Infant Feeding Practices and the Gut Microbiome: Impact on Infant Growth and Development in South Africa

Claudine Kanyana Nkera-Gutabara

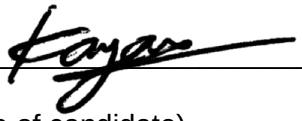
Student Number: 734113

A Dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

Johannesburg, 2020

DECLARATION

I, Claudine K. Nkera-Gutabara, declare that this Dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

7th day of June 20 20 in Johannesburg

This dissertation is dedicated

To my father, Dr Gihana Nkera-Gutabara,

For instilling the value of knowledge in me and guiding me to this achievement.

To my family and friends for their consistent support.

To God, for giving me the willpower and strength to reach the finish line.

PRESENTATIONS

- Poster presentation titled “Infant Feeding Practices and Secretor Status of Mother’s in an Urban South African Cohort”, at the 11th Annual Conference of the African Society of Human Genetics, hosted from 19th to 21st September 2018, in Kigali, Rwanda.
- Poster presentation titled “Infant Feeding Practices and Gut Microbiome Profiling: Impact on Infant Growth and Development in South Africa”, at the 18th Biennial South African Society of Human Genetics Congress, hosted from 4th to 6th August 2019, in Cape Town, South Africa.

ABSTRACT

In Africa, undernutrition is seen in approximately 45% of morbidity associated with the under five years' mortality rate. Children who are undernourished are at risk of dying, delayed or stunted growth and impaired brain development. Even if interventions are created, the negative effects of early poor nutrition seem to persist into adulthood, and warrant further research. There is strong support that the influence of microbial communities within the infant gut (gut microbiome) is significant in the utilisation of food nutrients for optimal growth. Alterations in the gut microbiome (dysbiosis) are associated with age, feeding practices and maternal host genetics. Maternal genotypes, specifically *α(1,2)-fucosyltransferase 2* (*FUT2*) rs601338 genotypes that determine secretor status, have been shown to alter breast milk composition. Little is known about these associations, the role of these variants, and their influence on the infant gut microbiome in non-Western populations.

The aim of this pilot study was three-fold: (1) to characterize the gut microbiome profile of South African infants living in an urban environment (Soweto) over time; (2) to assess whether the infant gut microbiome profiles are associated with exclusive breastfeeding, mixed feeding, and formula feeding practices; and, (3) to investigate the influence of maternal *FUT2* genotypes (secretor status) on the infant gut microbiome.

We recruited 46 mother-infant pairs via a collaboration with Wits Developmental Pathways for Health Research Unit. We then collected biological samples (infant stool sample and maternal saliva) from participants following informed consent. Infant stool samples were collected “at recruitment” (12-21 weeks/Time0), six months (Time6), nine months (Time9), and twelve months (Time12). We performed genotyping (TaqMan assay) of the *FUT2* gene mutation using DNA extracted from maternal saliva samples. We generated 16S rRNA sequencing data (via the sequencing facility at the National Institute for Communicable Diseases, Sandringham, Johannesburg) using DNA extracted from infant stool samples. Metagenomic analysis was performed in order to characterise the diversity and composition of microbiota present at different time points in infant stool samples using the QIIME2 pipeline, R and LEfSe. Association between infant gut taxonomic profiles at given time points (ages of the infants), feeding practice and maternal secretor status were investigated.

A total of 46 maternal saliva and 125 infant stool samples were collected. In this study, 83% of mothers participated in some form of breastfeeding – either exclusively or in combination with formula. Alpha diversity measures of Faith’s Phylogenetic Diversity revealed significant differences were specifically for Time12: versus Time0 (*P*/q-value=$0.44\text{x}10^{-3}$/$1.00\text{x}10^{-3}$); versus Time6 (*P*/q-value=0.01/0.02); and versus Time9 (*P*/q-value=$0.31\text{x}10^{-3}$/$1.00\text{x}10^{-3}$). Pielou’s Evenness revealed significant differences were specifically for the Time0: versus Time6 (*P*/q-value=0.01/0.03); and

versus Time9 (*P*/q-value=$2.00x10^{-3}$/0.02). There was also a higher diversity index in male samples compared to females for both alpha diversity measures. There were no distinct clusters when beta diversity was assessed using principle coordinates analysis. Beta diversity measures computed using PERMANOVA revealed statistically significant values for feeding practice (*P*=0.003) and time point analyses (*P*=0.001). Taxonomic analysis revealed that the greatest biodiversity is observed in exclusively breastfed infants and the least, in formula fed infants. Additionally, the ratio of the composition of the gut microbiome can be affected by the administration of medication at recruitment and nine months. This is between stages where feeding practices change and, the gut microbiome is dynamic and increases in diversity. Bacteria from the phylum Firmicutes make up the majority composition of the infant gut microbiome, and *Bifidobacteria* from the phylum Actinobacteria is uniquely associated with increased abundance in breast milk-fed infant microbiomes. Genotyping results revealed 79.1% of mothers were secretors, and 20.9% were non-secretors. Secretor status was not significantly associated with gut microbiome diversity - potentially due to the small sample size that excluded formula fed infants.

We were successful in using 16S rRNA sequencing and bioinformatic analysis to characterise the gut microbiome of clinically healthy infants from urban Soweto, South Africa. Future studies with a larger sample size in different geographical regions representing different communities within South Africa, would generate the necessary statistical power and enable the generalisability of these results to the South African population. The inclusion of infants adequately representing malnourished or stunted phenotypes, is necessary for a greater understanding of the relationship of this phenotype to the gut microbiome profile. This could address how the state of dysbiosis in those infants relates to these phenotypes, and would further describe the contributing factors that result in these poor health outcomes.

ACKNOWLEDGEMENTS

- I would like to thank my supervisors Dr Venesa Sahibdeen and Dr Robyn Kerr for their consistent guidance and encouragement throughout the research project and writing process.
- A special thanks to Dr Ovokeraye Oduaran and Dr Phelelani Mpangase for significant assistance with the bioinformatics analyses and Michaella Hulley for laboratory assistance.
- Thank you to the National Health Laboratory Service (Johannesburg) for provision of equipment and facilities necessary to carry out the research and funding for presentation of the research at local conferences, as well as to colleagues at the Division of Human Genetics and SBIMB for support.
- A big thank you to the Soweto Baby WASH team – Dr Rihlat Said-Mohamed, Doug Momberg, Hazel Pemba, Karabo Moloi, Nontuthuko Dube, Bongiwe Nkambule, Lutricia Moagi and Sostina Matina – for the assistance with participant recruitment and sample collections.
- This research was funded by an NRF Thuthuka Grant awarded to Dr V Sahibdeen, FRC Individual Research Grant and by a DST-NRF Innovation Masters Scholarship.

Table of Contents

LIST OF FIGURES

PAGE

LIST OF TABLES

PAGE

CHAPTER TWO

CHAPTER THREE

ABBREVIATIONS

%	Percentage
ABH	A, B, H blood group antigens
ARA	Arachidonic acid
BLAST	Basic Local Alignment Search Tool
CD	Crohn's disease
C-section	Caesarean section
DHA	Docosahexaenoic acid
DNA	Deoxyribonucleic acid
DPHRU	Developmental Pathways for Health Research Unit
DST	Department of Science and Technology
e.g.	For example
EtBr	Ethidium bromide
Faith's PD	Faith's Phylogenetic Diversity
FDA	Food and Drug Administration
FUT2	*Fucosyl transferase 2*
GC	Guanine-Cytosine
HMO	Human milk oligosaccharides
HMP	Human Microbiome Project
HREC	Human Research Ethics Committee
IDs	Identities
i.e.	In other words
Inc.	Incorporation
KGP	1000 Genomes Project
KW	Kruskal-Wallis
LDA	Linear Discriminant Analysis

LEfSe	Linear Discriminant Analysis Effect Size
µl	microliter
µM	micro molar
MAF	Minor allele frequency
NCBI	National Centre for Biotechnology Information
NEC	Necrotizing enterocolitis
ng	Nanogram
NGS	Next generation sequencing
NHLS	National Health Laboratory Service
NICD	National Institute for Communicable Diseases
OTU	Operational taxonomic unit
PCoA	Principal coordinates analysis
PCR	Polymerase chain reaction
PERMANOVA	Permutational multivariate analysis of variance
PROM	Premature rupture of membranes
Pty Ltd	Proprietary Limited
QIIME	Quantitative Insights into Microbial Ecology
R	Name of a programming language
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SA	South Africa
SBIMB	Sydney Brenner Institute for Molecular Biosciences
SBW	Soweto Baby WASH
SDGs	Sustainable Development Goals
SNP	Single nucleotide polymorphism
U	Symbol for ‘enzyme unit’ = micromole/min

UC	Ulcerative colitis
USA	United States of America
UNDP	United Nations Development Program
UNICEF	United Nations Children’s Emergency Fund
V3/V4	Variable region 3/variable region 4
w/v	Weight per volume
WASH	Water Sanitation and Hygiene
WHO	World Health Organization

CHAPTER ONE – INTRODUCTION

1.1 Rationale

Global statistical data has shown that more than half of mortality in children under five years is due to undernutrition or malnutrition (World Health Organization, 2017), and this has been described as the most important risk factor for illness and death in young children (Müller and Krawinkel, 2005). Over generations, chronic malnutrition has been a common occurrence in developing countries (Patel and Pettifor, 1992), although some progress has been made in countries such as South Africa to reduce the burden (UNICEF, 2019). Malnutrition can occur as a result of poor infant feeding and weaning practices, leading to a myriad of complications such as stunted growth, and delayed motor and mental development (Kruger and Gericke, 2002). Stunting is the most common outcome of undernutrition/malnutrition and is defined as "being too short for their age" and having a height-for-age (HAZ) index more than two standard deviations below the World Health Organization (WHO) Child Growth Standard median (UNICEF, 2013). Stunting associates with impaired physical growth, delayed cognitive development, lower productivity, and a greater risk of poor health including the development of cardiometabolic disease that may be transmitted to the next generation (Victora *et al*., 2008). In 2016, it was estimated that 155 million children under five years were stunted globally, with 34% in the WHO African region (WHO, 2017). The first 1000 days of life is the critical period when growth retardation and stunting start to develop (Kuklina *et al*., 2006).

The studies that have focused on malnutrition indicate that the determinants of stunting are multi-factorial including genetic and environmental influences. Stunting is mainly associated with: households of low socio-economic status and food security, instances of sickness during infancy, and maternal health before, during and after pregnancy. Twin studies have shown a higher similarity is observed in the "gut microbiota" (microbial species inhabiting the gastrointestinal tract) (Hagerty *et al*., 2020) composition between monozygotic twins than dizygotic twins, unrelated persons, marital couples and family members, pointing to a strong effect of host genetics (Zoetendal *et al*., 2001; Rajilic-Stojanovic, 2007). The short and long-term adverse effects of stunting on individuals, household and community levels highlight this as a public health burden and the need for reducing its prevalence (WHO, 2014). In the African context, the research done in rural communities has highlighted the issue of poor nutritional intake in infants (Ng'andu and Watts, 1990; Kruger and Gericke, 2003), and it has been suggested that the situation is exacerbated by behaviours such as sub-optimal breastfeeding and poor complementary feeding practices (Black *et al*., 2008).

The feeding practices undertaken in early age are fundamental for the appropriate development and health of young children. Exclusive breastfeeding has been shown to lower infant mortality by 13%, and is thus considered the most effective child health intervention (Fjeld *et al*., 2008). In a study by Kruger and Gericke (2003), assessing rural feeding and weaning practices in South Africa, it was found that most participants did not consider nutrition or actual food intake behaviour as being important in evaluating the health status of their babies. Exclusive breastfeeding was rarely practiced, and the introduction of solid foods was on average as early as two to three months of age. The foods used in substitution of breastfeeding were often overcooked or diluted e.g. maize meal washed and thinned with water (Kruger and Gericke, 2003). An earlier study in rural Zambia, assessed child growth over a period of two years, and concluded that the poor nutritional value of the weaning foods used were strongly associated with poor growth performance (Ng'andu and Watts, 1990).

Over the last two decades, Africa has experienced rapid transitions in nutrition and lifestyle. These are associated with an increased prevalence of non-communicable diseases; however the prevalence of stunting persists and may fuel the epidemiological trends of these diseases (Abrahams *et al*., 2011; Said-Mohamed *et al*., 2015). In the South African context, a study on infants admitted to the Chris Hani Baragwanath Academic Hospital in Soweto, reported an average prevalence of 26% of stunting, confirming its persistence as a public health burden (Said-Mohamed *et al*., 2015). Although interventions are utilised (Dewey and Adu-afarwuah, 2008), the negative effects of poor nutrition seem to persist into adulthood and warrant further research. This highlights the importance of alternative multi-disciplinary research on interventions being created that aim at making a difference in early life, ahead of the development of disease into adulthood. Such alternative examples of research are of microbiology and genetic studies. These have strongly supported the influence of microbial communities within the infant intestinal tract (colloquially referred to as "gut"). These communities are significant in the utilisation of food nutrients for optimal growth, where microbial alterations are associated with and may underlie malnutrition-associated pathologies (Ahmed *et al*., 2014; Laursen *et al*., 2017; Brewster *et al*., 2019).

1.2 Background of basic classification in Microbiology

Taxonomy is the science of classification, identification and nomenclature (naming) of microorganisms. The classification of microorganisms involves distinguishing one from another, and grouping similar organisms based on specific criteria of interest. The most important level of this classification is at the species level, whereby a species name should mean the same thing to everyone (Baron, 1996). A bacterial species is a distinct organism

with certain characteristic features, or a group of organisms that resemble one another closely in the most important features of their organization. These features include genetic criteria that can be used to define species in all groups of bacteria (Baron, 1996).

The names of microorganisms such as bacteria are underlined or written in italics, in order to identify the species-level or below level of classification. Other orders of classification – with some highlighted in Figure 1.1 (p. 3) – include subspecies, genera (groups of related species), families (groups of related genera), and other higher orders (Bergey and Williams, 1989; Baron, 1996).

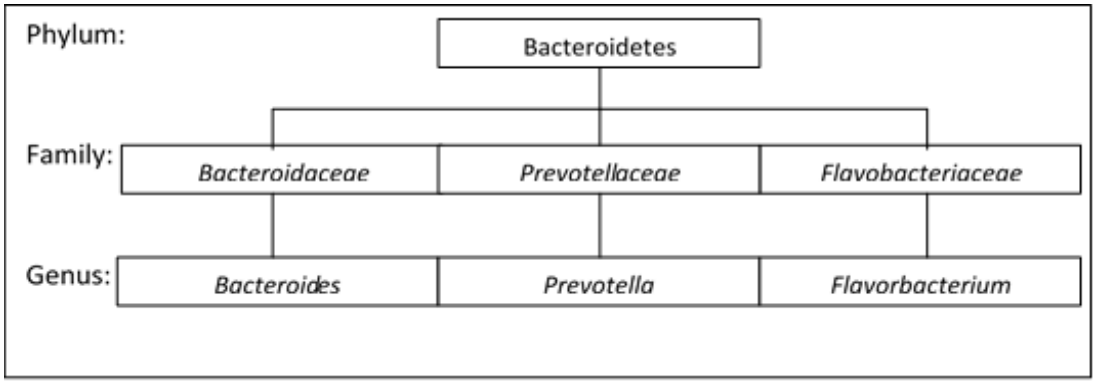


Figure 1.1: A taxonomic tree illustrating the classification of bacteria The phylum, family and genus level of classification of Bacteroidetes (Salguero *et al*., 2019).

Within a classified group, the bacteria can differ by the disease they produce, their environmental habitat and many other characteristics. The verification of the existing groups involves biochemical and phenotypic criteria – on the basis of a large number of characteristics – while DNA relatedness is used to group different species on the basis of overall genetic similarity (Baron, 1996). In practice, the approach to bacterial taxonomy should be polyphasic. This involves an initial step of phenotypic grouping of strains of species by morphological and biochemical characteristics, and other traits of interest. The following step would be to test for DNA relatedness to determine whether what is observed reflects the actual phylogenetic homogeneity (Baron, 1996). Phylogeny is a method for measuring the number of DNA sequences that organisms share and to estimate the percentage of divergence within their related but non-identical DNA sequences. The third step would be re-examining the biochemical characteristics of DNA related groups, enabling the determination of the diagnostic value of the group (Baron, 1996).

Classification below the species level usually involves designation of strains based on their common biochemical reactions, pathogenicity, or sensitivity to viral microorganisms (phage). Classification above the species level would ideally group strains into genera composed of

similar phenotypic and phylogenetic characteristics. However, when the phenotypic similarity is present and not the genetic similarity, scientists have used sequencing of ribosomal RNA (rRNA) genes. These have been highly conserved through evolution, and this technique falls under next-generation sequencing (NGS) (Baron, 1996).

1.3 The human gut microbiome

The human "microbiome" is a complex ecosystem that simultaneously exists as bacteria, viruses, fungi, and bacteriophages which interact with each other and their host (Tang *et al*., 2019). The collective microbial genes from the community are referred to as the microbiome where approximately 100 trillion microbial cells exist, containing almost 100-fold more unique genes than the human genome (Ahmed *et al*., 2014; Graham *et al*., 2015). The microbiome composition is unique to an individual and is established in early life (Kau *et al*., 2011). The microbiome displays site-specificity in composition, where distinct microbial populations ("microbiota") are found within each body site of an individual – for example the skin, mouth, vagina and gut (Ding and Schloss, 2014). Some examples of the distinct microbial composition of different body sites are – in the gut where enteric microbes produce numerous biologically active macromolecules to support host metabolism; and in the vagina where *Lactobacillus* species support a low pH which limits bacterial diversity and prevents bacteria from ascending to the uterus, which may ultimately result in infection of the amniotic fluid, placenta and foetus (Goldenberg *et al*., 2000; Turnbaugh *et al*, 2006; Ravel *et al*., 2011; van de Wijgert *et al*., 2014). Research has suggested that the development of the microbial ecosystem in early life plays a role in disease susceptibility throughout life (Scholtens *et al*., 2012). The relative contributions of the microbiota at each body site to disease susceptibility and health of the individual are not clearly defined, however they are likely to depend on the nature of the disease and the overall health of the individual (Zhang *et al*., 2015). In this study we focus on the microbiome composition of the human gut, particularly that of the infant.

The human gut microbiota is known to be involved in several metabolic, nutritional, physiological, and immunological processes (O'Hara and Shanahan, 2006). It is established after birth and starts out as dynamic and stabilizes over the first two to three years of life, whilst increasing in diversity and richness into adulthood (Scholtens *et al*., 2012). The gut is located at a crucial interface between exogenous dietary intake and internal nutrient metabolism (Tang *et al*., 2019). The microbiome composition in combination with diet is likely to play a major role in determining gut mucousal membrane permeability (Moreira *et al*., 2012). Whereby, specific dietary components act as food sources for microbial metabolism, shaping its composition and function (Wu *et al*., 2011). This serves as a source

of genetic diversity due to the representation of millions of genes and various gene functions within our gut microbial communities. The sampling of different populations across ages and various cultures allows for the discovery of how gut microbiomes evolve and vary between individuals, and how they respond to changing lifestyles, which research suggests has greater influence than genetic profiles and ethnicity (Yatsunenko *et al*., 2012).

Research conducted on characterising adult gut microbiomes, is largely based on "Western" studies from Europe and the USA (Biagi *et al*., 2010; Claesson *et al*., 2011; Wacklin *et al*., 2014; Karikari *et al*., 2015). The most consistent pattern observed in African populations (for example Burkina Faso, Central African Republic) is the inverse relationship between two members of the Bacteroidetes phylum, *Prevotella* and *Bacteroides* (Brewster *et al*., 2019). *Prevotella* is associated with diets rich in plant-based foods characteristic of non-industrialized populations, and are involved in breaking down/degradation of fibres as well as starch, hemicellulose and xylans (De Filippo *et al*., 2017). *Bacteroides* grow exclusively in the mammalian gut, suggesting a strong adaptation to this environment. These bacteria are associated with a higher intake of fats and proteins characteristic of industrialized populations, and plays a role in the metabolism of these compounds together with digestible sugars (De Filippo *et al*., 2010; Schnorr *et al*., 2014; Gomez *et al*., 2016; De Filippo *et al*., 2017; Ayeni *et al*., 2018). With a need for translational applications of research, it is important to establish the generalizability of findings of microbiome research. The research from Western populations show that lifestyle, diet and environmental exposures strongly alter and affect gut microbiome composition, thus significant biases are faced and non-western microbiome variability is poorly understood (Brewster *et al*., 2019).

The gut microbiome variability is additionally affected by changes in sanitation and hygiene practices, and administration of medication - particularly antibiotics (Brewster *et al*., 2019). Antibiotics can cause a pathological imbalance in gut microbial diversity, also referred to as gut microbial dysbiosis (Arrieta *et al*., 2014). Research has shown that in antibiotic-treated individuals, enrichment in bacteria carrying xenobiotic-processing genes is observed – which highlights a relation to exposure to medication or other foreign substances like food additives, and the increase in bacteria with functions to metabolize them (Gomez *et al*., 2016; Hansen *et al*., 2019). In a study on children in Western Africa, it was reported that a five day administration of antibiotics decreased the gut microbiome diversity and abundance and altered its composition (Doan *et al*., 2017; Oldenburg *et al*., 2018).

1.4 The infant gut microbiome

The colonization of the infant gut is influenced by several factors, including genetics, the mode of delivery, diet, gestational age, environment and antibiotic administration (Adlerberth

and Wold, 2009; Marques *et al*., 2010; Khodayar-Pardo *et al*., 2014). Research on the early gut colonization has a general consensus that the process involves four phases. Phase one is a sterile gut environment prior to birth where the maternal microbiota is separated from the developing foetus at the maternal-foetal interface of the placenta, chorio-amniotic sac, and immunological barriers. However, this has been challenged by the possibility that microbial colonization may begin *in utero*, due to evidence of bacterial genomic DNA within placental and chorio-amniotic tissues, and culturable microbes in the meconium i.e. the first faeces passed by the infant after birth (Jiménez *et al*., 2008; Aagaard *et al*., 2016). A recent study found that the human placenta did not have a microbiome, although it contained potential pathogens in a small proportion of samples (de Goffau *et al*., 2019). This study debates the existence of a foetal microbiota but does not provide definitive conclusions.

Phase two is the initial acquisition during delivery and the hospital environment. A wealth of evidence (Yatsunenko *et al*., 2012; Chu *et al*., 2017; Laursen *et al*., 2017; McDonald and McCoy, 2019) does in fact support the establishment at birth, and an influence of the mode of delivery. The dogmatic belief around the establishment of the infant microbiome is that at birth (phase two), the skin and mucousal surfaces of the new-born are exposed to microorganisms from the mothers body i.e. the maternal microbiota (McDonald and McCoy, 2019). It was argued that the mode of delivery of the baby strongly affects the composition of the microbiota, where babies born via caesarean-section (C-section) deliveries had microbiotas similar to the skin communities of the mothers, and those infants vaginally delivered resembled their mother's vaginal microbiotas (Dominguez-Bello *et al.,* 2010). Earlier research has demonstrated that C-section delivery results in a substantial reduction of *Bifidobacteria* – the predominant bacterial strain/species in the infant gut that also make up ~6% of the normal gut microbiota of adults (Lay *et al*., 2005) – in comparison to vaginal deliveries (Biasucci *et al*., 2008). *Bifidobacteria* assist in the modulation of the immune system development and govern early gut community structure (Ruiz *et al*., 2017). However, more recent observational studies of maternal-infant pairs argue that the mode of delivery has no discernible effect on the microbiome except in the very early neonatal period, and through the first six weeks of life the resulting microbial communities are diversified regardless of delivery route (Chu *et al*., 2017).

Phase three occurs during breastfeeding or formula feeding which produces different gut microbial compositions and levels of diversity, and phase four is where the start of solid foods results in a transition into an adult microbial flora which will be elaborated on further.

The early life - particularly the first 1000 days of life after birth - of an infant is significant, where the dietary richness and environmental exposures increase, and simultaneously

increase the richness and complexity of the gut microbiota (Yatsunenko *et al*., 2012). These first colonizing microbes of the gut are pivotal in immune system development and response and have been reported in the intestinal defence mechanisms of preterm babies, however the full functional acquisition of the microbiota is yet to be realized (Nanthakumar *et al*., 2000; Yatsunenko *et al*., 2012). Several notable taxa of the infant gut microbiome are *Bacteroides*, *Lactobacillus* and *Bifidobacterium*, and other observations in research highlight that *Escherichia* and *Klebsiella* are also predominant and highly specific for the infant gut (Koenig *et al*., 2011; Bäckhed *et al*., 2015; Chu *et al*., 2017). These bacteria are enriched in several microbial pathways that have been highlighted during infancy which include: the pentose phosphate pathway and phosphotransferase system involved in utilizing glucose for anabolism of amino acids and cell wall components; as well as pathways related to co-factor metabolism and biosynthesis of folate, biotin and vitamin B6, which are important for growth, metabolism and neurodevelopment (Chu *et al*., 2017; Kennedy, 2016).

In a longitudinal sampling study of infant gut microbiome composition, it was observed that at birth: appreciable levels (>1%) of *Lactobacillus*, *Bifidobacterium* and *Bacteroides* were present. Upon reaching six weeks of age where effects of mode of delivery would reportedly cease, *Lactobacillus* and *Bifidobacterium* were consistently present – which play different roles and dominate the guts of breastfed infants – while *Bacteroides* levels decreased or were undetectable (Chu *et al*., 2017). The findings of these studies from large cohorts – albeit from western populations – demonstrates a baseline for early infant microbiota composition, however due to the growing understanding of the variability of the microbiome, African specific research must be performed.

The African infant gut microbiome

In general, research on gut microbiome composition has shown that the most populated bacterial phyla observed in infants are Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria. However, when considering the differences that associate with influential factors of geographical origin, the differences in abundance of these phyla become apparent (Bäckhed *et al*., 2005; Qin *et al*., 2010). In a study with these phyla represented: Actinobacteria and Bacteroidetes phyla were more represented in an African cohort than in European infants, whereas Firmicutes and Proteobacteria were more abundant in Europeans than in African infants (De Filippo *et al*., 2010). When investigating these differences and accounting for additional factors that shape the biodiversity such as lifestyle, diet and mode of delivery, the variability in microbiome diversity across populations becomes all the more complex.

In a Central African study that investigated the early influence of delivery mode, an observation was made by Brazier and colleagues (2017), noting significant taxonomic differences in vaginal delivery (*Bacteroides* and *Collinsella*) versus C-section delivery (*Sarcina* and *Klebsiella*). Although these differences are noted by Chu *et al.* (2017) to have no discernible effect by six weeks of age, longitudinal studies over the course of an individual's life may in fact reveal that these early differences may affect physiological outcomes in adulthood. Another study (Kigbu *et al*., 2016) suggested the level of hygiene and sanitation practices of a delivery environment is a factor influencing microbial composition in an African versus European setting: where ~90% of one-day old infants in the African study were colonized by *Escherichia coli*, compared to the much lower rate of colonization in the corresponding European infants (Adlerberth *et al*., 1991; Adlerberth *et al*., 2006).

After birth the greatest influence on the early gut microbiome colonization are infant feeding practices, which are paramount in the transition from **aerobes or facultative anaerobes** (species that can create energy by respiration or fermentation in the presence or absence of oxygen) **to strictly anaerobic bacteria** (species that create energy by fermentation and die in the presence of oxygen). These facultative anaerobes create the environment necessary for the thriving of early strictly anaerobic colonizers that are involved in important developmental processes (Underwood *et al*., 2015). Studies that went further to investigate functional signatures in African cohorts, highlighted enrichment in genes related to differences in breast milk composition and earlier introductions to solid foods (De Filippo *et al*., 2017). When comparing the gut microbiome composition in Malawian and Finnish infants, a greater species diversity of *Lactobacillus* and higher levels of *Bifidobacterium, Clostridium histolyticum* and *Bacteroides-Prevotella* were observed by six months of age in the Malawian cohort (Grześkowiak *et al*., 2012; Aakko *et al*., 2015). Dietary influences on the gut microbiome persist up to three years of life, where the dynamic colonization process eventually stabilizes and forms an adult-like state in terms of composition and diversity (Koenig *et al*., 2011; Bergstrom *et al*., 2014). The process of weaning off of breast milk has been shown to increase the presence of Bacteroidetes and Firmicutes phyla, and is in line with the type of nutrition gained which in turn relates to the different diets amongst different geographical regions (Grzeskowiak *et al*., 2012; Aakko *et al*., 2015). Therefore, the different patterns of colonization and variation in African versus Western infant guts based on nutritional status may be important to further elucidate the outcome of host phenotypes observed later in life (Brewster *et al*., 2019).

1.5 Early nutrition and nutritional status

Since we have established that diet is an important influencing factor in the gut microbiome development and composition, the lack of access to adequate nutrition can have deleterious effects. The World Health Organization estimates that nearly 40% of all global cases of undernutrition – with regard to macro- and micronutrient deficiencies – are accounted for by sub-Saharan African countries (Bailey *et al*., 2015; WHO, 2015). The most common outcome of poor nutrition in the first 1000 days of a child's life is stunted growth, which is irreversible and associated with impaired cognitive ability and reduced school and work performance (WHO, 2015). In relation to the gut microbiome, a malnourished phenotype can cause negative alterations in the microbial composition (i.e. dysbiosis) and the gut microbiome in turn can contribute to, or worsen the individual's malnourished phenotype by reducing the ability to effectively utilize dietary nutrients (Kane *et al*., 2015). Studies have generated models of age-discriminatory bacteria that allow a microbiota-for-age Z score to be computed from undernourished infant microbiomes (Subramanian *et al*., 2014; Blanton *et al*., 2016). When applied to undernourished infants from Malawi, results of a persistent gut microbiota "immaturity" was observed (Blanton *et al.,* 2016). This meant the microbial community in these infants were "younger" than their chronologically age-matched healthy counterparts; highlighting a perturbation of the gut microbiome by malnourished phenotypes (Blanton *et al*, 2016).

As stated previously, the general consensus of various global research shows that *Bifidobacteria* species dominate the stool communities throughout the first year of life, and their proportional representation diminish into adulthood (Penders *et al*., 2006; Yatsunenko *et al*., 2012). Earlier studies have reported that breastfed infant microbiotas are dominated by *Bifidobacteria* and *Lactobacilli* (Sela *et al*., 2008; Adlerbeth and Wold, 2009; Sela and Mills, 2010). *Bifidobacteria* are characterized as beneficial commensals and possess immunomodulatory functions, with roles of beneficial metabolite production, and metabolism of diet-derived and non-digestible carbohydrates such as human milk oligosaccharides (HMOs i.e. the largest constituent of breast milk) (Bottacini *et al*., 2017).

Subspecies found in the infant gut, particularly *Bifidobacterium longum* subsp. *infantis*, typically carry a wide variety of genes that enable the sole use of HMOs for energy. These subspecies are involved in membrane transport and intracellular degradation of intact HMOs, while other subspecies rely partially on extracellular enzymes for HMO utilization (Sela *et al*., 2008; Sela and Mills, 2010; Sela, 2011; Garrido *et al*., 2016). *Lactobacilli* species contain genes encoding enzymes (fucosidases), and these degrading enzymes have a positive correlation to the breakdown of HMOs (Rodriguez-Diaz *et al*., 2011; Davis *et*

al., 2017). Studies also report that formula fed infants have higher proportions of *Bacteroides, Clostridium, Streptococcus, Enterobacteria*, and *Veillonella spp*. (Adlerberth and Wold, 2009; Fallani *et al*., 2010; Bezirtzoglou *et al*., 2011; Sahl *et al*., 2012; Vatanen *et al*., 2019). While formula feeds may have the nutritional value of breast milk, many of them lack HMOs – which are indigestible by the infant gut, but importantly act as a substrate for *Bifidobacteria* that seed that infant gut microbiome (Sela and Mills, 2008).

Bifidobacteria are mutualistic with their hosts, and these bacteria are commonly incorporated in probiotic products (Boesten and Vos, 2008). Probiotics are live beneficial bacteria which form naturally from fermentation of foods such as yoghurt, and can also be added to foods to help restore health to intestinal tracts. Existing interventions against malnutrition using probiotic microbes have been shown to decrease the incidence of infant diarrhoea due to host immune stimulation, lending support to the theory that the microbiomes play a role in immunity (Niers *et al*., 2009; Sjogren *et al*., 2009). These can work in conjunction with prebiotics, which are foods that help increase the beneficial bacteria already present in an individual's gut (Thomas *et al.*, 2010).

Other interventions involving nutrient replacement strategies have highlighted the issue of high rates of relapse and comorbidities, thus limiting the efficacy of such interventions (O'Sullivan *et al*., 2018). An example is the oral iron supplementation or fortification through micronutrient powders, for treating iron-deficiency anaemia (Pasricha *et al*., 2013). Many gut bacteria rely on iron as a growth-limiting factor (Andrews *et al*., 2003), and this results in the colonization of pathogenic microbes, with little effect on beneficial bacteria such as *Bifidobacteria* (Jaeggi *et al*., 2015). The side effects of such treatments can be negative in certain populations, due to the increase in dysbiosis and gut inflammation, where high-dose formulations of this supplement leads to increased hospitalizations, deaths and diarrhoea (Zlotkin *et al*., 2013). Another treatment strategy that is often administered is the use of antibiotics which have a disruptive effect on the gut microbiome, where decreases in diversity and dysbiosis have been well-established (Dethlefsen and Relman, 2011; Oldenburg *et al*., 2018). Using antibiotics in conjunction with other treatment strategies can further complicate the existing dysbiosis of the host, by increasing pathogens and decreasing beneficial bacterial abundance (Paganini *et al*., 2019).

1.5.1 The composition of formula food versus breast milk

Breastfeeding has nutritive and non-nutritive components (Figure 1.2, p. 11). Breast milk provides a continuous source of maternal microbes, as well as supplying nutritional and antimicrobial factors (McDonald and McCoy, 2019). Although strides have been made to ensure formula food content is as similar to that of breast milk as possible, science has

shown (Harmsen *et al*., 2000; Bezirtzoglou *et al*., 2011; Cabrera-Rubio *et al*., 2012; Underwood *et al*., 2015) differences in gut microbial composition, which indicates that even minute differences in their composition make an impactful difference.

In the South African context, various types of formula are available, although largely containing similar macromolecular content as required by the Food and Drug Administration (FDA) (Trumbo, 2016). The different formulas are as follows (Martin *et al*., 2016):

I) Regular formula is made from cow's milk and adapted to be suitable for babies. The proteins are altered to be easily digestible; milk sugar (i.e. lactose) is added to resemble breast milk; while vegetable oil is replaced for butterfat.

II) Hydrolysed formula is a cow's milk-based formula treated with enzymes to break down proteins that cause symptoms in allergenic infants.

III) Soy-based formula is made from soybeans supplemented with vitamins, minerals and nutrients, but is even more different from breast milk than cow's milk. This is usually recommended for infants with a cow's milk allergy.

IV) Special formula for special medical purposes such as; premature babies, cow's milk protein allergies, or metabolic disorders, are used at a paediatrician's recommendation.

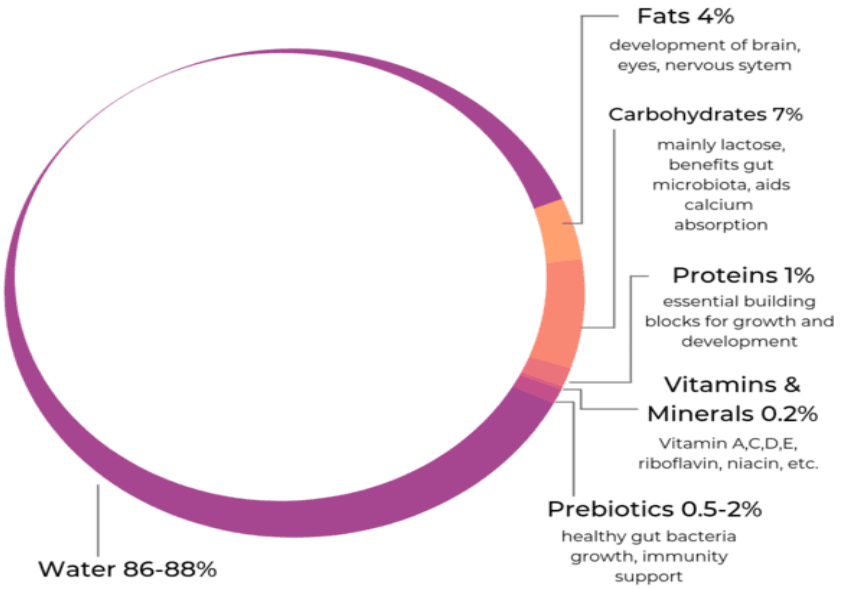


Figure 1.2: An illustration highlighting the composition of human breast milk This diagram shows the proportion of the different macro- and micronutrients contained in breast milk, and their relative significance to infant development. (% = total grams/100ml) (Finglas *et al*., 2014).

These formulas are available in different forms such as: ready-to-use, liquid concentrate, and powdered formula – with the latter being the most economical option. This requires a safe water supply to prepare and is suitable for occasional supplemental bottle-feeding, as

this occurs for example, with working mothers or when the amount of breast milk produced is insufficient (Martin *et al*., 2016; Nieuwoudt *et al.*, 2018).

The five main components of formula – in line with the essentials produced naturally in breast milk (Figure 1.2, p. 11) – are carbohydrates, proteins, fats, vitamins, minerals, and other nutrients in smaller quantities. For carbohydrate content, lactose is the main component in both breast milk and formula from cow's milk. In breast milk, the protein content is ~60% whey (by-product of milk curdling/straining) and ~40% casein (family of related phosphoproteins) with most formulas having similar content. For fat composition, breast milk has a blend of monounsaturated, polyunsaturated and saturated fat. Formulas use a variety of oils like soy, coconut, palm or sunflower oil to match the fat makeup of breast milk. Docosahexaenoic acid (DHA) and Arachidonic acid (ARA) are two long-chain fatty acids that are approved by the FDA as standard ingredients in formula. These fatty acids are supplied naturally by the mother during the third trimester and are important for brain and vision development during the first year of life. Numerous vitamins and minerals such as iron, vitamin C and vitamin B5 are included and listed on formula packages. Iron-fortified formula is generally recommended in non-exclusively breastfed infants because iron is important for developing a strong and healthy body. Nucleotides and amino acids are added to match what occurs naturally in breast milk and aid in immune system development. Formula is also supplemented with probiotics and prebiotics such as those stated above, to aid in digestion and gut microbial maturation (Martin *et al*., 2016).

1.5.2 Genetics and nutrition

Most of the gene functions assigned to early infant microbiomes are involved in the breakdown and metabolism of breast milk. Vaishampayan and colleagues (2010) have shown that during an exclusive breast milk diet, microbial genes facilitating the metabolism of breast milk are present. The active genes in bacteria are accompanied by proteins specialized in the transport and metabolism of human milk oligosaccharides (HMO) between infant and maternal microbiota samples. The HMOs are part of the innate defence system, and are the third largest solid constituent of human milk after lactose and lipids. HMOs are composed of various linkages of complex sugars (Morrow *et al*., 2004). Interestingly, HMOs are not digestible by the infant gut. This immediately then raises questions about the function of oligosaccharides in breast milk. It is postulated that these milk oligosaccharides serve as a food source instead for the bacteria that populate the infant gut, as highlighted earlier with reference to *Bifidobacteria* (Sela *et al*., 2008).

The HMOs secreted into milk vary, depending on, among other things, the mother's ABH and Lewis-histo blood group antigens (Morrow *et al*., 2004). Individuals with the A, B or H

blood group, contain A, B and H antigens, respectively. These antigens are present in mucous and other secretions and their expression is controlled by the enzyme fucosyltransferase 2 (*FUT2*), encoded by the *FUT2* gene. In “secretors”, *FUT2* converts glycan chains to H antigen, functioning as a precursor to ABH antigens. In “non-secretors”, a nonsense mutation in the *FUT2* gene (c.461G>A) leads to lack of expression of an active fucosyltransferase 2 enzyme, therefore ABH antigens are not expressed in mucous and other host secretions (Wacklin *et al*., 2011). These mucosal antigens serve as an energy source, and adhesion receptors for numerous microbes, highlighting their role in contributing to the host microbiota composition (Hoskins *et al*, 1985; Moulds *et al*., 1996).

A study by Wacklin and colleagues (2011), reported that the nonsense mutation in the *FUT2* gene (rs601338) determining the host secretor status is strongly associated with shaping the composition of *Bifidobacteria* in the human gut. The complex mixture within the HMOs acts as a bacterial food source and attracts mutualistic mucous-adapted species and HMO-adapted *Bifidobacteria* to the infant gut, that likely assist in breast milk and future solid food digestion. This highlights the significance of the mother's breast milk composition on the infant’s gut microbiome expression, influencing their growth and development (Schwartz *et al*., 2012). Genotyping of the rs601338 variant in different populations was performed by the 1000 Genomes Project (Figure 1.3, p. 13), identifying the global frequency of this variant (The 1000 Genomes Project Consortium, 2015). This variant has a global average frequency

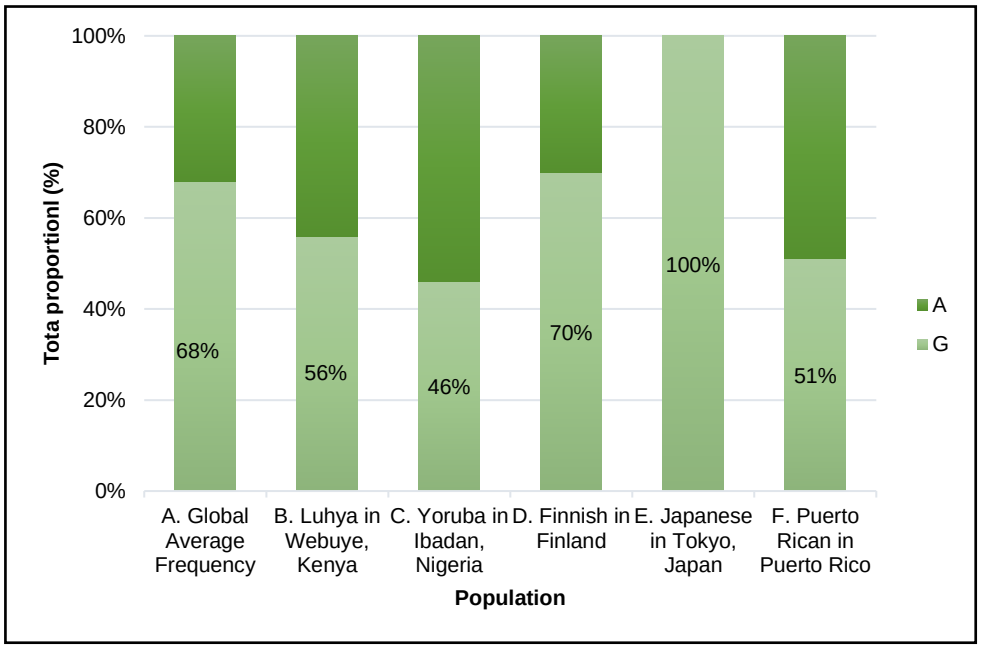


Figure 1.3: **The average global population frequencies of *FUT2*- rs601338 alleles** (“G” and “A”) The global average frequencies (A) of the alleles are 68% (“G”) and 32% (“A”). The most significant difference in allele frequencies can be seen in Japanese populations (E), where the mutant “A” allele has 0% frequency. The European population (D) frequency is similar to the global average frequency. In African populations (B, C) and the Puerto Rican population the “A” allele has a higher frequency than the global average. (The 1000 Genomes Project Consortium, 2015).

of 0.32: which is higher than in European populations (0.30, Finnish); lower than reported in African populations (0.44, Kenyan; 0.54, Nigerian).

A study based in rural Malawian birth-cohorts reported that HMOs contain sialylated (human milk oligosaccharide with a sialic acid moiety attached) and fucosylated (HMO with a fucose acid moiety attached) sugars, which are significantly less abundant in mothers with severely stunted infants (Charbonneau *et al*., 2017). The study showed that a secretor phenotype is identified by the *FUT2* gene producing a α1-2 linked fucosylated HMO; hence the fucosylated HMO content of breast milk samples collected from secretor mothers was significantly higher than non-secretors. Non-secretor phenotype mothers were deficient in fucosylated as well as sialylated HMOs, and had severely stunted children. This supports the significance of a functional *FUT2* gene that affects the composition of breast milk and influences infant growth. In this same study, germ-free mice were colonized with cultured bacterial strains from the stool microbiota of the severely stunted infants, and fed a diet with an addition of sialylated bovine (cattle-related) milk oligosaccharides. This resulted in a microbiota-dependent promotion of improved growth due to efficient utilization of nutrients in the infants (Charbonneau *et al*., 2017). These occurrences highlight the interrelationship between malnutrition, a need for developing treatment strategies, and the increased need for understanding the gut microbiome in a global and African context.

1.6 The burden of stunting in the global health context

There are several global development and sustainability goals which include; the eradication of malnutrition and mortality in children. The Sustainable Development Goals (SDGs), were adopted by all United Nations Member States in 2015 as a universal call to action to “end poverty, protect the planet and ensure that all people enjoy peace and prosperity by 2030” (UNDP, 2019). The United Nations International Children’s Emergency Fund (UNICEF) is responsible for seven global SDG indicators, which include “Under-5 mortality” and is a co-custodian for a further ten indicators, including “Stunting”. In order to track the development progress, measures of child undernutrition are taken to assess the overall progress towards the SDGs as illustrated in Figure 1.4 (p. 15). Sub-Saharan Africa is far behind other regions and the furthest from reaching a 50% reduction in stunting prevalence since 1990 (UNICEF, 2018). The progress in reducing child mortality has been accelerated in the 2000–2018 period, with the annual rate of reduction in the global under-five mortality rate increasing from 2% in 1990–2000 to almost to 4% in 2000–2018 (UNICEF, 2018). However, despite this progress, an estimated 5.3 million children under age five died in 2018 with about half of those occurring in sub-Saharan Africa, alone (Hug *et al*., 2019).

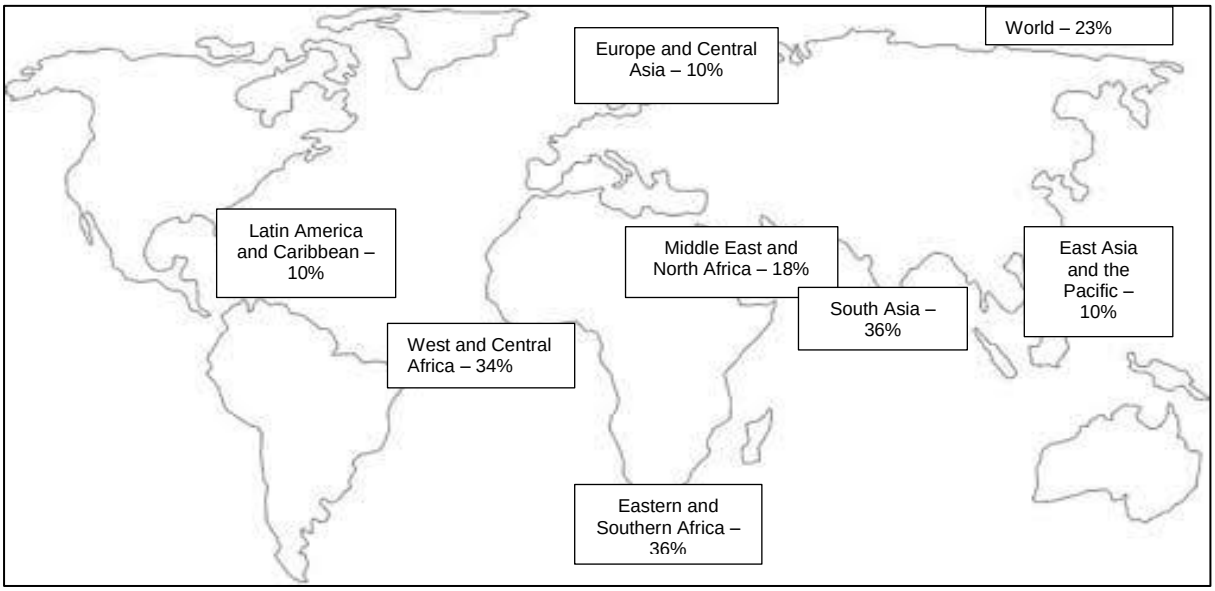


Figure 1.4: A world map illustrating the global frequency of undernutrition This contributes to nearly half of all deaths in children under five and is widespread in Asia and Africa. (UNICEF, WHO, World Bank Joint Child Malnutrition dataset, 2016 [https://data.unicef.org/topic/nutrition/ Accessed: 12:36pm 19 October 2019]).

The Department of Science and Technology (DST) of South Africa states that disease burdens the health system together with food insecurity and malnutrition, where 14% of the population are vulnerable to food insecurity and over 25% of children are stunted (Devereux and Waidler, 2017). The DST states that these conditions hamper the country's efforts to address its SDGs, particularly the goals of eradicating poverty and hunger, and reducing child mortality (Voluntary National Review Report, 2019). The persistent burden of disease and the abovementioned goal to address it highlights the need for extensive research into new, alternative and sustainable interventions. This reveals links to such alternative interventions, through new research using high-throughput technology such as next-generation sequencing. Overcoming the perpetuating cycle of poverty by investigating the role of the infant gut microbiome and nutrition to contribute to eradicating the burden of malnutrition, is in alignment to national interests.

1.7 Metagenomics

High-throughput comparative metagenomics has been enabled by next-generation sequencing (NGS) platforms, which advanced the understanding of the composition and function of microbial communities from various environments (Mardis, 2008; Caporaso *et al*., 2011; Huttenhower *et al*., 2014). NGS is one of the culture-independent techniques that has increased the feasibility and affordability of the characterization of microbial communities. NGS gives detailed information about the presence, identity, and amounts of microbes in different environments or samples (e.g. human faeces). This is performed without difficult and time-consuming cultivation techniques (Chen *et al*., 2013; Hart *et al*., 2015).

Collections of microbiome samples from different body sites must be considered when designing a study, and several issues to expand microbiome research are significant (Sinha *et al.,* 2015). One of the issues to consider is the method of collection which must preserve the microbial signature for each sample. Another consideration is that collection, transport and storage conditions must ensure the stability of the biological sample in suboptimal conditions. Storage of samples at -80 degrees Celsius is a standard protocol, however the long-term impact of freeze-thaw cycles on the intergrity of the sample is questionable (Sinha *et al*., 2015). Different sampling and collection tools are available, such as the QIAamp Powerfecal DNA kit (Qiagen, Hilden, Germany), which address the need for collection long-term stool storage and transport at room temperature conditions (Hart *et al*., 2015).

Several methods for the DNA extraction from stool samples from human hosts have been described, with a few in other species being reported (Li *et al*., 2003; Ferrand *et al*., 2014; Fiedorová *et al*., 2019). A study investigated the different common stool DNA extraction methods on different species and shows that the method of DNA extraction impacts: DNA concentration and purity; successful NGS amplification; and influences microbial communities seen in NGS output (Hart *et al*., 2015; Fiedorová *et al*., 2019). Limitations of the method of DNA extraction chosen can have adverse effects on the sequencing output by under- or over-representing specific microbial populations in different samples (Henderson *et al*., 2013; Kennedy *et al*., 2014). Faeces are known to contain various polymerase chain reaction (PCR) inhibitors that may affect the PCR reactions used to generate 16S rRNA amplicons; although these inhibitors vary and kit-based DNA extraction methods may resolve these issues (Fiedorová *et al*., 2019). This highlights the importance of DNA extraction methods used and the interpretation of data (Hart *et al*., 2015).

Positive and negative controls are important for use as standard reference materials in an experiment. Ideally, positive controls would include taxa representing site-specific or different

anatomic sites, to mimic a typical population's microbial community (Sinha *et al*., 2015). These controls would need to consider taxa with a wide variety of GC contents, staggered ratios and, absolute quantition of the number of genomes included. These considerations apply to negative controls, which additionally use blanks for extraction and sample handling. The benefits of using controls are the ability to normalize across batches and between experiments, enabling reproducibility of results (Sinha *et al.,* 2015). The lack of use of controls may lead to limitations in reliability of the data, whereby identifying and eliminating potential contaminants in reagents would not be possible. These contaminants sometimes vary in abundance and are able to mask true biological signals from low biomass samples (Sinha *et al.,* 2015; Fiedorová *et al*., 2019). Challenges to the development of adequate controls are met in a limited resource setting of a pilot study. In addition, the difficulty in the task of detection of these contaminants may be dependent on specific sequences, as well as the underlying community concentration, and varied bioinformatics analysis (Sinha *et al*., 2015).

