after a logarithmic-pareto normalization. The colours indicate the sampling times, either before (blue) or after (red) the introduction of solid foods.*

.

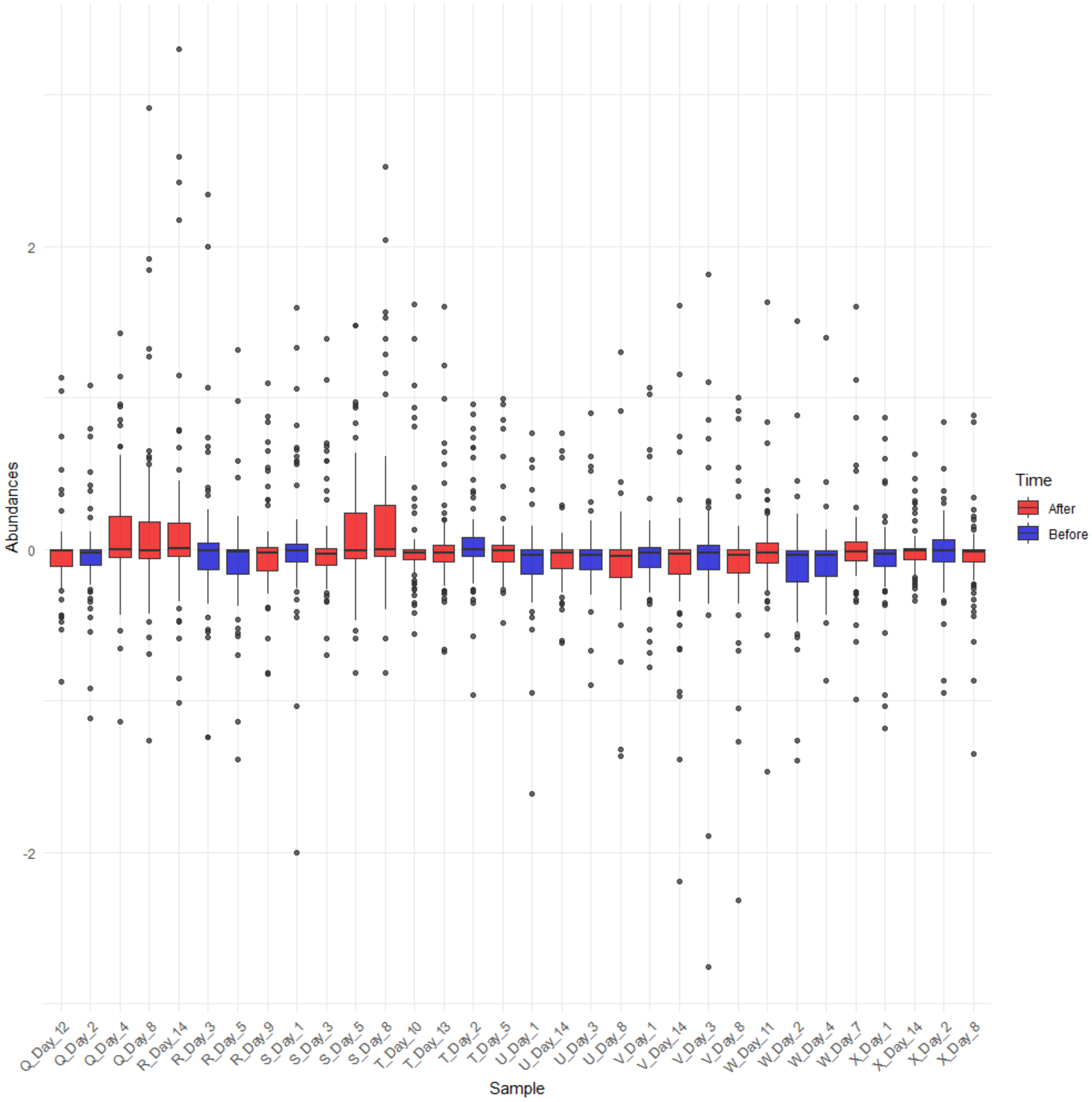

*Figure 41 Normalized metabolite distributions for UPLC metabolites after a logarithmic-pareto normalization. The colours indicate the sampling times, either before (blue) or after (red) the introduction of solid foods.*

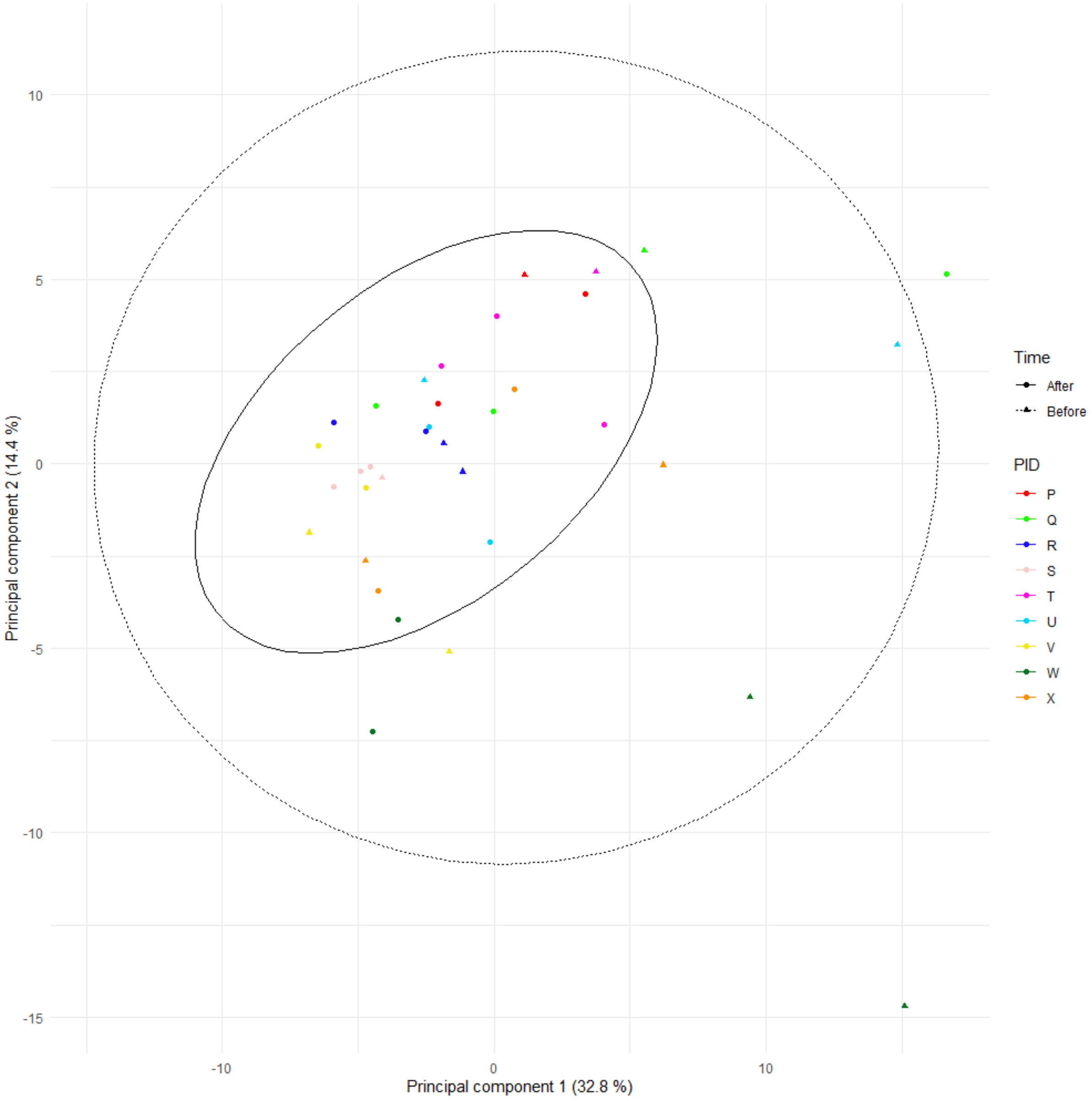

*Figure 42 PCA for the DIMS data set. Colours denote samples from the same patient. The ellipses and shape of the dots describe the data from either before or after the introduction of solid foods.*

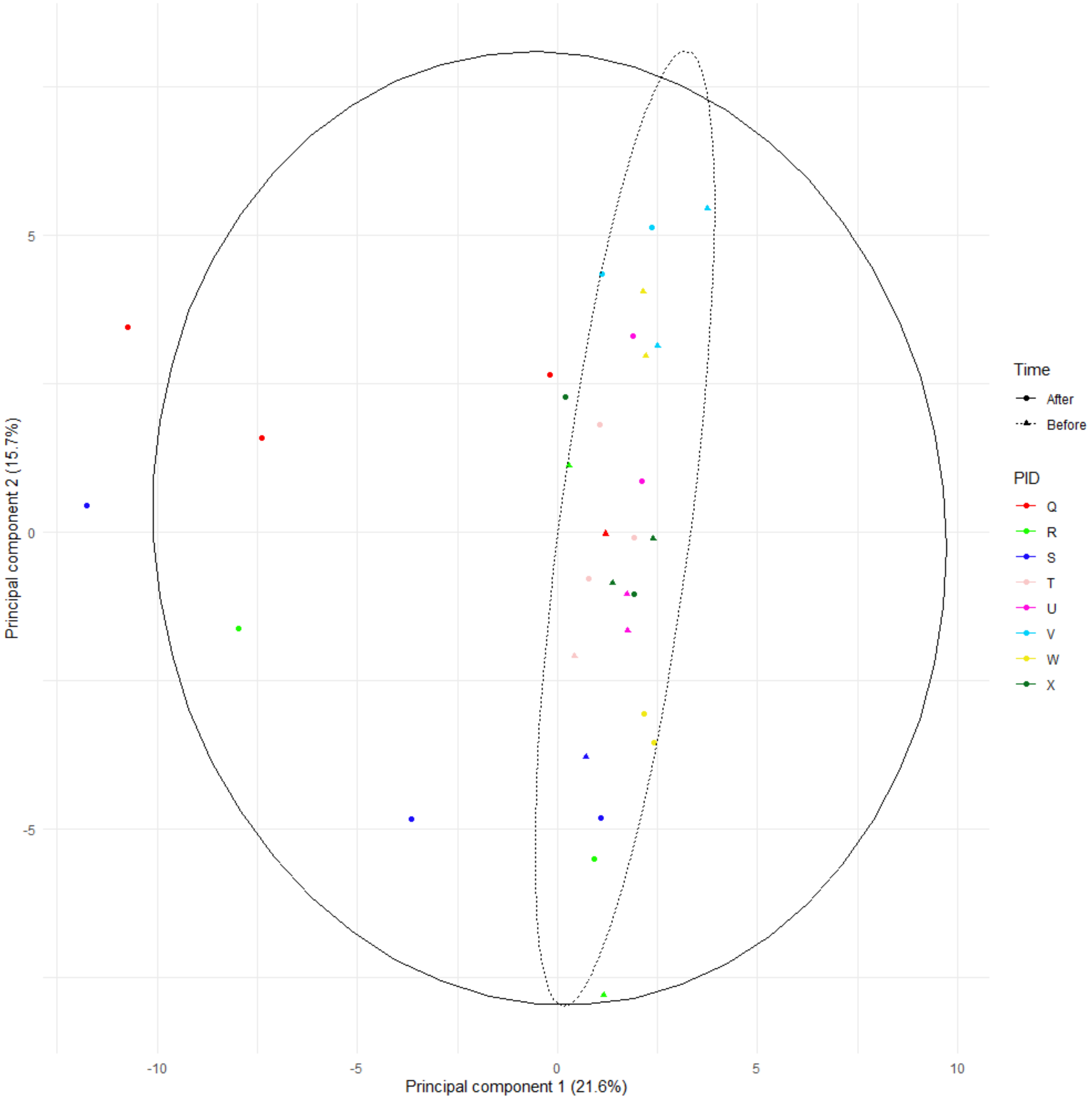

*Figure 43 PCA for the UPLC data set. Colours denote samples from the same patient. The ellipses and shape of the dots describe the data from either before or after the introduction of solid foods.*

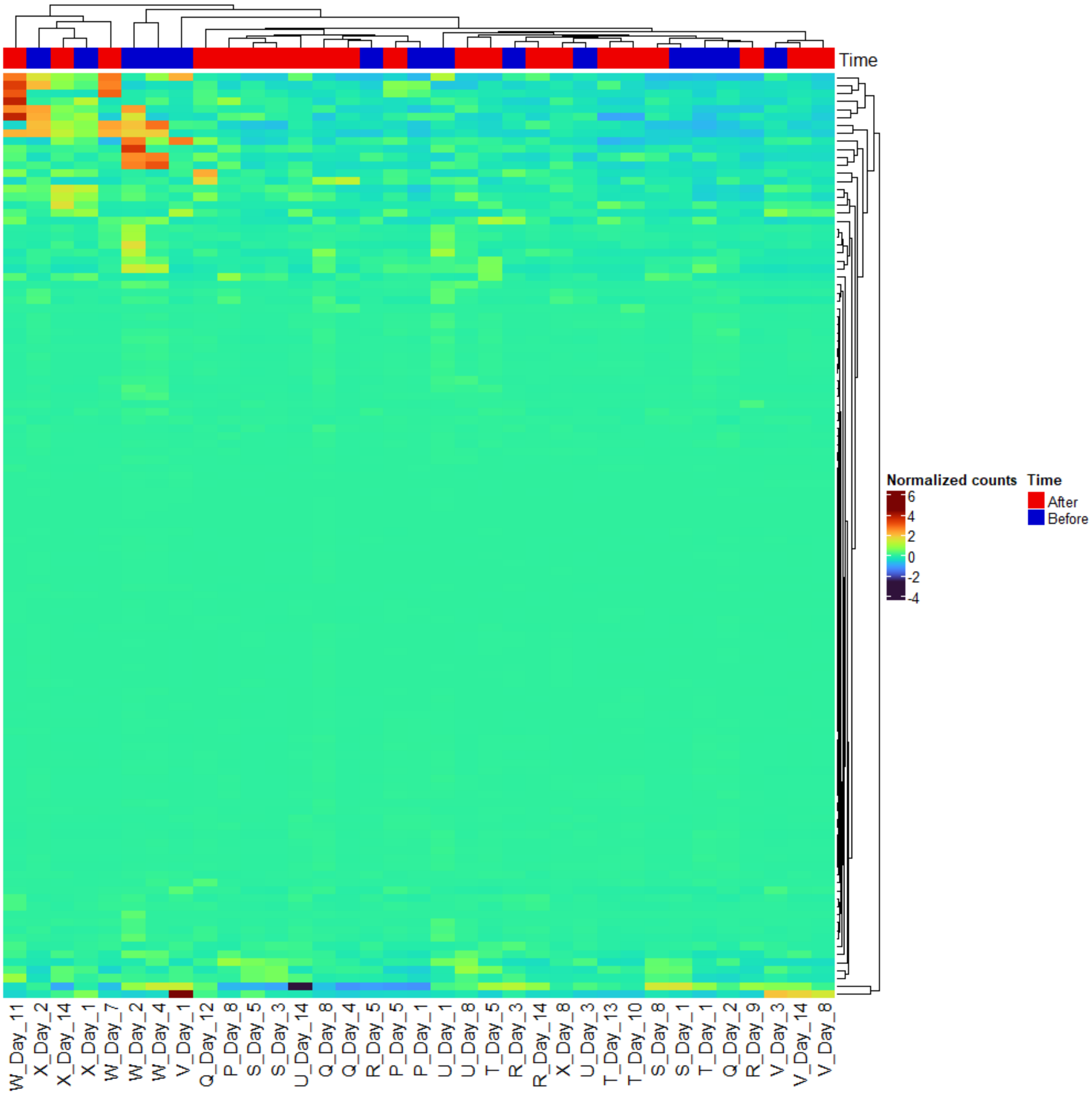

*Figure 44 Heatmap of the DIMS data set. The normalized counts are shown. Dendrograms show the clustering of metabolites in the rows and samples in the columns. A bar plot is constructed to indicate the sample period, before (blue) or after (red) the introduction solid foods.*

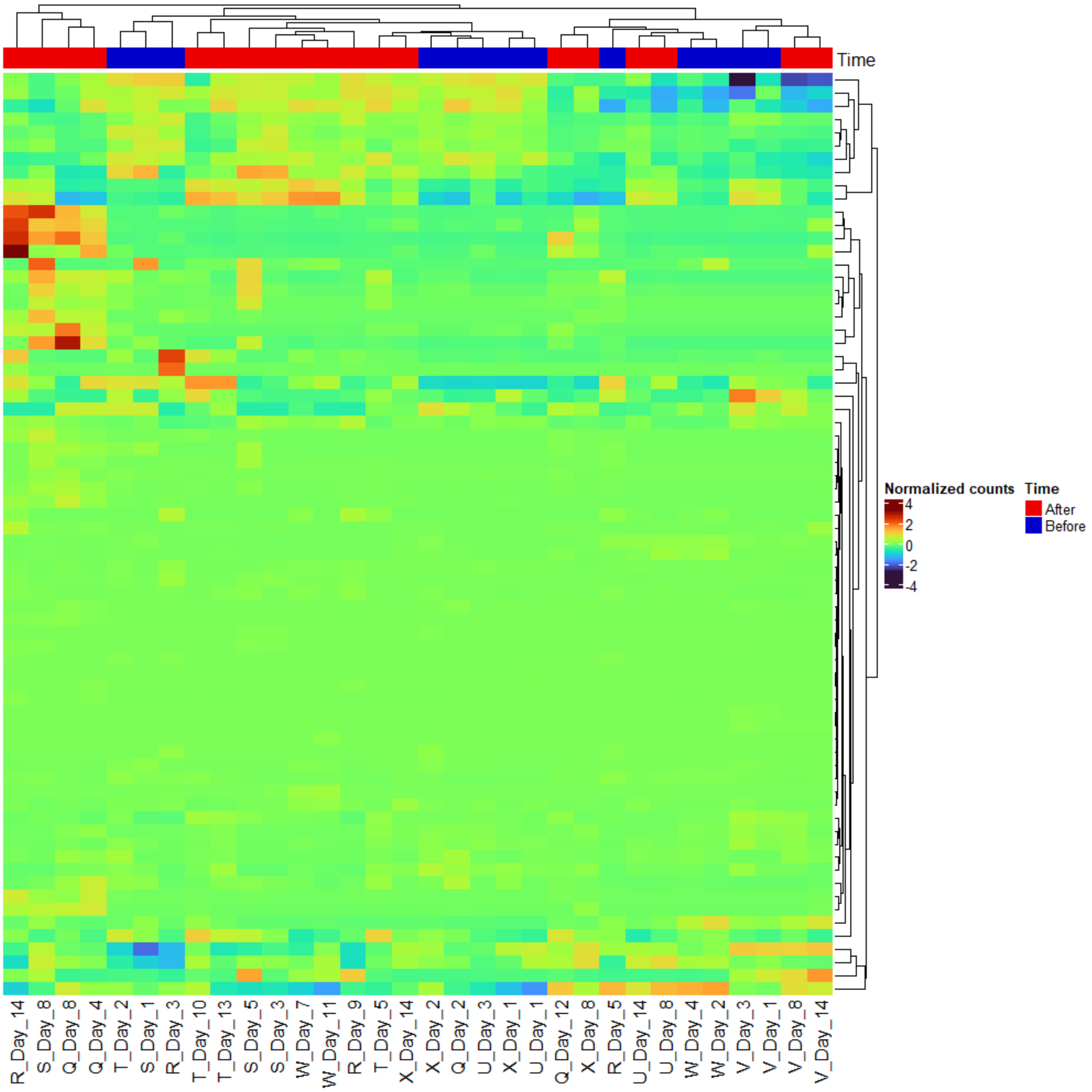

*Figure 45 Heatmap of the UPLC data set. The normalized counts are shown. Dendrograms show the clustering of metabolites in the rows and samples in the columns. A barplot is constructed to indicate the sample period, before (blue) or after (red) the introduction solid foods.*

