UNIVERSIDADE DE SÃO PAULO

INSTITUTE OF MATHEMATICS AND STATISTICS
INTERDEPARTMENTAL GRADUATE PROGRAM IN BIOINFORMATICS

Evaluation of the genetic influence on the infant gut microbiome through 16S rRNA sequence data analysis of triplets

Ondina Fonseca de Jesus Palmeira

Advisor: João Carlos Setubal

Co-advisor: Mayana Zatz

Co-Advisor: Júlia Pavan Soler

São Paulo

May, 2022

Ondina Fonseca de Jesus Palmeira

Evaluation of the genetic influence on the infant gut microbiome through 16S rRNA sequence data analysis of triplets

Revised Version

In fulfilment of the requirements for the degree of Master of Science in Bioinformatics by the Institue of Mathematics and Statistics, interdepartmental graduate program in bioinformatics at University of São Paulo.

Advisor: João Carlos Setubal

Co-advisor: Mayana Zatz

Co-Advisor: Júlia Pavan Soler

São Paulo

2022

Ficha catalog

Name: Ondina Fonseca de Jesus Palmeira

Title: Evaluation of the genetic influence on the infant gut microbiome through 16S rRNA sequence data analysis of triplets

In fulfilment of the requirements for the degree of Master of Science in Bioinformatics by the Institue of Mathematics and Statistics, interdepartmental graduate program in bioinformatics at University of São Paulo

Approved in: May 17th, 2022

Committee members

João Carlos Setubal

Institution: Institute of Chemistry, Universidade de São Paulo

Caroline Marcantonio Ferreira

Institution: Universidade Federal de São Paulo – Externo

Lygia da Veiga Pereira Carramaschi

Institution: Institute of Biosciences, Universidade de São Paulo

André Fujita

Institution: Institute of Mathematics and Statistics, Universidade de São Paulo

Acknowledgments

First and foremost, a special thanks to my advisor João C. Setubal and co-advisor Júlia . P. Solar who have been priceless sources of knowledge and inspiration. They always shared their expertise and steered me in the right direction.

I want to express my sincere gratitude for the members of the committee for providing important insights to this work.

I also thank the participants and their caregivers for donating their samples for this study. I thank Dr. Layla Farage Martins and Aline Maria da Silva from the Center for Advanced Technologies in Genomics, Institute of Chemistry, University of Sao Paulo, for guidance on amplicon library construction and Illumina sequencing. I thank Dr. Carla Taddei de Castro Neves, Dr. Andrew Maltez Thomas, M.S. Suzana Guima, M.S. Carlos Morais and Dr. Peter Kochunov for their precious technical support. I am also grateful to all members of the Setulab for providing me with support, tips and constructive criticism.

Finally, I want to express my deepest gratitude to my husband William for always supporting me on my academic journey and my son Nathan for bringing happiness to my life and thus hope to keep me going.

Collaborators and Affiliations

Human Genome and Stem Cell Research - Institute of Biosciences (IB), USP

Larissa Matos, PhD in Experimental Epidemiology applied to Zoonoses. Post-doctoral at

Genetics and Evolutionary Biology Dept., IB, USP

Michel Naslavsky, PhD in Genetics and Professor of Genetics and Evolutionary Biology, USP

Heloisa Bueno, PhD in Experimental Physiopathology, USP

This project was financed by CNPq and FAPESP

I thank CNPq for the finantial support through fellowship (Grant number: 381502/2018-0)

"Faith is taking the first step even when you don't see the whole staircase."

-Martin Luther King Jr.-

"Success begins when you take the first step even when you cannot see the whole staircase."

-Ondina Palmeira-

Abstract

Palmeira, O. F. J.(2022). *Evaluation of the genetic influence on the infant gut microbiome through 16S rRNA sequence data analysis of triplets* (Dissertation of Masters of Science). Institue of Mathematics and Statistics, interdepartmental graduate program in bioinformatics at University of São Paulo, SP

Studies have shown that the human microbiome plays an important role in physiology, from food digestion to mental diseases. Since the gut microbiota composes the greatest amount of microbial cells and genes outnumbering even our own cell and gene counts, it is expected that the gut microbiome would affect many biological functions, thus becoming key to maintaining homeostasis in the various biological processes. The structure of the gut microbiota is shaped by many factors, including the environment and host genetics. Understanding how these factors determine the gut microbiome during its development and establishment at the early stages of human life is crucial to infer commensal and pathological microbiome composition. The purpose of this study is to investigate how much host genetics and the environment influence the development and establishment of the gut microbiota profiles. For this purpose, five sets of dichorionic triplet babies (two monozygotic twins and one dizygotic twin) are followed during their first 3 years of life. By using Next-Generation Sequencing data (NGS) and Bioinformatic tools, such as specific pipelines for 16S amplicons, we will compare the triplets' gut microbiomes regarding presence and absence and relative abundances. We will also try to identify structure patterns, compare results with the literature and integrate the information on the genera associated or not with host genetics. All samples presented enough *reads* to identify all taxa up to the genus level. Phylogenetic alpha diversity increased in samples at later time points indicating time as a determinant factor. Monozygotic twins were significant more similar in beta diversity when compared to their dizygotic co-twins (DZs). Consistent with the literature, *Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria* and *Verrucomicrobia* were the dominant phyla in all sets. Analysis of relative abundance of Amplicon Sequence Variants (ASVs) by Correspondence Analysis (CA) showed that monozygotic twins (MZs) are more similar at time points 9, 11 and 13 months. Heritability test and CA results, as well as shared ASVs, revealed that ASVs of the genera *Veillonella* and *Bacteroides* are more similar in MZs.

Keywords: Microbiome, Microbiota, Host genetics, Heritability, Dichorionic triplet babies, Bioinformatic pipelines, 16S, ASVs, *Veillonella* and *Bacteroides*

Resumo

Palmeira, O. F. J.(2022). Avaliação da influência genética no microbioma intestinal de crianças trigêmeas através da análise de dados de sequencias de rRNA 16S. (Dissertação de Mestrado). Instituto de Matemática e Estatísticas, Programa Interdisciplinar de graduação em Bioinformática, Universidade of São Paulo, SP.

Alguns estudos têm demonstrado a importância do microbioma humano em processos fisiológicos envolvendo desde digestão até doenças mentais. Devido ao grande número de e genes bacterianos no intestino superando até mesmo o número de células e genes humanos, é plausível que o microbioma intestinal afete muitas funções biológicas, tornando-se assim essencial para manutenção da homeostase. A estrutura da microbiota intestinal é modelada por vários fatores, inclusive o ambiente e a genética do hospedeiro. Entender como esses fatores determinam o microbioma durante o desenvolvimento e estabelecimento do mesmo nos primeiros anos da vida humana é crucial para se inferir a composição microbioma relacionada às patologias ou ao bem-estar. O objetivo deste estudo é investigar quanto da genética do hospedeiro e quanto do ambiente influencia o desenvolvimento e estabelecimento do perfil da microbiota intestinal. Para essa finalidade, cinco conjuntos de bebês trigêmeos dicoriônicos (dois monozigóticos e um dizigótico) foram acompanhados durante seus três primeiros anos de vida. Utilizando a tecnologia de *Next-Generation Sequencing* (NGS) e ferramentas da bioinformática, tais como *pipeline* específicos para tratamento e análise de *amplicons* de 16S, iremos comparar os microbiomas dos bebês verificando ausência e presença de bactérias e também suas respectivas abundâncias relativas. Esforços também serão dedicados para identificar padrões estruturais, comparar os resultados com dados da literatura e integrar as informações de bactérias associadas ou não à genética do hospedeiro. Todas as amostras apresentaram quantidade suficiente de *reads* para identificar todas as *taxa* até ao nível de gênero. A diversidade filogenética aumentou nas amostras nos pontos no tempo mais avançados, indicando que o tempo é um fator determinante. Os gêmeos monozigóticos (MZs) foram significantemente mais similares em beta diversidade quando comparados com seus gêmeos dizigóticos. Consistente com a literatura, *Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria* e *Verrucomicrobia* foram os filos dominantes em todos os *sets*. As avaliações de abundância relativa das *Amplicon Sequence Variants* (ASVs), por Análise de Correspondência (CA), mostraram que os MZs são mais similares nos pontos de tempo 9, 11 e 13 meses. Os resultados do teste de herdabilidade, da Análise de Correspondência, bem como das ASVs compartilhadas, revelaram que os gêneros *Veillonella* e *Bacteroides* são mais semelhantes nos MZs.

Palavras-chave: Microbioma, Microbiota, Genética do hospedeiro, Herdabilidade, Bebês trigêmeos dicoriônicos, *Pipelines* de bioinformática, 16S, ASVs, *Veillonella* e *Bacteroide*

List of figures

List of tables

List of abbreviations and acronyms

ASV	Amplicon Sequence Variant
CA	Correspondence Analysis
C-section	Cesarean section
DNA	Deoxyribonucleic acid
DZ	Dizygotic
GMC	Gut microbiome composition
MZ	Monozygotic
NGS	Next-generation sequencing
OTU	Operational Taxonomic Unit
PCoA	Principal Coordinate Analysis
PCR	Polymerase chain reaction
PD	Phylogenetic diversity
QIIME	Quantitative Insights into Microbial Ecology
qPCR	quantitative Polymerase chain reaction
QTLs	Quantitative Trait Loci
RDP	Ribosomal Database Project
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SNPs	single nucleotide polymorphism
SOLAR	Sequential Oligogenic Linkage Analysis Routines
SPINGO	SPecies level IdentificatioN of metaGenOmic amplicons

Table of Contents

1. Introduction

1.1 Human microbiome in health and diseases

The human microbiome, which is the microbial genetic material extracted from samples, has been associated with a variety of biological processes such as the development and establishment of the immune system, metabolic pathways and interactions with the central nervous system(Cho & Blaser, 2012; Kim, Yun, Oh, & Choi, 2018; Lammert et al., 2018; Valdes, Walter, Segal, & Spector, 2018). These associations do not come as a surprise when we learn about human cell composition. Humans are composed of approximately 3.0 x 10^{13} human cells and 3.8x10^{13} bacterial cells(Sender, Fuchs, & Milo, 2016). Moreover, we carry 2 to ~10 million or more bacterial unique genes whereas we have about 20,000 human genes(Li et al., 2014; Ottman, Smidt, De Vos, & Belzer, 2012; Yang, Xie, Li, & Wei, 2009; Zhu, Wang, & Li, 2010). Such a high number of bacterial cells and genes poses a lot of effects in many physiological processes. The long-lasting interactions between the biological processes and the microbiome require microbial adaptation and establishment in the microenvironment, thus different sites on the human body harbor different bacterial communities which, in turn, also differ from individual to individual (**Figure 1**).

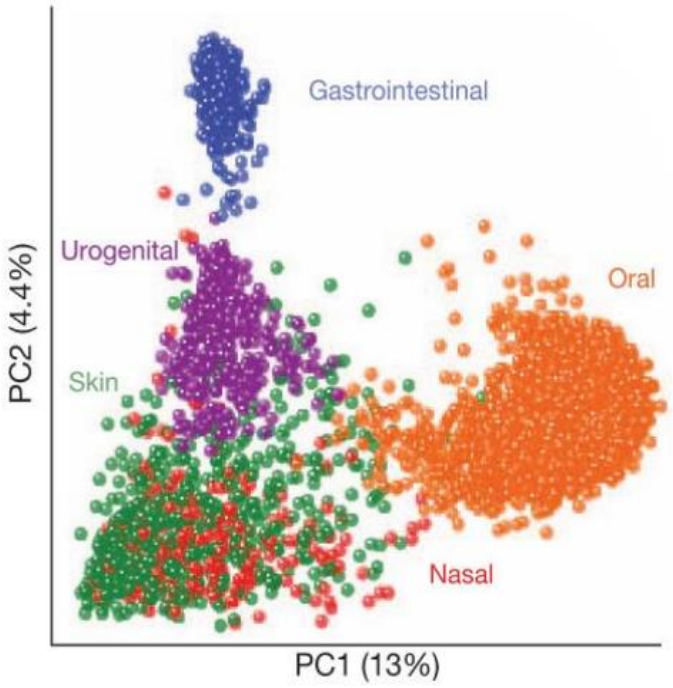

FIGURE 1: PRINCIPAL COORDINATES PLOT SHOWING VARIATION AMONG SAMPLES DEMONSTRATES THAT PRIMARY CLUSTERING IS BY BODY AREA, WITH THE ORAL, GASTROINTESTINAL, SKIN AND UROGENITAL HABITATS SEPARATE; THE NARES HABITAT BRIDGES ORAL AND SKIN HABITATS. (ADAPTED FROM HUTTENHOWER ET AL., 2012)

The main habitat variation of these communities is between the oral, skin, vaginal and gut(Huttenhower et al., 2012). In all these habitats the microbiota (community of microbes in an environment) plays important roles such as prevention of certain diseases and infections by pathogenic bacteria. For example, skin infection by the pathogen *Staphylococcus aureus* is prevented by *Staphylococcus epidermidis* that commensally inhabits the skin(Iwase et al., 2010). Furthermore, some bacteria can also increase the risk of acquiring diseases in the event of dysbiosis – disruption of the symbiotic relationship between microbes and their hosts(Cho & Blaser, 2012). For example, in the vaginal microbiota, enrichment in *Atopobium* and lower abundance in the *Faecalibacterium* genus was associated with human papilloma virus-positive (HPV+) persistence (Di Paola et al., 2017). Although the microbiota of all habitats might be involved in determining the boundaries between health and disease, the gut microbiota is the one that is most often associated with influences on human health (Rojo et al., 2017).

Despite the fact that the gut microbiota presents resilience, its structure can be highly affected by many factors such as maternal microbiome, environment and genetics(Drell et al., 2017; Goodrich et al., 2016; Murphy et al., 2015; Rojo et al., 2017). A healthy human gut microbiota presents *Actinobacteria, Firmicutes, Bacteroidetes, Proteobacteria* and *Verrucomicrobia* as the dominant/prevalent phyla with variation in their relative abundance between individuals (Huttenhower et al., 2012; King et al., 2018). Studies have shown that environmental factors and time are the main cause of interpersonal variation(Rothschild et al., 2018). For example, in the Global Gut Studies, healthy individuals whose age ranged from 0 to 17 years old and from different regions (Amazons of Venezuela, rural Malawi, and United States metropolitan areas) presented differences in their microbiota diversity regarding age, geography and individuality, i.e., microbiota phylogenetic composition of children around 3 years old become more similar to adult microbiota composition; there was a pronounced difference between the US and the Malawian and Amerindian; and finally individuals from the same region and age also presented differences among them (**Figure 2**) (Yatsunenko et al., 2012).

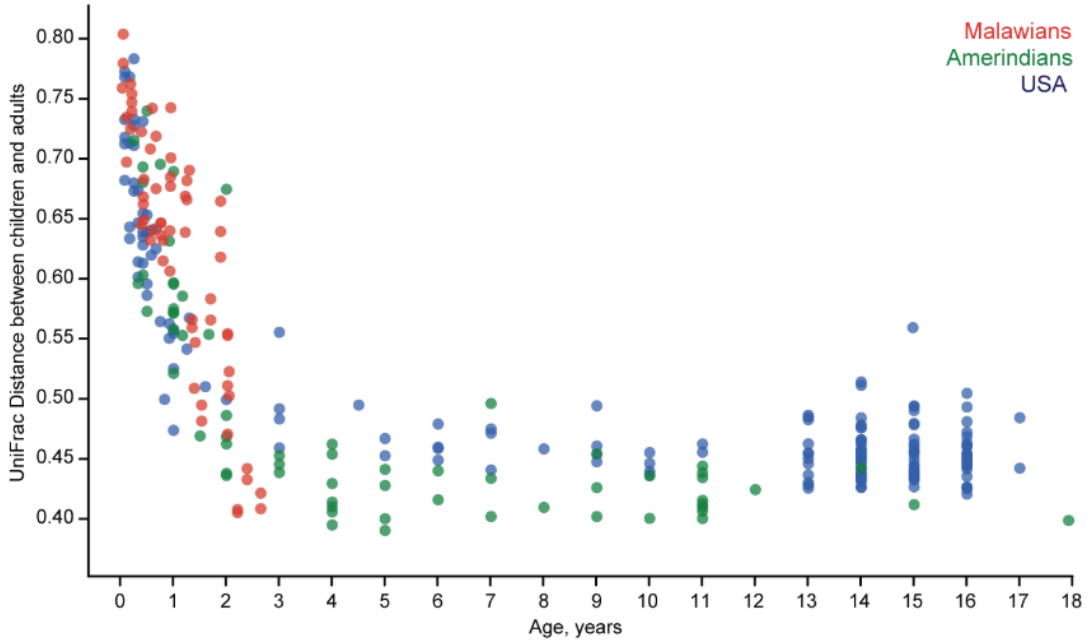

FIGURE 2: UNIFRAC DISTANCES BETWEEN CHILDREN AND ADULTS DECREASE WITH INCREASING AGE OF CHILDREN IN EACH POPULATION. EACH POINT SHOWS AN AVERAGE DISTANCE BETWEEN A CHILD AND ALL ADULTS UNRELATED TO THAT CHILD BUT FROM THE SAME COUNTRY. RESULTS ARE DERIVED FROM BACTERIAL V4-16S RRNA DATASETS. (ADAPTED FROM YATSUNENKO ET AL., 2012)

Regardless of the great variation associated with the above-mentioned factors, it has been possible to identify and distinguish some gut microbial groups that are beneficial to health and others that are associated with certain diseases. For example, the genus *Akkermansia* was negatively correlated with human body weight gain and positively correlated with the protection of the intestinal mucus layer(Belzer & de Vos, 2012; Derrien, Vaughan, Plugge, & de Vos, 2004; Everard et al., 2013). The groups *Faecalibacterium prausnitzii* and bifidobacteria have also been regarded as beneficial. The former, which is in low abundance in Crohn's disease patient's microbiota, showed great anti-inflammatory potential in human cells(Sokol et al., 2008); the latter was shown to be involved in colorectal cancer treatment and prevention, treatment of diarrhea caused by rotavirus, protection against pathogenic bacteria such as *Clostridium perfringens*, remission from ulcerative colitis, and many other benefits (O'Callaghan & van Sinderen, 2016). Conversely, groups, such as *Enterococcus* and *Clostridia,* have been considered as indigenous pathobionts, which, upon a broad-spectrum antibiotic treatment, can invade the bloodstream and generate pseudomembranous colitis, respectively(Chow, Tang, & Mazmanian, 2011).

The associations of the gut microbiome with health and diseases go beyond the scope of gut health status. Disorders such as cardiovascular, autoimmune, chronic kidney disease, to name a few, were all associated with the gut microbiome(Poesen et al., 2016; Rojo et al., 2015; Tang & Hazen, 2014). Furthermore, advanced researches have identified the gut-brain axis, which is bi-

directional communication between the central and the enteric nervous system. Microbial molecules, such as dopamine, γ -aminobutyric acid (GABA) and cortisol present in the human gut are some of the neuronal molecules to which the human host responds, thus being affected by them (Kim et al., 2018; Mayer & Hsiao, 2017; Sharon, Sampson, Geschwind, & Mazmanian, 2016; Valles-Colomer et al., 2019). Taken this into account, it has been shown that the microbiome has a considerable influence on our neurophysiology. For example, neuropsychiatric conditions such as Alzheimer's, Parkinson's, schizophrenia and autism have been linked to the gut microbiome (Chrobak, Nowakowski, & Dudek, 2016; Rojo et al., 2017; Sharon et al., 2016), not to mention the associations with human behavior modulation (Aatsinki et al., 2019; Rogers et al., 2016). For instance, one recent study investigated the relationship between the gut microbiota composition and infant's temperament. They could positively correlate surgency - high levels of activity and positive emotion - with a higher abundance of genera *Bifidobacterium* and *Streptococcus*, whereas negative emotionality and fear reactivity were associated positively with *Erwinia*, *Rothia* and *Serratia* genera (Aatsinki et al., 2019).

Awareness of the great significance of the gut microbiome, not only in the gut organ but also in many other physiological processes throughout the body, compels us to closely examine the human microbiome in all aspects, including influences of genetic variability of the host.

1.2 Gut microbiome and host genetics

It is widely known that the environment has a great impact on the gut microbiome composition (GMC), however, there is a growing number of studies indicating that there are also some host genetic influences. One of the first studies on humans associating host genetics with the gut microbiota was conducted by van de Merwe et al. (1983) who linked pieces of evidence of genetic predisposition of Crohn's disease (CD) with gut microbiota composition associated with CD. By using the faecal culture of monozygotic (MZ) – identical twins – and dizygotic twins – fraternal twins -, and identifying anaerobes, aerobes, gram-positive and gram-negative bacteria, the authors found that the microbiota compositions of MZ co-twins were more similar than those of DZs co-twins(van de Merwe, Stegeman, & Hazenberg, 1983). Later, Zoetendal et al. (2001), using a fingerprinting of bacterial gene approach, investigated the GMC of MZs and their marital partners. The authors found that MZs presented higher similarity indices between co-twins than those of unrelated individuals. In addition, the marital partners, despite sharing the same environment, did not show higher similarity when compared to unrelated individuals (Zoetendal, Akkermansia, Vliet, Visser, & Vos, 2001).

In addition to humans, some non-human models such as mice and chickens have also been used to examine GMC-host genetics associations(Benson et al., 2010; Zhao et al., 2013). One study with a murine model was able to identify 18 QTLs (Quantitative Trait Loci) associated with the relative abundance of some taxa. Two of these QTLs presented quite a significant linkage with taxa at the species level, *Lactobacillus johnsonii* and *L.gasseri*. The authors highlight the close physical association of these organisms to their host and suggest that genetic factors influence these interactions (Benson et al., 2010). Another study, also using QTL analysis, found 6 SNPs (Single Nucleotide Polymorphism) – a gene marker for QTL analysis – statistically significant associated with taxa abundance. Furthermore, the authors were able to replicate 4 SNP associations regarding genus relative abundance of the following genera: rs1394174 with *Faecalibacterium*, rs59846192 with *Lachnospira*, rs28473221 with *Eubacterium*, and rs62171178 with *Rikenellaceae* (Turpin et al., 2016).

Davenport et al. (2015), using both QTL and heritability (a measure of the proportion of genetic variation that accounts for the variation of a trait or phenotype in a population) analysis, investigated the Hutterites population, whose members have quite similar living habits with little differences in their environments (Davenport, 2016; Goodrich et al., 2016). Davenport and collaborators identified 8 bacterial taxa that were significantly associated with SNPs. One of the associations was between the SNP rs3747113 and genus *Lactococcus*, a taxon that was already demonstrated to be associated with QTL in mice by Benson et al. (2010). They also found *Akkermansia's* relative abundance to be associated with SNPs that are involved in signal transduction and trafficking. Furthermore, using 'chip heritability' – the proportion of genetic markers variation that accounts for phenotypic variation -, the authors found a heritability around 20% for the genera *Coprococcus, Lachnobacterium, Barnesiella* and *Veillonella*(Davenport et al., 2015).

A more recent and remarkable study was conducted on 1,126 twin pairs (MZs and DZs) from the United Kingdom by Goodrich et al. (2016). With this large sample size, the authors decreased the confidence intervals of some host genetic-microbiome associations that had already been reported in a previous study (Goodrich et al., 2016). They identified high heritability (42%) for the *Christensenellaceae* taxon followed by *Firmicutes, Actinobacteria, Tenericutes* and *Euryarchaeota*. The most abundant archeon in the gut, *Methanobrevibacter*, was also included in the list with 37% heritability. Among the 20 heritable taxa they identified, *Bifidobacterium* revealed association with the *LCT* gene in the QTL analysis. This was a validation from a previous work by Blekhman et al. (2015) (Blekhman et al., 2015). Some other taxa-SNPs (around genes) associations they reported are: *Bifidobacterium* with *RABGAP1* gene, *Erysipelotrichaceae Cc* 115 with *OR6A2*, *Blautia* with *CD36* and *Akkermansia* with *SIGLEC15*. Moreover, they also showed that these heritable taxa are highly stable throughout time (Goodrich et al., 2016). All these observations argue strongly for the existence of

genetic factors that shape the human gut microbiome to some extent. Nevertheless, Rothschild et al. (2018), by applying multiple statistical analyses on a large human cohort (1,045 individuals), found that there were no significant associations between the microbiome and individual SNPs nor genetic ancestry. They also calculated the overall heritability in 2,252 twins (same dataset used in Goodrich et al. (2016)) to be only 1.9% when accounting for the relative abundance of estimated heritable taxa. Conversely, they could infer 20% of the variance in beta diversity attributable to 95 environmental factors such as food frequency, self-reported median daily intake of calories, age and gender (Rothschild et al., 2018).