The 16S rRNA gene

Most studies conducted have elucidated bacterial population dynamics by using metagenomic approaches that produce a high resolution picture of bacteria-host interactions, which may be implicated in disease (Franzosa *et al*., 2015). Whole-genome shotgun analyses perform unrestricted sequencing of the genome of all microorganisms present in a sample; while the 16S rRNA gene can be used to sequence full-length or targeted domains of bacteria and archaea to make biological inferences (Janda and Abbott, 2007). Although shotgun metagenomics offers the advantage of species- and strain-level classification of bacteria, its high cost and more demanding analysis requirements, have made 16S sequencing a more cost-effective strategy for microbiome analysis on a wide-scale (Morgan and Huttenhower, 2012; Walsh *et al*., 2018).

The 16S gene forms part of the 30S subunit of the ribosomes found in prokaryotes (Figure 1.5, p. 18). This gene region encodes several highly conserved and hypervariable regions (Figure 1.6, p. 18), and hypervariable regions are useful in species-specific identification (Patel, 2001; Claesson *et al*., 2010). The exclusiveity of this gene to prokaryotes highlights the advantage of eukaryotes not being amplified by sequencing. This gene is approximately 1500 base pairs in length, and full length sequencing amplifies this entire gene. This approach provides increased taxonomic resolution and microbiota identification, and accurately resolves nucleotide substitutions. (Johnson and Spakowicz, 2019). However, it is less feasible in a resource limited setting as it carries a higher cost than the alternative of targeted sequencing.

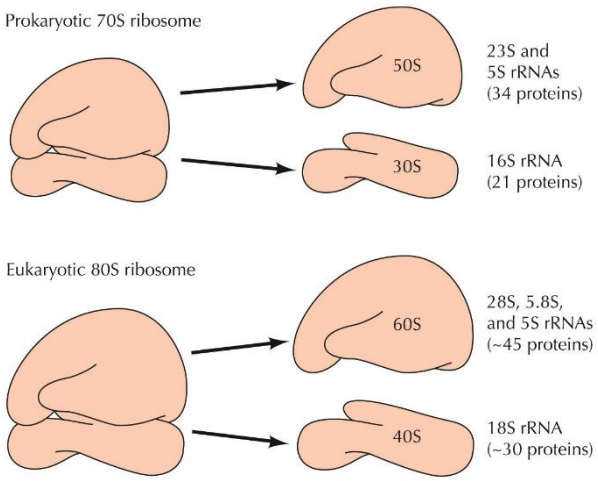


Figure 1.5: An image illustrating the composition of the prokaryotic and eukaryotic ribosome The 16S rRNA gene forms part of the 30S subunit of the prokaryotic ribosome, and is unique to this group.
(http://faculty.samford.edu/~djohnso2/44962w/405/_06translation.html).

The targeted approach sequences specific regions amongst the nine hypervariable regions (Figure 1.6, p. 18). The sequenced domains are generally the V3 and V4 regions, which function as the most taxonomically informative regions and as genetic markers for bacterial identification (Claesson *et al*., 2010). This region has a maximum sequence heterogeneity and, it shows the greatest discriminatory power compared to other hypervariable regions (Johnson, 2017).

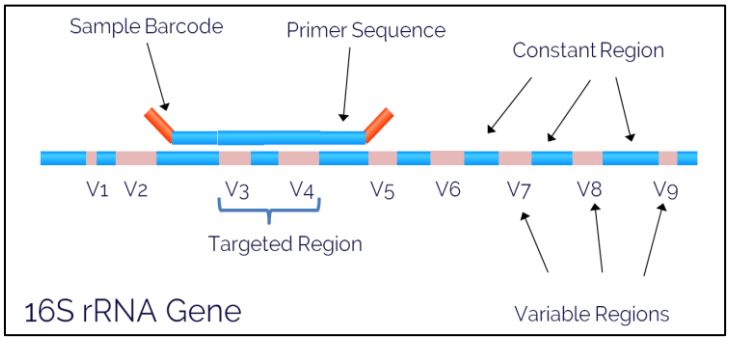


Figure 1.6: An image illustrating the 16S rRNA gene region. Constant regions refer to conserved regions, and variable regions refer to hypervariable regions. This image highlights the V3-V4 region targeted for sequencing of infant gut bacteria.
(https://www.lcsciences.com/discovery/applications/genomics/16s-mobile/).

Primers that target regions V1–V2 fail to amplify important bacterial groups, such as Bifidobacterium. A comparative analysis with full length Sanger sequencing of the 16S rRNA gene supports the use of primers that amplify the V3 and V4 region (Liu *et al*., 2008). In a

recent study, the number of core taxa in sequenced infant stool samples was higher in the V3-V4 region, and had an increased ability to delineate Bifidobacterium species (Biol-Aquino *et al*., 2019). The use of varied targeted regions and PCR primers in different microbiome studies represents a disadvantage, as it increases difficulty in making global comparisons (Johnson 2019).

Analysis of the data generated by these approaches requires computational methods and hardware resources (Gevers *et al*., 2012). This has been made accessible and feasible through open-source software of user-friendly bioinformatics tools and pipelines. This includes – but is not limited to – the MiSeq platform for 16S or metagenomics, where previous studies have demonstrated the strengths and weaknesses of this sequencing platform (Bokulich *et al*., 2013; Chen *et al*., 2013; Jovel *et al*., 2016; Biol-Aquino *et al*., 2019). With the data derived from the approach, one can perform taxonomical classification of bacterial sequences and assessments of bacterial diversity within (alpha diversity) and between (beta diversity) samples. As well as infer associations of the bacterial microbiome with the host and external influences.

For the analysis of NGS data, a series of steps are followed to ensure the derivation of biologically significant and reliable results, in a replicable manner. This requires the initial pre-processing of NGS data to remove uninformative data, such as low quality bases for the effective analyses of NGS sequences. A number of user-friendly tools have been developed for this purpose, with default parameters provided for setting a quality threshold such as Cutadapt (Martin, 2011) and fastqMcf (Aronesty, 2011). Taxonomic classification of bacterial sequences based on sequence alignments can represent a challenge because of the short NGS read lengths generated. Hence prior to classification, approaches to cluster the regions of bacterial 16S rRNA genes for example are performed by various methods.

These methods include: referencing to previously annotated databases by phylotypes (grouped by certain relative level of similarity of the rRNA gene sequence) (Liu *et al*., 2008); or by using Operational Taxonomic Units (OTUs) that are constructed by clustering sequences *de novo*, purely based on similarity (Schloss and Westcott, 2011; Sun *et al*., 2012). Hybrids of these methods are recommended and a 99% similarity threshold is generally accepted for discriminating between species. However, some closely related species are more difficult to differentiate and require higher resolution analytical tools to overcome this limitation (Stackerbrandt and Ebers, 2006; Eren *et al*., 2013; Tikhonov *et al*., 2015). An additional method using amplicon sequence variants (ASVs) is termed “zero noise OTUs” or “sub-OTUs” (Edgar, 2016; Amir *et al*., 2017). This controls errors derived from

OTUs sufficiently such that ASVs can be resolved exactly, down to the level of single-nucleotide differences over the sequenced gene region (Callahan *et al*., 2017).

Reference databases are used for the annotation of bacterial 16S rRNA genes and include the Greengenes database (DeSantis *et al*., 2006), the Ribosomal Database Project (Cole *et al*., 2014), and SILVA (Quast *et al*., 2013). Along with these databases that contain a vast amount of information, are a series of designated bioinformatics tools for the analyses of the NGS sequences. An approach used to increase the resolution of classified sequences and minimize the information, is using databases specific to a single environment under study, for example the human gut microbiota (Ritari *et al*., 2015; Forster *et al*., 2016). For subsequent analyses of 16S amplicon libraries generated, several pipelines are used: including the most widely adopted QIIME (Caporaso *et al*., 2010; Navas-Molina *et al*., 2013); as well as MiSeq Reporter v2.5 (developed by Illumina and MiSeq instrument). In a study by Jovel and colleagues (2016) evaluating the analysis of 16S amplicon libraries, the QIIME pipeline observed a higher relative abundance than expected. Therefore was highlighted as the most suitable over two other pipelines for subsequent analyses at the genus level.

Diversity is a standard measurable outcome used in microbiome research, defined as the number, abundance, and distribution of microorganisms (Gevers *et al*., 2012). The collective genes of gut bacteria can characterize the abundance and diversity of species (Turnbaugh *et al*., 2007), with human stool being the most viable biological sample for these investigations (Hagerty *et al*., 2020). Alpha and beta diversity captures the diversity of species within and between samples (Jost, 2007). Currently, there is not yet a gold standard for empirically based methodology in measuring diversity. Different metrics (e.g. observed OTUs, Shannon, Unweighted Unifrac distances) are used to observe: the richness (i.e. number of different kinds species) and evenness (i.e. the extent to which the species are in even abundance with each other) for alpha diversity (Hagerty *et al.*, 2020); and the dissimilarity between samples of communities of bacteria from different subjects (Chu *et al*., 2017). Mathematical expressions are used to derive indices that measure the statistical significance of these diversity measures (Chiarucci *et al*., 2011).

In addition to statistical significance of relative abundance measures, various visualizations of these tests are derived from these diversity measures. That includes: box-and-whisker plots, differentiating groups of data corresponding to their estimates of diversity; scatter plots illustrating each sequenced sample of data computed using different diversity metrics; and bar plots that highlight the most taxonomically differential abundance between groups of

data. These visualizations are derived according to the bioinformatics tools used and use of the pipelines stated above.

1.8 Considerations of confounding factors on microbial diversity

Confounding is a type of bias that can negatively influence the results of epidemiological studies, by potentially leading to erroneous conclusions. It can be considered the confusion of the effect of the exposure on the outcome, with the effects of other factors (confounders) that influence the outcome (Rothman *et al.,* 2008; Howards, 2018). External factors present an effect on the diversity microbiota, and contribute to the variability between individuals' gut microbiomes. Research has shown these confounders include gestational age at birth (Dogra *et al.*, 2015), mode of delivery (Jakobsson *et al.*, 2014), infant age at sampling, geography and culture (Yatsunenko *et al*., 2012), breastfeeding and early dietary patterns (Bergström *et al*., 2014; David *et al*., 2014), and the use of antibiotics (Dethlefsen *et al*., 2008).

Confounders in microbiome research are accounted for through procedures for reducing inter-individual variability and adjustments of statistical testing (Miyoshi *et al*., 2018; Chu *et al.,* 2017). Being aware of the assumptions and limitations of a statistical test, reduces the likelihood of more error-prone results – for example, the analysis of composition of microbiomes (ANCOM) test assumes that fewer than 25% of features differ between groups, and should not be used if a difference greater than 25% is expected (Mandal *et al*., 2015; Estaki *et al.,* 2020). The main assumption in all ordination methods for measuring diversity is that there are a limited number of factors that greatly influence distribution and relative abundance of species (Jovel *et al*., 2016).

An additional way to adjust for confounding is through multinomial regression designed for compositional microbiome data (Estaki *et al.,* 2020). This is used to explain the relationship between one nominal dependent variable and one or more independent variables. A tool such as Songbird (Morton *et al.*, 2019), identifies differentially abundant features while controlling for variables e.g. controlling for antibiotic exposure when identifying features that are significantly different between babies born vaginally versus through C-section (Estaki *et al*., 2020).

Limitations to not controlling for confounders introduce bias. Bias may preclude finding a true effect and may lead to an inaccurate estimate (underestimate or overestimate) of the true association between exposure and an outcome. Significance testing in itself does not take into account factors which may bias study results (Skelly *et al*., 2012). However, some

statistical tests adjust for multiple testing, which may increase the reliability of a highlighted relationship between an exposure and outcome.

1.9 Relevance of the research study

Although research has characterized the gut microbiome of various population groups and demographics, more is needed to determine whether there is a universal composition for determining a healthy status, or if microbial therapeutics should be individualized. Fundamental questions are emerging with the increase in gut microbiome research, due to the lack of clarity still remaining as to whether this information can be used to prevent and treat illness from an early onset. Such questions previously highlighted are, whether live microbial therapeutics are necessary for treating illness? Alternatively, are the by-products of bacterial metabolism sufficient for effective treatment? There are also ethical considerations of therapeutic modifications and the unforeseen consequences of potential adverse effects (McDonald and McCoy, 2019). The dilemma of informed consent with respect to such modifications is also a question in the African context, where the limitations of translating and disseminating biological research to local languages and the community arises.

Gut microbiome research is emerging as a field of significant inquiry in Africa, with consortia taking advantage of genomic technologies towards actionable improvement of health outcomes (Brewster *et al*., 2019). Given the need to address these challenging issues and gaps in the research, and to align scientific research in the South African context, an insight into the factors relating to the interactions of the gut microbiota – and linking them to the functionality of early intestinal colonisers to specific diets – can lead to a greater understanding. The data presented in this study aims to characterise the bacterial composition of the infant gut in a South African community. This is in an effort to identify early infant gut microbial communities consistent with different feeding practices and maternal secretor status at critical time points, which may significantly influence corresponding growth outcomes in the infant. This information could better inform further functional research to produce therapeutics (e.g. better formula feeds) which could be useful in improving stunting phenotypes and related growth at critical time points in South African infants.

1.10 Aims and objectives

The aim of this pilot study is: (1) to characterize the gut microbiome profile of South African infants living in an urban environment (Soweto) over time (2) to associate infant gut microbiome profiles with exclusive breastfeeding, mixed feeding, and formula feeding

practices (3) to investigate the influence of maternal *FUT2* genotypes (secretor status) on the infant gut microbiome.

This will be achieved by addressing the following objectives:

Objective 1: To recruit and enrol participants in the study following informed consent.

Objective 2: To collect biological samples (infant stool samples, maternal saliva samples) from participants and infants following informed consent.

Objective 3: To generate metagenomic data (16S rRNA) using DNA extracted from infant stool samples at four time points ("at recruitment" i.e. 12-21 weeks; six months; nine months; and twelve months).

Objective 4: To screen mothers for mutations in the *FUT2* gene (maternal genotype) and determine secretor status.

Objective 5: To perform a bioinformatic and data analysis in all infant stool samples (highlighting time points "at recruitment" and nine months old); and to associate feeding practices and maternal secretor status with the infant gut microbiome composition.

CHAPTER TWO – MATERIALS AND METHODS

A brief outline of the flow of research project method is shown in figure 2.1 (p. 24).

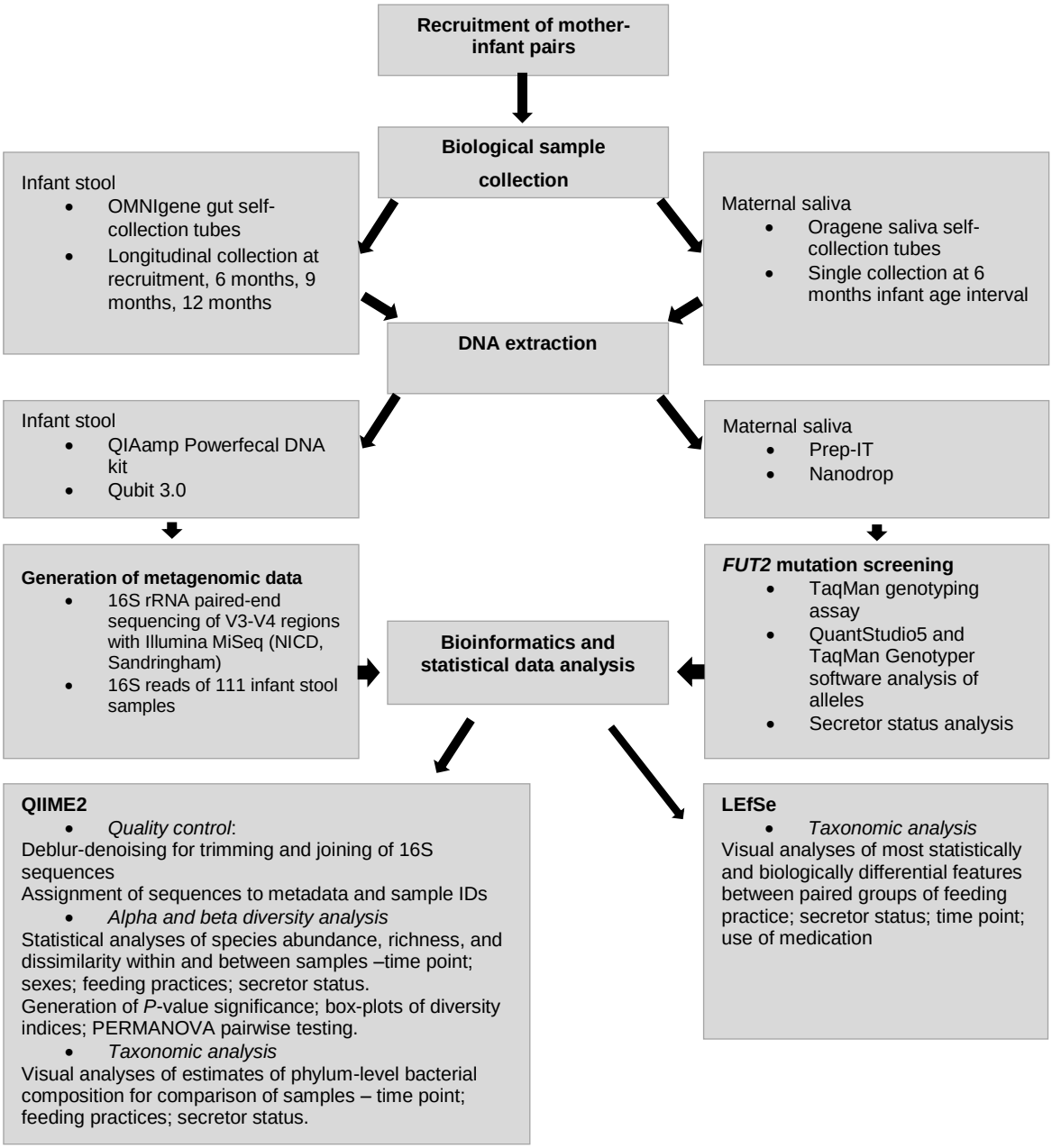


Figure 2.1: A flow chart indicating materials and methods briefly outlining the analyses performed and tools used.

2.1 Participant recruitment (Mother-infant pairs)

We collaborated with the Developmental Pathways for Health Research Unit (DPHRU) regarding the recruitment of mother-infant pairs to be enrolled in the microbiome study (Table 3.1, p 40). The mothers were recruited at the maternity services of the Chris Hani-

Baragwanath Academic Hospital in Soweto, Johannesburg, for the “Soweto Baby Water Sanitation and Hygiene (WASH) Study” (Ethics clearance certificate number M170753 – Title: *Interaction between nutrition, infection, household environment and care practices and their impact on growth and development in infants between birth and one year of age*). The recruitment was conducted post-delivery, with inclusion factors for the mothers following the WHO Multicentre Growth Reference Study criteria (de Onis *et al*., 2004) – which are mothers who were HIV-negative, non-smokers, without morbidity, with a term singleton pregnancy. Recruitment and sample collection for the microbiome study followed ethics clearance for the collection of biological data (Appendix N, p. 157).

The recruitment process was performed by the researcher and three research assistants – one hired and two from the Soweto Baby WASH (SBW) team. From June to August 2018, mothers in Soweto were recruited in their homes and enrolled after signing informed consent sheets (Appendix A, p.112). Given that this was a pilot study to assess the feasibility of microbiome research in an urban South African setting; we enrolled as many individuals as we could within the three month recruitment period. Forty-eight mothers were recruited, and two withdrew consent for a final n = 46. Information on the feeding practice – whether exclusively breastfeeding, mixed feeding, or formula feeding – and the duration of exclusive breastfeeding was gathered directly from the mothers through questioning by the researcher.

Details regarding use of medication was collected from the REDCap database (Harris *et al*., 2019), where questionnaires conducted by the SBW team were captured. The specific medication for instances of sickness reported in the child (e.g. coughing, fever, vomiting), were not captured from the REDCap database questionnaire. The reference to medication in this study includes simple medicine for flu (e.g. paracetamol/panadol), which were either prescribed at a local health clinic or given to infant by the mother. It also includes home remedies, vaccinations, and unspecified antibiotics. The most significant potential confounder on the diversity of the samples would be from antibiotics (Dethlefsen *et al*., 2008; Laursen *et al*., 2015; Azad and Ben, 2017). The research by Chu *et al*. (2017), demonstrated that mode of delivery only affects the early neonatal period (first six weeks). The youngest infant recruit was past this period, therefore this information was not captured for the study.

2.2 Mother-infant biological sample collections

The mothers provided consent for collection of biological samples for themselves and their child (Table 3.2, p. 41). Infant stool samples together with maternal saliva samples were collected for this study. These samples were anonymised by using the participant identification numbers assigned to mother-infant pairs, during recruitment for the SBW

Study. An example of how these samples were anonymised yet correlated between mother and infant was: *maternal ID – 00-001 or 00-002, infant ID – 00-101 or 00-102*, respectively.

2.2.1 Infant stool sample collection

Infant stool samples were collected using OMNIgene gut self-collection tubes (DNA Genotek Inc., Ottawa, Ontario Canada) at four time-points: at recruitment (12 - 21 weeks old), six months (~24 weeks), nine months (~36 weeks) and twelve months (~48 weeks). This was to enable the characterization of the infant gut microbiota profile over a longitudinal period. These time points differ from the information sheet (Appendix A, p. 115), due to ethics clearance for sample collection being obtained after the initial collection points stated in the approved forms for participant recruitment. As stated in the information sheet, collections of samples were done after explanation of procedures by a research assistant (and the researcher). Therefore mothers were informed of the adjusted time points before consent was obtained.

The tubes used enabled easy collection and contain a buffer for the stabilisation of stool samples for gut microbiome profiling; by rapidly homogenizing and stabilizing the sample at the time of collection, and preventing the growth of other microorganisms. It enables storage at room temperature for up to 60 days and was chosen to preserve microbiota DNA in the *in vivo* state, ensuring that high quality DNA could be obtained for 16S rRNA microbiome profiling. The recommended starting amount for each sample (0.25g of faeces – demonstrated as the equivalent of the size of a fingernail to participants) – was homogenized in a tube containing one metal bead for lysing of host cells and microbial cells as shown in Figure 2.2 (p. 27). A benefit associated with this method of sample collection is the improvement of donor compliance, due to the simple user-friendliness. As a result mothers were given kits and were able to collect samples by themselves in the privacy of their own homes and these samples would be picked up at a later visit.

Samples were to be collected without urine, however this is difficult to separate in an infant diaper and may have affected results. Mothers were instructed to keep samples in a cool place at room temperature conditions. However, room temperature may increase indoors, and participants may not have complied in the absence of observation, and this may have affected results which was beyond the scope of the study. Once collected by the researcher, the tubes were stored at room temperature at the DPHRU, and later transported to the National Health Laboratory Services (NHLS, Johannesburg) within the 60-day limit, where subsequent DNA extractions took place. Prior to a batch extraction, the stool samples were aliquoted into tubes and stored in a -80°C freezer – allowing a “snapshot” of the microbial

profiles. This step was performed in an attempt to minimize potential confounders of variability, so that prior to extractions all samples were treated uniformly.

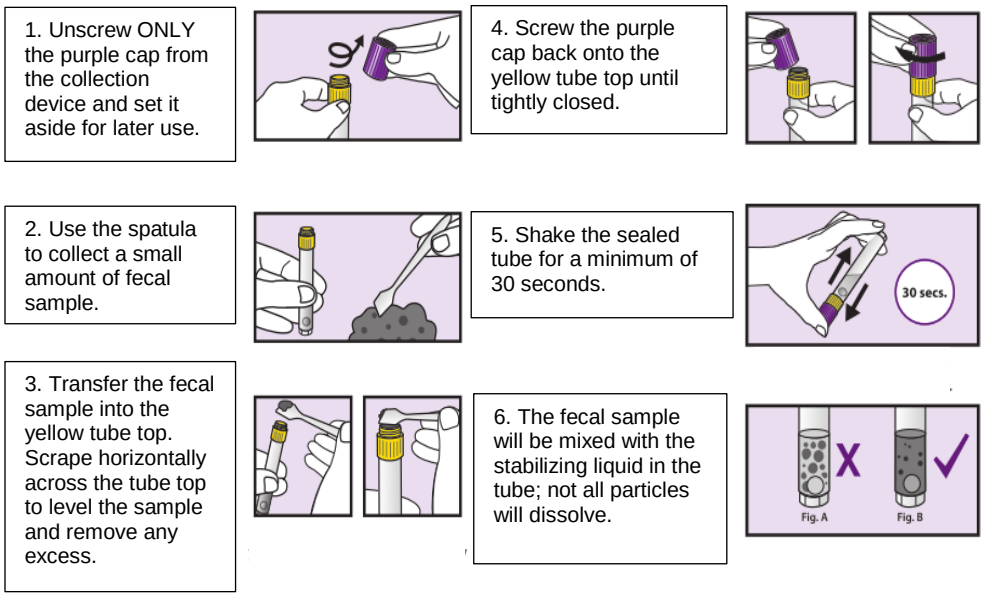


Figure 2.2: The method of stool sample collection and stabilization using the OMNIgene Gut self-collection tubes (DNA Genotek Inc., Ottawa, Ontario Canada, 2017). Infant faecal samples were collected by researchers and participants using these tubes and stored for subsequent analyses (https://www.dnagenotek.com/ROW/pdf/PD-BR-00159.pdf).

2.2.2 Maternal saliva sample collection

Saliva samples were collected from participating mothers using the Oragene saliva self-collection tubes (DNA Genotek Inc., Ottawa, Ontario Canada). The Oragene self-collection kits (Figure 2.3, p. 28) allow for the collection, preservation, transportation and purification of DNA from saliva. This collection was performed at a single time point (infant 6 months collection time point). This DNA sample collection kit (which is stable at room temperature) was favoured due to the high DNA yield, non-invasiveness, and quality of the DNA collected.

These collection kits were distributed to mothers, following the demonstration of how the samples were to be collected by the research team. Once the samples were collected, they were stored at room temperature at the DPHRU, and later transported to the NHLS for the subsequent DNA extractions.

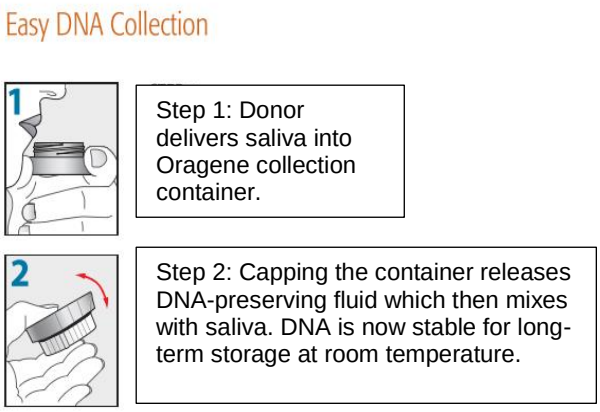


Figure 2.3: **Saliva sample collection using the Oragene saliva self-collection tubes** (DNA Genotek Inc., Ottawa, Ontario Canada, 2006). Maternal saliva samples were collected by participants using these tubes and stored for subsequent analyses (DNA Genotek, 2006).

2.3 Generation of metagenomic data

2.3.1 DNA extraction

The extraction of DNA from the infant stool samples was performed using the QIAamp PowerFaecal DNA kit (Qiagen, Hilden, Germany). This process involved the purification of microbial DNA using QIAamp spin columns, through the digestion of proteins, binding DNA to the silica membrane, washing away of impurities and the elution of pure genomic microbial DNA from the spin columns. By using the Quick-Start Protocol (Qiagen, Hilden, Germany, 2016) provided in the kit, the method was optimized for the extraction of DNA to increase the yield (Appendix B, p. 118). A limitation of this procedure was the lack of use of controls due to resource restrictions, as collection and extraction kits were expensive and each tube used in extraction procedures was needed for a sample. This limited identifying and eliminating potential contaminants in reagents; however consultation with colleagues from Standford University (USA) assured the quality use of these kits for microbiome research. Following DNA extraction a visual inspection of DNA quality was carried out by performing gel electrophoresis of the samples on a 1% (w/v) agarose/ethidium bromide (EtBr) gel at 180V/cm for one hour, visualized under UV light and photographed on the Vacutec Gel Doc Omega Fluor Aplegen System.

2.3.2 16S rRNA sequencing

Following the DNA extraction and quality check processes, samples were quantified and prepared for next-generation sequencing (NGS). The samples were quantified using the Qubit double-stranded DNA high-sensitivity assay kit (Qubit™, Thermo Fisher Scientific, Wilmington, Delaware USA), on the Qubit® 3.0 Fluorometer (Thermo Fisher Scientific,

Wilmington, Delaware USA, 2014). The Qubit concentrations of the samples were all normalised to a working concentration of 10ng/µl in a final volume of 12µl.

Amplification of the microbial 16S rRNA gene (Weisberg *et al*., 1991) (used in phylogenetic classifications of diverse microbial populations), was carried out at the National Institute for Communicable Diseases (NICD, Sandringham, Johannesburg) sequencing core facility. Here, total PCR amplification of the V3 and V4 genomic regions of the 16S rRNA gene was performed and sequenced on an Illumina MiSeq (Illumina Inc., San Diego, California USA), according to standard manufacturer protocols using the primers in Table 2.1 (p. 29). The sequencing coverage was 300X, and in paired-end sequencing format. The concentration and samples IDs were provided to NICD sequencing facility together with the samples. Following 16S rRNA sequencing, raw sequencing data was then uploaded via a shared Google Drive from the service providers. The raw sequencing data/reads were files in a fastq.gz format, which are compressed text files with biological sequence and quality scores, with sequence labels for forward and reverse reads.

Table 2.1: NICD Illumina MiSeq primer sequences used to amplify the V3-V4 region of the 16S rRNA gene

Primer	**Sequence**
Forward	5'- CCTACGGGNGGCWGCAG -3'
Reverse	5'- GACTACHVGGGTATCTAATCC -3'

2.4 *Fucosyl transferase 2* (rs601338) mutation screening

2.4.1 DNA extraction

DNA from the maternal saliva samples (n = 46) was extracted using the prep-IT (DNA Genotek Inc., Ottawa, Ontario Canada) kit in conjunction with the Oragene saliva self-collection kit. The prep-IT DNA solution was used to maximise the recovery of DNA from the saliva samples. A pre-purification check of the samples was performed as per the recommended guidelines (DNA Genotek Inc. prepIT•L2P protocol, 2015) – by measuring the total weight of the saliva in the tube and subtracting the weight of the unused Oragene tube – due to the amount of saliva collected being directly proportional to the amount of DNA recovered. DNA extraction was performed according to the manufacturers' instructions (DNA Genotek Inc. prepIT•L2P protocol, 2015), and the extracted DNA was then stored at -20°C.

To assess the quality of the DNA extracted, DNA samples (1µl) were electrophoresed on a 2% (w/v) agarose/ethidium bromide gel (SeaKem® LE Agarose, Lonza Rockland Inc.,

Rockland Maine, USA) at 90V/cm. The quality and concentrations of the DNA samples were assessed spectrophotometrically using the NanoDrop®ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, Delaware USA) (Desjardins and Conklin, 2010). Following this, dilutions of DNA samples were made to 10ng/µl as recommended for the downstream genotyping assay.

2.4.2 Genotyping assay and analysis

Genotyping of the *FUT2* gene mutation was carried out using a real-time PCR custom predesigned TaqMan SNP Genotyping Assay (Thermo Fisher Scientific, Wilmington, Delaware USA). This gene variant predicts the secretor status in mothers – determined by the presence of a single nucleotide polymorphism (SNP) rs601338 in either the alternative (G-allele) or polymorphic (A-allele) state. This assay uses a TaqMan 5'-nuclease chemistry for the amplification and detection of a 'G' or 'A' SNP in the maternal DNA samples. The predesigned TaqMan assay contained a *FUT2* sequence-specific forward, and reverse primer, as well as VIC-labelled (G) and FAM-labelled (A) minor groove binding probes with non-fluorescent quenchers. The primers sequences are proprietary, therefore the context sequence as per "minimum information for the publication of real-time quantitative PCR experiments (MIQE)" guidelines (Bustin *et al.,* 2011) is shown in Table 2.2 (p.30).

Table 2.2: The TaqMan assay Thermofisher context sequence used to amplify the *FUT2* (rs601338) SNP region

Sequence
CGCTTCACCGGCTACCCCTGCTCCT[A/G]GACCTTCTACCACCACCTCCGCCAG

A 96-well plate assay was prepared using the wet-DNA delivery method, and a minimum of 5ng/µl of DNA was aliquoted as per the manufacturer's instructions (Applied Biosystems, California USA, 2010). The analysis was performed at the Sydney Brenner Institute for Molecular Biosciences (SBIMB, Wits) using the QuantStudio S5 Real-Time PCR System (Life Technologies Corporation, California USA). The *FUT2* (rs601338) variant has a minor allele frequency (MAF) of 50% in the African subset of the Exome Aggregation Consortium Dataset (http://exac.broadinstitute.org/variant/19-49206674-G-A), so we expected to capture both secretors and non-secretors given our final sample size of 46 mothers. Inconclusive samples were repeated and reanalysed on the QuantStudio S3 Real-Time PCR System (Life Technologies Corporation, California USA) using the same cycling conditions. The resulting raw data (n = 43 conclusive samples) was analysed on the QuantStudio Design and Analysis Software and the TaqMan Genotyper Software (Thermo Fisher Scientific,

Wilmington, Delaware USA, 2015), for the auto-calling of data points and generation of allelic discrimination plots of the assayed samples, to determine the maternal genotypes.

2.4.3 Sanger sequencing

A subset of samples (10%) was selected for Sanger sequencing to validate the genotypes obtained from the Taqman assay. This subset represented three of each of the genotypes obtained from the genotyping assay. The primer sequences (Inqaba Biotechnical Industries (Pty) Ltd, Johannesburg, South Africa) that were to be used to amplify the *FUT2* gene region, with the SNP of interest, were designed using the online Primer1 program (Ye *et al*., 2001) and are listed in Table 2.3 (p. 31). The primer stock was made up to a concentration of 100µM.

Table 2.3: Sanger sequencing primer sequences used to amplify the *FUT2* (rs601338) gene region

Name	Sequence	Length(bp)	Tm min\max	GC % min\max
FUT2_F	GTGGACGATCAATGCAATAGGC	22	54.8	50.0
FUT2_R	CTTCCACACTTTTGGCATGACA	22	52.9	45.5

FUT2_F/R = fucosyltransferase2 forward/reverse; bp =base pairs; Tm = melting temperature; GC = guanine-cytosine content

PCR amplification was performed using the AmpliTaq Gold DNA Polymerase kit (Thermo Fisher Scientific, Wilmington, Delaware USA) with final PCR conditions given in Table 2.4 (p.32). The Thermo Scientific cycle sequencing protocol (2016) (BigDye v3.1) was used for cleaning up samples, and the final products (n=18 forward and reverse PCR products) were loaded onto a 96-well plate. Sanger sequencing was performed on the Genetic Analyser 3500xl using a Z- dye set for spectral calibration, and analysis was performed on the BioEdit software program (Hall, 1999).

Table 2.4: The AmpliTaq Gold DNA Polymerase kit components with optimized volumes for PCR amplification, cycle sequencing conditions and DNA visualization for Sanger sequencing of *FUT2* genotype samples

Component	**Final Concentration**	
10X PCR Buffer II	1X	
10mM dNTP Mix	0.2mM	
10µM F primer	0.2µM	
10µM R primer	0.2µM	
AmpliTaq Gold DNA Polymerase (5U/µl)	1.25U/50µl reaction	
25mM $MgCl_2$	1.5mM	
dH_2O	-	
Template DNA	<1µg/reaction	
Cycling conditions		
Initial Denaturation	95°C for 5 minutes	
Denature	95°C for 30 seconds	30X cycles
Anneal	58°C for 30 seconds	
Extend	72°C for 30 seconds	
Final extension	72°C for 5 minutes	
Hold	12°C indefinitely	
Visual inspection of DNA		
1% (w/v) agarose/ethidium bromide gel; 90V/cm; 1.15 hours; 5µl PCR product with 5µl Ficoll Dye		

2.5 Bioinformatics and data analysis

The bioinformatic analysis of the infant microbiome was performed for the characterization of gross taxonomic classifications of microbiota present. The 16S rRNA read data (n = 114; where n = 111 were sent for sequencing plus 3 repeated samples) received from the NICD was analysed using the "Quantitative Insights Into Microbial Ecology 2" (QIIME2) analysis pipeline software (Bolyen *et al*., 2019). By following the online user documentation (www.qiime2.org – version 2018.4) indicating the flow of microbial data processing, this pipeline enabled the pre-processing of reads, phylogenetic tree generation, and diversity measures of the infant microbiome samples. This was performed with the aim of investigating associations between the infant gut microbiome profile, with maternal secretor status, infant feeding practices, and microbiome diversity over time – specifically, samples collected "at recruitment" (12-21 weeks) and the nine month time points. These time points were highlighted due to transitions in infant feeding practices over time – from **liquid** (breast milk or formula feeds) **to solid foods** (more complex "adult-like" diet) – which was postulated to have an impact on gut microbial diversity and composition. The QIIME2

command-line interface and plug-ins were accessed for this process using the SBIMB/University of the Witwatersrand cloud server – Wits Cluster.

2.5.1 QIIME2

This software was navigated using the "Moving Pictures" tutorial, accessible at https://docs.qiime2.org/2018.4./tutorials/moving-pictures. In order to process the 16S sequences and test for associations, a metadata mapping file containing the relevant information was created for the study (Appendix C, p. 119). This tabular file was validated for use in QIIME2 using "Keemei" – a Google Sheets Add-on used in bioinformatics analyses (Rideout *et al*., 2016). The raw sequences generated by the NICD sequencing facility were imported to the Wits Cluster from the Google Drive folder, for use in QIIME2.

I. Quality control of sequenced data

Using the Casava 1.8 paired-end demultiplexed sequence format (Eren *et al.* 2013), the fastq.gz files of raw sequence data were imported and demultiplexed from the Wits Cluster to a QIIME2 "artifact" – this contains information about the type and source of the data, with a ".qza" extension. The summary information obtained of the demultiplexed forward and reverse reads determined how many sequences were obtained per sample, as well as the sequence quality at each position. A ".qzv" extension was generated from the artifact to visualize the data on QIIME2View (https://view.qiime2.org). Using the "cutadapt" plug-in (Martin, 2011) the primer sequences used for 16S sequencing were trimmed off the demultiplexed forward and reverse reads, and then corresponding reads were joined producing a visualization and interactive plot of quality scores (see Appendix D, p. 127).

To produce high quality sequence variant data, the Deblur workflow was used. Deblur (Amir *et al*., 2017) provides a rapid and sensitive method of investigating ecological patterns driven by differentiation of closely related microorganisms. The initial step of the workflow was the quality control of joined reads using a method of implementation of the quality filtering approach described by Bokulich *et al*. (2013). The quality filtering was based on q-score, using default parameters. This derived an output table of statistics of the trimmed-demultiplexed-paired-end-joined sequences, indicating whether the majority of sequences were retained following the read-joining (Appendix D, p. 127). The preceding step was the "denoising" option of the Deblur workflow for the elimination of non-similar length sequences. This uses a single parameter "p-trim-length n", where "n" is the position to truncate the trimmed-demultiplexed-paired-end-joined sequences. By use of the joined sequence interactive quality plot showing quality scores, this value was chosen as n=401, at the point where the median quality score began to drop too low i.e. ≤ 28, and was close to the position

at which the sequence quality became disorderly as illustrated in Figure 2.4 (p. 34, see Appendix D, p. 127). The data output from this process were artifacts of representative sequences, a feature table and related statistics of the retained reads.

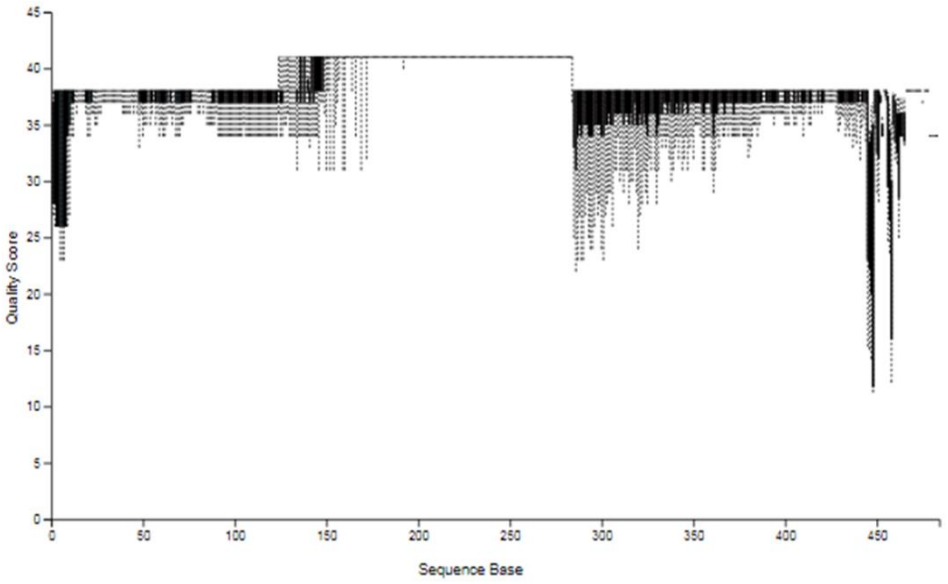


Figure 2.4: **An interactive quality plot illustrating the trimmed-demultiplexed-paired-end-joined sequences** These are important for choosing the truncating position "n" for denoising using Deblur. The plot shows that after 400 sequence bases, the quality score decreases dramatically and the reads are disorderly.

Feature Table and Feature Data

After completing the quality filtering step with Deblur, visualizations of the representative sequences and a feature table were created for information on how many sequences were associated with each sample and with each feature – by incorporating the metadata mapping file – histograms of those distributions, and some related summary statistics. This process also provided a mapping of feature identities (IDs) to sequences with links to easily use the Basic Local Alignment Search Tool (BLAST) on each representative sequence against the National Center for Biotechnology Information (NCBI) nucleotide database (Altschul *et al*., 1990).

Phylogenetic Tree generation

In order to observe meaningful patterns in microbial diversity in downstream analyses, a phylogenetic tree – relating the features of the data to one another – was generated. This was performed by using the pipeline through the alignment of multiple sequences in the Feature Data [Sequence] to create Feature Data [Aligned Sequence] using the *mafft* program (Katoh *et al*., 2002); masking the alignment to remove highly variable positions

which add noise to the tree; applying *FastTree* (Price *et al*., 2010) to generate the tree from the masking alignment; lastly, applying midpoint rooting to the unrooted tree previously generated. This phylogenetic tree is required for analysis of several diversity metrics which are presented in this research.

II. Alpha and beta diversity analysis

Using the QIIME2-diversity plugin, the core-metrics phylogenetic method rarefies Feature Table [Frequency] to a user-specified depth, and computes several alpha and beta diversity metrics. Rarefaction is a method of calculating the species richness from samples, allowing you to compare taxonomic richness of two samples of unequal size by 'rarefying' the larger sample to the size of the small sample; where curves are generated and indicate a number of species as a function of a number of samples. The artifacts created were used as an input to R software version 3.6.1 (R Core Team, 2016) for generating principle coordinates analysis (PCoA) plots, which will be described further under beta diversity measures. PCoA is a scaling method using a distance matrix of differences of composition between samples to produce a 3-dimensional graph, and was used with the aim of identifying possible grouping/clusters of samples based on metadata features.

The important parameter in this process was “p-sampling-depth” which is the even sampling (rarefaction) depth. This randomly subsamples the sequence frequency counts from each sample to the value chosen for this parameter. This parameter was chosen following the recommendation from the tutorial by evaluating which sequences had the lowest frequency counts and could be lost. A sampling depth of 9956 was chosen, meaning sequences with a frequency count below this (three in total) were lost from the diversity analyses (Appendix E, p. 132). In addition, one sample-ID from the metadata file was not returned with the same ID in the sequence 16S read data. Therefore, four samples were lost and this resulted in a sample size of n=110 going forward.

Alpha diversity statistical analysis

Alpha diversity (Whittaker, 1960) refers to a local measure of biodiversity *within a sample*. The diversity metrics computed by the above-mentioned method were Faith’s Phylogenetic Diversity “PD” (Faith, 1992) and Pielou’s Evenness (Pielou, 1966). Faith’s PD is a qualitative measure of the species richness found within a sample, whereas Pielou’s Evenness assesses the community evenness within a sample, accounting for under/overrepresentation of a singular species (Pielou, 1966; Faith, 1992). We explored the microbial composition of the samples in the context of the metadata file by viewing the association between categorical metadata columns (namely: time point, sex, feeding practice and secretor status)

and the two alpha diversity metrics, using the Kruskal-Wallis (Kruskal and Wallis, 1952) non-parametric statistical test. This assesses the differences among three or more independently sampled groups on a single, non-normally distributed continuous variable (McKight and Najab, 2010). The outputs generated were box-plots and statistical *P*- and q-values indicating significance and correction of multiple testing accounting for missing data of the metadata columns investigated of time point, sex, feeding practice and secretor status.

The following analysis carried out assessed alpha diversity as a function of sampling depth, by alpha rarefaction plotting. At “p-max-depth” value of 23000 – chosen based off of the tutorial recommendations – the rarefaction curves computed multiple diversity metrics at multiple sampling depths between one and 23000, namely: Shannon diversity; observed operational taxonomic units (OTUs)/amplicon sequence variants and Faith’s PD. In QIIME2-Deblur, the pipeline attempts to reconstruct the exact biological sequences present in the sample i.e. amplicon sequence variants (ASVs) or "zero noise OTUs" (Callahan et al., 2017). The rarefaction curves generally grow exponentially, as the most common species are found, but eventually plateau upon rare species sampling (Gotelli and Colwell, 2001). Shannon Diversity (Shannon, 1948) is metric of diversity including “richness” (the total number of species) and “evenness” (the relative proportions of all species).

Beta diversity statistical analysis

Beta diversity (Whittaker, 1960) refers to a regional measure of biodiversity *between samples*. There were several diversity metrics computed by the above-mentioned method, for use in the generation of principle coordinates analysis (PCoA) plots (Gouwer, 1966) and performing PERMANOVA statistical tests (Anderson, 2001). The features investigated were: feeding practice, secretor status, sex, all time points, and two time points (“at recruitment” i.e. 12-21 weeks or “0” months, and 9 months). PERMANOVA is a non-parametric multivariate test, used for testing whether the distances between samples from within a group are more similar than they are to distances from other sample groups. This statistical test was used for pairwise analysis of different metadata columns – namely feeding practice; secretor status; and time point – and an output of distance graphs and correlating significance values were generated.

The beta diversity metrics computed for use in the above analyses were Jaccard distance (Jaccard, 1902) and Bray-Curtis distance (Bray and Curtis, 1957) – which are qualitative and quantitative measures of community dissimilarity, respectively – where Jaccard is based on presence or absence of species, and Bray-Curtis considers taxa presence or absence as well as abundance of species. The additional metrics computed were Unweighted Unifrac

distance (Lozupone and Knight, 2005) and Weighted Unifrac distance (Lozupone *et al*., 2007) – which are qualitative and quantitative measures of community dissimilarity, respectively, that incorporate phylogenetic relationships between features – where Unweighted Unifrac is based on presence or absence of species, and Weighted Unifrac additionally considers the relative abundance of species shared between samples.

As stated some confounders of gut microbiome diversity between infants were tested in order to characterize the diversity of South African infants: namely age, feeding practice and genetics. These accounted for the confounder of geography and culture by sampling for a single demographic group in Soweto which reduces variability, and for potential exposure to antibiotics. However, the significance testing in itself did not take into account other factors which may have biased the study results (Skelly *et al*., 2012). The factors relevant to this study include gestational age at birth, mode of delivery and, use of antibiotics, accounting for their specific types. Due to the circumstances around being a nested project in an ongoing study, the age at sampling for "at recruitment" was large and also a confounder as the age gap was 12 weeks long and may mask or exaggerate potential drivers of diversity over time. In addition, the sample sizes at the different time points varied due to collections not being made at all time points, arising in the issue of missing data; only partially accounted for by alpha diversity testing. This presents limitations by potentially distorting true effects and generating inaccurate estimates (underestimates or overestimates) of the true associations between the metadata being observed and the gut microbial diversity.

2.5.2 Taxonomic analysis with QIIME2 and LEfSe

This step in the QIIME2 pipeline was crucial in exploring the gross taxonomic composition of the samples i.e. what species of bacteria were present, in relation to the sample metadata. This involved the assignment of taxonomy to the amplicon sequence variants using the q2-feature-classifier – a QIIME 2 plugin containing several novel machine-learning and alignment-based methods for taxonomy classification – trained on Naïve Bayes classifier (Bokulich *et al*., 2018), against the SILVA 132 99% Operational Taxonomic Unit (OTU) marker gene reference database. OTUs are used to cluster microbial communities based on sequence similarities. These can be used to estimate species, diversity, composition and richness of the cluster (Chen *et al*., 2013; Quast *et al*., 2013). This created a taxonomic metadata file of feature IDs and their relative taxonomic classification, along with confidence values for the assigned genus and/or species (Appendix F, p. 133). This table was used to generate taxonomic bar plots of the taxonomic metadata, for visualization of species and relative abundance within the different samples in context of the metadata mapping file.

The Linear Discriminant Analysis (LDA) Effect Size (LEfSe) (Segata *et al*., 2011) pipeline (Appendix G, p. 134) accessed via Galaxy 1.0 (https://huttenhower.sph.harvard.edu/galaxy/) was used to further corroborate representative taxa and additional analyses on medication taken at recruitment and nine months. LEfSe identifies genomic features or taxa characterizing the differences observed between groups – emphasizing both statistical significance and biological relevance that are consistent with biologically meaningful categories within the sample metadata. LEfSe functions by using the statistical tests including the Kruskal-Wallis and Wilcoxon rank-sum test (Mann and Whitney, 1947), to detect differentially abundant features, and to investigate biological significance in a pairwise set, respectively.

The taxonomy metadata was converted to a suitable tab-separated-values format for input into LEfSe via the QIIME2 tutorial for “exporting and modifying biom tables” (accessed at https://forum.qiime2.org/t/exporting-and-modifying-biom-tables-e-g-adding-taxonomy-annotations/3630). Once uploaded, the LEfSe pipeline was followed through various steps according to Segata *et al*. (2011) including: formatting the data; calculating the LDA Effect Size and providing input for the visualizations modules; plotting LEfSe results of biomarkers graphically with their effect sizes; plotting cladograms of the biomarkers in a taxonomic hierarchal tree; and plotting one or differential features as abundance histograms.

2.6 Correlation of diversity and taxonomy to stunting

Research has highlighted that healthy infant growth is partially dependent on "normal" postnatal development of gut microbiota, and perturbation of the expected trend is causally related to malnutrition (Blanton *et al*., 2016). Age- and growth-discriminatory bacterial strains have been identified in the normally developing microbiome, and represent therapeutic targets in stunted or malnourished infants (Blanton *et al*., 2016). The weight-for-age and length-for-age measurements were collected (Appendix C, p. 119), and compared to the WHO Child Growth Standards (de Onis and WHO, 2006). Due to the growth estimates falling within the threshold for normal child growth, a direct correlation to stunting could not be made. In order to potentially correlate the age- and growth-discriminatory bacteria with infant development, literature was reviewed on the role and function of these bacteria. By assessment of the statistical significance of the alpha and beta diversity estimates and, reviewing the taxonomic profile of the infants: we drew comparisons between the metadata and microbiome literature highlighting these bacteria at different time points; between different feeding practices; and secretor statuses; and with or without medication administered.

CHAPTER THREE - RESULTS

3.1 Recruited participants (Mother-infant pairs)

At the end of the recruitment period (August 2018), we had a final sample size of n= 46 mother-infant pairs. As indicated in Table 3.1 (p. 39), the first time-point described as “at recruitment” encompassed a range of ages of the infants; where the youngest infant participant was twelve weeks old, and the oldest was twenty-one weeks old. The recruitment age period is wide due to circumstances beyond our control, as this study was nested in the ongoing SBW Study; therefore fitting into their timeline was necessary. There were a total of 26 female infants and 20 male infants. A total of 96% (n = 44) mother-infant pairs self-identified as South African black, and 4% (n = 2) identified as individuals of mixed ancestry.

Table 3.1: Mother-infant pair participant demographics

Category	***N***
Female infants	26 (57%)
Male infants	20 (43%)
Mother-infant pairs • Secretors • Non-secretors	46 • 34 (79.1%)* • 9 (20.9%)*
Self-identified South African black	44 (96%)
Self-identified South African mixed ancestry	2 (4%)
Age-range (at recruitment)	12-21 weeks (mean age = 20.17 weeks)

* = the proportion of secretors and non-secretors out of 43 successfully genotyped mothers.