*Table 12 Results for the differential abundance testing of NMR metabolites.*

| Metabolite | Unadjusted P-value | Adjusted P-value |
|---|---|---|
| **Galactose** | 0,007 | 0,237 |
| **Butyrate** | 0,012 | 0,237 |
| **5-Aminopentanoate** | 0,038 | 0,393 |
| **Xylose** | 0,038 | 0,393 |
| **Uracil** | 0,072 | 0,562 |
| **Lactate** | 0,083 | 0,562 |
| **?-Alanine** | 0,099 | 0,562 |
| **Propionate** | 0,110 | 0,562 |
| **Acetate** | 0,143 | 0,651 |
| **Isobutyrate** | 0,213 | 0,816 |

| | | |
|---|---|---|
| **2-Oxoisocaproate** | 0,219 | 0,816 |
| **myo-Inositol** | 0,289 | 0,894 |
| **Methanol** | 0,304 | 0,894 |
| **Isovalerate** | 0,354 | 0,894 |
| **Cytosine** | 0,390 | 0,894 |
| **Dimethylamine** | 0,409 | 0,894 |
| **Glycerol** | 0,429 | 0,894 |
| **Tyramine** | 0,459 | 0,894 |
| **Acetoin** | 0,469 | 0,894 |
| **3-Hydroxyisovalerate** | 0,469 | 0,894 |
| **Fucose** | 0,469 | 0,894 |
| **phenylacetic acid** | 0,511 | 0,894 |
| **Fumarate** | 0,533 | 0,894 |
| **Propylene glycol** | 0,533 | 0,894 |
| **Ethanol** | 0,556 | 0,894 |
| **Succinate** | 0,567 | 0,894 |
| **Methylamine** | 0,590 | 0,896 |
| **4-Aminobutyrate** | 0,625 | 0,916 |
| **Aspartate** | 0,724 | 0,984 |
| **N-Acetylglutamate** | 0,788 | 0,984 |
| **Trimethylamine** | 0,827 | 0,984 |
| **Malonate** | 0,827 | 0,984 |
| **Pyroglutamate** | 0,827 | 0,984 |
| **Valerate** | 0,827 | 0,984 |
| **Xanthine** | 0,880 | 0,984 |
| **Cadaverine** | 0,880 | 0,984 |
| **4-Hydroxyphenylacetate** | 0,933 | 0,984 |
| **Citrate** | 0,960 | 0,984 |
| **Isopropanol** | 0,960 | 0,984 |
| **Pyruvate** | 0,960 | 0,984 |
| **Formate** | 1,000 | 1,000 |

*Table 13 Results for the differential abundance testing of DIMS metabolites.*

| Metabolite | Unadjusted P-value | Adjusted P-value |
|---|---|---|
| **LYSOC14:0** | 0,004 | 0,286 |
| **C18:1OH** | 0,005 | 0,286 |
| **PC40:1AA** | 0,020 | 0,368 |
| **LYSOC26:0** | 0,021 | 0,368 |
| **C5MDC** | 0,022 | 0,368 |
| **PC38:0AA** | 0,023 | 0,368 |
| **LYSOC20:3** | 0,027 | 0,368 |
| **PC40:6AA** | 0,034 | 0,368 |
| **PC32:2AA** | 0,035 | 0,368 |
| **PC40:6AE** | 0,038 | 0,368 |

| | | |
|---|---|---|
| **C14** | 0,045 | 0,368 |
| **C18:2** | 0,049 | 0,368 |
| **PC40:2AA** | 0,051 | 0,368 |
| **C4:1** | 0,053 | 0,368 |
| **C16** | 0,053 | 0,368 |
| **Asparagine** | 0,057 | 0,368 |
| **LYSOC28:1** | 0,057 | 0,368 |
| **C18:1** | 0,057 | 0,368 |
| **C16:2OH** | 0,064 | 0,386 |
| **C18** | 0,066 | 0,386 |
| **Glucose** | 0,072 | 0,386 |
| **C16:1** | 0,077 | 0,386 |
| **C12** | 0,077 | 0,386 |
| **LYSOC28:0** | 0,080 | 0,386 |
| **PC36:0AA** | 0,089 | 0,401 |
| **PC36:6AA** | 0,095 | 0,401 |
| **C16:1OH** | 0,096 | 0,401 |
| **C5OH** | 0,099 | 0,401 |
| **C12:1** | 0,110 | 0,401 |
| **LYSOC26:1** | 0,110 | 0,401 |
| **C0** | 0,110 | 0,401 |
| **LYSOC17:0** | 0,117 | 0,401 |
| **LYSOC18:0** | 0,117 | 0,401 |
| **C14:1** | 0,117 | 0,401 |
| **20:2SM** | 0,125 | 0,415 |
| **Methylhistidine** | 0,143 | 0,461 |
| **C16:2** | 0,152 | 0,465 |
| **C2** | 0,152 | 0,465 |
| **C6:1** | 0,157 | 0,468 |
| **22:2SMOH** | 0,167 | 0,485 |
| **LYSOC16:1** | 0,189 | 0,519 |
| **Glycine** | 0,195 | 0,519 |
| **C14:1OH** | 0,195 | 0,519 |
| **Putrescine** | 0,207 | 0,519 |
| **Methionine** | 0,207 | 0,519 |
| **C16OH** | 0,213 | 0,519 |
| **Threonine** | 0,219 | 0,519 |
| **Diacetylspermine** | 0,219 | 0,519 |
| **PC36:0AE** | 0,219 | 0,519 |
| **C5DC** | 0,232 | 0,537 |
| **LYSOC18:1** | 0,245 | 0,558 |
| **C14:2OH** | 0,252 | 0,563 |
| **LYSOC20:4** | 0,259 | 0,568 |
| **Serotonin** | 0,274 | 0,577 |
| **Creatinine** | 0,274 | 0,577 |

| | | |
|---|---|---|
| **C10** | 0,289 | 0,598 |
| **LYSOC16:0** | 0,297 | 0,599 |
| **C14:2** | 0,304 | 0,599 |
| **24:1SMOH** | 0,304 | 0,599 |
| **Leucine** | 0,337 | 0,608 |
| **C3OH** | 0,345 | 0,608 |
| **Trimethylamine N-oxide** | 0,349 | 0,608 |
| **C3** | 0,354 | 0,608 |
| **Glutamic acid** | 0,354 | 0,608 |
| **C12DC** | 0,363 | 0,608 |
| **Glutamine** | 0,363 | 0,608 |
| **Tryptophan** | 0,372 | 0,608 |
| **Proline** | 0,372 | 0,608 |
| **Tyrosine** | 0,372 | 0,608 |
| **18:1SM** | 0,372 | 0,608 |
| **22:1SMOH** | 0,372 | 0,608 |
| **Phenylalanine** | 0,400 | 0,644 |
| **C5:1DC** | 0,409 | 0,650 |
| **C7DC** | 0,429 | 0,663 |
| **PC38:6AA** | 0,429 | 0,663 |
| **Taurine** | 0,439 | 0,669 |
| **14:1SMOH** | 0,459 | 0,672 |
| **Methionine-sulfoxide** | 0,469 | 0,672 |
| **Serine** | 0,469 | 0,672 |
| **Betaine** | 0,469 | 0,672 |
| **16:1SMOH** | 0,469 | 0,672 |
| **Acetyl-Ornithine** | 0,489 | 0,692 |
| **C10:1** | 0,511 | 0,715 |
| **Histidine** | 0,556 | 0,747 |
| **16:1SM** | 0,556 | 0,747 |
| **18:0SM** | 0,556 | 0,747 |
| **Kynurenine** | 0,567 | 0,747 |
| **C4** | 0,567 | 0,747 |
| **Alanine** | 0,578 | 0,752 |
| **trans-hydroxy-Proline** | 0,590 | 0,752 |
| **C9** | 0,590 | 0,752 |
| **C3:1** | 0,601 | 0,758 |
| **Valine** | 0,625 | 0,780 |
| **Ornithine** | 0,649 | 0,793 |
| **16:0SM** | 0,649 | 0,793 |
| **LYSOC24:0** | 0,674 | 0,814 |
| **Choline** | 0,724 | 0,865 |
| **Citrulline** | 0,749 | 0,878 |
| **Creatine** | 0,749 | 0,878 |
| **C6** | 0,775 | 0,899 |

| | | |
|---|---|---|
| **Isoleucine** | 0,814 | 0,900 |
| **Spermine** | 0,814 | 0,900 |
| **Asymmetric dimethylarginine** | 0,814 | 0,900 |
| **C8** | 0,827 | 0,900 |
| **LYSOC18:2** | 0,827 | 0,900 |
| **C4OH** | 0,840 | 0,900 |
| **C5:1** | 0,840 | 0,900 |
| **Arginine** | 0,853 | 0,900 |
| **Total dimethylarginine** | 0,853 | 0,900 |
| **C5** | 0,853 | 0,900 |
| **Spermidine** | 0,920 | 0,949 |
| **Sarcosine** | 0,920 | 0,949 |
| **Aspartic acid** | 0,933 | 0,949 |
| **Lysine** | 0,933 | 0,949 |
| **alpha-Aminoadipic acid** | 0,946 | 0,954 |
| **C10:2** | 0,960 | 0,960 |

*Table 14 Results for the differential abundance testing of UPLC metabolites.*

| Metabolite | Unadjusted P-value | Adjusted P-value |
|---|---|---|
| **Lithocholic acid** | 0,014 | 0,252 |
| **Tauroallocholic acid** | 0,024 | 0,252 |
| **Taurocholic acid** | 0,027 | 0,252 |
| **Ursodeoxycholic acid-24-Glucuronide** | 0,028 | 0,252 |
| **Nordeoxycholic acid** | 0,034 | 0,252 |
| **Glycoursodeoxycholic acid** | 0,035 | 0,252 |
| **Beta-Muricholic acid** | 0,038 | 0,252 |
| **Isolithocholic acid-3-sulfate** | 0,038 | 0,252 |
| **Glycoursodeoxycholic acid-3-sulfate** | 0,039 | 0,252 |
| **Glycohyocholic acid** | 0,042 | 0,252 |
| **7-Ketodeoxycholic acid** | 0,042 | 0,252 |
| **Tauroursodeoxycholic acid/Taurohyodeoxycholic acid** | 0,046 | 0,252 |
| **Tauro-alpha-muricholic acid** | 0,047 | 0,252 |
| **Taurochenodeoxycholic acid** | 0,066 | 0,328 |
| **Isolithocholic acid** | 0,077 | 0,361 |
| **Omega-Muricholic acid** | 0,091 | 0,365 |
| **Deoxycholic acid-3-Glucuronide** | 0,094 | 0,365 |
| **Allocholic acid** | 0,099 | 0,365 |
| **Ursodeoxycholic acid-3-sulfate** | 0,099 | 0,365 |
| **Glycoallocholic acid** | 0,111 | 0,386 |
| **Chenodeoxycholic acid-3-Glucuronide** | 0,116 | 0,386 |
| **Glycochenodeoxycholic acid-3-sulfate** | 0,123 | 0,390 |
| **Glycochenodeoxycholic acid** | 0,135 | 0,409 |
| **Deoxycholic acid** | 0,145 | 0,422 |
| **Glycodeoxycholic acid** | 0,161 | 0,450 |

| | | |
|---|---|---|
| **Deoxycholic acid-3-sulfate** | 0,167 | 0,450 |
| **Taurochenodeoxycholic acid-3-sulfate** | 0,179 | 0,464 |
| **7-Ketolithocholic acid** | 0,192 | 0,480 |
| **Taurohyocholic acid** | 0,233 | 0,532 |
| **Lithocholic acid-3-sulfate** | 0,234 | 0,532 |
| **Apocholic acid** | 0,239 | 0,532 |
| **Glycocholic acid-3-sulfate** | 0,249 | 0,532 |
| **Glycolithocholic acid** | 0,252 | 0,532 |
| **Taurolithocholic acid** | 0,261 | 0,532 |
| **Glycocholic acid** | 0,266 | 0,532 |
| **Norursodeoxycholic acid** | 0,274 | 0,534 |
| **Tauroursodeoxycholic acid-3-sulfate** | 0,283 | 0,535 |
| **Ursocholic acid** | 0,300 | 0,553 |
| **Chenodeoxycholic acid-3-sulfate** | 0,337 | 0,597 |
| **Murocholic acid** | 0,357 | 0,597 |
| **GlycoLithocholic acid-3-sulfate** | 0,368 | 0,597 |
| **Tauro-omega-muricholic acid** | 0,376 | 0,597 |
| **Norcholic acid** | 0,378 | 0,597 |
| **Ursodeoxycholic acid-3-Glucuronide** | 0,382 | 0,597 |
| **3-Dehydrocholic acid** | 0,399 | 0,597 |
| **Glycohyodoxycholic acid-3-sulfate** | 0,401 | 0,597 |
| **Taurolithocholic acid-3-sulfate** | 0,401 | 0,597 |
| **12-Ketolithocholic acid** | 0,417 | 0,608 |
| **Dehydrolithocholic acid** | 0,445 | 0,636 |
| **Taurodeoxycholic acid** | 0,554 | 0,776 |
| **Glycodeoxycholic acid-3-sulfate** | 0,612 | 0,830 |
| **Tauro-beta-muricholic acid** | 0,617 | 0,830 |
| **Glycoallocholic acid-3-sulfate** | 0,645 | 0,852 |
| **Hyodeoxycholic acid** | 0,687 | 0,890 |
| **12-Ketochenodeoxycholic acid** | 0,701 | 0,892 |
| **Chenodeoxycholic acid** | 0,759 | 0,922 |
| **Taurodehydrocholic acid** | 0,762 | 0,922 |
| **Dehydrocholic acid** | 0,764 | 0,922 |
| **Alloisolithocholic acid** | 0,779 | 0,924 |
| **Glycohyodeoxycholic acid** | 0,800 | 0,928 |
| **Deoxycholic acid-24-Glucuronide** | 0,809 | 0,928 |
| **Isodeoxycholic acid** | 0,832 | 0,939 |
| **Dioxolithocholic acid** | 0,878 | 0,944 |
| **Lambda-Muricholic acid** | 0,878 | 0,944 |
| **Allocholic acid-3-sulfate** | 0,878 | 0,944 |
| **Chenodeoxycholic acid-24-Glucuronide** | 0,890 | 0,944 |
| **Cholic acid** | 0,939 | 0,981 |
| **Cholic acid-3-sulfate** | 0,969 | 0,998 |
| **Alpha-Muricholic acid** | 1,000 | 1,000 |
| **Ursodeoxycholic acid** | 1,000 | 1,000 |

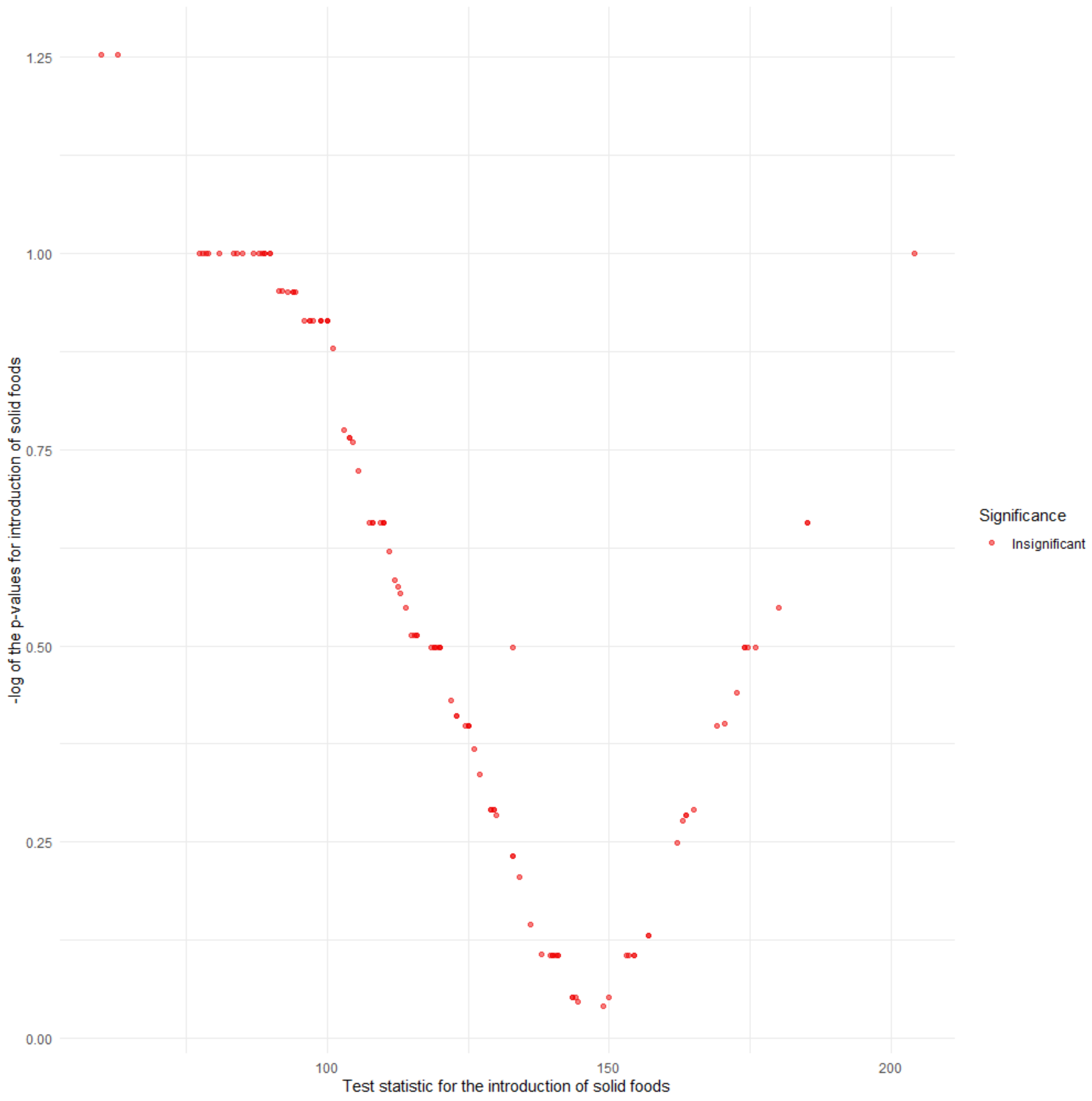

*Figure 46 Results for the differentially abundant metabolites after the introduction of solid foods in the DIMS data. Results are shown after adjustment for multiplicity. The x-axis denotes the test statistic for the introduction of solid food and the y-axis the negative logarithmic transformed p-values after adjustment of multiplicity. Colours denote the significance of the results, red being insignificant.*