These rather conflicting observations indicate that our knowledge about the microbiome is still limited and that further studies are needed to elucidate which and how factors structure the microbiome. It is important to note that there is a dearth of studies relating microbiome to genetic/environmental factors on infants; most of the studies have been conducted on adults. Towards this end, we propose to examine genetic/environmental variability roles in the development and establishment of the gut microbiota using triplet infants as models. To better contextualize this study, we give a short general introduction to the human gut microbiota development and then, we briefly discuss some aspects of twin studies.

1.3 Gut microbiota development and establishment

The human gut microbiota is shaped in early life and it is essential for a healthy individual's development in a world full of microorganisms. The gut is the main site where the immune system and metabolic physiology are trained and modulated, hence how initial colonization takes place can affect the individual's health throughout his/her life (Chong, Bloomfield, & O'Sullivan, 2018; Walker, 2017; Ximenez & Torres, 2017). When investigating the development and establishment of the gut microbiota, researchers have noticed that in the first 3 years of an individual's life there are approximately 3 - 5 phases of the microbiome development in which each is marked by specific changes in the structure (Christopher J. Stewart et al., 2018; Walker, 2017).

Phase 1: It was/is believed that the intrauterine environment was/is sterile, thus humans are born sterile. However, some studies have found bacteria in the placenta tissue. Particularly one study found intracellular bacteria in the placental basal plates (tissue layer adjacent to the myometrium) in 54% of spontaneous preterm and 26% of spontaneous term when investigating 195 patients (Stout et al., 2013). Later, another study revealed a low abundance of nonpathogenic commensal bacteria such as *Firmicutes, Tenericutes* and *Bacteroides* in the placenta. The authors also associated the presence of these bacteria with infection prior to delivery and preterm birth. In parallel, another study showed the presence of enteric and lactic acid bacteria in the newborn's

meconium (first intestinal discharge)(DiGiulio et al., 2008; Gosalbes et al., 2013) and commensal bacteria in the umbilical cord (Jiménez et al., 2005). In addition, non-pregnant woman's uterus revealed the presence of a unique microbiota dominated by *Bacteroides* in a study with 19 women of various health and geographical backgrounds (Verstraelen et al., 2016). One of the speculations attempting to explain these findings is that bacteria are transferred through the maternal blood to the placenta and then to the umbilical cord and reach the fetus (Wassenaar & Panigrahi, 2014), however, there is not enough evidence for this hypothesis. In fact, two recent studies claimed that there is no microbiome in the placenta. One of the studies found the pathogenic *Streptococcus agalactiae* in two cohorts of a total of 478 women including those with pre-eclampsia (de Goffau et al., 2019). However, they suggested an association between the presence of this pathogenic with neonatal disease in the subjects studied. Another study investigated fecal meconium samples and concluded that there is no microbiome in the meconium before birth (Kennedy et al., 2021). Whether there is a placental microbiome or not, there are pieces of evidence that the infant can be exposed to bacteria even before birth, however, we still cannot maintain that these exposures influence the infant's gut colonization.

Phase 2: During birth, the newborn is highly exposed to a community of bacteria whose composition depends on the mode of delivery. In vaginal delivery, most of the exposure is to the mother's vaginal and gut microbiota which starts with the rupture of the chorioamniotic membrane (Maria Gloria Dominguez-Bello, Godoy-Vitorino, Knight, & Blaser, 2019). Next, the passage through the birth canal promotes the infant's swallowing of the microbes of the mother's vaginal microbiota. In fact, studies have shown that the infant's faeces in the first week of life are significantly similar to their mother's vaginal and gut microbiota (Chong et al., 2018; Ferretti et al., 2018; Tanaka & Nakayama, 2017). Interestingly, one study suggested that the fitness of the species seems to be more relevant to the infant's gut colonization than the quantity since the mother-infant shared species were found in lower relative abundance in the mothers (Ferretti et al., 2018). In this same study, the researchers showed that the mother's gut microbiota was the main source for the infant's gut colonization, which gradually differed from their mothers with time, thus indicating a niche-specific selection in the infant's gut. *Lactobacillus, Prevotella* and *Sneathia* are some of the most abundant genus belonging to the vaginal microbiota that are transmitted to the newborn (Maria G Dominguez-Bello et al., 2010; Tanaka & Nakayama, 2017). In Brazilian infants from low-economic backgrounds, *Lactobacillus* was also detected from the second to the seventh days of newborns' lives. In addition, *Escherichia coli* and *Bifidobacterium spp*. were also abundant in this period (Talarico, Santos, Brandt, Martinez, & Taddei, 2017). *Lactobacillus,* which is a vaginal microbiota member, is proved to be beneficial for the infant's initial gut colonization because it binds to the epithelial cells

promoting mucin production. In addition, probiotics containing this organism is recommended for allergy treatment (Houghteling & Walker, 2015). Indeed, newborns delivered by Cesarean section (C-section) lack *Lactobacillus* (Chong et al., 2018) and they are more prone to dairy allergies (Sánchez-Valverde et al., 2009). These c-section born children have a low relative abundance of *Bacteroides fragilis* and a high relative abundance of *Clostridium difficile* (Penders et al., 2006). *Bacteroides* promote increased levels of Th1 chemokines, an important molecule for the immune system modulation (Jakobsson et al., 2014), whereas *C. difficile* is a pathogenic bacteria that can cause intestinal inflammation and diarrhea (Fordtran, 2006). Changes in the microbiome during this phase will also happen according to the mode of feeding. Infants that are exclusively breast-fed present a quite different microbiome when compared to those who are formula-fed included (Azad et al., 2013; Penders et al., 2006). The human milk contains complex oligosaccharides, live Bifidobacteria, soluble CD14 (pathogen co-receptor) and Toll-like Receptors 2 (sTLR2); all essential elements for proper immune system development. Breast-fed infants have high relative abundances of *Bacteroidetes, Bifidobacteria* and a low relative abundance of *Firmicutes* (Houghteling & Walker, 2015).

Phase 3 and 4: Another major change in the structure of the infant's gut microbiota is the introduction of liquids and solid food around the 5th – 12th month of life. At this phase, alpha diversity increases dramatically due to not only the new diet but also potential exposures to other environmental factors such as day-care (Thompson, Monteagudo-Mera, Cadenas, Lampl, & Azcarate-Peril, 2015). Complex carbohydrates like starches and plant's cell wall polysaccharides promote the establishment of fermenting bacteria, *Ruminococcus,* for instance. *Faecalibacterium prausnitzii* which is a butyrate producer that promotes anti-inflammatory responses increases in relative abundance during this phase(Koenig et al., 2011; Christopher J. Stewart et al., 2018). It is also noticed an increase of the relative abundance of adult-type microorganisms such as *Bacteroides, Prevotella, Ruminococcus, Clostridium,* and *Veillonella* (Koenig et al., 2011; Tanaka & Nakayama, 2017). In addition, it was detected genes in the infant's microbiome that are involved with the breakdown of xenobiotic compounds and vitamin biosynthesis, indicating a more adult-like microbiome(Koenig et al., 2011).

Phase 5: By age 1 – 3 years, the infant's microbiome starts to become more stable after dramatically increasing its diversity in the previous phase. Here, the microbiome reaches approximately 1000 species and is very similar to an adult's microbiome. The structure established here will probably remain in the individual's whole life(Walker, 2017). Occasional events such as antibiotics and diseases may cause some variation and disbalance, but the microbiome tends to be resilient and reestablishes its structure once the challenges have passed.

In summary, a typical full-term vaginally delivered and healthy singlet infant from a developed country initiates gut colonization by facultative anaerobes, then as oxygen is being consumed, the gut environment is more favorable to strictly anaerobic bacteria such as *Bifidobacterium, Bacteroides*, and *Clostridium* (Cong et al., 2016). Many factors may influence the infant's colonization such as mode of delivery, gestational age, mode of feeding in the first days, antibiotic treatment, socioeconomic status, geography, exposure to pets, maternal health and diet prior to and after delivery to name a few. All these factors can impact the microbiome composition during the critical period and may determine the development of diseases and conditions such as allergy, asthma, metabolic disorders and autoimmune diseases (Chong et al., 2018). The natural course of colonization is a gradual increase in diversity and richness promoting a balanced development of the immune system and homeostasis of the intestine and allowing proper response to stimuli during key events in the infant's life (Ximenez & Torres, 2017).

Here, we briefly introduced the main environmental factors studied so far that influence the development and establishment of the human gut microbiota. As mentioned in the previous section, little is known about whether or how genetic factors play any role in this process. However, there are a couple of studies with some interesting findings. A study showed that the diversity of genus *Eubacteria* in 4-month MZs was more similar when compared to non-twin siblings (J. A. Stewart, Chadwick, & Murray, 2005). Another study investigated a set of healthy triplets and found that the MZs had more similar microbiota profile between them than between their dizygotic co-twin (Murphy et al., 2015). Due to this lack of pieces of evidence for genetic factors, this study attempts to investigate the issue and contribute to the little knowledge of the subject with a twin studies approach.

1.4 Twin studies

You have probably heard of the twin brothers NASA (National Aeronautics and Space Administration) astronauts Scott Kelly and Mark Kelly who participated in a Twin study, recently. Scott spent 340 days in space and at the same time, his twin brother remained on earth to serve as a control. The aim of the study was to investigate the impact of microgravity on molecular physiological, and behavioral aspects. By comparing the twin's blood, urine, stool, cognitive and psychological test results, the research team could identify some significant differences during the inflight time points. Some of the changes observed were on gene expression and the microbiome composition which went back to baseline on post-flight time points. However, high frequencies of chromosomal translocations and inversions reported on inflight time points remained high on post-flight time points (Garrett-Bakelman et al., 2019). This study, albeit its extremely small sample size,

is one of the examples of the usefulness of twin studies which have been conducted for more than a century to examine the environmental and genetic influences on a myriad of pathological or non-pathological phenotypes (Garrett-Bakelman et al., 2019).

Twins are two individuals born from the same pregnancy and they can be either identical or non-identical or fraternal twins. Identical twins originate from one zygote, which, for some unknown reason, split into two zygotes, i.e., they came from the same egg fertilized by the same sperm and thus their DNA is roughly 100% identical which leads to the same sex. These twins are called monozygotic twins (MZs). Fraternal or non-identical twins are originated from two different zygotes, i.e., two eggs were fertilized, at the same time by different sperms, thus their DNA is around 50% identical just like any other siblings born from different pregnancies. They can be of the same or different sex. These are called dizygotic twins (DZs).

The fundamentals of using twins as models in study design are that they share the same environment and they have two levels of genetic relatedness – 100% or 50% similar. Therefore, the variation of the environment can be reduced while one can examine the differences between the twins. One of the premises of the twin studies is that if a pair of monozygotic twins respond (phenotypic variance) more similarly to a stimulus or variable in comparison to a pair of dizygotic twins, whose response differ from one another, then the response to the variable is more likely to be associated with genetic factors. In contrast, if MZs and DZs responses are all different, then genetics plays little or no role in the response which might be associated with environmental factors (Felson, 2014; Fisher, 1919; Liew, Elsner, Spector, & Hammond, 2005; Rende, Plomin, & Vandenberg, 1990; Sahu & Prasuna, 2016).

Twin studies have been applied for more than a century now and in various fields such as behavioral genetics, intelligence, metagenome and pathological and physiological conditions (Goodrich et al., 2014; Haworth, Dale, & Plomin, 2008; Keller, Medland, & Duncan, 2010). Although the technology of molecular biology has greatly advanced, twin studies are sometimes preferable over DNA sequencing to investigate genetic associations, or twin studies have also been conducted as validation or complement of DNA sequencing. Some of the reasons that explain such facts are that twin studies yield a higher estimation of the proportion of the phenotypic variance than does molecular data since this is limited in capturing all the variation on the DNA, not to mention that molecular data collection is still more expensive (Coventry & Keller, 2005; Felson, 2014; Keller et al., 2010; Lakhani et al., 2019). Nevertheless, DNA sequencing approach to study microbes captures much more information from the microbiomes than studying the microbes per se since some bacteria are difficult to be cultivated. We will develop this theme in more detail in

the next section.

1.5 Next-generation sequencing (NGS)

Microbial communities have long been studied by cultivating bacteria in the laboratory. Despite the many advances in the study, culture technique poses certain limitations such as providing and maintaining media composed of specific and essential Physico-chemical properties for certain bacteria growth making it difficult to cultivate them (Kallmeyer, Pockalny, Adhikari, Smith, & D'Hondt, 2012). Therefore, characterization of the entire community was not possible (Malla et al., 2019; Rosario & Breitbart, 2011). Nowadays, DNA/RNA sequencing approach has overcome some of the culture-dependent approach limitations. By collecting samples directly from the environment and been able to identify its microbiome composition, sequencing approach allows one to better characterize richness and diversity. However, identification of low abundant microbes and certain species or strains are still challenges in the DNA sequencing, thus culture-dependent methods should not be totally disregarded, but applied as a complement to sequencing techniques when appropriate(Hiergeist, Gläsner, Reischl, & Gessner, 2015; Malla et al., 2019).

Sequencing was first developed by Sanger et al. (1975) and was widely used for many years helping in the development of the Human Genome Project in 2001(Collins & Fink, 1995; Sanger & Coulson, 1975). Today, the most performed sequencing technology is next-generation sequencing (NGS) which is a high throughput approach. Its fundamentals are to sequence, at once and parallel, millions of small fragments of the isolated DNA without prior knowledge of the gene or fragment under investigation, and therefore new microbes can be discovered. The technology is cost-effective, can generate higher sequencing depth and yields results in hours or days (Behjati & Tarpey, 2013; D'Argenio, 2018; Panek et al., 2018).

From the many different NGS platforms, Illumina (GAIIx, MiSeq and HiSeq), Ion Torrent, Roche 454 GS FLX and Oxford Nanopore are some of the most used ones. The main differences between them are the generation of *read* length, sequencing depth and error rate(Behjati & Tarpey, 2013; Forbes, Knox, Ronholm, Pagotto, & Reimer, 2017; Malla et al., 2019).

We will briefly describe only Illumina MiSeq principles, since it was the platform used in this study. It is a sequencing-by-synthesis (SBS) approach where fragments of DNA bound to their primers are attached to a microscope slide-like flow cell by adapters and cycles of amplification occurs generating clusters of clones (amplicons) of the targeted fragments. Next, the amplicons serve as templates for the sequencing run where there is an addition of fluorescent reversible terminator nucleotides, images are captured, there is the removal of reversible terminator

nucleotides, the addition of regenerated 3' hydroxyl group nucleotides, and addition of reversible terminator nucleotides again to repeat the cycle. For each image captured, there is a *read* (identification of the color-coded nucleotide) of the added nucleotides. With the sequential images, the sequence of the nucleotides is calculated. This platform generates 25 million of 300bp- length reads (D'Argenio, 2018; Malla et al., 2019; Pfeiffer et al., 2018).

Because several copies of a single sequence are polymerized at the same time in a cluster, some unwanted events can occur such as incomplete wash out of the reversible terminator which can cause the polymerization to lag behind while others advance in the polymerization. This weakens the signal and can cause errors at the time of reading the nucleotide(base calling)(Ewing & Green, 1998; Pfeiffer et al., 2018). For each base called, an error probability is calculated resulting in Phred 33 scores that can range from 0 to 40. Score 20 means that 1 in 100 bases is incorrect, i.e. there is a certainty of 99% that the base is correct(Ewing & Green, 1998).

In metagenomics, which is a technique of extracting genetic information from an environment, there are basically two approaches. One can sequence the whole genome of the organisms present in a sample (shotgun metagenomics) or sequence only marker genes such as 16S small subunit of RNA ribosomal gene (targeted metagenomics)(Malla et al., 2019; Panek et al., 2018).

Shotgun approaches consist of deep sequencing of fragments of the sampled DNA and assembling by overlapping regions. Then, one can recover entire genomes or infer taxonomy or functions by analysis of long sequences called *contigs* (Malla et al., 2019; Quince, Walker, Simpson, Loman, & Segata, 2017). Since shotgun yields a great deal of genetic information, this approach brings some advantages and disadvantages over the target metagenomics. Some of the advantages are more reliable functional inferences, sequencing DNA of various organisms belonging to different kingdoms at the same time in a sample and better identification and characterization of the microbes present in a sample. Conversely, sequencing all DNA present in a sample may pose challenges to distinguish genomes, especially if they are at low abundance. Furthermore, it requires more powerful computational performance, thus leading to higher costs and more time consuming (Behjati & Tarpey, 2013; Forbes et al., 2017).

Targeted amplicon metagenomics, small subunit of 16S ribosomal RNA gene marker, which is the approach applied in this study, will be discussed in more detail in the next section.

1.6 Microbiome analysis through targeted 16S gene sequencing

Targeted amplicons of 16S ribosomal RNA gene (16S) is the most widely used method of NGS to study microbiome composition. The marker 16S is highly conserved in all bacteria and archaea

since it is a housekeeping gene which is essential for bacterial protein expressions. This gene is comprised of approximately 1500 base pairs which are divided into 9 shorter regions from V1 to V9(**Figure 3**). These regions are called hypervariable because each one has certain variation between the different bacterial taxa, i.e., they can be species-specific(Bukin et al., 2019; D'Argenio, 2018; Liu, DeSantis, Andersen, & Knight, 2008). The entire 16S can be studied, as well as just a fraction of it, by using primers specific to the regions of interest. The choice of regions will depend on the goal of the study, for example, studies suggest that the entire 16S would better represent the microbe community, while others argue that only one to three regions would suffice to address one's question (Forbes et al., 2017; Yarza et al., 2014; Zhang et al., 2018). One study assessed selected microbes that had clinical relevance from the human gastrointestinal tract and they could accurately identify and quantify all of them with the 16S approach (Almonacid et al., 2016). Interestingly, in another study when evaluating PCR and qPCR positive *Clostridium difficile* samples, they could better identify this species with 16S than with shotgun approach(Zhou et al., 2016). A recent study indicated that the most reliable regions regarding community composition were V1 – V2 and V1 - V3. On the other hand, bacterial diversity was underestimated by these regions, unlikely by V4 (Zhang et al., 2018). Another study suggested that phyla abundance detected with V2-V3 or V3-V4 did not differ, while at class and family levels differences were larger (Bukin et al., 2019). In contrast, V2 and V3 were showed to be the most satisfactory regions to distinguish all bacterial organisms at the genus level (Chakravorty, Helb, Burday, Connell, & Alland, 2007). Furthermore, V2-V3 showed the highest resolution at genus and species levels when compared to V3-V4(Bukin et al., 2019).

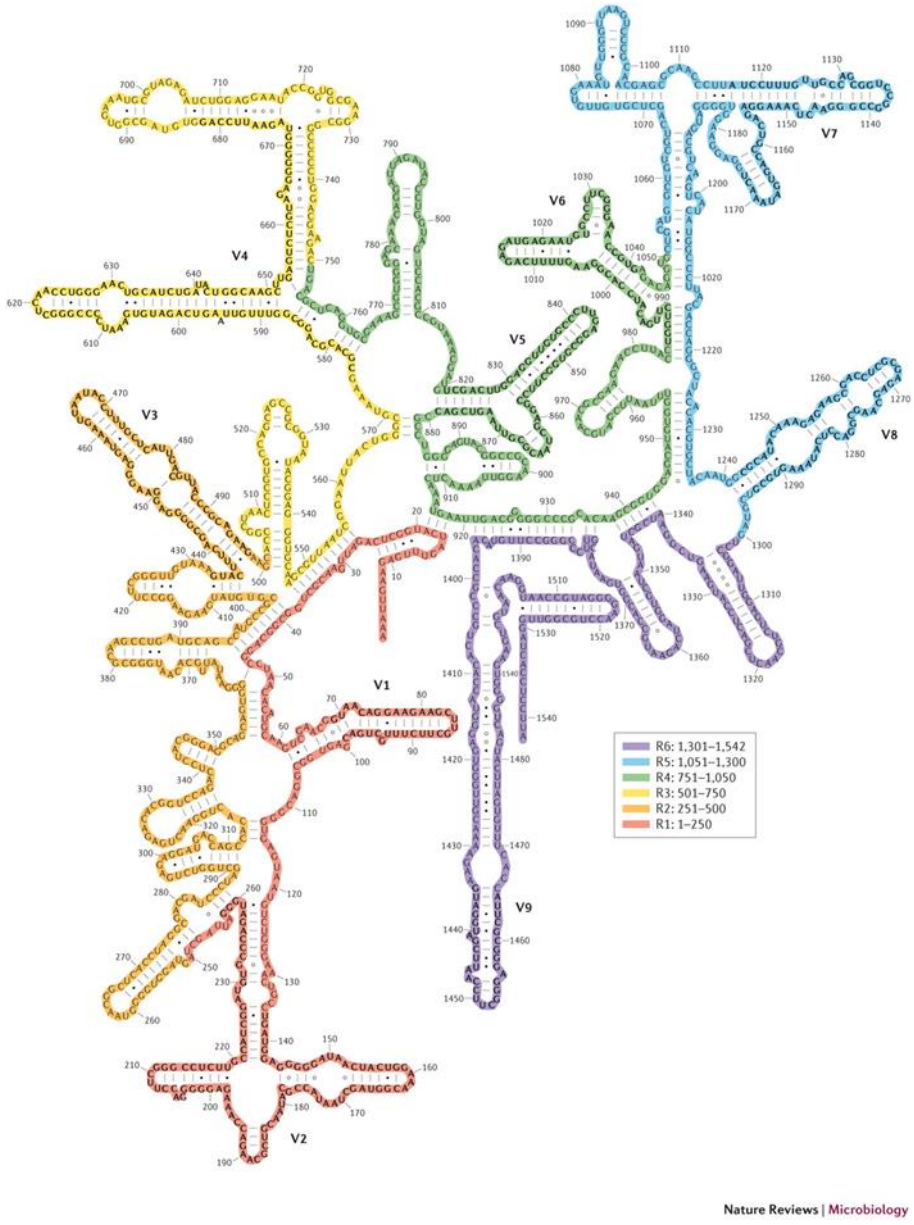

***FIGURE 3: SECONDARY STRUCTURE OF THE 16S RRNA OF ESCHERICHIA COLI, AS GENERATED USING THE XRNA PROGRAM. FOR OUR ANALYSIS, SIX R FRAGMENTS OF ~250 NUCLEOTIDES WERE DESIGNED ACCORDING TO THE KNOWN V REGIONS. IN RED, FRAGMENT R1 INCLUDING REGIONS V1 AND V2; IN ORANGE, FRAGMENT R2 INCLUDING REGION V3; IN YELLOW, FRAGMENT R3 INCLUDING REGION V4; IN GREEN, FRAGMENT R4 INCLUDING REGIONS V5 AND V6; IN BLUE, FRAGMENT R5 INCLUDING REGIONS V7 AND V8; AND IN PURPLE, FRAGMENT R6 INCLUDING REGION V9.* (ADAPTED FROM YARZA ET AL., 2014)**

All these observations suggest the difficulty to pinpoint which region would be a better representation for the microbiome composition and structure as a whole. While one can increase reliability regarding richness, one may lose true estimation of abundance. Moreover, different organisms are better identified by different regions, such is the case of V3 that showed to accurately distinguish species within *Staphylococcus* sp. and *Haemophilus*, but was limited at the family *Enterobacteriaceae* (Chakravorty et al., 2007). Nevertheless, V3 and V4 regions are widely sequenced, either in conjunction or separately. What is more, V4 was indicated to yield results similar to the

full-length 16S (Youssef et al., 2009). In fact, several studies on human gut microbiome associating host genetics used V3 or V4 on their investigations (Davenport et al., 2015; Goodrich et al., 2014; Knights et al., 2014; Murphy et al., 2015; Org et al., 2015; Rothschild et al., 2018; C. J. Stewart et al., 2013; Subramanian et al., 2014; Turpin et al., 2016; Zhao et al., 2013). As a means to compare our results with the literature, we chose to use V3-V4 in our investigation as well.

Regardless of the chosen hypervariable region, taxonomic classification is usually performed by the alignment of the sequences to 16S-specific databases such as Ribosomal Database Project (RDP)(Cole et al., 2014), Greengenes(DeSantis et al., 2006), or Silva(Yilmaz et al., 2014) after processing the reads of the amplicons. Note that different databases can also contribute to the divergent above-mentioned results of 16S regions due to the database's different limited annotations (D'Argenio, 2018).

Despite the fact that 16S approach has yet many challenges, it still offers some advantages over shotgun sequencing. It can provide more sensibility to certain species; there are more accurate 16S specific databases; changes in relative abundance can be assessed and compared between different samples; if sequence depth is sufficient, it can provide reliable abundance of rare taxa; it can be performed on most sequencing platforms; it is cost-effective by requiring less computational manipulation and thus less time consuming (Donkor, 2013; Forbes et al., 2017; Hiergeist et al., 2015). Manipulation and processing of the amplicons will be further explored in the next section.