3.2 Mother-infant biological samples collected

A total of n =171 maternal saliva and infant stool samples were collected following informed consent (refer to Table 3.2, p. 40). This project demonstrated the feasibility of biological sample collection in an urban South African setting. However, the n = 46 stool sample ideal at all the time points was not possible due to numerous circumstances. This included the loss/damage of stool collection kits, and the mothers’ unavailability for delivery of kits before or at time of collection.This resulted in a loss of 59 potential samples for analysis. Only n = 12 participants delivered stool samples across all four sample collection points. This was partially due to some participants being recruited at the six months time point (n = 12), but also a reduction in compliance; highlighting a limitation of longitudinal collection in this setting.

Table 3.2: Counts of mother and infant biological samples collected, secretor allele frequency, feeding practices, and administered medication at four time points in field and at the Developmental Pathways for Health Research Unit (DPHRU)

Sample	At recruitment (in field)			6 months (DPHRU)			9 months (in field)			12 months (DPHRU)			*N* all time points			*N* 16S sequenced		
Maternal saliva	-			46			-			-			46			-		
G allele frequency				-									0.62					
A allele frequency				-									0.38					
	F	**M**	***N***	**F**	**M**	***N***	**F**	**M**	***N***	**F**	**M**	***N***	**F**	**M**	***N***	**F**	**M**	***N***
Infant stool samples	14	16	30	21	17	38	15	13	28	15	14	29	65	60	125	59	51	111
-With known secretor status	8	13	21	14	13	27	12	11	23	13	13	26	47	50	97	42	41	83
-With known non-secretor status	5	1	6	7	2	9	2	1	3	2	1	3	16	5	21	15	5	20
Exclusive breastfeeding up to 6 months	8	8	16	10	9	19	9	7	16	12	6	18	39	30	69	33	27	60
-Secretors	6	6	12	8	6	14	8	6	14	10	5	15	32	23	55	27	20	47
-Non-secretors	2	1	3	2	2	4	1	1	2	2	1	3	7	5	12	6	5	11
Mixed feeding	5	7	12	8	6	14	4	6	10	2	6	8	19	25	44	19	22	41
-Secretors	2	6	8	4	5	9	3	5	8	2	6	8	11	22	33	11	18	29
-Non-secretors	3	0	3	4	0	4	1	0	1	0	0	0	8	0	8	8	0	8
Formula feeding	1	1	2	3	2	5	2	0	2	1	2	3	7	5	12	7	3	10
-Secretors	N/A	1	1	2	2	4	1	0	1	1	2	3	4	5	9	4	3	7
-Non-secretors	N/A	0	0	1	0	1	0	0	0	0	0	0	1	0	1	1	0	1
Infants administered medication	3	6	9	-	-	-	9	4	13	-	-	-	12	10	22	12	10	22

F = females; M = males; DPHRU = Developmental Pathways for Health Research Unit; *N* = sum total

A total of n = 125 infant stool samples were collected over the four time points. This comprised of 52% (n = 65) samples from females, and 48% (n = 60) from males. The greatest proportion of 30.4% samples were from the sixth month collection. Samples from exclusive breast feeders made up 55.2% (n = 69), mixed feeders were 35.2% (n = 44), and formula feeders were 9.6% (n = 12). Stool samples collected from secretor mothers made up 77.60% (n = 97), and non-secretors were 16.80% (n = 21) of the total. Samples sent for sequencing (n = 111) excluded n = 14 samples from twelve months that were collected after the timeline set for sending samples to NICD. The samples received from the sequencing

facility accounted for n = 60 exclusive, n = 43 mixed, and n = 10 formula feeding infants, and an unlabelled sample (total n = 114). Of these, n = 82 were secretor samples, and 21 non-secretor samples. Infants that had medication that could act as a confounder on diversity at the two time points of interest, Time0 ("at recruitment") and Time9, made up 19.82% (n = 22) of the total samples sequenced (Table 3.2, p. 40).

3.3 Generated metagenomic data

DNA extraction and 16S rRNA sequencing

We were successful in extracting DNA from infant stool samples using the QIAamp PowerFaecal DNA kit (Qiagen, Hilden, Germany) following a modification and optimisation of the protocol (Appendix B, p. 118). DNA concentrations using the DNA HS-Kit (Qubit™, Thermo Fisher Scientific, Wilmington, Delaware USA) and the Qubit Fluorometer had a range in concentration (0.166ng/µl to 224ng/µl) and quality, corresponding to band intensity visible on the agarose gel image indicated in Figure 3.1 (p. 41). A total of 125 stool samples were collected and DNA extracted with a particular focus for this research project on the first time point (12 - 21 weeks; n = 30) and at nine months (~36 weeks; n = 28). Sequencing reads were generated for n = 111/125 (refer to Methods section 2.4) samples were sent for sequencing with fourteen samples remaining (time point 12). The remaining samples

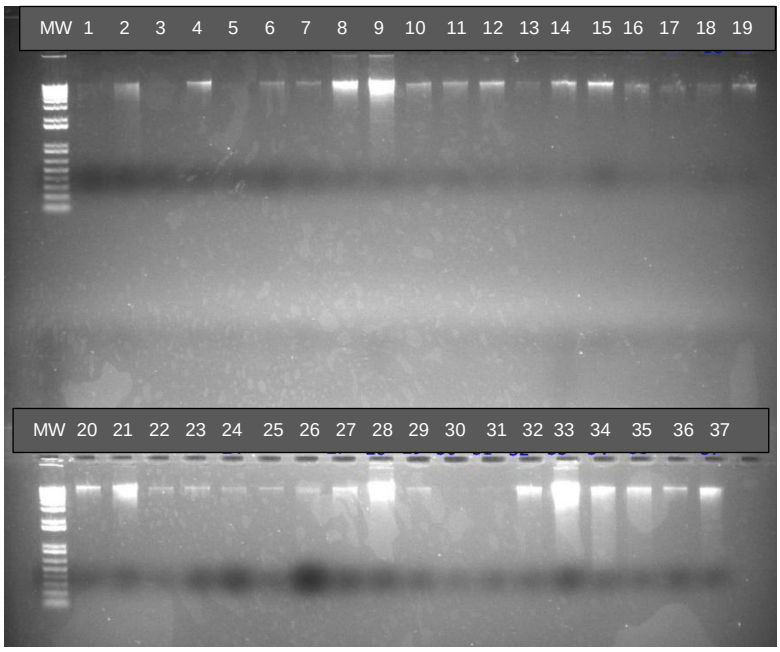


Figure 3.1: **An ethidium-bromide (Et-Br) stained agarose gel (1% w/v) image illustrating DNA extractions from stool samples (n=37) collected at the six month time point** This demonstrates the range in intensity of DNA bands from low to high, which corresponds to the a range in DNA concentrations. MW= 1kb molecular weight DNA ladder. The 38th six month sample was collected at a later date and analysed with a different batch of samples.

collected later than the other samples due to an infant age gap, were not sent for sequencing due to time and resource constraints.

3.4 *Fucosyl transferase 2* (rs601338) mutation screened

In this study, 83% (n = 36) of the genotyped mothers (n = 43) participated in some form of breastfeeding – either exclusively (n = 24/36) or in combination with formula (n = 12/36). In figure 3.2 (p. 42) the feeding practice performed by the entire cohort of mothers is shown. The presence or absence of the *FUT2* mutation in mothers is only relevant in ~93% of infant stool samples collected (n = 82 secretor stool samples, n = 21 non-secretor stool samples).

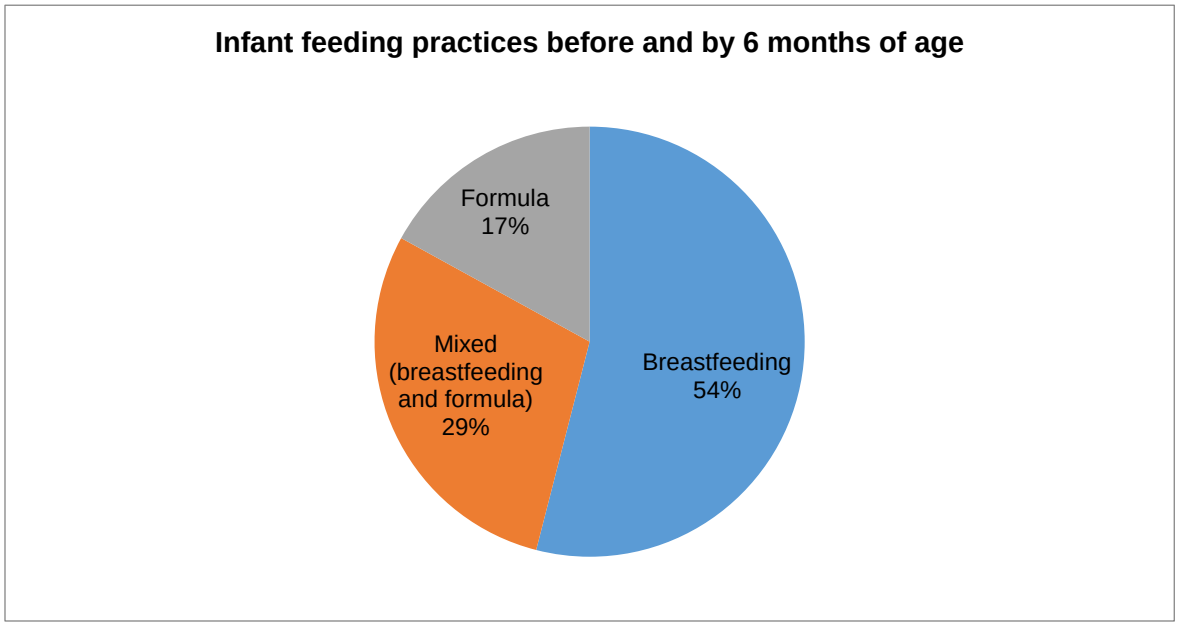


Figure 3.2: A pie chart demonstrating the proportion of feeding practices performed by all mothers before and up to six months of age Formula feeding was practiced in 17% (n = 8) of mothers. ~83% of mothers breastfed either exclusively (n = 25, 54%) or in combination with formula feeding (n = 13, 29%). Therefore, maternal secretor status is relevant in ~93% (n = 103) of infant samples.

3.4.1 DNA extraction

The DNA extraction performed on the maternal saliva samples yielded DNA of good quality and quantity as evidenced by the NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific Inc., Wilmington, Delaware USA) results, with DNA concentrations ranging from 29.6 - 1706.9ng/µl. The EtBr-stained agarose gel electrophoresed on the samples showed DNA of good quality and integrity, as shown in Figure 3.3 (p. 43).

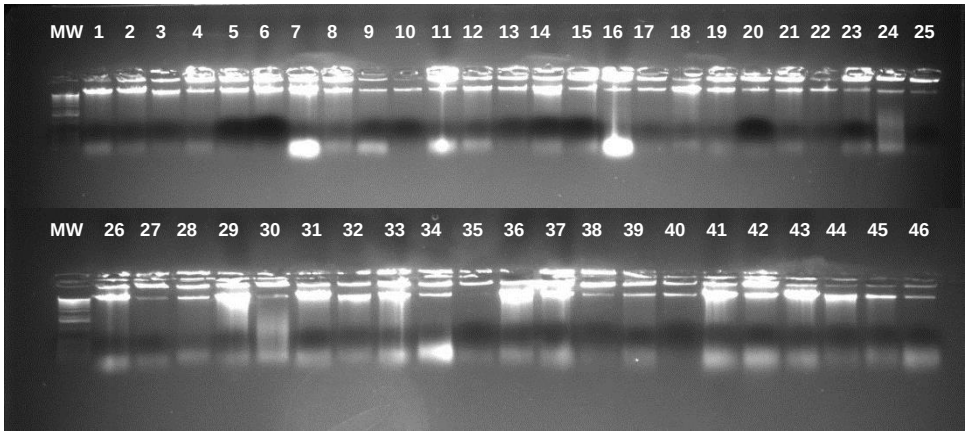


Figure 3.3: An ethidium bromide (EtBr) stained agarose gel (1% w/v) image illustrating the quality of maternal saliva DNA extractions Bright bands (n=46 samples) correspond to the good quality of DNA from the Oragene saliva self-collection kit, and the prep-IT L2P extraction protocol. The first column of the top and bottom rows are 1kb molecular weight (MW) DNA ladders, and the remaining columns (numbered 1-46) are the extracted saliva samples.

3.4.2 Genotype assay and analysis

The different genotypes ("GG", "GA" or "AA") corresponding to mothers present in the study, were called by generating an allelic discrimination plot using the TaqMan Genotyper software (Figure 3.4, p. 44). The "G" allele frequency was 61.60% and "A" was 38.40%. These genotypes were in Hardy-Weinberg equilibrium.

The genotype analysis indicated that 44.20% had a GG genotype, 34.90% had a GA genotype, and 20.90% were AA genotype. As shown in Figure 3.5 (p. 44) a total of 79.10% (n = 34/43) of mothers were secretors (GG and GA frequencies were combined), and 20.90% (n = 9/43) were non-secretors (only the AA frequency). From the 34 maternal secretors, 82 infant stool samples were collected (46 exclusively breastfed, 30 mixed fed, and 6 formula fed); and from the 9 maternal non-secretors, 21 samples were collected (11 exclusively breastfed, 9 mixed fed, and 1 formula fed).

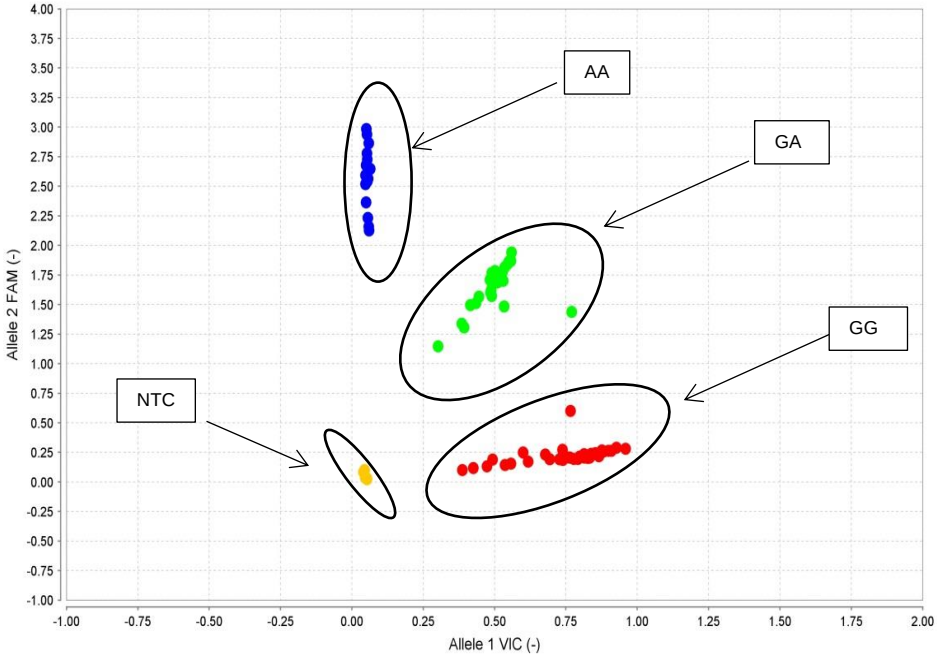


Figure 3.4: An allelic discrimination plot illustrating the TaqMan assay maternal samples This distinguishes the different genotypes observed for the rs601338 SNP. Genotype frequencies: "GG" (red; 44.20%, n = 19), "GA" (green; 34.90%, n = 15), and "AA" (blue; 20.90%, n = 9). "GG" and "GA" indicate the secretors (n = 34) of the study, while AA (n = 9) indicates the non-secretors. NTC= non-template controls.

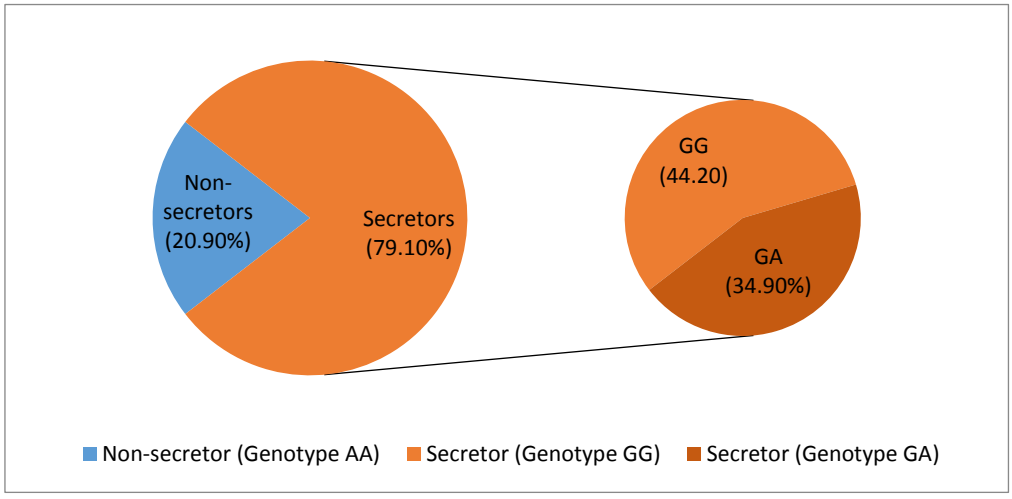


Figure 3.5: A pie chart illustrating the maternal genotype frequency of secretors and non-secretors in the sample of n = 43 mothers (three samples were inconclusive). Frequency of non-secretors and genotype frequency of "AA" is 20.90% (n = 9, blue). Secretor frequency (79.10%, n = 34) is a total of GG (44.20%, n = 19) and GA (34.90%, n = 15) genotype frequencies.

3.4.3 Sanger sequencing

The results of the Sanger sequencing were inconsistent due to the difficulty of PCR optimization and amplification (Figure 3.6, p. 45). Time and resource constraints permitting, we repeated the experiment several times, however insufficient product was generated to proceed to the next step in Sanger sequencing verification of the samples.

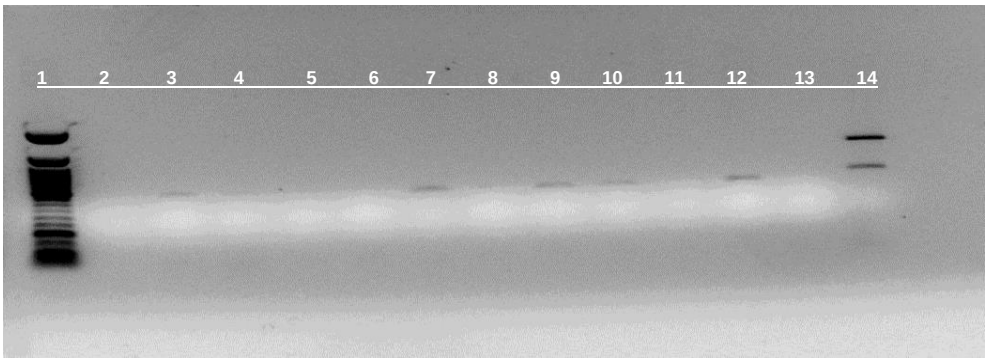


Figure 3.6: **An ethidium bromide (Et-Br) stained agarose gel (1% w/v) image illustrating PCR amplification output post optimization** This illustrates the final 18th experimental attempt at attaining three sufficient PCR products of each genotype for Sanger sequencing, and can be seen as faint bands or not visible at all. Column labels: **1**- 1kb molecular weight DNA ladder, **2**- blank, **3 to 5**- 3x samples of "GG" genotype, **6 to 8**- 3x samples of "GA" genotype, **9 to 11**- 3x samples of "AA" genotype, **12** and **14**- positive control, **13**- negative control.

3.5 Bioinformatics and data analysis

3.5.1 QIIME2

I. Quality control of sequenced data

Following importation of fastq.gz files from the Wits Cluster into QIIME2, there were n = 228 (n = 114 pairs) forward and reverse paired-end sequences in total. This accounted for the 110 samples sent for sequencing, an unlabelled sample and the additional three sequences rerun for quality purposes (Appendix H, p. 135). The demultiplexed sequencing output indicated that a mean of ~139 000 sequences were obtained per sample and 15 917 621 sequences were obtained in total. Following the trimming and joining of reads using Cut-adapt, the mean number of sequences per sample decreased to ~85 776, and the quality of the joined sequences decreased after 400 bases. To ensure quality of sequences, the "n" trim length value was chosen at 401, of the joined and trimmed sequences using Deblur-denoising, and a plot was generated using a random sampling of 9 966 out of 9 778 501 sequences without replacement.

The denoising quality control statistics showed a maximum of 131 740 reads that matched a reference sequence, indicating that high quality sequence variant data was produced. This

process additionally resulted in: 1 384 features/ representative-sequences, where the mean frequency of features per sample was >28 800; frequency per feature was >2 000; and we retained more than three million sequences in 113 samples when sampling depth was zero. Here, the samples were n = 113 because the statistics output started a sample count at zero, therefore there were actually more than three million sequences in n = 114 samples.

II. Alpha and beta diversity analysis

Alpha diversity statistical analysis

Using the core-metrics phylogenetic method, alpha diversity boxplots were generated in order to identify significant associations between the different columns of data (categories: time point, child sex, secretor status and feeding practice) in the metadata mapping file (Appendix C, p. 119), and alpha diversity metrics – Faith's Phylogenetic Diversity and Pielou's Evenness. Sequences returned from NICD for Time0, 6, 9, 12 accounted for n = 30, 38, 27 and 15, respectively; due to a single nine month 16S read sample that did not match the sample-ID (00-012-9m) in the metadata mapping file. Therefore at nine months that total sample count is n = 27. Returned sequences corresponding to females and males accounted for n = 60 and 53, respectively. Quality control led to loss of two female samples, and the nine month sample previously mentioned was male; therefore n = 58 and 52. For exclusive, mixed and formula feeding, the corresponding sequence counts were n = 60, 43 and 10, respectively. After quality control and matching sample IDs, each group had one less sample count of n = 59, 42, and 9. In secretor status analyses, an input of n = 110 samples were used, but the analysis was only done on 101 samples accounting for missing data and formula feeders. The nine month sample previously mentioned and quality control process was a loss of two secretor samples; therefore n= 80 and 21 for secretors and non-secretors.

Faith's Phylogenetic Diversity

A significant difference in biodiversity richness indices was found within the samples for comparisons of time points (twelve months vs at recruitment; six; and nine months) and child sex (male vs females). This is indicated by the box-plots (Figure 3.7, Figure 3.8, p. 47) and Kruskal-Wallis (KW) *P*-value and q-value (false positive rate, and the adjusted *P*-value controlling the positive false discovery rate, respectively) being **less than 0.05** (Table 3.3, Table 3.4, p. 48).

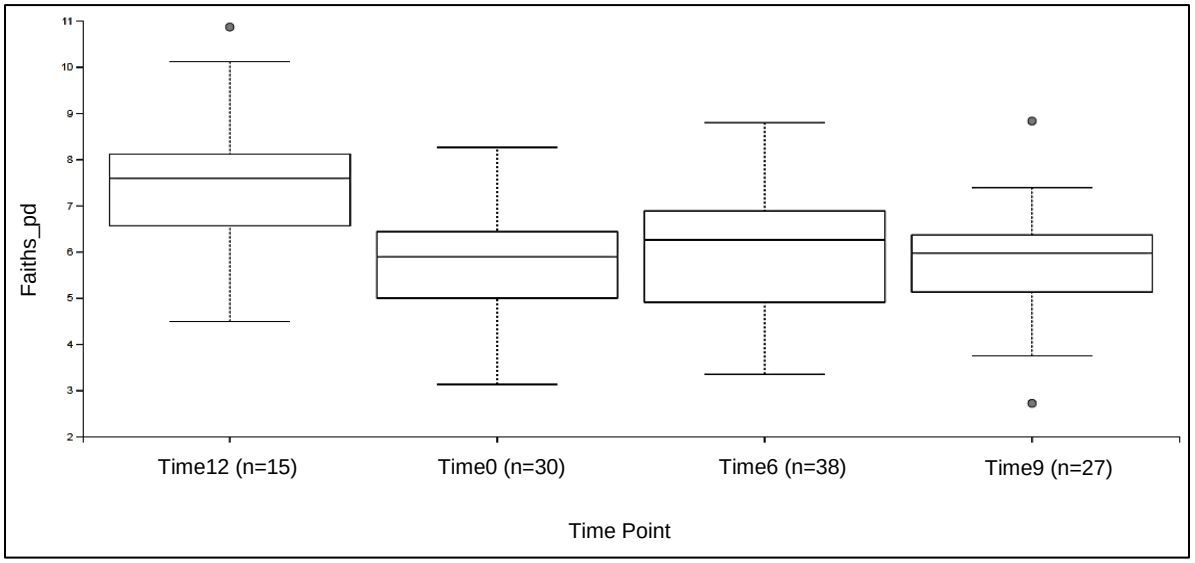


Figure 3.7: Box-plots illustrating taxonomic phylogenetic indices (Faith's PD) of different time points (Box plots automatically generated from left: 1st= Time12; 2nd= Time0; 3rd= Time6; 4th= Time9). Time12 has a high Faith's PD index compared to the other time points that have similar indices, corresponding with the significant KW statistical test values shown in Table 3.3, relating to a greater degree of species richness within Time12 samples.

- The significant differences were specifically for twelve months: versus at recruitment (*P*-value/q-value=$0.44x10^{-3}$/$1.00x10^{-3}$); versus six months (*P*-value/q-value = 0.01/0.02); versus nine months (*P*-value/q-value = $0.31x10^{-3}$/$1.00x10^{-3}$).
- There was also a higher diversity index in male samples compared to females at all sample collection points (*P*-value/q-value = 0.02).

Figure 3.8: Box-plots illustrating taxonomic phylogenetic indices (Faith's PD) between males and females (left =female; right= male). The samples from all collection points from males have a higher Faith's PD index than females, corresponding with the significant KW statistical test values shown in Table 3.4, relating to a greater degree of species richness within the male samples.

However, there was no significant difference in Faith's PD indices for other time point comparisons (Time0 vs Time 6, Time9; Time 6 vs Time9 – Table 3.3, p.48), as well as comparison of feeding practices (exclusive vs mixed; exclusive vs formula; formula vs mixed) and secretor status (Appendix I, p. 136).

Table 3.3: Kruskal-Wallis statistical test results of significance for comparisons among all time point groups and pairwise tests for phylogenetic diversity using Faith's Phylogenetic Diversity metric

Kruskal-Wallis (all groups)				
	Result			
H	14.73			
P-value	2.00×10^{-3}			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
Time12 (n=15)	Time0 (n=30)	12.36	0.44×10^{-3}	1.00×10^{-3}
	Time6 (n=38)	6.39	0.01	0.02
	Time9 (n=27)	13.03	0.31×10^{-3}	1.00×10^{-3}
Time0 (n=30)	Time6 (n=38)	1.44	0.23	0.35
	Time9 (n=27)	0.01	0.91	0.91
Time6 (n=38)	Time9 (n=27)	0.99	0.31	0.38

Time0= at recruitment (12-21 weeks); Time6 (6 months); Time9 (9 months); Time12 (12 months).

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures.

Table 3.4: Kruskal-Wallis statistical test results of significance for comparisons between male and female infants at all sample collection points and pairwise tests for phylogenetic diversity using Faith's Phylogenetic Diversity metric

Kruskal-Wallis (all groups)				
	Results			
H	5.74			
P-value	0.02			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
F (n=58)	M (n=52)	5.74	0.02	0.02

F= females; M= males.

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures

Pielou's Evenness

A significant difference in community Evenness indices was observed within the samples for time points (at recruitment vs six; and nine months) and child sex (males vs females). This is indicated by the box-plots (Figure 3.9, p. 49; Figure 3.10, p. 50) and Kruskal-Wallis (KW) *P*-value and q-value being **less than 0.05** (Table 3.5, p. 50; Table 3.6, p. 51)**.**

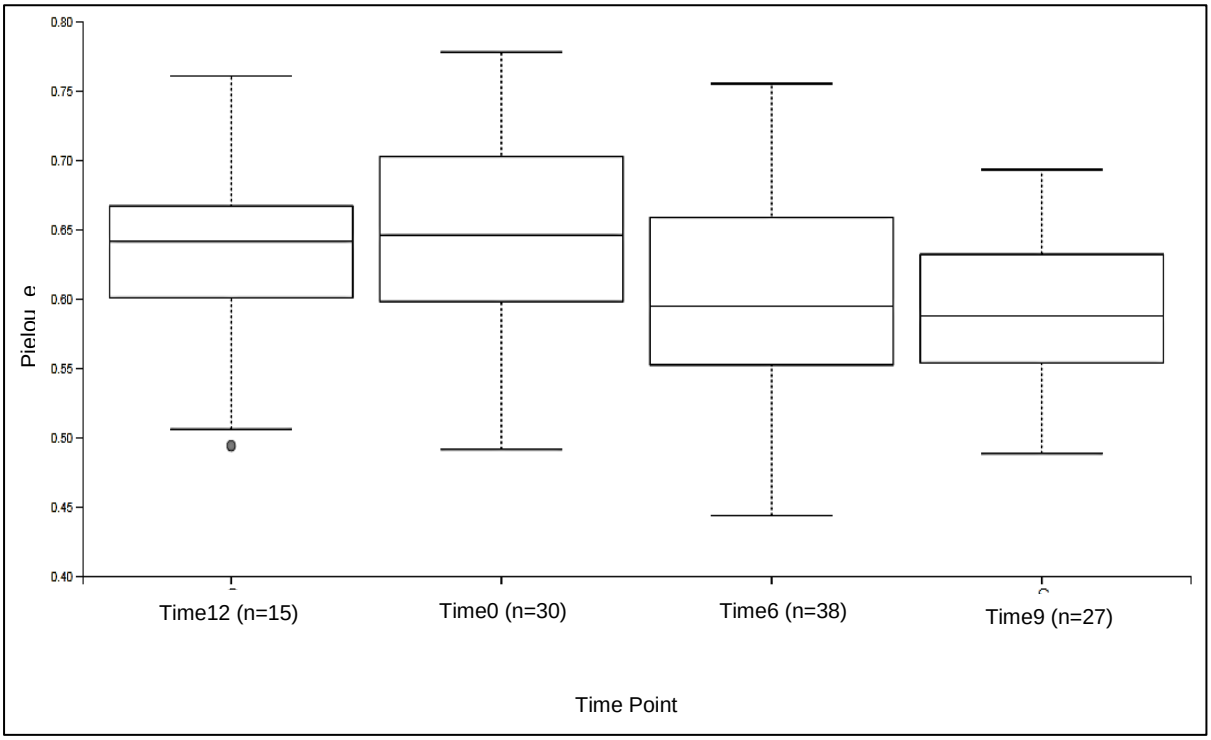


Figure 3.9: **Box-plots illustrating taxonomic phylogenetic indices (Evenness) highlighting different time points** (Box plots automatically generated from left: 1st= Time12; 2nd= Time0; 3rd= Time6; 4th= Time9. Time0 has a high Evenness index compared to the other time points that have similar indices, corresponding with the significant KW statistical test values in Table 3.5, relating to a greater degree of community evenness.

- The significant differences were specifically for the at recruitment time point: versus six months (*P*-value/q-value=0.01/0.03); and versus nine months (*P*-value/q-value=$2.00x10^{-3}$/0.02).
- There was also a higher diversity index observed for community evenness in male samples compared to females at all sample collection points (*P*-value/q-value = 0.02).
- However, there was no significant difference in Evenness indices for other time point comparisons (Table 3.5, p. 47), as well as comparison of feeding practices and secretor status (Appendix I, p. 136).

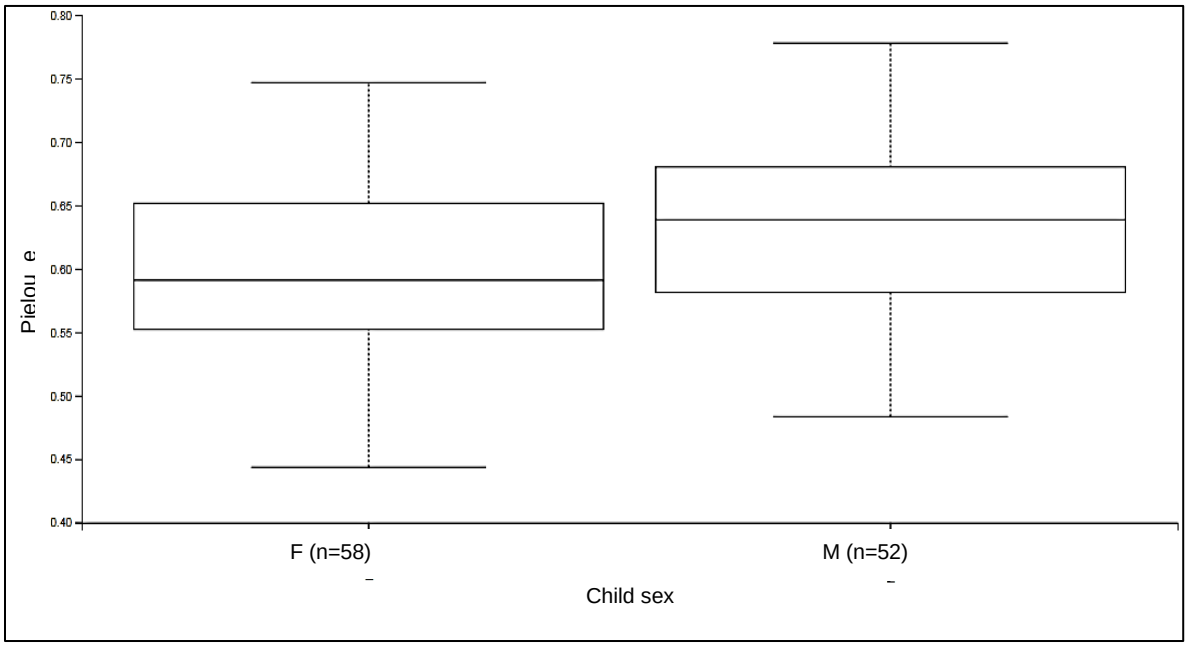


Figure 3.10: **Box-plots illustrating taxonomic phylogenetic indices (Evenness) between males and females** (left =female; right= male). The samples from all collection points from males have a higher Evenness index than females, corresponding with the significant KW statistical test values in Table 3.5, relating to a greater degree of community evenness.

Table 3.5: Kruskal-Wallis statistical test results of significance for comparisons among all time point groups and pairwise tests for phylogenetic diversity using Pielou's Evenness metrics

Kruskal-Wallis (all groups)				
	Result			
H	11.57			
P-value	$9.00x10^{-3}$			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
Time12 (n=15)	Time0 (n=30)	0.59	0.44	0.53
	Time6 (n=38)	2.31	0.13	0.19
	Time9 (n=27)	3.52	0.06	0.12
Time0 (n=30)	Time6 (n=38)	6.73	0.01	0.03
	Time9 (n=27)	9.12	$2.00x10^{-3}$	0.02
Time6 (n=38)	Time9 (n=27)	$0.01x10^{-2}$	0.99	0.99

Time0= at recruitment (12-21 weeks); Time6 (6 months); Time9 (9 months); Time12 (12 months).

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures.

Table 3.6: Kruskal-Wallis statistical test results of significance for comparisons between male and female infants at all sample collection points and pairwise tests for phylogenetic diversity using Pielou's Evenness metrics

Kruskal-Wallis (all groups)				
	Results			
H	5.68			
P-value	0.02			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
F (n=58)	M (n=52)	5.68	0.02	0.02

F= females; M= males.

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures.

Alpha rarefaction plotting

We investigated alpha diversity as a function of sampling depth with the QIIME2 diversity alpha-rarefaction visualizer, and computed several metrics – Shannon diversity, Amplicon sequence variants and Faith's PD – at multiple sampling depths between a minimum depth (1) and maximum depth (23 000). This resulted in rarefaction curves of average diversity values plotted for each sample based on columns of metadata, and indicated that the species richness was fully observed or sequenced, based on the initial exponential growth and gradual levelling out of the curves. This indicates that the collection of additional sequences beyond that sampling depth would not be likely to result in the observation of additional features.

The metadata columns of time point, secretor status, child sex and feeding practice were all plotted against the three metrics, where the top visualization is the alpha rarefaction plot indicating the full observation of species richness, and the bottom plot indicates the number of samples that remain in each metadata group when the feature is sub-sampled between one and 23 000 (Figure 3.11, p. 52; see also Appendix J for additional plots, p. 139).

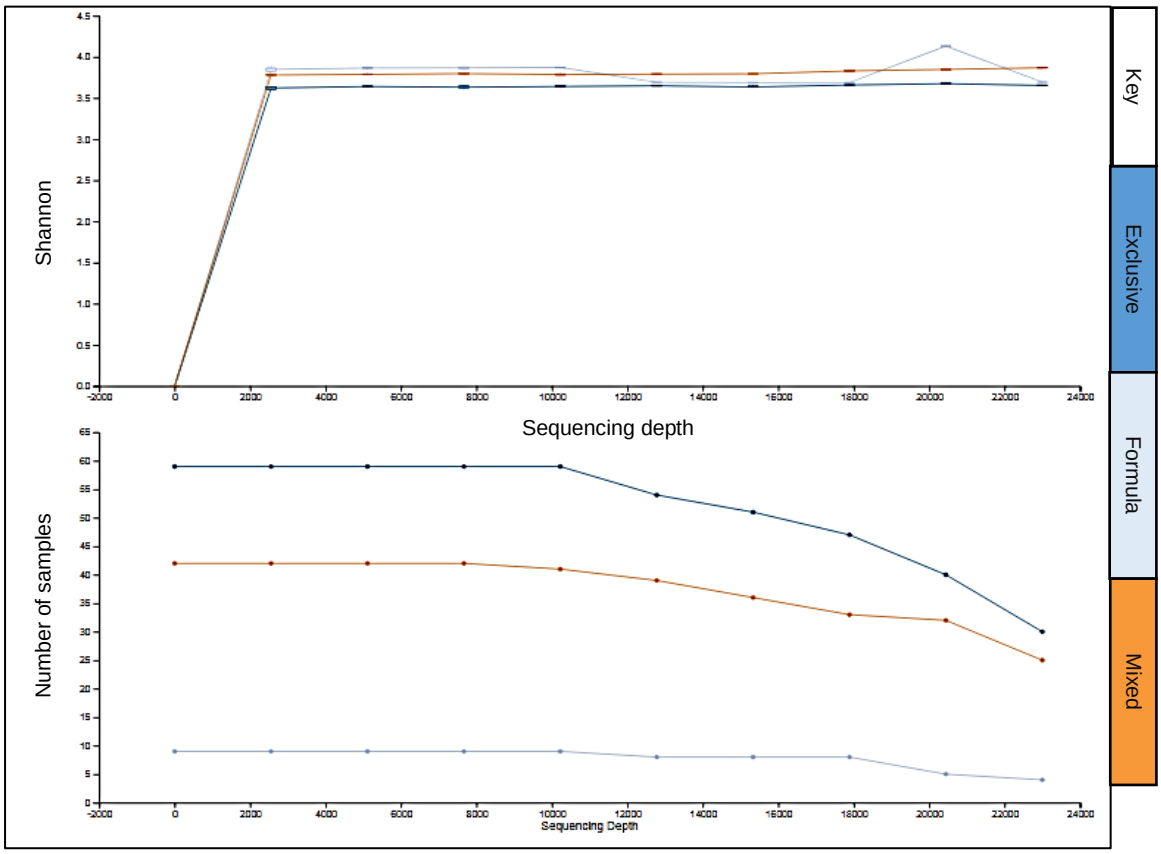


Figure 3.11: Alpha rarefaction curves illustrating the different feeding practices computed with Shannon Diversity metrics (y-axis top plot). The exponential growth and plateau of the curves show species richness is fully observed at a maximum index of 4 (formula – light blue) and sequencing depth of 2 000. The bottom plot indicates the number of samples (y-axis bottom plot) rarefied to the sequencing depth (x-axis) decreases after 12 000.

Beta diversity statistical analysis

Principle Coordinates Analysis (PCoA) of metadata

Using the core-metrics phylogenetic method, beta diversity PCoA plots (Figure 3.12, p. 54) were generated with R in order to view the distances between groups of categorical metadata in ordination space, with the aim of identifying clusters, using beta diversity metrics – Jaccard distance, Bray-Curtis distance, Weighted Unifrac distance, and Unweighted Unifrac distance. Distinct clusters - according to the metadata columns of time point, Time0 versus Time9, feeding practice, secretor status, and sex of the infant – were unidentifiable in the PCoA plots that were generated (Appendix K, p. 150). The x- and y-axes indicated the percentage of variation in distance measures that the principle coordinates (PC) 1 and 2 accounted for, respectively. Although distinct clusters were not illustrated, the loading scores

on the plot indicate what variable has a larger effect, and varied according to the metric computed.

- *Jaccard distance measures of metadata Time0 versus Time9 illustrated that:* PC1 accounts for 6% and PC2 accounts for 4% of the qualitative variation in distance between groups. Time9 has a larger effect indicated by the greater loading score (>0.2) on both PC1 and PC2 axes (Figure 3.12 A, p. 54).
- *Bray-Curtis distance measures of the same metadata illustrated that*: PC1 accounts for 13% of the quantitative variation in distances between the groups, and Time9 has a larger effect indicated by the greater loading score (>0.2) along the x-axis. PC2 accounts for 6% variation and Time0 has larger effect indicated by the greater loading score (>0.3) along the y-axis (Figure 3.12 B, p. 54).
- *Weighted Unifrac distance measures for the same metadata illustrated that:* PC1 accounts for 39% and PC2 for 11% of the quantitative variation in distance between groups. Similarly, Time9 has a larger effect for both axes (>0.2) (Figure 3.12 C, p. 54)
- *Unweighted Unifrac distance illustrated that* PC1 accounts for 13%, and PC2 for 8% of the qualitative variation in distance between groups, where Time9 has a larger effect (~0.3) for both axes (Figure 3.12 D, p. 54).

The plots generated for the remaining metadata columns (Appendix K, p. 150) have similar axes and loading scores generated, and the same rationale can be deduced from the plots.

PERMANOVA in QIIME2

We analysed the distances within and between sample groups, using PERMANOVA (in QIIME2) computed with Unweighted Unifrac distances against: feeding practice, time point and secretor status. A statistics summary of the test performed and the overall *P*-value was generated (Table 3.7, p. 55), and the output of distances was generated as box-plots of distance values against each group (Figure 3.13, p. 56; Appendix L, p. 154).

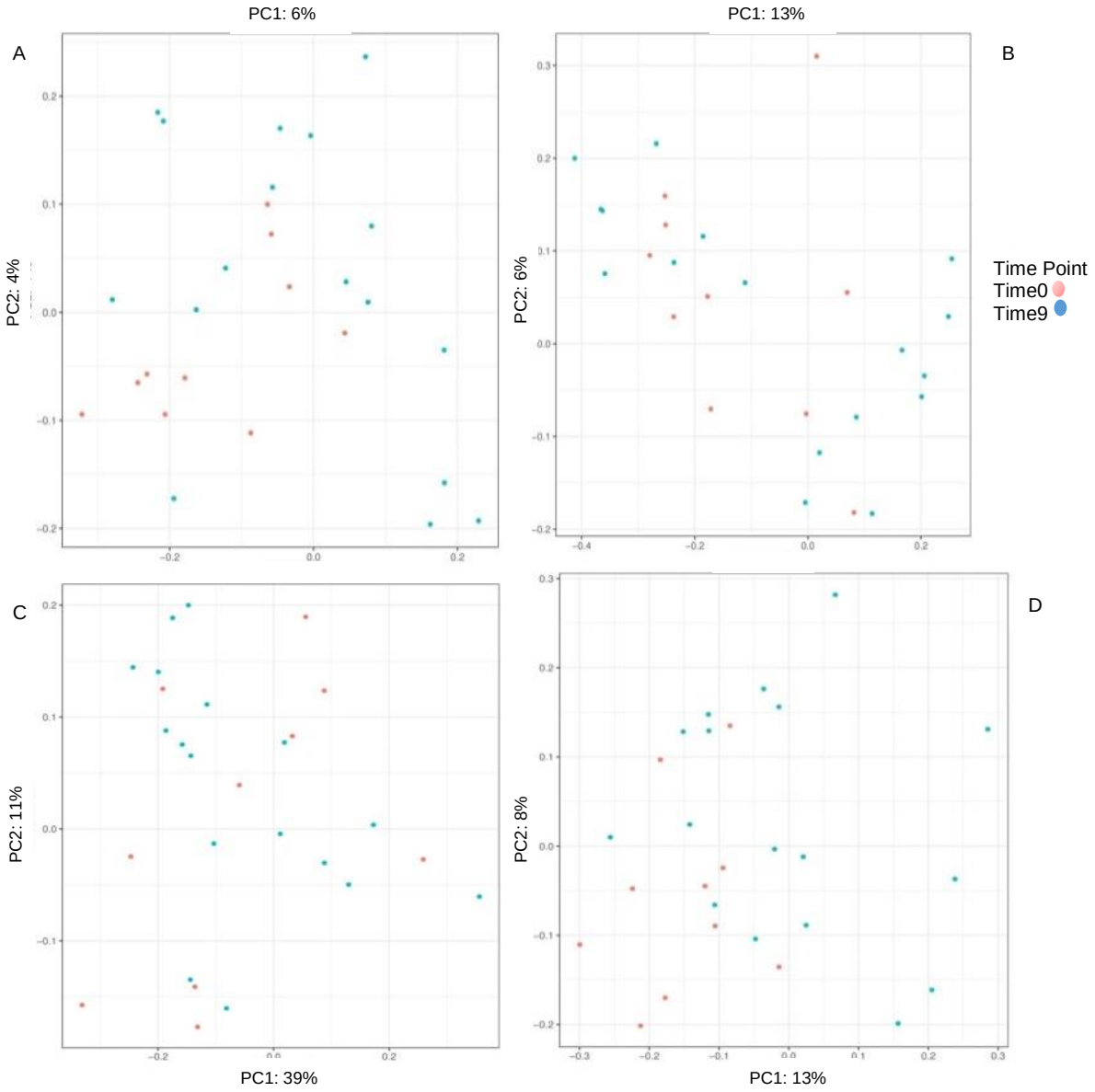


Figure 3.12: **Principle Coordinates Analysis plots illustrating Time0 (red) versus Time9 (blue) for beta diversity metrics** A) Jaccard Distance B) Bray-Curtis Distance C) Weighted Unifrac D) Unweighted Unifrac; is showing that distinct clusters are not formed for the different time points. PC= principle coordinates 1 and 2 show the percentage of variation accounted for, and the x- and y-axis values are loading scores indicating which variable has a larger effect.

*Statistically significant values were generated for feeding practice (P-*value= *0.003) (Table 3.7, p. 55), where:*

- **Mixed and formula feeding versus exclusive breastfeeding distances were significant** – exclusive versus mixed (*P*-value/q-value = 0.01/0.02); exclusive versus formula (*P*-value/q-value= 0.03/0.04). However, when mixed was compared to

formula feeding, there was no statistical significance (*P*-value/q-value= 0.14/0.14) (Table 3.8, p. 55).

Table 3.7: PERMANOVA (in QIIME2) statistics summary highlighting the analysis of feeding practices including overall *P*-value significance

	PERMANOVA results
Method name	PERMANOVA
Test statistic name	Pseudo-F
Sample size	110
Number of groups	3
Test statistic	1.79
P-value	$3.00x10^{-3}$
Number of permutations	999

*Statistical significance was also observed in time point analyses (P-*value= *$1.00x10^{-3}$) (Table 3.9, p. 57), where:*

- **Time12 versus all other time points generated *P*-values of significance** (below 0.05) as well as Time9 versus Time0 (Table 3.9, p. 57).

Table 3.8: PERMANOVA statistical values highlighting each pairwise test of the analysis of exclusive, formula and mixed feeding practices

Group 1	**Group 2**	**Sample size**	**Permutations**	**pseudo-F**	***P*-value**	**q-value**
Exclusive	Formula	68	999	1.72	0.03	0.04
Exclusive	Mixed	101	999	2.05	0.01	0.02
Formula	Mixed	51	999	1.34	0.14	0.14

Secretor status was not found to have statistically significant differences in distances between secretor and non-secretor sample groups (Appendix L, p. 154).

Calculating associated n in PERMANOVA box-plots for “Distances to exclusive”

- Sub-total of exclusive = the sum of all integers less than exclusive sample size n,

  Where, n = 59; Therefore, (58(58+1)) / 2 = 1711
- Sub-total of formula = formula sample size n x exclusive sample size n,

  Which is, 9 x 59 = 531
- Sub-total of mixed = mixed sample size n x exclusive sample size n,

  Which is, 42 x 59 = 2478

All n values in the PERMANOVA box-plots presented were calculated with this method.

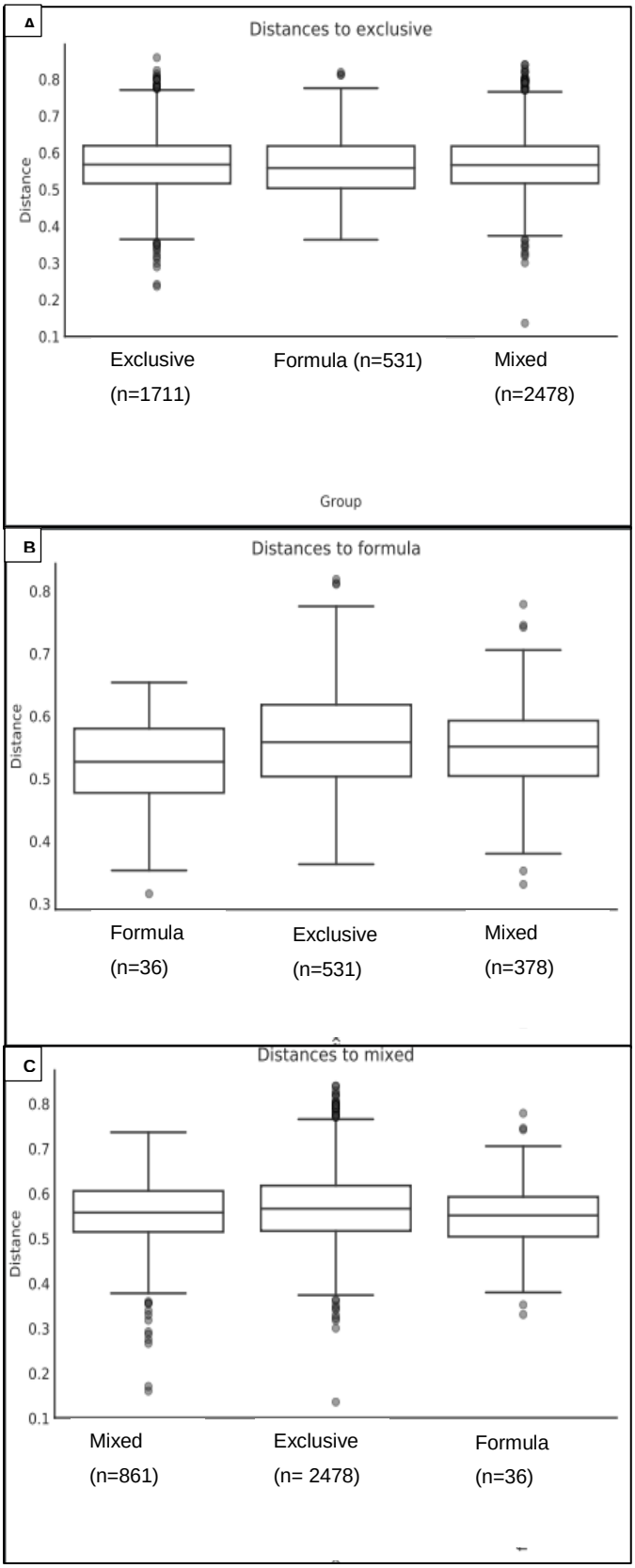


Figure 3.13: Box-plots illustrating Unweighted Unifrac distances of feeding practices using PERMANOVA (QIIME2) A) Distances to exclusive, B) Distances to formula, C) Distances to mixed feeding. This shows pairwise distances in each graph for within groups (the first box-plot) and between groups (the subsequent box-plots). A, C) the smallest distances among samples are seen **between groups** for formula feeding; B) the smallest distances among samples is seen **within groups** for formula feeding. Refer to page 55 for calculation of n for each group box-plot.