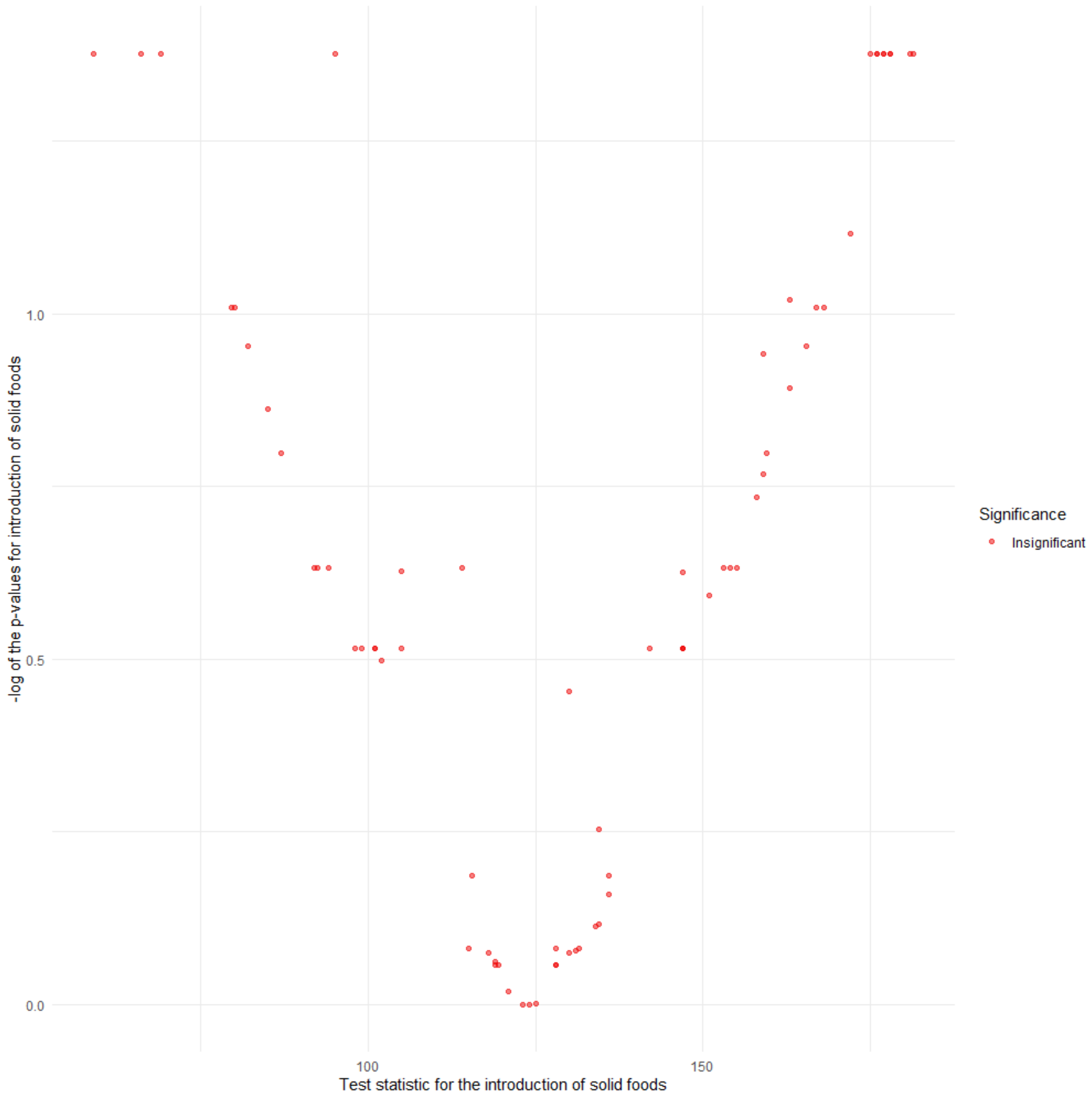

*Figure 47 Results for the differentially abundant metabolites after the introduction of solid foods in the UPLC data. Results are shown after adjustment for multiplicity. The x-axis denotes the test statistic for the introduction of solid food and the y-axis the negative logarithmic transformed p-values after adjustment of multiplicity. Colours denote the significance of the results, red being insignificant.*

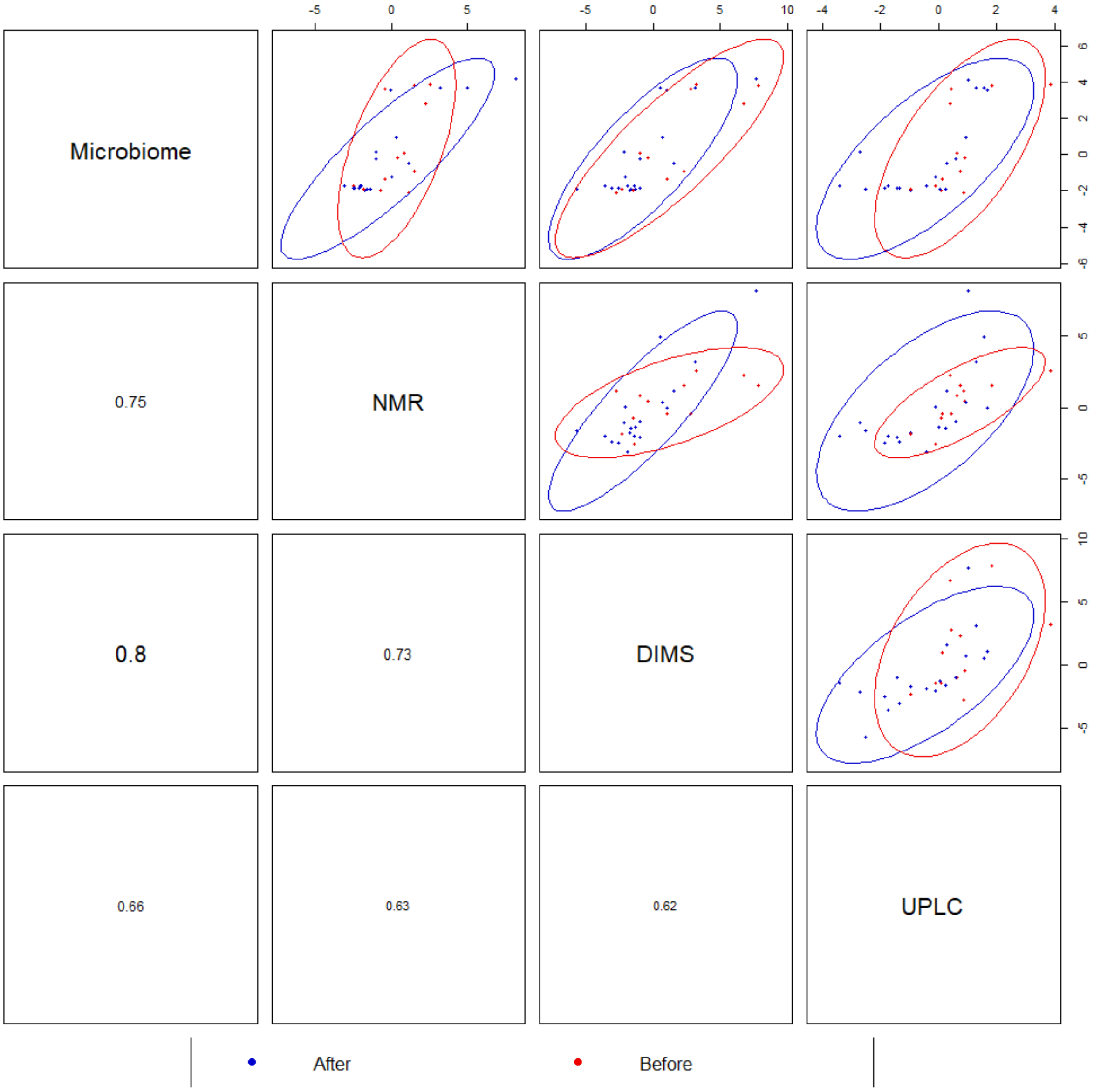

*Figure 48 Correlation plot of the correlation between the second principal component of each data set. Dots denote samples and the colour shows the sampling time, either before the introduction of solid foods in red or after in blue. 95% confidence ellipses are added.*

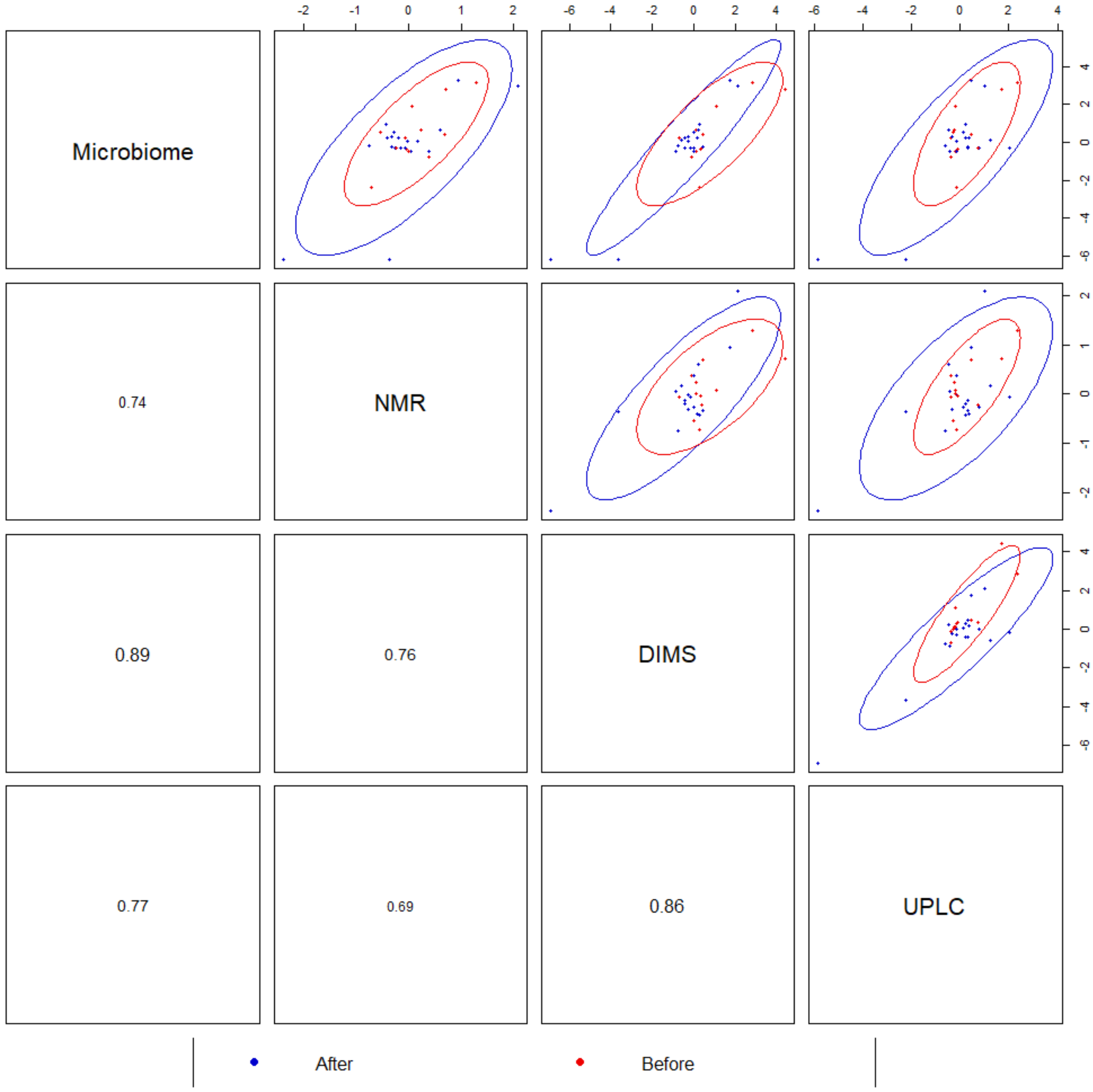

*Figure 49 Correlation plot of the correlation between the third principal component of each data set. Dots denote samples and the colour shows the sampling time, either before the introduction of solid foods in red or after in blue. 95% confidence ellipses are added.*

*Table 15 All correlations computed by DIABLO higher than 0.7 or lower than -0.7.*

| Feature 1 | Feature 2 | Correlation |
|---|---|---:|
| C16:1 | Beta-Muricholic acid | -0,88769203 |
| C16:1 | 7-Ketodeoxycholic acid | -0,882213841 |
| C18:1 | Beta-Muricholic acid | -0,870726108 |
| PC36:6AA | Beta-Muricholic acid | -0,869141362 |
| PC36:6AA | 7-Ketodeoxycholic acid | -0,862738085 |
| C18:1 | 7-Ketodeoxycholic acid | -0,862125096 |
| C14 | Beta-Muricholic acid | -0,860232182 |
| C18:2 | Beta-Muricholic acid | -0,854848119 |
| C14 | 7-Ketodeoxycholic acid | -0,850418315 |
| C18:2 | 7-Ketodeoxycholic acid | -0,846520871 |
| C16:1OH | Beta-Muricholic acid | -0,824894444 |
| C16:2 | Beta-Muricholic acid | -0,824215993 |
| C16:1OH | 7-Ketodeoxycholic acid | -0,821744379 |
| C18:1OH | Beta-Muricholic acid | -0,821296233 |
| C16:2 | 7-Ketodeoxycholic acid | -0,820452509 |
| C18:1OH | 7-Ketodeoxycholic acid | -0,815386147 |
| C0 | Beta-Muricholic acid | -0,814648621 |
| C0 | 7-Ketodeoxycholic acid | -0,80836201 |
| C16:2OH | Beta-Muricholic acid | -0,802497806 |
| C16:2OH | 7-Ketodeoxycholic acid | -0,798966732 |
| Veillonellaceae Veillonella dispar | ?-Alanine | -0,789317942 |
| Lachnospiraceae [Ruminococcus] gnavus | ?-Alanine | -0,789317942 |
| Enterobacteriaceae Proteus | ?-Alanine | -0,789317942 |
| C5OH | Beta-Muricholic acid | -0,776994913 |
| C10:1 | Beta-Muricholic acid | -0,774338219 |
| C10:1 | 7-Ketodeoxycholic acid | -0,772491425 |
| C5OH | 7-Ketodeoxycholic acid | -0,768206613 |
| Ruminococcaceae Ruminococcus | Taurohyocholic acid | -0,767399389 |
| Erysipelotrichaceae | Taurohyocholic acid | -0,765364663 |
| Veillonellaceae Veillonella dispar | Uracil | -0,762794531 |
| Lachnospiraceae [Ruminococcus] gnavus | Uracil | -0,762794531 |
| Enterobacteriaceae Proteus | Uracil | -0,762794531 |
| C16OH | Beta-Muricholic acid | -0,753138986 |
| C16OH | 7-Ketodeoxycholic acid | -0,750543155 |
| PC32:2AA | Beta-Muricholic acid | -0,749937238 |
| Creatinine | Beta-Muricholic acid | -0,746142098 |
| PC32:2AA | 7-Ketodeoxycholic acid | -0,745112136 |
| Creatinine | 7-Ketodeoxycholic acid | -0,738308935 |
| C7DC | Beta-Muricholic acid | -0,722354549 |