1.7 Bioinformatics pipelines for 16S data manipulation and analysis

Amplicon sequences obtained from samples are computationally represented by *reads* generated during sequencing. In order to infer the microbiota composition in a sample with maximum confidence, those *reads* need to be accurate or almost accurate representations of the sequences present in a sample. Towards this end, one must perform gold-standard and rigorous processes on the pool of reads before attempting to infer which microorganisms were present in a sample. Such processes include trimming, filtering by quality score, alignment against a database, clusterization or feature assignment, generation of phylogenetic trees and taxonomic classification. Currently, there are several well designed and 16S specific pipelines and tools that perform all or part of those processes. To name a few, mothur (Schloss et al., 2009), Integrated Microbiome Analysis Pipeline (iMAP)(Buza et al., 2019), Phyloseq(McMurdie & Holmes, 2013), MetaAmp(Dong et al., 2017), MAPseq (Matias Rodrigues, Schmidt, Tackmann, & von Mering, 2017), Quantitative Insights into Microbial Ecology (QIIME)(Caporaso et al., 2010) and Quantitative Insights into Microbial

Ecology (QIIME 2) (Bolyen et al., 2019). Some of these and other tools and pipelines can be used as complements of one another as long as they share file formats, or one can convert output files to an appropriate format to input in another tool or pipeline.

Here, we will focus on QIIME 2 (pronounced "chime") since it is the main pipeline used in this study. Briefly, the advantages offered by this pipeline over the others are: its open-source property allows easy and constant improvements of the codes; it is developed by a community that could include anyone who can truly contribute to the software evolution; it is constantly updated; it has a forum where users and developers can interact and finally, it is free(Bolyen et al., 2019; Caporaso et al., 2010). QIIME 2 is the only Python-based software for microbiome analysis, which is also an advantage since Python is free and has user-friendly data structures that can be used and run in most operational systems.

A study benchmarking mothur, MAPseq and QIIME 2 pipelines showed that the latter performed the best in recall and F-scores at family and genus levels (Almeida, Mitchell, Tarkowska, & Finn, 2018).

Basically, QIIME 2 pipeline is composed of plugins that execute all those important above-mentioned processes on the reads. The pipeline also provides plugins for qualitative and quantitative analysis. Here, the plugins we used are mentioned in the Materials and Methods section. Yet, it is important to introduce some of the concepts and algorithms of the processes. In general, after quality control of the reads, the sequences are clustered by similarities. Clusterization reduces the number of reads to be analyzed, control for non-biological sequences, and control for recent divergent evolution. One way to cluster is to perform OTU picking. OTU stands for "operational taxonomic unit". Essentially, sequences that are somewhat similar – usually >97% similarity – are cluster together making it a unit that represents a certain taxon(Nguyen, Warnow, Pop, & White, 2016). Existing OTU picking approaches are closed, *de novo* and open reference. Closed-reference approach is database-dependent in which all sequences are aligned to a database and all the hits are kept, while sequences without a hit are discarded. *De novo* approach is database-independent in which sequences within the sample are aligned to each other and clustered according to their similarities. Open-reference approach is both a closed-reference and *de novo* approach in which all sequences without a hit on the database are clustered together according to their similarities (Edgar, 2017). Many studies have shown some problems with the OTU picking process (Callahan, McMurdie, & Holmes, 2017; D'Argenio, 2018; Mysara et al., 2017; Nguyen et al., 2016). They argue that OTUs might bring bias by overestimating or underestimating phylogenetic diversity. One problem is the incorrect clusterization by applying a general cut-off to all sequences. This happens

because genes do not evolve at the same rate within different taxa. For example, on the one hand, some organisms might have more than 98% similarities in the entire 16S rRNA gene, yet they are completely different species. On the other hand, some families have great variability in certain 16S regions, yet they are from the same family (Edgar, 2017; Panek et al., 2018). Another problem is the reproducibility since the de novo OTU picking is a sample property, i.e., OTUs will differ from sample to sample. Moreover, closed OTU picking does not allow for the discovery of new organisms (Mysara et al., 2017).

In an attempt to solve these issues, new algorithms have been developed such as DADA 2 (Callahan et al., 2016) and Deblur (Amir et al., 2017). QIIME 2 provides both algorithms and they mainly differ on the run time of process as well as the percentage of sequence retention (Nearing, Douglas, Comeau, & Langille, 2018). Since Deblur was our algorithm of choice, we will focus only on it in this introduction. Considering that the human gut microbiome is widely studied and thus more curated databases are available, we chose to apply a more stringent algorithm in order to decrease false positives, i.e., at this stage of the study we are not interested in discovering new organisms, but characterizing the ones we have learned from the literature at early human life. Deblur algorithm takes one sample per time and generates error profiles by calculating the probability of indels and error rate based on upper error rate bound. With this, it predicts spurious sequences and subtracts all their Hamming distance neighboring reads from the sample. Reads whose frequency drops to zero is discarded. After chimeric removal and alignment with the Greengenes database, the remaining sequences are the representative sequences. This algorithm can differentiate sequences by a single nucleotide, termed amplicon sequence variants (ASVs) or sub-operational taxonomic unit (sOTUs), and predicts if they are the same taxon or if they are in fact different (Amir et al., 2017). Thus, the ASV assignment outperforms OTU methods by promoting sensitivity and specificity to infer microbes from 16S amplicons, and they can be considered as true biological sequences (Caruso, Song, Asquith, & Karstens, 2019). Recent studies have suggested and recommended that ASV should replace OTU methods because it is not only independent of a database but it can also be compared across samples and datasets (Callahan et al., 2017; Caruso et al., 2019; Knight et al., 2018; Nearing et al., 2018).

Lastly but not least, proper analysis of the microbiome requires an assessment of alpha and beta diversity besides taxonomic classification. Alpha diversity is a measure of richness (number of different organisms) and abundance of organism (frequency of each organism) within a sample, while beta diversity is a measure of richness and abundance between samples, i.e., measure the overall change (Wagner et al., 2018; Whittaker, 1972). There are several methods to measure alpha diversity. The evaluation might be quantitative or qualitative. Quantitative measures can be done

by applying Shannon index (H') which is essentially a statistical measure based on probability distribution, i.e.,

$$H' = -\sum p_i \ln(p_i)$$

where *i* is the *ith* species, *pi* is the proportion of the *ith* species and *ln* is the natural logarithm (Morris et al., 2014; Shannon, 1948). A qualitative measure of alpha diversity can be done simply by counting the number of different organisms present in a sample (community). These two above mentioned methods do not account for the relatedness between the organisms. Faith diversity is a qualitative method that considers this relatedness, thus is termed as phylogenetic diversity (PD). Faith PD is based on cladistic information (shared derived characteristics) between the taxa (Faith, 1992).

To calculate beta diversity, one can apply Bray Curtis dissimilarity (BC_{ij}), which measures the presence and absence as well as abundance of each organism. This method takes the sum of the lowest value of all common taxa in each pair of sample and divides by the total count of taxa in the pair of sample, i.e.,

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where, *i* and *j* are two different samples and S is the sum of all species in sample *i* or *j* (Bray & Curtis, 1957; Primicerio, 2013). Another measure of beta diversity is Jaccard index ($J_{X,Y}$), a coefficient of matching (similarity) between a pair of samples relative to the number of organisms present in at least one sample of the pair. It counts the co-presence ignoring the co-absence (Jaccard, 1912):

$$J_{x,y} = \frac{x \cap y}{x \cup y}$$

This is a qualitative measure of diversity between samples. Again, these two methods do not account for phylogenetic diversity. There is a relatively new method developed by Lozupone et Knight (2005) that takes phylogeny into consideration. It measures the fraction of the tree where there are organisms from a particular sample and it is unique to that sample. This measure is called the Unifrac distance. If two samples cover the same fraction of the tree, their unifrac distance is zero, that is, they are phylogenetically similar, whereas if they cover totally different fraction of the tree (their fraction is unique to one or another), then they are phylogenetically different. Unifrac can be applied taking into consideration only the presence and absence

(unweighted Unifrac), as well as the abundance (weighted Unifrac) of each bacterial lineage (C. Lozupone & Knight, 2005). Note that all beta diversity measures mentioned here result in a distance matrix which can be visualized with ordination technique such as PCoA (Principal Coordinate Analysis). PCoA is one of the techniques of dimensionality reduction that enables one to visualize how close (similar) or distant (dissimilarity) one sample is from another regarding the targeted multi variables (organisms)(Groenen, 2005). QIIME 2 pipeline provides plugins that not only calculate alpha and beta diversity but also provides a platform for visualizations of these results. For each diversity assessment, appropriate statistical tests should be applied to evaluate the power of the results, but at this stage of the study we applied minimal statistical tests and only descriptive statistics was performed due to the small sample size. The various methods applied for ecological and compositional data analysis will not be introduced or discussed here since it is beyond the scope of the study at the present moment.

2. Aims

The aim of this study is to investigate what role host genetics plays in the establishment and development of the gut microbiota in infants based on longitudinal rRNA 16S data fromdichorionic triplets.

2.1 Specific aims

Our aims were to answer the following questions:

1. Do MZ twins differ significantly from their DZ sibling in terms of microbiota diversity?
2. Do MZ twins differ significantly from their DZ sibling in terms of ASV composition?
3. If there are significant differences, can they be explained by specific ASVs?
4. Compare results obtained from the present study with those in the literature for singleton, twin and triplet infants

3. Materials and methods

3.1 Basic computational processing on the datasets

3.1.1. Participants and sample collection

Faecal samples were collected from 5 sets of infant triplets ranging from 1 to 37 months old by Dr. Larissa Matos. Some samples from the sets were collected at the same time points, but others were collected at different time points, thus the samples are unbalanced (Figure 4). There were a total of 111 samples, but 12 were discarded because one, two, or all three triplets in a triplet set were taking or took antibiotics within the previous 30 days. This resulted in 99 accepted samples (Table 1). The children's parents authorized the collection and investigation by written informed consents. This project was approved by Plataforma Brazil with the number CAAE: 15291119.2.0000.5464. About 30g of all samples were collected by either the infants' caregivers or the mothers from diapers in tubes universal DNA collectors and/or preservative tubes Omnigene-Gut OMR-200 GenoTeck® at their own homes. Samples were immediately stored at 4C or -20C for up to 24 hours, then transported in ice to the facility where they were stored at -20C or -80C until processing. For every collection, Dra Larissa applied questionnaires for clinical follow-up. Clinical data about the individuals is described in Table 1. In this document the sets were named as A, B, C, D and E. In all sets there were 2 monozygotic twins and 1 dizygotic twin who we identified as MZ1 (monozygotic twin 1) MZ2 (monozygotic twin 2) and DZ (dizygotic twin). MZs are 100% genetically similar, while DZ have 50% genetically similarity with his/her siblings (Figure 5).

Time points with valid samples per set

T	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	...	36
A		✓				✓	✓		✓		✓	✓	✓											✓
B						✓			✓		✓	✓	✓	✓				✓		✓	✓	✓		
C																✓			✓		✓	✓		
D									✓		✓		✓					✓						
E	✓	✓				✓	✓		✓		✓		✓											

FIGURE 4: TIME POINTS WHEN SAMPLES WERE COLLECTED PER SET

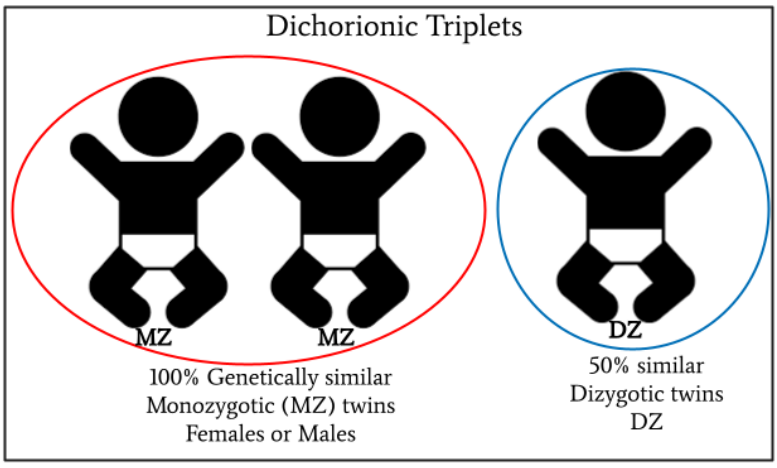

FIGURE 5: DICHORIONIC TRIPLETS WHERE TWO MZS ARE IN THE SAME CHORION AND THE SIBLING DZ IS IN A SEPARATE CHORION. MZS ARE GENETICALLY IDENTCAL AND DZ IS THE FRATERNAL SIBLING.

3.1.2 Sample processing and 16S rRNA gene sequencing

DNA extraction, PCR and sequencing were performed by Dr. Larissa Matos. For bacterial DNA extraction QIAmp DNA Stool MiniKit (QIAGEN ®, UK) was used. The extracted bacterial DNA were quantified by both fluorometric quantification using Qubit Assay (ThermoFisher Scientific, Ireland) or Quant-iT ™ PicoGreen ™ dsDNA (ThermoFisher Scientific, Ireland) and by absorbance in NanoDrop ND-1000 Visible UV Spectrophotometer (ThermoFisher Scientific, Ireland) Amplicons of the V3-V4 region of 16 ribosomal RNA gene were generated by Polymerase Chain reaction (PCR) with U341F primer, 5'-CCTACGGGRSGCAGCAG-3' (17 bases) and and 806R (GGACTACHVGGGTATCTAAT). The samples were normalized and pooled for 4nM following sequencing at 12pM with 20% PhiX as control with 500 cycles per run using the Illumina MiseqTM 2500. Instructions and reagents lists were obtained from the "16S Metagenomic Sequencing Library Preparation protocol (Part #15044223 Rev. B)" provided by Illumina Inc.

3.1.3. Quality filtering and reads processing

We processed the 16S reads with QIIME2 (Quantitative Insights Into Microbial Ecology) version 2019.10(Bolyen et al., 2019). Forward and Reverse read files were imported to Qiime2 separately. First, we trimmed off primers from forward and reverse reads by applying the *q2_cutadapt* plugin(Martin, 2011) with the *trim-single* method and *p-front/p-adapter* option, which removes adapters at the 5'/3' end and any preceding bases. Attempting to merge forward and reverse reads were performed with qiime *vsearch join-pairs* (Rognes T, 2016.) and *bbmerge* (Bushnell, Rood, & Singer, 2017) before and after trimming, but the quality of the reverse reads

were so low that the merged reads could not reach our threshold of Phred quality scores of 20 in more than 50% of the positions. Thus, we proceeded with our bioinformatics processing with the forward reads only. Reads that were retained after quality filtering with plugin *quality-filter q-score* (parameters: $q = 20$, $r = 3$, $p = 0.75$ and $n = 0$)(Bokulich et al., 2013) were processed for ASVs (Amplicon Sequence Variants) assignment with Deblur(Amir et al., 2017). The reads were trimmed at position 230 base pair within the algorithm of Deblur. Generation of phylogenetic tree was performed with plugin *q2-phylogeny,* where the alignment was made with *MAFFT* program (Katoh, Misawa, Kuma, & Miyata, 2002) and *fasttree2* (Price MN, 2010) to build the tree.

3.2. Standard analyses on each individual dataset

3.2.1 Diversity analysis

To evaluate alpha diversity - diversity within samples - , we first examine if the number of denoised reads was representative of the microbial richness in each sample. To this end, we ran *ggrare* with step size = 10 from the *ggplot2* R package to generate Alpha rarefaction curves. Then, we performed alpha diversity measures (Shannon index, Faith diversity(Faith, 1992) and Observed ASVs) within the pipeline QIIME2 (*core-metrics-phylogenetic*) with *p-sampling-depth* = 10,000 (normalization). Graphs of alpha diversities measures were plotted with *geom_boxplot* from *ggplot2* R package. We used the R package *phyloseq* version 1.36.0 to normalize the data by rarefying to 10,000 reads sampling depth without replacement and to perform beta diversity analysis by Jaccard distance, Bray Curtis dissimilarity, weighted and unweighted UniFrac (C. Lozupone & Knight, 2005; C. A. Lozupone, Hamady, Kelley, & Knight, 2007). PCoA graphs were generated with R package *phyloseq* version 1.36.0 and ellipses were calculated by *ggforce:: geom_mark_ellipse*. Permutation tests on alpha and beta diversities were performed to check for statistical significance on mean difference, μ_D, based on alpha = 0.05. We used as models

$$\mu_D = (y_{m_1} - y_{m_2}) - \frac{(y_{m_1} - y_d) + (y_{m_2} - y_d)}{2}$$

for alpha diversities and

$$\mu_D = (y_{m_1 m_2}) - \frac{(y_{m_1 d}) + (y_{m_2 d})}{2}$$

for beta diversities, where *y* is the mean or the median of the values of the measures

(Shannon, Faith, Observed ASVs, Jaccard, Bray Curtis, weighted and unweighted UniFrac); m_1 is the value of monozygotic twin 1; m_2 is the value of the monozygotic twin 2; d is the value of the dizygotic twin; m_1m_2 is the distance/dissimilarity between monozygotic twin 1 and monozygotic twin 2; m_1d is the distance/dissimilarity between monozygotic twin 1 and dizygotic twin; m_2d is the distance/dissimilarity between monozygotic twin 2 and dizygotic twin. The model calculates the mean or median differences between the groups considering the relationship between monozygotic-monozygotic twins and monozygotic-dizygotic twins. Since our sets are unbalanced, we calculated the mean difference per set as follows:

$$\mu_D = \frac{\left(y_{m_1} + y_{m_2}\right)}{2} - y_d$$

for alpha diversity and

$$\mu_D = y_{m_1 m_2} - \frac{y_{m_1 d} + y_{m_2 d}}{2}$$

for beta diversity, taking into account the longitudinal profile by set. Then, we permuted 10,000 times the original values of the groups m_1, m_2 and d (or m_1m_2, m_1d, m_2d) and sampled with replacement. *P*-values were obtained by calculating the number of times the mean or median from our models were present within the distribution generated by the permutations, i.e, if the values from the models which considered the different relationships between the groups (m_1, m_2, d) were rarely (0.05 or lower) present in a randomized distribution, we concluded that there was a significant difference between the monozygotic and dizygotic groups. We counted and identified the ASVs in each member of the sets and descriptively compared common ASVs between the groups of monozygotic twins and their dizygotic co-twins using standard commands in R. Visualization of the common ASVs was made with Venn diagrams using the R package *VennDiagram* version 1.6.17. In order to check if the fractions of ASVs in the Venn Diagram were significant larger in MZs, we applied a statistical test, *Z*, for the comparison of two proportions in a multinomial model using R:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{(\hat{p}_1 + \hat{p}_2) - (\hat{p}_1 - \hat{p}_2)^2}{\sum_i^n \hat{p}_i}}}$$

where $\hat{p}$ is the estimated probability of sharing ASVs between 1 = MZ1 and MZ2, 2 = MZ1 and DZ, 3 = MZ2 and DZ in each set, and $\sum_i^n \hat{p}_i$ is the sum of all estimated probabilities in

the Venn Diagram in the corresponding set. Overlapping of 95% confidence intervals – Goodman for the probabilities of occurrence were calculated with the function GM and p-values calculated with the function *pchisq*; both functions from the *CoinMinD* R package. We adopted alpha = 0.05.

3.2.2 Taxonomic classification

We exported the ASVs and their abundance table produced by Deblur out of QIIME Environment and classified the ASVs with RDP(Ribosomal Database Project) classifier version 2.12(Wang, Garrity, Tiedje, & Cole, 2007) and SPINGO (SPecies level IdentificatioN of metaGenOmic amplicons) version 1.3(Allard, Ryan, Jeffery, & Claesson, 2015), using the databases RDP version 11.5(Cole et al., 2014), setting a minimum confidence score of 0.85 in both classifiers. Output taxonomy files were merged (merge script in R created by Andrew Thomas (Personal communication)) followed by addition to the table BIOM (Biological Observation Matrix) file (McDonald et al., 2012). The resulting file was imported back to QIIME environment or to RStudio to proceed with the analysis. Histograms for both Phyla and Genus relative abundances were created with the R package *ggplot2* version 3.3.3.

3.2.3 CA analysis

Correspondence analysis (CA) in each set was performed in the R studio. We used the absolute abundance of ASVs present in each set to check for associations between the ASVs and the samples. First, we performed the CA with the 50 most abundant ASVs in each set allowing for clear visualization on the graph and avoiding overlap of variables which would make it difficult to read. We applied the function *fviz_ca_biplot* from the R package *factoextra* (version 1.0.5) setting the map to *rowprincipal* which generates an asymmetric plot with rows (samples) in principal coordinates and columns (variables - ASVs) in standard coordinates, i.e, samples with similar abundance cluster together and the variables that contribute to the clustering are positioned in a low angle with the samples revealing their associations. Then, we calculated the Person´s chi square, χ^2, test with simulated *p*-value based on 2,000 replicates to overcome the problem of very low expected values or zeros on the dataset. Since the contingency tables of all sets had large sample size (N > 450 cells), we calculated the coefficient of contingency by:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

We performed the CA again later with those 50 most abundant ASVs adding all the ASVs whose heritability (explained below) was significant, regardless of their abundance. We selected time points 9, 11 and 13, where we noticed clusterization from previous CA and where the heritability tests were applied. Here, we calculated the log2 likelihood ratio

$$ratio = \frac{observed\ value}{expected\ value}$$

as an index to interpret the association between the samples and the ASVs. The ratio calculation serves to interpret the ASVs-samples significance of association. The calculation of similarity of ratios was done by subtracting the ratios of the pairs compared (MZ1 vs MZ2, MZ1 vs DZ, MZ2 vs DZ).

Any process performed in R was done in R version 4.1.1, 4.1.0, 4.0.5, 4.0.4, 4.0.3, and 3.5.2, and RStudio version 1.4.1717.

3.2.4 Heritability

Heritability tests were applied to sets A, B, D and E at time points 9, 11 and 13, and it was made by testing H_0: $\sigma_g^2 = 0$ against H_1: $\sigma_g^2 > 0$ through the likelihood ratio statistic, which is asymptotically distributed as a 1/2:1/2 mixture of Chi-square with 0 and 1 degree of freedom. SOLAR(Sequential Oligogenic Linkage Analysis Routines) Eclipse version software package version 8.5.1 (beta)(Kochunov et al., 2015) (http://www.solar-eclipse-genetics.org/) was used for the test with default settings. SOLAR software applies variance component models under maximum likelihood estimation approach using the following matrix to take into account family dependence among individuals:

$$\Omega = 2\Phi\sigma_g^2 + I\sigma_e^2$$

where, Ω is the covariance matrix for a pedigree of individuals, Φ is the kinship matrix of pair-wise kinship coefficient among all individuals, σ_g^2 is the genetic variance component due to additive genetic factors, I is the identity matrix, and σ_e^2 is the variance component due to

specific environmental effects of each individual, such that

$$H2r = \frac{\sigma_g^2}{\sigma_p^2}$$

is the narrow sense heritability and

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2$$

is the total phenotypic variance (Kochunov et al., 2015).

First, we excluded from our analysis the ASVs that were not present in 50% of the samples of the sets following criteria in the literature (Goodrich et al., 2016; Goodrich et al., 2014; Xie et al., 2016). Then, we filtered in, per time point, ASVs that were present in at least 3 sets and at least in one monozygotic and dizygotic co-twin in each set. The phenotype file containing relative abundance responses was input into the software and normalized with *inorm* before the heritability test. Environmental factors such as diet, antibiotic intake, diseases and milestones (crawling and walking) were registered for each set of triplets, however, those variables could not be tested for heritability because there were very few variabilities, that is, the babies presented very similar profiles within the sets and between the sets. The phenotype sex was modeled as covariate in the adjusted models. Graph comparing heritability estimates was generated in R package *ggplot2*.

4. Results

4.1. Triplets data sets and read quality

The collection of fecal samples of five sets of triplet infants, named A through E, aging from 1 to 37 months in a longitudinal manner was performed by Dr. Larissa Matos. All five sets of triplets lived in the city of São Paulo. We obtained 111 samples, but 12 were removed from the study because they were collected at the time of the infant antibiotics intake within the previous 30 days. Thus, 99 samples were considered valid for this study. Information about the zygosity of the triplets was obtained by their mother´s declaration. No DNA test was performed to confirm zygosity. Clinical data of the sets are described in Table 1. See also Figure 4.