Table 3.9: PERMANOVA statistical values highlighting each pairwise test of the analysis of Time0, Time6, Time9, and Time12 time points

Group 1	Group 2	Sample size	Permutations	pseudo-F	*P*-value	q-value
Time12	Time0	45	999	2.83	$1.00x10^{-3}$	$3.00x10^{-3}$
Time12	Time6	53	999	2.57	$1.00x10^{-3}$	$3.00x10^{-3}$
Time12	Time9	42	999	2.02	$4.00x10^{-3}$	$8.00x10^{-3}$
Time6	Time0	68	999	1.05	0.36	0.36
Time6	Time9	65	999	1.47	0.06	0.07
Time9	Time0	57	999	1.85	0.01	0.02

Time0= at recruitment (12-21 weeks); Time6 (6 months); Time9 (9 months); Time12 (12 months).

3.5.2 Taxonomic analysis with QIIME2 and LEfSe

The taxonomic composition of the infant gut samples were characterized to investigate any marked differences between infants of mothers with differing secretor status, varied feeding practices, and possible changes over time – specifically at the at recruitment (Time0 = 12-21 weeks) and nine months (Time9) collection points. Taxonomic bar-plots generated using the taxonomy metadata in QIIME2, illustrates up to seven levels of classification (Kingdom to Species) of data in relation to the sample metadata mapping file. For QIIME2 analyses we focused on phylum-level classification, and observed estimates of abundance through the percent composition of each phylum in a sample. Additional statistical analyses were not performed to generate *P*-values. The taxonomy metadata from the QIIME2 pipeline was used for investigating differential abundance at genus-level for LEfSe analyses. The marked taxa presented statistically significant differences, and were presented as linear discriminant analysis scores, and not additional *P*-values.

I. Taxonomic analysis relative to secretor status

Results presented include the taxonomic analysis of infant gut microbiomes where the infants have been exclusively breastfed (n = 24 infants) or where mothers used mixed feeding practices (n = 12 infants), and the influence of their secretor status. Only n = 43/46 maternal samples were successfully genotyped. Microbiome analysis of formula fed infants (n = 7 infants) was excluded from this part of the analysis, due to the fact that secretor status is only relevant in infants that have been breastfed (n = 34 secretors and n = 9 non-secretors).

Phylum-level analyses conducted in QIIME2 (Figure 3.14 A, p. 59) showed that, samples from both secretors and non-secretors have similar counts of certain phylum groups, although the samples with the higher percentage of these phyla are found in secretors. The different phylum groups accounting for majority of the biodiversity in the samples are:

- Firmicutes (70% non-secretors; 72% secretors), Bacteriodetes (58%; 69%), Proteobacteria (47%; 54%), and Actinobacteria (24%; 27%). Notably, in a minority of samples, there are counts of Verrucomicrobia (~10% in non-secretors; ~28% in secretors), and Fusobacteria (~13% non-secretors; ~34% secretors).

In the **LEfSe genus-level analyses** (Figure 3.14 B, p. 59) several clades of bacteria are highlighted as differentially abundant between secretors and non-secretors. The LEfSe scores shown indicate the degree of consistent difference in relative abundance between features in the two groups.

- The highest score is at 3.5 for *Lachnoclostridium* and the lowest at -4.5 for *Veillonellaceae*, representing the most statistically abundant and depleted clades, respectively, between secretor status groups. Both of these genera fall under the phylum Firmicutes accounting for all the differentially abundant groups in secretors, whereas the phyla Proteobacteria (*Rhodospirillales*) and Cyanobacteria (*Citrus maxima*) are additionally observed in differentially depleted groups of non-secretors. This corresponds with observations in the phylum-level QIIME2 analyses.

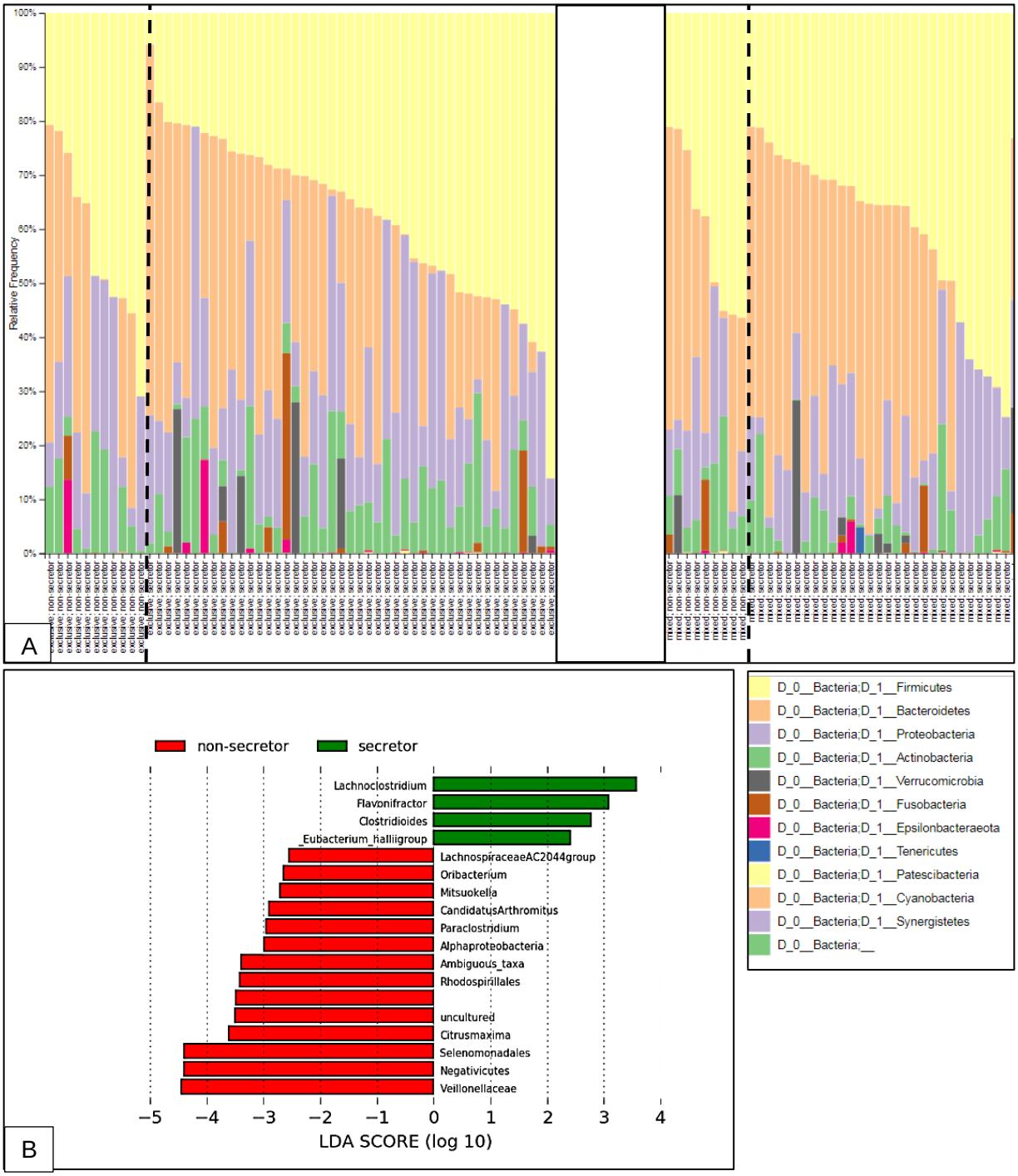


Figure 3.14: **Taxonomic analysis relative to secretor status** A) Phylum-level taxonomic bar-plots and key of samples corresponding to secretor status of mothers for infants that are exclusively breastfed or have mixed feeding. Non-secretors and secretors are separated by the dotted black lines illustrated on the left graph for exclusive feeders, and the right graph for mixed feeders. The overall trend highlights the similarity in composition and abundance between secretor status samples; however, some secretor samples have a greater diversity illustrated by more colours corresponding to additional phyla. B) A histogram of the LDA scores computed for features differentially abundant between non-secretors and secretors identified fifteen bacterial clades. For secretors, the most differentially abundant (*Lachnoclostridium*) and underrepresented (*Veillonellaceae*) clades are both from the phylum Firmicutes, also accounting for the majority phyla represented. Species from the phylum Proteobacteria (*Rhodospirillales*) and Cyanobacteria *(Citrus maxima)* are also presented as abundant in non-secretors.

II. Taxonomic analysis relative to feeding practice

The taxonomic bar plots representing the different feeding practices (Figure 3.15 A, p. 62) had varying biodiversity composition at the phylum level – particularly between the feeding groups involving some form of breastfeeding (exclusive and mixed feeding) in comparison to formula feeding. The greatest abundance of the phyla representing the majority of the biodiversity were observed in samples from infants with exclusive breast milk diets, and the lowest levels were observed in formula feeders as illustrated in Table 3.10 (p. 60).

Table 3.10: Estimates of phyla observed in majority of the samples with regards to different feeding practices highlighting exclusive breast milk diets associate with high biodiversity abundance

	Phylum	Feeding practice		
		Exclusive (%)	Mixed (%)	Formula (%)
Majority of samples have abundance of these phyla	**Firmicutes**	86	74	63
	Bacteroidetes	68	69	59
	Proteobacteria	71	42	40
	Actinobacteria	27	24	8
	Verrucomicrobia	36	28	7
	Fusobacteria	34	13	0.8

The LEfSe genus-level analysis was performed for pairwise comparisons of exclusive breastfeeding versus mixed feeding, exclusive breastfeeding versus formula feeding, and formula feeding versus mixed feeding.

In exclusive versus mixed feeding analyses illustrated in Figure 3.15 B (p. 62):

- The bacterial clades found to be most abundant in mixed feeding were phylum Firmicutes, such as *Clostridia* and *Ruminococcaceae* being observed with LDA scores greater than 4 – accounting for the greatest statistical and biological difference seen in the community. The bacterial clades most underrepresented in mixed feeding, yet abundant in exclusive feeding, were from the phylum Actinobacteria, specifically, *Bifidobacteria.*

In exclusive versus formula feeding analyses illustrated in Figure 3.15 C (p. 62):

- The most differentially abundant species for formula feeding were from phyla Proteobacteria (*Klebsiella*), Bacteroidetes and Firmicutes.

- In the infants that were exclusively breastfed, different genera from the phylum Firmicutes, such as *Lactobacillus* were abundant, as well as Actinobacteria being observed as underrepresented in samples from formula feeders.

In formula versus mixed feeding analyses illustrated in Figure 3.15 D (p. 62):

- The most differentially abundant bacterial clades were from the phyla Firmicutes and Bacteroidetes for both feeding practices. However the taxa differed, where mixed feeders had high representation of clades such as *Lactobacillus* and *Rikenellaceae*, and formula feeders had *Lachnospiraceae* and *Alloprevotella*.

III. Taxonomic analysis relative to time point and medication

The QIIME2 analyses focused on all phylum-level abundance across the samples according to the different time points (Time0, Time6, Time9, and Time12 – Figure 3.16 A, p. 64). The biodiversity between the different groups showed that:

- Firmicutes showing an increased abundance level from 55% at Time0, to 86% at Time12.
- Proteobacteria also showed a change in abundance from 71% at Time0 to 54% at Time12, and even lower estimates at Time6 (48%) and Time9 (41%).
- Fusobacteria, a phylum present in a minority of samples, had increase in abundance observed from 1% at Time0, to the highest abundance observed of 34% at Time9 (Table 3.11, p. 61).

Table 3.11: Estimates of phyla observed in majority of the samples highlighting the dynamic nature of the infant gut microbiome with changes in abundance over time

	Phylum	Time point			
		Time0 (%)	Time6 (%)	Time9 (%)	Time12 (%)
Majority of samples have abundance of these phyla	**Firmicutes**	55	74	60	86
	Bacteroidetes	60	68	58	69
	Proteobacteria	71	48	41	54
	Actinobacteria	27	26	22	26
	Verrucomicrobia	27	36	27	28
	Fusobacteria	1	13	34	18

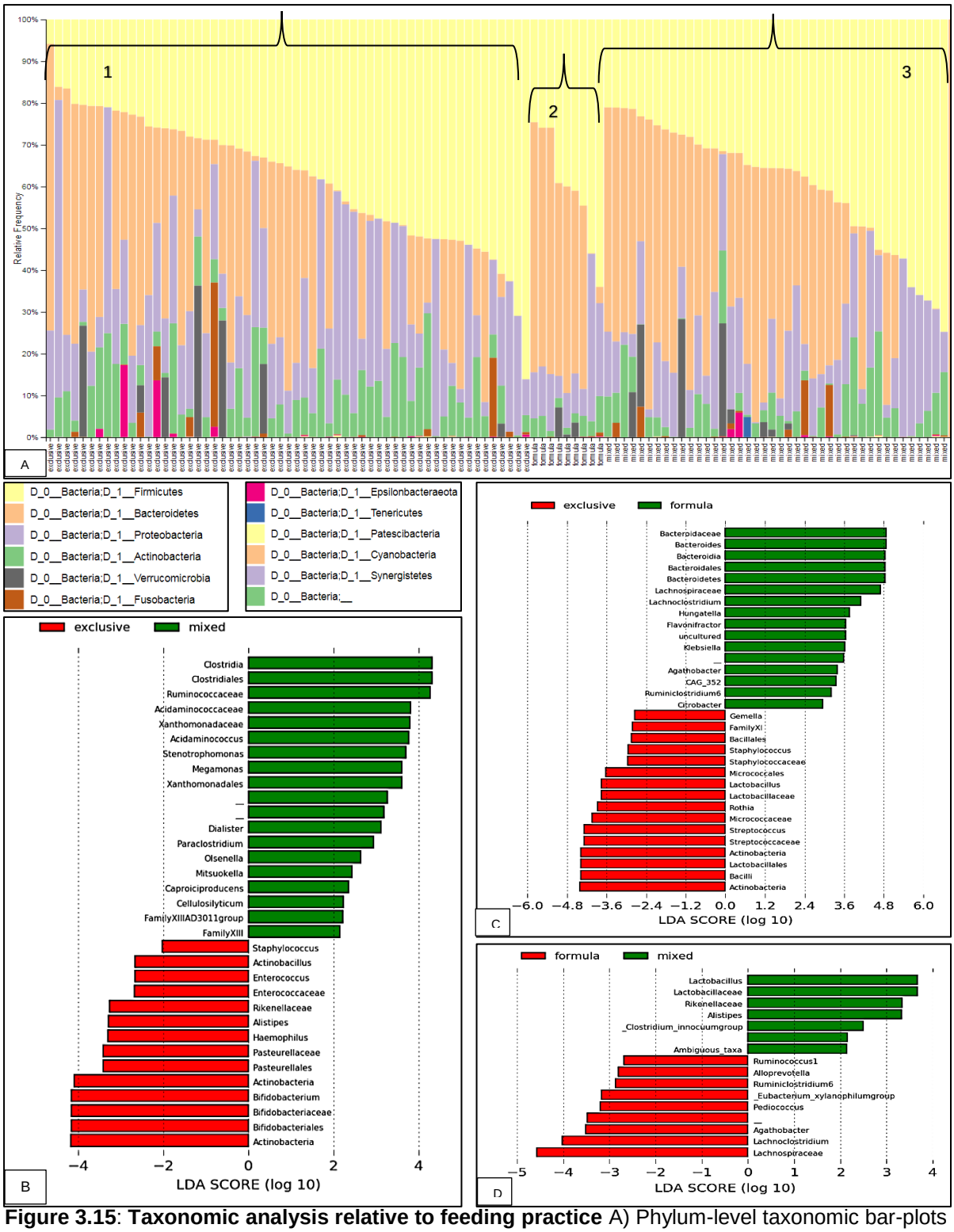


Figure 3.15: **Taxonomic analysis relative to feeding practice** A) Phylum-level taxonomic bar-plots corresponding to exclusive (1), formula (2) and mixed-feeding (3) practices, showing that the lowest microbial diversity is within formula-fed infants. Genus-level histograms of the LDA scores computed for features most differentially abundant between: B) **exclusive versus mixed**: where Firmicutes phyla are most abundant in mixed samples, and *Bifidobacteria* are most differentially abundant in exclusive samples; C) **exclusive versus formula**: where formula samples are abundant in Proteobacteria, Bacteroidetes and Firmicutes, and exclusive samples are abundant in different Firmicutes genera and Actinobacteria; D) **formula versus mixed**: where bacteria from Firmicutes and Bacteroidetes are the most differentially abundant.

The LEfSe genus-level analysis illustrated the most differentially abundant bacterial clades between time points of interest: at recruitment (Time0: 12-21 weeks) and nine months (Time9).

In Time0 versus Time9 analysis illustrated in Figure 3.16 B (p. 64):

- Differing genera that fell under the phyla Bacteroidetes and Firmicutes were found to be either overly abundant or underrepresented in the different groups.
- At Time0, taxa under these phyla such as *Prevotellaceae* (Bacteroidetes), and *Agathobacter* (Firmicutes), were found to be the most differentially abundant.
- At Time9 some of the differentially abundant taxa highlighted were *Erysipelatoclostridium* (Firmicutes), *Oscillibacter* (Firmicutes), and *Bacteroidaceae*.
- The phylum found to be unique in its differential abundance was Actinobacteria, where taxa such as *Actinomyces* and *Eggerthellaceae* were highlighted for Time0.

Additional analyses were performed using LEfSe, to investigate differential abundance between samples from infants that had the potential confounding effect of medication administered at recruitment and the nine month collection ("Yes"), and those that did not have any medication ("No"). A histogram, indicating LDA scores was generated – as with previous analyses – as well as a cladogram highlighting the phylogenetic relationships between the differentially abundant features (Figure 3.17 A, B, p. 65; see also Appendix M, p. 156 for additional plots).

In medication administration analyses illustrated in Figure 3.17 A, p. 65:

- Proteobacteria were differentially abundant in the samples from infants with medication administered. Firmicutes (*Clostridia*) were differentially abundant in samples from infants that did not have medication administered.

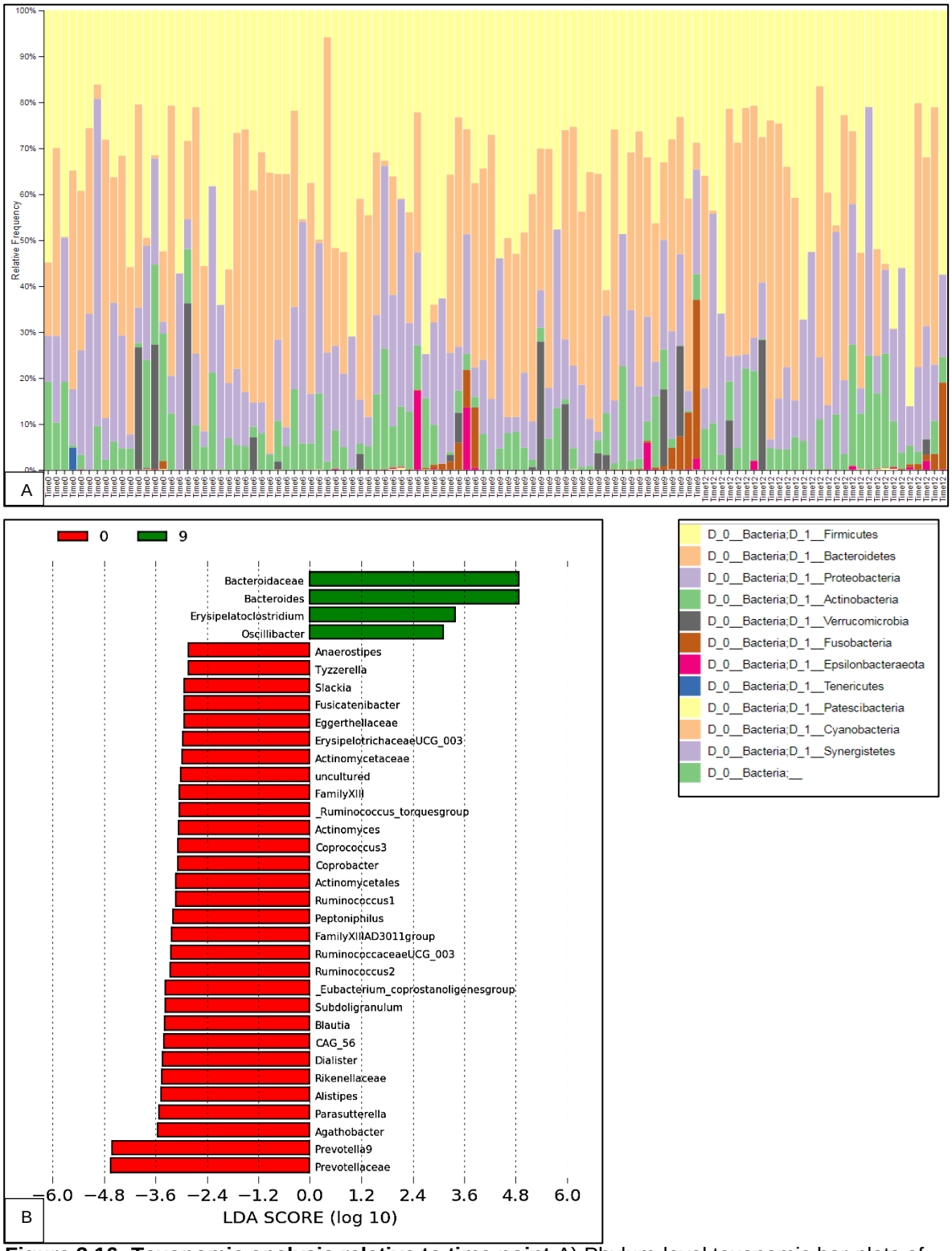


Figure 3.16: Taxonomic analysis relative to time point A) Phylum-level taxonomic bar-plots of samples corresponding to Time0, Time6, Time9, and Time12: showing the dynamic change in microbial composition and abundance over time, with Firmicutes being the most represented. B) Genus-level histograms of the LDA scores computed for features differentially abundant between Time0 and Time9: showing the main phyla – both over and underrepresented by different genera – as Bacteroidetes and Firmicutes at both time points. Actinobacteria is uniquely highlighted as abundant at Time0.

Figure 3.17: Taxonomic analysis relative to administration of medication LEfSe analyses on medication taken (Yes - green) or not taken (No - red) at Time0 and Time9 sample collection. A) Cladogram of the differentially abundant clades in the No and Yes group; B) Histogram of the differentially abundant clades in the No and Yes group – showing that Proteobacteria are differentially abundant in medicated groups, and Firmicutes (*Clostridia*) are differentially abundant in non-medicated groups.

CHAPTER FOUR - DISCUSSION

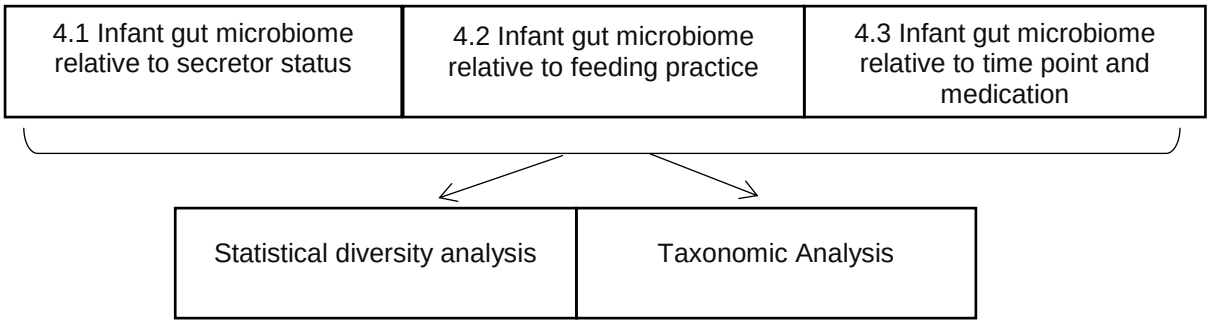


Figure 4.1: **General overview of the discussion section** highlighting the statistical diversity analysis and taxonomic analysis of the research objectives.

In this project, we assessed:

- the dynamic change in the infant microbiome over a time period
- the impact of different feeding practices
- the impact of maternal genetics

On the diversity and composition of infant gut microbiome profiles in South African infants living in Soweto, South Africa. In the following sections, a brief overview of results followed by a discussion is presented.

4.1 Infant gut microbiome relative to secretor status

Overview

Both secretors (n = 34) and non-secretors (n = 9) (*FUT2* -rs601338) were observed (Figure 3.4, p. 44). The frequency of the variant "A" allele in this cohort was higher than the global average (Soweto cohort: 38% vs global average: 32%; Figure 1.3, p.13; Table 3.2, p. 40). This allele was observed at a lower frequency in this cohort than in other African populations (Figure 1.3, p. 13) investigated by the 1000 Genomes Project (KGP) (The 1000 Genomes Project Consortium, 2015). Alpha and beta diversity measures indicate that maternal secretor status does not appear to influence microbiome diversity in this cohort (Appendix I, p. 136). The taxonomic phylum-level analysis revealed similar levels of abundance between infants of secretors and non-secretors, in line with the statistical measures (Figure 3.14 A, p. 59). Genus-level LEfSe analysis (Figure 3.14B, p. 59) illustrated clades of bacteria from the same phylum were differentially abundant.

In this study n = 43/46 mothers were successfully screened for the *FUT2* gene mutation (rs601338), using Oragene saliva self-collection kits and TaqMan genotyping assay. The

genotype results (Figure 3.4, p. 44; Figure 3.5, p. 44) show that the “G” allele frequency accounts for 61.6% of the cohort, and “A” is in 38.4%. This is in line with the KGP global frequencies, where the “G” and “A” allele is present on average in 68% and 32% of populations, respectively (Figure 1.3, p. 13). In the KGP African population frequencies, “G” is present in 51%, and “A” in 49% of the population (https://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=19:48702917-48703917;v=rs601338;vdb=variation;vf=45189331#population_freq_AFR Accessed: November 2019). Thus, this cohort shows a higher frequency of the alternative allele compared to other African populations (e.g. Nigerian, Gambian). The KGP African sub-population frequencies do not include results from the South African population however, the West African Mende population from Sierra Leone show similar frequencies to this cohort at “G” in 61% and “A” in 39%.

Sanger sequencing results were inconsistent and the different sequence variants could not be deduced from the electropherograms. This was as a result of PCR amplification failure potentially caused by high GC-content (60%) (Yip *et al*., 2007).

When analysing the genotype data, shown as distinctive clusters in Figure 3.14 (p. 44), the results from the assay demonstrated that 79.1% of the cohort are secretors (GG and GA) and 20.9% are non-secretors (AA) (Figure 3.5, p. 44). The secretor status of mothers is relevant due to the strong associations of maternal genetics shaping the composition of the infant gut microbiome diversity (Kumar *et al*., 2015; Charbonneau *et al*., 2017). Host secretor status is particularly evident in shaping the *Bifidobacteria* content in infants that breastfeed (Wacklin *et al*., 2011). This is due to breast milk acting as a probiotic (Boesten and Vos, 2008), where the transfer of bacteria from mother to infant occurs in feeding; as well as a prebiotic, where HMOs complexed with sugars such as fucose in breast milk act as a substrate for present *Bifidobacteria* and other early bacterial colonizers (Thomas and Greer, 2010).

Research has shown that mutations in the *FUT2* gene determining secretor status can modulate immune responses, and may play a role in evolutionary survival of hosts to pathogenic outbreaks (Linden *et al*., 2008; Anstee, 2010). The non-secretor phenotype, seen at a comparatively lower frequency in this cohort, is genetically associated with increased risks to disease such as Crohn's disease, as well as susceptibility to several infectious diseases (McGovern *et al*., 2010). In contrast to these vulnerabilities, it may also confer a reduced risk to diarrhoea caused by certain genotypes of Norovirus, which is a contagious pathogen most implicated in intestinal infections in the world (Larsson *et al*.,

2006). The research highlighting associations between secretor status and susceptibility to the aforementioned diseases, has also been linked to the role of shaping the composition of gut biodiversity (Wacklin *et al*., 2011). The analysis of gut microbiota diversity in host samples containing *Bifidobacteria* or a strain mixture of this bacteria, have shown that it has health promoting effects on humans, whereas reduced *Bifidobacteria* abundance has been connected to intestinal disorders (Boesten and Vos, 2008).

The interrelationship between the gut microbiome, malnutrition and maternal *FUT2* gene profiles has been highlighted in research, demonstrating a microbiota-dependent promotion of improved infant growth (Charbonneau *et al*., 2017). However, the relatively significant frequency of non-secretors does not directly correlate with the global prevalence of stunting (20.8% - https://www.who.int/data/gho/data/indicators/indicator-details/GHO/stunting-prevalence Accessed 21 May 2020). Charbonneau and colleagues (2017) showed that among secretor mothers, there were no significant differences in HMO abundances observed between those with healthy infants and with stunted infants. This may be interpreted as some non-secretor mothers being unable to compensate for deficiencies in fucosylated HMOs through the increased output of other HMOs. This results in breast milk that is less supportive of positive infant growth outcomes (Charbonneau *et al*., 2017). Therefore, it can be postulated that in global populations of non-secretor mothers with healthy infants, the compensation of HMO production through other (e.g. sialylated) HMOs may account for the lower frequency of stunted infants.

In this study, the infant growth measurements (Appendix C, p. 119) were within range of WHO Child Growth Standards (de Onis and WHO, 2006). This may be as a result of optimal breast milk for growth conditions in secretor mothers that are breastfeeding their child. As reported, secretor status is strongly associated with the growth of *Bifidobacteria* in the infant gut (Wacklin *et al*., 2011), by acting as a source of bacteria and substrate for resident bacteria (Boesten and Vos, 2008; Thomas and Greer, 2010). Stunting tends to be associated with a non-secretor phenotype: however due to the research highlighted (Charbonneau *et al*., 2017), mothers in this cohort may have compensated for the lack of fucosylated-HMOs by producing other HMOs. Therefore, a direct correlation to stunting could not be concluded. Investigating the infants in this cohort over a longer period of time – approximately up to three years of age – may reveal more observable stunting phenotypes at a later stage. In addition, analysing mothers for the presence of other substrates such as sialylated-HMOs, may reveal their protective role in the South African population.

Other similar research (Davis *et al*., 2017) has shown that relative total sialylation was not a predictor of infant growth, as not all HMO structures were associated with better growth. It is yet to be determined in non-secretor mothers: whether up-regulation of other genes involved in breast milk composition, or other compositional changes in breast milk are responsible for protective mechanisms that ensure energy resources are being directed towards healthy infant growth outcomes. In the following sections, we discuss the results of secretor status associated with infant gut microbiome diversity, and whether (in infants) the maternal secretor phenotype displays greater taxonomic diversity compared to maternal non-secretors.

4.1.1 Secretor status statistical diversity analysis

The assessment of alpha diversity using Faith's PD and Pielou's Evenness yielded *P*- and q-values well above the threshold of significance of $P \leq 0.05$ (Appendix I, p. 136). The Kruskal-Wallis pairwise statistical tests showed that there was no significant association between infants from mothers with a different secretor status, with phylogenetic indices observed on box-plots resulting in the same values for secretor and non-secretor groups (Appendix I, p. 136). The alpha rarefaction curves generated for additional diversity measures show that all species richness and evenness was fully observed using the Shannon Diversity metric (Appendix J, p. 139). When computing the curve using Faith's PD and amplicon sequence variants as the diversity metric, the continuous growth of the curve suggests that with the addition of sequences, more rare species may be observed (Appendix J, p. 139; Figure J 3.4, p. 137).

PCoA plots illustrating the relationship of beta diversity with secretor versus non-secretor status samples did not illustrate distinct clusters of these samples. However, in all but one of the beta diversity metrics used to account for distance measures, the secretor samples represented the greatest percentage of variation in distance measures on the x- and y-axes, by having a higher loading score corresponding to the variable with the larger effect (Appendix K, p. 150). The statistical tests carried out using PERMANOVA and Unweighted Unifrac distances yielded a *P*- and q-value of 0.054 (Appendix L, p. 154), which is trending towards significance ($\alpha \leq 0.05$). Thus, the results from beta diversity estimates did not indicate that the distance measures between the secretors sample group was more similar than distance measures between the non-secretor samples. Therefore, the biodiversity between the groups was not significant in this cohort. We hypothesize that the relatively small sample size (n = 101) for secretor status analysis – in order to only account for mothers that performed some form of breastfeeding and excluding inconclusive sequencing

results – was a limitation for identifying a statistically significant effect on infant gut microbiome diversity as reported in literature.

4.1.2 Secretor status taxonomic analysis

In the phylum-level analysis of infant gut microbiome composition between secretor and non-secretor groups, samples were analysed to determine the highest percentage levels of abundance of different bacterial groups present in the cohort. This analysis highlighted the four groups of phyla most represented in gut microbiome research – Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria – as well as an additional two phyla of interest due to their noticeable levels of abundance in some samples, namely Verrucomicrobia and Fusobacteria (Figure 3.14 A, p. 59) The abundance levels seen amongst the samples with the highest proportions of the four main phyla were relatively similar between secretor and non-secretor groups, which seem to align with the statistical diversity estimates not resulting in *P*-values of significance (Appendix I, p. 136). The most abundant phyla observed is that of Firmicutes (~71%), which at the genus-level of analysis of secretor status, are most differentially abundant and depleted by *Lachnoclostridium* and *Veillonellaceae,* in secretor and non-secretor groups respectively (Figure 3.14 B. p. 59).

The presence of bacteria from the phyla Firmicutes correlate to the ability of a host to extract energy from their diet (Turnbaugh *et al*., 2006). Firmicutes abundance in infants has been shown to be associated with enrichment in glycan and urease metabolic pathways, which are involved in the use of nitrogen and glycans as an energy source from breast milk (Arrieta *et al*., 2014). The most differentially abundant taxa in secretors observed as *Lachnoclostridia* have been highlighted to have a positive association with dietary metals and processed foods, such as sodium and mannitol. Research has shown its involvement in the metabolism of amino acids, energy production, nucleotide synthesis, and xenobiotic metabolism of dietary nutrients, which are all related to important processes for development of the infant (Tang *et al*., 2019).

The over-abundance of this taxa in secretors compared to non-secretors may relate to the added advantage of early diets of fucose containing breast milk in infants of secretor mothers, which can act as substrates for beneficial bacteria and aid in the persistence of these bacteria as complementary feeding practices commence. *Veillonellaceae* are the most differentially depleted bacteria in the secretor group and are also lactose fermenting bacteria; however, they persist at a lower abundance than the more significantly highlighted *Bifidobacteria (*Arrieta *et al*., 2014). A study (Charrier *et al*., 2006) showed that although functional differences amongst groups of infants were not significantly identifiable, the

functional abilities of the gut microbiome increase over time. One of these functions are demonstrated by *Veillonellaceae*. These bacteria are associated with butyrate metabolism and they utilize larger concentrations of *Bifidobacterial*-derived acetate that is present in the gut. Abundance of *Veillonellaceae* likely reflects a maturation of the microbiome by selective pressures, resulting in the growth of communities of better-adapted species (Davis *et al*., 2017).

In a similar study by Wacklin and colleagues (2014), the intestinal microbiota of infants of non-secretors and secretors were dominated by Firmicutes and significantly influenced by secretor phenotype, with similar abundances at the phylum level, but differences noted at the genus level e.g. *Clostridia*. However, when the *P*-values for the statistical tests were corrected for the number of tests performed, the significance of these associations was lost. This shows that although associations are observed between maternal genetics and infant gut microbiome diversity, additional influences may need to be considered to identify the full extent of contributing factors. For instance, Firmicutes are hypothesized to reflect climate where its dominance is associated with colder climates necessitating increased body fat percentage. It is also associated with increased maturation of the microbiome after cessation of breast milk (Stewart *et al*., 2018). More studies are showing that the composition of breast milk seems to be dictated by maternal genetics, as well as the environment, and lifestyle can confer a more protective microbiome profile associated with lower rates of infection or inflammation, allowing the infant to invest energy in growth (Davis *et al*., 2017).

4.2 Infant gut microbiome relative to feeding practice

Overview

The association of feeding practice with gut microbiome composition showed that a 56% (n = 24) majority of genotyped mothers practiced exclusive breastfeeding for up to six months in this cohort. Results showed that a greater microbiome diversity was observed in breastfed (exclusive or mixed) infants, compared to formula fed infants which showed the lowest levels of diversity (Table 3.10, p. 60). This finding was further supported by statistically significant measures of beta diversity in infants with exclusive breast milk diets compared to formula fed infants (Table 3.7, p. 55). Taxonomic analyses showed that Actinobacteria were consistently represented as abundant in exclusive breast milk diets (Figure 3.15B & C, p. 62). Genera within the Bacteroidetes phylum were represented as differentially abundant for formula feeding (Figure 3.15C & D, p. 62).

With regard to feeding practises within our entire cohort – 54% (n = 25/46) practiced exclusive breastfeeding up to six months of age, 29% (n = 13/46) practised mixed feeding which involves a combination of breast milk and formula, and only 17% (n = 8/46) practised formula feeding (Figure 3.2, p. 42; Table 3.2, p. 40). By nine months (Time9) most mothers reported that complementary feeding practices had been introduced. This involved the introduction of solid foods like porridges – either commercial starch based foods such as oatmeal, homemade purees such as mealie meal, fruit and vegetable blends, and more complex carbohydrates and animal proteins. These complementary foods associated with complex food groups and mixtures of "adult-like" foods, which would potentially cause a shift in the gut microbiome composition. In the following sections, the difference in microbial diversity in association with different feeding practices of exclusive, mixed and formula feeding are discussed. This is in order to characterize and identify the feeding practices associated with increased microbial diversity.

4.2.1 Feeding practice statistical diversity analysis

In alpha diversity analyses for Faith's PD and Evenness there were no statistically significant differences found between feeding groups (Appendix I, p. 136). The alpha rarefaction plot, used as a quality control measure for the sequencing data computed using the Shannon diversity metric, identified that all species were captured although no significant differences were identified (Figure 3.11, p. 52). The plots computed using amplicon sequence variants and Faith's PD show a slight gradual and continual growth of the curve (Appendix J, p. 139). This may indicate that with an increased sampling depth (>24 000), rarer or less abundant species may have been identified that could have resulted in statistically significant differences between feeding practices (Appendix I, p. 136). The accuracy of the identification of bacteria using NGS is greatly influenced by the use of controls and the DNA extraction method employed, which may affect the purity of the DNA and subsequent downstream NGS analyses (Hart *et al*., 2015). An over- or under-representation of different bacteria can occur and influence the diversity analyses, by affecting the capturing of rarer or less abundant species by rarefaction.

The recorded alpha diversity results are inconsistent for different studies. One study demonstrated that the average alpha diversity of the gut community from infants receiving breast milk was higher than those not receiving breast milk (i.e. formula fed), with *P*-values < 0.001 (Cong *et al*., 2016). However, another study also found no differences in alpha diversity for breastfed versus mixed fed infants. They hypothesize that this may be due to 27.7% of gut bacteria from breastfed infants being acquired from breast milk, where gut microbial changes are associated with breast milk intake in a dose-dependent manner

(Pannaraj *et al*., 2017). Therefore, a lack of quantifiable data of daily milk intake may eliminate the microbial differences between the feeding practices recorded (Liu *et al*., 2019).

Beta diversity measures were evaluated from PCoA, and although distinct clusters were not observed in the PCoA plots between feeding practices (Appendix K, p. 150), the variable that had the largest effect on the degree of distances between the groups was exclusive breastfeeding. Beta diversity analysis was carried out using PERMANOVA (in QIIME2), where the degree of "clumping together" within and between samples of different feeding groups were assessed using Unweighted Unifrac distances (Table 3.8, p. 55). Statistically significant distance measures were identified between exclusively breastfed and mixed fed infants ($P = 0.008$), and exclusively breastfed and formula fed infants ($P = 0.025$). This highlights a greater degree of diversity within exclusively breastfed infants that separates them from the other feeding practices, when phylogenetic features are computed. The data from the box-plots generated using PERMANOVA for beta diversity measures illustrate that when outliers are included in the distance measures between groups, samples from formula feeders have the smallest distances (Figure 3.13, p. 56). There is also a greater degree of similarity of species for within and between formula sample analyses. This further highlights the greater gut microbial diversity observed within exclusive breast feeders and breast milk-containing diets.

4.2.2 Feeding practice taxonomic analysis

Taxonomic analysis of the different feeding practices identified the six previously mentioned phyla as the most representative of gut microbiome diversity (Table 3.10, p. 60). Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Verrucomicrobia and Fusobacteria have the highest abundance in exclusively breastfed infants, and the lowest in formula fed infants. Similar levels of Bacteroidetes and Actinobacteria are observed between exclusively breastfed and mixed fed groups, and similar levels of Proteobacteria are seen between mixed fed and formula fed infants (Figure 3.15, p. 62).

In mixed-fed and exclusively breastfed sample analyses (Figure 3.15B, p. 62), the most differentially abundant phyla were from Firmicutes and Actinobacteria, respectively. The bacterial clades identified in mixed feeding are *Clostridia* and *Ruminococcaceae* – which fall under the same phylogenetic grouping – and *Bifidobacteria* within exclusive breastfeeding. Research has shown that during the first month of life, *Bifidobacteria* dominates the gut, and that the introduction of alternative and complementary foods leads to an increase of *Clostridia* taxa from the family *Ruminococcaceae* (Arrieta *et al*., 2014; Bergstrom *et al*., 2014). These complementary foods result in increases in dietary fibre from bread, or protein

from dairy, cheese and meat (Zhu *et al*., 2012). The stool samples collected were at different time points, and these bacteria potentially highlight the change in feeding practice over time, as different type of foods were being introduced as the infant aged. *Bifidobacteria* are essential for the breakdown of particular oligosaccharides present in breast milk (i.e. human milk oligosaccharides/HMOs). They play an important role in the child's immune system development with reduced incidence to infections. Thus, its presence is beneficial for child health and is strongly associated with breastfeeding in this cohort (Sela and Mills, 2010; Underwood *et al*., 2014; Victora *et al*., 2016).

In the analyses assessing exclusive breastfeeding versus formula feeding (Figure 3.15C, p. 62), the most differentially abundant phyla in formula feeding are Proteobacteria, Bacteroidetes and Firmicutes. The species highlighted from Proteobacteria is *Klebsiella*, which are facultative anaerobes (which can synthesize energy in the presence or absence of oxygen), and is typical of the early infant gut. This species is a common opportunistic pathogen in hospitals. They may be associated with formula feeders that potentially have less of the protective microbial profile associated with *Bifidobacteria* (Actinobacteria), which is abundant in breastfed infants. *Klebsiella* have been flagged as clinically important due to its propensity for translocation into the blood stream which may cause adverse health outcomes (Kigbu *et al*., 2016).

In contrast to the negative implications of these differentially abundant bacteria, the biosynthetic pathway for cobalamin (vitamin B12) is well represented in the genomes of various bacteria including *Bifidobacteria* (Yatsunenko *et al*., 2012). This is a positive presentation of the significance of these bacteria in developmental processes. *Lactobacillus* species which are differentially abundant from the Firmicutes phylum in exclusive breast-feeders are known to dominate breastfed infant guts as well (Adlerberth and Wold, 2009). This may be due to a positive correlation between fucosylation of breast milk content and *Lactobacillus* that was highlighted in a study (Davis *et al*., 2017). Specifically, that the *Lactobacillus* species contain genes encoding fucosidases that degrade fucosylated oligosaccharides in breast milk (Rodriguez-Diaz *et al*., 2011). These dietary influences that impact the composition of potentially pathogenic or beneficial bacteria persist until eighteen to thirty-six months of age; thereafter the gut microbiome metabolizes and forms an adult-like state (Brewster *et al*., 2019).

In mixed- versus formula-feeding analyses (Figure 3.15D, p. 62), the taxa unique to this comparison were *Rikenellaceae* in mixed feeders, and *Lachnospiraceae* and *Alloprevotella* in formula feeders. A previous African-related study (De Filippo *et al*., 2017) showed that

higher abundances of *Rikenellaceae* (Bacteroidetes) were seen in infants living in the more urban capital city area, with similar abundance levels to European children. *Lachnospiraceae* (Firmicutes) are responsible for metabolism of diverse plant polysaccharides, and *Alloprevotella* (Bacteroidetes) with the metabolism of secondary bile acids (Tang *et al*., 2019). There is a need for additional studies to further elucidate the role of these taxa in infant guts, and uncover the potential cause of the differences between these feeding groups, which can persist for up to 14 months (Stewart *et al*., 2018). The ratio between Bacteroidetes and Firmicutes – the corresponding phyla – correlates to the ability of the host to extract energy from their diet (Turnbaugh *et al*., 2006). Species from Bacteroidetes tend to be associated with diets consisting of plant-derived polysaccharides typical of more rural African populations. A higher ratio of Firmicutes to Bacteroidetes – as observed in this cohort – is associated with a westernized diet with more animal and dairy-derived nutrients (De Filippo *et al*., 2017). This ratio may reflect the introduction of complementary foods as the infant aged, and highlights the differences between infant diets from an urban setting.

4.3 Infant microbiome relative to time point and medication

Overview

For the objective of characterizing the infant gut microbiome profile – particularly over time for “at recruitment” (Time0; 12-21 weeks) and nine months (Time9) time points – there was an overall increase in diversity. Measures of alpha diversity highlighted significant differences for comparisons of time points (Figure 3.7, p. 47; Figure 3.9, p.49) and infant sexes (Figure 3.8, p. 47; Figure 3.10, p. 50). Beta diversity estimates highlighted significant differences for time points when comparing Time12 to Time0, Time6 and Time9 (Table 3.9, p. 57). Taxonomic phylum-level analysis of the time points (Table 3.11, p. 61) showed a general increase in abundance: with Firmicutes being the most abundant, but Proteobacteria showed a decrease in abundance over time. The most differentially presented phyla were Bacteroidetes and Firmicutes with different genera representing the groups, and Actinobacteria were uniquely abundant at Time0 (Figure 3.16, p. 64). At Time0 and Time9 there was a higher abundance of Proteobacteria in infants with medication administered, and a higher abundance of Firmicutes in those without medication.

The infant gut microbiome samples that were profiled revealed 1 384 features of representative sequences. These indicate the proportion of different species of bacteria identified after quality control processes (Appendix D, p. 127). There were 114 samples (refer to section 3.5.1, p. 46) corresponding to over three million sequences – this is expected given that there are millions of gene sequences belonging to microbial organisms

within the human gut. Of these, over 130 000 sequences matched to a reference sequence within the 16S rRNA gene database (SILVA) for identifying bacteria – the focus of this project. The sequencing, quality control processes used, and the subsequent output is comparable to other groups that have done similar studies (Arrieta *et al*., 2014; Jovel *et al*., 2016; Stewart *et al*., 2018; Liu *et al*., 2019). However, great variability is seen in DNA extraction methods, where adjustment of protocols to increase DNA yield can alter the microbial profiles seen, and affect the ability of downstream NGS applications to efficiently and accurately identify all bacteria present in a sample. The lack of use of positive and negative controls to ensure validity of the results is of note, as a masking or signalling of contaminated samples and untrue biological sequences may occur. In similar research, the QIAamp PowerFaecal DNA kit – used in this study – demonstrated efficiency in stool DNA extraction and subsequent gut microbiome profiling (Hart *et al*., 2015, Liu *et al*., 2019).

4.3.1 Time point statistical diversity analysis

In alpha diversity analysis using the Faith's Phylogenetic Diversity (Faith's PD) metric, a significant difference in species richness was observed in Time12 versus Time0, Time6 and Time19 (Figure 3.7, p. 47). The most significant *P*-values were observed for Time12 versus Time0 and Time9 (*P* and q = 1.00×10^{-3} – Table 3.3, p. 48). This highlights the dynamic nature of the infant gut over time with different species being required for varying roles, as the feeding practices change, and the infant grows and develops. This also indicates interactions with their environments, as they begin to explore their surroundings and are exposed to more microbes (Laursen *et al*., 2015). However, there were no significant differences seen for comparison between the time points when Time12 was not included (Table 3.3, p. 48). The greatest species richness similarity indicated by an insignificant *P*-value, was observed for comparisons between Time0 and Time9 (*P* = 0.910). This observation may be due to the advent of changes in feeding practices, resulting in a growth of newly better-adapted species for complementary foods. Therefore, the species richness at Time0 and Time9 may be similar but with different bacteria.

A higher diversity index was seen in males versus female infants (*P* = 0.016; Figure 3.8, p. 47). This may relate to a hormonal change in the body as the male and female infant develops (Gomez *et al*., 2015). This relates to a stage of infancy called "minipuberty" during the first 3-6 months of life, where the hypothalmic-pituitary-gonadal (HPG) axis is activated. This leads to a rise in levels of testosterone and oestradiol hormones for the maturation of sexual organs, and thereafter remains dormant till puberty (Lanciotti *et al*., 2018). However, due to the analysis of male and female infants occurring over all the time point sample

collections, it would be difficult to conclude the significant changes observed were driven by sex differences.

Since Faith's PD incorporates phylogenetic relationships between the features represented, more bacterial groups that are present and related to each other are identified as the gut microbiome matures. This may increase the species richness likely leading to an increase in redundancy of functions of similar microbes present. When alpha diversity was measured with Pielou's Evenness metrics, a significant difference was observed for Time0 versus Time6 and Time9, with a more significant *P*-value seen between Time10 versus Time9 (*P*=0.002; Table 3.5, p. 50). As alluded to earlier, the species richness levels may have been similar, but the present community abundance was observed to be different. A higher community evenness index was also observed in males over female infants (Figure 3.10, p. 50). This analysis was also performed across all time points, making it difficult to conclude the driving factor was the difference in sex. However, these results are supportive of an increase in microbial community – although likely redundant in features (HMP Consortium, 2012). Similar microbes may be present (richness) but in different proportions (evenness), which likely increase in abundance simultaneously with age and developmental processes.

A study by Cong and colleagues (2016) showed that male infants begin with a lower diversity index than females shortly after birth, and gradually appear similar between the sexes over time with gut microbial diversification. Further contrasts in diversity estimates between sexes in different studies may be present due to the added effect of clinical factors that may influence the microbiome development such as: mode of delivery, premature rupture of membranes (PROM), antibiotic use, and feeding type (Cong *et al*., 2016). In this research, the use of medication and feeding practices were investigated and represented relatively equally between the sexes (Table 3.2, p. 40). However, their effect on biasing the diversity estimates were tested over all sample collection points and may mask the confounding effect of age at sampling. In addition, the other clinical factors and additional environmental exposures as the infant develops may influence the differences in species richness and evenness.

Differences in the gut biodiversity observed in sexes is further complicated by the fact that most human studies that have reported sex-specific differences in gut microbial diversity are few and inconsistent. These studies speculate the variation in the microbiome is driven by hormone-immune-microbe interactions and genetic traits (Karunasena *et al*., 2014; Gomez *et al*., 2015), which may explain the differences in male and female gut microbiome profiles but may not be relevant in prepubescent individuals. A study investigating arsenic exposure

and sex-specific effects highlighted that males are more susceptible to the effects of the poison on their microbiota than their female counterparts (Hoen *et al*., 2018). However, another study investigating the same exposure in mice showed that females were more susceptible to microbiome perturbation (Chi *et al*., 2016). These stark contrasts in studies of humans and animal models demonstrates the complexity of the gut microbiome that is yet to be understood through more research. These should aim to uncover the gene functions and profile of the diversified gut in different sexes.

Research involving African populations and focusing on age-related influence on the gut microbiome, has shown that the phylogenetic composition of bacterial communities evolves toward an adult-like configuration from birth up to three years of life (Yatsunenko *et al*., 2012). This is in line with the results shown here that, bacterial diversity increases with age (Table 3.11, p. 61). The study also demonstrated that when the major driving influence of a particular bacterial species was excluded – the age-related changes were found to involve the metabolism of vitamins B12 (cobalamin) and folate (Yatsunenko *et al*., 2012). Folate is a dietary supplement primarily synthesized by microbes and plants, and cobalamin primarily by microbes (Krautler, 2005). The gut microbiomes of babies in this study were enriched in genes for *de novo* biosynthesis of folate, which decrease with age, while the proportional representation of genes involved in cobalamin biosynthesis increase with age (Yatsunenko *et al*., 2012). This shows evidence that over time different bacteria are present with different functions and levels of abundance, and play a role in the developmental processes required in an infant for optimal growth. Thus, the microbiome should be considered when assessing the nutritional needs of children at these different ages which are associated with different stages of development.

4.3.2 Time point and medication taxonomic analysis

The taxonomic analysis of samples at the different time points identified 11 phyla present in the cohort, however only six phyla are observed at appreciable levels (i.e. >1% of the sample composition). These six phyla in order of overall highest abundance over time are: Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Verrucomicrobia and Fusobacteria (Table 3.11, p. 61). These findings are in line with previous research, describing such phyla as contributing to the majority of gut microbial composition in African and global populations (Arrieta *et al*., 2014; Stewart *et al*., 2018; Brewster *et al*., 2019). In this study we show that the phyla Firmicutes and Bacteroidetes, have a high abundance and show an increase over time. At Time0, the phylum with the highest abundance is Proteobacteria (71%), but at Time6 and Time9 it shows an overall decrease by 30%. From Time9 to Time12 there is an increase in levels from 30 to 54% (Table 3.11, p. 61).