| | | |
|---|---|---|
| Veillonellaceae Veillonella dispar | Cadaverine | -0,71930027 |
| Lachnospiraceae [Ruminococcus] gnavus | Cadaverine | -0,71930027 |
| Enterobacteriaceae Proteus | Cadaverine | -0,71930027 |
| C7DC | 7-Ketodeoxycholic acid | -0,717466922 |
| Bifidobacteriaceae Bifidobacterium | Taurohyocholic acid | -0,717297256 |
| C8 | Beta-Muricholic acid | -0,714560717 |
| C8 | 7-Ketodeoxycholic acid | -0,712845839 |
| LYSOC20:4 | Beta-Muricholic acid | -0,70798938 |
| 22:2SMOH | Beta-Muricholic acid | -0,705149839 |
| 22:2SMOH | 7-Ketodeoxycholic acid | -0,702468005 |
| Streptococcaceae Streptococcus infantis | Beta-Muricholic acid | -0,70236118 |
| ?-Alanine | Taurohyocholic acid | -0,702266986 |
| Ruminococcaceae Ruminococcus | Tauro-omega-muricholic acid | -0,701966348 |
| C2 | Beta-Muricholic acid | -0,701779805 |
| Erysipelotrichaceae | Tauro-omega-muricholic acid | -0,701311202 |
| Isobutyrate | Aspartic acid | 0,700023239 |
| Bifidobacteriaceae Bifidobacterium bifidum | C4:1 | 0,70188278 |
| Streptococcaceae Streptococcus infantis | C0 | 0,702560254 |
| Isobutyrate | Dehydrolithocholic acid | 0,70314199 |
| Streptococcaceae Streptococcus infantis | PC36:6AA | 0,705699271 |
| Histidine | Dehydrolithocholic acid | 0,707736847 |
| Veillonellaceae Veillonella dispar | Tauro-omega-muricholic acid | 0,707762079 |
| Lachnospiraceae [Ruminococcus] gnavus | Tauro-omega-muricholic acid | 0,707762079 |
| Enterobacteriaceae Proteus | Tauro-omega-muricholic acid | 0,707762079 |
| Bifidobacteriaceae Bifidobacterium | Uracil | 0,710321949 |
| Ruminococcaceae Ruminococcus | Uracil | 0,716377456 |
| Streptococcaceae Streptococcus infantis | C16:1 | 0,717215246 |
| Erysipelotrichaceae | Uracil | 0,719442528 |
| Erysipelotrichaceae | Cadaverine | 0,721798187 |
| Ruminococcaceae Butyricicoccus pullicaecorum | C4OH | 0,722147489 |
| Ruminococcaceae Ruminococcus | Cadaverine | 0,723421631 |
| Erysipelotrichaceae | Tyramine | 0,725244397 |
| Bifidobacteriaceae Bifidobacterium bifidum | C3OH | 0,728653604 |
| Bifidobacteriaceae Bifidobacterium | ?-Alanine | 0,732051047 |
| Isobutyrate | Murocholic acid | 0,735927157 |
| Bacteroidaceae Bacteroides caccae | C4OH | 0,737804245 |
| Ruminococcaceae Ruminococcus | Tyramine | 0,738471544 |

| | | |
|---|---|---:|
| Putrescine | Isolithocholic acid | 0,739839138 |
| Bacteroidaceae Bacteroides ovatus | C4OH | 0,740276092 |
| Ruminococcaceae Butyricicoccus pullicaecorum | Aspartic acid | 0,740963338 |
| Bacteroidaceae Bacteroides ovatus | C4OH | 0,741976235 |
| Bacteroidaceae Bacteroides ovatus | Isobutyrate | 0,746515642 |
| Bacteroidaceae Bacteroides ovatus | C4OH | 0,75247486 |
| Bacteroidaceae Bacteroides ovatus | Isobutyrate | 0,752752909 |
| Bacteroidaceae Bacteroides caccae | Aspartic acid | 0,756204061 |
| Bacteroidaceae Bacteroides ovatus | Aspartic acid | 0,758230455 |
| Bacteroidaceae Bacteroides ovatus | Dehydrolithocholic acid | 0,759832374 |
| *Bacteroidaceae Bacteroides ovatus* | Aspartic acid | 0,759876152 |
| *Erysipelotrichaceae* | ?-Alanine | 0,76082977 |
| *Ruminococcaceae Ruminococcus* | ?-Alanine | 0,760938006 |
| *Bacteroidaceae Bacteroides ovatus* | Dehydrolithocholic acid | 0,761011879 |
| *Bacteroidaceae Bacteroides caccae* | Isobutyrate | 0,763906616 |
| *Veillonellaceae Veillonella dispar* | Taurohyocholic acid | 0,764620733 |
| *Lachnospiraceae [Ruminococcus] gnavus* | Taurohyocholic acid | 0,764620733 |
| *Enterobacteriaceae Proteus* | Taurohyocholic acid | 0,764620733 |
| Isobutyrate | Putrescine | 0,764811128 |
| *Ruminococcaceae Butyricicoccus pullicaecorum* | Isobutyrate | 0,765340698 |
| *Bacteroidaceae Bacteroides ovatus* | Isobutyrate | 0,767140866 |
| *Ruminococcaceae Butyricicoccus pullicaecorum* | Dehydrolithocholic acid | 0,768685196 |
| *Bacteroidaceae Bacteroides ovatus* | Aspartic acid | 0,771347441 |
| *Bacteroidaceae Bacteroides ovatus* | Dehydrolithocholic acid | 0,773190982 |
| *Bacteroidaceae Bacteroides caccae* | Dehydrolithocholic acid | 0,780218175 |
| Histidine | Isolithocholic acid | 0,790448713 |
| *Bacteroidaceae Bacteroides ovatus* | Isolithocholic acid | 0,81319899 |
| *Bacteroidaceae Bacteroides ovatus* | Isolithocholic acid | 0,817977192 |
| Isobutyrate | Histidine | 0,820963253 |
| *Ruminococcaceae Butyricicoccus pullicaecorum* | Putrescine | 0,822471693 |
| *Bacteroidaceae Bacteroides ovatus* | Isolithocholic acid | 0,828928165 |
| *Ruminococcaceae Butyricicoccus pullicaecorum* | Isolithocholic acid | 0,832302718 |
| *Bacteroidaceae Bacteroides ovatus* | Putrescine | 0,836633228 |
| Isobutyrate | Isolithocholic acid | 0,836857337 |
| *Bacteroidaceae Bacteroides caccae* | Putrescine | 0,838642035 |
| *Bacteroidaceae Bacteroides ovatus* | Putrescine | 0,83872407 |
| *Bacteroidaceae Bacteroides caccae* | Isolithocholic acid | 0,839877831 |
| *Bacteroidaceae Bacteroides ovatus* | Putrescine | 0,846225341 |

|  |  |  |
|---|---|---|
| *Ruminococcaceae Butyricicoccus pullicaecorum* | Histidine | 0,860613214 |
| *Bacteroidaceae Bacteroides ovatus* | Histidine | 0,871923142 |
| *Bacteroidaceae Bacteroides ovatus* | Histidine | 0,874901277 |
| *Bacteroidaceae Bacteroides caccae* | Histidine | 0,876022024 |
| *Bacteroidaceae Bacteroides ovatus* | Histidine | 0,882641746 |

# Addendum R-code

## Microbiome data analysis

```
# Libraries and colour palette
library(tidyverse)
library(phyloseq)
library(skimr)
library(data.table)
library(microbiome)
library(ggridges)
library(microViz)
library(ggtext)
library(matrixStats)
library(metagMisc)
library(lme4)
library(nlme)
library(ANCOMBC)
library(DT)
library(MESS)
ColourPalette <- c("#FB0000", "#1CFC00", "#1C0DFA", "#F9C8C8", "#FF0DDF",
"#00D1FB", "#F0E816", "#0D7222", "#FC8D00", "#5D3581", "#BB005A",
"#35FDDA", "#D06CFE", "#7F3B1C", "#FF95E5", "#668C92", "#AAF580",
"#819BFE", "#948626", "#D8C7FF", "#CBE7BD", "#FE8D88", "#E74738",
"#922663", "#FFC24F", "#FC0DAD", "#A2009F", "#584242", "#2271A6",
"#0DFA98", "#98EEF5", "#5C2EB9", "#FB007C", "#B17DA2", "#9F2A00",
"#FE7AB2")

# Reading the data
Metadata <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metadata/Metadata.csv", sep = ";", row.names = 1)
MicrobiomeData <- read.csv(file =
"C:/Users/Gebruiker/Documents/School/Master of Statistics - Bioinformatics
(2021-2022)/Master Thesis Bioinformatics/Data/Microbiome
data/Code_LucKi_mergtab_nochim_LucKi_v34_transposed_ISS.csv", sep = ";",
row.names = 1)
TaxaData <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Microbiome data/taxa.csv", sep = ";", row.names = 1)

# Summarizing the data
##Skimming the data
skim(Metadata)
## Selecting the most important variables (According to me)
Metadata %>%
  group_by(Infant.nr) %>%
  dplyr::select(Infant.nr, PID, PID_Day, Sex.F.M, Length_cm, Weight_gramm,
Age_days_SF_intro) %>%
  summarise(mean_length = mean(Length_cm),
            mean_weight = mean(Weight_gramm),
            mean_age_intro = mean(Age_days_SF_intro)) -> Metadata_summary
Metadata_summary
table(Metadata$Infant.nr)
table(Metadata$Sex.F.M)
mean(Metadata_summary$mean_length)
sd(Metadata_summary$mean_length)
mean(Metadata_summary$mean_weight)
sd(Metadata_summary$mean_weight)
mean(Metadata_summary$mean_age_intro)
sd(Metadata_summary$mean_age_intro)
```

```
# Microbiome data
## Phyloseq object creation
###Looking into making a phyloseq object first
# Selecting a taxa indicator first
#TaxaInf <- TaxaData[1:7]
# Transforming the metadata before - after
Metadata %>% mutate(across(Time, factor, levels=c("Before","After"))) ->
Metadata
# Transforming into matrices for Phyloseq
MicrobiomeInf <- as.matrix(MicrobiomeData)
TaxaInf <- as.matrix(TaxaData)
MetaInf <- as.matrix(Metadata)

# Creating Phyloseq objects
ASV <- otu_table(MicrobiomeInf, taxa_are_rows = TRUE)
TAX <- tax_table(TaxaInf)
SampleData <- sample_data(Metadata)
PhySeqObj <- phyloseq(ASV, TAX, SampleData)

# Testing
PhySeqObj

## Missing taxa case
###Only 8330 taxa present? Original taxa data contained data on 9313 ASV
and microbiome data had 8787 rows/ASV
# Reason
Count_ASVs <- data.frame(rownames(MicrobiomeData))
Taxa_ASVs <- data.frame(rownames(TaxaData))
# Look at the length of the intersect
Intersection <- intersect(Count_ASVs$rownames.MicrobiomeData.,
Taxa_ASVs$rownames.TaxaData.)
length(Intersection)
# This indicates that there isn't any taxonomic information on 457 ASVs
present in the original original dataset so Phyloseq filters them out.
# Identify the taxa!
Unidentified_Taxa <- setdiff(Count_ASVs$rownames.MicrobiomeData.,
Intersection)
# Check their counts in the original count data
Microbiome_2 <- MicrobiomeData
Microbiome_2$ASVs <- rownames(Microbiome_2)
Microbiome_2 %>% filter(ASVs %in% Unidentified_Taxa) -> Microbiome_2
Microbiome_2$Sums <- rowSums(Microbiome_2[1:87])
# Amount of non-zers
Microbiome_2 %>% filter(Sums != 0) -> Microbiome_2
# Contains 31 samples
openxlsx::write.xlsx(Microbiome_2, file =
"C:/Users/Gebruiker/Documents/School/Master of Statistics - Bioinformatics
(2021-2022)/Master Thesis Bioinformatics/Data/Microbiome
data/Microbiome_2.xlsx")

## Filtering
# Filter based on prevalence
PhySeqObjFiltered_1 <- phyloseq_filter_prevalence(PhySeqObj, prev.trh =
0.05, abund.trh = NULL)
# Filter based on relative abundance = 0.01%
minTotRelAbun = 1e-4
x = taxa_sums(PhySeqObjFiltered_1)
keepTaxa = (x / sum(x)) > minTotRelAbun
PhySeqObjFiltered = prune_taxa(keepTaxa, PhySeqObjFiltered_1)
PhySeqObjFiltered
```

```r
### Effects of filtering
# Extract the abundances
Abundances_unfiltered <- abundances(PhySeqObj)
Abundances_filtered <- abundances(PhySeqObjFiltered)
# Transpose them to calculate the library sizes per sample
Abundances_unfiltered_transposed <- data.frame(t(Abundances_unfiltered))
Abundances_filtered_transposed <- data.frame(t(Abundances_filtered))
# Library sizes
Abundances_unfiltered_transposed %>%
  mutate(LibrarySizes = rowSums(Abundances_unfiltered_transposed)) %>%
  dplyr::select(LibrarySizes) -> Abundances_unfiltered_LibSizes
Abundances_filtered_transposed %>%
  mutate(LibrarySizes = rowSums(Abundances_filtered_transposed)) %>%
  dplyr::select(LibrarySizes) -> Abundances_filtered_LibSizes
# Indicator variables
Abundances_unfiltered_LibSizes$Filter <- "Unfiltered"
Abundances_unfiltered_LibSizes$Sample <-
row.names(Abundances_unfiltered_LibSizes)
Abundances_filtered_LibSizes$Filter <- "Filtered"
Abundances_filtered_LibSizes$Sample <-
row.names(Abundances_filtered_LibSizes)
# Merge all
Abundances_total <- Abundances_unfiltered_LibSizes
Abundances_total <- rbind(Abundances_total, Abundances_filtered_LibSizes)
# Plot
Abundances_total %>%
  ggplot(aes(x = LibrarySizes, fill = Filter)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("mediumblue", "red2")) +
  theme_minimal() +
  xlab("Library sizes")

### Fixing the taxa
PhySeqObjFiltered %>% tax_fix(min_length = 4) -> PhySeqObjFiltered

## Family level object
# Create a family Phyloseq object
Family <- tax_glom(PhySeqObjFiltered, taxrank = "Family")
PhySeqObjFam <- Family
PhySeqObjFam

## Exploratory data analysis
###Also calculating relative abundances to make comparisons possible
# TSS
PhySeqObjFilteredRA <- transform_sample_counts(PhySeqObjFiltered,
function(x) { x/sum(x)})
PhySeqObjFamRA <- transform_sample_counts(PhySeqObjFam, function(x) {
x/sum(x)})
# CLR transformation
PhySeqObjFilteredRA_CLR <- transform(PhySeqObjFiltered, "clr")
PhySeqObjFamRA_CLR <- transform(PhySeqObjFam, "clr")

### PCA Using CoDa
####Scree-plot
# PCA via phyloseq
ord_clr <- phyloseq::ordinate(PhySeqObjFilteredRA_CLR, "RDA")
#Plot scree plot
plot_scree(ord_clr) +
  geom_bar(stat="identity", fill = "blue") +
  labs(x = "\nPrincipal component", y = "Proportion of Variance\n") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, size = 6))
#Examine eigenvalues and % prop. variance explained
head(ord_clr$CA$eig)
sapply(ord_clr$CA$eig[1:5], function(x) x / sum(ord_clr$CA$eig))
#### PCA-plot
clr1 <- ord_clr$CA$eig[1] / sum(ord_clr$CA$eig)
clr2 <- ord_clr$CA$eig[2] / sum(ord_clr$CA$eig)
plot_ordination(PhySeqObjFilteredRA_CLR, ord_clr, type="samples",
color="PID", shape = "Time") +
  geom_point(size = 2) +
  coord_fixed(clr2 / clr1) +
  theme_minimal() +
  stat_ellipse(aes(group = Time, linetype = Time)) +
  scale_colour_manual(values = ColourPalette)


### Alpha diversities
#At ASV level
plot_richness(PhySeqObjFiltered, measures=c("Observed", "Shannon",
"InvSimpson"), x="Time", color = "PID") +
  theme_minimal() +
  scale_colour_manual(values = ColourPalette)
#### Further alpha diversity research
# Extract all richness estimates
Richness_PhySeq <- estimate_richness(PhySeqObjFiltered,
measures=c("Observed", "Shannon", "InvSimpson"))
Rownames_Richness_Physeq <- rownames(Richness_PhySeq)
Richness_PhySeq$ID <- Rownames_Richness_Physeq
# Extract meaningful variables from Metadata
Rownames_Metadata <- rownames(Metadata)
Metadata$ID <- Rownames_Metadata
Metadata %>% dplyr::select(ID, Infant.nr, PID, PID_Day, Sex.F.M, Age_days,
Time, Substudy_timepoint_days, Length_cm, Weight_gramm,
Additional_comments) -> Metadata_Small
merge(Richness_PhySeq, Metadata_Small) -> Metadata_New
# Transform additional comments
Metadata_New %>%
  mutate(Disease = if_else(Additional_comments == "None", 0, 1)) ->
Metadata_New
Metadata_New %>% gather(key = "Alpha Diversity", value = "Value", Observed,
Shannon, InvSimpson) -> Metadata_New
##### Longitudinal trends
Metadata_New %>%
  ggplot(aes(x = Substudy_timepoint_days, y = Value, group = PID, colour =
PID)) +
  geom_line() +
  theme_minimal() +
  scale_colour_manual(values = ColourPalette) +
  facet_wrap(.~`Alpha Diversity`, scales = "free") +
  xlab("Timepoint (in days)")
##### Age
Metadata_New %>%
  ggplot(aes(x = Age_days, y = Value, group = PID, colour = PID)) +
  geom_point(aes(shape = Time)) +
  geom_smooth(se = FALSE) +
  theme_minimal() +
  scale_colour_manual(values = ColourPalette) +
  scale_shape_manual(values=c(15, 16)) +
  facet_wrap(.~`Alpha Diversity`, scales = "free") +
  xlab("Age (in days)")
##### Length (in cm)
```

```
Metadata_New %>%
  ggplot(aes(x = Length_cm, y = Value, group = PID, colour = PID)) +
  geom_point(aes(shape = Time)) +
  theme_minimal() +
  scale_colour_manual(values = ColourPalette) +
  scale_shape_manual(values=c(15, 16)) +
  facet_wrap(.~`Alpha Diversity`, scales = "free") +
  xlab("Length (in cm)")
##### Weight (in gramm)
Metadata_New %>%
  ggplot(aes(x = Weight_gramm, y = Value, group = PID, colour = PID)) +
  geom_point(aes(shape = Time)) +
  theme_minimal() +
  scale_colour_manual(values = ColourPalette) +
  scale_shape_manual(values=c(15, 16)) +
  facet_wrap(.~`Alpha Diversity`, scales = "free") +
  xlab("Weight (in gram)")
##### Gender
Metadata_New %>%
  ggplot(aes(x = Sex.F.M, y = Value, fill = Sex.F.M)) +
  geom_boxplot(alpha = 0.75) +
  scale_fill_manual(name = "Gender", labels = c("Female", "Male"), values =
c("red2", "mediumblue")) +
  theme_minimal() +
  facet_wrap(.~`Alpha Diversity`, scales = "free") +
  xlab("Gender")
##### Disease
Metadata_New %>%
  ggplot(aes(x = as.factor(Disease), y = Value, fill = as.factor(Disease)))
+
  geom_boxplot(alpha = 0.75) +
  scale_fill_manual(name = "Diseased", labels = c("No", "Yes"), values =
c("red2", "mediumblue")) +
  theme_minimal() +
  facet_wrap(.~`Alpha Diversity`, scales = "free") +
  xlab("Diseased") +
  scale_x_discrete(labels = c("0" = "No", "1" = "Yes"))
### MicroViz visualizations
#### Barplots
PhySeqObjFiltered %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant P
PhySeqObjFiltered %>%
  ps_filter(PID == "P") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant Q
PhySeqObjFiltered %>%
  ps_filter(PID == "Q") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant R
PhySeqObjFiltered %>%
  ps_filter(PID == "R") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant S
PhySeqObjFiltered %>%
  ps_filter(PID == "S") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant T
PhySeqObjFiltered %>%
  ps_filter(PID == "T") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant U
PhySeqObjFiltered %>%
  ps_filter(PID == "U") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant V
PhySeqObjFiltered %>%
  ps_filter(PID == "V") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant W
PhySeqObjFiltered %>%
  ps_filter(PID == "W") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### Infant X
PhySeqObjFiltered %>%
  ps_filter(PID == "X") %>%
  comp_barplot(
    tax_level = "Family",
    n_taxa = 19,
    palette = distinct_palette(n = 19, add = "grey90"),
    merge_other = FALSE,
    bar_outline_colour = "darkgrey",
    label = "PID_Day") +
  coord_flip() +
  facet_wrap(vars(Time), nrow = 1, scales = "free") +
  theme(axis.text.y = element_text(size = 5))
##### heatmap ASV lvl
cols <- c("mediumblue", "red2")
names(cols) <- c("Before", "After")
htmp_asv <-
  PhySeqObjFiltered %>%
  ps_mutate(Time = as.character(Time)) %>%
  tax_transform("clr") %>%
  comp_heatmap(
    taxa = tax_top(PhySeqObjFiltered, n = 121),
    grid_col = NA,
    name = "CLR",
    sample_names_show = TRUE,
    colors = heat_palette(palette = viridis::turbo(n = 121), sym = TRUE),
    show_row_names = FALSE,
    row_dend_side = "right",
    row_labels = "PID_Day",
    sample_side = "bottom",
    sample_anno = sampleAnnotation(
      Time = anno_sample("Time"),
      col = list(Time = cols),
      border = FALSE
    )
  )
ComplexHeatmap::draw(
  object = htmp_asv, annotation_legend_list = attr(htmp_asv,
"AnnoLegends"),
  merge_legends = TRUE
)


# Statistical analysis plan
## Differential abundance ASVs
Results <- ancombc(phyloseq = PhySeqObjFiltered,
                   formula = "Age_days + Time",
                   p_adj_method = "BH",
                   zero_cut = 0.95,
                   lib_cut = 0,
                   group = "Infant.nr",
                   struc_zero = FALSE,
                   neg_lb = FALSE,
                   tol = 1e-5,
                   max_iter = 100,
                   conserve = TRUE,
                   alpha = 0.05,
                   global = FALSE)

Results_extracted <- Results$res
## Coefficients
Coefficients <- Results_extracted$beta
Col_name <- c("Age (in days)", "Time")
colnames(Coefficients) <- Col_name
### Standard errors
SE <- Results_extracted$se
colnames(SE) <- Col_name
### Test Statistic
TS <- Results_extracted$W
colnames(TS) <- Col_name
### p-values
PValue <- Results_extracted$p_val
colnames(PValue) <- Col_name
PValue %>%
  ggplot(aes(x = Time)) +
  geom_histogram(fill = "red2", alpha = 0.75, binwidth = 0.05) +
  xlab("Unadjusted p-values for the introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
PValue %>%
  ggplot(aes(x = `Age (in days)`)) +
  geom_histogram(fill = "red2", alpha = 0.75, binwidth = 0.05) +
  xlab("Unadjusted p-values for age (in days)") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
### Adjusted p-values
PValueAdj <- Results_extracted$q_val
colnames(PValueAdj) <- Col_name
PValueAdj %>%
  ggplot(aes(x = Time)) +
  geom_histogram(fill = "mediumblue", alpha = 0.75, binwidth = 0.05) +
  xlab("Adjusted p-values for time of introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
PValueAdj %>%
  ggplot(aes(x = `Age (in days)`)) +
  geom_histogram(fill = "mediumblue", alpha = 0.75, binwidth = 0.05) +
  xlab("Adjusted p-values for age (in days)") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
####  Significant age taxa
sum(PValueAdj$`Age (in days)` < 0.05)
### Differentially abundant taxa
DiffAbbTax <- Results_extracted$diff_abn
colnames(DiffAbbTax) <- Col_name
## Volcano plot
# Add the SE to the adjust p-values
PValueAdj$SE_Age <- SE$`Age (in days)`
PValueAdj$SE_Time <- SE$Time
# Add test statistic to the adjusted p-values
PValueAdj$Stat_Age <- TS$`Age (in days)`
PValueAdj$Stat_Time <- TS$Time
# Plot
PValueAdj %>%
  mutate(Significance = if_else(`Age (in days)` < 0.05, "Significant",
"Insignificant")) %>%
  ggplot(aes(x = Stat_Age, y = -log(`Age (in days)`))) +
  geom_point(aes(size = SE_Age, colour = Significance), alpha = 0.5) +
  theme_minimal() +
  xlab("Test statistic for age (in days)") +
  scale_colour_manual(values = c("red2", "mediumblue")) +
  ylab("-log of the p-values of Age (in days)")
PValueAdj %>%
  mutate(Significance = if_else(Time < 0.05, "Significant",
"Insignificant")) %>%
  ggplot(aes(x = Stat_Time, y = -log(Time))) +
  geom_point(aes(size = SE_Time, colour = Significance), alpha = 0.5) +
  theme_minimal() +
  xlab("Test statistic for the introduction of solid food") +
  scale_colour_manual(values = c("red2", "mediumblue")) +
  ylab("-log of the p-values of the introduction of solid food")

#### Looking into the differentially abundant taxa for age
DiffAbbTax %>% select(`Age (in days)`) %>% filter(`Age (in days)` == TRUE)
-> DiffAbbTaxAge
DiffAbbTaxAge <- rownames(DiffAbbTaxAge)
TaxaDataFiltered <- as.data.frame(tax_table(PhySeqObjFiltered))
TaxaDataFiltered$ASV <- row.names(TaxaDataFiltered)
TaxaDataFiltered %>% filter(ASV %in% DiffAbbTaxAge) -> TaxaAge
TaxaAge %>%
  ggplot(aes(x = Family)) +
  geom_bar(fill = "mediumblue", alpha = 0.75) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Family") +
  ylab("Significant ASVs")

## Extract all p-values to create the final results table
# Raw p-values
Final_table <- Results_extracted$p_val
colnames(Final_table) <- c("Age (in days) p-value", "Time p-value")
```

```
# Adjusted p-values
Final_table2 <- Results_extracted$q_val
colnames(Final_table2) <- c("Age (in days) p-value adj.", "Time p-value
adj.")
# Create 1 large table
Final_table <- merge(x = Final_table, y = Final_table2, by = "row.names")
# Set row names correct again
row.names(Final_table) <- Final_table$Row.names
Final_table %>% dplyr::select(-c(Row.names)) -> Final_table
# Set row names to taxonomic names
TaxaData2 <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Microbiome data/taxa_LUC1-897_v34_gg2013.csv", sep =
";", row.names = 1)
TaxaData2 %>% dplyr::select(Family_genus_species_ASV, X.2) -> TaxaData2
TaxaData2$Concatenated <- paste(TaxaData2$Family_genus_species_ASV,
TaxaData2$X.2, sep = "_")
TaxaData2 %>% dplyr::select(Concatenated) -> TaxaData2
Final_table <- merge(x = Final_table, y = TaxaData2, by = "row.names")
row.names(Final_table) <- Final_table$Concatenated
Final_table %>% dplyr::select(-c(Row.names, Concatenated)) -> Final_table
openxlsx::write.xlsx(Final_table, file =
"C:/Users/Gebruiker/Documents/School/Master of Statistics - Bioinformatics
(2021-2022)/Master Thesis Bioinformatics/Data/Microbiome
data/Final_table.xlsx", colNames = TRUE, rowNames = TRUE)
```