Table 1: Clinical data of the study subjects

Triplets set[a]	Infant-zygosity	Sex	In Vitro Fertilization (IVF)	Gestation (weeks)	Birth Weight (g)	Mode of feeding	Neonatal ICU Admission (days)	Antibiotic Use in ICU	Life Months with Antibiotic Use	Valid Sample Months (total # of samples)
A	Mz1	F	Yes	33	1775	Formula	21	No	None	2,6,7,9,11,12, 13,36 (24)
A	Mz2	F	Yes	33	2060	Formula	21	No	None	
A	Dz	M	Yes	33	2200	Formula	20	No	37	
B	Mz1	M	Yes	28	815	EBM + Formula	89	Yes	None	6,9,11,12,13, 14,18,20,21,22, (30)
B	Mz2	M	Yes	28	750	EBM + Formula	96	Yes	None	
B	Dz	F	Yes	28	1100	EBM + Formula	80	Yes	None	
C	Mz1	M	Yes	32	1870	Formula	45	No	20	16,19,21,22 (12)
C	Mz2	M	Yes	32	1710	Formula	46	No	20	
C	Dz	M	Yes	32	1915	Formula	47	No	20	
D	Mz1	M	Yes	30	1235	EBM + Formula	81	Yes	16	9,11,13,18 (12)
D	Mz2	M	Yes	30	1430	EBM +Formula	70	Yes	16	
D	Dz	F	Yes	30	1135	EBM + Formula	70	Yes	16	
E	Mz1	F	No[b]	34	1895	EBM + Formula	0	No	None	1,2,6,7,9,11,13 (21)
E	Mz2	F	No[b]	34	1755	EBM + Formula	0	No	None	
E	Dz	M	No[b]	34	2175	EBM + Formula	0	No	12	

[a]All triplets were born from c-section; Mz1 = Monozygotic 1; Mz2 = Monozygotic 2; DZ = Dizygotic;
[b]Use of ovulatory stimulant; EBM = Expressed Breast Milk

Bacterial 16S amplicon sequencing was performed by Dr. Larissa Matos to obtain forward and reverse reads, however, the latter were falling below our threshold of quality (20) in more than 50% of the base pair positions, thus we only consider the forward reads for this study. They resulted in 22,101,869 reads and after pre-processing, we obtained 6,672,051 forward reads with a length of 230 bp with a minimum quality value of 20 for each base, resulting in an average of 61,394 reads per sample (Table 2).

Table 2: Description of processing of 16S sequences from fecal samples in each triplet set.

Triplet sets	Num. of samples	Total num. raw reads	Num. of ASVs	Total num. of processed reads	Mean reads per sample	Mean reads per ASVs	Highest num. reads in a sample	Lowest num. reads in a sample
A	27	3.982.550	1.364	1.251.106	46.337	917	140.443	13.109
B	30	5.813.571	1.343	1.758.163	58.605	1.309	93.279	21.489
C	15	3.713.794	1.178	964.244	64.283	819	114.698	46.084
D	15	3.442.790	992	1.012.249	67.483	1.020	130.203	44.995
E	24	5.149.164	1.285	1.686.289	70.262	1.312	137.098	36.550
Total	111[a]	22.101.869	4.168	6.672.051	NA	NA	NA	NA
Mean	22	4.420.374	1.232	1.334.410	61.394	1.075	123.144	32.445

[a]Twelve samples were removed from further analyses due antibiotics intake

4.2. ASV assignment and distribution

The reads were assigned to 4,168 Amplicon Sequence Variants (ASVs) with a length of 230 bp. The mean number of ASVs per set was 1,232 with a mean of 1,075 reads per ASVs. The ASVs were identified by number from 1 to 4,168. The most abundant ASV throughout the sets was the ASV-18 followed by ASV-1.

In order to ascertain that the samples reached enough sequencing depth, i.e., sampling captured all or most species from the population, we rarefied them at the level of ASVs. All samples in all sets reached a plateau suggesting that we have obtained enough species to be analyzed. We also observed a trend of increasing number of ASV as time goes by (**Figure 4)**

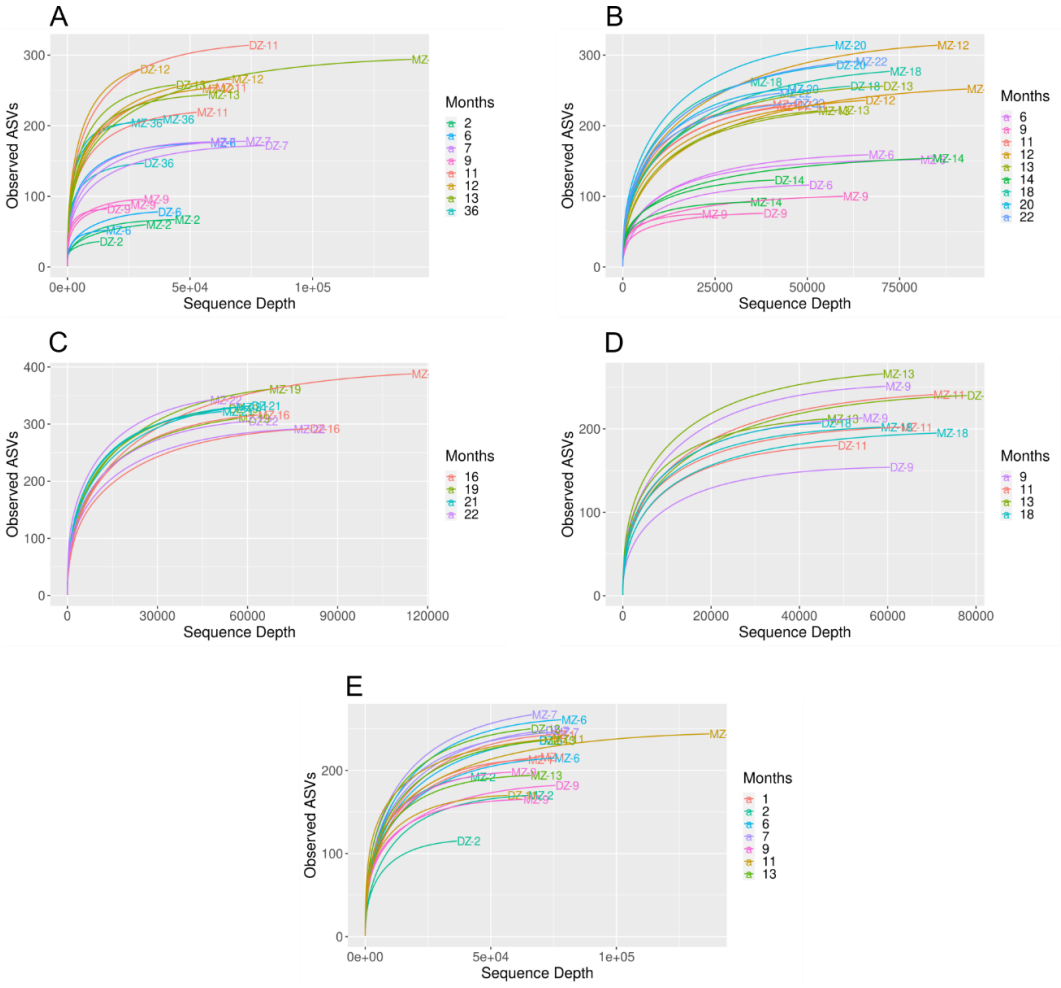

FIGURE 6: ALL SAMPLES IN ALL SETS REACHED A PLATEAU

A) SET A, (B) SET B, (C) SET C, (D) SET D AND E SET E. SAMPLES REACHED A PLATEAU INDICATING SUFFICIENT SEQUENCE DEPTH. EACH CURVE IS A SAMPLE THAT ARE COLOR-CODED BY TIME POINT AT THE FAR RIGHT OF EACH GRAPH. MZ1 = MONOZYGOTIC 1, MZ2 = MONOZYGOTIC 2, DZ = DIZYGOTIC, NUMBERS ARE THE TIME POINTS IN MONTHS

4.3 Diversity analysis

The ASV tables (each ASVs and their corresponding frequency) were normalized by rarefaction to 10,000 reads in all sets in order to account for the inter-sample variation in read number before diversity analysis was performed. Next, we calculated alpha and beta diversity with QIIME *core-metrics-phylogenetic* pipeline and R packages.

4.3.1 Alpha diversity

Although MZs appear to be more similar in almost all sets in Faith diversity analysis (**Figure 5),** the mean and median difference between MZs and DZ are not statistically significant as

shown by permutation tests ($p > 0.05$) based on 10,000 permuted results (Table 3). Likewise, observed ASVs and Shannon diversity are not significantly different between the triplets in all sets (**Table 3 and Figure 5**).

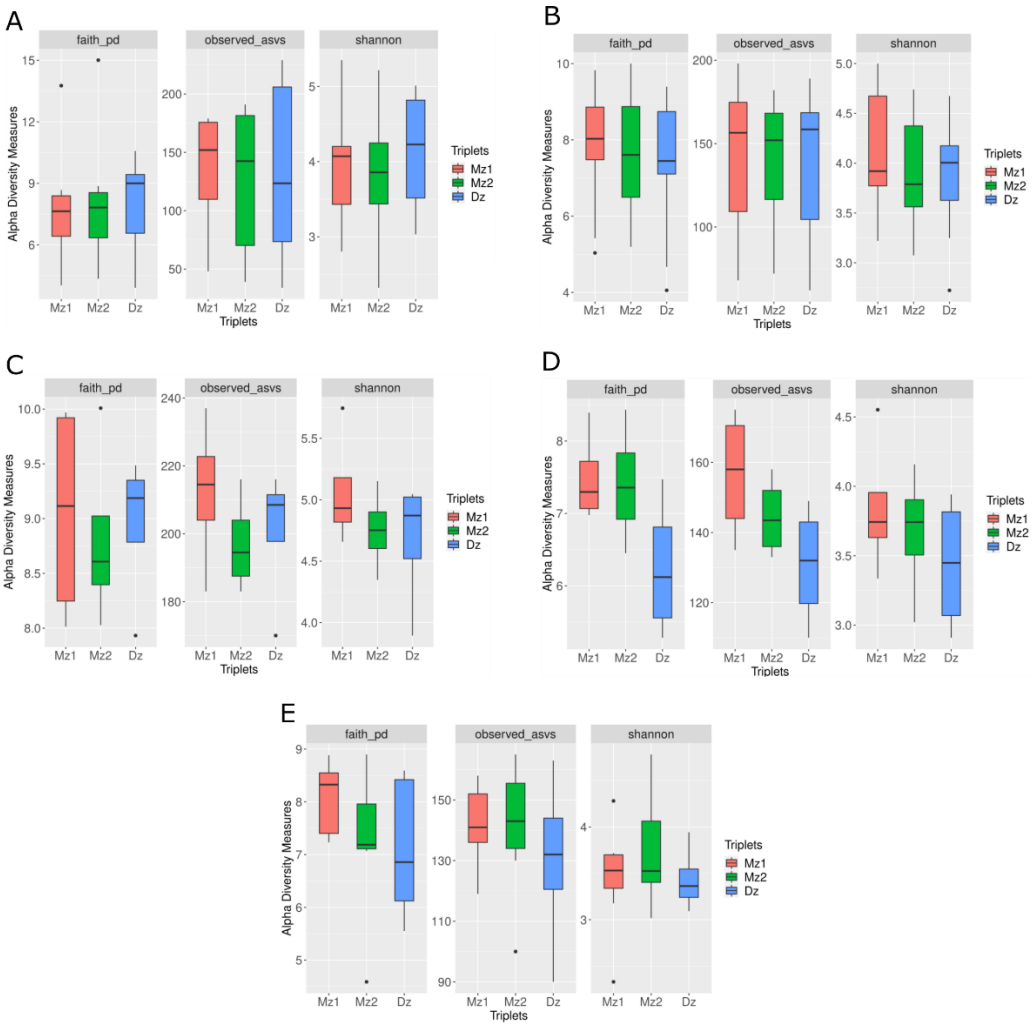

FIGURE 7: ALPHA DIVERSITY OF THE TRIPLETS ARE NOT SIGNIFICANTLY DIFFERENT

SETS A – E ARE REPRESENTED BY THE LETTERS ABOVE THE FIGURES. THE THICK MIDDLE LINE OF BOX PLOT REPRESENTS THE MEDIAN OF EACH TRIPLET IN ALL TIME POINTS.

Table 3: Permutation test of mean difference and Multinomial model for proportions

			m1.m2 means and medians difference[a]	*p*-value[b]
Permutation test (10,000) of mean difference	Faith	All Sets	mean(-0.1260)	0,2677
			median(0.2163)	0,6527
		By Set[c]	mean(0.3512)	0,9627
			median(0.3536)	0,9523
	Shannon	All Sets	mean(-0.0066)	0,4637
			median(-0.1993)	0,0735
		By Set[c]	mean(0.1202)	0,8808
			median(0.2097)	0,9658
	Observed ASVs	All Sets	mean(1.7424)	0,6703
			median(-5.5)	0,2832
		By Set[c]	mean(6.9760)	0,9544
			median(4.25)	0,8198
	Unweighted UniFrac	All Sets	mean(-0.0247)	0,0221*
			median(-0.0334)	0,0261*
		By Set[c]	mean(-0.0303)	0,0168*
			median(-0.0249)	0,0238*
	Weighted UniFrac	All Sets	mean(-0.0106)	0,3124
			median(-0.0287)	0,3107
		By Set[c]	mean(-0.0300)	0,0914
			median(0.0022)	0,5432
	Bray-Curtis	All Sets	mean(-0.0915)	0,0016**
			median(-0.0770)	0,0511
		By Set[c]	mean(-0.1179)	<0.001***
			median(-0.0844)	0,0108*
	Jaccard	All Sets	mean(-0.0367)	0,0009***
			median(-0.0397)	0,0003***
		By Set[c]	mean(-0.0419)	<0.0001***
			median(-0.0432)	0,0013***
Multinomial Model for Proportions	Shared ASVs	Set A	MZ1and MZ2 vs. MZ1 and DZ[d]	0.0444*
			MZ1and MZ2 vs. MZ2 and DZ[d]	0.0183*
		Set B	MZ1and MZ2 vs. MZ1 and DZ[d]	0.3889
			MZ1and MZ2 vs. MZ2 and DZ[d]	0.1496
		Set C	MZ1and MZ2 vs. MZ1 and DZ[d]	0.1860
			MZ1and MZ2 vs. MZ2 and DZ[d]	0.0283*
		Set D	MZ1and MZ2 vs. MZ1 and DZ[d]	0.1278
			MZ1and MZ2 vs. MZ2 and DZ[d]	0.0527
		Set E	MZ1and MZ2 vs. MZ1 and DZ[d]	0.3859
			MZ1and MZ2 vs. MZ2 and DZ[d]	0.1664

[a]Measured between monozygotic infant 1 and monozygotic infant 2
[b]statistical significance if *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$
[c]Takes into account the weight (size) of each set
[d]Comparison of proportions between the pairs

4.3.2 Beta diversity

For beta diversity analysis we applied four measures: Jaccard distance, Bray Curtis dissimilarities, Weighted and Unweighted Unifrac distance. Permutation tests on the mean and median difference of the distances between MZ1 and MZ2, MZ1 and DZ, and MZ2 and DZ in all sets

resulted in significantly higher similarities between the MZ1 and MZ2 in all measures when compared to the other two pairs, except for Weighted Unifrac distance where p values > 0.05 (Table 3). However, MZs clustered together on PCoA in this measure (**Figure 6)**. Bray Curtis PCoA also presented clusterization of MZs, except in set A (**Figure 7).** PCoA of Jaccard distance showed clusterization of MZs in at least two coordinates in all sets **(Figure 8**) . Unweighted Unifrac PCoA was the only measure that presented clusterization of MZs in all first coordinates in all sets (**Figure 9**), and MZs similarities were significantly greater than between MZ1 and DZ, and between MZ2 and DZ.

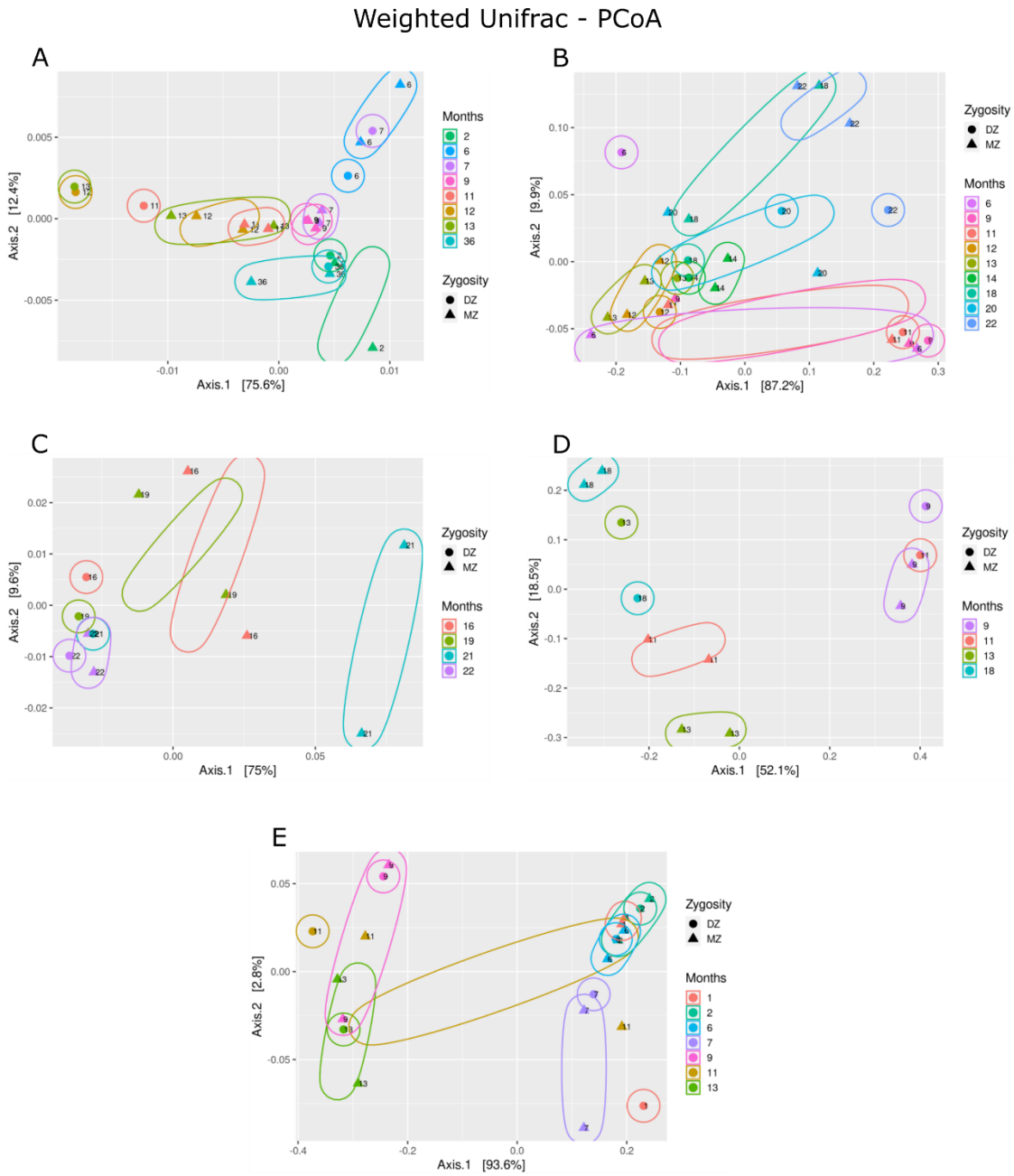

FIGURE 8: WEIGHTED UNIFRAC DISTANCE ON PCOA

SETS A-E ARE PRESENTED BY THE LETTERS ABOVE THE GRAPHS. FIRST PRINCIPAL COORDINATE WAS ENOUGH TO EXPLAIN MORE THAN 50% OF VARIANCE ON THE SAMPLES IN ALL SETS. MZS ARE CLUSTERED IN TIME POINTS 6, 7, 11, 12 AND 13 MONTHS (SET A), IN 6, 13 AND 18 (SET B), IN 21 MONTHS (SET C), IN 9, 11, 13 AND 18 MONTHS (SET D) AND IN 1 MONTHS IN SET E.

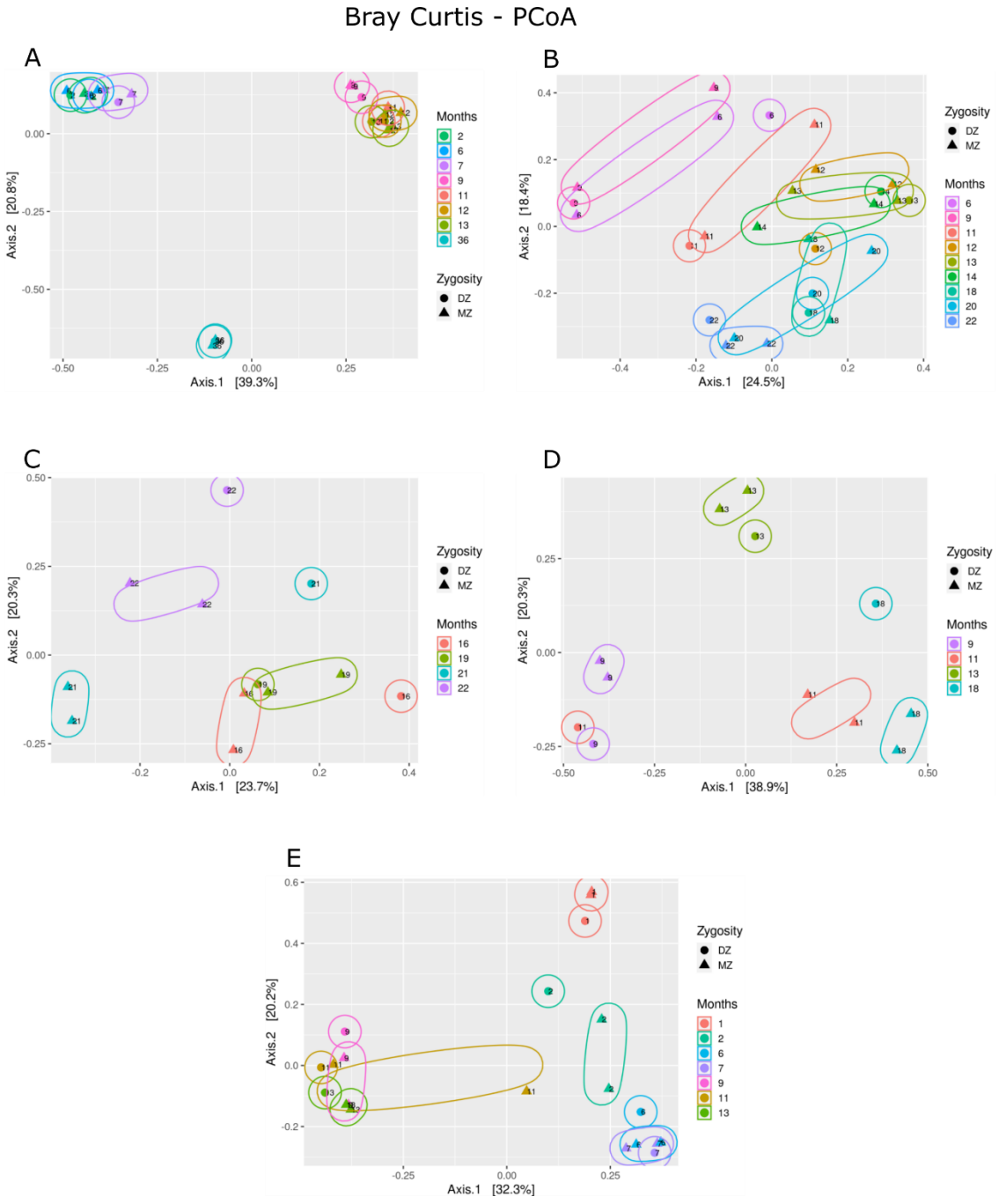

FIGURE 9: PCOA OF BRAY CURTIS DISSIMILARITY

SETS A-E ARE PRESENTED BY THE LETTERS ABOVE THE GRAPHS. MORE THAN 2 COORDINATES WERE NECESSARY TO EXPLAIN MORE THAN 50% OF VARIANCE IN ALL SAMPLES. THERE WAS NO CLUSTERIZATION OF MZS IN SET A, BUT WE NOTICED CLUSTERIZATION OF TIME POINTS (2 – 7 MONTHS, THEN 9 -13 MONTHS AND 36 MONTHS). MZS CLUSTERED AT 6 AND 12(SET B), AT 16, 21 AN 22 MONTHS (SET C), AT 9, 11, 13 AND 18 MONTHS (SET D), AT 2 AND 6 (SET E).