The phylum Proteobacteria is associated with the colonization of the gut by potentially pathogenic microorganisms, and has a higher correlation with dominance in Western guts and African infants living in urban areas (Cong *et al*., 2016; De Filippo *et al*., 2017). In other research characterizing phenotypes of individuals in relation to their gut microbiome, it was found that this phylum is associated with disease phenotypes where increases in its abundance are associated with the onset of disease. Such diseases include necrotizing enterocolitis (NEC), ulcerative colitis (UC) and Crohn's disease (CD) – associating inflamed bowels with dysbiosis, due to the pro-inflammatory nature of species from Proteobacteria (Arrieta *et al*., 2014). High levels of Proteobacteria causing dysbiosis in infants has been linked to preterm infant mortality and morbidity (Zhou *et al*., 2015). Research has also suggested that the different proportions between Proteobacteria and the other dominating phyla such as Bacteriodetes can contribute to the phenotype of a malnourished child: where a higher proportion of potentially pathogenic Proteobacteria in comparison to a lower proportion of energy extracting Bacteroidetes may explain their poor health outcomes (Ottman *et al*., 2012). However, in this cohort we observed a decrease in Proteobacteria over time – concomitantly with a lower proportion of potentially pathogenic bacteria – which implies a healthy diversifying gut profile (Table 3.11, p. 61).

The phylum with the highest abundance at three time points (Time6, Time9 and Time12) is Firmicutes which generally increased over time, however it temporarily decreased in abundance at Time9 (Table 3.11, p. 61). The presence and increased abundance of this phylum is typical of transition to an adult-like microbiota composition, and studies suggest that this trend along with Bacteroidetes is due to increased contact with the environment and other individuals (Laursen *et al*., 2015). Studies of rural African populations (De Filippo *et al*., 2010; De Filippo *et al*., 2017) show a higher proportion of Bacteroidetes to Firmicutes. In urban settings, similar to our study cohort, there is an opposing observation with a higher proportion of Firmicutes to Bacteroidetes. In this study by De Filippo and colleagues (2017), Firmicutes was more abundant in children living in the capital city, similar to the microbial estimates of European children. The observation of a shift of the microbial diversity in other African populations to a more westernized composition as in our cohort, is in accordance with the similar dietary habits and an environment indicative of a more urbanized society (De Filippo *et al*., 2017). In line with other related research our study shows that as the infant ages – particularly around one to two years of age – there is a concomitant increase in the presence of Firmicutes and Bacteriodetes as the microbiome further diversifies (Aakko *et al*., 2015).

In order to further elucidate the potential mechanisms involved in the establishment of the gut microbiome profile, we carried out taxonomic analysis of infants that had or had not taken medication at Time0 and Time9 (Figure 3.17, p. 65). This ranged from simple flu medication (e.g. paracetamol) to full courses of antibiotic treatments, which research has shown has a profound effect on the diversity of the gut microbiome (Fallani *et al*., 2010). Different antibiotics can affect clustering of microbial profiles and may mask the true grouping of infants according to the different pairwise analyses that were performed. The specific treatments taken in this cohort were not captured, and presents a limitation on establishing the precise confounder on this diversity. Studies analysing gut microbial genomic function show that there is enrichment in xenobiotic-processing genes. This is potentially due to therapeutic medication from exposure to antibiotics from food and medicine, increasing the presence of antibiotic-resistance genes in the gut microbiome (i.e. the resistome) (Gomez *et al*., 2016; Brewster *et al*., 2019).

In our sub-set of samples from infants (n = 22) that were administered medication (Table 3.2, p. 40) there was a differentially abundant presence of Proteobacteria, in comparison to a higher presence of Firmicutes in infants that did not take medication (Figure 3.17, p. 65). This finding is in line with previous research that demonstrated that infants exposed to antibiotics have higher proportions of Proteobacteria than unexposed infants, and this effect lasts for up to four weeks after treatment (Fouhy *et al*., 2012). This finding may explain the high abundance of Proteobacteria at Time0 and sudden decrease in abundance of Firmicutes at Time9. The average abundance levels of Firmicutes in the entire sample collection of n = 111 was captured, thus samples from infants not exposed to antibiotics may not have this pattern of decreased levels of Firmicutes. The decrease observed in the sub-set of antibiotic-exposed infants is potentially due to the suppressive effects of an increase of Proteobacteria that tend to rise in abundance with gut dysbiosis. Although Time12 is approximately three months after the administration of medication in this cohort, this could also explain the increased presence of Proteobacteria at Time0 and from Time9 to Time12 – where the average abundance levels of this phylum in the entire cohort was captured (Table 3.11, p. 61). Research has shown that the ecology of the microbial population can be severely altered if exposed to antibiotics too early in development, and for long periods of time (Fouhy *et al*., 2012). Thus, this finding highlights an additional contributing factor to the characterized profile of the infant gut microbiome within this cohort, and its potential limitation in obscuring the true baseline microbiome profile.

Comparison of taxonomic profiles at different time points – Time0 (at recruitment) versus Time9 (nine months)

A total of twenty-five bacterial clades were identified as differentially abundant between Time0 (at recruitment) and Time9 (nine months), with twenty-one of those represented at Time0 (Figure 3.16B, p. 64). All these taxa fall under the phyla Bacteriodetes and Firmicutes in line with other research (Mariat *et al*., 2009). This relates back to the identification of these phyla and their changing abundance levels, as being significant in the establishment of the gut microbiome over time (Table 3.11, p. 61).

At Time0 the most differentially abundant taxa from phylum Bacteroidetes were *Prevotellaceae* (Figure 3.16, p. 64). An abundance of these bacteria is apparently higher in less urbanized regions (e.g. Venezuela, Malawi) with lower cases of immune-mediated diseases, highlighting a possible protective effect that associates with immunity (Arrieta *et al*., 2014). *Prevotellaceae* has been implicated in several metabolic processes such as bile acid metabolism and the breakdown of dietary nutrients, paramount for the adequate use of resources for the growth and development of an infant. In various research (Arrieta *et al*., 2014; De Filippo *et al*., 2017; Brewster *et al*., 2019), this group of bacteria illustrates the variation of the microbiome profile between African and European populations. A higher level of abundance of *Prevotellaceae* is associated with populations from rural areas or small towns in comparison to more urban settings (De Filippo *et al*., 2017).

Although in an urban setting, there were visible contrasts in the housing and hygiene facility availability amongst some of the participants (Nieuwoudt *et al*., 2018). Some mother-infant pairs resided in “modern” homes, made with adequate material (e.g. brick and cement) able to withstand harsh weather conditions that could result in sickness. They also had personal indoor plumbing which would decrease risks of an unhygienic environment. However, other participants had less adequate housing and communal plumbing, which could confer greater susceptibility to disease. This contrast may explain the abundance of *Prevotellaceae* in this cohort, possibly resulting in the infant’s need for increased immune defence mechanisms against these infections (Arrieta *et al*., 2014).

Agathobacter is also highlighted as differentially abundant at Time0 (Figure 3.16, p. 64). These are butyrate-producing bacteria from phylum Firmicutes that are negatively associated with the metabolism of xenobiotics in comparison to *Prevotella*. *Agathobacter* bacteria are known to increase in abundance during a period characterized by changes in feeding practices, and are associated with an increase in diversity over time, forming an adult-like microbiota profile (Arrieta *et al*., 2014; Laursen *et al*., 2017; Tang *et al*., 2019).

Taxa that are found to be unique at Time0 to the phylum Actinobacteria are *Actinomyces* and *Eggerthellaceae*. These bacteria have both been found to also be associated with early age and feeding practices – specifically as taxa associated with breastfeeding (Azad and Ben, 2017; Stewart *et al*., 2018).

At Time9, differentially abundant taxa from *Bacteroidaceae* and Firmicutes were highlighted, with the genera *Erysipelaclostridium* and *Oscillibacter* both falling under the latter (Figure 3.16, p. 64). In a previous study, *Erysipelaclostridium* were associated with the most age discriminatory bacterial taxa identified within the first fifteen months of life: where a high abundance is seen in normal human gut microbiota associated with dairy intake (Stewart *et al*., 2018). *Oscillibacter* is also associated with the metabolism of dietary nutrients such as vitamins and minerals, and is involved in the synthesis of nucleotides and xenobiotic metabolism (Tang *et al*., 2019). This finding may relate to the administration of medication, where the *Oscillibacter* present may be involved in the breakdown and metabolism of antibiotics in the sub-cohort of infants exposed to this medication. This is likely since xenobiotic-processing genes are known to be enriched for the metabolism of foreign substances (such as antibiotics) in the gut microbiome.

Research on European and African populations have noted the differential distribution of Firmicutes and Bacteroidetes between these populations: where Firmicutes are twice as abundant in Europeans, than in Africans (De Filippo *et al*, 2010). These discordant ratios of Firmicutes/Bacteroidetes have been implicated in health and obesity, where studies involving African cohorts demonstrate that increases in Bacteroidetes abundance accompanies healthier outcomes and associated weight loss over time (Grzeskowiak *et al*., 2012). This may be due to abundant species from the Bacteroidetes phylum triggering a protective response from the immune system after delivery (Stout *et al*., 2013). The administration of antibiotics can reduce the beneficial effect of Bacteroidetes before the age of one year, resulting in a relative increase in Firmicutes, sometimes without the complete recovery of the initial microbial profile (Jernberg *et al*., 2010). Although cautious to conclude due to lack of account of confounders and controls, in this cohort we observed results concordant with the published literature.

Limitations of this study

A power calculation was not made, due to the pilot nature of this study with the aim of assessing the feasibility of characterizing the South African infant gut microbiome in an urban setting. Statistically significant differences in diversity profiles in this cohort were reported for time points and feeding practices though we had a relatively small sample size

(due to this being a pilot study). However, this small sample size may have affected identifying associations between secretor status and the gut microbiome as reported in the literature (Wacklin *et al*., 2012; Wacklin *et al*., 2014; Kumar *et al*., 2015). More participants could have been recruited had we had more time, however due to time and funding constraints caused by delays with ethics approval, and the necessity to remain within the budget period, this was not possible. These constraints – along with being nested in an ongoing study pivotal for our recruitment process – were beyond our control.

In addition, the analyses of different groups – such as the child sex analyses – over all the time point sample collections, makes it difficult to distinguish if the significant differences observed are driven by the comparison being made (Estaki *et al,* 2020). The delay in ethics approval and a need to recruit along with the SBW Study, also led to the necessity of grouping a baseline of infant gut samples over a wide range of ages (12-21 weeks, average age = 20.16 weeks). The age range is almost double, and significant changes in microbial diversity from an earlier composition at three months to one at five months was most likely missed.

The use of 16S rRNA sequencing is a cost effective and valuable method for gut microbiome profiling, however it is prone to amplification biases during the PCR process (Kennedy *et al*., 2014). The lack of use of positive and negative controls in this study during extraction, may have resulted in the masking or signalling of potentially contaminated samples. Taxonomic analyses in this study revealed clades of “uncultured bacteria”. This refers to microorganisms that have not been grown – or are difficult to grow – under laboratory conditions and potentially lack a sequenced genome (Fodor *et al*., 2012). Due to the difficulty of growing in external environments, they may be rarer and in smaller quantities in samples. This may end up reflecting as a failure to align to reference genomes during taxonomic analyses, as a result of amplification biases during PCR. These biases may have occurred from inefficient stool DNA extraction methods which affects: DNA concentration and purity; successful NGS amplification; and influences microbial communities seen in NGS output (Hart *et al*., 2015). This can cause adverse effects on the sequencing output by under- or over-representing specific microbial populations in different samples (Henderson *et al*., 2013; Kennedy *et al*., 2014). This may also have an adverse effect on diversity analyses by rarefaction where some rarer or less abundant phyla may not have been captured. Faeces are also known to contain various PCR inhibitors that may affect the PCR reactions for downstream sequencing. However, these inhibitors vary and kit-based DNA extraction methods (such as the QIAamp PowerFaecal DNA kit used in this study) may resolve these issues in approximately 87%-100% of extractions (Hart *et al*., 2015).

An alternative issue causing failure to identify certain bacteria in the sample may be existing reference databases that are well characterized at the higher classification levels, but may not fully capture the gut microbial diversity in African populations, as there is still a substantial but undetermined degree of unclassified microbial diversity within the gut ecosystem (Almeida *et al*., 2019). Metagenomics (shotgun) sequencing is an effective approach that could address the technical issues experienced with 16S rRNA sequencing, since there is characterization of the microbiome and additional analysis of the genomic functions of the microorganisms identified (Hagemann, 2015). With whole-genome shotgun sequencing, the entire genome of the microorganism is sequenced – unlike a specific rRNA gene region in 16S sequencing – and this can result in greater accuracy and precision when identifying the sequenced microorganism by alignment to reference genomes.

The sample sizes for the different analyses were unequal, due to previously mentioned issues with collection of used stool sample kits, and inevitable differences in feeding practices and sexes of recruited participants. Although these differences in sizes were not significantly large, this could have resulted in biases by overrepresentation of bacteria from groups with greater numbers, over those with smaller numbers. The use of antibiotics is also a great limitation that can cause dysbiosis and distort the resident gut microbial profile, affecting clustering of samples in different groups. Due to a lack of statistical adjusting, minimal accouting for missing data and potential confounders in the study, this may cause the significance to be unreliable, since significance testing in itself does not take confounding into account (Skelly *et al*., 2012).Additionally, the exact contents of formula used in feeding practices were unknown and likely non-homogenous between the infants within the cohort. Thus, inferring gut microbial profiles off of the content of the formula diet was not directly addressed.

Future prospects of infant gut microbiome research

Publishing of this research in reputable and relevant journals for maternal and child health, could assist in the communication of key points derived from this study. The reporting of this research in communities and clinics, could help support the declarations made for promotion of exclusive breastfeeding for at least six months. Potential research going forward would be moving beyond observational correlations to define causal relationships and incorporate the functional interactions that exist between microbes and the host. This can be achieved by employing shotgun metagenomic sequencing methods for more comprehensive microbial classification and characterization, in relation to gene function predictions and involvement in host metabolic pathways that are paramount for optimal growth and development. The

recruitment of larger mother-infant cohorts and analysis of infants over a longer period (at least three years old where the gut microbiome stabilizes and forms an adult-like state), could help us address relations to infant growth outcomes where stunting phenotypes are more clearly observed. Additionally this could provide a basis for continuous research in order to further unpack the associations between gut microbiomes and the human host and health outcomes in African populations.

CHAPTER FIVE - CONCLUSION

For this purpose of this pilot study, we demonstrated the feasibility of infant gut microbiome research in the urban South African setting of Soweto, Johannesburg. By collaborating with the Wits DPHRU we were able to recruit mother-infant pairs and facilitate a biospecimen collection. The collection of biological samples was demonstrated to be simple and effective in the setting of the participant's home, and by the participant, demonstrating a collaboration of social science with biological sciences. We managed to extract good quality DNA from the samples collected, and the outsourcing of sequencing (due to unavailable facilities in-house) of the infant stool samples to the NICD was effective with quick turnaround times and good quality sequencing data.

In summary, we were able to effectively characterize the South African infant gut microbiome using longitudinal 16S rRNA data at: Time0 (at recruitment, 12-21 weeks), Time6 (six months), Time9 (nine months) and Time12 (twelve months) time points, to assess the variability with different feeding practices and address the potential impact of maternal genetics on gut microbial composition. Alpha diversity measures using Faith's Phylogenetic Diversity and Pielou's Evenness revealed statistical significance for age and sex analyses. Beta diversity measures were significantly different in phylogenetic features for age and the types of feeding practices. The effect of maternal genetics in terms of secretor status and breast milk composition did not reveal statistically significant results potentially due to the small sample size of this cohort. Taxonomic analysis was in line with previously published data and revealed that the greatest biodiversity among infant samples is seen in exclusively breastfed infants and is lowest in only formula fed infants. Additionally, with time the gut microbiome is dynamic and diversity profiles increase and the ratio of this composition can be affected by the administration of medication. Bacteria from the phylum Firmicutes make up the majority composition of the infant gut microbiome. *Bifidobacteria* from the phylum Actinobacteria is uniquely associated with increased abundance in breast milk-associated diets.

In conclusion, we were successful in using next-generation sequencing and bioinformatics analysis for the characterization of the gut microbiome of clinically healthy infants. Future studies with a larger sample cohort in different geographical regions and adjustment for confounders would enable the generalizability of these results to the South African population. The inclusion of infants representing malnourished or stunted phenotypes is necessary to lead to a greater understanding of the gut microbiome profile associated with these phenotypes. This could address how the state of dysbiosis in those infants relates to

these phenotypes, and further describe the contributing factors that result in these poor health outcomes.

REFERENCES

Aagaard, K., Ma, J., Antony, KM., Ganu, R., Petrosino, J. and Versalovic, J. (2016). The Placenta Harbors a Unique Microbiome. *Sci Transl Med*. **6**: 1–22 [doi: 10.1126/scitranslmed.3008599].

Aakko, J., Endo, A., Mangani, C., Maleta, K., Ashorn, P., Isolauri, E. et al. (2015). Distinctive intestinal Lactobacillus communities in 6-month-old infants from rural Malawi and Southwestern Finland. *J. Pediatr. Gastroenterol. Nutrit*. **61**: 641–648.

Abrahams, Z., McHiza, Z. and Steyn, NP. (2011). Diet and mortality rates in Sub-Saharan Africa: stages in the nutrition transition. *BMC Public Health*.**13**: 801.

Adlerberth, I. and Wold, AE. (2009). Establishment of the gut microbiota in Western infants. *Acta Paediatr.* **98**: 229–238 [doi: 10.1111/j.1651-2227.2008. 01060.x].

Adlerberth, I., Carlsson, B., de Man, P., Jalil, F., Khan, S., Larsson, P. et al. (1991). Intestinal colonization with Enterobacteriaceae in Pakistani and Swedish hospital-delivered infants. *Acta Paediatr. Scand*. **80**: 602–610.

Adlerberth, I., Lindberg, E., Aberg, N., Hesselmar, B., Saalman, R., Strannegård, IL., et al. (2006). Reduced enterobacterial and increased staphylococcal colonization of the infantile bowel: an effect of hygienic lifestyle? *Pediatr. Res*. **59**: 96–101.

Ahmed, T., Auble, D., Berkley, JA., Black, R., Ahern, PP., Hossain, M. et al. (2014). An evolving perspective about the origins of childhood undernutrition and nutritional interventions that includes the gut microbiome. *Ann. N.Y. Acad. Sci.* 1–17. https://doi.org/10.1111/nyas.12487

Almeida, A., Mitchell, AL., Boland, M., Forster, SC., Gloor, GB., Tarkowska, A. et al. (2019). A new genomic blueprint of the human gut microbiota. *Nature*. [doi: 10.1038/s41586-019-0965-1].

Altschul, SF., Gish, W., Miller, W., Myers, EW. and Lipman, DJ. (1990). Basic local alignment search tool. *J. Mol. Biol*. **215**: 403-410. [doi: 10.1016/S0022-2836(05)80360-2]

Amir, A., McDonald, D., Navas-Molina, JA., Kopylova, E., Morton, JT., Zech Xu, Z. et al. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*. **2**: e00191-16 [doi: 10.1128/mSystems.00191-16].

Anderson, MJ. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**: 32-46 [https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x].

Andrews, SC., Robinson, AK. and Rodríguez-Quiñones, F. et al. (2003). Bacterial iron homeostasis. *FEMS Microbiol. Rev*. **27**: 215–237.

Anstee, DJ. (2010). The relationship between blood groups and disease. *Blood*. **115**: 4635–4643.

Applied Biosystems. (2010). TaqMan SNP Genotyping Assays Protocol. Part Number 4332856 Rev. D.

Aronesty, E. (2011). Command-Line Tools for Processing Biological Sequencing Data ea-utils. *Expression Analysis*. [http:// code.google.com/p/ea-utils].

Arrieta, M., Stiemsma, LT., Amenyogbe, N., Brown, EM., Finlay, B., Finlay, B. et al. (2014). The intestinal microbiome in early life : health and disease. *Frontiers in Immunology*. **5:** 427 [https://doi.org/10.3389/fimmu.2014.00427].

Ayeni, FA., Biagi, E., Rampelli, S., Fiori, J., Soverini, M., Audu, HJ. et al. (2018). Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. *Cell Rep*. **23**: 3056–3067.

Azad, SK. and Ben, X. (2017). Pediatrics & Therapeutics The Development and Role of Micro Biome in Infant Gut : An Overview of Recent Literature. *Pediatrics & Therapeutics.* **7**: 2 [https://doi.org/10.4172/2161-0665.1000320].

Bäckhed, F., Ley, RE., Sonnenburg, JL., Peterson, DA. and Gordon, JI. (2005). Host-bacterial mutualism in the human intestine. *Science.* **307**: 1915–1920.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P. et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe*. **17**: 852 [PubMed: 26308884].

Bailey, RL., West, KP., Jr, and Black, RE. (2015). The epidemiology of global micronutrient deficiencies. *Annals of nutrition & metabolism*. **66***: Suppl 2*, 22–33 [https://doi.org/10.1159/000371618].

Baron, EJ. (1996). Chapter 3: Classification*. Medical Microbiology*. 4th Ed. [https://www.ncbi.nlm.nih.gov/books/NBK8406/].

Bergey, DH. and Williams, ST. (1989). *Bergey's manual of systematic bacteriology*. **Vol. 4**.

Bergström, A., Skov, TH., Bahl, MI., Roager, HM., Christensen, LB., Ejlerskov, KT. et al. (2014). Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of Danish infants. *Appl. Environ. Microbiol*. **80**: 2889–2900 [doi: 10.1128/AEM.00342-14].

Bezirtzoglou, E., Tsiotsias, A. and Welling, GW. (2011). Microbiota profile in feces of breast- and formula-fed newborns by using fluorescence in situ hybridization (fish). *Anaerobe*. **17**: 478–82 [doi:10.1016/j.anaerobe.2011.03.009].

Biasucci, G., Benenati, B., Morelli, L., Bessi, E., and Boehm, G. (2008). Cesarean delivery may affect the early biodiversity of intestinal bacteria. *J. Nutr*. **138**: 1796S–1800S.

Biagi, E., Nylund, L., Candela, M., Ostan, R., Bucci, L., Pini, E. et al. (2010). Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS ONE.* **5**: e10667 [doi: 10.1371/journal.pone.0010667].

Biol-Aquino, MA., Perdiz, CJ., Borlagdan, M., Alcantara, JD., and Mallillin, A. (2019). Differences in the bacterial profiles of infant gut by birth process, milk diet, and choice of 16S rRNA gene target region. *Human Microbiome Journal*. **13**: 100062 [https://doi.org/10.1016/j.humic.2019.100062].

Black, RE., Allen, LH., Bhutta, ZA., Caulfield, LE., de Onis, M., Ezzati, M. et al. (2008). Maternal and child undernutrition: Global and regional exposures and health consequences. *Lancet*. **371**: 243–260.

Blanton, LV., Barratt, MJ., Charbonneau, MR., Ahmed, T., and Gordon, JI. (2016). Childhood undernutrition, the gut microbiota, and microbiota-directed therapeutics. *Science*. **352**: 1533-7 [https://doi.org/10.1126/science.aad9359].

Blanton, LV., Charbonneau, MR., Salih, T., Barratt, MJ., Venkatesh, S., Ilkaveya, O. et al. (2016). Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children. *Science.* **351** [https://doi.org/10.1126/science.aad331].

Boesten, RJ. and de Vos, WM. (2008). Interactomics in the human intestine: Lactobacilli and bifidobacteria make a difference. *J Clin Gastroenterol.* **42**: Suppl 3 Pt 2: S163–7.

Bokulich, NA., Kaehler, BD., Rideout, JR., Dillon, M., Bolyen, E., Knight, R. et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome.* **6**: 90 [doi: 10.1186/s40168-018-0470-z].

Bokulich, NA., Subramanian, S., Faith, JJ., Gevers, D., Gordon, J. I., Knight, R. et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*. **10**: 57–59 [https://doi.org/10.1038/nmeth.2276].

Bolyen, E., Rideout, JR., Dillon, MR., Bokulich, NA., Abnet, CC., Al-Ghalith, GA. et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**: 852–857 [https://doi.org/10.1038/s41587-019-0209-9].

Bottacini, F., van Sinderen, D. and Ventura, M. (2017). Omics of bifidobacteria: research and insights into their health-promoting activities. *Biochem*. *J*. **474**: 4137–4152.

Bray, JR. and Curtis, JT. (1957). An ordination of upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27**: 325-349.

Brazier, L., Elguero, E., Koumavor, CK., Renaud, N., Prugnolle, F., Thomas, F. et al. (2017). Evolution in fecal bacterial/viral composition in infants of two central African countries (Gabon and Republic of the Congo) during their first month of life. *PLoS One.* **12**: e0185569.

Brewster, R., Tamburini, FB., Asiimwe, E., Oduaran, O., Hazelhurst, S. and Bhatt, AS. (2019). Surveying Gut Microbiome Research in Africans: Toward Improved Diversity and Representation. *Trends Microbiol*. **27**: 824–835 [doi: 10.1016/j.tim.2019.05.006].

Bustin, SA., Benes, V., Garson, JA., Hellemans, J., Huggett, J., Kubista, M. et al. (2011). Primer sequence disclosure: a clarification of the MIQE guidelines. *Clinical chemistry*. **57**: 919–921 [https://doi.org/10.1373/clinchem.2011.162958].

Cabrera-Rubio, R., Collado, MC., Laitinen, K., Salminen, S., Isolauri, E., and Mira, A. (2012). The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. *The American journal of clinical nutrition*. **96**: 544–551 [https://doi.org/10.3945/ajcn.112.037382].

Callahan, BJ., McMurdie, PJ., and Holmes, SP. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*. **11**: 2639–2643 [https://doi.org/10.1038/ismej.2017.119].

Caporaso, JG., Lauber, CL., Walters, WA., Berg-Lyons, D., Lozupone, CA., Turnbaugh, P. J. et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A*. **108**: 4516–4522 [doi: 10.1073/pnas.1000080107].

Charbonneau, MR., Donnell, DO., Blanton, LV, Totten, SM., Davis, JCC., Barratt, MJ. et al. (2017). Growth in Models of Infant Undernutrition. *Cell*. **164**: 859–871 [https://doi.org/10.1016/j.cell.2016.01.024].

Charrier, C., Duncan, GJ., Reid, MD., Rucklidge, GJ., Henderson, D., Young, P. et al. (2006). A novel class of CoA-transferase involved in short-chain fatty acid metabolism in butyrate-producing human colonic bacteria. *Microbiology*. **152**: 179–185 [doi: 10.1099/mic.0.28412-0].

Chen, W., Zhang, CK., Cheng, Y., Zhang, S. and Zhao, H. (2013). A comparison of methods for clustering 16S rRNAsequences intoOTUs. *PLoS ONE*. **8**: e70837 [doi: 10.1371/journal.pone.0070837].

Chi, L., Bian, X., Gao, B., Ru, H., Tu, P. and Lu, K. (2016). Sex-Specific Effects of Arsenic Exposure on the Trajectory and Function of the Gut Microbiome. *Chem Res Toxicol*. **29**: 949–951 [https://doi.org/10.1021/acs.chemrestox.6b00066].

Chiarucci, A., Bacaro, G. and Scheiner, SM. (2011). Old and new challenges in using species diversity for assessing biodiversity. *Philos TransR Soc B Biol Sci*. **366**: 2426–2437.

Chu, DM., Ma, J., Prince, AL., Antony, KM., Seferovic, MD., and Aagaard, KM. (2017). Maturation of the Infant Microbiome Community Structure and Function Across Multiple Body Sites and in Relation to Mode of Delivery Derrick. *Nature Medicine*. **23**: 314–326 [https://doi.org/10.5588/ijtld.16.0716.Isoniazid].

Claesson, MJ., Cusack, S., O'Sullivan, O., Greene-Diniz, R., De Weerd, H., Flannery, E. et al. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci*.**108**: 4586–4591.

Claesson, MJ., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, JR., Ross, RP., et al. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16s rrna gene regions. *Nucleic Acids Res*. **38**: e200 [doi:10.1093/nar/gkq873].

Cong, X., Xu, W., Janton, S., Henderson, WA., Matson, A., McGrath, JM., et al. (2016). Gut microbiome developmental patterns in early life of preterm infants: Impacts of feeding and gender. *PLoS ONE*. **11**: 1–19 [doi: 10.1371/journal.pone.0152751].

Cotillard, A., Kennedy, SP., Kong, LC., Prifti, E., Pons, N., Le Chatelier, E., et al. (2013). Dietary intervention impact on gut microbial gene richness. *Nature.* **500**: 585–588 [doi: 10.1038/nature12480].

David, LA., Maurice, CF., Carmody, RN., Gootenberg, DB., Button, JE., Wolfe, BE., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature.* **505**: 559–563 [doi: 10.1038/nature12820].

Davis, JCC., Lewis, ZT., Krishnan, S., Bernstein, RM., Moore, SE., Pren, AM. et al. (2017). Growth and Morbidity of Gambian Infants are Influenced by Maternal Milk Oligosaccharides and Infant Gut Microbiota. *Nature Scientific Reports*. **4**: 40466 [DOI: 10.1038/srep40466]. https://doi.org/10.1038/srep40466

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, JB.,Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc.Natl. Acad. Sci.* **107**: 14691–14696 [doi: 10.1073/pnas.1005963107].

De Filippo, C., Di Paola, M., Ramazzotti, M., Albanese, D., Pieraccini, G., Banci, E. et al. (2017). Diet, environments, and gut microbiota. A preliminary investigation in children living in rural and urban Burkina Faso and Italy. *Front. Microbiol*. **8**: 1979 [doi: 10.3389/fmicb.2017.01979].

de Goffau, MC., Lager, S., Sovio, U., Gaccioli, F., Cook, E., Peacock, SJ. et al. (2019). Human placenta has no microbiome but can contain potential pathogens. *Nature*. **572**: 329–334.

de Onis, M., Garza, C., Victora, CG., Onyango, AW., Frogillo, ED. and Martines, J. (2004). The WHO Multicentre Growth Reference Study: Planning, Study Design, and Methodology. *Food Nutr. Bull.* **25**: S15-26. [10.1177/15648265040251S103].

DeSantis, TZ., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, EL., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARApplied, B., and environmental. *Microbiology*. **72**: 5069–5072 [doi: 10.1128/AEM.03006-05].

Desjardins, P. and Conklin, D. (2010). NanoDrop Microvolume Quantitation of Nucleic Acids. *JoVE*. **45**: 2565 [doi: 10.3791/2565].

Dethlefsen, L. and Relman, DA. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci*. **108:** 4554–4561.

Dethlefsen, L., Huse, S., Sogin, ML. and Relman, DA. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol*. **6**: e280.

Devereux, S. and Waidler, J. (2017). Why does malnutrition persist in South Africa despite social grants? *Food Security SA Working Paper***. Series No.001.** DST-NRF Centre of Excellence in Food Security, South Africa.

Dewey, KG., and Adu-Afarwuah, S. (2008). Systematic review of the efficacy and effectiveness of complementary feeding interventions in developing countries. *Maternal & child nutrition*. **4***: 2*4–85 [https://doi.org/10.1111/j.1740-8709.2007.00124.x].

Ding, T. and Schloss, PD. (2014). Dynamics and associations of microbial community types across the human body. *Nature.* **509**: 357–360 [doi: 10.1038/ nature13178].

DNA Genotek (2015). PrepIT•L2P. Laboratory protocol for manual purification of DNA from whole sample. PD-PR-015.

DNA Genotek. (2006). Oragene DNA Self Collection Kit. DNA from Saliva. BR002 Rev 1.1.

DNA Genotek. (2017). OMNIGene Gut. Collect and stabilize DNA for quantitative gut microbiome profile analysis. PD-BR-00159 Issue 8/2017-02.

Doan, T., Arzika, AM., Ray, KJ., Cotter, SY., Kim, J., Maliki, R. et al. (2017). Gut microbial diversity in antibiotic-naive children after systemic antibiotic exposure: a randomized controlled trial. *Clin. Infect. Dis*. **64**: 1147–1153.

Dogra, S., Sakwinska, O., Soh, SE., Ngom-Bru, C., Brück, WM., Berger, B. et al. (2015). Dynamics of infant gut microbiota are influenced by delivery mode and gestational duration and are associated with subsequent adiposity. *Mbio*. **6**. [DOI: 10.1128/mBio.02419-14].

Dominguez-Bello, MG., Costello, EK., Contreras, M., Magris, M., Hidalgo, G., Fierer, N. et al. (2010). Deliverymode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci*. **107**: 11971–11975.

Du Plessis, L., Peer, N., Honikman, S. and English, R. (2016). Breastfeeding in South Africa: Are we making progress? *Health Systems Trust*. **South African Health Review**: 109-123.

Edgar, RC. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioR- xiv*. **081257** [https://doi.org/10.1101/081257].

Eren, AM., Maignien, L., Sul, WJ., Murphy, LG., Grim, SL., Morrison, HG. et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4**: 1111–1119 [doi: 10.1111/2041210X.12114].

Eren, AM., Vineis, JH., Morrison, HG. and Sogin, ML. (2013). Correction: A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLOS ONE.* **8**: e66643 [https://doi.org/10.1371/annotation/afa5c40d-c604-46ae-84c4-82cb92193a5e].

Estaki, M., Jiang, L., Bokulich, NA., McDonald, D., González, A., Kosciolek, T. et al. (2020). QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols in Bioinformatics*. **70**: 1–46 [https://doi.org/10.1002/cpbi.100].

Faith, DP. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv*. **61**: 1–10.

Fallani, M., Young, D., Scott, J., Norin, E., Amarri, S., Adam, R., et al. (2010). Intestinal microbiota of 6-week-old infants across Europe: geographic influence beyond delivery mode, breast-feeding, and antibiotics. *J. Pediatr. Gastroenterol. Nutr*. **51**: 77–84 [doi: 10.1097/MPG.0b013e3181d1b11e].

Ferrand, J., Patron, K., Legrand-Frossi, C., Frippiat, JP., Merlin, C., Alauzet, C. et al. (2014). Comparison of seven methods for extraction of bacterial DNA from fecal and cecal samples of mice. *J. Microbiol. Methods.* **105**: 180–185 [doi: 10.1016/j.mimet.2014.07.029 PMID: 25093756].

Fiedorová, K., Radvanský, M., Němcová, E., Grombiříková, H., Bosák, J., Černochová, M. et al. (2019). The impact of DNA extraction methods on stool bacterial and fungal microbiota community recovery. *Frontiers in Physiology*. **10**: 1–11 [https://doi.org/10.3389/fmicb.2019.00821].

Finglas, P., Roe, M., Pinchen, H., Berry, R., Church, S. Dodhia, S. et al. (2015). McCance and Widdowson's The Composition of Foods. **7**.

Fjeld, E., Siziya, S., Katepa-bwalya, M., Kankasa, C., Moland, K. M., & Tylleskär, T. (2008). "No sister, the breast alone is not enough for my baby": a qualitative assessment of

potentials and barriers in the promotion of exclusive breastfeeding in southern Zambia. *International Breasfeeding Journal*. **3**: 1–12 [https://doi.org/10.1186/1746-4358-3-26].

Fodor, AA., DeSantis, TZ., Wylie, KM., Badger, JH., Ye, Y., Hepburn, T., et al. (2012). The "most wanted" taxa from the human microbiome for whole genome sequencing. *PloS one*. **7**: e41294 [https://doi.org/10.1371/journal.pone.0041294].

Forster, SC., Browne, HP., Kumar, N., Hunt, M., Denise, H., Mitchell, A. et al. (2016). HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res*. **44**: D604–D609 [doi: 10.1093/nar/gkv1216].

Fouhy, F., Guinane, CM., Hussey, S., Wall, R., Ryan, CA., Dempsey, EM., et al. (2012). High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrob Agents Chemother*. **56**: 5811–20 [doi:10.1128/AAC.00789-12].

Franzosa, EA., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, XC. et al. (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol*. **13**: 360–372 [doi: 10.1038/nrmicro3451].

Garrido D., Ruiz-Moyano, S., Kirmiz, N., Davis, JC., Totten, SM., Lemay, DG. (2016). A novel gene cluster allows preferential utilization of fucosylated milk oligosaccharides in Bifidobacterium longum subsp. longum SC596. *Sci Rep.* **6**: 35045 [doi: 10.1038/srep35045].

Gevers, D., Knight, R., Petrosino, JF., Huang, K., McGuire, AL., Birren, BW., et al. (2012). The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. *PLOS Biol*. **10**: e1001377 [https://doi.org/10.1371/journal.pbio.1001377].

Goldenberg, RL., Hauth, JC. and Andrews, WW. (2000). Intrauterine infection and preterm delivery. *The New England journal of medicine*. **342**: 1500–1507 [https://doi.org/10.1056/NEJM200005183422007].

Gomez, A., Luckey, D. and Taneja, V. (2015). The gut microbiome in autoimmunity: Sex matters. *Clin Immunol*. **159**: 154–62 [doi: 10.1016/j.clim.2015.04.016 PMID: 25956531].

Gomez, A., Petrzelkova, KJ., Burns, MB., Yeoman, CJ., Amato, KR., Vlckova, K. et al. (2016). Gut microbiome of coexisting BaAka pygmies and Bantu reflects gradients of traditional subsistence patterns. *Cell Rep.* **14**: 2142–2153 [doi: 10.1016/j.celrep.2016.02.013]

Gotelli, NJ. and Colwell, RK. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol*. **4**: 379–391 [doi:10.1046/j.1461-0248.2001.00230.x].

Gouwer, JC. (1966). Some distance properties of latent roots and vector methods used in multivariate analysis. *Biometrika*. **53**: 325-338.

Graham, DB. and Root, DE. (2015). Resources for the design of CRISPR gene editing experiments. *Genome Biol.* **16:** 260 [https://doi.org/10.1186/s13059-015-0823-x].

Grześkowiak, Ł., Collado, MC., Mangani, C., Maleta, K., Laitinen, K., Ashorn, P. et al. (2012) Distinct gut microbiota in South- eastern African and Northern European infants. *J. Pediatr. Gastroenterol. Nutr*. **54**: 812–816.

Hagemann, IS. (2015). Overview of Technical Aspects and Chemistries of Next-Generation Sequencing. *Clinical Genomics*. 3-19 [10.1016/B978-0-12-404748-8.00001-0].

Hagerty, SL., Hutchison, KE., Lowry, CA., and Bryan, AD. (2020). An empirically derived method for measuring human gut microbiome alpha diversity: Demonstrated utility in predicting healthrelated outcomes among a human clinical sample. *PLoS ONE*. **15**: 1–21 [https://doi.org/10.1371/journal.pone.0229204].

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp.* **41**: 95-98.

Hansen, MEB., Rubel, MA., Bailey, AG., Ranciaro, A., Thompson, SR1., Campbell, MC. et al. (2019). Population structure of human gut bacteria in a diverse cohort from rural Tanzania and Botswana. *Genome Biol*. **20**: 16.

Harmsen, HJ., Wildeboer-Veloo, AC., Raangs, GC., Wagendorp, AA., Klijn, N., Bindels, JG., and Welling, GW. (2000). Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. *Journal of pediatric gastroenterology and nutrition*. **30**: 61–67 [https://doi.org/10.1097/00005176-200001000-00019].

Harris, PA., Taylor, R., Minor, BL., Elliott, V., Fernandez, M., et al. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of biomedical informatics*. **95**: 103208 [https://doi.org/10.1016/j.jbi.2019.103208].

Hart, ML., Meyer, A., Johnson, PJ., and Ericsson, AC. (2015). Comparative evaluation of DNA extraction methods from feces of multiple host species for downstream next-generation sequencing. *PLoS ONE*. **10**: e0143334 [https://doi.org/10.1371/journal.pone.0143334].

Henderson, G., Cox, F., Kittelmann, S., Miri, VH., Zethof, M., Noel, SJ. et al. (2013). Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PLoS One.* **8**: e74787 [doi: 10.1371/journal.pone.0074787 PMID: 24040342].

Hoen, AG., Madan, JC., Li, Z., Coker, M., Lundgren, SN., Morrison, HG. et al. (2018). Sex-specific associations of infants' gut microbiome with arsenic exposure in a US population. *Scientific Reports*. **8**: 12627 [https://doi.org/10.1038/s41598-018-30581-9].

Hoskins, LC., Agustines, M., McKee, WB., Boulding, ET., Kriaris, M. and Niedermeyer, G. (1985). Mucin degradation in human colon ecosystems. Isolation and properties of fecal strains that degrade ABH blood group antigens and oligosaccharides from mucin glycoproteins. *The Journal of clinical investigation*. **75**: 944–953 [https://doi.org/10.1172/JCI111795].

Howards, PP. (2018). An overview of confounding. Part 2: how to identify it and special situations. *Acta Obstetricia et Gynecologica Scandinavica*. **97**: 400–406 [https://doi.org/10.1111/aogs.13293].

Hug, L., Alexander, M., You, D., Alkema, L. and Estimation, UN. et al. (2019). National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *Lancet Glob. Health*. **7**: e710–e720 [doi: 10.1016/S2214-109X(19)30163-9].

HMP Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature.* **486**: 207–14 [doi:10.1038/ nature11234].

Huttenhower, C.,Knight, R., Brown, CT., Caporaso, JG., Clemente, JC.,Gevers, D. et al. (2014). Advancing the microbiome research community. *Cell.* **159** : 227–230 [doi: 10.1016/j.cell.2014.09.022].

Jaccard, P. (1902). Lois de distribution florale dans la zone alpine. *Bull. Soc. Vaud. Sci. Nat*. **38**: 69.

Jaeggi, T., Kortman, GA., Moretti, D., Chassard, C., Holding, P., Dostal, A. et al. (2015). Iron fortification adversely affects the gut microbiome, increases pathogen abundance and induces intestinal inflammation in Kenyan infants. *Gut.* **64**: 731–742.

Jakobsson, HE., Abrahamsson, TR., Jenmalm, MC., Harris, K., Quince, C., Jernberg, C. et al. (2014). Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut*. **63**: 559–66.

Janda, JM., and Abbott, SL. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol*. **45**: 2761–2764 [doi: 10.1128/JCM.01228-07].

Jernberg, C., Lofmark, S., Edlund, C. and Jansson, JK. (2010). Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology*. **156**: 3216-3223.

Jiménez, E., Marin, ML., Martin, R., Odriozola, JM., Olivares, M., Xaus, J. et al. (2008). Is meconium from healthy newborns actually sterile? *Microbiol. Res*. **159**: 187-193.

Johnson, J. (2017). Full-length 16S rRNA gene sequencing. *The Jackson Laboratory for Genomic Medicine*. ***2***: 1–125.

Johnson, J. and Spakowicz, LPD. (2019). Evolution of 16s rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*. **10**: 1–11.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*. **88**: 2427–2439 [https://doi.org/10.1890/06-1736.1 PMID: 18027744].

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T. et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*. **7**: 1–17 [https://doi.org/10.3389/fmicb.2016.00459].

Kane, AV., Dinh, DM. and Ward, HD. (2015). Childhood malnutrition and the intestinal microbiome. *Pediatr. Res*. **77**: 256–262.

Karikari, TK., Quansah, E. and Mohamede, WMY. (2015) Developing expertise in bioinformatics for biomedical research in Africa. *Appl. Translat. Genom*. **6**: 31–34 [https://doi.org/10.1016/j.atg.2015.10.002].

Karunasena, E., McMahon, KW., Chang, D. and Brashears, MM. (2014). Host responses to the pathogen Mycobacterium avium subsp. paratuberculosis and beneficial microbes exhibit host sex specificity. *Appl. Environ. Microbiol*. **80**: 4481–90 [PMID: 24814797].

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. **30**: 3059-66.

Kau, AL., Ahern, PP., Griffin, NW., Goodman, AL. and Gordon, JI. (2011). Human nutrition, the gut microbiome and the immune system. *Nature*. **474**: 327–336 [doi: 10.1038/nature10213].

Kennedy, DB. (2016). Vitamins and the Brain: Mechanisms, Dose and Efficacy—A Review. *Nutrients.* **8**: 68.

Kennedy, NA., Walker, AW., Berry, SH., Duncan, SH., Farquarson, FM., Louis, P. et al. (2014). The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One.* **9**: e88982 [doi: 10.1371/journal.pone.0088982 PMID: 24586470].

Khodayar-Pardo, P., Mira-Pascual, L., Collado, M. C. and Martinez-Costa, C. (2014). Impact of lactation stage, gestational age and mode of delivery on breast milk microbiota. *J. Perinatol*. **34**: 599–605 [doi: 10.1038/jp.2014.47].

Kigbu, A., Orimadegun, AE., Tongo, OO., Odaibo, GN., Olaleye, DO., Akinyinka, OO. et al. (2016). Intestinal bacterial colonization in the first 2 weeks of life of Nigerian neonates using standard culture methods. *Front. Pediatr*. **4**: 139.

Koenig, JE., Spor, A., Scalfone, N., Fricker, AD., Stombaugh, J., Knight, R. et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci*. **108**: 4578–4585 [PubMed: 20668239].

Krautler, B. (2005). Vitamin B12: chemistry and biochemistry. *Biochem Soc Trans*. **33**: 806–810 [PubMed: 16042603].

Kruger, R., & Gericke, GJ. (2003). A qualitative exploration of rural feeding and weaning practices, knowledge and attitudes on nutrition. *Public Health Nutrition*. **6**: 217–223 [https://doi.org/10.1079/PHN2002419].

Kruskal, WH. and Wallis, WA. (1952). Use of Ranks in One-Criterion Variance Analysis. *JASA.* **47**: 583-621.

Kuklina, EV., Ramakrishnan, U., Stein, AD., Barnhart, HH. and Martorell, R. (2006). Early childhood growth and development in rural Guatemala. *Early. Hum*. *Dev*. **82**: 425–433.

Kumar, H., Wacklin, P., Nakphaichit, M., Loyttyniemi, E., Chowdhury, S., Shouche, Y. et al. (2015). Secretor status is strongly associated with microbial alterations observed during pregnancy. *PLoS ONE*. **10**: e0134623 [https://doi.org/10.1371/journal.pone.0134623].

Lanciotti, L., Cofini, M., Leonardi, A., Penta, L., and Esposito, S. (2018). Up-To-Date Review About Minipuberty and Overview on Hypothalamic-Pituitary-Gonadal Axis Activation in Fetal and Neonatal Life. *Frontiers in endocrinology*. **9**: 410 [https://doi.org/10.3389/fendo.2018.00410].

Laursen, MF., Bahl, MI., Michaelsen, KF. and Licht, TR. (2017). First foods and gut microbes. *Frontiers in Microbiology*. **8**: 1–8 [https://doi.org/10.3389/fmicb.2017.00356].

Laursen, MF., Zachariassen, G., Bahl, MI., Bergström, A., and Høst, A. (2015). Having older siblings is associated with gut microbiota development during early childhood. *BMC Microbiology*. **15:**154 [https://doi.org/10.1186/s12866-015-0477-6].

Larsson, MM., Rydell, GE., Grahn, A., Rodriguez-Diaz, J., Akerlind, B., Hutson, AM. et al. (2006.) Antibody prevalence and titer to norovirus (genogroup II) correlate with secretor (FUT2) but not with ABO phenotype or lewis (FUT3) genotype. *J Infect Dis*. **194**: 1422–1427.

Lay, C., Rigottier-Gois, L., Holmstrom, K., Rajilic, M., Vaughan, EE., de Vos, WM. et al. (2005) Colonic microbiota signatures across five northern european countries. *Appl Environ Microbiol.* **71**: 4153–4155.

Ley, RE., Turnbaugh, PJ., Klein, S. and Gordon, JI. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nature*. **444**: 1022–1023.

Li, M., Gong, J., Cottrill, M., Yu, H., de Lange, C., Burton, J. et al. (2003). Evaluation of QIAamp DNA Stool Mini Kit for ecological studies of gut microbiota. *J Microbiol Methods.* **54**: 13–20 [PMID: 12732417].

Life Technologies. (2014). Qubit 3.0 Fluorometer User Guide. MAN0010866. Rev A.0.

Linden, S., Mahdavi, J., Semino-Mora, C., Olsen, C., Carlstedt, I., Borén, T. et al. (2008). Role of ABO secretor status in mucosal innate immunity and H. pylori infection. *PLoS Pathog.* **4**: e2.

Liu, Y., Qin, S., Song, Y., Feng, Y., Lv, N., Xue, Y. et al. (2019). The perturbation of infant gut microbiota caused by cesarean delivery is partially restored by exclusive breastfeeding. *Front. Microbiol*. **10**: 1–11 [doi: 10.3389/fmicb.2019.00598].

Liu, Z., Desantis, TZ., Andersen, GL. and Knight, R. (2008). Accurate taxonomy assignments from 16s rrna sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res*. **36**: e120 [doi:10.1093/nar/gkn491].

Lozupone, CA. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*. **71**: 8228-35.

Lozupone, CA., Hamady, M., Kelley, ST. and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol.* **73**: 1576-85.

Mackie, RI., Sghir, A., and Gaskins, HR. (1999). Developmental microbial ecology of the neonatal gastrointestinal tract. *The American journal of clinical nutrition*. **69**: 1035S–1045S [https://doi.org/10.1093/ajcn/69.5.1035s].

Mandal, S., Van Treuren, W., White, RA., Eggesbø, M., Knight, R. and Peddada, SD. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microbial Ecology in Health and Disease*. **26**: 27663 [doi: 10.3402/mehd.v26. 27663].

Mann, HB. and Whitney, DR. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat*. **18**: 50–60 [doi:10.1214/aoms/1177730491].

Mardis, ER. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet*. **9** : 387–402 [doi: 10.1146/annurev.genom.9.081307. 164359].

Mariat, D., Firmesse, O., Levenez, F., Guimarães, VD., Sokol, H., Doré, J. et al. (2009). The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol.* **6**: 1–6 [doi: 10.1186/1471-2180-9-123].

Marques, TM., Wall, R., Ross, RP., Fitzgerald, GF., Ryan, CA. and Stanton, C. (2010). Programming infant gut microbiota: influence of dietary and environmental factors. *Curr. Opin. Biotechnol*. **21**: 149–156 [doi: 10.1016/j. copbio.2010.03.020].

Martin, CR., Ling, PR., and Blackburn, GL. (2016). Review of Infant Feeding: Key Features of Breast Milk and Infant Formula. *Nutrients*. ***8***: 279 [https://doi.org/10.3390/nu8050279].

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

Mary (2019). Can a mother's diet change the nutritional quality of breast milk? *Nanobébé.* [Accessed: 08:45pm 18 February 2020 https://www.nanobebe.com/can-a-mothers-diet-change-the-nutritional-quality-of-breast-milk/].

McDonald, B. and McCoy, KD. (2019). Maternal microbiota in pregnancy and early life. *Science*. **365**: 984–985 [https://doi.org/10.1126/science.aay0618].

McGovern, DP., Jones, MR., Taylor, KD., Marciante, K., Yan, X., Dubinsky, M. et al. (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with crohn's disease. *Hum Mol Genet*. **19**: 3468–3476 [doi: 10.1093/hmg/ddq248].

McKight, PE. and Najab, J. (2010). Kruskal-Wallis Test. *The Corsini Encyclopedia of Psychology.* **4.** [doi:10.1002/9780470479216.corpsy0491].

Miyoshi, J., Leone, V., Nobutani, K., Musch, MW., Martinez-Guryn, K. et al. (2018). Minimizing confounders and increasing data quality in murine models for studies of the gut microbiome. *PeerJ*. **6**: e5166 [https://doi.org/10.7717/peerj.5166].

Moreira, AP., Texeira, TF., Ferreira, AB., Peluzio Mdo, C. and Alfenas Rde, C. (2012). Influence of a high-fat diet on gut microbiota, intestinal permeability and metabolic endotoxaemia. *Br. J. Nutr*. **108**: 801–809 [doi: 10.1017/S0007114512001213].

Morgan, XC. and Huttenhower, C. (2012). Chapter 12: human microbiome analysis. *PLoS Comput Biol*. **8**: e1002808.

Morrow, AL., Ruiz-Palacios, GM., Altaye, M., Jiang, X., Lourdes Guerrero, M., Meinzen-Derr, J. K. et al. (2004). Human milk oligosaccharides are associated with protection against diarrhea in breast-fed infants. *Journal of Pediatrics*. **145**: 297–303 [https://doi.org/10.1016/j.jpeds.2004.04.054].

Morton, JT., Marotz, C., Washburne, A., Silver- man, J., Zaramela, LS., Edlund, A. et al. (2019). Establishing microbial composition measurement standards with reference frames. *Nature Communications*. **10**: 2719 [doi: 10. 1038/s41467-019-10656-5].