## Metabolome data analysis

### *NMR data*

```
# Libraries and data
## Settings
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
## Libraries
library(tidyverse)
library(factoextra)
library(ComplexHeatmap)
library(viridis)
library(oligo)
library(nlme)
library(IMIFA)
## Data
Metadata <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metadata/Metadata.csv", sep = ";", row.names = 1)
NMR <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metabolomics data/NMR.csv", sep = ";", check.names =
FALSE, row.names = 1)
## Colourpalette
ColourPalette <- c("#FB0000", "#1CFC00", "#1C0DFA", "#F9C8C8", "#FF0DDF",
"#00D1FB", "#F0E816", "#0D7222", "#FC8D00", "#5D3581", "#BB005A",
"#35FDDA", "#D06CFE", "#7F3B1C", "#FF95E5", "#668C92", "#AAF580",
"#819BFE", "#948626", "#D8C7FF", "#CBE7BD", "#FE8D88", "#E74738",
"#922663", "#FFC24F", "#FC0DAD", "#A2009F", "#584242", "#2271A6",
"#0DFA98", "#98EEF5", "#5C2EB9", "#FB007C", "#B17DA2", "#9F2A00",
"#FE7AB2")
## Select meaningfull columns
Metadata %>% dplyr::select(Other_Sample_ID_Stearns, PID, PID_Day, Time) ->
Metadata_New
Metadata_New %>% filter(Other_Sample_ID_Stearns %in% colnames(NMR)) ->
Metadata_New
```

```r
# Data prep
# #Prepare data
NMR <- as.data.frame(t(NMR))
NMR$Other_Sample_ID_Stearns <-row.names(NMR)
NMR <- left_join(x = NMR, y = Metadata_New, by = "Other_Sample_ID_Stearns")
row.names(NMR) <- NMR$PID_Day
NMR %>%
  dplyr::select(-c(Other_Sample_ID_Stearns, PID, PID_Day, Time)) -> NMR
NMR <- as.data.frame(t(NMR))

# Data normalization
## Before normalization
NMR %>%
  mutate(Metabolite = row.names(NMR)) %>%
  gather(key = "PID_Day", value = "Abundances", -Metabolite) -> NMR_Long
NMR_Long <- left_join(x = NMR_Long, y = Metadata_New, by = "PID_Day")
NMR_Long %>%
  ggplot(aes(x = PID_Day, y = Abundances, fill = Time, )) +
  geom_boxplot(alpha = 0.75) +
  theme_minimal() +
  scale_fill_manual(values = c("red2", "mediumblue")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Sample")
## Normalization
NMR_Norm_Pareto <- as.data.frame(pareto_scale(t(log(NMR + 1)), centering =
TRUE))
NMR_Norm_Pareto_Wide <- as.data.frame(t(NMR_Norm_Pareto))
### After normalization
NMR_Norm_Pareto_Wide %>%
  mutate(Metabolite = row.names(NMR_Norm_Pareto_Wide)) %>%
  gather(key = "PID_Day", value = "Abundances", -Metabolite) -> NMR_Long2
NMR_Long2 <- left_join(x = NMR_Long2, y = Metadata_New, by = "PID_Day")
NMR_Long2 %>%
  ggplot(aes(x = PID_Day, y = Abundances, fill = Time, )) +
  geom_boxplot(alpha = 0.75) +
  theme_minimal() +
  scale_fill_manual(values = c("red2", "mediumblue")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Sample")

# PCA
## Scree plot
NMR_Norm_Pareto_PCA <- prcomp(NMR_Norm_Pareto, scale. = TRUE)
fviz_eig(NMR_Norm_Pareto_PCA, addlabels = TRUE)
## PCA
PCA_Results <- as.data.frame(NMR_Norm_Pareto_PCA$x)
PCA_Results$PID_Day <- row.names(PCA_Results)
PCA_Results <- left_join(x = PCA_Results, y = Metadata_New, by = "PID_Day")
PCA_Results %>%
  ggplot(aes(x = PC1, y = PC2, colour = PID, shape = Time)) +
  geom_point() +
  theme_minimal() +
  stat_ellipse(aes(group = Time, linetype = Time)) +
  xlab("Principal component 1 (21.0%)") +
  ylab("Principal component 2 (13.7%)") +
  scale_colour_manual(values = ColourPalette)

# Heatmap
## Prepare data
NMR_Norm_Pareto$PID_Day <-row.names(NMR_Norm_Pareto)
```

```
NMR_Norm_Pareto <- left_join(x = NMR_Norm_Pareto, y = Metadata_New, by =
"PID_Day")
## Heatmap
col = list(Time = c("Before" = "mediumblue", "After" = "red2"))
ha <- HeatmapAnnotation(Time = NMR_Norm_Pareto$Time, col = col)
colnames(NMR_Norm_Pareto_Wide) <- as.vector(NMR_Norm_Pareto$PID_Day)
Heatmap(as.matrix(NMR_Norm_Pareto_Wide),
        name = "Normalized counts",
        col = turbo(35),
        top_annotation = ha,
        show_row_names = TRUE,
        row_dend_side = "right",
        row_names_side = "left")

# Differential abundance testing
## Make back-up
NMR_Norm_Pareto_Wilc <- NMR_Norm_Pareto
## Transformation of the data set
NMR_Norm_Pareto_Wilc <- gather(NMR_Norm_Pareto_Wilc, key = "Metabolite",
value = "Abundance", -c("Other_Sample_ID_Stearns", "PID", "PID_Day",
"Time"))
## Testing
Metabolite_names <- as.vector(unique(NMR_Norm_Pareto_Wilc$Metabolite))
Results <- data.frame()
Results2 <- data.frame()
Results3 <- data.frame()
Results4 <- data.frame()
for(metabolite in Metabolite_names){
  NMR_Norm_Pareto_Wilc %>% filter(Metabolite == metabolite) ->
Temp_NMR_Data
  wilcox <- wilcox.test(Abundance ~ Time, exact = FALSE, data =
Temp_NMR_Data)
  Results[metabolite, 1] <- metabolite
  Results[metabolite, 2] <- wilcox$statistic[[1]]
  Results[metabolite, 3] <- wilcox$p.value[[1]]
  t.test <- t.test(Abundance ~ Time, data = Temp_NMR_Data)
  Results2[metabolite, 1] <- metabolite
  Results2[metabolite, 2] <- t.test$statistic[[1]]
  Results2[metabolite, 3] <- t.test$p.value[[1]]
  anova_temp <- aov(Abundance ~ Time, data = Temp_NMR_Data)
  Results3[metabolite, 1] <- metabolite
  Results3[metabolite, 2] <- summary(anova_temp)[[1]][["F value"]][1]
  Results3[metabolite, 3] <- summary(anova_temp)[[1]][["Pr(>F)"]][1]
  lmm <- lme(Abundance ~ Time ,random=~1|PID, data = Temp_NMR_Data)
  Results4[metabolite, 1] <- metabolite
  Results4[metabolite, 2] <- anova(lmm)[2,3]
  Results4[metabolite, 3] <- anova(lmm)[2,4]
  print(metabolite)
}
colnames(Results) <- c("Metabolite", "Statistic", "P-Value")
Results$p.adjusted <- p.adjust(Results$`P-Value`, method = "fdr")
colnames(Results2) <- c("Metabolite", "Statistic", "P-Value")
Results2$p.adjusted <- p.adjust(Results2$`P-Value`, method = "fdr")
colnames(Results3) <- c("Metabolite", "Statistic", "P-Value")
Results3$p.adjusted <- p.adjust(Results3$`P-Value`, method = "fdr")
colnames(Results4) <- c("Metabolite", "Statistic", "P-Value")
Results4$p.adjusted <- p.adjust(Results4$`P-Value`, method = "fdr")

## Visualizations
Results %>%
  ggplot(aes(x = `P-Value`)) +
  geom_histogram(fill = "red2", alpha = 0.75, binwidth = 0.05) +
  xlab("P-values for the change in abundance before and after the
introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
Results %>%
  ggplot(aes(x = p.adjusted)) +
  geom_histogram(fill = "mediumblue", alpha = 0.75, binwidth = 0.05) +
  xlab("Adjusted p-values for the change in abundance before and after the
introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
Results %>%
  mutate(Significance = if_else(p.adjusted < 0.05, "Significant",
"Insignificant")) %>%
  ggplot(aes(x = Statistic, y = -log(p.adjusted))) +
  geom_point(aes(colour = Significance), alpha = 0.5) +
  theme_minimal() +
  xlab("Test statistic for the introduction of solid foods") +
  scale_colour_manual(values = c("red2", "mediumblue")) +
  ylab("-log of the p-values for introduction of solid foods")

## Final table
openxlsx::write.xlsx(x = Results, file = "Final_tableNMR.xlsx", colNames =
TRUE)
```