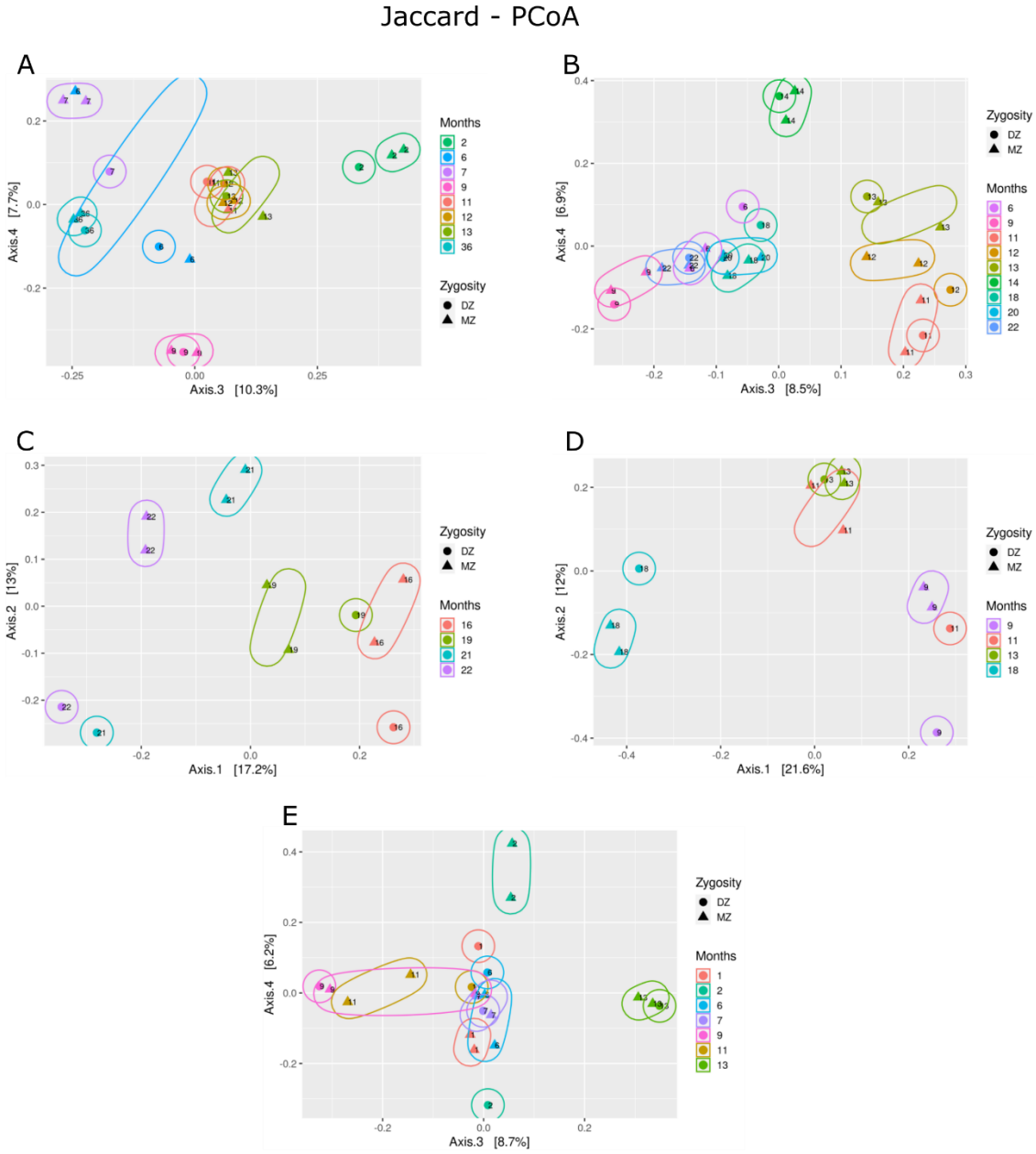

FIGURE 10: PCOA OF JACCARD DISTANCE.

SETS A, B AND E DID NOT PRESENT CLUSTERIZATION IN AXIS 1 OR 2 AND IT WAS NECESSARY MORE THAN 3 COORDINATES TO EXPLAIN 50% OF VARIANCE. IN SET A, MZS CLUSTERED AT TIME POINTS 2 AND 7. IN SET B, MZS CLUSTERED AT 6, 12, 18 MONTHS. IN SET C, MZS CLUSTERED IN ALL TIME POINTS. IN SET D, MZS CLUSTERED AT 9, 11 AND 18 MONTHS. IN SET E, MZS CLUSTERED AT 1 AND 2 MONTHS

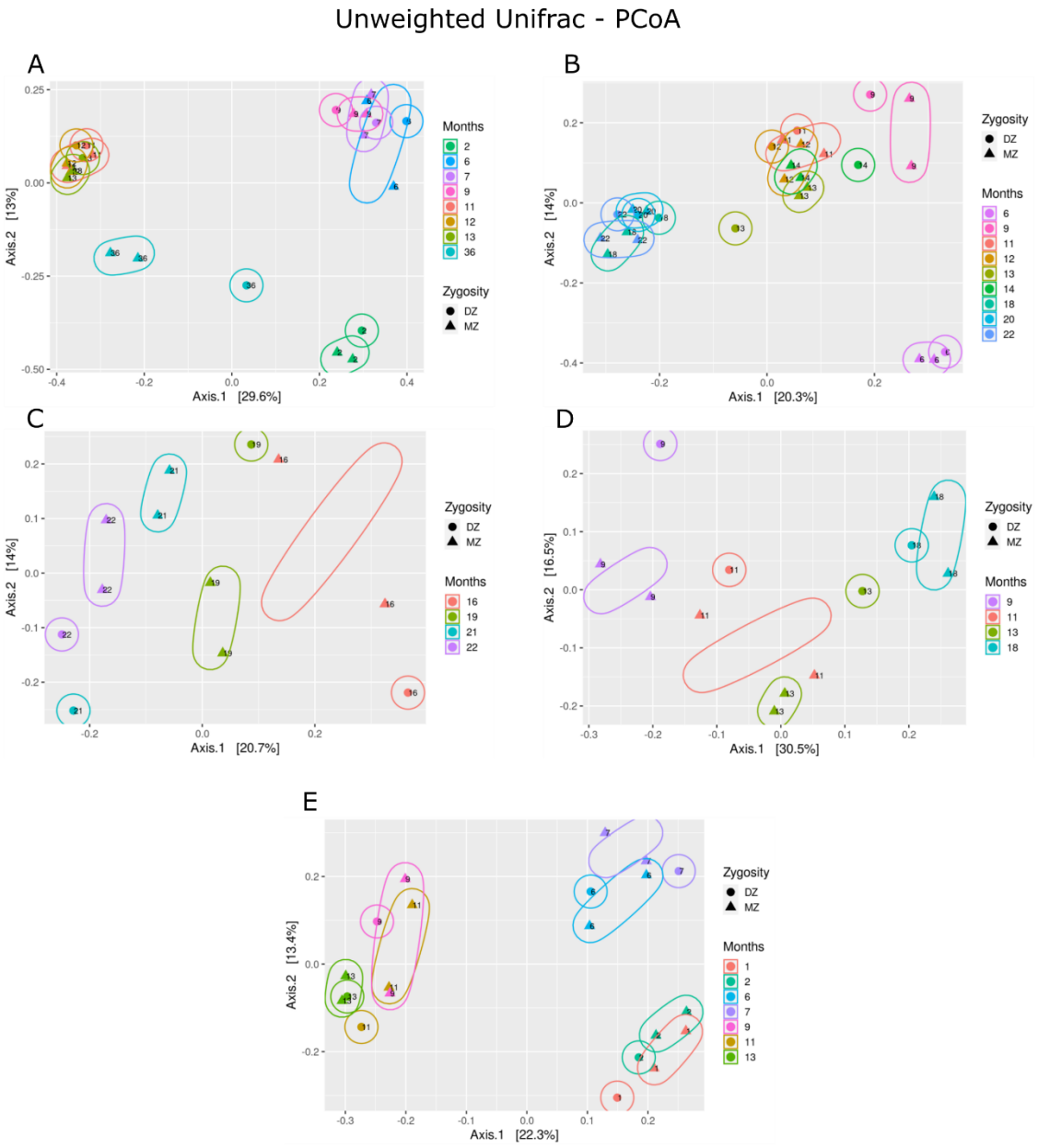

FIGURE 11: UNWEIGHTED UNIFRAC DISTANCE ON PCOA

IT WAS NECESSARY MORE THAN 2 PRINCIPAL COORDINATES TO EXPLAIN MORE THAN 50% VARIANCE OF THE SAMPLES, OVERALL. MZS CLUSTERED IN ALL SETS ON THE FIRST COORDINATE. IN SET A, MZS CLUSTERED AT TIME POINT 36. IN SET B, THEY CLUSTERED AT 9, 13, 14 AND 18 MONTHS. IN SET C, MZS CLUSTERED AT 19, 21 AND 22. IN SET D, THEY CLUSTERED AT 13. IN SET E, MZS CLUSTERED AT 1 AND 11 MONTHS.

4.4 Taxonomic classification

Taxonomic classification was performed in each set separately. In set A, 1,364 ASVs were classified into 11 phyla. The phylum Bacteroidetes and the genus Bacteroides had the highest abundance, while Fusobacteria and Hydrogenoanaerobacterium had the lowest abundance of all in the set. In set B, there were 1,343 ASVs classified into 9 phyla. The highest abundant phylum and genus were Firmicutes and Bacteroides, respectively. There were 1,178 ASVs in set C that were classified into 9 phyla. Firmicutes and Bacteroides had the highest abundance, whereas Fusobacteria and Clostridium XI had the lowest. Set D had 992 ASVs that were classified into 9 phyla. Proteobacteria and Escherichia/Shigella had the highest abundance whereas Fusobacteria and Novosphingobium had the lowest. Finally, in set E we obtained 1,285 that were classified into 12 phyla. Firmicutes and Bacteroides were the highest abundant phylum and genus, respectively. Planctomycetes and unclassified Oxalobacteraceae had the lowest abundance in this set (Table 4).

Table 4 Summary of ASV - classifications and absolute abundances

		Most abundant	Absolute abundance	Number of ASVs	Least abundant	Absolute abundance	Number of ASVs
A	Phylum	Bacteroides	457.776	181	Fusobacteria	19	3
	Genus	Bacteroides	343.631	98	Hydrogenoanaerobacterium	4	2
B	Phylum	Bacteroides	571.144	139	Enhydrobacter	4	1
	Genus	Firmicutes	614.935	703	Verrucomicrobia	133	2
C	Phylum	Bacteroides	288.147	149	Clostridium_XI	4	1
	Genus	Firmicutes	378.444	637	Fusobacteria	109	9
D	Phylum	Escherichia/Shigella	190.566	32	Novosphingobium	4	1
	Genus	Proteobacteria	287.662	211	Fusobacteria	4	1
E	Phylum	Bacteroides	457.626	107	unclassified_Oxalobacteraceae	4	1
	Genus	Firmicutes	557.435	693	Planctomycetes	12	1
Total number of ASVs throughout the sets A-E			4,168				
Total number of different classified Phyla			13				

Overall, Firmicutes and Bacteroidetes had the highest abundance in all time points together. Bacteroides were the most abundant genus and Bifidobacterium represented most of the phylum Actinobacteria (**Figure 10**). In a timely manner, Firmicutes did not vary substantially while Proteobacteria decreased and Bacteroidetes increased. At the genus level, it is more difficult to discern trends in relative abundance, but it does seem that the increase in abundance for the Bacteroidetes phylum is driven by increase in Bacteroides. In addition, Bifidobacterium, Clostridium XIVa and Veillonella seemed to be stable over time (**Figure 11**).

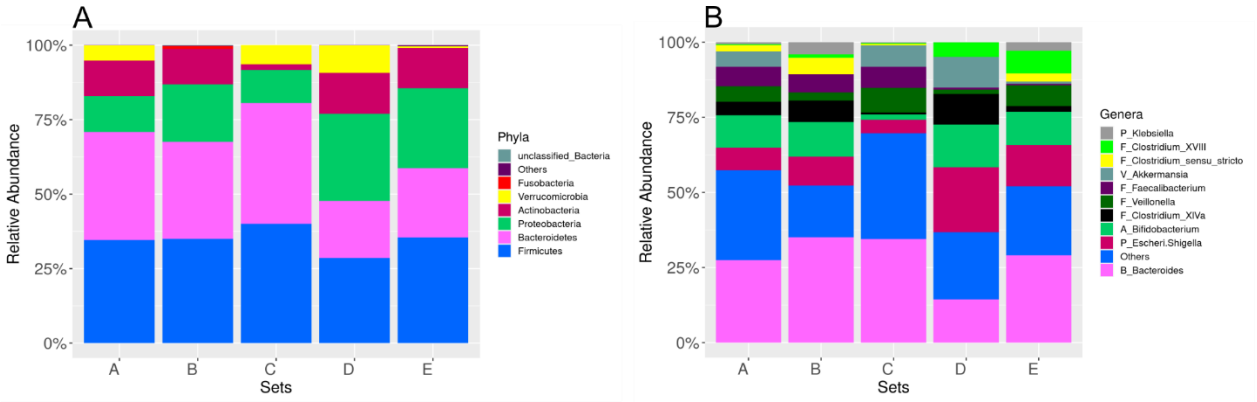

FIGURE 12: RELATIVE ABUNDANCE OF PHYLA (A) AND GENUS(B) WITH ALL TIME POINTS COMBINED PER SET

PHYLA AND GENERA ARE PLOTTED IN INCREASING ORDER OF RELATIVE ABUNDANCE FROM BOTTOM TO TOP. (A) ONLY THE SEVEN MOST ABUNDANT PHYLA ARE SHOWN AND THE REMAINING SIX ARE GROUPED IN THE "OTHERS" CATEGORY. THE ORDERS OF RELATIVE ABUNDANCE ARE SIMILAR IN ALL SETS WITH SLIGHT VARIATION IN THE PHYLA WITH LOW ABUNDANCES. (B) ONLY THE TEN MOST ABUNDANT GENERA ARE SHOWN AND THE REMAINING GENERA ARE GROUPED IN THE "OTHERS" CATEGORY. THE MOST ABUNDANT GENUS IS BACTEROIDES FOLLOWED BY OTHERS. VARIATION OF RELATIVE ABUNDANCE DISTRIBUTION IS HIGHER IN LOW ABUNDANT GENERA. EACH BAR REPRESENTS A SET WITH ALL TIME POINTS WHEN SAMPLES WERE COLLECTED IN THE SET.

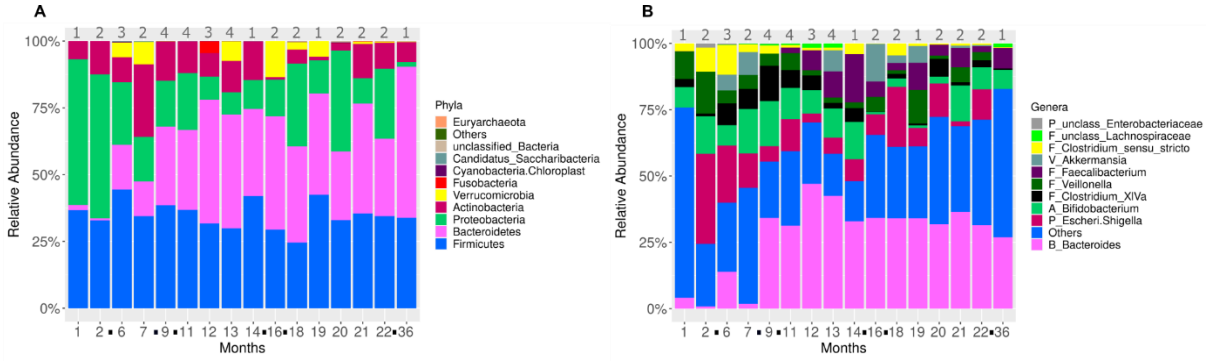

FIGURE 13: RELATIVE ABUNDANCE OF PHYLA(A) AND GENUS (B) WITH ALL SETS COMBINED PER TIME POINTS

PHYLA AND GENERA ARE PLOTTED IN INCREASING ORDER OF RELATIVE ABUNDANCE FROM BOTTOM TO TOP. (A) ONLY THE SIX MOST ABUNDANT PHYLA ARE SHOWN AND THE REMAINING SEVEN ARE GROUPED IN THE "OTHERS" CATEGORY. FIRMUTES IS THE MOST ABUNDANT PHYLA FOLLOWED BY PROTEOBACTERIA AND BACTEROIDES WHICH PRESENT INVERSE TREND OVER TIME. (B) ONLY THE TEN MOST ABUNDANT GENERA ARE SHOWN AND THE REMAINING GENERA ARE GROUPED IN THE "OTHERS" CATEGORY. THE MOST ABUNDANT GENERA IS IN THE OTHERS CATEGORY FOLLOWED BY BACTEROIDES WHOSE RELATIVE ABUNDANCE INCREASE WITH TIME. VARIATION OF RELATIVE ABUNDANCE DISTRIBUTION IS HIGHER IN LOW ABUNDANT GENERA. EACH BAR REPRESENTS A TIME POINT WHEN SAMPLES OF ALL SETS AND ALL TRIPLETS WERE COLLECTED. NUMBERS ON THE TOP OF EACH BAR REPRESENT THE QUANTITY OF SETS IN THE RESPECTIVE TIME POINT. TIME POINT IN MONTHS IS NOT CONTINUOUS. THE DOTS REPRESENT THE GAP BETWEEN TWO TIME POINTS.

4.5 ASVs Analysis

4.5.1 Shared ASVs

In order to evaluate ASVs shared by the triplets, we generated a Venn diagram in each set separately (**Figure 12**). In all sets, the estimated probabilities of MZ1 and MZ2 sharing the same ASVs were bigger than the ones sharing between MZ1 and DZ, and between MZ2 and DZ. However, only in set A these estimated probabilities were significantly bigger when analyzed by a multinomial model ($p = 0.044$ when compared to MZ1 and DZ, and $p = 0.018$ when compared to MZ2 and DZ). In addition, MZ1 and MZ2 significantly shared more ASVs when compared to MZ2 and DZ in set C ($p = 0.028$), and in set D there was a trend for significance ($p = 0.052$), but not when compared to MZ1 and DZ ($p = 0.1860$ in C and $p = 0.1278$ in D). In sets B and E, MZ1 and MZ2 did not significantly share more ASVs than MZ1 and DZ or MZ2 and DZ (Table 3). Most ASVs shared exclusively between MZs throughout the sets were classified as Bifidobacterium (60 ASVs – 7%), Bacteroides (52 ASVs – 6%) and Veillonella (50 ASVs – 6%) (Table 5).

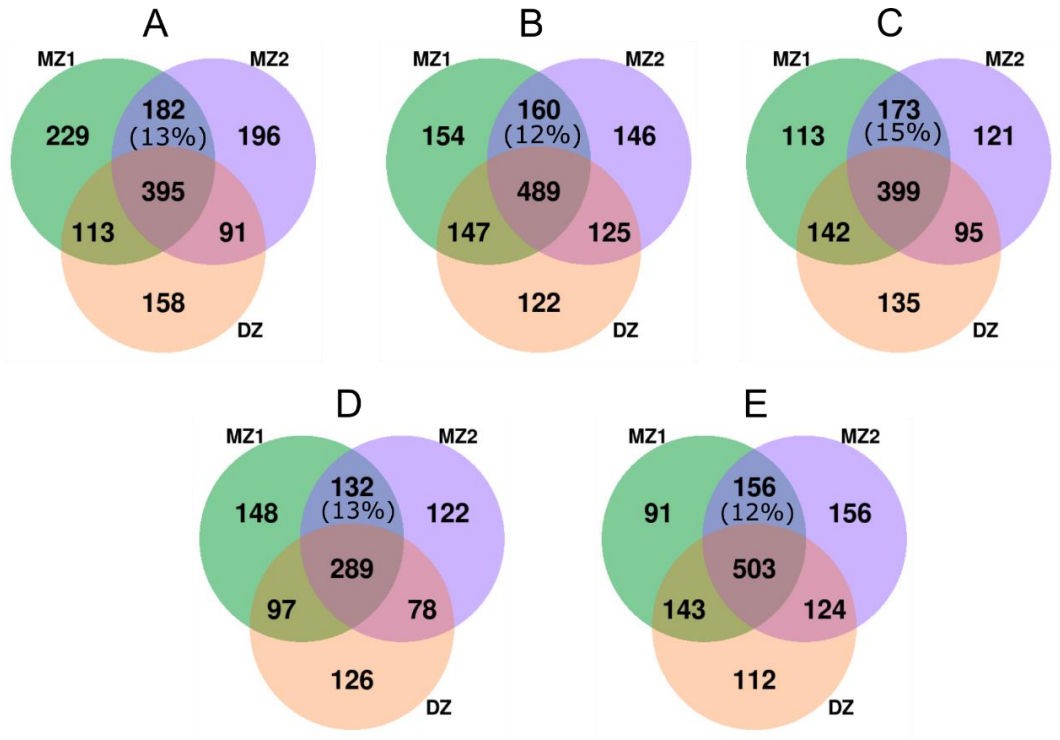

FIGURE 14: AMOUNT OF SHARED ASVS BETWEEN TRIPLETS IN EACH SET

SETS ARE REPRESENTED BY THE LETTERS A-E ABOVE THE GRAPHS. MZS APPEARED TO SHARED MORE ASVS BETWEEN THEM THAN BETWEEN THEIR DZ CO-TWINS IN ALL SETS. PERCENTAGE VALUES IS THE SHARED ASVS BETWEEN THE MZS OVER THE TOTAL NUMBER OF ASVS IN THE SET. MZ1 = MONOZYGOTIC TWIN 1, MZ2 = MONOZYGOTIC TWIN 2 AND DZ = DIZYGOTIC TWIN

Table 5 : Taxa whose ASVs are shared exclusively between the MZs throughout the sets