Moulds, JM., Nowicki, S., Moulds, JJ. and Nowicki, BJ. (1996). Human blood groups: Incidental receptors for viruses and bacteria. *Transfusion*. **36**: 362–374.

Müller, O. and Krawinkel, M. (2005). Malnutrition and health in developing countries. *CMAJ* **173**: 279-286 [http://dx.doi.org/10.1503/cmaj.050342].

Nanthakumar, NN., Fusunyan, RD., Sanderson, I., and Walker, WA. (2000). Inflammation in the developing human intestine: a possible pathophysiologic contribution to necrotizing enterocolitis. *Proc. Natl. Acad. Sci. U.S.A*. **97**: 6043–6048.

Navas-Molina, JA., Peralta-Sánchez, JM., González, A., McMurdie, PJ., Vázquez-Baeza, Y., Xu, Z., et al. (2013). Advancing our understanding of the human microbiome using QIIME. *Meth. Enzymol*. **531**: 371–444 [doi: 10.1016/B978-0-12-407863-5.00019-8].

Nearing, JT., Douglas, G.M., Comeau, AM. and Langille, M. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*. ***6***: e5364 [doi:10.7717/peerj.5364].

Ng'andu, NH. and Watts, TEE. (1990). Child growth and duration of breast feeding in urban Zambia. *Journal of Epidemiology and Community Health*. **44:** 281–285.

Niers, LEM., Martin, R., Rijkers, GT., Sengers, F., Timmerman, HM.,van Uden, NO. et al. (2009). The effects of selected probiotic strains on the development of eczema (The PandA study). *Allergy.* **64**: 1349–1358.

Nieuwoudt, S., Manderson, L., and Norris, SA. (2018). Infant feeding practices in Soweto, South Africa: Implications for healthcare providers. *South African Medical Journal*. **108**: 756–762 [https://doi.org/10.7196/SAMJ.2018.v108i9.13358].

O'Hara, AM. and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep*. **7**: 688–693.

O'Sullivan, NP., Lelijveld, N., Rutishauser-Perera, A., Kerac, M., James, P. (2018). Follow-up between 6 and 24 months after discharge from treatment for severe acute malnutrition in children aged 6-59 months: A systematic review. *PLoS One*. **13**: e0202053.

Oldenburg, CE., Sié, A., Coulibaly, B., Ouermi, L., Dah, C., Tapsoba, C. et al. (2018). Effect of commonly used pediatric antibiotics on gut microbial diversity in preschool children in Burkina Faso: a randomized clinical trial. *Open Forum Infect. Dis*. **5**: ofy289.

Ottman, N., Smidt, H., Vos, WM. De, and Belzer, C. (2012). The function of our microbiota : who is out there and what do they do ? *Frontiers in Cellular and Infection Microbiology*. **2**: 104. https://doi.org/10.3389/fcimb.2012.00104

Paganini, D., Uyoga, MA., Kortman, GAM., Cercamondi, CI., Winkler, HC., Boekhorst, J. et al. (2019). Iron-containing micronutrient powders modify the effect of oral antibiotics on the infant gut microbiome and increase post-antibiotic diarrhoea risk: a controlled study in Kenya. *Gut*. **68**: 645–653.

Pannaraj, PS., Li, F., Cerini, C., Bender, JM., Yang, S., Rollie, A., et al. (2017). Association between breast milk bacterial communities and establishment and development of the infant gut microbiome. *JAMA Pediatr*. **171**: 647–654 [doi: 10.1001/jamapediatrics.2017.0378].

Pasricha, SR., Hayes, E., Kalumba, K., Biggs, BA. et al. (2013). Effect of daily iron supplementation on health in children aged 4–23 months: a systematic review and meta-analysis of randomised controlled trials. *Lancet Glob. Health.* **1**: e77–e86.

Patel, DN. and Pettifor, JM. (1992). Malnutrition in South Africa. *S. Afr. J. Food Sci. Nutr*. **4**: 22–23.

Patel, JB. (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular diagnosis.* **6**: 313–321 [https://doi.org/10.1054/modi.2001.29158].

Penders, J., Thijs, C., Vink, C., Stelma, FF., Snijders, B., Kummeling, I. et al. (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* **118**: 511–521 [doi: 10.1542/peds.2005-2824].

Pielou, EC. (1966).The measurement of diversity in different types of biological collections*. J. Theor. Biol.* **13**: 131–144 [doi:10.1016/0022-5193(66)90013-0].

Price, MN., Dehal, PS. and Arkin, AP. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*. **5**: e9490 [https://doi.org/10.1371/journal.pone.0009490].

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, KS., Manichanh, C. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* **464**: 59–65.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. **41**: D590–D596 [doi:10.1093/nar/gks1219].

Rajilić-Stojanović, M., Smidt, H., and de Vos, WM. (2007). Diversity of the human gastrointestinal tract microbiota revisited. *Environmental microbiology*. **9**: 2125–2136. [https://doi.org/10.1111/j.1462-2920.2007.01369.x].

Ravel, J., Gajer, P., Abdo, Z., Schneider, GM., Koenig, SS., McCulle, SL. et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci*. **108:** 4680–4687 [PubMed: 20534435].

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. [https://www.R-project.org/].

Rideout, JR., Chase, JH., Bolyen, E., Ackermann, G., González, A., Knight, R. et al. (2016). Keemei: cloud-based validation of tabular bioinformatics file formats in Google Sheets. *GigaScience*. **5**: 27.

Ritari, J., Salojärvi, J., Lahti, L., and de Vos, WM. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics.* **16**: 1056 [doi: 10.1186/s12864-015-2265-y].

Rodriguez-Diaz, J., Monedero, V. and Yebra, MJ. (2011). Utilization of Natural Fucosylated Oligosaccharides by Three Novel a-L-Fucosidases from a Probiotic Lactobacillus casei Strain. *Appl. Environ. Microbiol.* **77**: 703–705 [doi: 10.1128/AEM.01906-10].

Roe, M., Pinchen, H., Church, S. and Finglas, P. (2015), McCance and Widdowson's The Composition of Foods Seventh Summary Edition. *Nutrition Bulletin*. **40**: 36-39 [doi:10.1111/nbu.12124].

Rothman, KJ., Greenland, S. and Lash, TL. (2008). *Modern Epidemiology*. **3rd Ed**. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Ruiz, L., Delgado, S., Ruas-Madiedo, P., Sánchez B. and Margolles, A. (2017). Bifidobacteria and their molecular communication with the immune system. *Front. Microbiol*. **8**: 2345.

Sahl, JW., Matalka, MN. and Rasko, DA. (2012). Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. *Appl Environ Microbiol.* **78**: 4884–92 [doi:10.1128/AEM.00929-12].

Said-Mohamed, R., Micklesfield, LK., Pettifor, JM., and Norris, SA. (2015). Has the prevalence of stunting in South African children changed in 40 years? A systematic review. *BMC Public Health*. **15**: 1–10 [https://doi.org/10.1186/s12889-015-1844-9].

Salguero, MV., Al-Obaide, MA., Singh, R., Siepmann, T. and Vasylyeva, TL. (2019). Dysbiosis of Gram-negative gut microbiota and the associated serum lipopolysaccharide exacerbates inflammation in type 2 diabetic patients with chronic kidney disease. *Experimental and Therapeutic Medicine*. **18**, 3461-3469 [https://doi.org/10.3892/etm.2019.7943].

Schloss, PD., and Westcott, SL. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol*. **10**: 3219–3226 [doi: 10.1128/AEM.028 10-10]

Schnorr, SL., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G. et al. (2014). Gut microbiome of the Hadza hunter- gatherers. *Nat. Commun*. **5**: 3654.

Scholtens, PA., Oozeer, R., Martin, R., Amor, KB. and Knol, J. (2012). The early settlers: intestinal micro- biology in early life. *Annu. Rev. Food Sci. Technol*. **3**: 425–447.

Schwartz, S., Friedberg, I., Ivanov, I. V., Davidson, LA., Goldsby, JS., Dahl, DB., Herman, D., Wang, M., Donovan, SM., and Chapkin, R. (2012). A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biol*. **13**: R32.

Segata, N., Izard, J., Walron, L., Gevers, D., Miropolsky, L., Garrett, W. et al. (2011). Metagenomic Biomarker Discovery and Explanation. *Genome Biol.* **12**: R60.

Sela, DA. (2011). Bifidobacterial utilization of human milk oligosaccharides. *Int. J. Food Microbiol*. **149**: 58–64.

Sela, DA. and Mills, DA. (2010). Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. *Trends Microbiol*. **18**: 298–307 [doi: 10.1016/j.tim.2010.03.008].

Sela, DA., Chapman, J., Adeuya, A., Kim, JH., Chen, F., Whitehead, TR. et al. (2008). The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci.* **105**: 18964-9 [doi: 10.1073/pnas.0809584105].

Shannon, CE. (1948). A Mathematical Theory of Communication. *Bell Syst.* **27:** 379-423 [https://doi.org/10.1002/j.1538-7305.1948.tb01338.x].

Shastri, P., McCarville, J., Kalmokoff, M., Brooks, SP. and Green-Johnson, JM. (2015). Sex differences in gut fermentation and immune parameters in rats fed an oligofructose-supplemented diet. *Biol. Sex Differ*. **6**: 13. [doi: 10.1186/s13293-015-0031-0 PMID: 26251695].

Sinha, R., Abnet, CC., White, O., Knight, R., and Huttenhower, C. (2015). The microbiome quality control project : baseline study design and future directions. *Genome Biology*. **16**: 276. https://doi.org/10.1186/s13059-015-0841-8

Sjogren, YM., Jenmalm, MC., Bottcher, MF., Bjorksten, B., and Sverremark-Ekstrom, E. (2009). Altered early infant gut microbiota in children developing allergy up to 5 years of age. *Clin.Exp.Allergy*. **39**: 518–526.

Skelly, AC., Dettori, JR. and Brodt, ED. (2012). Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*. **3**: 9–12 [https://doi.org/10.1055/s-0031-1298595].

Stackebrandt, E., and Ebers, J. (2006). Molecular taxonomic parameters: tarnished gold standards. *Microbiol. Today.* **33**: 152–155 [doi: 10.1038/msb.2013.22]

Stewart, CJ., Ajami, NJ., O'Brien, JL., Hutchinson, DS., Smith, DP., Wong, MC. et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*. **562**: 583–588 [https://doi.org/10.1038/s41586-018-0617-x].

Stout, MJ., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q. et al. (2013). Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *Am J Obstet Gynecol*. **208**: 226.

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A. et al. (2014). Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*. **510**: 417–421 [https://doi.org/10.1038/nature13421].

Sun, Y., Cai, Y., Huse, SM., Knight, R., Farmerie, WG., Wang, X. et al. (2012). A large-scale benchmark study of existing algorithms for taxonomy- independent microbial community analysis. *Brief. Bioinformatics.* **13**: 107–121 [doi: 10.1093/bib/bbr009].

Tang, Z., Chen, G., Hong, Q., Huang, S., Smith, HM., Shah, RD. et al. (2019). Multi-Omic Analysis of the Microbiome and Metabolome in Healthy Subjects Reveals Microbiome-Dependent Relationships Between Diet and Metabolites. *Front. Genet*. **10**. [doi: 10.3389/fgene.2019.00454].

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature.* **526**: 68-74 [doi:10.1038/nature15393].

Thermo Fisher Scientific. (2015). QuantStudio 3 and 5 Real-Time PCR Systems: Installation, Use and Mainternance. MAN0010407. Rev C.

Thermo Fisher Scientific. (2016). BigDye Terminator v3.1 Cycle Sequencing Kit: User Guide. 4337035. Rev C.

Thomas, DW., Greer, FR., American Academy of Pediatrics Committee on Nutrition, and American Academy of Pediatrics Section on Gastroenterology, Hepatology, and Nutrition (2010). Probiotics and prebiotics in pediatrics. *Pediatrics*. **126:** 1217–1231 [https://doi.org/10.1542/peds.2010-2548].

Tikhonov, M., Leach, RW., and Wingreen, N. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J*. **9**: 68–80 [doi: 10.1038/ismej.2014.117].

Trumbo, PR. (2016). FDA regulations regarding iodine addition to foods and labeling of foods containing added iodine. *The American Journal of Clinical Nutrition*. **104:** 864S–867S [https://doi.org/10.3945/ajcn.115.110338].

Tun, HM., Bridgman, SL., Chari, R., Field, CJ., Guttman, DS., Becker, AB. et al. (2017). Roles of Birth Mode and Infant Gut Microbiota in Intergenerational Transmission of Overweight and Obesity From Mother to Offspring. *JAMA Pediatr.* **172**:368-377 [doi: 10.1001/jamapediatrics.2017.5535.]

Turnbaugh, PJ., Ley, RE., Hamady, M., Fraser-Liggett, CM., Knight, R., Gordon, JI. (2007). The Human Microbiome Project. *Nature*. **449**: 804–810 [https://doi.org/10.1038/nature06244].

Turnbaugh, PJ., Ley, RE., Mahowald, MA., Magrini, V., Mardis, ER. and Gordon, JI. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature.* **444**:1027–31 [doi:10.1038/nature05414].

Underwood, MA., Gaerlan, S., De Leoz, ML., Dimapasoc, L., Kalanetra, KM., Lemay, DG. et al. (2015). Human milk oligosaccharides in premature infants: absorption, excretion, and influence on the intestinal microbiota. *Pediatr Res*. **6:** 670-677 [doi: 10.1038/pr.2015.162].

Underwood, MA., German, JB., Lebrilla, CB., and Mills, DA. (2014). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatr. Res*. **77**: 229–235 [doi: 10.1038/pr.2014.156].

UNDP. (2019). Sustainable Development Goals. [https://www.za.undp.org/content/south_africa/en/home/sustainable-development-goals.html Accessed October 2019].

UNICEF, Division of Data Research and Policy. (2019). UNICEF Global Databases: Overlapping Stunting, Wasting and Overweight.

UNICEF, WHO, The World Bank Group, UN Population Division. (2018). Levels & trends in child mortality: report 2018. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation. [https://childmortality.org/wp-content/uploads/2018/12/UN-IGME- Child-Mortality-Report-2018.pdf].

United Nations Children's Fund (UNICEF). (2013). Improving Child Nutrition. The achievable imperative for global progress. New York: Division of Communication, UNICEF. [www.unicef.org/publications/index_ 68661.html].

Vaishampayan, PA., Kuehl, JV., Froula, JL., Morgan, JL., Ochman, H., and Francino, MP. (2010). Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biol. Evol*. **2**: 53–66.

van de Wijgert, JH., Borgdorff, H., Verhelst, R., Crucitti, T., Francis, S., Verstraelen, H. et al. (2014). The vaginal microbiota: what have we learned after a decade of molecular characterization? *PloS one*. **9**: e105998 [https://doi.org/10.1371/journal.pone.0105998].

Vatanen, T., Plichta, DR., Somani, J., Münch, PC., Arthur, TD., Hall, AB. et al. (2019). Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nature Microbiology*. **4**: 470–479 [https://doi.org/10.1038/s41564-018-0321-5].

Victora, CG., Adair, L., Fall, C., Hallal, PC., Martorell, R., Richter, L. et al. (2008). Maternal and child undernutrition: consequences for adult health and human capital. *Lancet.* **371**: 340–357 [https://doi.org/10.1016/S0140-6736(07)61692-4].

Victora, CG., Bahl, R., Barros, AJD., França, GVA., Horton, S., Krasevec, J. et al. (2016). Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *Lancet.* **387**: 475–490 [doi: 10.1016/S0140-6736(15)01024-7].

Voluntary National Review Report. (2019). South Africa's Implementation Of the 2030 Agenda For Sustainable Development: "Solving Complex Challenges Together". [https://sustainabledevelopment.un.org/content/documents/23402SOUTH_AFRICA_RSA_Voluntary_National_Review_Report_Final__14_June_2019.pdf].

Wacklin, P., Mäkivuokko, H., Alakulppi, N., Nikkilä, J., Tenkanen, H., Räbinä, J. et al. (2011). Secretor genotype (FUT2 gene) is strongly associated with the composition of bifidobacteria in the human intestine. *PLoS ONE*. **6**. [https://doi.org/10.1371/journal.pone.0020113].

Walsh, AM., Crispie, F., O'Sullivan, O., Finnegan, L., Claesson, MJ. and Cotter PD. (2018). Species classifier choice is a key consideration when analysing low-complexity food microbiome data. Microbiome. **6**: 50.

Weisburg, WG., Barns, SM., Pelletier, DA. and Lane, DJ. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol*. **173**: 697–703 [doi:10.1128/jb.173.2.697-703.1991].

Whittaker, RH. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monogr*. **30**: 279–338 [doi:10.2307/19435633].

de Onis, M. and World Health Organization. (2006). WHO Child Growth Standards. Methods and development. *Department of Nutrition for Health and Development.*

World Health Organization. (2014). Global Nutrition targets 2025: Stunting Policy Brief (WHO/NMH/NHD/14.3). Geneva: World Health Organization.

World Health Organization. (2017). Monitoring health for the SDGs, Sustainable Development Goals. *World health statistics.*

World Health Organization. (2015). *Levels and trends in child malnutrition*. UNICEF, W. B. Group: 1–6.

Wu, GD., Chen, J., Hoffmann, C., Bittinger, K., Chen, YY., Keilbaugh, SA. et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science.* **334**: 105–108.

Yatsunenko, T., Rey, FE., Manary, MJ., Trehan, I., Dominguez-Bello, MG., Contreras, M. et al. (2012). Human gut microbiome viewed across age and geography. *Nature*. **486**: 222–227 [https://doi.org/10.1038/nature11053.Human].

Ye, S., Dhillon, S., Ke, X., Collins, AR. and Day, INM. (2001). An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acid Res*. **29**: 17; E88-8.

Yip, SP., Lai, SK. and Wong, ML. (2007). Systematic sequence analysis of the human fucosyltransferase 2 (*FUT2*) gene identifies novel sequence variations and alleles. *Transfusion*. **47**: 1369-1380 [doi:10.1111/j.1537-2995.2007.01280.x].

Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D. et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med*. **21**: 895–905 [doi: 10.1038/nm. 3914].

Zhou, Y., Shan, G., Sodergren, E., Weinstock, G., Walker, WA., and Gregory, KE. (2015). Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. *PloS one*. **10**: e0118632 [https://doi.org/10.1371/journal.pone.0118632].

Zhu, Y., Lin, X., Zhao, F., Shi, X., Li, H., Li, Y., et al. (2015). Meat, dairy and plant proteins alter bacterial composition of rat gut bacteria. *Sci. Rep*. **5**:15220 [doi: 10.1038/srep15220].

Zlotkin, S., Newton, S., Aimone, AM., Azindow, I., Amenga-Etego, S., Tchum, K. et al. (2013). Effect of iron fortification on malaria incidence in infants and young children in Ghana: a randomized trial. *JAMA.* **310**: 938.

Zoetendal, EG., Akkermans, ADL., Akkermans-van Vliet, WM., de Visser, JAGM. and de Vos, WM. (2001) The host genotype affects the bacterial community in the human gastronintestinal tract. *Microbial Ecology in Health and Disease.* **13**: 129–134 [https://doi.org/10.1080/089106001750462669].

Appendix A - Information and consent sheets

Soweto Baby Wash study- Microbiome study

Information sheet and Consent

Hello, my name is Dr Venesa Sahibdeen, and I am a Researcher at the Department of Human Genetics at the University of the Witwatersrand. Together with other colleagues from the University we will be conducting a research study to understand the effect of bacteria (small organisms that can be found everywhere) on health. Although some bacteria are bad for us, many bacteria are good and we depend on them to help us digest our food and to protect us from bad bacteria and viruses (germs). We particularly want to study the bacteria that live in our gut (intestines, colon). Sometimes babies don't' grow very well for their age, this type of growth is called stunted growth and this is maybe because the nutrients from food are not reaching where it needs to go because the bacteria in our gut are not working well. Your baby's first contact with the bacteria is from the birth canal during delivery, your breast milk, and contact with your skin during breast feeding, and of course the surroundings that we grow up in. This is the first way of building our bacteria supply. Before you decide to participate, we would like you to understand why the research is being done, and what it would involve for you and your baby.

What does this study involve?

Firstly, the collection of samples for the study will be conducted at your home. There will be a research assistant who will explain how the samples are collected with your permission. We would like to collect samples from you (at one visit) and your baby at five visits- when baby is three weeks, 6 weeks, 3 months, 6 months and 12 months of age.

If you agree to take part in the study we would like to collect the following samples:

Sample collection involving the Mother

-Your DNA sample from Saliva

From the saliva samples (about a teaspoon-full) that we will collect from you following your permission and consent we will extract DNA (the inherited material in your cells) so that this DNA can be used for studies that look for genes/changes in DNA that can give us clues

about how bacteria in the gut break down milk. Please remember that you cannot eat, drink, smoke or chew gum, 30 minutes before collecting the sample.

-Your breast milk sample

You will be asked to provide a small sample at only one time (1-5ml) of your breast milk during feeding, to take advantage of the milk flow in the privacy of your home, with your permission. The research assistant will provide you with a collection tube. Your sample will provide helpful information about the type of nutrients that you are feeding your baby and help us learn about the overall growth and development of your baby.

Sample collection involving the infant:

- **Stool/poo/faecal samples from your baby**

You will collect this yourself or with the help of the research assistant at home. Using a small stick that comes with the kit you will need to get a small part of your baby's stool (see the picture below) – the easiest way to get this may be from the baby's nappy, before baby passes urine. You will put this sample in a tube we will give you for safekeeping. Our research assistant will explain the procedures to you and collect the tube a when they come to see you at home.

Possible risks:

There are no physical risks to you or your baby regarding sample collection and the procedures are completely non-invasive. There is a possibility that you may not feel comfortable in supplying a breast milk sample or taking a poo sample (the research assistant can assist with sample collection), but we hope to reassure you that there are no risks and there is no pressure to supply a sample.

Cost to you:

Participating in this research will involve no costs to you.

Voluntary participation in research:

You have the right to agree or to refuse to participate in this research. The DNA belongs to you and your baby and we are the keepers of the DNA. You may withdraw your sample/data or your baby's microbiome sample/data at any time.

Records of your participation in this research:

Ethics approval:
This study protocol has been submitted to the University of the Witwatersrand's Human Research Ethics Committee (HREC), and that committee has granted written approval. Ethics Clearance number: M170753. An amendment for the collection of these samples is being submitted.

DNA samples:
Once the data has been extracted from your saliva and baby's stool samples they will be stored in a safe place at the Division of Human Genetics, University of the Witwatersrand under strict control access. All samples will be coded to ensure privacy.

If at some future date we realise there are other analyses that might involve us sending small amounts of the DNA out of the country where tests will be done that we cannot easily do in South Africa, but that would give us valuable information for our studies; such tests would only be performed if permission is given to us by the Human Research Ethics Committee of the University of the Witwatersrand on your behalf.

Data generated:
The scientists who do the research will generate information (data), which will be placed in databases and the outcomes may appear in scientific publications. These scientists will not have access to your name or your baby's name and therefore the information will not be linked back to you.

Benefits:
The discoveries that come from the studies on your DNA or your baby's microbiome DNA will not be of direct benefit to you and will not be communicated back to you. The discoveries may lead to information that will help us in the future to diagnose diseases and how the bacteria in the gut may be having an effect on the growth and well-being of babies.
YOU WILL HAVE A COPY OF THIS INFORMATION SHEET TO KEEP

If you are happy to take part in the study please read and sign the attached consent form and contact us to confirm your participation.

Your signature on the consent form certifies the following:

· You have read the information provided in this consent form

· You have received answers to all of your questions.

· You have freely decided to participate in this research.

· You understand that you are not giving up any of your legal rights.

If you have any complaints or problems with this research you may report them to the Chair of the Human Research Ethics Committee at the University of the Witwatersrand on this telephone number – (011) 717 1234.

Tick sheet of samples collected

	Collected 3wk	Collected 5wk	Collected (tick) 6Wk	Collected 12wk/3 months	Collected 24wk/6 months	Collected 48wk/12 months
Mother						
Saliva samples collected from mother (3wk or 5wk visit)						
Breast milk sample collected from mother (3 wk or 5wk visit)						
Infant						
Infant faecal sample collected						

Soweto Baby Wash study- Sample Consent Sheet

I agree to myself and my child being participants in the study. I understand that the study will involve the collection of samples and that the details and purposes of this study have been explained to me. I understand that I have the right to refuse to participate in the study.

I acknowledge that all procedure/tests on the samples have been or will be approved by the Human Research Ethics Committee of the University of the Witwatersrand.

I agree to giving a breast milk sample

YES ☐ NO ☐

I agree to giving a saliva sample

YES ☐ NO ☐

I agree to the collection of infant faecal samples

YES ☐ NO ☐

MOTHER OF CHILD

_____

Printed name Signature/Mark or Thumbprint
Date and Time

RESEARCH ASSISTANT:

_____

Printed Name Signature/Mark or Thumbprint
Date and Time

Soweto Baby Wash study- DNA Consent Sheet

I agree to myself and my child being participants in the study. I acknowledge that all procedure/tests on the stored DNA samples have been or will be approved by the Human Research Ethics Committee of the University of the Witwatersrand.

YES ☐ NO ☐

I am in agreement that my DNA may be stored and used for the purposes described above.

YES ☐ NO ☐

I am in agreement that the data generated from my DNA may be made available in a public domain without any identifiers.

YES ☐ NO ☐

I agree that a small bit of my DNA may be sent out of the country if the research cannot easily be done in South Africa.

YES ☐ NO ☐

I understand that every time a new study is done on my DNA, permission will be obtained from the ethics committee for the study to make sure that it is used only for the purposes stated above.

YES ☐ NO ☐

I understand that I will not benefit directly from the research done on my DNA.

YES ☐ NO ☐

I understand that I may withdraw from the study at any time.

YES ☐ NO ☐

MOTHER OF CHILD

_____

Printed name Signature/Mark or Thumbprint
Date and Time
RESEARCH ASSISTANT:

Printed Name Signature/Mark or Thumbprint
Date and Time

Appendix B – Infant stool DNA extraction protocol optimization

The steps that were adjusted on the Qiagen Quick Start protocol for stool DNA extraction are as follows:

- Step 1: Pipette set at 300µl (less sample used = better yield). 1000µl pipettes were used, and the tips were cut off to ensure stool could be inserted.
- Step 3: After addition of C1 solution, tubes were vortexed **and** inverted several times before heating.
- Step 5: Thirty second oscillations for five minutes.
- Step 7: Transferred 850µl of supernatant.
- Step 10: Transferred 600µl of supernatant.
- Step 13: Pipette set at 750µl for transfer of supernatant.
- Step 19: 50µl of elution buffer solution C6 used (minimum required).

Appendix C – QIIME2 metadata mapping file

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
#q2:types	cat	cat	cat	cat	cat	cat	cat	num	num	num	num
00-010-12m	Time12	F	mixed	non-secretor	subject1_0	subject1	2	7.4	8.5	64.8	70
00-010-6m	Time6	F	mixed	non-secretor	subject1_6	subject1	2	7.4	8.5	64.8	70
00-010-9m	Time9	F	mixed	non-secretor	subject1_9	subject1	2	7.4	8.5	64.8	70
00-010-X	Time0	F	mixed	non-secretor	subject1_12	subject1	2	7.4	8.5	64.8	70
01-024-12m	Time12	M	mixed	secretor	subject2_0	subject2	2	8.8	10.7	67.4	71.4
01-024-6m	Time6	M	mixed	secretor	subject2_6	subject2	2	8.8	10.7	67.4	71.4
01-024-9m	Time9	M	mixed	secretor	subject2_9	subject2	2	8.8	10.7	67.4	71.4
01-024-X	Time0	M	mixed	secretor	subject2_12	subject2	2	8.8	10.7	67.4	71.4
01-025-12m	Time12	F	exclusive	secretor	subject3_0	subject3	2	6.7	9.3	60.9	71.3
01-025-6m	Time6	F	exclusive	secretor	subject3_6	subject3	2	6.7	9.3	60.9	71.3
01-025-9m	Time9	F	exclusive	secretor	subject3_9	subject3	2	6.7	9.3	60.9	71.3
01-025-X	Time0	F	exclusive	secretor	subject3_12	subject3	2	6.7	9.3	60.9	71.3
01-075-12m	Time12	M	exclusive	secretor	subject4_0	subject4	2	7.4	9	65	71.6
01-075-6m	Time6	M	exclusive	secretor	subject4_6	subject4	2	7.4	9	65	71.6
01-075-9m	Time9	M	exclusive	secretor	subject4_9	subject4	2	7.4	9	65	71.6
01-075-X	Time0	M	exclusive	secretor	subject4_12	subject4	2	7.4	9	65	71.6
01-066-12m	Time12	M	mixed	secretor	subject5_0	subject5	2	6.7	8.6	60.6	65
01-066-6m	Time6	M	mixed	secretor	subject5_6	subject5	2	6.7	8.6	60.6	65
01-066-9m	Time9	M	mixed	secretor	subject5_9	subject5	2	6.7	8.6	60.6	65
01-066-Xa	Time0	M	mixed	secretor	subject5_12	subject5	2	6.7	8.6	60.6	65
01-054-12m	Time12	M	mixed	secretor	subject6_0	subject6	2	6.4	7.6	60.1	76.6
01-054-6m	Time6	M	mixed	secretor	subject6_6	subject6	2	6.4	7.6	60.1	76.6
01-054-9m	Time9	M	mixed	secretor	subject6_9	subject6	2	6.4	7.6	60.1	76.6

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
01-054-X	Time0	M	mixed	secretor	subject6_12	subject6	2	6.4	7.6	60.1	76.6
01-043-12m	Time12	F	exclusive	secretor	subject7_0	subject7	2	5.9	6.6	60.9	69
01-043-6m	Time6	F	exclusive	secretor	subject7_6	subject7	2	5.9	6.6	60.9	69
01-043-9m	Time9	F	exclusive	secretor	subject7_9	subject7	2	5.9	6.6	60.9	69
01-043-X	Time0	F	exclusive	secretor	subject7_12	subject7	2	5.9	6.6	60.9	69
01-093-12m	Time12	M	mixed		subject8_0	subject8	2	6.6	7.9	61	65.5
01-093-6m	Time6	M	mixed		subject8_6	subject8	2	6.6	7.9	61	65.5
01-093-9m	Time9	M	mixed		subject8_9	subject8	2	6.6	7.9	61	65.5
01-093-X	Time0	M	mixed		subject8_12	subject8	2	6.6	7.9	61	65.5
01-0111-12m	Time12	F	mixed	secretor	subject9_0	subject9	2	6.6	8.9	59.6	69.1
01-0111-6m	Time6	F	mixed	secretor	subject9_6	subject9	2	6.6	8.9	59.6	69.1
01-0111-9m	Time9	F	mixed	secretor	subject9_9	subject9	2	6.6	8.9	59.6	69.1
01-0111-X	Time0	F	mixed	secretor	subject9_12	subject9	2	6.6	8.9	59.6	69.1
01-057-12m	Time12	F	exclusive	secretor	subject10_0	subject10	2	7.7	8.4	64.6	69.3
01-057-6m	Time6	F	exclusive	secretor	subject10_6	subject10	2	7.7	8.4	64.6	69.3
01-057-9m	Time9	F	exclusive	secretor	subject10_9	subject10	2	7.7	8.4	64.6	69.3
01-057-X	Time0	F	exclusive	secretor	subject10_12	subject10	2	7.7	8.4	64.6	69.3
00-012-12m	Time12	M	exclusive	secretor	subject11_0	subject11	2	6.4	6.8	62.5	65.6
00-012-6m	Time6	M	exclusive	secretor	subject11_6	subject11	2	6.4	6.8	62.5	65.6
00-012-9m	Time9	M	exclusive	secretor	subject11_9	subject11	2	6.4	6.8	62.5	65.6
00-012-X	Time0	M	exclusive	secretor	subject11_12	subject11	2	6.4	6.8	62.5	65.6
01-045-12m	Time12	M	exclusive	secretor	subject12_0	subject12	2	7.1	8	68.2	70.5
01-045-6m	Time6	M	exclusive	secretor	subject12_6	subject12	2	7.1	8	68.2	70.5
01-045-9m	Time9	M	exclusive	secretor	subject12_9	subject12	2	7.1	8	68.2	70.5

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
01-045-X	Time0	M	exclusive	secretor	subject12_12	subject12	2	7.1	8	68.2	70.5
01-015-12m	Time12	M	exclusive	secretor	subject13_0	subject13	2	6.3	7.6	62.5	68.4
01-015-6m	Time6	M	exclusive	secretor	subject13_6	subject13	2	6.3	7.6	62.5	68.4
01-015-9m	Time9	M	exclusive	secretor	subject13_9	subject13	2	6.3	7.6	62.5	68.4
01-015-X	Time0	M	exclusive	secretor	subject13_12	subject13	2	6.3	7.6	62.5	68.4
01-087-12m	Time12	F	exclusive	non-secretor	subject14_0	subject14	2	6.1	9.6	60	70.4
01-087-6m	Time6	F	exclusive	non-secretor	subject14_6	subject14	2	6.1	9.6	60	70.4
01-087-9m	Time9	F	exclusive	non-secretor	subject14_9	subject14	2	6.1	9.6	60	70.4
01-087-X	Time0	F	exclusive	non-secretor	subject14_12	subject14	2	6.1	9.6	60	70.4
01-086-12m	Time12	F	formula		subject15_0	subject15	2	8.7	9.7	65.4	69
01-086-6m	Time6	F	formula		subject15_6	subject15	2	8.7	9.7	65.4	69
01-086-9m	Time9	F	formula		subject15_9	subject15	2	8.7	9.7	65.4	69
01-086-Xa	Time0	F	formula		subject15_12	subject15	2	8.7	9.7	65.4	69
01-085-12m	Time12	M	exclusive	secretor	subject16_0	subject16	2	5.4	6.8	59.6	67.1
01-085-6m	Time6	M	exclusive	secretor	subject16_6	subject16	2	5.4	6.8	59.6	67.1
01-085-9m	Time9	M	exclusive	secretor	subject16_9	subject16	2	5.4	6.8	59.6	67.1
01-085-X	Time0	M	exclusive	secretor	subject16_12	subject16	2	5.4	6.8	59.6	67.1
01-002-12m	Time12	F	exclusive	secretor	subject17_0	subject17	2	8.8	10.2	62.9	70.5
01-002-6m	Time6	F	exclusive	secretor	subject17_6	subject17	2	8.8	10.2	62.9	70.5
01-002-9m	Time9	F	exclusive	secretor	subject17_9	subject17	2	8.8	10.2	62.9	70.5
01-002-X	Time0	F	exclusive	secretor	subject17_12	subject17	2	8.8	10.2	62.9	70.5
01-006-12m	Time12	F	mixed	secretor	subject18_0	subject18	2	7	8.3	62.5	66.6
01-006-6m	Time6	F	mixed	secretor	subject18_6	subject18	2	7	8.3	62.5	66.6
01-006-9m	Time9	F	mixed	secretor	subject18_9	subject18	2	7	8.3	62.5	66.6
01-006-X	Time0	F	mixed	secretor	subject18_12	subject18	2	7	8.3	62.5	66.6
00-030-12m	Time12	M	mixed	secretor	subject19_0	subject19	2	7.8	9.3	63.9	72.2

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
00-030-6m	Time6	M	mixed	secretor	subject19_6	subject19	2	7.8	9.3	63.9	72.2
00-030-9m	Time9	M	mixed	secretor	subject19_9	subject19	2	7.8	9.3	63.9	72.2
00-030-X	Time0	M	mixed	secretor	subject19_12	subject19	2	7.8	9.3	63.9	72.2
00-004-12m	Time12	M	mixed	secretor	subject20_0	subject20	2	7.3	8.6	62.5	66.1
00-004-6m	Time6	M	mixed	secretor	subject20_6	subject20	2	7.3	8.6	62.5	66.1
00-004-9m	Time9	M	mixed	secretor	subject20_9	subject20	2	7.3	8.6	62.5	66.1
00-004-X	Time0	M	mixed	secretor	subject20_12	subject20	2	7.3	8.6	62.5	66.1
01-011-12m	Time12	F	exclusive	secretor	subject21_0	subject21	9		9.8		68.6
01-011-6m	Time6	F	exclusive	secretor	subject21_6	subject21	9		9.8		68.6
01-011-9m	Time9	F	exclusive	secretor	subject21_9	subject21	9		9.8		68.6
01-011-X	Time0	F	exclusive	secretor	subject21_12	subject21	9		9.8		68.6
01-026-12m	Time12	F	exclusive	secretor	subject22_0	subject22	9		7.8		64.5
01-026-6m	Time6	F	exclusive	secretor	subject22_6	subject22	9		7.8		64.5
01-026-9m	Time9	F	exclusive	secretor	subject22_9	subject22	9		7.8		64.5
01-026-X	Time0	F	exclusive	secretor	subject22_12	subject22	9		7.8		64.5
00-037-12m	Time12	M	exclusive		subject23_0	subject23	0	6.8		59.3	
00-037-6m	Time6	M	exclusive		subject23_6	subject23	0	6.8		59.3	
00-037-9m	Time9	M	exclusive		subject23_9	subject23	0	6.8		59.3	
00-037-X	Time0	M	exclusive		subject23_12	subject23	0	6.8		59.3	
00-014-12m	Time12	M	exclusive	non-secretor	subject24_0	subject24	9		7.1	66.4	
00-014-6m	Time6	M	exclusive	non-secretor	subject24_6	subject24	9		7.1	66.4	
00-014-9m	Time9	M	exclusive	non-secretor	subject24_9	subject24	9		7.1	66.4	
00-014-X	Time0	M	exclusive	non-secretor	subject24_12	subject24	9		7.1	66.4	
00-001-12m	Time12	M	exclusive	secretor	subject25_0	subject25	9		10.4		69.6
00-001-6m	Time6	M	exclusive	secretor	subject25_6	subject25	9		10.4		69.6
00-001-9m	Time9	M	exclusive	secretor	subject25_9	subject25	9		10.4		69.6

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
00-001-X	Time0	M	exclusive	secretor	subject25_12	subject25	9		10.4		69.6
01-044-12m	Time12	F	mixed	non-secretor	subject26_0	subject26	0	7.1		61	
01-044-6m	Time6	F	mixed	non-secretor	subject26_6	subject26	0	7.1		61	
01-044-9m	Time9	F	mixed	non-secretor	subject26_9	subject26	0	7.1		61	
01-044-X	Time0	F	mixed	non-secretor	subject26_12	subject26	0	7.1		61	
01-089-12m	Time12	F	exclusive	non-secretor	subject27_0	subject27	0	7		62.3	
01-089-6m	Time6	F	exclusive	non-secretor	subject27_6	subject27	0	7		62.3	
01-089-9m	Time9	F	exclusive	non-secretor	subject27_9	subject27	0	7		62.3	
01-089-X	Time0	F	exclusive	non-secretor	subject27_12	subject27	0	7		62.3	
01-029-12m	Time12	F	formula	secretor	subject28_0	subject28	9		8.4		65.3
01-029-6m	Time6	F	formula	secretor	subject28_6	subject28	9		8.4		65.3
01-029-9m	Time9	F	formula	secretor	subject28_9	subject28	9		8.4		65.3
01-029-X	Time0	F	formula	secretor	subject28_12	subject28	9		8.4		65.3
01-031-12m	Time12	F	exclusive	secretor	subject29_0	subject29	0	6.4		62.6	
01-031-6m	Time6	F	exclusive	secretor	subject29_6	subject29	0	6.4		62.6	
01-031-9m	Time9	F	exclusive	secretor	subject29_9	subject29	0	6.4		62.6	
01-031-X	Time0	F	exclusive	secretor	subject29_12	subject29	0	6.4		62.6	
01-009-12m	Time12	F	mixed	secretor	subject30_0	subject30	9		9.1		70
01-009-6m	Time6	F	mixed	secretor	subject30_6	subject30	9		9.1		70
01-009-9m	Time9	F	mixed	secretor	subject30_9	subject30	9		9.1		70
01-009-X	Time0	F	mixed	secretor	subject30_12	subject30	9		9.1		70
01-067-12m	Time12	F	exclusive	secretor	subject31_0	subject31	0				
01-067-6m	Time6	F	exclusive	secretor	subject31_6	subject31	0				
01-067-9m	Time9	F	exclusive	secretor	subject31_9	subject31	0				
01-067-Xa	Time0	F	exclusive	secretor	subject31_12	subject31	0				
01-046-12m	Time12	M	formula	secretor	subject32_0	subject32	0	6.7		64.8	

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
01-046-6m	Time6	M	formula	secretor	subject32_6	subject32	0	6.7		64.8	
01-046-9m	Time9	M	formula	secretor	subject32_9	subject32	0	6.7		64.8	
01-046-X	Time0	M	formula	secretor	subject32_12	subject32	0	6.7		64.8	
01-078-12m	Time12	M	mixed	secretor	subject33_0	subject33	0	7.4		63.5	
01-078-6m	Time6	M	mixed	secretor	subject33_6	subject33	0	7.4		63.5	
01-078-9m	Time9	M	mixed	secretor	subject33_9	subject33	0	7.4		63.5	
01-078-X	Time0	M	mixed	secretor	subject33_12	subject33	0	7.4		63.5	
01-022-12m	Time12	F	exclusive	secretor	subject34_0	subject34	9		6.5		67
01-022-6m	Time6	F	exclusive	secretor	subject34_6	subject34	9		6.5		67
01-022-9m	Time9	F	exclusive	secretor	subject34_9	subject34	9		6.5		67
01-022-X	Time0	F	exclusive	secretor	subject34_12	subject34	9		6.5		67
00-020-12m	Time12	M	exclusive	non-secretor	subject35_0	subject35	0	7.5		65.7	
00-020-6m	Time6	M	exclusive	non-secretor	subject35_6	subject35	0	7.5		65.7	
00-020-9m	Time9	M	exclusive	non-secretor	subject35_9	subject35	0	7.5		65.7	
00-020-X	Time0	M	exclusive	non-secretor	subject35_12	subject35	0	7.5		65.7	
01-055-12m	Time12	M	exclusive	secretor	subject36_0	subject36	0	9.6		67.2	
01-055-6m	Time6	M	exclusive	secretor	subject36_6	subject36	0	9.6		67.2	
01-055-9m	Time9	M	exclusive	secretor	subject36_9	subject36	0	9.6		67.2	
01-055-X	Time0	M	exclusive	secretor	subject36_12	subject36	0	9.6		67.2	
01-0115-12m	Time12	F	mixed	non-secretor	subject37_0	subject37	0	4.7		55.5	
01-0115-6m	Time6	F	mixed	non-secretor	subject37_6	subject37	0	4.7		55.5	
01-0115-9m	Time9	F	mixed	non-secretor	subject37_9	subject37	0	4.7		55.5	
01-0115-X	Time0	F	mixed	non-secretor	subject37_12	subject37	0	4.7		55.5	
01-036-12m	Time12	F	exclusive	secretor	subject38_0	subject38	9				

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
01-036-6m	Time6	F	exclusive	secretor	subject38_6	subject38	9				
01-036-9m	Time9	F	exclusive	secretor	subject38_9	subject38	9				
01-036-X	Time0	F	exclusive	secretor	subject38_12	subject38	9				
01-047-12m	Time12	M	formula	secretor	subject39_0	subject39					
01-047-6m	Time6	M	formula	secretor	subject39_6	subject39					
01-047-9m	Time9	M	formula	secretor	subject39_9	subject39					
01-047-X	Time0	M	formula	secretor	subject39_12	subject39					
01-041-12m	Time12	F	mixed	secretor	subject40_0	subject40					
01-041-6m	Time6	F	mixed	secretor	subject40_6	subject40					
01-041-9m	Time9	F	mixed	secretor	subject40_9	subject40					
01-041-X	Time0	F	mixed	secretor	subject40_12	subject40					
00-029-12m	Time12	M	formula	secretor	subject41_0	subject41					
00-029-6m	Time6	M	formula	secretor	subject41_6	subject41					
00-029-9m	Time9	M	formula	secretor	subject41_9	subject41					
00-029-X	Time0	M	formula	secretor	subject41_12	subject41					
01-018-12m	Time12	F	exclusive	secretor	subject42_0	subject42					
01-018-6m	Time6	F	exclusive	secretor	subject42_6	subject42					
01-018-9m	Time9	F	exclusive	secretor	subject42_9	subject42					
01-018-X	Time0	F	exclusive	secretor	subject42_12	subject42					
01-020-12m	Time12	F	formula	non-secretor	subject43_0	subject43					
01-020-6m	Time6	F	formula	non-secretor	subject43_6	subject43					
01-020-9m	Time9	F	formula	non-secretor	subject43_9	subject43					
01-020-X	Time0	F	formula	non-secretor	subject43_12	subject43					
01-012-12m	Time12	F	formula	secretor	subject44_0	subject44					
01-012-6m	Time6	F	formula	secretor	subject44_6	subject44					
01-012-9m	Time9	F	formula	secretor	subject44_9	subject44					

Sample ID	Time Point	Child Sex	Feeding Practice	Secretor Status	Subject Time	Subject	Sample Present	Weight At0	Weight At9m	Length At0	Length At9m
01-012-X	Time0	F	formula	secretor	subject44_12	subject44					
01-040-12m	Time12	F	exclusive	secretor	subject45_0	subject45					
01-040-6m	Time6	F	exclusive	secretor	subject45_6	subject45					
01-040-9m	Time9	F	exclusive	secretor	subject45_9	subject45					
01-040-X	Time0	F	exclusive	secretor	subject45_12	subject45					
01-010-12m	Time12	F	exclusive	non-secretor	subject46_0	subject46					
01-010-6m	Time6	F	exclusive	non-secretor	subject46_6	subject46					
01-010-9m	Time9	F	exclusive	non-secretor	subject46_9	subject46					
01-010-X	Time0	F	exclusive	non-secretor	subject46_12	subject46					
NC											
01-067-X	Time0	F	exclusive	secretor	subject31_12	subject31	0				
01-086-X	Time0	F	formula		subject15_12	subject15	2	8.7	9.7	65.4	69

*X = at recruitment; num =numeric; cat = categorical

Appendix D – QIIME2 quality control visualizations

Demultiplexed sequence data

Table D 1 Demultiplexed sequence counts summary

Minimum	261
Median	115150.50
Mean	139628.25
Maximum	742794
Total	15917621

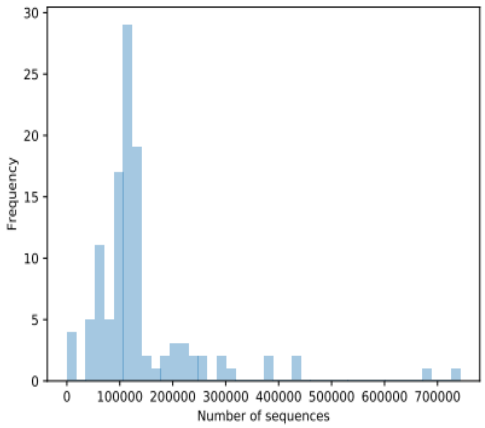


Figure D 1: A histogram showing the distribution of demultiplexed sequence counts.

Cutadapt and read-joining data

Table D 2 Trimmed and joined demultiplexed sequence counts summary

Minimum	12
Median	69721.5
Mean	85776.32
Maximum	454882
Total	9778501

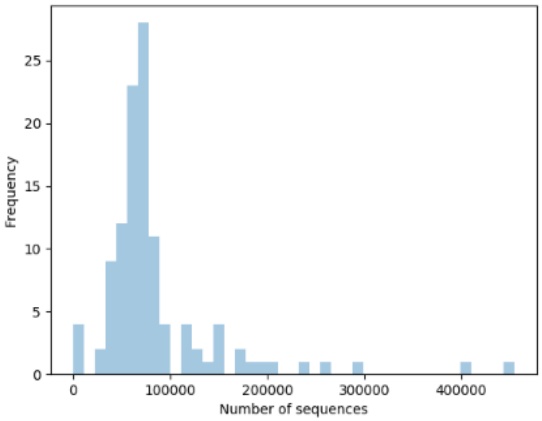


Figure D 2: A histogram showing the distribution of demultiplexed sequences after trimming and read-joining.

Deblur workflow

Deblur denoising statistics

sample-id #q2:types	total-input-reads numeric	total-retained-reads numeric	reads-truncated numeric	reads-too-short-after-truncation numeric	reads-exceeding-maximum-ambiguous-bases numeric
00-001-6m	88607	88607	0	0	0
00-001-9m	76185	76185	0	0	0
00-004-6m	77345	77345	0	0	0
00-004-9m	44530	44477	42	42	11
00-004-X	135267	135266	0	0	1
00-010-6m	76175	76175	0	0	0
00-010-9m	55375	55375	0	0	0
00-010-X	113855	113855	0	0	0
00-010-x	44025	43969	39	39	17
00-012-6m	71747	71747	0	0	0
00-012-X	79541	79541	0	0	0
00-012-x	39397	39352	33	33	12
00-014-6m	77915	77915	0	0	0
00-014-9m	61898	61898	0	0	0
00-020-6m	85401	85401	0	0	0

Figure D 3: A section of the quality filter data of the joined reads from the initial step of the Deblur workflow. This data showed that the majority of reads were retained post truncation and read-joining.