### *DIMS data*

```
# Libraries and data
# Settings
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
# Libraries
library(tidyverse)
library(factoextra)
library(ComplexHeatmap)
library(viridis)
library(oligo)
library(nlme)
library(IMIFA)
# Data
Metadata <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metadata/Metadata.csv", sep = ";", row.names = 1)
DIMS <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metabolomics data/DIMS.csv", sep = ";", check.names =
FALSE, row.names = 1)
# Colourpalette
ColourPalette <- c("#FB0000", "#1CFC00", "#1C0DFA", "#F9C8C8", "#FF0DDF",
"#00D1FB", "#F0E816", "#0D7222", "#FC8D00", "#5D3581", "#BB005A",
"#35FDDA", "#D06CFE", "#7F3B1C", "#FF95E5", "#668C92", "#AAF580",
"#819BFE", "#948626", "#D8C7FF", "#CBE7BD", "#FE8D88", "#E74738",
"#922663", "#FFC24F", "#FC0DAD", "#A2009F", "#584242", "#2271A6",
"#0DFA98", "#98EEF5", "#5C2EB9", "#FB007C", "#B17DA2", "#9F2A00",
"#FE7AB2")
# Data prep
# Prepare data
DIMS <- as.data.frame(t(DIMS))
DIMS$Other_Sample_ID_Stearns <-row.names(DIMS)
DIMS <- left_join(x = DIMS, y = Metadata_New, by =
"Other_Sample_ID_Stearns")
row.names(DIMS) <- DIMS$PID_Day
```

```
DIMS %>%
  dplyr::select(-c(Other_Sample_ID_Stearns, PID, PID_Day, Time)) -> DIMS
DIMS <- as.data.frame(t(DIMS))

# Data normalization
## Boxplot before normalization
DIMS %>%
  mutate(Metabolite = row.names(DIMS)) %>%
  gather(key = "PID_Day", value = "Abundances", -Metabolite) -> DIMS_Long
DIMS_Long <- left_join(x = DIMS_Long, y = Metadata_New, by = "PID_Day")
DIMS_Long %>%
  ggplot(aes(x = PID_Day, y = Abundances, fill = Time, )) +
  geom_boxplot(alpha = 0.75) +
  theme_minimal() +
  scale_fill_manual(values = c("red2", "mediumblue")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Sample")

## Normalization
DIMS_Norm_Pareto <- as.data.frame(pareto_scale(t(log(DIMS + 1)), centering
= TRUE))
DIMS_Norm_Pareto_Wide <- as.data.frame(t(DIMS_Norm_Pareto))


## Boxplots after normalization
DIMS_Norm_Pareto_Wide %>%
  mutate(Metabolite = row.names(DIMS_Norm_Pareto_Wide)) %>%
  gather(key = "PID_Day", value = "Abundances", -Metabolite) -> DIMS_Long2
DIMS_Long2 <- left_join(x = DIMS_Long2, y = Metadata_New, by = "PID_Day")
DIMS_Long2 %>%
  ggplot(aes(x = PID_Day, y = Abundances, fill = Time, )) +
  geom_boxplot(alpha = 0.75) +
  theme_minimal() +
  scale_fill_manual(values = c("red2", "mediumblue")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Sample")

# PCA
## Scree plot
DIMS_Norm_Pareto_PCA <- prcomp(DIMS_Norm_Pareto, scale. = TRUE)
fviz_eig(DIMS_Norm_Pareto_PCA, addlabels = TRUE)

## PCA
PCA_Results <- as.data.frame(DIMS_Norm_Pareto_PCA$x)
PCA_Results$PID_Day <- row.names(PCA_Results)
PCA_Results <- left_join(x = PCA_Results, y = Metadata_New, by = "PID_Day")
PCA_Results %>%
  ggplot(aes(x = PC1, y = PC2, colour = PID, shape = Time)) +
  geom_point() +
  theme_minimal() +
  stat_ellipse(aes(group = Time, linetype = Time)) +
  xlab("Principal component 1 (32.8 %)") +
  ylab("Principal component 2 (14.4 %)") +
  scale_colour_manual(values = ColourPalette)

# Heatmap
# Prepare data
DIMS_Norm_Pareto$PID_Day <-row.names(DIMS_Norm_Pareto)
DIMS_Norm_Pareto <- left_join(x = DIMS_Norm_Pareto, y = Metadata_New, by =
"PID_Day")
# Heatmap
```

```
col = list(Time = c("Before" = "mediumblue", "After" = "red2"))
ha <- HeatmapAnnotation(Time = DIMS_Norm_Pareto$Time, col = col)
colnames(DIMS_Norm_Pareto_Wide) <- as.vector(DIMS_Norm_Pareto$PID_Day)
Heatmap(as.matrix(DIMS_Norm_Pareto_Wide),
        name = "Normalized counts",
        col = turbo(35),
        top_annotation = ha,
        show_row_names = FALSE,
        row_dend_side = "right",
        row_names_side = "left")

# Statistical testing
# Make back-up
DIMS_Norm_Pareto_Wilc <- DIMS_Norm_Pareto
# Transformation of the data set
DIMS_Norm_Pareto_Wilc <- gather(DIMS_Norm_Pareto_Wilc, key = "Metabolite",
value = "Abundance", -c("Other_Sample_ID_Stearns", "PID", "PID_Day",
"Time"))
# Statistical testing
Metabolite_names <- as.vector(unique(DIMS_Norm_Pareto_Wilc$Metabolite))
Results <- data.frame()
Results2 <- data.frame()
Results3 <- data.frame()
Results4 <- data.frame()
for(metabolite in Metabolite_names){
  DIMS_Norm_Pareto_Wilc %>% filter(Metabolite == metabolite) ->
Temp_DIMS_Data
  wilcox <- wilcox.test(Abundance ~ Time, exact = FALSE, data =
Temp_DIMS_Data)
  Results[metabolite, 1] <- metabolite
  Results[metabolite, 2] <- wilcox$statistic[[1]]
  Results[metabolite, 3] <- wilcox$p.value[[1]]
  t.test <- t.test(Abundance ~ Time, data = Temp_DIMS_Data)
  Results2[metabolite, 1] <- metabolite
  Results2[metabolite, 2] <- t.test$statistic[[1]]
  Results2[metabolite, 3] <- t.test$p.value[[1]]
  anova_temp <- aov(Abundance ~ Time, data = Temp_DIMS_Data)
  Results3[metabolite, 1] <- metabolite
  Results3[metabolite, 2] <- summary(anova_temp)[[1]][["F value"]][1]
  Results3[metabolite, 3] <- summary(anova_temp)[[1]][["Pr(>F)"]][1]
  lmm <- lme(Abundance ~ Time ,random=~1|PID, data = Temp_DIMS_Data)
  Results4[metabolite, 1] <- metabolite
  Results4[metabolite, 2] <- anova(lmm)[2,3]
  Results4[metabolite, 3] <- anova(lmm)[2,4]
  print(metabolite)
}
colnames(Results) <- c("Metabolite", "Statistic", "P-Value")
Results$p.adjusted <- p.adjust(Results$`P-Value`, method = "fdr")
colnames(Results2) <- c("Metabolite", "Statistic", "P-Value")
Results2$p.adjusted <- p.adjust(Results2$`P-Value`, method = "fdr")
colnames(Results3) <- c("Metabolite", "Statistic", "P-Value")
Results3$p.adjusted <- p.adjust(Results3$`P-Value`, method = "fdr")
colnames(Results4) <- c("Metabolite", "Statistic", "P-Value")
Results4$p.adjusted <- p.adjust(Results4$`P-Value`, method = "fdr")

## Visualizations
Results %>%
  ggplot(aes(x = `P-Value`)) +
  geom_histogram(fill = "red2", alpha = 0.75, binwidth = 0.05) +
  xlab("P-values for the change in abundance before and after the
introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
Results %>%
  ggplot(aes(x = p.adjusted)) +
  geom_histogram(fill = "mediumblue", alpha = 0.75, binwidth = 0.05) +
  xlab("Adjusted p-values for the change in abundance before and after the
introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
Results %>%
  mutate(Significance = if_else(p.adjusted < 0.05, "Significant",
"Insignificant")) %>%
  ggplot(aes(x = Statistic, y = -log(p.adjusted))) +
  geom_point(aes(colour = Significance), alpha = 0.5) +
  theme_minimal() +
  xlab("Test statistic for the introduction of solid foods") +
  scale_colour_manual(values = c("red2", "mediumblue")) +
  ylab("-log of the p-values for introduction of solid foods")

## Final table
openxlsx::write.xlsx(x = Results, file = "Final_tableDIMS.xlsx", colNames =
TRUE)
```

### *UPLC data*

```
# Libraries and data

# Settings
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
# Libraries
library(tidyverse)
library(factoextra)
library(ComplexHeatmap)
library(viridis)
library(oligo)
library(nlme)
library(IMIFA)
# Data
Metadata <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metadata/Metadata.csv", sep = ";", row.names = 1)
UPLC <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metabolomics data/UPLC.csv", sep = ";", check.names =
FALSE, row.names = 1)
# Colourpalette
ColourPalette <- c("#FB0000", "#1CFC00", "#1C0DFA", "#F9C8C8", "#FF0DDF",
"#00D1FB", "#F0E816", "#0D7222", "#FC8D00", "#5D3581", "#BB005A",
"#35FDDA", "#D06CFE", "#7F3B1C", "#FF95E5", "#668C92", "#AAF580",
"#819BFE", "#948626", "#D8C7FF", "#CBE7BD", "#FE8D88", "#E74738",
"#922663", "#FFC24F", "#FC0DAD", "#A2009F", "#584242", "#2271A6",
"#0DFA98", "#98EEF5", "#5C2EB9", "#FB007C", "#B17DA2", "#9F2A00",
"#FE7AB2")
# Data transformation
# Select meaningfull columns
Metadata %>% select(Other_Sample_ID_Stearns, PID, PID_Day, Time) ->
Metadata_New
Metadata_New %>% filter(Other_Sample_ID_Stearns %in% colnames(UPLC)) ->
Metadata_New
# Transform UPLC data
UPLC_colnames <- colnames(UPLC)
```

```
UPLC[UPLC_colnames] <- lapply(UPLC[UPLC_colnames], gsub, pattern = ",",
replacement = ".")
UPLC %>% mutate(across(1:32, as.numeric)) -> UPLC
# Remove zero columns
UPLC <- UPLC[ rowSums(UPLC)!=0, ]
# Transformations
UPLC_Long <- UPLC
UPLC_Long$Metabolite <- row.names(UPLC_Long)
UPLC_Long %>% gather(key = "Infant", value = "Value", -Metabolite) ->
UPLC_Long
is.numeric(UPLC_Long$Value)
# Data prep
# Prepare data
UPLC <- as.data.frame(t(UPLC))
UPLC$Other_Sample_ID_Stearns <-row.names(UPLC)
UPLC <- left_join(x = UPLC, y = Metadata_New, by =
"Other_Sample_ID_Stearns")
row.names(UPLC) <- UPLC$PID_Day
UPLC %>%
  dplyr::select(-c(Other_Sample_ID_Stearns, PID, PID_Day, Time)) -> UPLC
UPLC <- as.data.frame(t(UPLC))


# Data transformation
## Boxplots before normalization
UPLC %>%
  mutate(Metabolite = row.names(UPLC)) %>%
  gather(key = "PID_Day", value = "Abundances", -Metabolite) -> UPLC_Long
UPLC_Long <- left_join(x = UPLC_Long, y = Metadata_New, by = "PID_Day")
UPLC_Long %>%
  ggplot(aes(x = PID_Day, y = Abundances, fill = Time, )) +
  geom_boxplot(alpha = 0.75) +
  theme_minimal() +
  scale_fill_manual(values = c("red2", "mediumblue")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Sample")
# Normalization
UPLC_Norm_Pareto <- as.data.frame(pareto_scale(t(log(UPLC + 1)), centering
= TRUE))
UPLC_Norm_Pareto_Wide <- as.data.frame(t(UPLC_Norm_Pareto))
## Boxplots after normalization
UPLC_Norm_Pareto_Wide %>%
  mutate(Metabolite = row.names(UPLC_Norm_Pareto_Wide)) %>%
  gather(key = "PID_Day", value = "Abundances", -Metabolite) -> UPLC_Long2
UPLC_Long2 <- left_join(x = UPLC_Long2, y = Metadata_New, by = "PID_Day")
UPLC_Long2 %>%
  ggplot(aes(x = PID_Day, y = Abundances, fill = Time, )) +
  geom_boxplot(alpha = 0.75) +
  theme_minimal() +
  scale_fill_manual(values = c("red2", "mediumblue")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Sample")

# PCA
## Scree plot
# Biplot prep
UPLC_Norm_Pareto_PCA_Prep <- UPLC_Norm_Pareto
UPLC_Norm_Pareto_PCA_Prep$PID_Day <-row.names(UPLC_Norm_Pareto_PCA_Prep)
UPLC_Norm_Pareto_PCA_Prep <- left_join(x = UPLC_Norm_Pareto_PCA_Prep, y =
Metadata_New, by = "PID_Day")
```

```
row.names(UPLC_Norm_Pareto_PCA_Prep) <-
as.vector(UPLC_Norm_Pareto_PCA_Prep$PID_Day)
# PCA
UPLC_Norm_Pareto_PCA <- prcomp(UPLC_Norm_Pareto_PCA_Prep[, 1:70], scale. =
TRUE)
fviz_eig(UPLC_Norm_Pareto_PCA, addlabels = TRUE)
## PCA
PCA_Results <- as.data.frame(UPLC_Norm_Pareto_PCA$x)
PCA_Results$PID_Day <- row.names(PCA_Results)
PCA_Results <- left_join(x = PCA_Results, y = Metadata_New, by = "PID_Day")
PCA_Results %>%
  ggplot(aes(x = PC1, y = PC2, colour = PID, shape = Time)) +
  geom_point() +
  theme_minimal() +
  stat_ellipse(aes(group = Time, linetype = Time)) +
  xlab("Principal component 1 (21.6%)") +
  ylab("Principal component 2 (15.7%)") +
  scale_colour_manual(values = ColourPalette)
## Biplot
fviz_pca_biplot(UPLC_Norm_Pareto_PCA,
                repel = TRUE,
                col.ind = UPLC_Norm_Pareto_PCA_Prep$Time,
                label = "var",
                legend.title = "Introduction of solids",
                palette = c("mediumblue", "red2"),
                select.var = list(contrib = 5)
                )

# Heatmap
# Prepare data
UPLC_Norm_Pareto$PID_Day <-row.names(UPLC_Norm_Pareto)
UPLC_Norm_Pareto <- left_join(x = UPLC_Norm_Pareto, y = Metadata_New, by =
"PID_Day")
# Heatmap
col = list(Time = c("Before" = "mediumblue", "After" = "red2"))
ha <- HeatmapAnnotation(Time = UPLC_Norm_Pareto$Time, col = col)
colnames(UPLC_Norm_Pareto_Wide) <- as.vector(UPLC_Norm_Pareto$PID_Day)
Heatmap(as.matrix(UPLC_Norm_Pareto_Wide),
        name = "Normalized counts",
        col = turbo(35),
        top_annotation = ha,
        show_row_names = FALSE,
        row_dend_side = "right",
        row_names_side = "left")


# Statistical testing
# Make back-up
UPLC_Norm_Pareto_Wilc <- UPLC_Norm_Pareto
# Transformation of the data set
UPLC_Norm_Pareto_Wilc <- gather(UPLC_Norm_Pareto_Wilc, key = "Metabolite",
value = "Abundance", -c("Other_Sample_ID_Stearns", "PID", "PID_Day",
"Time"))
# Testing
Metabolite_names <- as.vector(unique(UPLC_Norm_Pareto_Wilc$Metabolite))
Results <- data.frame()
Results2 <- data.frame()
Results3 <- data.frame()
Results4 <- data.frame()
for(metabolite in Metabolite_names){
  UPLC_Norm_Pareto_Wilc %>% filter(Metabolite == metabolite) ->
Temp_UPLC_Data
  wilcox <- wilcox.test(Abundance ~ Time, exact = FALSE, data =
Temp_UPLC_Data)
  Results[metabolite, 1] <- metabolite
  Results[metabolite, 2] <- wilcox$statistic[[1]]
  Results[metabolite, 3] <- wilcox$p.value[[1]]
  t.test <- t.test(Abundance ~ Time, data = Temp_UPLC_Data)
  Results2[metabolite, 1] <- metabolite
  Results2[metabolite, 2] <- t.test$statistic[[1]]
  Results2[metabolite, 3] <- t.test$p.value[[1]]
  anova_temp <- aov(Abundance ~ Time, data = Temp_UPLC_Data)
  Results3[metabolite, 1] <- metabolite
  Results3[metabolite, 2] <- summary(anova_temp)[[1]][["F value"]][1]
  Results3[metabolite, 3] <- summary(anova_temp)[[1]][["Pr(>F)"]][1]
  lmm <- lme(Abundance ~ Time ,random=~1|PID, data = Temp_UPLC_Data)
  Results4[metabolite, 1] <- metabolite
  Results4[metabolite, 2] <- anova(lmm)[2,3]
  Results4[metabolite, 3] <- anova(lmm)[2,4]
  print(metabolite)
}
colnames(Results) <- c("Metabolite", "Statistic", "P-Value")
Results$p.adjusted <- p.adjust(Results$`P-Value`, method = "fdr")
colnames(Results2) <- c("Metabolite", "Statistic", "P-Value")
Results2$p.adjusted <- p.adjust(Results2$`P-Value`, method = "fdr")
colnames(Results3) <- c("Metabolite", "Statistic", "P-Value")
Results3$p.adjusted <- p.adjust(Results3$`P-Value`, method = "fdr")
colnames(Results4) <- c("Metabolite", "Statistic", "P-Value")
Results4$p.adjusted <- p.adjust(Results4$`P-Value`, method = "fdr")

## Visualizations
Results %>%
  ggplot(aes(x = `P-Value`)) +
  geom_histogram(fill = "red2", alpha = 0.75, binwidth = 0.05) +
  xlab("P-values for the change in abundance before and after the
introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
Results %>%
  ggplot(aes(x = p.adjusted)) +
  geom_histogram(fill = "mediumblue", alpha = 0.75, binwidth = 0.05) +
  xlab("Adjusted p-values for the change in abundance before and after the
introduction of solid foods") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1, 0.10), expand = c(0,0))
Results %>%
  mutate(Significance = if_else(p.adjusted < 0.05, "Significant",
"Insignificant")) %>%
  ggplot(aes(x = Statistic, y = -log(p.adjusted))) +
  geom_point(aes(colour = Significance), alpha = 0.5) +
  theme_minimal() +
  xlab("Test statistic for the introduction of solid foods") +
  scale_colour_manual(values = c("red2", "mediumblue")) +
  ylab("-log of the p-values for introduction of solid foods")

## Final table
openxlsx::write.xlsx(x = Results, file = "Final_tableUPLC.xlsx", colNames =
TRUE)
```