Taxa(Phylum_Genus)	Amount of ASVs in each set					Total of ASVs
	A	B	C	D	E	
Actinobacteria_Bifidobacterium	10	6	2	29	13	60
Bacteroidetes_Bacteroides	6	17	14	7	8	52
Firmicutes_Veillonella	10	18	5	8	9	50
Proteobacteria_unclassified_Enterobacter	5	4	9	10	16	44
Firmicutes_unclassified_Lachnospiraceae	15	14	4	5	4	42
Bacteroidetes_Prevotella	1	2	36	0	1	40
Firmicutes_unclassified_Ruminococcacea	13	3	11	2	7	36
Firmicutes_Faecalibacterium	22	1	7	6	0	36
Proteobacteria_Haemophilus	6	17	5	0	2	30
Firmicutes_Clostridium_sensu_stricto	2	6	12	3	6	29
Firmicutes_Clostridium_XlVa	6	5	3	2	3	19
Firmicutes_Streptococcus	1	4	3	2	7	17
Proteobacteria_Klebsiella	2	12	2	1	0	17
Firmicutes_Blautia	4	1	3	3	4	15
Firmicutes_unclassified_Peptostreptococc	3	4	0	1	7	15
Firmicutes_Dialister	0	0	13	0	0	13
Firmicutes_unclassified_Clostridiales	4	1	5	2	0	12
Actinobacteria_Actinomyces	2	1	0	4	4	11
Verrucomicrobia_Akkermansia	3	0	1	2	5	11
Firmicutes_Clostridium_IV	4	4	0	1	1	10
Firmicutes_Clostridium_XVIII	0	2	2	2	4	10
Proteobacteria_Escherichia/Shigella	3	1	1	2	2	9
Firmicutes_Ruminococcus2	4	1	3	0	1	9
Firmicutes_unclassified_Clostridiaceae_1	0	7	1	0	1	9
Firmicutes_Flavonifractor	3	2	1	2	0	8
Firmicutes_Roseburia	2	2	0	2	2	8
Firmicutes_Erysipelotrichaceae_incertae_	2	0	1	3	2	8
Firmicutes_Lachnospiracea_incertae_sed	6	1	0	0	0	7
Firmicutes_Gemmiger	1	1	1	2	1	6
Actinobacteria_Eggerthella	0	0	2	1	3	6
Firmicutes_Enterococcus	1	1	0	3	1	6
unclassified_Bacteria_unclassified_Bacter	5	1	0	0	0	6
Firmicutes_Clostridium_XI	1	2	0	1	2	6
Firmicutes_Oscillibacter	2	0	0	2	2	6
Firmicutes_Ruminococcus	1	0	0	1	4	6
Bacteroidetes_Dysgonomonas	5	0	0	0	1	6
Firmicutes_Butyricicoccus	1	0	4	0	0	5
Fusobacteria_Fusobacterium	0	1	1	0	2	4
Firmicutes_Granulicatella	1	1	0	1	1	4
Actinobacteria_unclassified_Bifidobacteria	2	0	0	0	2	4
Bacteroidetes_Parabacteroides	1	1	0	0	2	4
Firmicutes_Abiotrophia	0	1	2	0	1	4
Candidatus_Saccharibacteria_	0	0	1	0	3	4
Firmicutes_Anaerostipes	2	0	1	1	0	4
Bacteroidetes_Alistipes	2	0	1	1	0	4
Actinobacteria_Collinsella	0	0	0	0	4	4
Firmicutes_Lactobacillus	2	0	0	0	2	4
Proteobacteria_Proteus	1	0	0	2	0	3
Actinobacteria_Olsenella	3	0	0	0	0	3
Firmicutes_Megasphaera	1	0	0	1	1	3
Cyanobacteria/Chloroplast_Streptophyta	0	2	0	1	0	3
Actinobacteria_unclassified_Actinomycetal	1	1	0	0	1	3
Proteobacteria_Providencia	0	0	0	3	0	3
Firmicutes_unclassified_Peptoniphilaceae	0	0	0	0	3	3
Firmicutes_Turicibacter	0	1	0	0	1	2
Firmicutes_Lachnospira	0	2	0	0	0	2
Firmicutes_Phascolarctobacterium	0	1	0	1	0	2
Firmicutes_unclassified_Clostridia	1	1	0	0	0	2
Firmicutes_Romboutsia	0	1	1	0	0	2
Proteobacteria_Neisseria	0	1	1	0	0	2
Proteobacteria_Mesorhizobium	0	0	2	0	0	2
Proteobacteria_Cupriavidus	0	0	2	0	0	2
Proteobacteria_Rhizobium	0	0	1	1	0	2
Bacteroidetes_unclassified_Bacteroidales	1	0	1	0	0	2
Firmicutes_Gemella	1	0	1	0	0	2
Firmicutes_Anaerotruncus	0	0	0	2	0	2
Proteobacteria_Acinetobacter	0	0	0	1	1	2
Actinobacteria_Corynebacterium	0	0	0	1	1	2
Firmicutes_Paenibacillus	0	0	0	0	2	2
Proteobacteria_Citrobacter	2	0	0	0	0	2
Firmicutes_Dorea	0	0	0	0	1	1
Actinobacteria_Propionibacterium	0	0	1	0	0	1
Actinobacteria_Varibaculum	0	0	0	0	1	1
Firmicutes_Peptostreptococcus	0	1	0	0	0	1
Proteobacteria_unclassified_Pasteurellaci	0	1	0	0	0	1
Fusobacteria_Leptotrichia	0	1	0	0	0	1
Actinobacteria_Atopobium	0	1	0	0	0	1
Proteobacteria_unclassified_Neisseriacea	0	1	0	0	0	1
Proteobacteria_Caulobacter	0	0	1	0	0	1
Actinobacteria_Leifsonia	0	0	1	0	0	1
Bacteroidetes_Heliimonas	0	0	1	0	0	1
Proteobacteria_Cardiobacterium	0	0	1	0	0	1
Firmicutes_Oribacterium	0	0	1	0	0	1
Firmicutes_unclassified_Firmicutes	0	0	1	0	0	1
Proteobacteria_Buttiauxella	0	0	0	1	0	1
Proteobacteria_Delftia	0	0	0	1	0	1
Proteobacteria_Salmonella	0	0	0	1	0	1
Firmicutes_unclassified_Lactobacillaceae	0	0	0	1	0	1
Firmicutes_Coprobacillus	0	0	0	1	0	1
Proteobacteria_Cronobacter	0	0	0	1	0	1
Proteobacteria_Morganella	0	0	0	1	0	1
Firmicutes_Staphylococcus	0	0	0	0	1	1
Proteobacteria_Stenotrophomonas	0	0	0	0	1	1
Proteobacteria_Bosea	0	0	0	0	1	1
Proteobacteria_Methylobacterium	0	0	0	0	1	1
Proteobacteria_Sphingomonas	1	0	0	0	0	1
Proteobacteria_Serratia	1	0	0	0	0	1
Bacteroidetes_Barnesiella	1	0	0	0	0	1
Euryarchaeota_Methanobrevibacter	1	0	0	0	0	1
Firmicutes_Intestinimonas	1	0	0	0	0	1
Total	182	160	173	132	156	803

4.5.2 Correspondence Analysis (CA)

We performed CA in the 50 most abundant ASVs in each set and observed that samples of the same time points tended to group together on the graph, specifically in set A and E in which there was clusterization of time points 9, 11 and 13 months (**Figure 13**). Then, CA was performed again on those time points only, however, we added ASVs that also showed significant heritability regardless of their abundance (results of heritability test are in the next section) making it 53 ASVs in set A, 59 in set B, 57 in set D and 53 ASVs in set E (**Figure 14**). Set C was not included in this analysis due to the lack of samples on the three time points, however, there was clusterization of MZs in 2 (16 and 21 months) out of 4 time points in this set (**Figure 13C**). Chi square test in all CAs returned $p < 0.05$. To account for the large sample size of the contingency tables, we calculated the coefficients of contingency and they were all greater than 0.99 in all sets showing that there are strong relationships between the samples and the ASVs. We also observed twenty-two ASVs (1, 3, 8, 18, 19, 21, 31, 43, 44, 60, 63, 64, 80, 107, 116, 130, 150, 204, 213, 714, 758, 860) common to all sets associated with the samples from those time points (**Figure 15**). In order to interpret the contribution of sample-ASVs for the significance of the association, we calculated the observed values divided by the expected values obtained from the Chi square test, i.e., the log2 of the likelihood ratio (for simplicity, we will call it "ratio" in this document) We compared the ratios of relative abundance of ASVs in all sets combined (total of 111 ASVs) between the MZ1, MZ2 and DZ from all three time points, making it three pairs of comparisons. We observed that the ratios of 34 out of 111 ASVs were more similar between MZ1 and MZ2, and between MZ1 and DZ, whereas MZ2 and DZ had more similar ratios in 20 ASVs. Ratios of 23 ASVs were equally similar between two or the three pairs compared. Eleven ASVs (3, 8, 18, 19, 31, 43, 44, 63, 64, 80 and 714) were present in all samples of each set. Ratios of ASVs 43 and 63 were more similar between the MZs, ASVs 8, 18, 44 and 80 showed more similar ratios between MZ1 and DZ, while ASVs 3 and 714 ratios were more similar between MZ2 and DZ. Between MZs and between MZ2 and DZ ASV 19 presented similar ratios, and ASV 31 were equally similar in all three pairs compared.

FIGURE 15: CORRESPONDENCE ANALYSIS IN EACH SET SHOWS CLUSTERIZATION OF TIME POINTS

SETS ARE REPRESENTED BY THE LETTERS ABOVE THE GRAPHS. SAMPLES OF THE SAME TIME POINTS TEND TO CLUSTER. TIME POINTS 9, 11 AND 13 IN BOTH SETS A AND E CLUSTERED IN THE LOW QUADRANTS. MZS CLUSTERED IN ALL TIME POINTS IN SET D. SAMPLES ARE IN BLUE AND ASV, HERE AS S#, ARE COLORED BY CONTRIBUTION TO THE COORDINATES. ASVS IN WINE COLOR ARE OUTLIERS. M1: MZ1 (MONOZYGOTIC TWIN 1); M2:MZ2 (MONOZYGOTIC TWIN 2) AND D:DZ (DIZYGOTIC TWIN). P VALUES OF THE CHI-SQUARE ARE ABOVE THE GRAPH.

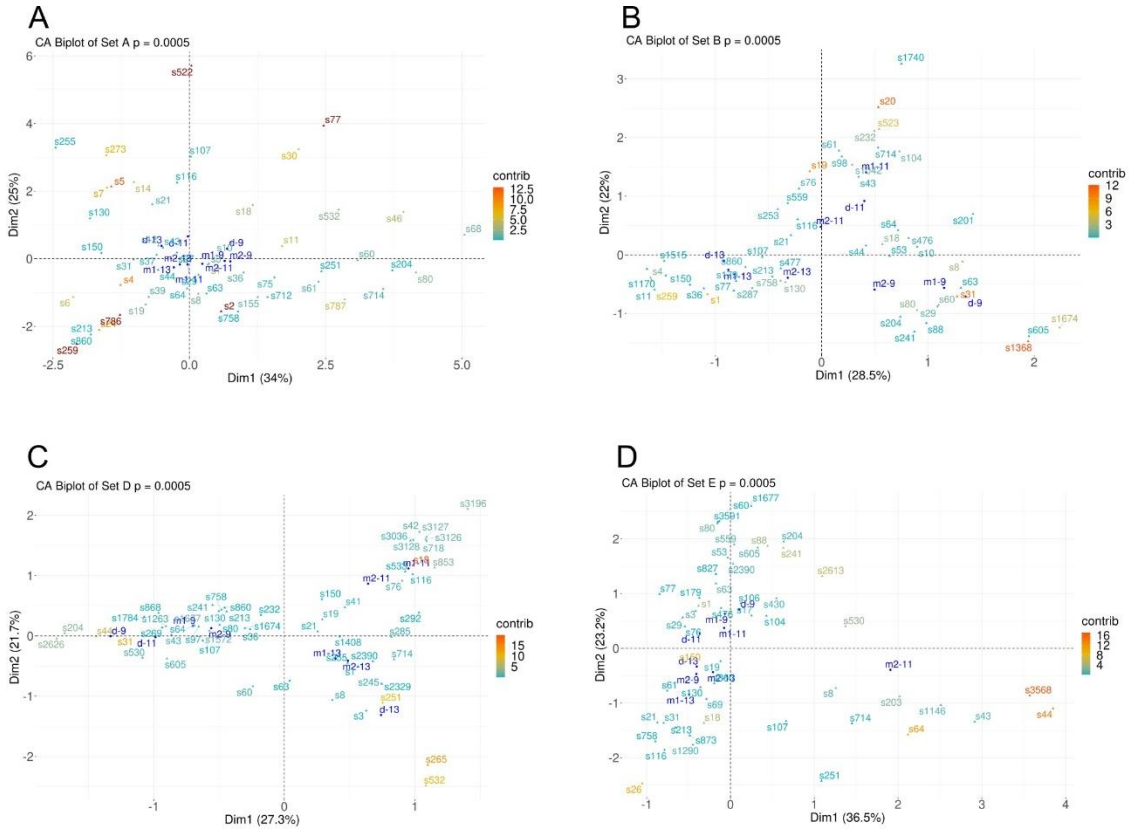

FIGURE 16: CORRESPONDENCE ANALYSIS IN EACH SET AT TIME POINTS 9, 11 AND 13 MONTHS

SET A IS PRESENTED BY A, SET B IS REPRESENTED BY B, SET D IS REPRESENTED BY C AND SET E IS REPRESENTED BY D. SAMPLES ARE IN BLUE AND ASVS (HERE CODED WITH S#) IN A GRADIENT COLOR REPRESENTING THEIR CONTRIBUTION TO THE PRINCIPAL COMPONENTS OF THE SAMPLES. SAMPLES GROUPED TOGETHER ARE MORE SIMILAR TO ONE ANOTHER THAN SAMPLES FURTHER AWAY. SAMPLES ON THE OPPOSITE QUADRANT ARE MORE DIFFERENT THAN THE ONES IN THE SAME QUADRANT. ASSOCIATIONS BETWEEN ASVS AND THE SAMPLES CAN BE EVALUATED BY THE ANGLE BETWEEN THEM FORMED FROM THE ORIGIN AND THE DISTANCE OF THE ASVS FROM THE ORIGIN. THE FURTHER FROM THE ORIGIN AND THE SMALLER THE ANGLE, THE STRONGER THE ASSOCIATION. IN SET A (A), MZS ARE ALL IN THE LOWER QUADRANTS AND DZS ARE IN THE UPPER QUADRANTS. IN SET B (B), SAMPLES ARE SEPARATED BY TIME POINTS AND MZS ARE NOT CLUSTERED; EACH TIME POINT IS IN DIFFERENT QUADRANTS. IN SET D(C), ALL MZS ARE MORE SIMILAR WHEN COMPARED TO THEIR DZ CO-TWINS. IN SET E(D), THERE WAS NO CLUSTERIZATION OF MZS. SET C CA IS IN FIGURE S3. M1: MZ1; M2:MZ2 AND D: DZ

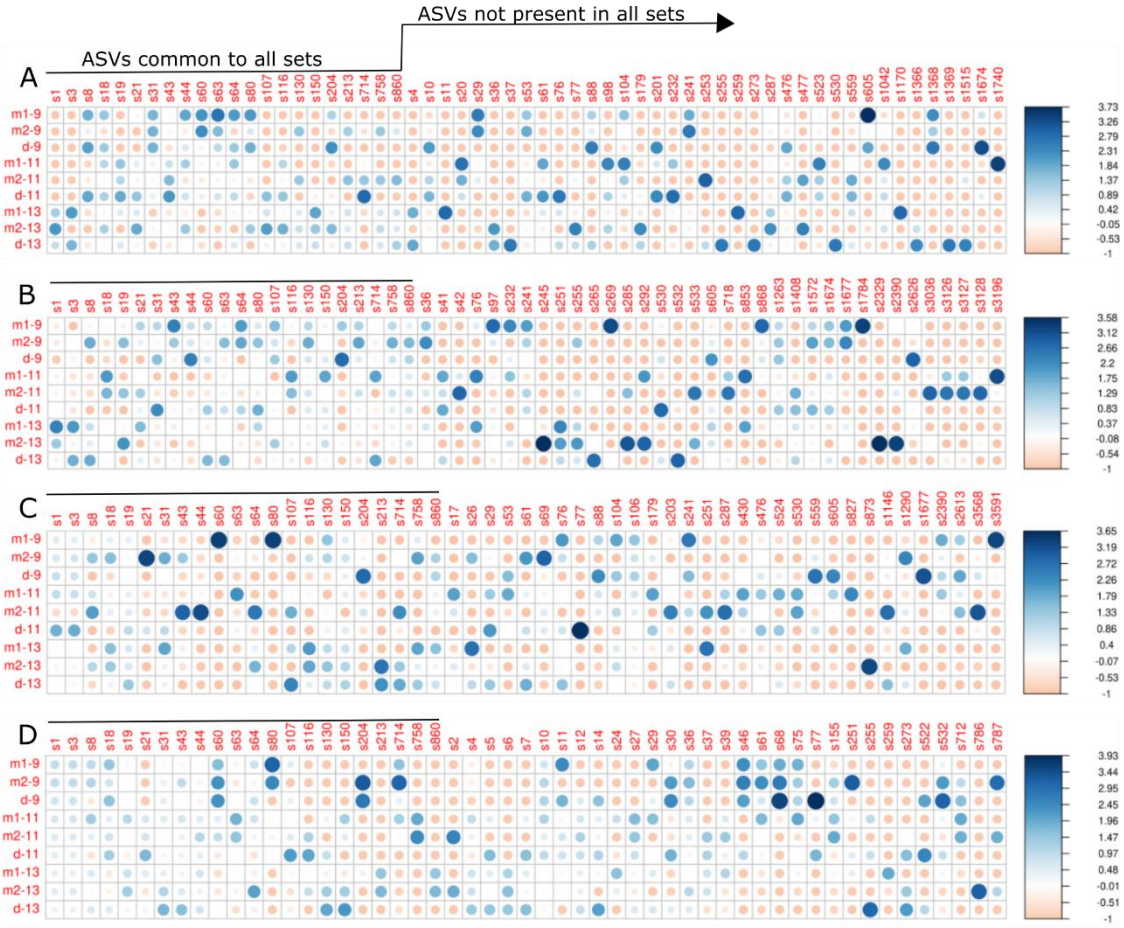

FIGURE 17: CORRELATION PLOT OF ASSOCIATIONS BETWEEN SAMPLES AND ASVS.

ASVS ARE IDENTIFIED AS S#. A) SET A; B) SET B; C) SET D AND D) SET E. M1 IS THE MONOZYGOTIC TWIN 1, M2 IS THE MONOZYGOTIC TWIN 2, AND D IS THE DIZYGOTIC TWIN AT TIME POINTS 9, 11 AND 13. ASSOCIATIONS ARE GRADED BY COLORS. BLUE COLOR REPRESENTS ASSOCIATION OF ATTRACTION, RED COLOR REPRESENTS ASSOCIATION OF REPULSION. NUMBERS NEXT TO THE COLOR GRADIENT BAR ARE THE VALUES OF LOG2 OF LIKELIHOOD RATIOS (OBSERVED VALUE/EXPECTED VALUE) AS INDEX OF ASSOCIATION. VALUES NEXT TO ZERO INDICATE NO ASSOCIATION, POSITIVE VALUES INDICATE ASSOCIATION OF ATTRACTION AND NEGATIVE VALUES INDICATE ASSOCIATION OF REPULSION. BLACK BAR ON THE TOP OF EACH GRAPH INDICATES THE TWENTY-TWO ASVS THAT ARE COMMON TO ALL FOUR SETS. ASVS THAT ARE NOT PRESENT IN ALL SETS ARE INDICATED BY THE BLACK ARROW.

4.5.3 Heritability

We performed heritability tests of ASVs abundances on sets A, B, D and E in which samples of time points 9, 11 and 13 were present, and these are the only time points where we had a large enough sample size to obtain convergence in the analysis performed (n = 4: sets A, B, D, and E). There was no collection of samples in those time points in set C, so we did not consider this set for this test. Criteria of filtering ASVs is explained in the Methods section. After filtering, we obtained 20, 28 and 32 ASVs at time points 9, 11 and 13 months, respectively. Heritability test

with the phenotype sex modeled as covariate was applied at each time point separately, resulting in nineteen distinct ASVs with significant ($p<0.05$) heritability, five at time point 9, seven ASVs at time point 11 and ten ASVs at time point 13 months (**Figure 16 and Table 6**). ASV-1 presented significant heritability at all three time points and ASV-63 at time points 9 and 11. The average value of heritability throughout the time points was 75%. The highest value was 90% for ASV-31 at 11 months and ASV-1 at 13 months, and the lowest value was 57% for ASV-1 at time point 13.

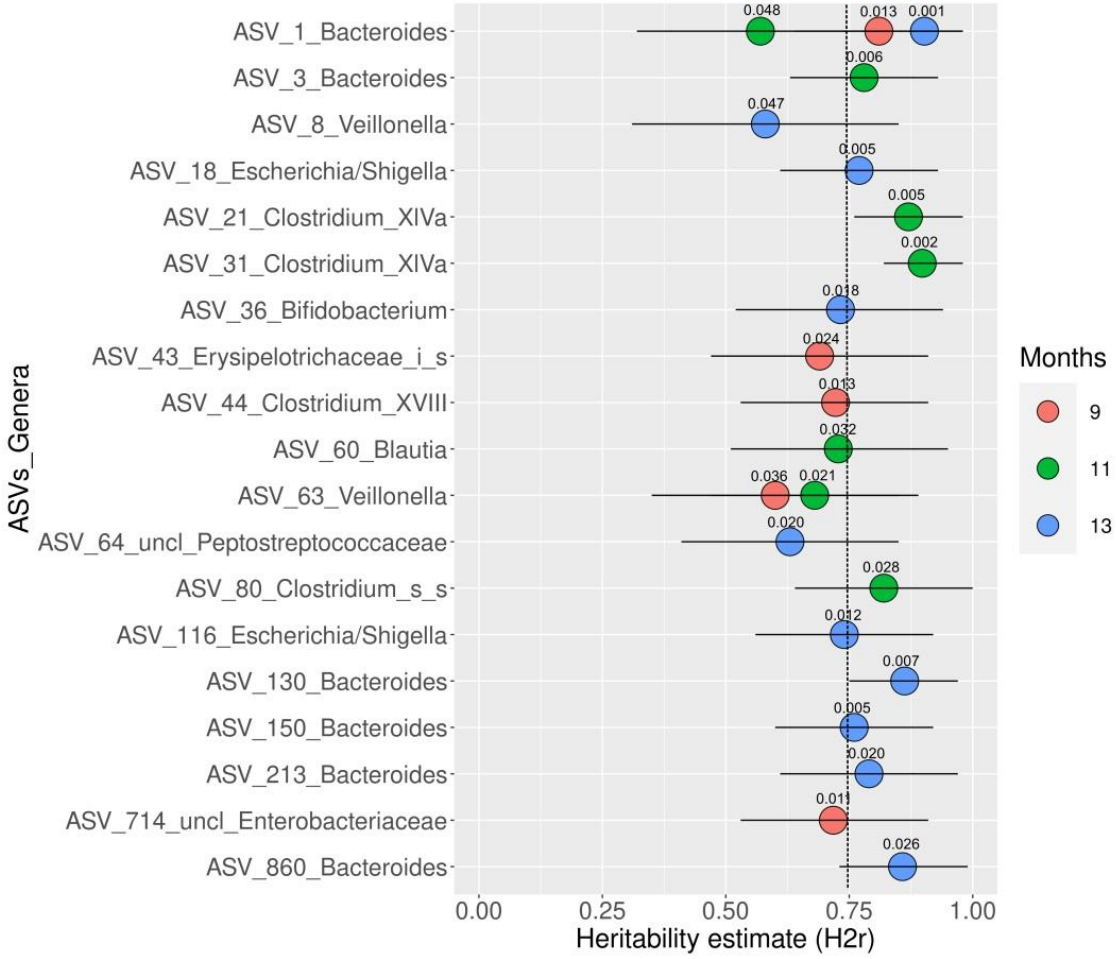

FIGURE 18: HERITABILITY ESTIMATES (H2R) OF ASVS RELATIVE ABUNDANCE AT TIME POINTS 9, 11 AND 13 MONTHS

ASVS OF BACTEROIDES AND VEILLONELLA HAVE SIGNIFICANT HERITABILITIES AT TIME POINTS 9, 11 AND 13 MONTHS. DOTS REPRESENT THE HERITABILITY ESTIMATES AND BARS ARE THE STANDARD ERRORS OF THE ESTIMATES. SHOWN HERE ARE ONLY ASVS WHOSE HERITABILITY WAS STATISTICALLY SIGNIFICANT ($P \leq 0.05$). NUMBERS ABOVE THE DOTS ARE THE P-VALUES OF THE TEST. VERTICAL DOTTED LINE IS THE MEAN OF H2R OF ALL ASVS REPRESENTED HERE.

Table 6: Heritability test on ASVs relative abundance by SOLAR-Eclipse

Time Point	ASVs	Heritability (Hr2)	Taxa	p-value	Sex covariate (beta values)	std. error of Hr2 estimate	#N (families)[a]
9	ASV_1	0,81	Bacteroidetes_Bacteroides	0,0134	0,432	0,167	3
	ASV_44	0,72	Firmicutes_Clostridium_XVIII	0,0133	0,301	0,195	4
	ASV_714	0,72	Proteobacteria_unclassified_Enterobacteriaceae	0,0107	0,000	0,189	4
	ASV_43	0,69	Firmicutes_Erysipelotrichaceae_incertae_sedis	0,0243	0,023	0,225	4
	ASV_63	0,60	Firmicutes_Veillonella	0,0357	0,004	0,251	4
11	ASV-31	0,90	Firmicutes_Clostridium_XlVa	0,0025	0,083	0,082	4
	ASV-21	0,87	Firmicutes_Clostridium_XlVa	0,0048	0,051	0,113	4
	ASV-80	0,82	Firmicutes_Clostridium_sensu_stricto	0,0281	0,034	0,181	3
	ASV-3	0,78	Bacteroidetes_Bacteroides	0,0060	NE[b]	0,150	4
	ASV-60	0,73	Firmicutes_Blautia	0,0320	0,400	0,217	3
	ASV-63	0,68	Firmicutes_Veillonella	0,0209	0,107	0,208	4
	ASV-1	0,57	Bacteroidetes_Bacteroides	0,0476	0,040	0,251	4
13	ASV-1	0,90	Bacteroidetes_Bacteroides	0,0007	0,003	0,079	4
	ASV-130	0,86	Bacteroidetes_Bacteroides	0,0073	0,000	0,115	4
	ASV-860	0,86	Bacteroidetes_Bacteroides	0,0265	0,003	0,133	4
	ASV-213	0,79	Bacteroidetes_Bacteroides	0,0201	NE[b]	0,182	3
	ASV-18	0,77	Proteobacteria_Escherichia/Shigella	0,0053	0,018	0,156	4
	ASV-150	0,76	Bacteroidetes_Bacteroides	0,0049	0,368	0,158	4
	ASV-116	0,74	Proteobacteria_Escherichia/Shigella	0,0120	0,025	0,180	4
	ASV-36	0,73	Actinobacteria_Bifidobacterium	0,0184	0,001	0,206	4
	ASV-64	0,63	Firmicutes_unclassified_Peptostreptococcaceae	0,0200	0,147	0,215	4
	ASV-8	0,58	Firmicutes_Veillonella	0,0466	0,014	0,268	4

[a]Number of families sets analysed
[b]NE: not estimated due to instability

4.6 Shared ASVs, CA and Heritability results comparison

By comparing the results of the tests Shared ASVs, CA and Heritability, we found that ASV-1, which was classified as Bacteroides, was shared between the MZs, had a more similar ratio in CA between the MZs when compared to the ratios of their co-twin DZ and presented highly significant heritability at time point 9 months. At time point 11 the ASV-21, ASV-31, ASV-60, ASV-63 and ASV-80 (classified as Clostridium_XlVa, Clostridium_XlVa, Blautia, Veillonella, Clostridium_sensu_stricto, respectively) were shared between the MZs, presented similar ratios in MZs and had significant heritabilities. ASV-1, ASV-18, ASV-860 (Bacteroides, Proteobacteria_Escherichia/Shigella, Bacteroides, respectively) at time point 13 were shared between the MZs, showed to have more similar ratios between the MZs and also presented significant heritabilities (**Figure 17**).