Joined-reads quality filter data

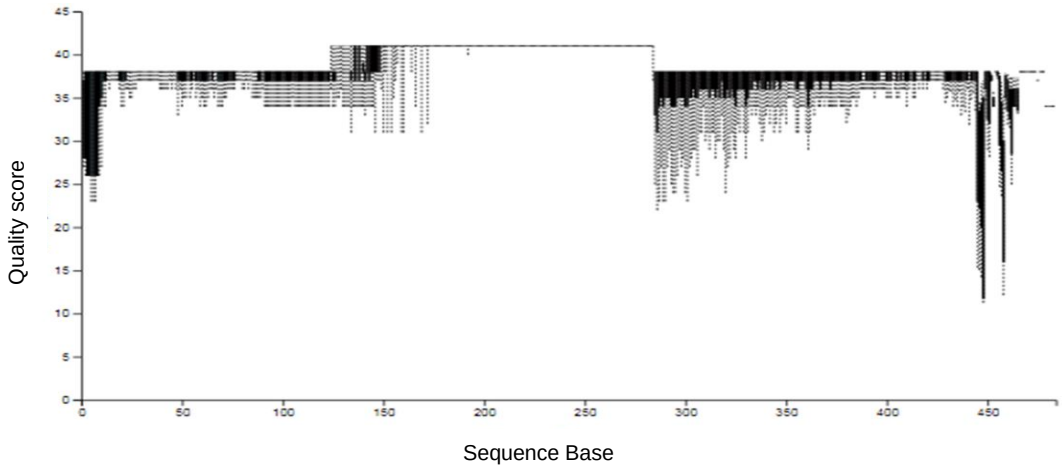
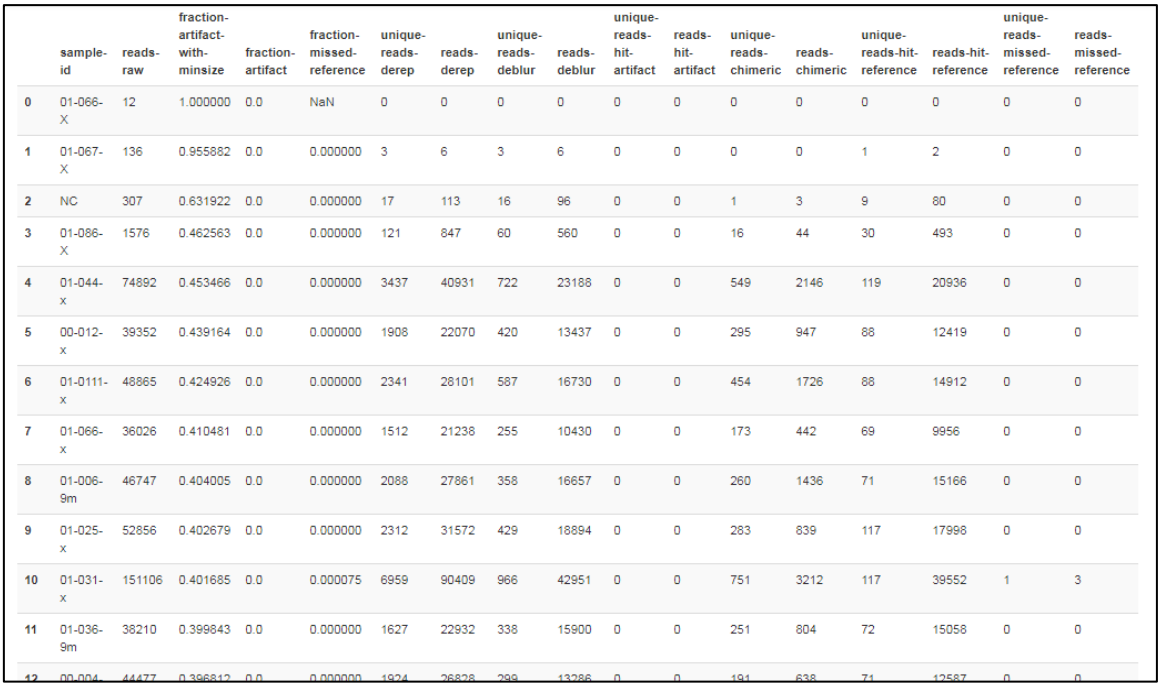


Figure D 4: The interactive quality plot of the trimmed and joined demultiplexed sequences. This shows an estimate of the position for truncating the sequences for denoising, in order to remove erroneous sequences.



	sample-id	reads-raw	fraction-artifact-with-minsize	fraction-artifact	fraction-missed-reference	unique-reads-derep	reads-derep	unique-reads-deblur	reads-deblur	unique-reads-hit-artifact	reads-hit-artifact	unique-reads-chimeric	reads-chimeric	unique-reads-hit-reference	reads-hit-reference	unique-reads-missed-reference	reads-missed-reference
0	01-066-X	12	1.000000	0.0	NaN	0	0	0	0	0	0	0	0	0	0	0	0
1	01-067-X	136	0.955882	0.0	0.000000	3	6	3	6	0	0	0	0	1	2	0	0
2	NC	307	0.631922	0.0	0.000000	17	113	16	96	0	0	1	3	9	80	0	0
3	01-086-X	1576	0.462563	0.0	0.000000	121	847	60	560	0	0	16	44	30	493	0	0
4	01-044-x	74892	0.453466	0.0	0.000000	3437	40931	722	23188	0	0	549	2146	119	20936	0	0
5	00-012-x	39352	0.439164	0.0	0.000000	1908	22070	420	13437	0	0	295	947	88	12419	0	0
6	01-0111-x	48865	0.424926	0.0	0.000000	2341	28101	587	16730	0	0	454	1726	88	14912	0	0
7	01-066-x	36026	0.410481	0.0	0.000000	1512	21238	255	10430	0	0	173	442	69	9956	0	0
8	01-006-9m	46747	0.404005	0.0	0.000000	2088	27861	358	16657	0	0	260	1436	71	15166	0	0
9	01-025-x	52856	0.402679	0.0	0.000000	2312	31572	429	18894	0	0	283	839	117	17998	0	0
10	01-031-x	151106	0.401685	0.0	0.000075	6959	90409	966	42951	0	0	751	3212	117	39552	1	3
11	01-036-9m	38210	0.399843	0.0	0.000000	1627	22932	338	15900	0	0	251	804	72	15058	0	0
12	00-004-	44477	0.396812	0.0	0.000000	1924	26828	299	13286	0	0	191	638	71	12587	0	0

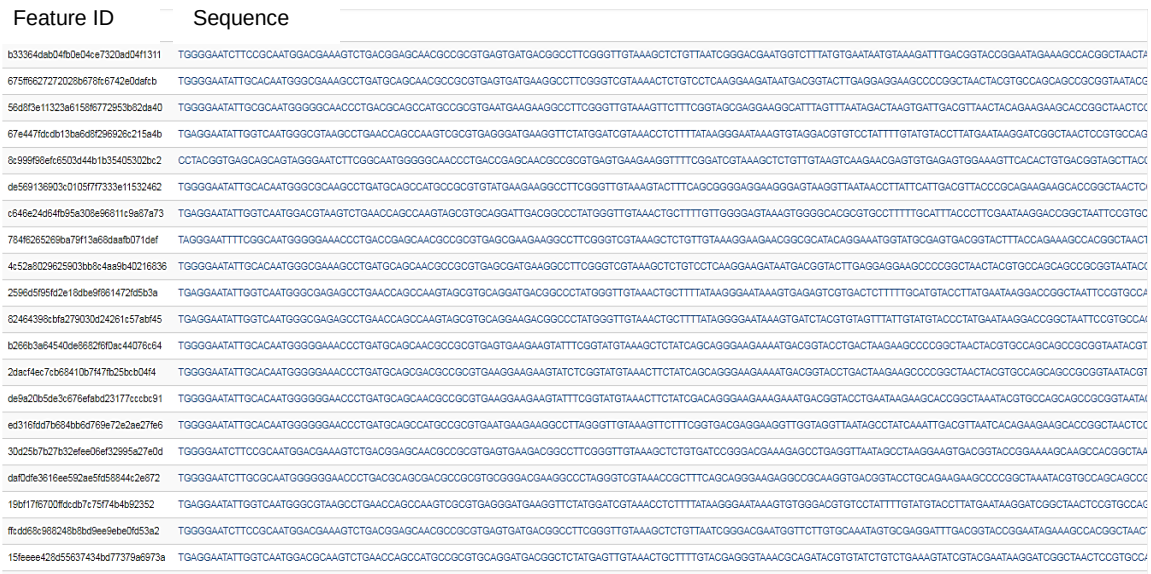
Figure D 5: A section of the denoising statistics output from the Deblur workflow, where the joined reads were truncated at position 401. These statistics indicate that high quality sequence data was produced from the quality filtering process.

Deblur denoising statistics

Table D 3: The detailed quality score summary for n=401 truncating position in relation to the interactive quality plot

Parametric seven-number summary for position 401		
Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	23
Lower Whisker	9th	35
Bottom of Box	25th	37
Middle of Box	50th (median)	38
Top of Box	75th	38
Upper Whisker	91st	38
(Not shown in box plot)	98th	38

Feature table and data summaries



Feature ID	Sequence
b33364dab04fb0e04ce7320ad04f1311	TGGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGACGGCCTTCGGGTTGTAAAGCTCTGTTAATCGGGACGAATGGTCTTTATGTGAATAATGTAAAGATTTGACGGTACCGGAATAGAAAGCCACGGCTAACTA
675ff6627272028b676fc6742e0dafcb	TGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCAACGCCGCGTGAGTGATGAAGGCCTTCGGGTCGTAAAACTCTGTCCTCAAGGAAGATAATGACGGTACTTGAGGAGGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACG
56d8f3e11323a6158f6772953b82da40	TGGGGAATATTGCGCAATGGGGGCAACCCTGACGCAGCCATGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAAGTTCTTTCGGTAGCGAGGAAGGCATTTAGTTTAATAGACTAAGTGATTGACGTTAACTACAGAAGAAGCACCGGCTAACTCC
67e447fdcdb13ba6d8f296926c215a4b	TGAGGAATATTGGTCAATGGGCGTAAGCCTGAACCAGCCAAGTCGCGTGAGGGATGAAGGTTCTATGGATCGTAAACCTCTTTTATAAGGGAATAAAGTGTAGGACGTGTCCTATTTTGTATGTACCTTATGAATAAGGATCGGCTAACTCCGTGCCAG
8c999f96efc6503d44b1b35405302bc2	CCTACGGTGAGCAGCAGTAGGGAATCTTCGGCAATGGGGGCAACCCTGACCGAGCAACGCCGCGTGAGTGAAGAAGGTTTTCGGATCGTAAAGCTCTGTTGTAAGTCAAGAACGAGTGTGAGAGTGGAAAGTTCACACTGTGACGGTAGCTTACC
de569136903c0105f7f7333e11532462	TGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAGGTTAATAACCTTATTCATTGACGTTACCCGCAGAAGAAGCACCGGCTAACTC
c646e24d64fb95a308e96811c9a87a73	TGAGGAATATTGGTCAATGGACGTAAGTCTGAACCAGCCAAGTAGCGTGCAGGATTGACGGCCCTATGGGTTGTAAACTGCTTTTGTTGGGGAGTAAAGTGGGGCACGCGTGCCTTTTTGCATTTACCCTTCGAATAAGGACCGGCTAATTCCGTGC
784f6265269ba79f13a68daafb071def	TAGGGAATTTTCGGCAATGGGGGAAACCCTGACCGAGCAACGCCGCGTGAGCGAAGAAGGCCTTCGGGTCGTAAAGCTCTGTTGTAAAGGAAGAACGGCGCATACAGGAAATGGTATGCGAGTGACGGTACTTTACCAGAAAGCCACGGCTAACT
4c52a8029625903bb8c4aa9b40216836	TGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCAACGCCGCGTGAGCGATGAAGGCCTTCGGGTCGTAAAGCTCTGTCCTCAAGGAAGATAATGACGGTACTTGAGGAGGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACC
2596d5f95fd2e18dbe9f661472fd5b3a	TGAGGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTAGCGTGCAGGATGACGGCCCTATGGGTTGTAAACTGCTTTTATAAGGGAATAAAGTGAGAGTCGTGACTCTTTTTGCATGTACCTTATGAATAAGGACCGGCTAATTCCGTGCCA
82464396cbfa279030d24261c57abf45	TGAGGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTAGCGTGCAGGAAGACGGCCCTATGGGTTGTAAACTGCTTTTATAGGGGAATAAAGTGATCTACGTGTAGTTTATTGTATGTACCCTATGAATAAGGACCGGCTAATTCCGTGCCA
b266b3a64540de8682f6f0ac44076c64	TGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCGCGTGAGTGAAGAAGTATTTCGGTATGTAAAGCTCTATCAGCAGGGAAGAAAATGACGGTACCTGACTAAGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGT
2dacf4ec7cb68410b7f47fb25bcb04f4	TGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGAAGGAAGAAGTATCTCGGTATGTAAACTTCTATCAGCAGGGAAGAAAATGACGGTACCTGACTAAGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGT
de9a20b5de3c676efabd23177cccbc91	TGGGGAATATTGCACAATGGGGGAACCCTGATGCAGCAACGCCGCGTGAAGGAAGAAGTATTTCGGTATGTAAACTTCTATCGACAGGGAAGAAAGAAATGACGGTACCTGAATAAGAAGCACCGGCTAAATACGTGCCAGCAGCCGCGGTAATAC
ed316fdd7b684bb6d769e72e2ae27fe6	TGGGGAATATTGCACAATGGGGGAACCCTGATGCAGCCATGCCGCGTGAATGAAGAAGGCCTTAGGGTTGTAAAGTTCTTTCGGTGACGAGGAAGGTTGGTAGGTTAATAGCCTATCAAATTGACGTTAATCACAGAAGAAGCACCGGCTAACTCC
30d25b7b27b32efee06ef32995a27e0d	TGGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGAAGACGGCCTTCGGGTTGTAAAGCTCTGTGATCCGGGACGAAAGAGCCTGAGGTTAATAGCCTAAGGAAGTGACGGTACCGGAAAAGCAAGCCACGGCTAA
daf0dfe3616ee592ae5fd58844c2e872	TGGGGAATCTTGCGCAATGGGGGAACCCTGACGCAGCGACGCCGCGTGCGGGACGAAGGCCCTAGGGTCGTAAACCGCTTTCAGCAGGGAAGAGGCCGCAAGGTGACGGTACCTGCAGAAGAAGCCCCGGCTAAATACGTGCCAGCAGCCG
19bf17f6700ffdcdb7c75f74b4b92352	TGAGGAATATTGGTCAATGGGCGTAAGCCTGAACCAGCCAAGTCGCGTGAGGGATGAAGGTTCTATGGATCGTAAACCTCTTTTATAAGGGAATAAAGTGTGGGACGTGTCCTATTTTGTATGTACCTTATGAATAAGGATCGGCTAACTCCGTGCCAG
ffcdd68c9682248b8bd9ee9ebe0fd53a2	TGGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGACGGCCTTCGGGTTGTAAAGCTCTGTTAATCGGGACGAATGGTTCTTGTGCAAATAGTGCGAGGATTTGACGGTACCGGAATAGAAAGCCACGGCTAACT
15feeee428d55637434bd77379e6973a	TGAGGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGCAGGATGACGGCTCTATGAGTTGTAAACTGCTTTTGTACGAGGGTAAACGCAGATACGTGTATCTGTCTGAAAGTATCGTACGAATAAGGATCGGCTAACTCCGTGCCA

Figure D 6: A section of the representative sequences from the study sample, where each feature ID has a sequence with a link to BLAST against the NCBI nucleotide database.

Table D 4: A summary of the feature table from the Deblur workflow

Metric	Sample
Number of samples	113
Number of features	1384
Total frequency	3258.47

Table D 5: The frequency per sample of the feature data from the Deblur workflow

	Frequency
Minimum frequency	2.0
1st quartile	18516.0
Median frequency	23658.0
3rd quartile	29852.0
Maximum frequency	131740.0
Mean frequency	28835.99

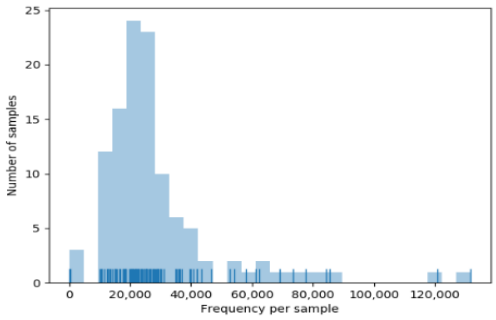


Figure D 7: A histogram showing the frequency per sample (x-axis) versus the number of samples (y-axis) for the feature data.

Table D 6: The frequency per feature of the feature data from the Deblur workflow

	Frequency
Minimum frequency	10
1st quartile	21
Median frequency	81.50
3rd quartile	598.25
Maximum frequency	344539
Mean frequency	2354.38

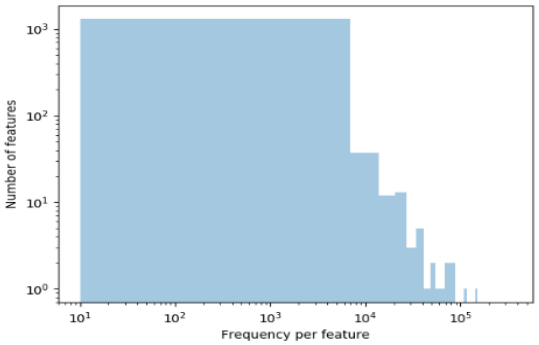


Figure D 8: A histogram showing the frequency per feature (x-axis) versus the number of features (y-axis) for the feature data.

Appendix E – Sequences lost after rarefaction

00-012-x	12,419
01-022-9m	11,654
01-089-x	10,730
00-030-9m	10,707
01-057-X	10,249
01-066-x	9,956
01-086-X	493
NC	80
01-067-X	2

Figure E 1: A section of the frequency per sample data generated from a csv link from the Deblur denoising output (Figure D 5). The red box shows the three sequences lost when the p-sampling depth of 9956 is chosen for rarefaction. NC is unlabelled sample from NICD with no metadata reference.

Appendix F – Taxonomy metadata

Table F 1: A section of the taxonomic classification table of the sequences found within the study sample and the confidence values for their assignment

Feature ID	Taxon	Confidence
#q2:types	categorical	categorical
b33364dab04fb0e04ce7320ad04f1311	*D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Veillonella;Ambiguous_taxa;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.91
675ff6627272028b678fc6742e0dafcb	*D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Peptostreptococcaceae;D_5__Clostridioides;D_6__Clostridioides difficile;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.78
56d8f3e11323a6158f6772953b82da40	*D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pasteurellales;D_4__Pasteurellaceae;D_5__Haemophilus*	0.99
67e447fdcdb13ba6d8f296926c215a4b	*D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Tannerellaceae;D_5__Parabacteroides;Ambiguous_taxa;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.85
8c999f98efc6503d44b1b35405302bc2	*D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Streptococcaceae;D_5__Streptococcus;D_6__Streptococcus salivarius subsp. thermophilus;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.90
de569136903c0105f7f7333e11532462	*D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Enterobacteriales;D_4__Enterobacteriaceae*	0.99
c646e24d64fb95a308e96811c9a87a73	*D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;D_6__uncultured Prevotella sp.;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.93
784f6265269ba79f13a68daafb071def	*D_0__Bacteria;D_1__Firmicutes;D_2__Erysipelotrichia;D_3__Erysipelotrichales;D_4__Erysipelotrichaceae;D_5__Catenibacterium;D_6__uncultured bacterium;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.99
4c52a8029625903bb8c4aa9b40216836	*D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Peptostreptococcaceae;D_5__Paeniclostridium;D_6__uncultured bacterium;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.88
2596d5f95fd2e18dbe9f861472fd5b3a	*D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 9*	0.99
82464398cbfa279030d24261c57abf45	*D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;Ambiguous_taxa;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.98
b266b3a64540de8682f6f0ac44076c64	*D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae*	0.99
2dacf4ec7cb68410b7f47fb25bcb04f4	*D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Blautia*	0.99
de9a20b5de3c676efabd23177cccbc91	*D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Tyzzerella 3;D_6__uncultured bacterium;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.97
ed316fdd7b684bb6d769e72e2ae27fe6	*D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pasteurellales;D_4__Pasteurellaceae;D_5__Actinobacillus;Ambiguous_taxa;D_7__;D_8__;D_9__;D_10__;D_11__;D_12__;D_13__;D_14__*	0.99

Appendix G – LEfSe overview

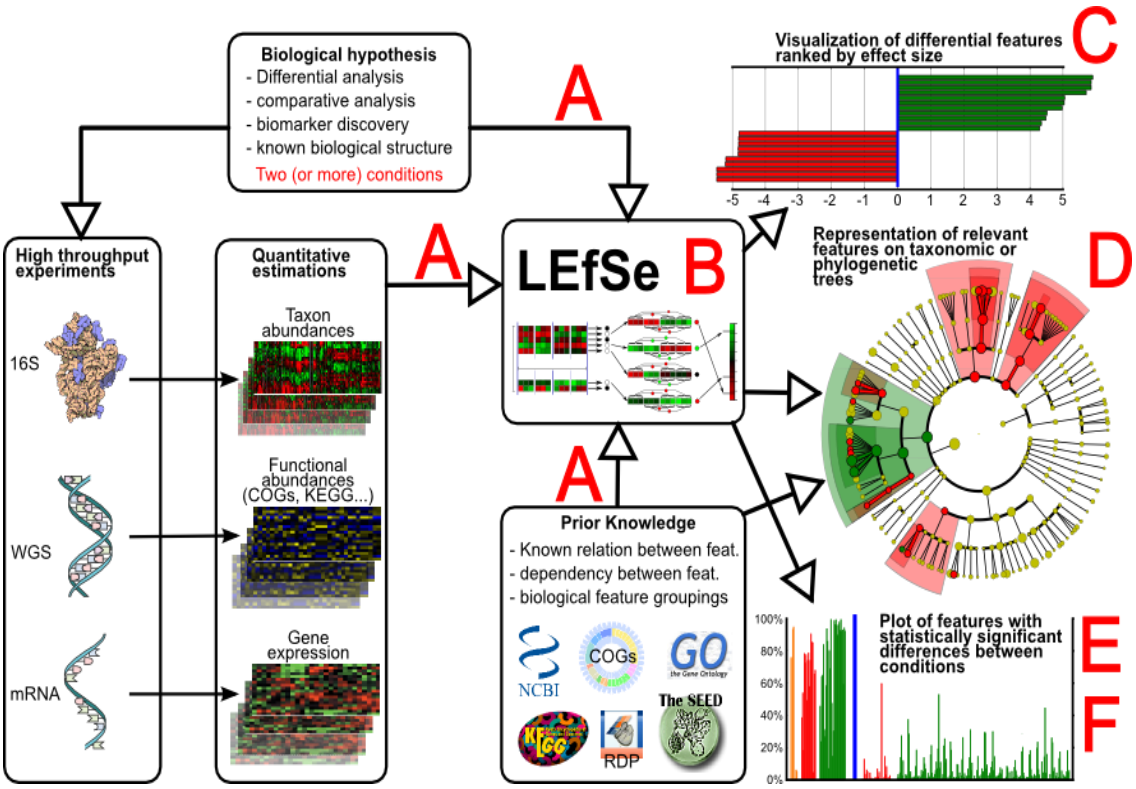


Figure G 1: An overview of the Linear Discriminant Analysis of Effect Size pipeline, showing the assumptions, inputs and output data derived (https://bytebucket.org/biobakery/galaxy_lefse/wiki/lefse_ove.png).

Appendix H – 16S rRNA sequencing data

Table H 1: Output of the 16S rRNA total read counts with the corresponding sample IDs from the NICD showing good sequencing data

	Sample ID	Total reads	Sample ID	Total Reads	Sample ID	Total Reads	Sample ID	Total reads	Sample ID	Total reads
1	01-066_X	139857	00-010_X	419895	01-006_X	883806	01-018_6m	633157	01-025_9m	213677
2	01-067_X	119235	01-024_X	508024	00-030_X	221740	01-020_6m	308148	01-075_9m	202808
3	01-086_X	107152	01-025_X	268020	00-004_X	443226	00-012_6m	237231	01-066_9m	225374
4	01-022_9m	97195	01-075_X	1366491	00-020_X	214792	01-031_6m	245263	00-014_9m	221039
5	01-006_9m	139361	00-037_X	276562	01-055_X	175556	01-045_6m	224035	00-001_9m	248514
6	00-030_9m	80895	01-066_X	510	01-0115_X	243988	01-009_6m	217629	01-054_9m	249884
7	00-004_9m	115373	01-054_X	217597	00-010_6m	263207	01-012_6m	190091	01-043_9m	269677
8	01-036_9m	103892	01-043_X	480900	01-011_6m	381042	01-015_6m	482748	01-093_9m	410715
9	00-010_x	112396	01-044_X	233565	01-026_6m	221262	01-067_6m	770020	01-029_9m	197039
10	01-024_x	777135	01-093_X	190824	01-075_6m	204911	01-046_6m	257725	01-0111_9m	194808
11	01-025_x	139751	01-089_X	320355	00-037_6m	157229	01-087_6m	180019	01-057_9m	402078
12	01-075_x	499369	01-0111_X	270920	01-066_6m	571068	01-078_6m	255016	01-002_9m	180656
13	00-037_x	129582	01-057_X	112544	00-014_6m	248769	01-085_6m	238474	01-045_9m	271367
14	01-066_x	97230	00-012_X	269040	01-041_6m	240673	01-002_6m	232322	01-009_9m	186106
15	01-054_x	229718	01-031_X	206888	00-001_6m	271462	01-006_6m	259669	01-015_9m	258490
16	01-043_x	385485	01-045_X	259127	01-054_6m	201214	00-030_6m	230856	01-087_9m	205198
17	01-044_x	228513	01-015_X	217979	01-043_6m	241279	00-004_6m	242808	01-086_9m	249597
18	01-093_x	263522	01-067_X	1721	01-044_6m	1485600	00-020_6m	282650	01-085_9m	590766
19	01-089_x	106054	01-046_X	173230	01-093_6m	454923	01-0115_6m	865256	01-026_9m	213135
20	01-0111_x	135889	01-087_X	206951	01-089_6m	216322	01-010_6m	243669	01-057_6m	230920
21	01-057_x	195701	01-078_X	312610	00-029_6m	237581	00-010_9m	177915	01-002_X	169150
22	00-012_x	107081	01-086_X	5769	01-029_6m	213517	01-011_9m	159333	NC	1169
23	01-031_x	459825	01-085_X	234490	01-0111_6m	209692	01-024_9m	181899		

Small "X"= at recruitment (12-21 weeks, time point 0); capital "x" = time point 12; 6m (time point 6); 9m (time point 9). NC= unlabelled sample.

Yellow highlights correspond to samples with repeated sequencing for quality purposes.

Appendix I – Alpha diversity analysis

1. Alpha diversity using Faith's Phylogenetic Diversity metric

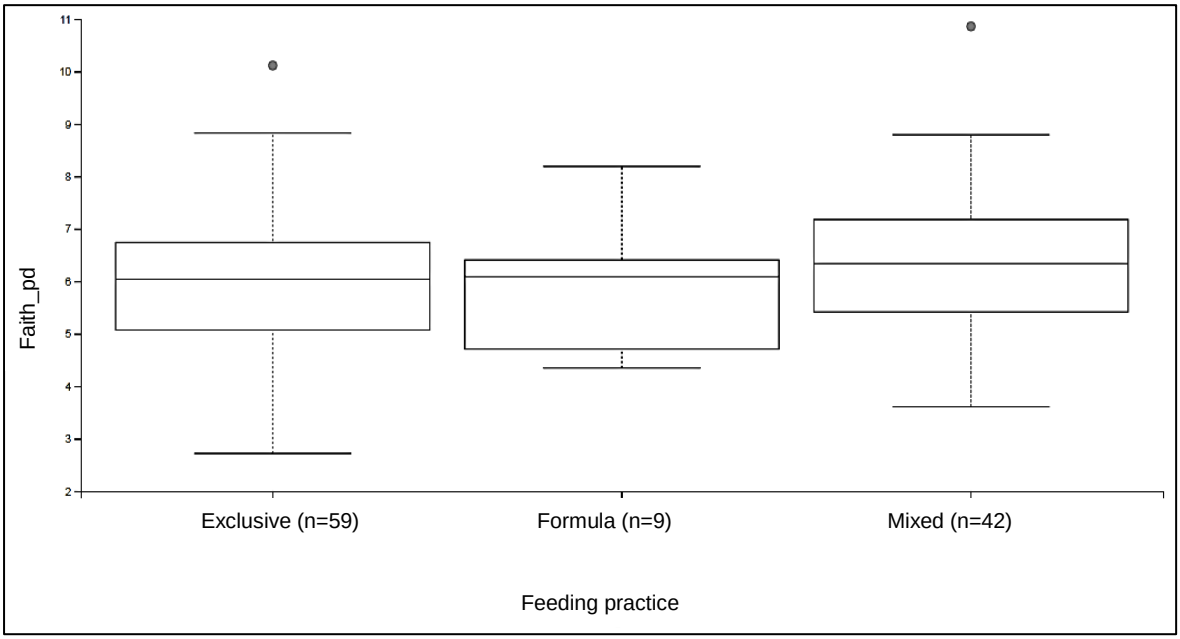


Figure I 1.1: Box plot visualizations of the taxonomic phylogenetic diversity indices (Faith's Phylogenetic Diversity) for different feeding practices (left=exclusive; middle= formula; right= mixed). This indicates diversity indices are relatively similar and correspond to the insignificant statistical KW values obtained in Table I 1.1.

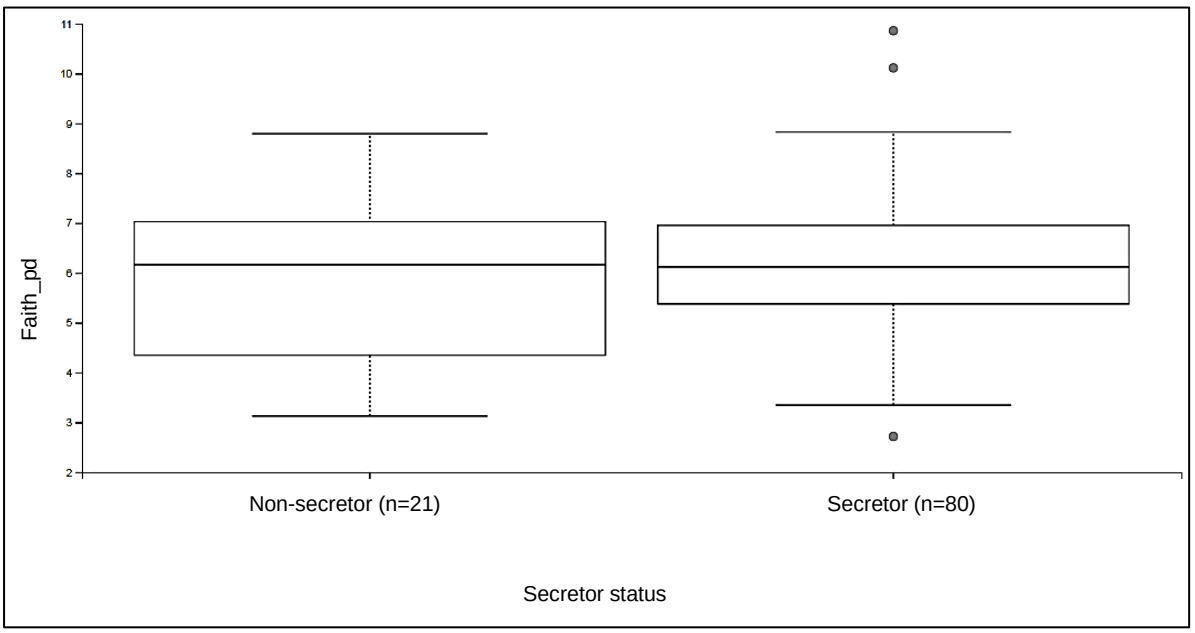


Figure I 1.2: Box-plot visualizations of taxonomic phylogenetic indices (Faith's Phylogenetic Diversity) for secretor status (left=non-secretor; right= secretor). This indicates diversity indices are relatively similar and correspond to the insignificant statistical KW values obtained in Table I 1.2

Table I 1.1: Kruskal-Wallis statistical test results of significance for comparisons between all feeding practice groups and pairwise tests for phylogenetic diversity using Faith's Phylogenetic Diversity metric

Kruskal-Wallis (all groups)				
	Result			
H	1.52			
P-value	0.47			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
Exclusive (n=59)	Formula (n=9)	0.08	0.78	0.78
	Mixed (n=42)	1.08	0.30	0.47
Formula (n=9)	Mixed (n=42)	1.03	0.31	0.47

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures.

Table I 1.2: Kruskal-Wallis statistical test results of significance for comparisons between all secretor status groups and pairwise tests for phylogenetic diversity using Faith's Phylogenetic Diversity

Kruskal-Wallis (all groups)				
	Result			
H	0.23			
P-value	0.63			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
Non-secretor (n=21)	Secretor (n=80)	0.23	0.63	0.63

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures.

2. Alpha diversity using Pielou's Evenness metrics

Table I 2.1: Kruskal-Wallis statistical test results of significance for comparisons between all feeding practice groups and pairwise tests for phylogenetic diversity using Pielou's Evenness metrics

Kruskal-Wallis (all groups)				
	Result			
H	1.82			
P-value	0.40			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
Exclusive (n=59)	Formula (n=9)	0.01	0.94	0.94
	Mixed (n=42)	1.70	0.19	0.58
Formula (n=9)	Mixed (n=42)	0.55	0.46	0.69

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures.

Table I 2.2: Kruskal-Wallis statistical test results of significance for comparisons between all secretor status groups and pairwise tests for phylogenetic diversity using Pielou's Evenness

Kruskal-Wallis (all groups)				
	Result			
H	0.07			
P-value	0.79			
Kruskal-Wallis (pairwise)				
		H	*P*-value	q-value
Group 1	Group 2			
Non-secretor (n=21)	Secretor (n=80)	0.07	0.79	0.79

H= H-statistic and is the test statistic used to determine whether the medians of two or more groups are different. This value is generated using the core-phylogenetic methods script from the QIIME2 pipeline for alpha diversity measures.

Appendix J – Alpha rarefaction plots

1. Shannon Diversity alpha rarefaction curves

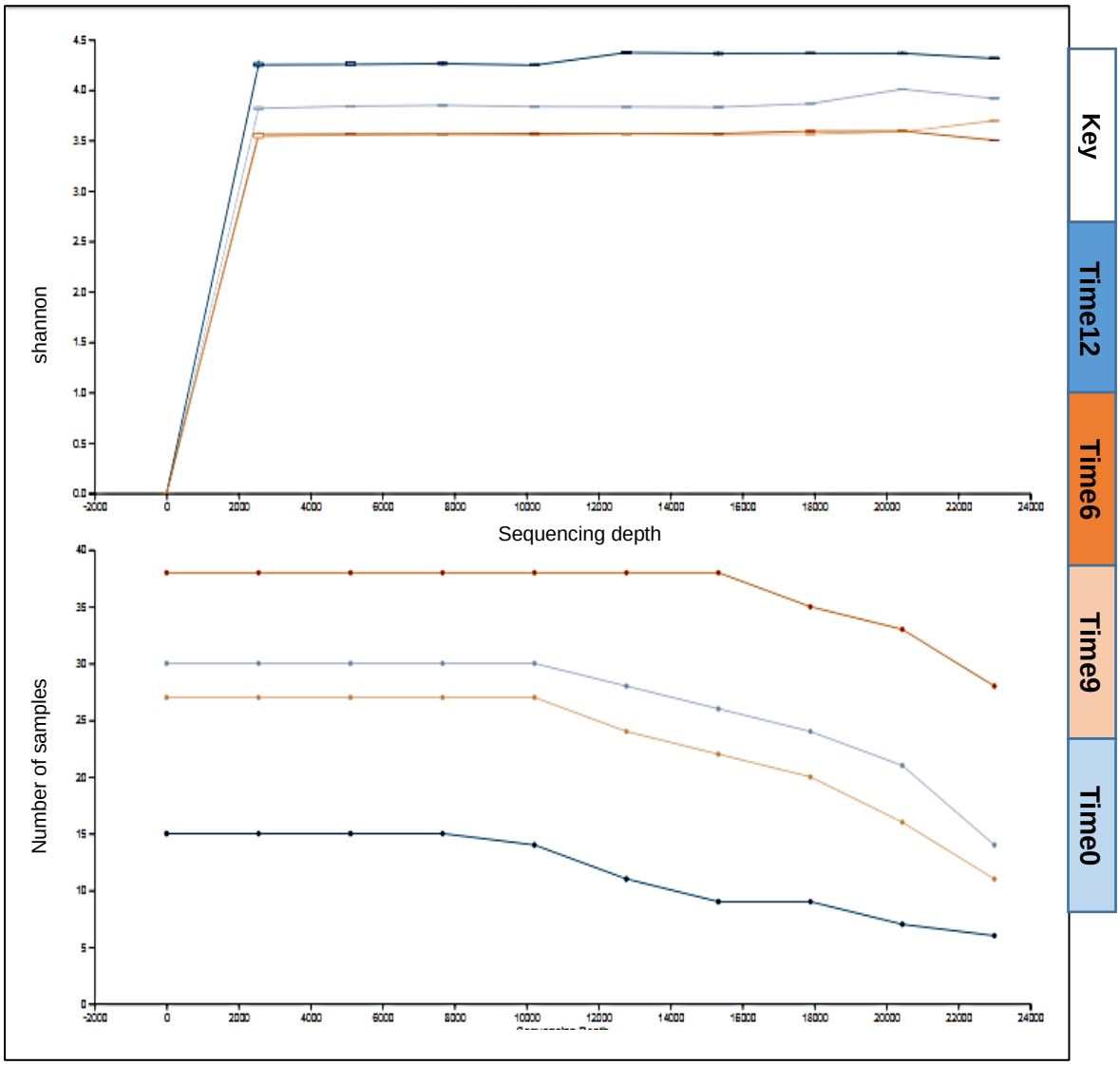


Figure J 1.1: Alpha rarefaction curves of the different time points computed with Shannon Diversity metrics. The exponential growth and plateau of the curves show species richness is fully observed at a maximum index of 4.5 (Time12 – dark blue) and sequencing depth of 2 000. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000 for all time points, except Time6 at 16 000. The key on the right relates the colours of the curves with the time points.

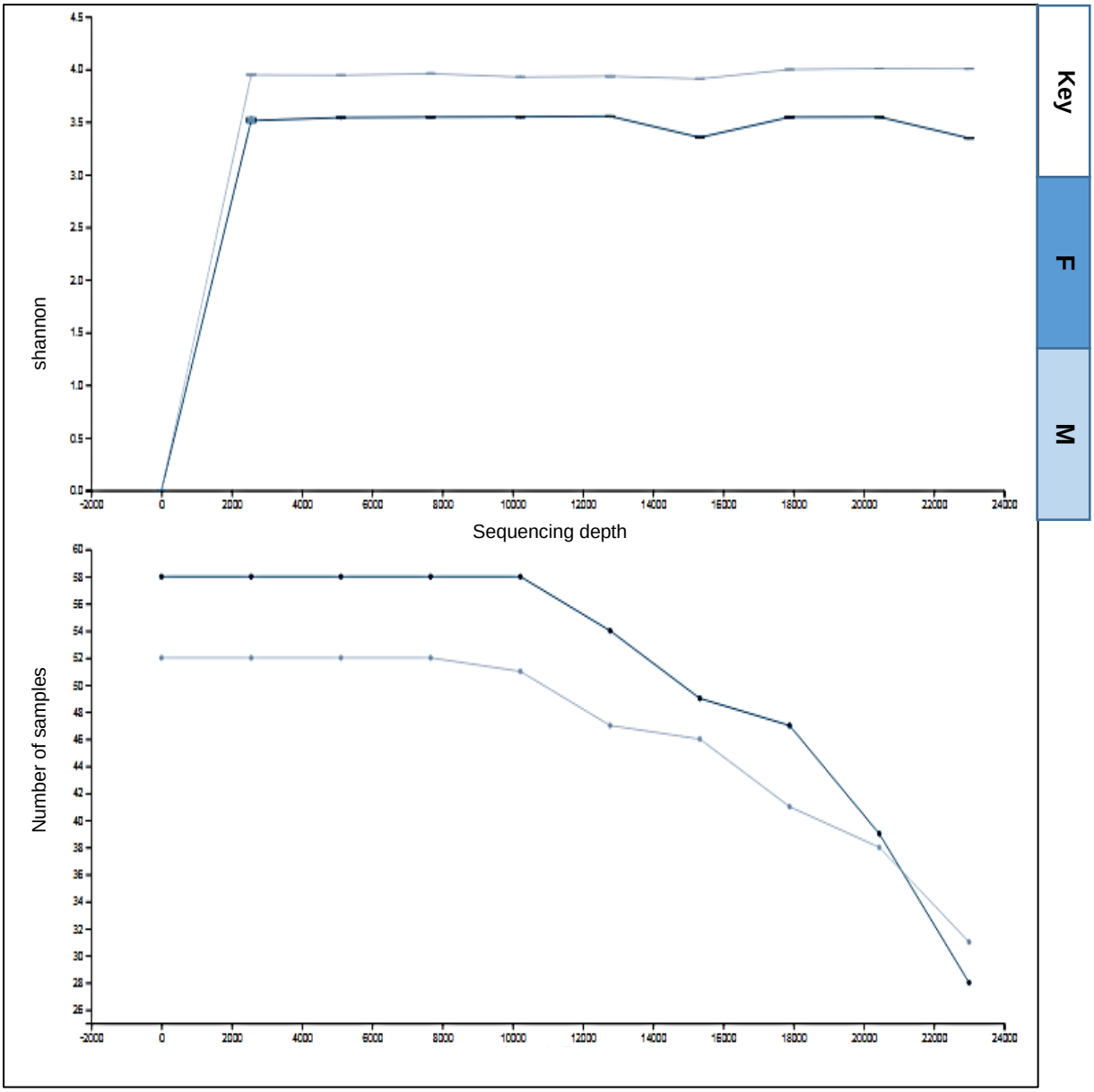


Figure J 1.2: Alpha rarefaction curves of male (M) and female (F) infants computed with Shannon Diversity metrics. The exponential growth and plateau of the curves show species richness is fully observed at a maximum index of 4 (males – light blue) and sequencing depth of 2 000. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000. The key on the right relates the colours of the curves with the sex of the infant.

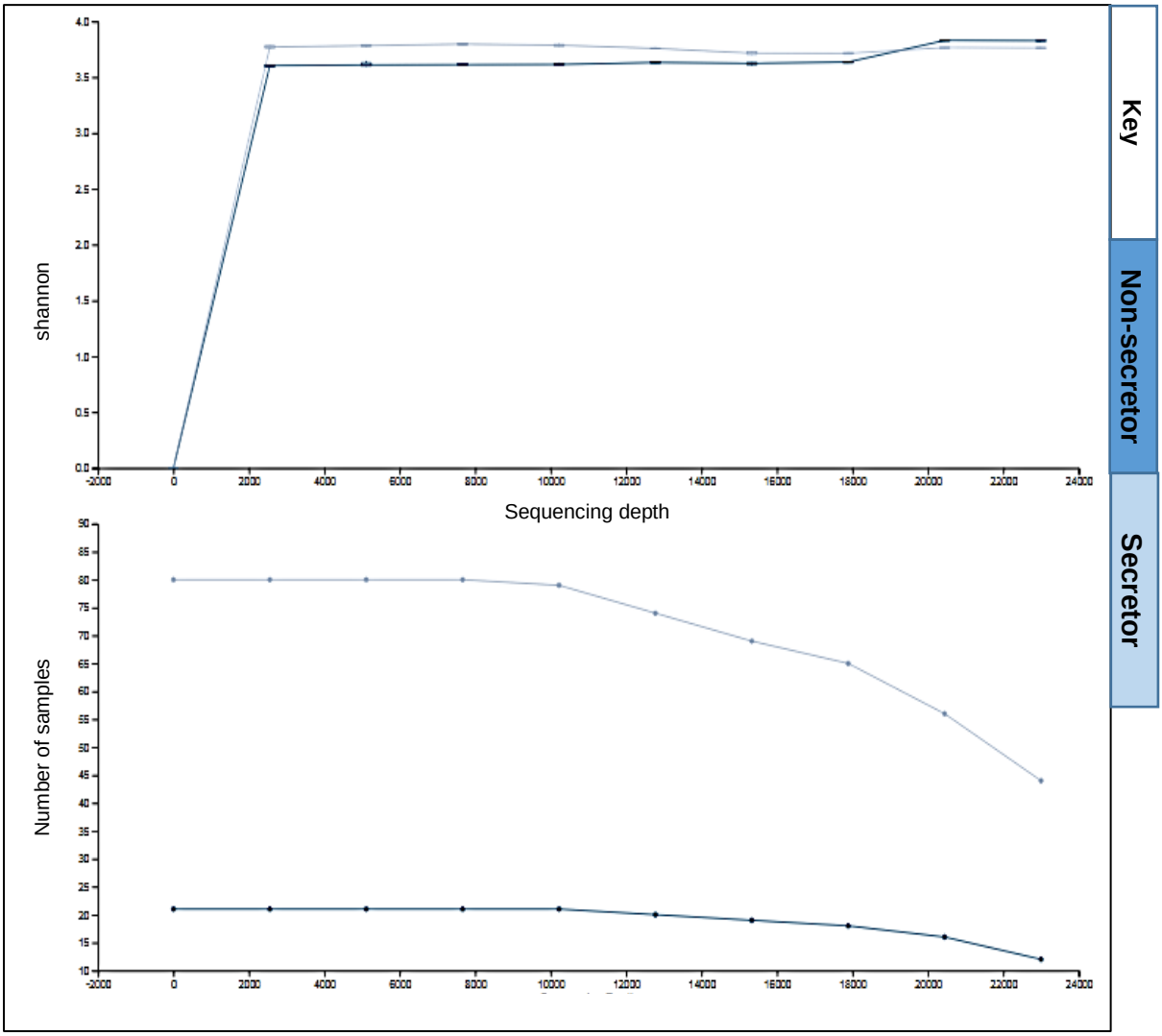


Figure J 1.3: Alpha rarefaction curves of secretor status computed with Shannon Diversity metrics. The exponential growth and plateau of the curves show species richness is fully observed at a maximum index of 4 (non-secretors – dark blue) and sequencing depth of 2 000. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000. The key on the right relates the colours of the curves with the secretor status.

2. Amplicon sequence variants alpha rarefaction curves

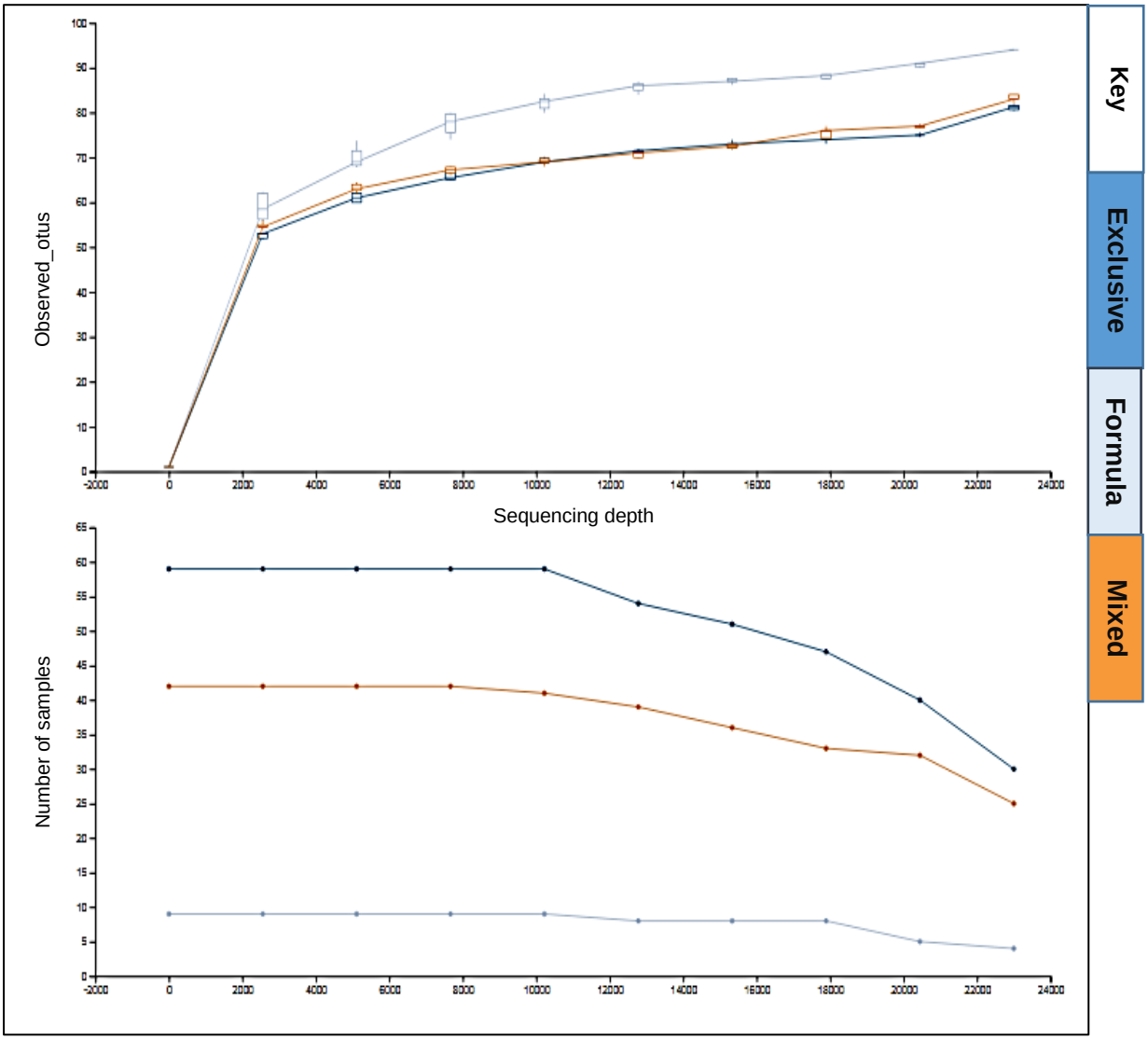


Figure J 2.1: Alpha rarefaction curves of the different feeding practices computed with amplicon sequence variants counts (labelled as “observed_otus” on the y-axis). Exponential growth of the curves is visible until a sequencing depth of 2 000 and 50 ASVs, indicating common species richness is fully observed. As rarer species are subsampled the curves begin to level out, and collecting additional sequences may result in additional features being observed. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000 for exclusive and mixed feeding, and after 18 000 for formula feeding. The key on the right relates the colours of the curves with the feeding practices.

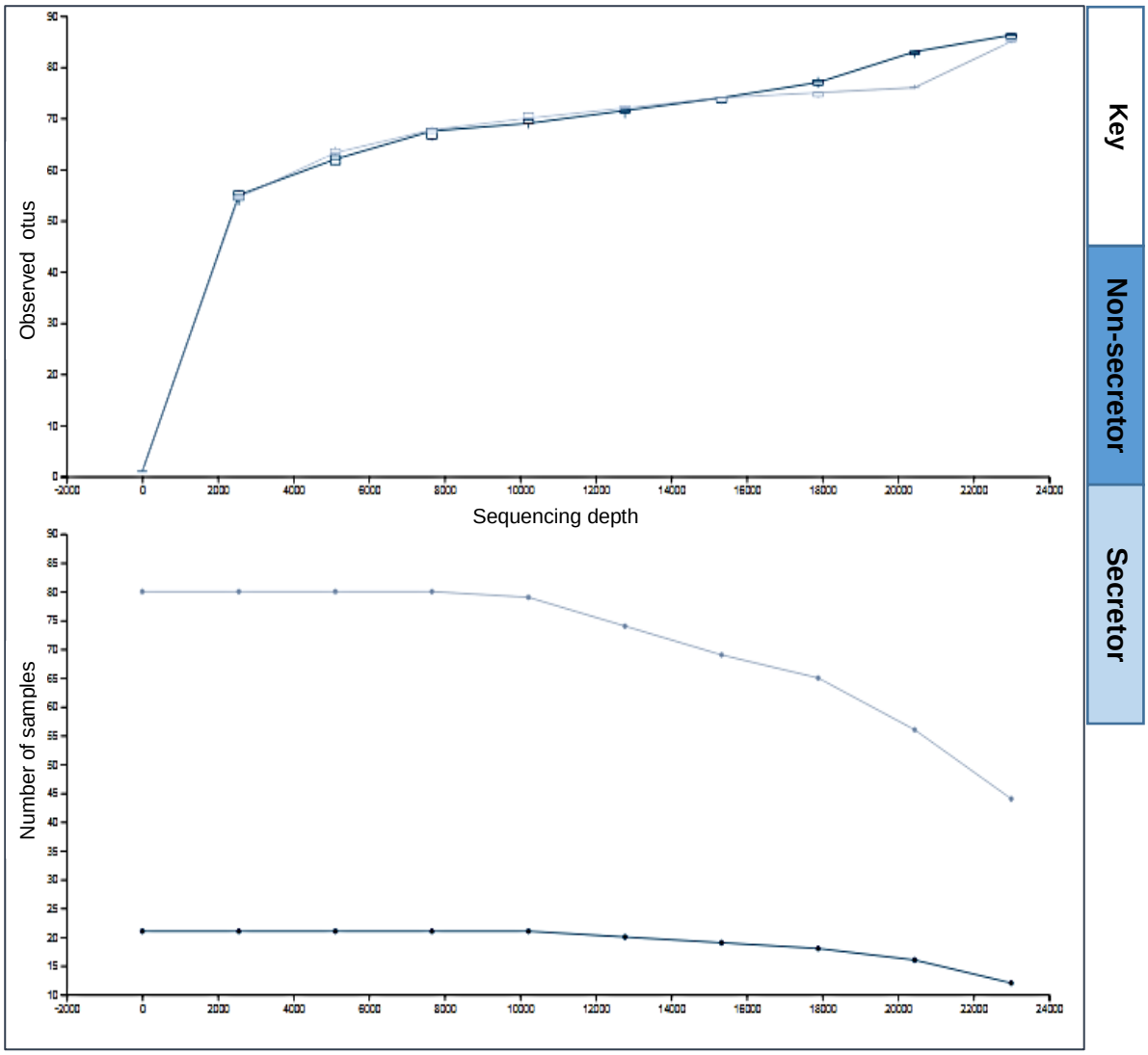


Figure J 2.2: Alpha rarefaction curves of secretor status computed with amplicon sequence variants counts (labelled as “observed_otus” on the y-axis). Exponential growth of the curves is visible until a sequencing depth of ~2 000 and ~55 ASVs, indicating common species richness is fully observed. As rarer species are subsampled the curves begin to level out, and collecting additional sequences may result in additional features being observed. The bottom plot indicates the number of samples rarefied to the sequencing depth rapidly decreases after 10 000 for secretors and gradually for non-secretors. The key on the right relates the colours of the curves with the secretor status.