## Association study

```
# Libraries
library(tidyverse)
library(phyloseq)
library(microbiome)
library(IMIFA)
library(metagMisc)
library(microViz)
library(mixOmics)
library(BiocParallel)
library(matrixStats)
library(viridis)
set.seed(1996)


# Data and preparation
## Microbiome
### Reading the data
Metadata <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metadata/Metadata.csv", sep = ";", row.names = 1)
MicrobiomeData <- read.csv(file =
"C:/Users/Gebruiker/Documents/School/Master of Statistics - Bioinformatics
(2021-2022)/Master Thesis Bioinformatics/Data/Microbiome
data/Code_LucKi_mergtab_nochim_LucKi_v34_transposed_ISS.csv", sep = ";",
row.names = 1)
TaxaData <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Microbiome data/taxa.csv", sep = ";", row.names = 1)
### PhySeq
#### Transforming into matrices for Phyloseq
MicrobiomeInf <- as.matrix(MicrobiomeData)
TaxaInf <- as.matrix(TaxaData)
MetaInf <- as.matrix(Metadata)
#### Creating Phyloseq objects
ASV <- otu_table(MicrobiomeInf, taxa_are_rows = TRUE)
TAX <- tax_table(TaxaInf)
SampleData <- sample_data(Metadata)
PhySeqObj <- phyloseq(ASV, TAX, SampleData)
### Filtering
#### Filter based on prevalence
PhySeqObjFiltered <- phyloseq_filter_prevalence(PhySeqObj, prev.trh = 0.05,
abund.trh = NULL)
#### Filter based on relative abundance = 0.01%
minTotRelAbun = 1e-4
x = taxa_sums(PhySeqObjFiltered)
keepTaxa = (x / sum(x)) > minTotRelAbun
PhySeqObjFiltered = prune_taxa(keepTaxa, PhySeqObjFiltered)
### Fixing the taxa
PhySeqObjFiltered %>% tax_fix(min_length = 4) -> PhySeqObjFiltered
### Normalization
PhySeqObjFiltered_CLR <- transform(PhySeqObjFiltered, "clr")
## NMR
### Reading data
NMR <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metabolomics data/NMR.csv", sep = ";", check.names =
FALSE, row.names = 1)
### Normalization
NMR_Norm_Pareto <- as.data.frame(pareto_scale(t(log(NMR + 1)), centering =
TRUE))
```

```
## DIMS
### Reading data
DIMS <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metabolomics data/DIMS.csv", sep = ";", check.names =
FALSE, row.names = 1)
### Normalization
DIMS_Norm_Pareto <- as.data.frame(pareto_scale(t(log(DIMS + 1)), centering
= TRUE))
## UP-LC
### Reading data
UPLC <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Metabolomics data/UPLC.csv", sep = ";", check.names =
FALSE, row.names = 1)
### Transform UPLC data
UPLC_colnames <- colnames(UPLC)
UPLC[UPLC_colnames] <- lapply(UPLC[UPLC_colnames], gsub, pattern = ",",
replacement = ".")
UPLC %>% mutate(across(1:32, as.numeric)) -> UPLC
### Remove zero columns
UPLC <- UPLC[ rowSums(UPLC)!=0, ]
### Normalization
UPLC_Norm_Pareto <- as.data.frame(pareto_scale(t(log(UPLC + 1)), centering
= TRUE))
## Setting all variables correct
### Creating analysis data sets
Microbiome_analysis <- as.data.frame(t(abundances(PhySeqObjFiltered_CLR)))
NMR_analysis <- NMR_Norm_Pareto
DIMS_analysis <- DIMS_Norm_Pareto
UPLC_analysis <- UPLC_Norm_Pareto
### Setting name variables right for microbiome data
#### Setting sample names rights
Metadata %>%
  dplyr::select(PID_Day) -> Metadata_microbiome
Microbiome_analysis <- merge(Microbiome_analysis, Metadata_microbiome, by =
"row.names")
Microbiome_analysis <-
Microbiome_analysis[order(Microbiome_analysis$PID_Day),]
row.names(Microbiome_analysis) <- Microbiome_analysis$PID_Day
Microbiome_analysis %>%
  dplyr::select(-c(Row.names, PID_Day)) -> Microbiome_analysis
#### Setting ASVs with shorter names
TaxaData2 <- read.csv(file = "C:/Users/Gebruiker/Documents/School/Master of
Statistics - Bioinformatics (2021-2022)/Master Thesis
Bioinformatics/Data/Microbiome data/taxa_LUC1-897_v34_gg2013.csv", sep =
";", row.names = 1)
TaxaData2 %>% dplyr::select(Family_genus_species_ASV, X.2) -> TaxaData2
TaxaData2$Concatenated <- paste(TaxaData2$Family_genus_species_ASV,
TaxaData2$X.2, sep = "_")
TaxaData2 %>% dplyr::select(Concatenated) -> TaxaData2
Microbiome_analysis <- as.data.frame(t(Microbiome_analysis))
Microbiome_analysis <- merge(x = Microbiome_analysis, y = TaxaData2, by =
"row.names")
row.names(Microbiome_analysis) <- Microbiome_analysis$Concatenated
Microbiome_analysis %>% dplyr::select(-c(Row.names, Concatenated)) ->
Microbiome_analysis
Microbiome_analysis <- as.data.frame(t(Microbiome_analysis))
### Setting name variables right for NMR data
NMR_analysis$Other_Sample_ID_Stearns <- row.names(NMR_analysis)
Metadata %>%
  dplyr::select(Other_Sample_ID_Stearns, PID_Day) -> Metadata_NMR
NMR_analysis <- left_join(NMR_analysis, Metadata_NMR, by =
"Other_Sample_ID_Stearns")
NMR_analysis <- NMR_analysis[order(NMR_analysis$PID_Day),]
row.names(NMR_analysis) <- NMR_analysis$PID_Day
NMR_analysis %>%
  dplyr::select(-c(Other_Sample_ID_Stearns, PID_Day)) -> NMR_analysis
### Setting name variables right for DIMS data
DIMS_analysis$Other_Sample_ID_Stearns <- row.names(DIMS_analysis)
Metadata %>%
  dplyr::select(Other_Sample_ID_Stearns, PID_Day) -> Metadata_DIMS
DIMS_analysis <- left_join(DIMS_analysis, Metadata_DIMS, by =
"Other_Sample_ID_Stearns")
DIMS_analysis <- DIMS_analysis[order(DIMS_analysis$PID_Day),]
row.names(DIMS_analysis) <- DIMS_analysis$PID_Day
DIMS_analysis %>%
  dplyr::select(-c(Other_Sample_ID_Stearns, PID_Day)) -> DIMS_analysis
### Setting name variables right for UPLC data
UPLC_analysis$Other_Sample_ID_Stearns <- row.names(UPLC_analysis)
Metadata %>%
  dplyr::select(Other_Sample_ID_Stearns, PID_Day) -> Metadata_UPLC
UPLC_analysis <- left_join(UPLC_analysis, Metadata_UPLC, by =
"Other_Sample_ID_Stearns")
UPLC_analysis <- UPLC_analysis[order(UPLC_analysis$PID_Day),]
row.names(UPLC_analysis) <- UPLC_analysis$PID_Day
UPLC_analysis %>%
  dplyr::select(-c(Other_Sample_ID_Stearns, PID_Day)) -> UPLC_analysis
### Filtering for common samples
Microbiome_samples <- row.names(Microbiome_analysis)
NMR_samples <- row.names(NMR_analysis)
DIMS_samples <- row.names(DIMS_analysis)
UPLC_samples <- row.names(UPLC_analysis)
Uniques <- intersect(Microbiome_samples, NMR_samples)
Uniques <- intersect(Uniques, DIMS_samples)
Uniques <- intersect(Uniques, UPLC_samples)
Microbiome_analysis <- subset(Microbiome_analysis,
row.names(Microbiome_analysis) %in% Uniques)
NMR_analysis <- subset(NMR_analysis, row.names(NMR_analysis) %in% Uniques)
DIMS_analysis <- subset(DIMS_analysis, row.names(DIMS_analysis) %in%
Uniques)
UPLC_analysis <- subset(UPLC_analysis, row.names(UPLC_analysis) %in%
Uniques)
### Creating outcome
Metadata %>%
  filter(PID_Day %in% Uniques) %>%
  dplyr::select(PID_Day, Time) -> Outcome
Outcome <- Outcome[order(Outcome$PID_Day),]
row.names(Outcome) <- Outcome$PID_Day
Outcome %>%
  dplyr::select(Time) -> Outcome
Outcome <- Outcome$Time
### Creating a dataset
data <- list(Microbiome = as.matrix(Microbiome_analysis),
             NMR = as.matrix(NMR_analysis),
             DIMS = as.matrix(DIMS_analysis),
             UPLC = as.matrix(UPLC_analysis))
### Selecting ASVs and metabolites
Microbiome_ASVs <- colnames(Microbiome_analysis)
NMR_metabolites <- colnames(NMR_analysis)
DIMS_metabolites <- colnames(DIMS_analysis)
UPLC_metabolites <- colnames(UPLC_analysis)
```

```
## Cleaning the environment
rm(MicrobiomeData, MicrobiomeInf, ASV, PhySeqObj, PhySeqObjFiltered,
PhySeqObjFiltered_CLR, TaxaData, TaxaInf, TAX, keepTaxa, minTotRelAbun, x,
NMR, NMR_Norm_Pareto, MetaInf, Metadata_microbiome, Metadata_NMR,
SampleData, DIMS, DIMS_Norm_Pareto, Metadata_DIMS, UPLC, UPLC_colnames,
UPLC_Norm_Pareto, Metadata_UPLC, Uniques, DIMS_samples, Microbiome_samples,
NMR_samples, UPLC_samples, Metadata, TaxaData2)
# Correlation analysis
## Hyperparameter tuning
### Covariance matrix
pls1 <- spls(Microbiome_analysis, NMR_analysis, ncomp = 1)
pls2 <- spls(Microbiome_analysis, DIMS_analysis, ncomp = 1)
pls3 <- spls(Microbiome_analysis, UPLC_analysis, ncomp = 1)
pls4 <- spls(NMR_analysis, DIMS_analysis, ncomp = 1)
pls5 <- spls(NMR_analysis, UPLC_analysis, ncomp = 1)
pls6 <- spls(DIMS_analysis, UPLC_analysis, ncomp = 1)
cor(pls1$variates$X, pls1$variates$Y)
cor(pls2$variates$X, pls2$variates$Y)
cor(pls3$variates$X, pls3$variates$Y)
cor(pls4$variates$X, pls4$variates$Y)
cor(pls5$variates$X, pls5$variates$Y)
cor(pls6$variates$X, pls6$variates$Y)
design <- matrix(0, nrow = length(data), ncol = length(data), dimnames =
list(names(data), names(data)))
design[1,2] <- 0.8241704
design[2,1] <- 0.8241704
design[1,3] <- 0.6923896
design[3,1] <- 0.6923896
design[1,4] <- 0.866391
design[4,1] <- 0.866391
design[2,3] <- 0.8399946
design[3,2] <- 0.8399946
design[2,4] <- 0.6672782
design[4,2] <- 0.6672782
design[3,4] <- 0.7526392
design[4,3] <- 0.7526392
### Principal components
# form basic DIABLO model
basic.diablo.model = block.splsda(X = data, Y = Outcome, ncomp = 25, design
= design)
# run component number tuning with repeated CV
perf.diablo2 = perf(basic.diablo.model, validation = "loo", progressBar =
TRUE)
plot(perf.diablo2) # plot output of tuning
### Feature selection
#### Done in steps
# set grid of values for each component to test
Grid = list (Microbiome = seq(5, 121, 30),
             NMR = seq(5, 41, 10),
             DIMS = seq(5, 116, 30),
             UPLC = seq(5, 70, 20))
Clusters <- BiocParallel::SnowParam(progressbar = TRUE)
# run the feature selection tuning
tune.DIABLO1 <- tune.block.splsda(X = data,
                               Y = Outcome,
                               ncomp = 3,
                               test.keepX = Grid,
                               design = design,
                               validation = "loo",
                               progressBar = TRUE)
```

```
list.keepX = tune.DIABLO1$choice.keepX # set the optimal values of features
to retain
list.keepX
# set grid of values for each component to test
Grid2 = list (Microbiome = seq(5, 65, 15),
             NMR = seq(5, 35, 5),
             DIMS = seq(5, 100, 15),
             UPLC = seq(5, 10, 1))

Clusters <- BiocParallel::SnowParam(progressbar = TRUE)
# run the feature selection tuning
tune.DIABLO2 <- tune.block.splsda(X = data,
                               Y = Outcome,
                               ncomp = 3,
                               test.keepX = Grid2,
                               design = design,
                               validation = "loo",
                               progressBar = TRUE)
list.keepX = tune.DIABLO2$choice.keepX # set the optimal values of features
to retain
list.keepX
# set grid of values for each component to test
Grid3 = list (Microbiome = seq(5, 30, 4),
             NMR = seq(5, 21, 4),
             DIMS = seq(5, 70, 10),
             UPLC = seq(5, 10, 1))
Clusters <- BiocParallel::SnowParam(progressbar = TRUE)
# run the feature selection tuning
tune.DIABLO3 <- tune.block.splsda(X = data,
                               Y = Outcome,
                               ncomp = 3,
                               test.keepX = Grid3,
                               design = design,
                               validation = "loo",
                               progressBar = TRUE)
list.keepX = tune.DIABLO3$choice.keepX # set the optimal values of features
to retain
list.keepX
# set grid of values for each component to test
Grid4 = list (Microbiome = seq(5, 15, 2),
             NMR = seq(5, 25, 4),
             DIMS = seq(5, 55, 5),
             UPLC = seq(5, 10, 1))
Clusters <- BiocParallel::SnowParam(progressbar = TRUE)
# run the feature selection tuning
tune.DIABLO4 <- tune.block.splsda(X = data,
                               Y = Outcome,
                               ncomp = 3,
                               test.keepX = Grid4,
                               design = design,
                               validation = "loo",
                               progressBar = TRUE)
list.keepX = tune.DIABLO4$choice.keepX # set the optimal values of features
to retain
list.keepX
# set grid of values for each component to test
Grid5 = list (Microbiome = seq(5, 10, 1),
             NMR = seq(5, 25, 2),
             DIMS = c(seq(5, 23, 2), seq(25,60,5)),
             UPLC = seq(5, 10, 1))
Clusters <- BiocParallel::SnowParam(progressbar = TRUE)
```

```
# run the feature selection tuning
tune.DIABLO5 <- tune.block.splsda(X = data,
                               Y = Outcome,
                               ncomp = 3,
                               test.keepX = Grid5,
                               design = design,
                               validation = "loo",
                               progressBar = TRUE)
list.keepX = tune.DIABLO5$choice.keepX # set the optimal values of features
to retain
list.keepX
list.keepX <- list(Microbiome = c(10, 8, 5),
                   NMR = c(19, 21, 5),
                   DIMS = c(13, 60, 13),
                   UPLC = c(7, 8, 6))


# set the optimised DIABLO model
final.diablo.model = block.splsda(X = data, Y = Outcome, ncomp = 3,
                          keepX = list.keepX, design = design)
## Plotting the principal components for the final diablo model
plotDiablo(final.diablo.model,
           ncomp = 1,
           col = c('mediumblue', 'red2'))
plotDiablo(final.diablo.model,
           ncomp = 2,
           col = c('mediumblue', 'red2'))
plotDiablo(final.diablo.model,
           ncomp = 3,
           col = c('mediumblue', 'red2'))

## Circos plots per component and full circos plot
circosPlot(final.diablo.model, cutoff = 0.75, line = TRUE,
           color.blocks= c('darkorchid', 'brown1', 'lightgreen', "yellow"),
           color.cor = c("chocolate3","grey20"),
           size.labels = 1.5,
           size.variables = 0.5,
           comp = 1)
circosPlot(final.diablo.model, cutoff = 0.75, line = TRUE,
           color.blocks= c('darkorchid', 'brown1', 'lightgreen', "yellow"),
           color.cor = c("chocolate3","grey20"),
           size.labels = 1.5,
           size.variables = 0.5,
           comp = 2)
circosPlot(final.diablo.model, cutoff = 0.75, line = TRUE,
           color.blocks= c('darkorchid', 'brown1', 'lightgreen', "yellow"),
           color.cor = c("chocolate3","grey20"),
           size.labels = 1.5,
           size.variables = 0.5,
           comp = 3)
circosPlot(final.diablo.model, cutoff = 0.7, line = TRUE,
           color.blocks= c('darkorchid', 'brown1', 'lightgreen', "yellow"),
           color.cor = c("mediumblue","red2"),
           size.labels = 1.5,
           size.variables = 0.5,
           comp = 1:3)
### Extracting correlations for the final table
circosPlot(final.diablo.model, cutoff = 0.7, line = TRUE,
           color.blocks= c('darkorchid', 'brown1', 'lightgreen', "yellow"),
           color.cor = c("mediumblue","red2"),
           size.labels = 1.5,
           size.variables = 0.5,
           comp = 1:3) -> Correlations
heatmap(Correlations, col = turbo(152))
Correlations[lower.tri(Correlations, diag = TRUE)] <- NA
Correlations <- as.data.frame(Correlations)
Correlations$"Feature 1" <- row.names(Correlations)
Correlations <- gather(data = Correlations, key = "Feature 2", value =
"Correlation", -"Feature 1")
Correlations %>% na.omit() -> Correlations
Filtered_correlations <- matrix(ncol = 3, nrow = 0)
# Loop to build table
for(i in 1:nrow(Correlations)){
  Feature1 <- Correlations[i, 1]
  Feature2 <- Correlations[i, 2]
  Correlation <- Correlations[i, 3]
  if(Feature1 %in% Microbiome_ASVs & Feature2 %in% Microbiome_ASVs){
    return
  } else if(Feature1 %in% NMR_metabolites & Feature2 %in% NMR_metabolites){
    return
  } else if(Feature1 %in% DIMS_metabolites & Feature2 %in%
DIMS_metabolites){
    return
  } else if(Feature1 %in% UPLC_metabolites & Feature2 %in%
UPLC_metabolites){
    return
  } else if(abs(Correlation) >= 0.7){
    temp <- c(Feature1, Feature2, Correlation)
    Filtered_correlations <- rbind(Filtered_correlations, temp)
  }
}
Filtered_correlations <- as.data.frame(Filtered_correlations)
colnames(Filtered_correlations) <- c("Feature 1", "Feature 2",
"Correlation")
writexl::write_xlsx(Filtered_correlations, path =
"C:/Users/Gebruiker/Documents/School/Master of Statistics - Bioinformatics
(2021-2022)/Master Thesis Bioinformatics/Data/Correlations.xlsx")

## Building the final heatmap
circosPlot(final.diablo.model, cutoff = 0, line = TRUE,
           color.blocks= c('darkorchid', 'brown1', 'lightgreen', "yellow"),
           color.cor = c("mediumblue","red2"),
           size.labels = 1.5,
           size.variables = 0.5,
           comp = 1:3) -> Correlations2
ComplexHeatmap::Heatmap(matrix = Correlations2,
                        col = turbo(100),
                        show_row_names = FALSE,
                        show_column_names = FALSE,
                        heatmap_legend_param = list(title = "Correlation"))
```

# Bibliography

[1] A. Shreiner, J. Kao and V. Young, “The gut microbiome in health and in disease.,” *Current opinion in gastroenterology,* vol. 31, no. 1, pp. 69-75, 2015.

[2] Z. Z. Tang, G. Chen, Q. Hong, S. Huang, H. M. Smith, R. D. Shar, M. Scholz and J. F. Ferguson, “Multi-Omic Analysis of the Microbiome and Metabolome in Healthy Subjects Reveals the Microbiome-Dependent Relationships Between Diet and Metabolites,” *Frontiers in Genetics,* no. 10, 2019.