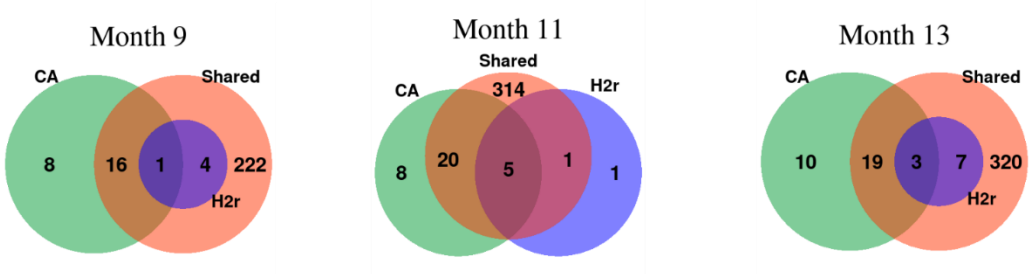

FIGURE 19: AMOUNT OF ASVS THAT RESULTED MZS MORE SIMILAR IN THREE TESTS.

COMPARISON OF THE RESULTS OBTAINED IN CORRESPONDENCE ANALYSIS (CA) WHERE MZS WERE MORE SIMILAR, NON EXCLUSIVELY SHARED ASVS BETWEEN THE MZS (SHARED) AND HERITABILITY (H2R) PER TIME POINTS. 251, 349, 359 ASVS WERE ANALYZED AT TIME POINTS 9, 11 AND 13, RESPECTIVELY.

5. Discussion

In this study, we investigated the microbiome of five sets of triplets and compared the composition of microbiota between triplets within a set. This triplet model, which is composed of two monozygotic and one dizygotic infant, allows us to investigate genetic factors while environmental factors variation is reduced to the minimum, i.e., the environment to which the triplets are exposed is quite similar. We hypothesized that genetics plays a part in developing and establishing the microbiome composition, thus we predicted that monozygotic twins would present a more similar composition between them than between their dizygotic co-twins. To this end, we analyzed bacterial sequences of the V3-V4 region of 16S small subunit ribosomal RNA gene extracted from the triplets faeces.

It is important to note the limitations of this study before the discussion of the results. The two major limitations are: (1)small sample size; although the number of multiple gestations is increasing in western populations due to later maternity and in vitro fertilization techniques, the number was still approximately 0.0005% of all births in Brazil between 2015-2019(DATASUS). For this reason, finding volunteers to participate in this study is one of the major limitations. (2)16S rRNA gene sequences; amplicons from 16S rRNA gene have widely been used to study bacterial and archaea diversity due to their evolutionary highly conservative character in the bacteria and archaea kingdom and regions different enough to distinguish between bacteria and archaea. However, resolution decreases in lower taxonomic levels such as species and strains(Poretsky, Rodriguez-R, Luo, Tsementzi, & Konstantinidis, 2014; Turnbaugh et al., 2010). Thus, we will carefully interpret our results here, even though sequence analysis was done based on ASVs, which are more precise than OTUs (Callahan et al., 2017). The results of this study should be used to observe and perhaps infer certain population trends based on these five particular sets investigated here.

The quality of the forward reads generated by sequencing was considerably high. The total number of reads per sample enabled us to reach enough sequence depth where all or almost all species could be captured from the population. We attempted to maintain the high quality of the sequences by applying a clustering process different from that of "OTU " (operational taxonomic unit). The algorithm used here maximizes the biological meaning of the sequences regardless of the reference database used. Moreover, since ASVs assignment does not involve centroid methods, these sequences can be compared to those of other datasets. Any two ASVs may be differed by as few as one base pair, thus improving resolution during classification(Callahan et al., 2017; Caruso et al., 2019).

Before we attempt to answer the questions on our specific aims section, we compared the

microbiome structure found on our samples to that of the literature as a validation proposal of our findings.

Taxonomic classification revealed a microbiome structure in infants similar to those reported in the literature. Firmicutes, Proteobacteria, Bacteroidetes, Actinobacteria, and Verrucomocrobia were the most abundant in all sets. Previous studies have also showed these phyla as the most abundant in infants whose ages ranged from 1 to 3 years old(Azad et al., 2013; Bokulich et al., 2016; Chew et al., 2020; Hill et al., 2017; Yassour et al., 2016). By combining samples of all sets per time point, we found that relative abundance of Firmicutes showed less variation along the time. On the other hand, the relative abundance of Proteobacteria decreases, while Bacteroidetes increases with time starting around 6 months of age. Earlier reports showed similar results in European children born by cesarean(Hill et al., 2017; McGeachie et al., 2016; Christopher J. Stewart et al., 2018; Yassour et al., 2016). Previously, it has been shown that species belonging to Proteobacteria and Bacteroidetes have important roles in human health in early ages. Bacteroidetes species prepare the gut for colonization by the strict anaerobes, helping develop the initial microbiome development into a healthy and mature one(Moon, Young, Maclean, Cookson, & Bermingham, 2018). Proteobacteria species metabolize polysaccharides and oligosaccharides that provide nutrition and vitamins to the host(Zafar & Saier, 2021). We noted that Escherichia/Shigella´s abundance was high in the first months of life, but presented a small decrease as time went by. Studies on Brazilian newborns showed similar patterns(Brandt et al., 2012; Taddei et al., 2014), despite the fact that these babies were born vaginally and belonged to low socioeconomic backgrounds. *Bifidobacterium* was among the most abundant genus with little variation throughout the time points, and weaning did not seem to affect this pattern as it was reported previously (Mancabelli et al., 2020). Similarly, in c-section preterm European infants, bifidobacteria presented low relative abundance in the first week of life, increased in week 4 and presented little variation in the subsequent weeks (Hill et al., 2017). The abundance of this genus has long been associated with breastfeeding and aging, (Blekhman et al., 2015; Davenport et al., 2015). Our subjects were breastfed in the first months of life, but formula milk was introduced in the very beginning and continued until after weaning which might explain the little variation of this genus. Some studies also showed the early introduction of formula was associated with higher bifidobacterial carriage when compared to exclusively breast-fed babies(Harmsen et al., 2000; Nagpal et al., 2017). These results and comparisons to the literature suggest that the microbiome composition and its variation in our sets of triplets replicate the findings for similar profile children from western countries. This may be explained at least in part by the fact that the diet of people inhabiting large urban centers in Brazil (such as São Paulo) is generally similar to that of people in European cities(Santos & Conde, 2020). One alternative

explanation is that the establishment of the microbiota in infants is primarily determined by the human genome, with the environment being a secondary factor, but further analysis on the level of transcriptome is necessary to test this hypothesis.

5.1. Do MZ twins differ significantly from their DZ sibling in terms of microbiota diversity?

In this study, there was no significant difference between the MZs and their DZ co-twins in the alpha diversity permutation tests. On the other hand, a longitudinal study of Irish dichorionic triplets showed that DZ had higher alpha diversity than their MZs co-twins (Murphy et al., 2015), however, no statistical test results were provided by the authors. In beta diversity analysis, MZs were significantly more similar than their co-twin DZs in the measures Bray Curtis, Jaccard and Unweighted Unifrac, but not in Weighted Unifrac distance (Table 3). Similar results of weighted and unweighted unifrac were found previously in a study with 2,731 individuals whose mean age was 60 years old (Goodrich et al., 2016; Goodrich et al., 2014) and in another study of one set of dichorionic infant triplets (Murphy et al., 2015). PCoA on these beta diversity measures revealed some clusterization of MZs, and up to 4 principal coordinates were necessary to account for more than 50% variance of all samples.

5.2. Do MZ twins differ significantly from their DZ sibling in terms of ASV composition?

In order to answer this question we decided to give special attention to the ASVs due to the advantages explained above, thus we performed two main analyses on the ASV level: Estimated probability of shared ASVs by a multinomial model and Correspondence Analysis (CA) with log2 of the likelihood ratio. Venn Diagram graphs of shared ASVs showed that the estimated probability of shared ASVs between the MZs were bigger than between the MZs1/2 and their DZ co-twins in all sets, however these differences were not statistically significant as shown by the multinomial test, overall. We speculate that the sample size might have caused the divergent results between the graphs and the multinomial tests. Nevertheless, we considered specific (those ASVs that yielded positive results for similarities between the MZs) shared ASVs for further comparison with the results of the other two analyses. Initial CA revealed specific time points (9, 11 and 13 months) when the abundance of ASVs are more similar between the MZs. Thus, we focused our analysis

on these three time points and performed the CA on them only, and we also added all ASVs (explored below) that yielded significant heritability results to this analysis. It is known that time is a key factor that influences microbiome abundance, especially in early life (Hill et al., 2017; La Rosa et al., 2014; Martinez et al., 2018; Turroni, Milani, Ventura, & van Sinderen, 2022; Yassour et al., 2016) so we hypothesized that genetic factors influencing the development of the microbiome might be more evident in a time window. Although we identified twenty-two ASVs common to all sets at time points 9, 11 and 13 months, only two ASVs had the ratios of abundance more similar between the MZs when considering these specific time points.

5.3. If there are significant differences, can they be explained by specific ASVs?

After running heritability tests on the ASVs, we compared the results with the ones we obtained from the tests above. Three ASVs (1, 8 and 63) were especially noteworthy as their relative abundance/presence/absence were more similar between the MZs throughout the time points. These ASVs belonged to the genus Bacteroides (ASV-1) and Veillonella (ASV-8 and 63), and although Bifidobacteria was among the most abundant genus, its ASVs did not show positive results on the analysis regarding distribution among the triplets. Previous reports (Goodrich et al., 2016; Goodrich et al., 2014; Singh et al., 2017; Turnbaugh et al., 2009) have shown that gut bacteria of the genus Bacteroides are particularly sensitive to environmental factors. The evidence presented here suggests that bacteria of this genus may also be susceptible to host genetics. Veillonella was on the top three most abundant genera that had the most number (50) of ASVs that were exclusively shared between the MZs in all sets of triplets (Table 5). Previous reports on heritability of the human gut microbiome showed that the genus Veillonella had high heritability in an adult population of Hutterites (Davenport et al., 2015). In another gut microbiome study of adult Chinese (40–75 years old), the family Veillonellaceae was reported to have high heritability and its abundance to be negatively associated with some diseases predictable by host genetics (Xu et al., 2020). In a study of the oral microbiome of 485 Australian dizygotic and monozygotic twins aged five to eleven (Gomez et al., 2017), it was reported that Veillonella was one of the most heritable taxa as well as one of the most dominant. A recent longitudinal study on baboons demonstrated significant heritability in a large list of phenotypes based on 16S rRNA sequences, including the genus Veillonella as a whole and some ASVs classified as belonging to this genus(Grieneisen et al., 2021). The authors showed correlation between the heritability of traits in baboons and humans. This body of literature and our own results suggest that members of the genus Veillonella in the

gut may be susceptible to host genetic influence, especially early in life. Recent studies demonstrated some health benefits provided by *Veillonella*. *V. atypica* improves athletes performance by turning lactate into short-chain fatty acid propionate which counteracts inflammation and provides energy for the body (Scheiman, 2019). *V. parvula* was identified as a co-participant of the innate immune system modulation by increasing *IL-8, IL-6, IL-10*, and *TNF*-α responses which are involved in the regulation of inflammation process(van den Bogert B, 2014).

Taken together, these findings suggest that the host genetic-gut microbiome interaction can be very dynamic, especially in early life, when the microbiota is developing. Here we could analyze only three time points when host genetics revealed to be influenced by two genus, but we believe that there might be other time windows during infancy/childhood when these same or different genus are associated with host genetics.

6. Conclusion

As far as we know, our study is the first to conduct a descriptive analysis of the gut microbiome on five sets of triplets and test heritability on four of them. Overall, our findings are consistent with previous studies regarding the main phyla and genera and some patterns of abundance in the development of the human gut microbiota during the first months of life. As it was found in the literature, our individuals also showed shifts of the microbiota displaying higher diversity as time passed by. Although the sample size was one of the limitations for inferential statistical analysis in this study, we could describe some patterns of similarities between monozygotic twins. Our findings revealed that there are some bacteria that, at least at the sequence level, seem to be affected by genetics such as *Veillonella* and *Bacteroides*. We have shown that there is a time window when this effect occurs, thus capturing it is very challenging. Since signals of genetic effects on the gut microbiome might be subtle and temporary, we believe further studies using the experimental design described here should have as a priority an increase in the number of triplet sets, as well as increase in the number of shared time points between the triplet sets and the number of samples during the first six months of sampled babies' lives.

A manuscript describing the results of this dissertation has been publised by the journal *iScience* on January 28th, 2022.

Ondina Palmeira, Larissa R.B. Matos, Michel S. Naslavsky, Heloisa M.S. Bueno, Júlia P. Soler, João C. Setubal, Mayana Zatz**, Longitudinal 16S rRNA gut microbiota data of infant triplets show partial susceptibility to host genetics**, iScience,Volume 25, Issue 3, 2022, 103861, ISSN 2589-0042, https://doi.org/10.1016/j.isci.2022.103861

7. References

Aatsinki, A.-K., Lahti, L., Uusitupa, H.-M., Munukka, E., Keskitalo, A., Nolvi, S., . . . Karlsson, L. (2019). Gut microbiota composition is associated with temperament traits in infants. *Brain, Behavior, and Immunity, 80*, 849-858. doi:https://doi.org/10.1016/j.bbi.2019.05.035

Allard, G., Ryan, F. J., Jeffery, I. B., & Claesson, M. J. (2015). SPINGO: a rapid species-classifier for microbial amplicon sequences. *16*(1). doi:10.1186/s12859-015-0747-1

Almeida, A., Mitchell, A. L., Tarkowska, A., & Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience, 7*(5), giy054. doi:10.1093/gigascience/giy054

Almonacid, D. E., Kraal, L., Ossandon, F. J., Budovskaya, Y. V., Cardenas, J. P., Richman, J., & Apte, Z. S. J. b. (2016). 16S rRNA Gene Sequencing as a Clinical Diagnostic Aid for Gastrointestinal-related Conditions. 084657.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., . . . Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems, 2*(2), e00191-00116. doi:https://doi.org/10.1128/msystems.00191-16

Azad, M. B., Konya, T., Maughan, H., Guttman, D. S., Field, C. J., Chari, R. S., . . . Kozyrskyj, A. L. (2013). Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. *Canadian Medical Association Journal, 185*(5), 385-394. doi:https://doi.org/10.1503/cmaj.121189

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of disease in childhood. Education and practice edition, 98*(6), 236-238. doi:10.1136/archdischild-2013-304340

Belzer, C., & de Vos, W. M. (2012). Microbes inside—from diversity to function: the case of Akkermansia. *The ISME Journal, 6*, 1449. doi:10.1038/ismej.2012.6

https://www.nature.com/articles/ismej20126#supplementary-information

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., . . . Pomp, D. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences, 107*(44), 18933-18938. doi:10.1073/pnas.1007028107

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., . . . Clark, A. G. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biology, 16*(1). doi:10.1186/s13059-015-0759-1

Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., . . . Blaser, M. J. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine, 8*(343), 343ra382-343ra382. doi:https://doi.org/10.1126/scitranslmed.aad7121

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., . . . Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods, 10*(1), 57-59. doi:https://doi.org/10.1038/nmeth.2276

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., . . . Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology, 37*(8), 852-857. doi:https://doi.org/10.1038/s41587-019-0209-9

Brandt, K., Taddei, C. R., Takagi, E. H., Oliveira, F. F., Duarte, R. T. D., Irino, I., . . . Carneiro-Sampaio, M. (2012). Establishment of the bacterial fecal community during the first month of life in Brazilian newborns. *Clinics (Sao Paulo, Brazil), 67*(2), 113-123. doi:https://doi.org/10.6061/clinics/2012(02)05

Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *27*(4), 325-349. doi:10.2307/1942268

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data, 6*, 190007. doi:10.1038/sdata.2019.7

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *Plos One, 12*(10), e0185056. doi:10.1371/journal.pone.0185056

Buza, T. M., Tonui, T., Stomeo, F., Tiambo, C., Katani, R., Schilling, M., . . . Kapur, V. (2019). iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. *BMC Bioinformatics, 20*(1), 374. doi:10.1186/s12859-019-2965-4

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal, 11*(12), 2639-2643. doi:10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*, 581. doi:10.1038/nmeth.3869

https://www.nature.com/articles/nmeth.3869#supplementary-information

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*, 335. doi:10.1038/nmeth.f.303

https://www.nature.com/articles/nmeth.f.303#supplementary-information

Caruso, V., Song, X., Asquith, M., & Karstens, L. (2019). Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *mSystems, 4*(1), e00163-00118. doi:10.1128/mSystems.00163-18

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods, 69*(2), 330-339. doi:10.1016/j.mimet.2007.02.005

Chew, C., Barros, K. V., Weffort, V. R. S., Maranhão, H. d. S., Laranjeira, M., Knol, J., . . . de Morais, M. B. (2020). Gut Microbiota of Young Children Living in Four Brazilian Cities. *Frontiers in Pediatrics, 8*(783), 573815. doi:https://doi.org/10.3389/fped.2020.573815

Cho, I., & Blaser, M. J. (2012). APPLICATIONS OF NEXT-GENERATION SEQUENCING The human microbiome: at the interface of health and disease. *Nature Reviews Genetics, 13*(4), 260-270. doi:10.1038/nrg3182

Chong, C., Bloomfield, F., & O'Sullivan, J. (2018). Factors Affecting Gastrointestinal Microbiome Development in Neonates. *Nutrients, 10*(3), 274. doi:10.3390/nu10030274

Chow, J., Tang, H., & Mazmanian, S. K. (2011). Pathobionts of the gastrointestinal microbiota and inflammatory disease. *Current opinion in immunology, 23*(4), 473-480. doi:10.1016/j.coi.2011.07.010

Chrobak, A. A., Nowakowski, J., & Dudek, D. (2016). Interactions between the gut microbiome and the central nervous system and their role in schizophrenia, bipolar disorder and depression. *Archives of Psychiatry and Psychotherapy, 18*(2), 5-11. doi:10.12740/APP/62962

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., . . . Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research, 42*(Database issue), D633-D642. doi:https://doi.org/10.1093/nar/gkt1244

Collins, F. S., & Fink, L. (1995). The Human Genome Project. *Alcohol health and research world, 19*(3), 190-195.

Cong, X., Xu, W., Janton, S., Henderson, W. A., Matson, A., McGrath, J. M., . . . Graf, J. (2016). Gut Microbiome Developmental Patterns in Early Life of Preterm Infants: Impacts of Feeding and Gender. *Plos One, 11*(4), e0152751. doi:10.1371/journal.pone.0152751

Coventry, W. L., & Keller, M. C. (2005). Estimating the Extent of Parameter Bias in the Classical Twin Design: A Comparison of Parameter Estimates From Extended Twin-Family and Classical Twin Designs. *Twin Research and Human Genetics, 8*(3), 214-223. doi:10.1375/twin.8.3.214

D'Argenio, V. (2018). Human Microbiome Acquisition and Bioinformatic Challenges in Metagenomic Studies. *International Journal of Molecular Sciences, 19*(2), 383. doi:10.3390/ijms19020383

DATASUS. Nascidos Vivos - desde 1994. Retrieved from http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinasc/cnv/nvsp.def

Davenport, E. R. (2016). Elucidating the role of the host genome in shaping microbiome composition. *Gut Microbes, 7*(2), 178-184. doi:10.1080/19490976.2016.1155022

Davenport, E. R., Cusanovich, D. A., Michelini, K., Barreiro, L. B., Ober, C., & Gilad, Y. (2015). Genome-Wide Association Studies of the Human Gut Microbiota. *Plos One, 10*(11), e0140301. doi:https://doi.org/10.1371/journal.pone.0140301

de Goffau, M. C., Lager, S., Sovio, U., Gaccioli, F., Cook, E., Peacock, S. J., . . . Smith, G. C. S. (2019). Human placenta has no microbiome but can contain potential pathogens. *Nature*. doi:10.1038/s41586-019-1451-5

Derrien, M., Vaughan, E. E., Plugge, C. M., & de Vos, W. M. (2004). Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *54*(5), 1469-1476. doi:doi:10.1099/ijs.0.02873-0

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., . . . Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology, 72*(7), 5069-5072. doi:10.1128/AEM.03006-05

Di Paola, M., Sani, C., Clemente, A. M., Iossa, A., Perissi, E., Castronovo, G., . . . Torcia, M. G. (2017). Characterization of cervico-vaginal microbiota in women developing persistent high-risk Human Papillomavirus infection. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-09842-6

DiGiulio, D. B., Romero, R., Amogan, H. P., Kusanovic, J. P., Bik, E. M., Gotsch, F., . . . Relman, D. A. J. P. o. (2008). Microbial prevalence, diversity and abundance in amniotic fluid during preterm labor: a molecular and culture-based investigation. *3*(8), e3056.

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *107*(26), 11971-11975.

Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R., & Blaser, M. J. (2019). Role of the microbiome in human development. *Gut, 68*(6), 1108-1114. doi:10.1136/gutjnl-2018-317503

Dong, X., Kleiner, M., Sharp, C. E., Thorson, E., Li, C., Liu, D., & Strous, M. (2017). Fast and Simple Analysis of MiSeq Amplicon Sequencing Data with MetaAmp. *Frontiers in Microbiology, 8*, 1461-1461. doi:10.3389/fmicb.2017.01461

Donkor, E. S. (2013). Sequencing of bacterial genomes: principles and insights into pathogenesis and development of antibiotics. *Genes, 4*(4), 556-572. doi:10.3390/genes4040556

Drell, T., Štšepetova, J., Simm, J., Rull, K., Aleksejeva, A., Antson, A., . . . Mändar, R. (2017). The Influence of Different Maternal Microbial Communities on the Development of Infant Gut and Oral Microbiota. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-09278-y

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ, 5*, e3889-e3889. doi:10.7717/peerj.3889

Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J. P., Druart, C., Bindels, L. B., . . . Cani, P. D. (2013). Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proceedings of the National Academy of Sciences of the United States of America, 110*(22), 9066-9071. doi:10.1073/pnas.1219451110

Ewing, B., & Green, P. J. G. r. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *8*(3), 186-194.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation, 61*(1), 1-10. doi:https://doi.org/10.1016/0006-3207(92)91201-3

Felson, J. (2014). What can we learn from twin studies? A comprehensive evaluation of the equal environments assumption. *Social Science Research, 43*, 184-199. doi:https://doi.org/10.1016/j.ssresearch.2013.10.004

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., . . . Segata, N. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host & Microbe, 24*(1), 133-145.e135. doi:10.1016/j.chom.2018.06.005

Fisher, R. A. (1919). The Genesis of Twins. *Genetics, 4*(5), 489-499.

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., & Reimer, A. (2017). Metagenomics: The Next Culture-Independent Game Changer. *Frontiers in Microbiology, 8*, 1069-1069. doi:10.3389/fmicb.2017.01069

Fordtran, J. S. (2006). Colitis due to Clostridium difficile toxins: underdiagnosed, highly virulent, and nosocomial. *Proceedings (Baylor University. Medical Center), 19*(1), 3-12. doi:10.1080/08998280.2006.11928114

Garrett-Bakelman, F. E., Darshi, M., Green, S. J., Gur, R. C., Lin, L., Macias, B. R., . . . Turek, F. W. (2019). The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science, 364*(6436), eaau8650. doi:10.1126/science.aau8650

Gomez, A., Espinoza, J. L., Harkins, D. M., Leong, P., Saffery, R., Bockmann, M., . . . Nelson, K. E. (2017). Host Genetic Control of the Oral Microbiome in Health and Disease. *Cell Host & Microbe, 22*(3), 269-278.e263. doi:https://doi.org/10.1016/j.chom.2017.08.013

Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., . . . Ley, R. E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host & Microbe, 19*(5), 731-743. doi:https://doi.org/10.1016/j.chom.2016.04.017

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., . . . Ley, R. E. (2014). Human Genetics Shape the Gut Microbiome. *Cell, 159*(4), 789-799. doi:https://doi.org/10.1016/j.cell.2014.09.053

Gosalbes, M., Llop, S., Valles, Y., Moya, A., Ballester, F., Francino, M. J. C., & Allergy, E. (2013). Meconium microbiota types dominated by lactic acid or enteric bacteria are differentially associated with maternal eczema and respiratory problems in infants. *43*(2), 198-211.