[3] C. L. Sears, “A dynamic partnership: celebrating our gut flora,” *Anaerobe,* no. 11, pp. 247-251, 2005.

[4] S. Minot, R. Sinha, J. Chen, H. Li, S. A. Keilbaugh, G. D. Wu and et al., “The human gut virome: inter-individual variation and dynamic response to diet,” *Genome Res.,* no. 21, pp. 1616-1625, 2011.

[5] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes and C. Huttenhower, “Microbial Co-occurrence Relationships in the Human Microbiome,” *PLoS Computational Biology,* vol. 8, no. 7, 2012.

[6] T. Ding and P. Schloss, “Dynamics and associations of microbial community types across the human body,” *Nature,* vol. 509, pp. 357-360, 2014.

[7] A. L. Kau, P. P. Ahern, N. W. Griffin, A. L. Goodman and J. I. Gordon, “Human nutrition, the gut microbiome and the immune system.,” *Nature,* no. 474, pp. 327-336, 2011.

[8] C. L. Maynard, C. O. Elson, R. D. Hatton and C. T. Weaver, “Reciprocal interactions of the intestinal microbiota and immune system.,” *Nature,* no. 489, pp. 231-241, 2012.

[9] A. L. Goodman and J. I. Gordon, “Our unindicted coconspirators: human metabolism from a microbial perspective,” *Cell Metab.,* no. 12, pp. 111-116, 2010.

[10] S. Jandhyala, R. Talukdar, C. Subramanyam, H. Vuyyuru, M. Sasikala and D. Nageshwar Reddy, “Role of the normal gut microbiota,” *World journal of gastroenterology,* vol. 21, no. 29, pp. 8787-8803, 2015.

[11] G. Galazzo, N. van Best, L. Bervoets, L. Dapaah, P. Savelkoul, M. Hornef, GI-MDH Consortium, S. Lau, E. Hamelmann and J. Penders, “Development of the Microbiota and Associations with Birth Mode, Diet and Atopic Disorders in a Longitudinal Analysis of Stool Samples, Collected from Infancy through Early Childhood.,” *Gastroenterology,* vol. 158, pp. 1584-1596, 2020.

[12] P. Turnbaugh and J. Gordon, “The core gut microbiome, energy balance and obesity.,” *Journal of Physiology,* vol. 587, pp. 4153-4158, 2009.

[13] K. E. Fujimura, A. R. Sitarik, S. Havstad, D. L. Lin, S. Levan, D. Fadrosh and et al., “Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation,” *Nature Medicine,* pp. 1187-1191, 2016.

[14] A. D. Kostic, D. Gevers, H. Siljander, T. Vatanen, T. Hyotylainen, A. M. Hamalainen and et al., “The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes,” *Cell Host Microbe,* pp. 260-273, 2015.

[15] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley and et al., “A core gut microbiome in obsese and lean twins,” *Nature,* pp. 480-484, 2009.

[16] G. Sharon, T. R. Sampson, D. H. Geschwind and S. K. Mazmanian, “The Central Nervous System and the Gut Microbiome,” *Cell,* pp. 915-932, 2016.

[17] V. C. Harris, G. Armah, S. Fuentes, K. E. Korpela, U. Parashar, J. C. Victor and et al., “Significant Correlation Between the Infant Gut Microbiome and Rotavirus Vaccine Response in Rural Ghana,” *Journal of Infectious Diseases,* pp. 34-41, 2017.

[18] N. R. Klatt, R. Cheu, K. Birse, A. S. Zevin, M. Perner, L. Noel-Romas and et al., “Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women,” *Science,* pp. 938-945, 2017.

[19] N. Hasan and H. Yang, “Factors affecting the compositions of the gut microbiota and its modulation.,” *Peer Journal,* vol. 7, 2019.

[20] J. Zimmer, N. Lange, J. Frick, H. Sauer, K. Zimmermann, A. Schwiertz, K. Rusch, S. Klosterhalfen and P. Enck, “A vegan or vegetarian diet substantially alters the human colonic faecal microbiota.,” *European Journal of Clinical Nutrition,* vol. 66, pp. 53-60, 2012.

[21] F. Backhed, J. Roswall, Y. Peng, Q. Feng, H. Jia, P. Kovatcheva-Datchary, Y. Li, Y. Xia, H. Xie, H. Zhing and et al., “Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life,” *Cell host & microbe,* vol. 17, pp. 690-703, 2015.

[22] E. Callaway, “C-section babies are missing key microbes,” *Nature,* 2019.

[23] J. Ma, Z. Li, W. Zhang and et al., “Comparison of gut microbiota in exclusively breast-fed and formula-fed babies: a study of 91 term infants,” *Scientific Reports,* 2020.

[24] F. Turroni, C. Milani, S. Duranti and et al., “The infant gut microbiome as a microbial organ influencing host well-being,” *Italian Journal of Pediatrics,* 2020.

[25] S. Ainonen, M. V. Tejesvi, M. R. Mahmud and et al., “Antibiotics at birth and later antibiotic courses: effects on gut microbiota,” *Pediatric Research,* pp. 154-162, 2022.

[26] S. Matamoros, C. Gras-Leguen, F. Le Vacon, G. Potel and M. de La Cochetiere, “Development of intestinal microbiota in infants and its impact on health,” *Trends in Microbiology,* vol. 21, pp. 167-173, 2013.

[27] C. M. Homann, C. A. J. Rossel, S. Dizzell, L. Bervoets, J. Simioni, J. Li, E. Gunn, M. G. Surette, R. J. de Souza, M. Mommers, E. K. Hutton, K. M. Morrison, J. Penders, N. van Best and J. C. Stearns, “Infants' First Solid Foods: Impact on Gut Microbiota Development in Two Intercontinental Cohorts,” *Nutrients,* 2021.

[28] M. Fallani, S. Amarri, A. Uusijarvi, R. Adam, S. Khanna, M. Aguilera, A. Gil, J. Vieites, E. Norin, D. Young and et al., “Determinants of the humqn infant intestinal microbiota after the

introduction of first complementary foods in infant samples from five European centres.," *Microbiology,* vol. 157, pp. 1385-1392, 2011.

[29] H. Pearson, "Meet the human metabolome," *Nature,* vol. 446, no. 8, 2007.

[30] M. J. Gibney, M. Walsh, L. Brennan, H. M. Roche, B. German and B. van Ommen, "Metabolomics in human nutrition: Opportunities and Challenges," *American Journal of Clinical Nutrition,* pp. 497-503, 2005.

[31] P. J. Turnbaugh and J. I. Gordon, "An invitation to the marriage of metagenomics and metabolomics," *Cell,* pp. 708-713, 2008.

[32] K. Oliphant and E. Allen-Vercoe, "Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health," *Microbiome,* p. 91, 2019.

[33] A. Heinken and I. Thiele, "Systems biology of host–microbe metabolomics," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine,* pp. 195-219, 2015.

[34] M. Li, B. Wang, M. Zhang, M. Rantalainen, S. Wang, H. Zhou and et al., "Symbiotic gut microbes modulate human metabolic phenotypes," *Proceedings of the National Academy of Sciences,* pp. 2117-2122, 2008.

[35] J. K. Nicholson, E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia and et al., "Host-gut microbiota metabolic interactions," *Science,* pp. 1262-1267, 2012.

[36] J. Zierer, M. A. Jackson, G. Kastenmüller, M. Mangino, T. Long, A. Telenti and et al., "The fecal metabolome as a functional readout of the gut microbiome," *Nature Genetics,* pp. 790-795, 2018.

[37] Q. P. Nguyen, M. R. Karagas, J. C. Madan and et al., "Associations between the gut microbiome and metabolome in early life," *BMC Microbiology,* 2021.

[38] D. de Korte-de Boer, M. Mommers, H. Creemers, E. Dompeling, F. Feron, C. Gielkens-Sijstermans, M. Jaminon, S. Mujakovic, O. van Schayk, C. Thijs and et al., "LucKi Birth Cohort Study: rationale and design," *BMC Public Health,* 2015.

[39] S. Dizzel, J. C. Stearns, J. Li, N. van Best, L. Bervoets, M. Mommers, J. Penders, K. M. Morrison and E. K. Hutton, "Investigating colonization patterns of the infant gut microbiome during the introduction of solid food and weaning from breastmilk: A cohort study protocol," *PLOS ONE,* 2021.

[40] J. M. Janda and S. L. Abbott, "16s rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls," *Journal of Clinical Microbiology,* vol. 45, no. 9, pp. 2761-2764, 2007.

[41] "16S rRNA Gene Sequencing for identification, classification and quantitation of microbes," LC Sciences - Technologies for Genomics & Proteomics Discoveries, 30 September 2021. [Online]. Available: https://lcsciences.com/16s-rrna-gene-sequencing-for-identification-classification-and-quantitation-of-microbes/. [Accessed 17 February 2022].

[42] J. C. Stearns, J. Simioni, E. Gunn, H. McDonald, A. C. Holloway, L. Thebane and et al., “Intrapartum antibiotics for GBS prohylaxis alter colonization patterns in the early infant gut microbiome of low risk infants,” *Sci Rep.,* vol. 7, no. 1, 2017.

[43] A. K. Bartram, M. D. Lynch, J. C. Stearns, G. Moreno-Hagelsieb and J. D. Neufeld, “Generation of multimillionsequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads.,” *Appl Environ Micriobiology,* vol. 77, pp. 3846-3852, 2011.

[44] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads.,” *EMBnet Journal,* vol. 17, no. 1, p. 10, 2011.

[45] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson and S. P. Holmes, “DADA2: High-resolution sample inference from illumina amplicon data,” *Nature methods,* vol. 13, no. 7, pp. 581-583, 2016.

[46] R Core Team, *R: A language and environment for statistical computing.,* Vienna, Austria: R Foundation for Statistical Computing, 2021.

[47] B. Callahan, P. McMurdie and S. Holmes, “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis,” *ISME Journal,* pp. 2639-2643, 2017.

[48] K. Thiele, D. Yeates, K. Abrams and N. Wilson, “It's not the science of tax, and five other things you should know about taxonomy,” *The conversation,* 7 July 2017.

[49] P. J. McMurdie and S. Holmes, “phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data,” *PLoS ONE,* 2013.

[50] Q. Cao, X. Sun, K. Rajesh, N. Chalasani, K. Gelow, B. Katz, V. H. Shah, A. J. Sanyal and E. Smirnova, “Effects of Rare Microbiome Taxa Filtering on Statistical Analysis,” *Fronties in Microbiology,* 2021.

[51] D. J. Lahr and L. A. Katz, “Reducing the impact of pcr-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity dna polymerase,” *Biotechniques,* pp. 857-866, 2009.

[52] D. Knights, J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman and et al., “Bayesian community-wide culture-independent microbial source tracking,” *Nature Methods,* pp. 761-763, 2011.

[53] J. Ravel, P. Gajer, Z. Abdo, G. Schneider, S. K. Koenig, S. McCulle and et al., “Vaginal microbiome of reproductive-age women,” *Proceedings of the National Academy of Sciences of the United States of America,* pp. 4680-4687, 2011.

[54] J. Fettweis, M. Serrano, N. Sheth, C. Mayer, A. Glascock, J. Brooks, K. Jefferson and et al., “Species-level classification of the vaginal microbiome,” *BMC Genomics,* pp. 1-9, 2012.

[55] R. Sinha, C. C. Abnet, O. White, R. Knight and C. Huttenhower, “The microbiome quality control project: baseline study design and future directions,” *Genome Biology,* 2015.

[56] J. T. Nearing, G. M. Douglas, M. G. Hayes and et al., “Microbiome differential abundance methods produce different results across 38 datasets,” *Nature communications,* 2022.

[57] D. Barnett and et al., “microViz: an R package for microbiome data visualization and statistics,” *Journal of Open Source Software,* 2021.

[58] H. Lin and S. D. Peddada, “Analysis of microbial compositions: a review of normalization and differential abundance analysis,” *npj Biofilms and Microbiomes,* 2020.

[59] J. Aitchison, The statistical analysis of compositional data, London: Chapman and Hall ltd. , 1986.

[60] Y. Xia, J. Sun and D. Chen, Statistical Analysis of Microbiome Data with R, Chicago, USA: Springer, 2018.

[61] H. Li, “Microbiome, metagenomics, and high-dimensional compositional data analysis,” *Anual review of Statistics and Its Application,* pp. 73-94, 2015.

[62] K. G. van den Boogaart and R. Tolosana-Delgado, Analyzing compositional data with R., Berlin, Germany: Springer, 2013.

[63] Y. S. Kim, T. Unno, B. Y. Kim and M. S. Park, “Sex Differences in Gut Microbiota,” *The world Journal of Men's Health,* pp. 48-60, 2020.

[64] M. Reyman, M. A. van Houten, D. van Baarle, A. A. T. M. Bosch, W. Ho Man and et al., “Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life,” *Nature Communications,* 2019.

[65] L. W. J. van den Elsen, J. Garssen, R. Burcelin and V. Verhasselt, “Shaping the Gut Microbiota by Breastfeeding: The Gateway to Allergy Prevention?,” *Frontiers in Pediatrics,* 2019.

[66] X. Gao, M. Zhang, J. Xue, J. Huang, R. Zhuang, X. Zhou, H. Zhang, Q. Fu and Y. Hao, “Body Mass Index Differences in the Gut Microbiota Are Gender Specific,” *Frontiers in Microbiology,* 2018.

[67] I. T. Jolliffe and J. Cadima, “Principal component analysis: A review ad recent developments,” *The Royal Society Publishing,* 2016.

[68] J. H. Ward, “Hierarchical Grouping to Optimize an Objective Function,” *Journal of the American Statistical Association,* pp. 236-244, 1963.

[69] G. Martin and S. D. Larson, “Analysis of Variance,” *Circulation,* pp. 115-121, 2008.

[70] A. D. Willis, “Rarefaction, Alpha Diversity, and Statistics,” *Frontiers in Microbiology,* 2019.

[71] M. Cox, W. Cookson and M. Moffatt, “Sequencing the human microbiome in health and disease.,” *Human Molecular Genetics,* pp. 88-94, 2013.

[72] E. H. Simpson, “Measurement of diversity,” *Nature,* p. 688, 1949.

[73] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal,* pp. 379-423; 623-656, 1948.

[74] H. Lin and S. D. Peddada, “Analysis of compositions of microbiomes with bias correction.,” *Nature Communications,* 2020.

[75] S. Weiss and et al., “Normalization and microbial differential abundance strategies depend upon data characteristics,” *Microbiome,* 2017.

[76] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society,* pp. 289-300, 1995.

[77] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes and C. Steinbeck, “ChEBI in 2016: Improved services and an expanding collection of metabolites.,” *Nucleic Acids Research,* pp. 1214-1219, 2015.

[78] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde and M. J. van der Werf, “Centering, scaling, and transformations: improving the biological information content of metabolomics data,” *BMC Genomics,* 2006.

[79] O. M. Kvalheim, F. Brakstad and Y. Liang, “Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise,” *Analytical Chemistry,* pp. 43-51, 1994.

[80] L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, “Scaling. Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS),” *Umetrics,* pp. 213-225, 1999.

[81] Z. Huang and C. Wang, “A Review on Differential Abundance Analysis Methods for,” *Metabolomics,* p. 305, 2022.

[82] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin,* pp. 80-83, 1945.

[83] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt and K. L. Cao, “DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays,” *Bioinformatics,* pp. 3055-3062, 2019.

[84] R. Tibshirani, “Regresson Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society B,* pp. 267-288, 1996.

[85] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.

[86] K. Lê Cao and et al., “A sparse PLS for variable selection when integrating omics data,” *Statistical Applications in Genetics and Molecular Biology,* pp. 1-29, 2008.

[87] A. Sankhya, “Mahalanobis, P.C. "On the Generalised Distance in Statistics.",” *Metrics,* pp. 1-7, 1936.

[88] F. Rohart, B. Gautier, A. Singh and K. A. Lê Cao, “mixOmics: An R package for ‘omics feature selection and multiple data integration,” *PLOS Computational Biology,* 2017.

[89] I. González and et al., “Visualising associations between paired ‘omics’ data sets,” *BioData Mining,* p. 19, 2012.

[90] E. A. Franzosa, K. Huang, J. F. Meadow and et al., “Identifying personal microbiomes using metagenomic codes,” *PNAS,* pp. 2930-2938, 2015.

[91] E. Rinninella, P. Raoul and et al., “What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases,” *Microorganisms,* vol. 7, no. 1, p. 14, 2019.

[92] E. R. Leeming, A. J. Johnson, T. D. Spector and C. I. Le Roy, “Effect of Diet on the Gut Microbiota: Rethinking Intervention Duration,” *Nutrients,* p. 2862, 2019.

[93] B. A. Petriz, A. P. Castro, J. A. Almeida and et al., “Exercise induction of gut microbiota modifications in obese, non-obese and hypertensive rats,” *BMC Genomics,* vol. 15, no. 1, 2014.

[94] D. S. Wishart, D. Tzur, C. Knox and et al., “HMDB: the Human Metabolome Database.,” *Nucleic Acids Research,* vol. 35, no. Database issue, pp. 521-526, 2007.

[95] H. Zafar and M. H. Saier, “Gut Bacteroides species in health and disease,” *Gut Microbes,* p. 13, 2021.

[96] S. I. Sayin, A. Wahlström, J. Felin and et al., “Gut Microbiota Regulates Bile Acid Metabolism by Reducing the Levels of Tauro-beta-muricholic Acid, a Naturally Occurring FXR Antagonist,” *Cell Metabolism,* vol. 17, no. 2, pp. 225-235, 2013.

[97] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes.,” *Nucleic Acids Res.,* vol. 28, pp. 27-30, 2000.

[98] A. Geirnaert, J. Wang, M. Tinck and et al., “Interindividual differences in response to treatment with butyrate-producing Butyricicoccus pullicaecorum 25–3T studied in an in vitro gut model,” *FEMS Microbiology Ecology,* vol. 91, no. 6, 2015.

[99] F. D. Ihekweazu, M. A. Engevik, W. Ruan and et al., “Bacteroides ovatus Promotes IL-22 Production and Reduces Trinitrobenzene Sulfonic Acid-Driven Colonic Inflammation,” *American Journal of Pathology,* vol. 191, no. 4, pp. 704-719, 2021.

[100] R. L. Brown, M. L. Y. Larkinson and T. B. Clarke, “Immunological design of commensal communities to treat intestinal infection and inflammation,” *PLOS Pathogens,* vol. 17, no. 1, 2021.