Grieneisen, L., Dasari, M., Gould, T. J., Björk, J. R., Grenier, J.-C., Yotova, V., . . . Blekhman, R. (2021). Gut microbiome heritability is nearly universal but environmentally contingent. *Science, 373*(6551), 181. doi:https://doi.org/10.1126/science.aba5483

Groenen, I. B. P. J. F. (2005). *Modern Multidimensional Scaling*. ZUMA, Mannheim, Germany.

Harmsen, H. J., Wildeboer-Veloo Ac Fau - Raangs, G. C., Raangs Gc Fau - Wagendorp, A. A., Wagendorp Aa Fau - Klijn, N., Klijn N Fau - Bindels, J. G., Bindels Jg Fau - Welling, G. W., & Welling, G. W. (2000). Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. (0277-2116 (Print)).

Haworth, C. M. A., Dale, P., & Plomin, R. (2008). A Twin Study into the Genetic and Environmental Influences on Academic Performance in Science in nine-year-old Boys and Girls. *International journal of science education, 30*(8), 1003-1003. doi:10.1080/09500690701324190

Hiergeist, A., Gläsner, J., Reischl, U., & Gessner, A. J. I. j. (2015). Analyses of intestinal microbiota: culture versus sequencing. *56*(2), 228-240.

Hill, C. J., Lynch, D. B., Murphy, K., Ulaszewska, M., Jeffery, I. B., O'Shea, C. A., . . . Stanton, C. (2017). Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort. *Microbiome, 5*, 4. doi:https://doi.org/10.1186/s40168-016-0213-y

Houghteling, P. D., & Walker, W. A. (2015). Why is initial bacterial colonization of the intestine important to infants' and children's health? *Journal of pediatric gastroenterology and nutrition, 60*(3), 294-307. doi:10.1097/MPG.0000000000000597

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., . . . Earl, A. M. e. a. (2012). Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), 207-214. doi:10.1038/nature11234

Iwase, T., Uehara, Y., Shinji, H., Tajima, A., Seo, H., Takada, K., . . . Mizunoe, Y. (2010). Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization. *Nature, 465*(7296), 346-349. doi:10.1038/nature09074

Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *11*(2), 37-50. doi:10.1111/j.1469-8137.1912.tb05611.x

Jakobsson, H. E., Abrahamsson, T. R., Jenmalm, M. C., Harris, K., Quince, C., Jernberg, C., . . . Andersson, A. F. J. G. (2014). Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *63*(4), 559-566.

Jiménez, E., Fernández, L., Marín, M. L., Martín, R., Odriozola, J. M., Nueno-Palop, C., . . . Rodríguez, J. M. J. C. m. (2005). Isolation of commensal bacteria from umbilical cord blood of healthy neonates born by cesarean section. *51*(4), 270-274.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *109*(40),

16213-16216. doi:10.1073/pnas.1203849109 %J Proceedings of the National Academy of Sciences

Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30*(14), 3059-3066. doi:https://doi.org/10.1093/nar/gkf436

Keller, M. C., Medland, S. E., & Duncan, L. E. (2010). Are Extended Twin Family Designs Worth the Trouble? A Comparison of the Bias, Precision, and Accuracy of Parameters Estimated in Four Twin Family Models. *Behavior Genetics, 40*(3), 377-393. doi:10.1007/s10519-009-9320-x

Kennedy, K. M., Gerlach, M. J., Adam, T., Heimesaat, M. M., Rossi, L., Surette, M. G., . . . Braun, T. (2021). Fetal meconium does not have a detectable microbiota before birth. *Nature Microbiology, 6*(7), 865-873. doi:10.1038/s41564-021-00904-0

Kim, N., Yun, M., Oh, Y. J., & Choi, H. J. (2018). Mind-altering with the gut: Modulation of the gut-brain axis with probiotics. *Journal of Microbiology, 56*(3), 172-182. doi:10.1007/s12275-018-8032-4

King, C. H., Desai, H., Sylvetsky, A. C., LoTempio, J., Ayanyan, S., Carrie, J., . . . Mazumder, R. (2018). Baseline human gut microbiota profile in healthy people and standard reporting template. doi:10.1101/445353

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., . . . Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology, 16*(7), 410-422. doi:10.1038/s41579-018-0029-9

Knights, D., Silverberg, M. S., Weersma, R. K., Gevers, D., Dijkstra, G., Huang, H., . . . Xavier, R. J. (2014). Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Medicine, 6*(12). doi:10.1186/s13073-014-0107-1

Kochunov, P., Jahanshad, N., Marcus, D., Winkler, A., Sprooten, E., Nichols, T. E., . . . Van Essen, D. C. (2015). Heritability of fractional anisotropy in human white matter: A comparison of Human Connectome Project and ENIGMA-DTI data. *Neuroimage, 111*, 300-311. doi:https://doi.org/10.1016/j.neuroimage.2015.02.050

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., . . . Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences, 108*(Supplement 1), 4578. doi:10.1073/pnas.1000081107

La Rosa, P. S., Warner, B. B., Zhou, Y. J., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., . . . Tarr, P. I. (2014). Patterned progression of bacterial populations in the premature infant gut (vol 111, pg 12522, 2014). *Proceedings of the National Academy of Sciences of the United States of America, 111*(48), 17336-17336. doi:10.1073/pnas.1420817111

Lakhani, C. M., Tierney, B. T., Manrai, A. K., Yang, J., Visscher, P. M., & Patel, C. J. (2019). Repurposing large health insurance claims data to estimate genetic and environmental

contributions in 560 phenotypes. *Nature Genetics, 51*(2), 327-334. doi:10.1038/s41588-018-0313-7

Lammert, C. R., Frost, E. L., Bolte, A. C., Paysour, M. J., Shaw, M. E., Bellinger, C. E., . . . Lukens, J. R. (2018). Cutting Edge: Critical Roles for Microbiota-Mediated Regulation of the Immune System in a Prenatal Immune Activation Model of Autism. *The Journal of Immunology, 201*(3), 845. doi:10.4049/jimmunol.1701755

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., . . . Meta, H. I. T. C. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology, 32*(8), 834-841. doi:10.1038/nbt.2942

Liew, S. H. M., Elsner, H., Spector, T. D., & Hammond, C. J. (2005). The First 'Classical' Twin Study? Analysis of Refractive Error Using Monozygotic and Dizygotic Twins Published in 1922. *Twin Research and Human Genetics, 8*(3), 198-200. doi:10.1375/twin.8.3.198

Liu, Z., DeSantis, T. Z., Andersen, G. L., & Knight, R. J. N. a. r. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *36*(18), e120-e120.

Lozupone, C., & Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology, 71*(12), 8228. doi:https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and Environmental Microbiology, 73*(5), 1576. doi:https://doi.org/10.1128/AEM.01996-06

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., & Abd Allah, E. F. (2019). Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. *Frontiers in immunology, 9*, 2868-2868. doi:10.3389/fimmu.2018.02868

Mancabelli, L., Tarracchini, C., Milani, C., Lugli, G. A., Fontana, F., Turroni, F., . . . Ventura, M. (2020). Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). *Computational and Structural Biotechnology Journal, 18*, 2480-2493. doi:https://doi.org/10.1016/j.csbj.2020.08.028

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011, 17*(1), 3 %J EMBnet.journal. doi:10.14806/ej.17.1.200

Martinez, I., Maldonado-Gomez, M. X., Gomes-Neto, J. C., Kittana, H., Ding, H., Schmaltz, R., . . . Walter, J. (2018). Experimental evaluation of the importance of colonization history in early-life gut microbiota assembly. *Elife, 7*. doi:10.7554/eLife.36521

Matias Rodrigues, J. F., Schmidt, T. S., Tackmann, J., & von Mering, C. J. B. (2017). MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *33*(23), 3808-3810.

Mayer, E. A., & Hsiao, E. Y. (2017). The Gut and Its Microbiome as Related to Central Nervous System Functioning and Psychological Well-being: Introduction to the Special Issue of Psychosomatic Medicine. *Psychosomatic Medicine, 79*(8), 844-846. doi:10.1097/PSY.0000000000000525

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., . . . Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience, 1*(1). doi:10.1186/2047-217X-1-7 %J GigaScience

McGeachie, M. J., Sordillo, J. E., Gibson, T., Weinstock, G. M., Liu, Y. Y., Gold, D. R., . . . Litonjua, A. (2016). Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Scientific Reports, 6*, 20359. doi:https://doi.org/10.1038/srep20359

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *Plos One, 8*(4), e61217. doi:10.1371/journal.pone.0061217

Moon, C. D., Young, W., Maclean, P. H., Cookson, A. L., & Bermingham, E. N. (2018). Metagenomic insights into the roles of Proteobacteria in the gastrointestinal microbiomes of healthy dogs and cats. *MicrobiologyOpen, 7*(5), e00677-e00677. doi:10.1002/mbo3.677

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., . . . Rillig, M. C. (2014). Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and evolution, 4*(18), 3514-3524. doi:10.1002/ece3.1155

Murphy, K., O' Shea, C. A., Ryan, C. A., Dempsey, E. M., O' Toole, P. W., Stanton, C., & Ross, R. P. (2015). The Gut Microbiota Composition in Dichorionic Triplet Sets Suggests a Role for Host Genetic Factors. *Plos One, 10*(4), e0122561. doi:https://doi.org/10.1371/journal.pone.0122561

Mysara, M., Vandamme, P., Props, R., Kerckhof, F.-M., Leys, N., Boon, N., . . . Monsieurs, P. (2017). Reconciliation between operational taxonomic units and species boundaries. *Fems Microbiology Ecology, 93*(4), fix029. doi:10.1093/femsec/fix029

Nagpal, R., Kurakawa, T., Tsuji, H., Takahashi, T., Kawashima, K., Nagata, S., . . . Yamashiro, Y. (2017). Evolution of gut Bifidobacterium population in healthy Japanese infants over the first three years of life: a quantitative assessment. *Scientific Reports, 7*(1), 10097. doi:10.1038/s41598-017-10711-5

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ, 6*, e5364-e5364. doi:10.7717/peerj.5364

Nguyen, N.-P., Warnow, T., Pop, M., & White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *npj Biofilms and Microbiomes, 2*(1), 16004. doi:10.1038/npjbiofilms.2016.4

O'Callaghan, A., & van Sinderen, D. (2016). Bifidobacteria and Their Role as Members of the Human Gut Microbiota. *Frontiers in Microbiology, 7*. doi:10.3389/fmicb.2016.00925

Org, E., Parks, B. W., Joo, J. W. J., Emert, B., Schwartzman, W., Kang, E. Y., . . . Lusis, A. J. (2015). Genetic and environmental control of host-gut microbiota interactions. *Genome research, 25*(10), 1558-1569. doi:10.1101/gr.194118.115

Ottman, N., Smidt, H., De Vos, W. M., & Belzer, C. (2012). The function of our microbiota: who is out there and what do they do? *Frontiers in Cellular and Infection Microbiology, 2*. doi:10.3389/fcimb.2012.00104

Panek, M., Čipčić Paljetak, H., Barešić, A., Perić, M., Matijašić, M., Lojkić, I., . . . Verbanac, D. (2018). Methodology challenges in studying human gut microbiota – effects of collection, storage, DNA extraction and next generation sequencing technologies. *Scientific Reports, 8*(1), 5143. doi:10.1038/s41598-018-23296-4

Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., . . . Stobberingh, E. E. (2006). Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. *Pediatrics, 118*(2), 511. doi:10.1542/peds.2005-2824

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports, 8*(1), 10950. doi:10.1038/s41598-018-29325-6

Poesen, R., Windey, K., Neven, E., Kuypers, D., De Preter, V., Augustijns, P., . . . Meijers, B. (2016). The Influence of CKD on Colonic Microbial Metabolism. *Journal of the American Society of Nephrology, 27*(5), 1389. doi:10.1681/ASN.2015030279

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *Plos One, 9*(4), e93827. doi:10.1371/journal.pone.0093827

Price MN, D. P., Arkin AP. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One, 5*(3), e9490. doi:https://doi.org/10.1371/journal.pone.0009490

Primicerio, M. G. a. R. (2013). *Multivariate Analysis of Ecological Data*. Bilbao, Spain: Fundación BBVA.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology, 35*(9), 833-844. doi:10.1038/nbt.3935

Rende, R. D., Plomin, R., & Vandenberg, S. G. (1990). Who discovered the twin method? *Behavior Genetics, 20*(2), 277-285. doi:10.1007/BF01067795

Rogers, G. B., Keating, D. J., Young, R. L., Wong, M. L., Licinio, J., & Wesselingh, S. (2016). From gut dysbiosis to altered brain function and mental illness: mechanisms and pathways. *Molecular Psychiatry, 21*, 738. doi:10.1038/mp.2016.50

Rognes T, F. T., Nichols B, Quince C, Mahé F. (2016.). VSEARCH: a versatile open source tool for metagenomics. *PeerJ, 4*(e2584). doi:https://doi.org/10.7717/peerj.2584

Rojo, D., Hevia, A., Bargiela, R., López, P., Cuervo, A., González, S., . . . Ferrer, M. (2015). Ranking the impact of human health disorders on gut metabolism: Systemic lupus erythematosus and obesity as study cases. *Scientific Reports, 5*, 8310. doi:10.1038/srep08310

https://www.nature.com/articles/srep08310#supplementary-information

Rojo, D., M´endez-Garc´ıa, C., Raczkowska, B. A., Bargiela, R., Moya, A. e., Ferrer, M., & Barbas, C. (2017). Exploring the human microbiome from multiple perspectives: factors altering its composition and function. *Oxford-FEMS, 41*, 453–478. doi:10.1093/femsre/fuw046

Rosario, K., & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology, 1*(4), 289-297. doi:https://doi.org/10.1016/j.coviro.2011.06.004

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., . . . Segal, E. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature, 555*(7695), 210-215. doi:https://doi.org/10.1038/nature25973

Sahu, M., & Prasuna, J. G. (2016). Twin Studies: A Unique Epidemiological Tool. *Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine, 41*(3), 177-182. doi:10.4103/0970-0218.183593

Sánchez-Valverde, F., Gil, F., Martinez, D., Fernandez, B., Aznal, E., Oscoz, M., & Olivera, J. E. (2009). The impact of caesarean delivery and type of feeding on cow's milk allergy in infants and subsequent development of allergic march in childhood. *Allergy, 64*(6), 884-889. doi:https://doi.org/10.1111/j.1398-9995.2008.01931.x

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology, 94*(3), 441-448. doi:https://doi.org/10.1016/0022-2836(75)90213-2

Santos, I. K. S. d., & Conde, W. L. (2020). Tendência de padrões alimentares entre adultos das capitais brasileiras. *Revista Brasileira de Epidemiologia, 23*(13), e200035. doi:https://doi.org/10.1590/1980-549720200035

Scheiman, J., Luber, J.M., Chavkin, T.A. et al. . (2019). Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. *Nat Med, 25*, 1104–1109. doi:https://doi.org/10.1038/s41591-019-0485-4

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Robinson, C. J. J. A. E. M. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *75*(23), 7537-7541.

Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology, 14*(8), e1002533. doi:10.1371/journal.pbio.1002533

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379-423. doi:10.1002/j.1538-7305.1948.tb01338.x

Sharon, G., Sampson, T. R., Geschwind, D. H., & Mazmanian, S. K. (2016). The Central Nervous System and the Gut Microbiome. *Cell, 167*(4), 915-932. doi:10.1016/j.cell.2016.10.027

Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., . . . Liao, W. (2017). Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine, 15*(1), 73. doi:https://doi.org/10.1186/s12967-017-1175-y

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L. G., Gratadoux, J.-J., . . . Langella, P. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences, 105*(43), 16731. doi:10.1073/pnas.0804812105

Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., . . . Petrosino, J. F. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature, 562*(7728), 583-588. doi:https://doi.org/10.1038/s41586-018-0617-x

Stewart, C. J., Marrs, E. C. L., Nelson, A., Lanyon, C., Perry, J. D., Embleton, N. D., . . . Berrington, J. E. (2013). Development of the Preterm Gut Microbiome in Twins at Risk of Necrotising Enterocolitis and Sepsis. *Plos One, 8*(8), e73465. doi:https://doi.org/10.1371/journal.pone.0073465

Stewart, J. A., Chadwick, V. S., & Murray, A. (2005). Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *Journal of Medical Microbiology, 54*(12), 1239-1242. doi:https://doi.org/10.1099/jmm.0.46189-0

Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., . . . Mysorekar, I. U. (2013). Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *American Journal of Obstetrics and Gynecology, 208*(3), 226.e221-226.e227. doi:10.1016/j.ajog.2013.01.018

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A., . . . Gordon, J. I. (2014). Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature, 510*(7505), 417-+. doi:10.1038/nature13421

Taddei, C. R., Oliveira, F. F., Duarte, R. T. D., Talarico, S. T., Takagi, E. H., Carvalho, I. I. R., . . . Martinez, M. B. (2014). High Abundance of Escherichia During the Establishment of Fecal Microbiota in Brazilian Children. *Microbial Ecology, 67*(3), 624-634. doi:https://doi.org/10.1007/s00248-014-0381-x

Talarico, S. T., Santos, F. E., Brandt, K. G., Martinez, M. B., & Taddei, C. R. (2017). Anaerobic bacteria in the intestinal microbiota of Brazilian children. *Clinics, 72*(3), 154-160. doi:10.6061/clinics/2017(03)05

Tanaka, M., & Nakayama, J. (2017). Development of the gut microbiota in infancy and its impact on health in later life. *Allergology International, 66*(4), 515-522. doi:10.1016/j.alit.2017.07.010

Tang, W. H. W., & Hazen, S. L. (2014). The contributory role of gut microbiota in cardiovascular disease. *The Journal of Clinical Investigation, 124*(10), 4204-4211. doi:10.1172/JCI72331

Thompson, A. L., Monteagudo-Mera, A., Cadenas, M. B., Lampl, M. L., & Azcarate-Peril, M. A. (2015). Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *5*(3). doi:10.3389/fcimb.2015.00003

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., . . . Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature, 457*(7228), 480-U487. doi:https://doi.org/10.1038/nature07540

Turnbaugh, P. J., Quince, C., Faith, J. J., McHardy, A. C., Yatsunenko, T., Niazi, F., . . . Gordon, J. I. (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proceedings of the National Academy of Sciences of the United States of America, 107*(16), 7503-7508. doi:10.1073/pnas.1002355107

Turpin, W., Espin-Garcia, O., Xu, W., Silverberg, M. S., Kevans, D., Smith, M. I., . . . Croitoru, K. (2016). Association of host genome with intestinal microbial composition in a large healthy cohort. *Nature Genetics, 48*(11), 1413-1417. doi:10.1038/ng.3693

Turroni, F., Milani, C., Ventura, M., & van Sinderen, D. (2022). The human gut microbiota during the initial stages of life: insights from bifidobacteria. *Current Opinion in Biotechnology, 73*, 81-87. doi:https://doi.org/10.1016/j.copbio.2021.07.012

Valdes, A. M., Walter, L., Segal, E., & Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *Bmj-British Medical Journal, 361*. doi:10.1136/bmj.j2179

Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E. F., Wang, J., Tito, R. Y., . . . Raes, J. (2019). The neuroactive potential of the human gut microbiota in quality of life and depression. *Nature Microbiology, 4*(4), 623-632. doi:10.1038/s41564-018-0337-x

van de Merwe, J. P., Stegeman, J. H., & Hazenberg, M. P. (1983). The resident faecal flora is determined by genetic characteristics of the host. Implications for Crohn's disease? *Antonie van Leeuwenhoek, 49*(2), 119-124. doi:10.1007/BF00393669

van den Bogert B, M. M., Zoetendal EG, Wells JM, Kleerebezem M. . (2014). Immunomodulatory Properties of Streptococcus and Veillonella Isolates from the Human Small Intestine Microbiota. *Plos One, 9*(12), e114277. doi:https://doi.org/10.1371/journal.pone.0114277

Verstraelen, H., Vilchez-Vargas, R., Desimpel, F., Jauregui, R., Vankeirsbilck, N., Weyers, S., . . . Van De Wiele, T. (2016). Characterisation of the human uterine microbiome in non-

pregnant women through deep sequencing of the V1-2 region of the 16S rRNA gene. *PeerJ, 4*, e1602-e1602. doi:10.7717/peerj.1602

Wagner, B. D., Grunwald, G. K., Zerbe, G. O., Mikulich-Gilbertson, S. K., Robertson, C. E., Zemanick, E. T., & Harris, J. K. J. F. i. m. (2018). On the use of diversity measures in longitudinal sequencing studies of microbial communities. *9*, 1037.

Walker, W. A. (2017). The importance of appropriate initial bacterial colonization of the intestine in newborn, child, and adult health. *Pediatric Research, 82*(3), 387-395. doi:10.1038/pr.2017.111

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology, 73*(16), 5261. doi:https://doi.org/10.1128/AEM.00062-07

Wassenaar, T. M., & Panigrahi, P. (2014). Is a foetus developing in a sterile environment? *Letters in Applied Microbiology, 59*(6), 572-579. doi:10.1111/lam.12334

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

Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., . . . Jia, H. (2016). Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Systems, 3*(6), 572-584.e573. doi:https://doi.org/10.1016/j.cels.2016.10.004

Ximenez, C., & Torres, J. (2017). Development of Microbiota in Infants and its Role in Maturation of Gut Mucosa and Immune System. *Archives of Medical Research, 48*(8), 666-680. doi:10.1016/j.arcmed.2017.11.007

Xu, F., Fu, Y., Sun, T.-y., Jiang, Z., Miao, Z., Shuai, M., . . . Zheng, J.-S. (2020). The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for complex human diseases. *Microbiome, 8*(1), 145. doi:https://doi.org/10.1186/s40168-020-00923-9

Yang, X., Xie, L., Li, Y., & Wei, C. (2009). More than 9,000,000 Unique Genes in Human Gut Bacterial Community: Estimating Gene Numbers Inside a Human Body. *Plos One, 4*(6), e6074. doi:10.1371/journal.pone.0006074

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., . . . Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology, 12*, 635. doi:10.1038/nrmicro3330

https://www.nature.com/articles/nrmicro3330#supplementary-information

Yassour, M., Vatanen, T., Siljander, H., Hamalainen, A. M., Harkonen, T., Ryhanen, S. J., . . . Xavier, R. J. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine, 8*(343), 343ra381. doi:https://doi.org/10.1126/scitranslmed.aad0917

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., . . . Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature, 486*(7402), 222-227. doi:10.1038/nature11053

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., . . . Glöckner, F. O. (2014). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research, 42*(D1), D643-D648. doi:10.1093/nar/gkt1209

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A., & Elshahed, M. S. (2009). Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys. *Applied and Environmental Microbiology, 75*(16), 5227. doi:10.1128/AEM.00592-09

Zafar, H., & Saier, M. H. (2021). Gut Bacteroides species in health and disease. *Gut Microbes, 13*(1), 1-20. doi:10.1080/19490976.2020.1848158

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., . . . Lu, Z. (2018). Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *Science of The Total Environment, 618*, 1254-1267. doi:https://doi.org/10.1016/j.scitotenv.2017.09.228

Zhao, L., Wang, G., Siegel, P., He, C., Wang, H., Zhao, W., . . . Meng, H. (2013). Quantitative Genetic Background of the Host Influences Gut Microbiomes in Chickens. *Scientific Reports, 3*, 1163. doi:10.1038/srep01163

https://www.nature.com/articles/srep01163#supplementary-information

Zhou, Y., Wylie, K. M., El Feghaly, R. E., Mihindukulasuriya, K. A., Elward, A., Haslam, D. B., . . . Weinstock, G. M. (2016). Metagenomic Approach for Identification of the Pathogens Associated with Diarrhea in Stool Specimens. *Journal of Clinical Microbiology, 54*(2), 368. doi:10.1128/JCM.01965-15

Zhu, B., Wang, X., & Li, L. (2010). Human gut microbiome: the second genome of human body. *Protein & Cell, 1*(8), 718-725. doi:10.1007/s13238-010-0093-z

Zoetendal, E. G., Akkermansia, A. D. L., Vliet, W. M. A.-v., Visser, J. A. G. M. d., & Vos, W. M. d. (2001). The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. *Microbial Ecology in Health and Disease, 13*(3), 129-134. doi:https://doi.org/10.1080/089106001750462669