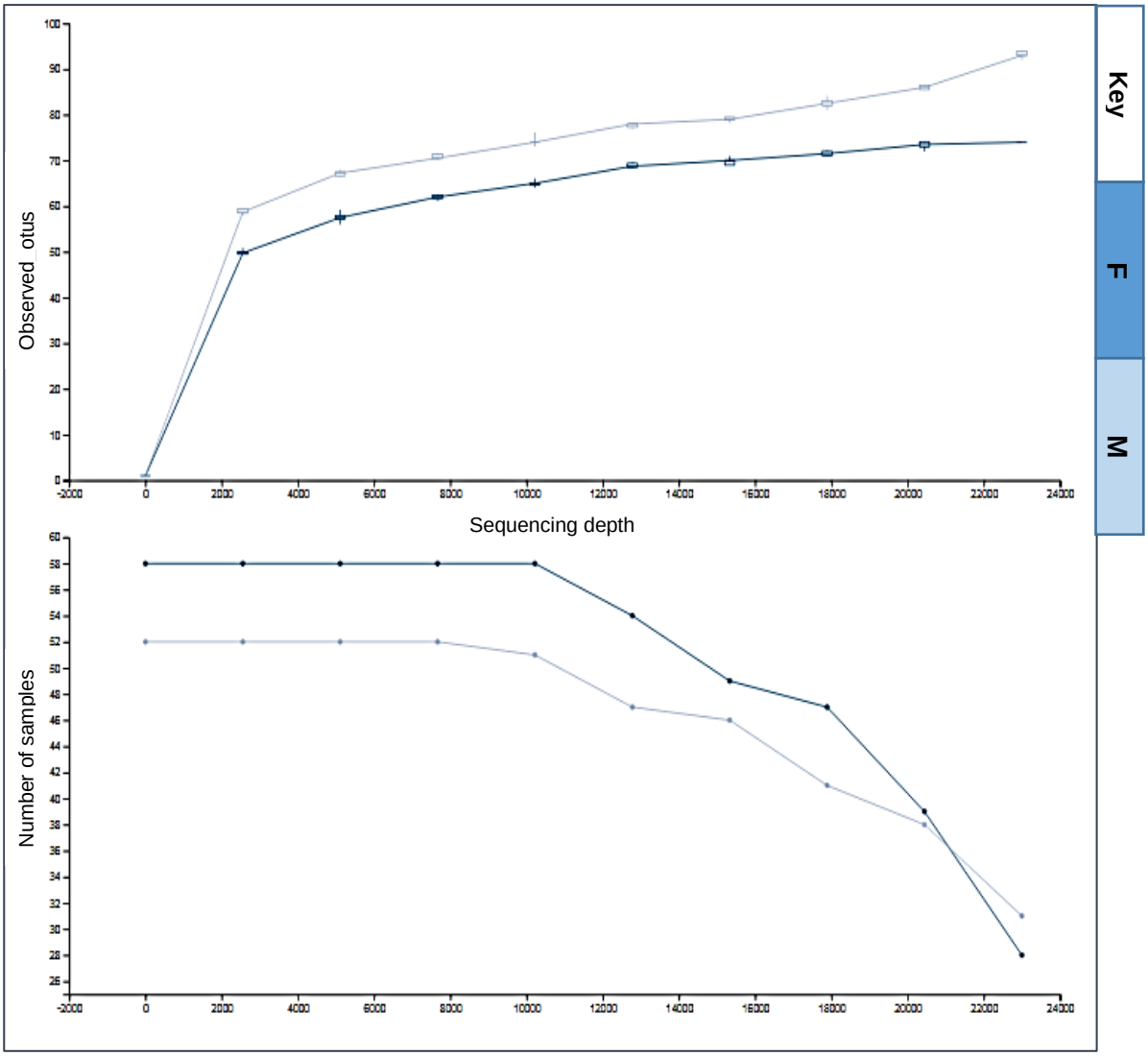


Figure J 2.3: Alpha rarefaction curves of male (M) and female (F) infants computed with amplicon sequence variants counts (labelled as “observed_otus” on the y-axis). Exponential growth of the curves is visible until a sequencing depth of ~2 000 and ~60 ASVs (males – light blue), indicating common species richness is fully observed. As rarer species are subsampled the curves begin to level out, and collecting additional sequences may result in additional features being observed. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000. The key on the right relates the colours of the curves with the sex of the infant.

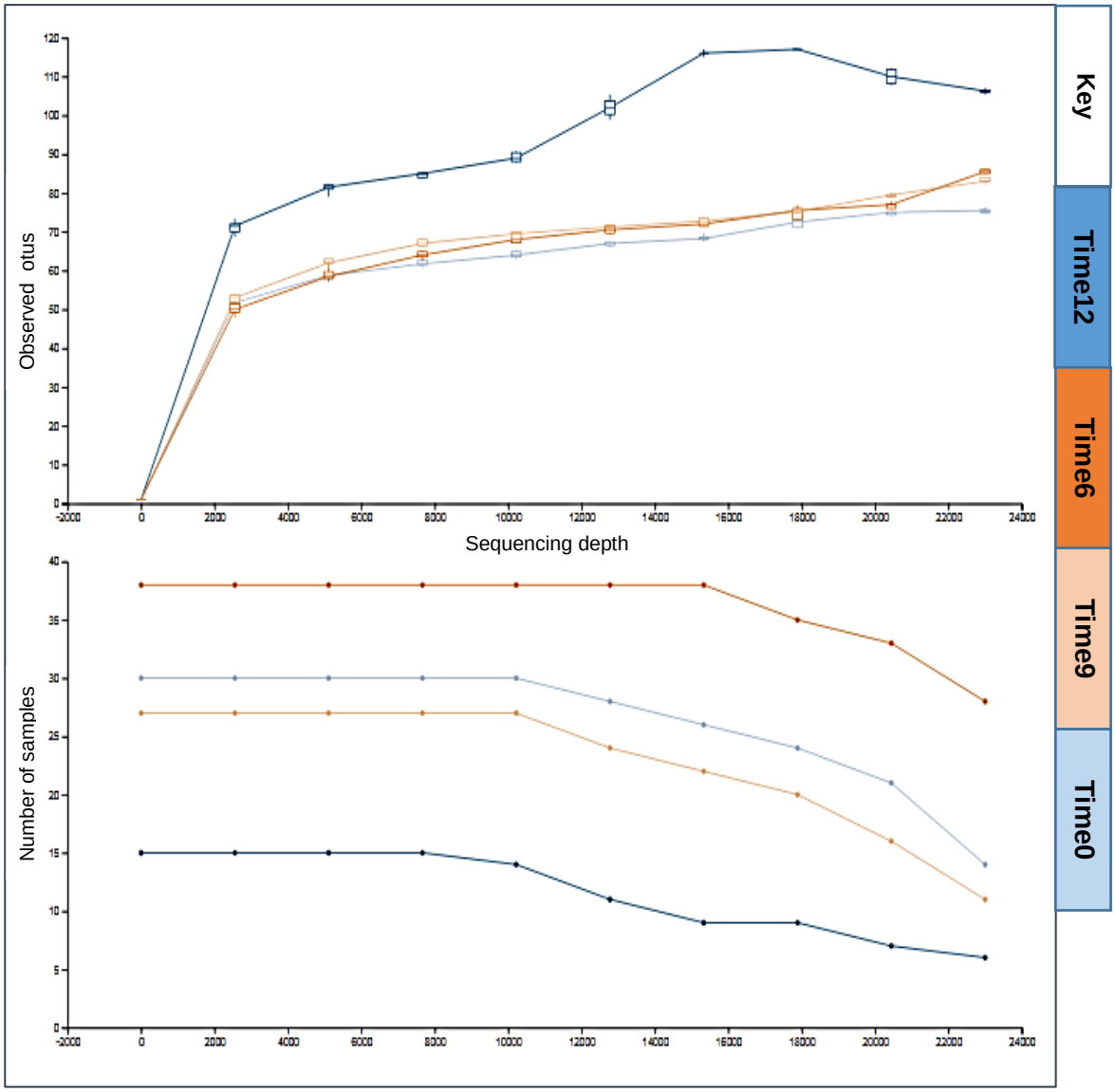


Figure J 2.4: Alpha rarefaction curves of the different time points computed with amplicon sequence variants counts (labelled as “observed_otus” on the y-axis). Exponential growth of the curves is visible until a sequencing depth of ~2 000 and ~70 ASVs (Time12 – dark blue), and the curves then level out indicating common species richness is fully observed. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000 for all time points except Time6 at 16 000. The key on the right relates the colours of the curves with the time points.

3. Faith's Phylogenetic Diversity alpha rarefaction curves

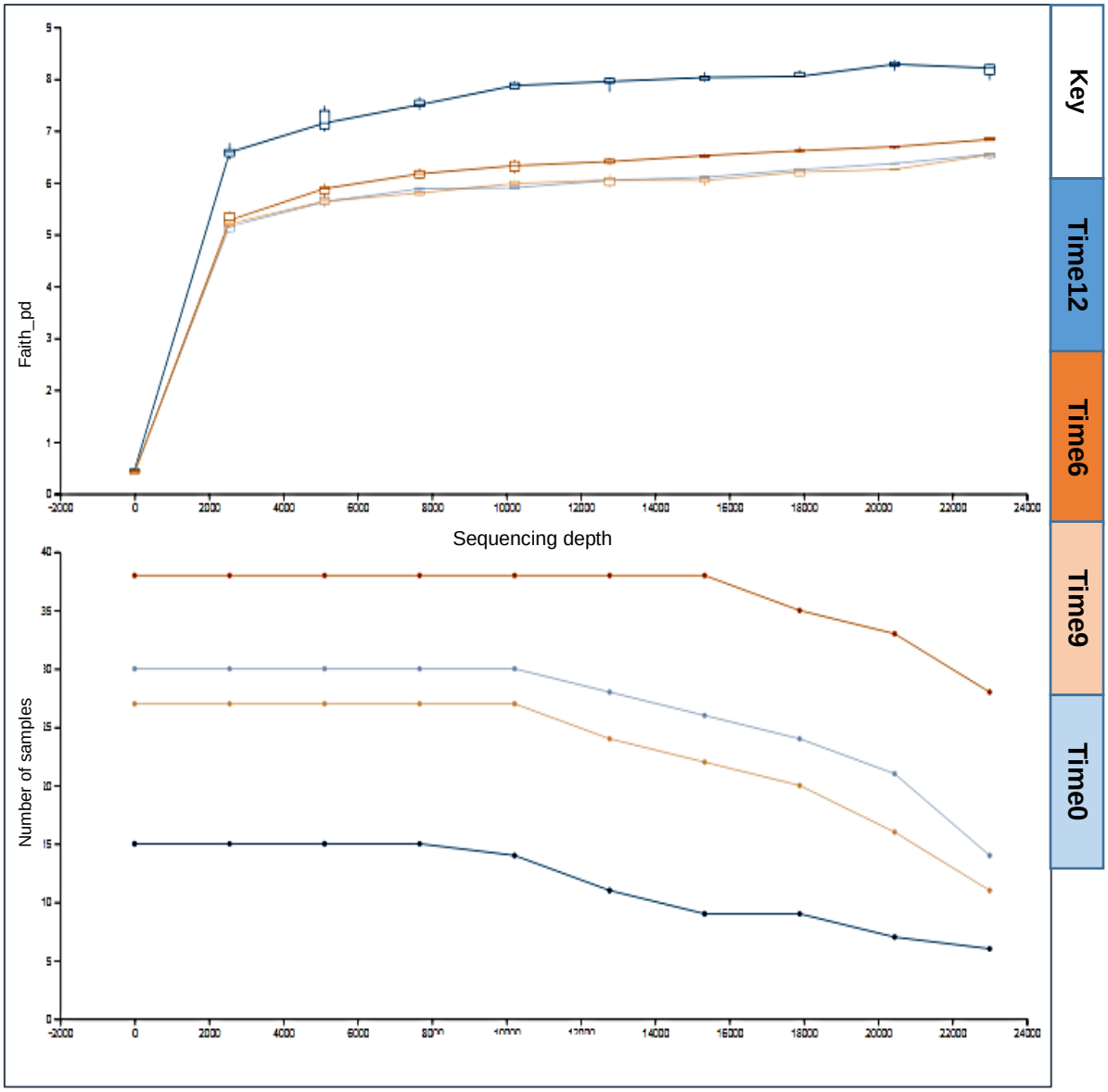


Figure J 3.1: Alpha rarefaction curves of the different time points computed using Faith's PD. Exponential growth of the curves is visible until just after a sampling depth of 2 000. The highest species richness index is observed at Time12 (dark blue) at ~6.5, and the curves level out indicating species richness is fully observed. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000 for all time points except Time6 at 16 000. The key on the right relates the colours of the curves with the time points.

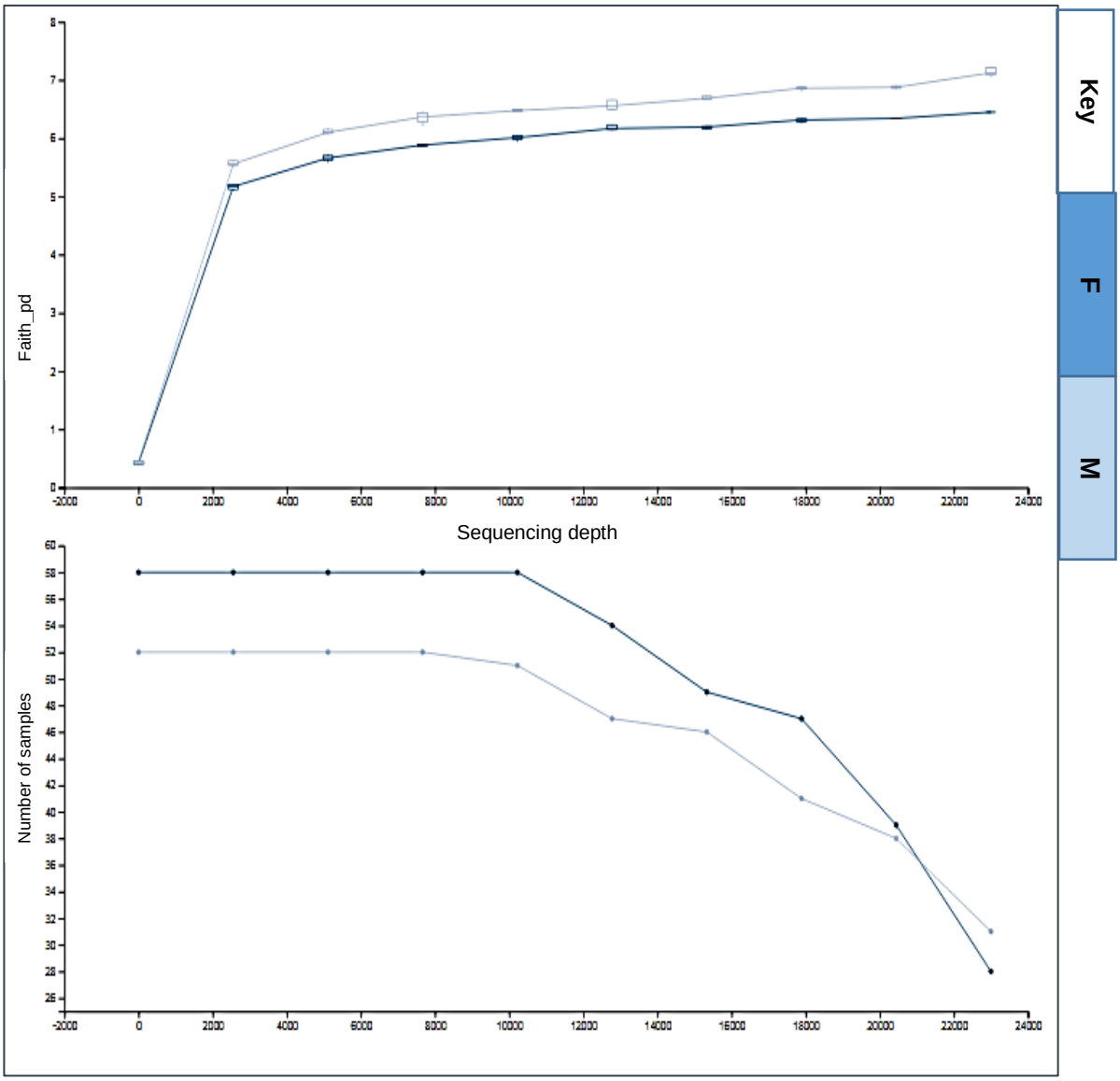


Figure J 3.2: Alpha rarefaction curves of male (M) and female (F) infants computed using Faith's PD. Exponential growth of the curves is visible until a sampling depth of 2 000. The highest species richness index is observed in males (light blue) at ~5.5, and the curves level out indicating species richness is fully observed. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000. The key on the right relates the colours of the curves with the sex of the infant.

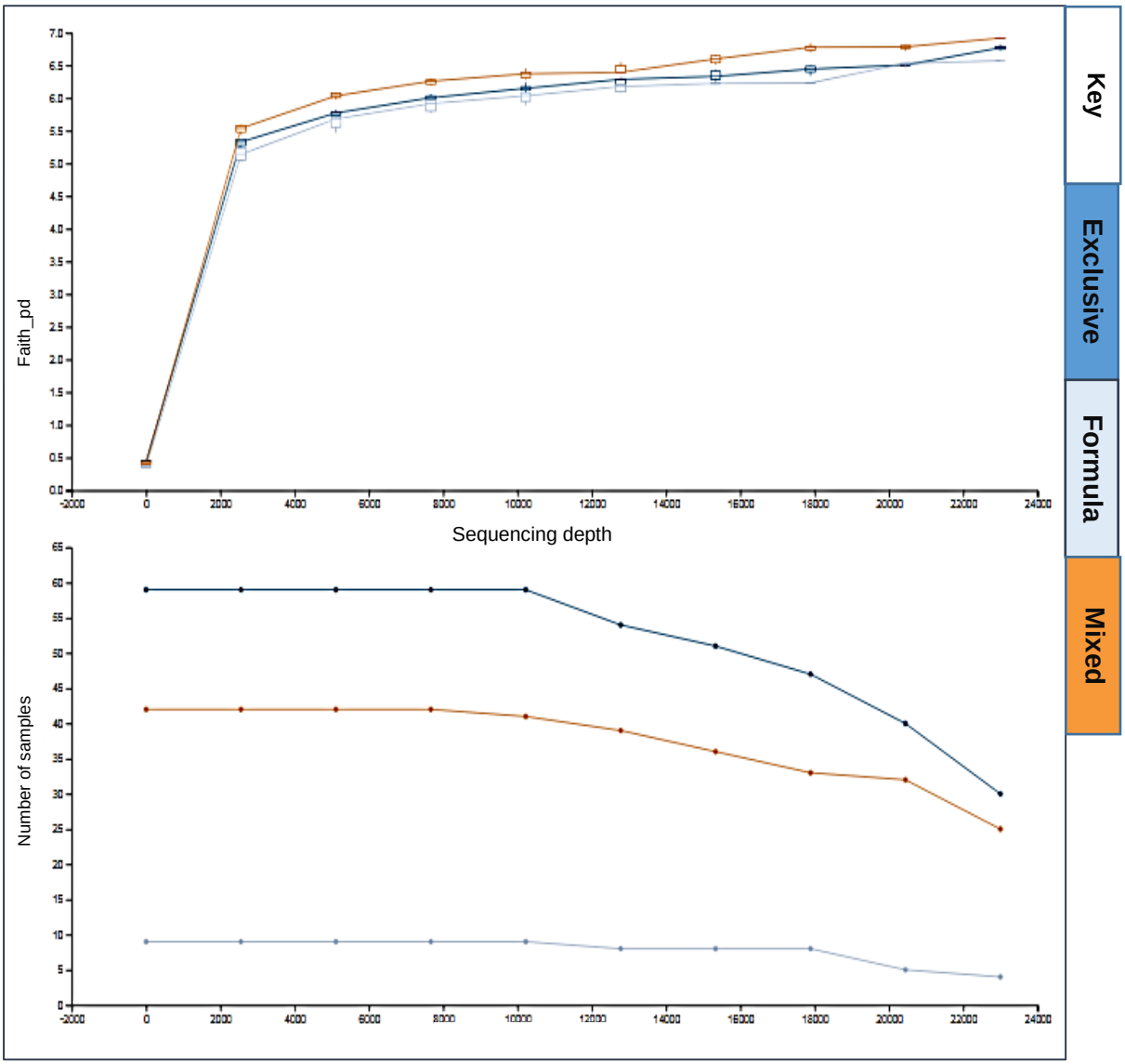


Figure J 3.3: Alpha rarefaction curves of the different feeding practices computed using Faith's PD. Exponential growth of the curves is visible until a sampling depth of 2 000. The highest species richness index is observed in mixed feeding practices (orange) at ~5.5, and the curves level out indicating species richness is fully observed. The bottom plot indicates the number of samples rarefied to the sequencing depth decreases after 10 000 for exclusive and mixed feeding, and 18 000 for formula feeding. The key on the right relates the colours of the curves with the feeding practices.

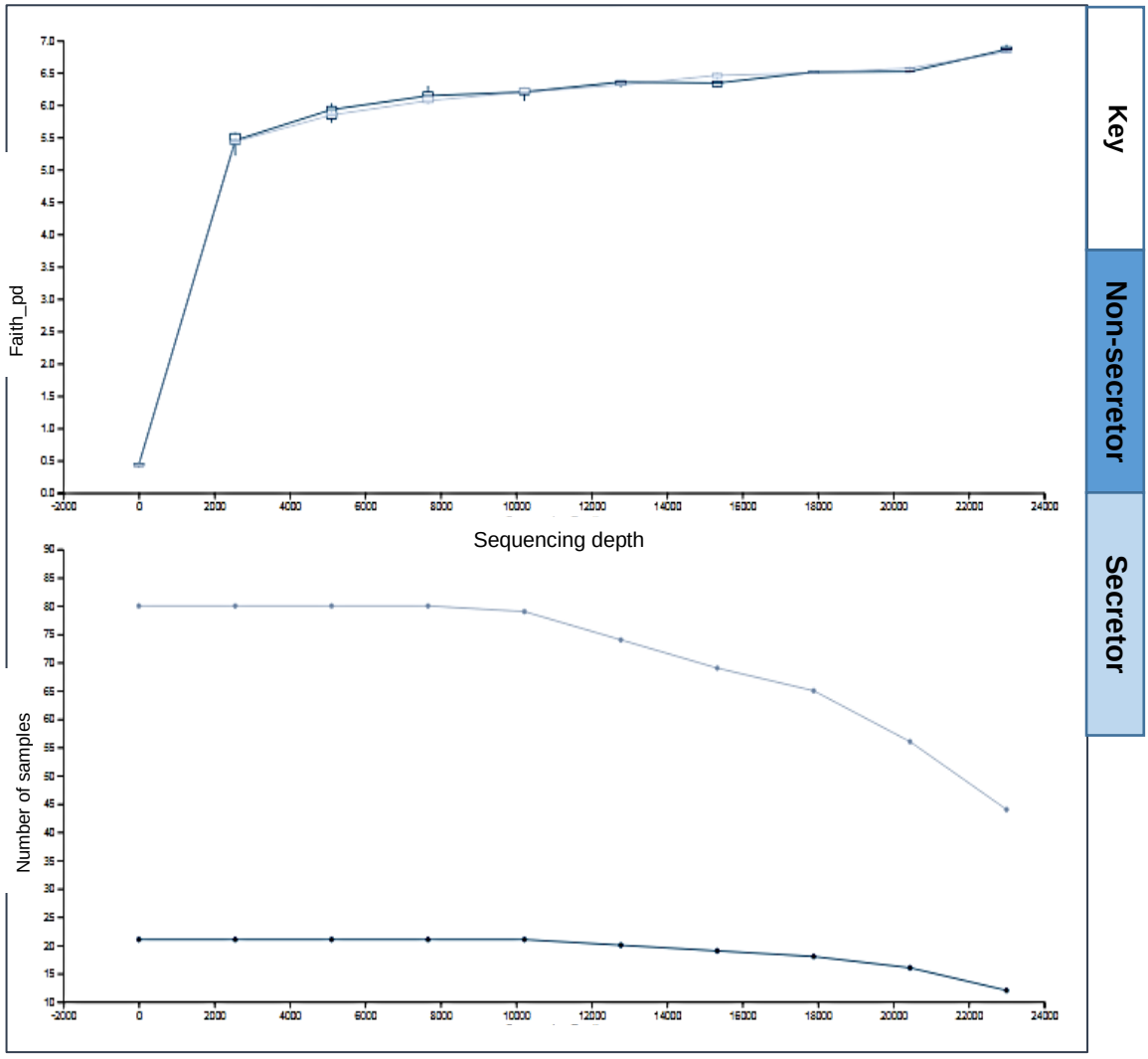


Figure J 3.4: Alpha rarefaction curves of secretor status computed using Faith's PD. Exponential growth of the curves is visible until a sampling depth of 2 000. The highest species richness index is observed at ~5.5, and as rarer species are subsampled the curves begin to level out, and collecting additional sequences may result in additional features being observed. The bottom plot indicates the number of samples rarefied to the sequencing depth rapidly decreases after 10 000 for secretors and gradually for non-secretors. The key on the right relates the colours of the curves with the secretor status.

Appendix K – Beta diversity PCoA plots

1. Beta diversity metrics computed on all time points

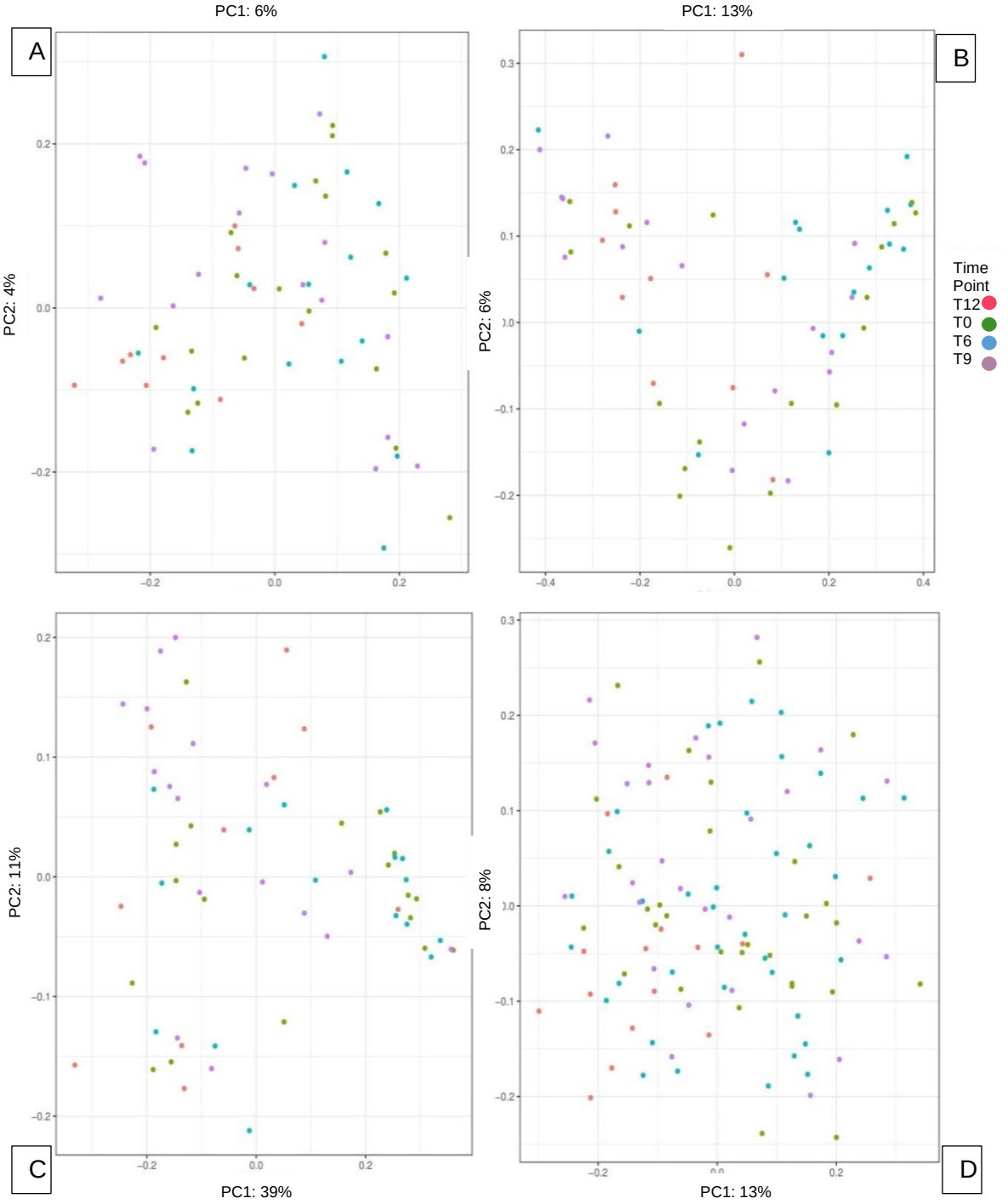


Figure K 1: Principle Coordinates Analysis plots for all time points for Beta diversity metrics: A) Jaccard Distance B) Bray-Curtis Distance C) Weighted Unifrac D) Unweighted Unifrac is showing that distinct clusters are not formed for the different time points. PC= principle coordinates 1 and 2 show the percentage of variation accounted for, and the x- and y-axis values are loading scores indicating which variable has a larger effect.

2. Beta diversity metrics computed on feeding practices

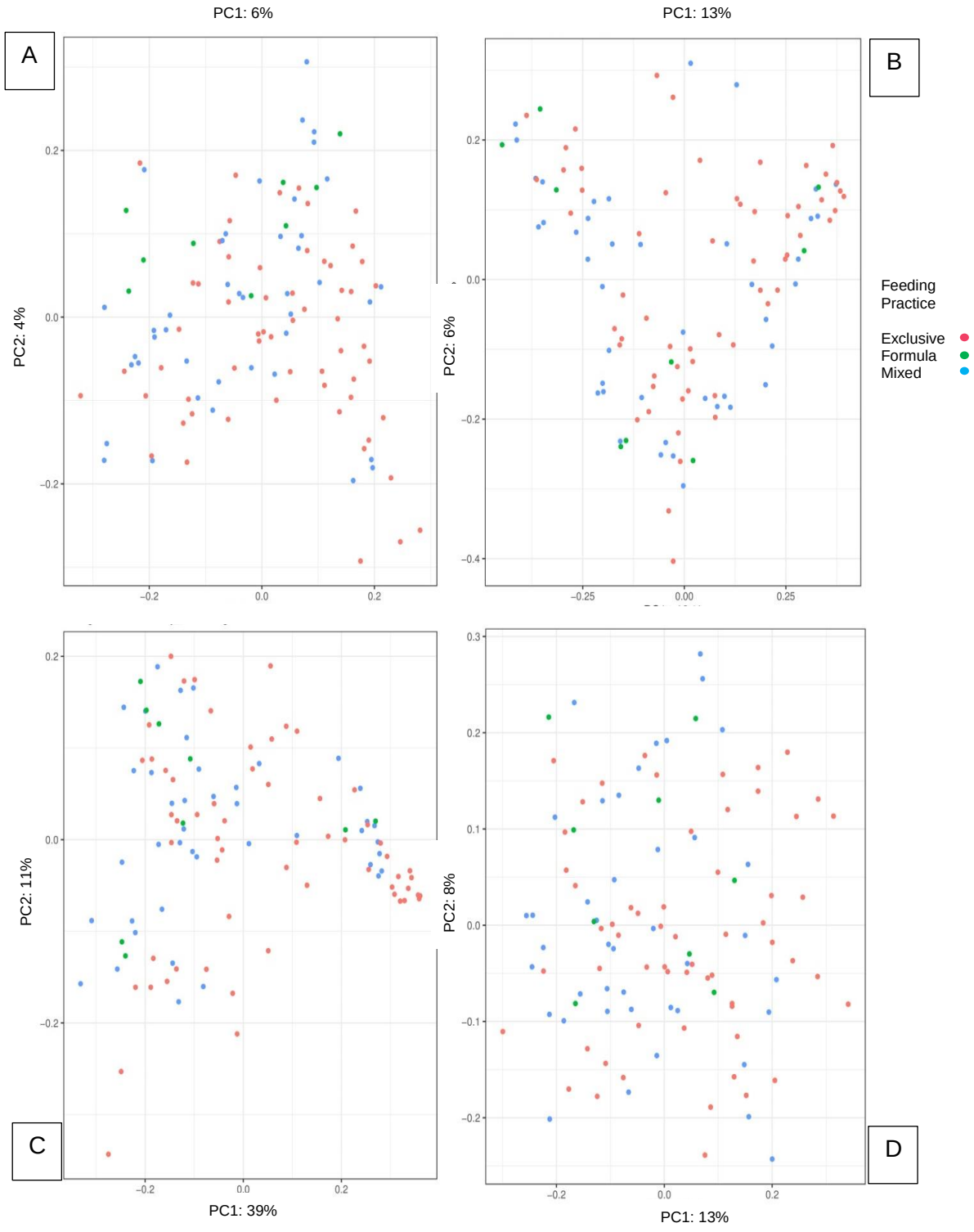


Figure K 2: Principle Coordinates Analysis plots for feeding practices for Beta diversity metrics: A) Jaccard Distance B) Bray-Curtis Distance C) Weighted Unifrac D) Unweighted Unifrac is showing that distinct clusters are not formed for the different feeding practices. PC= principle coordinates 1 and 2 show the percentage of variation accounted for, and the x- and y-axis values are loading scores indicating which variable has a larger effect.

3. Beta diversity metrics computed on infant sexes

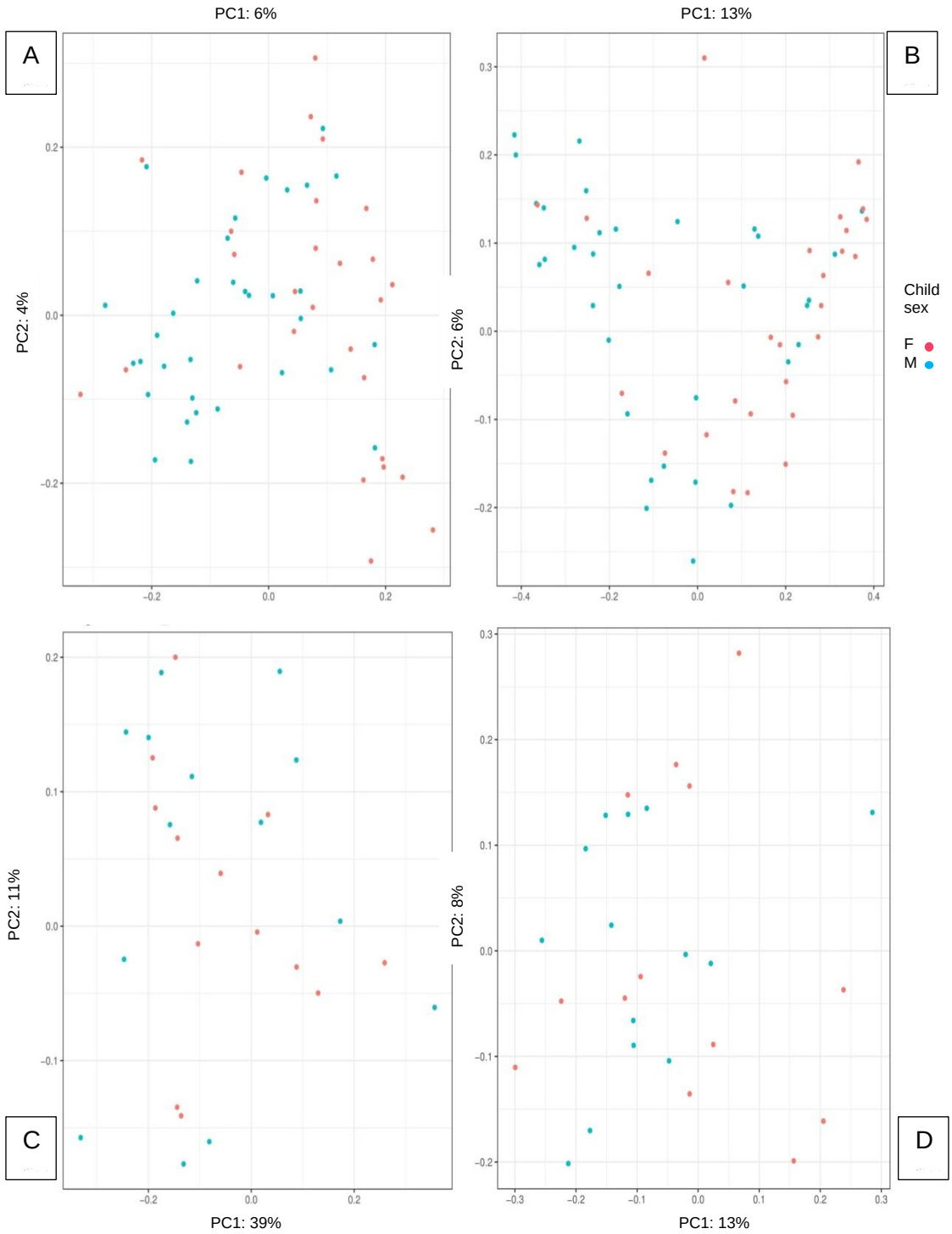


Figure K 3: Principle Coordinates Analysis plots for sex of the infant for Beta diversity metrics: A) Jaccard Distance B) Bray-Curtis Distance C) Weighted Unifrac D) Unweighted Unifrac is showing that distinct clusters are not formed for the different infant sexes. PC= principle coordinates 1 and 2 show the percentage of variation accounted for, and the x- and y-axis values are loading scores indicating which variable has a larger effect.

4. Beta diversity metrics computed on secretor status

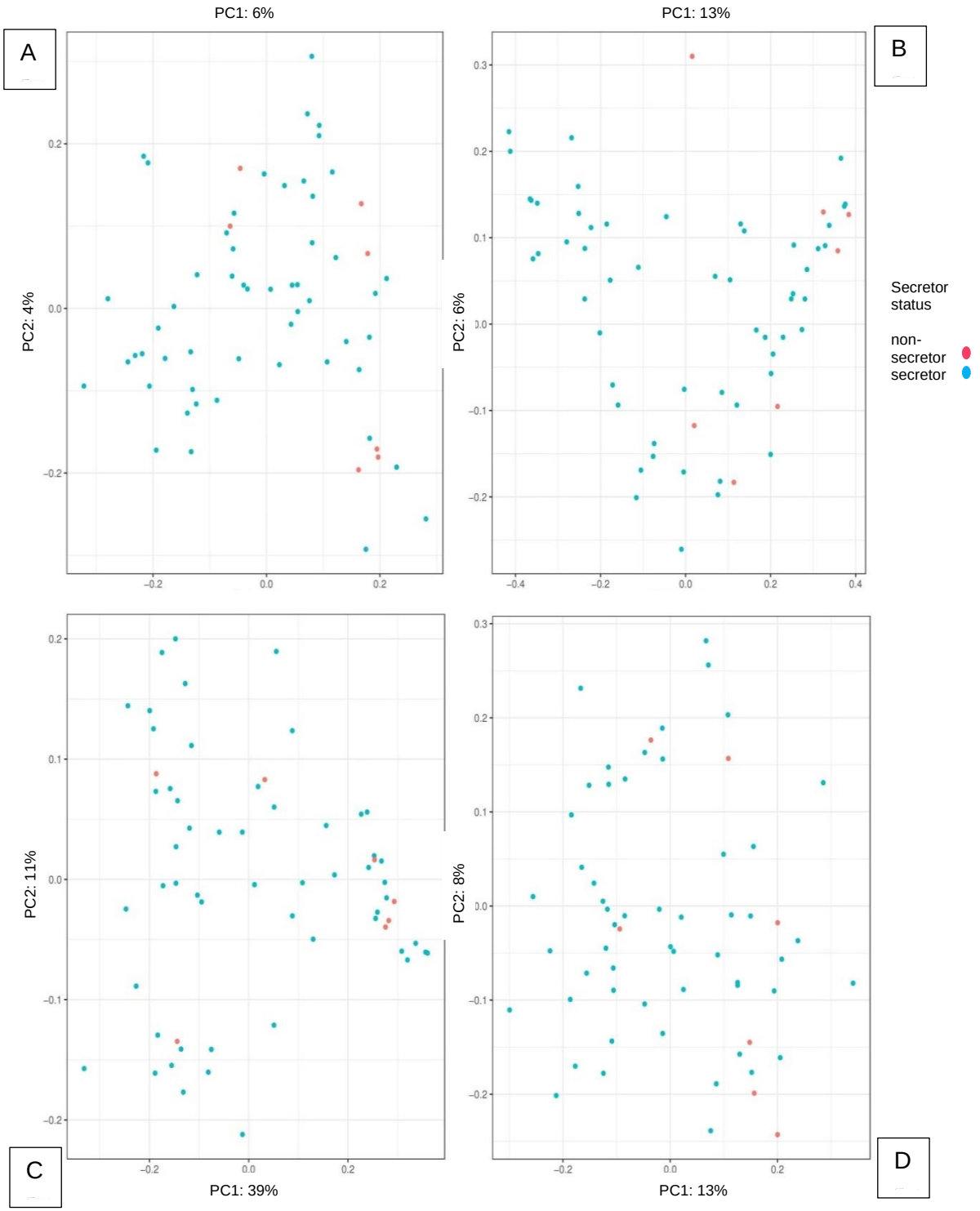


Figure K 4: Principle Coordinates Analysis plots on secretor status for Beta diversity metrics: A) Jaccard Distance B) Bray-Curtis Distance C) Weighted Unifrac D) Unweighted Unifrac is showing that distinct clusters are not formed for the different secretor status. PC= principle coordinates 1 and 2 show the percentage of variation accounted for, and the x- and y-axis values are loading scores indicating which variable has a larger effect.

Appendix L – Beta diversity using PERMANOVA (in QIIME2)

1. Secretor status testing with PERMANOVA

Table L 1.1: PERMANOVA (in QIIME2) statistics summary of the analysis of secretor status including overall *P*-value showing statistical insignificance

Method name	PERMANOVA
Test statistic name	Pseudo-F
Sample size	101
Number of groups	2
Test statistic	1.52
P-value	0.06
Number of permutations	999

Table L 1.2: PERMANOVA statistical values of the pairwise test of the analysis of secretor versus non-secretor status

Group 1	Group 2	Sample size	Permutations	pseudo-F	*P*-value	q-value
non-secretor	secretor	101	999	1.52	0.05	0.05

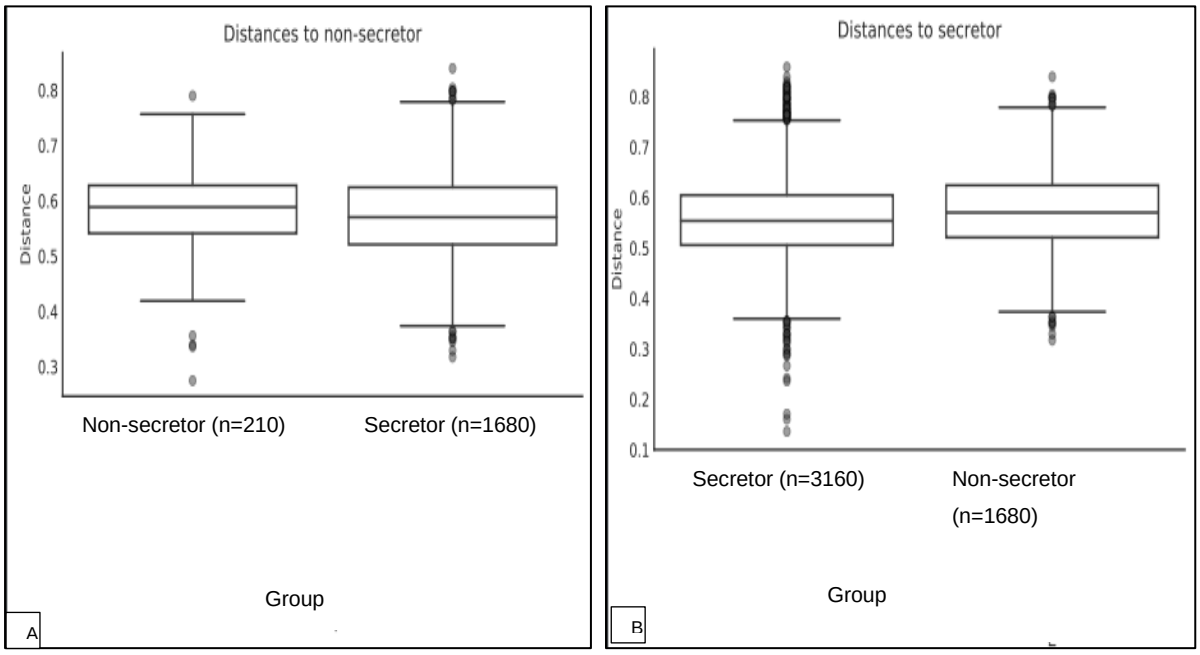


Figure L 1.1: Unweighted Unifrac distances of secretor status using PERMANOVA (QIIME2) for A) Distances to non-secretor, B) Distances to secretor. This shows pairwise distances in each graph for within groups (the first box-plot) and between groups (the second box-plot) are similar for non-secretor and secretor status.

2. Time point testing with PERMANOVA

Table L 2.1: PERMANOVA (in QIIME2) statistics summary of the analysis of all time points including overall *P*-value showing statistical significance

Method name	PERMANOVA
Test statistic name	Pseudo-F
Sample size	110
Number of groups	4
Test statistic	1.88
P-value	1.00×10^{-3}
Number of permutations	999

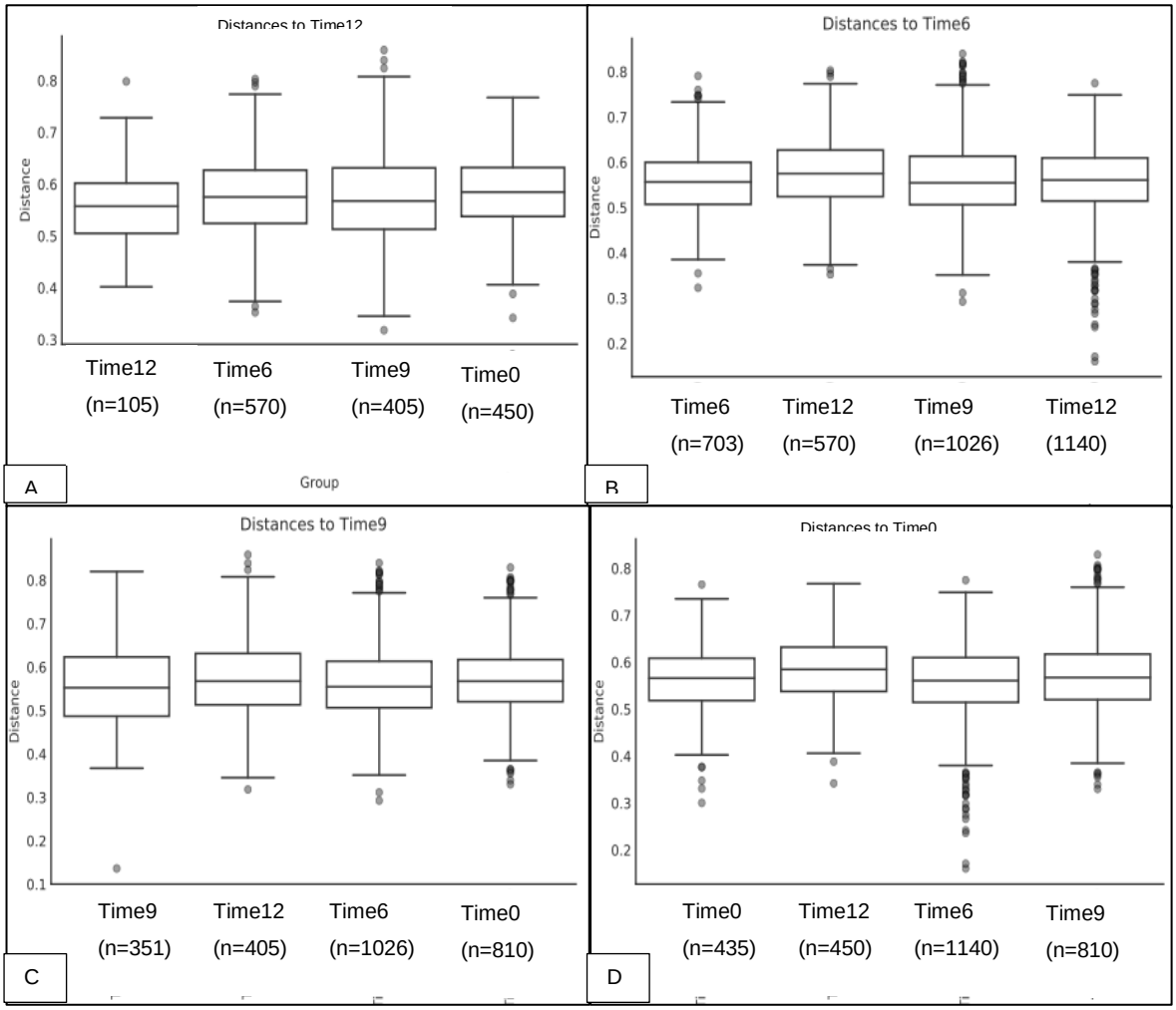


Figure L 2.1: Unweighted Unifrac distances of time points using PERMANOVA (QIIME2) for A) Distances to Time0, B) Distances to Time6, C) Distances to Time9, D) Distances to Time12. This shows pairwise distances in each graph for within groups (the first box-plot) and between groups (the subsequent box-plots). The smallest distance is seen **within** the group for Time0 and Time9 (A, C), and **between** groups for Time6 and Time12 when compared to Time0 (B, D).

Appendix M – Taxonomic analysis on feeding practice using LEfSe

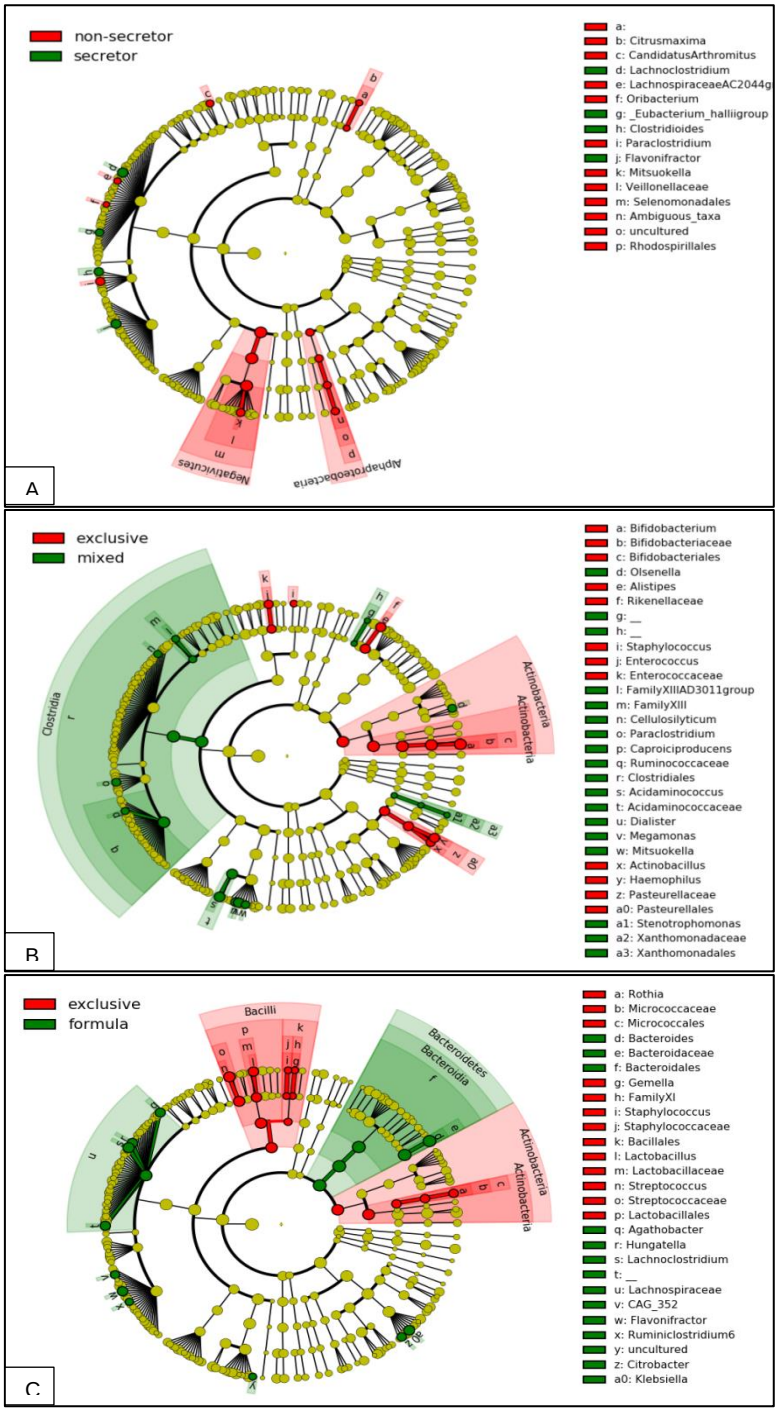


Figure M 1: Cladograms of LEfSe genus-level analysis showing: A) **Secretor status** groups highlighting that Firmicutes and Proteobacteria are differentially abundant in non-secretors. B) **Exclusive versus mixed feeding** groups highlights that Firmicutes (*Clostridia*) are differentially abundant in mixed feeders, and Actinobacteria are differentially abundant in exclusive feeders. C) **Exclusive versus formula** feeding groups highlights that, Actinobacteria and Firmicutes (*Bacilli*) are differentially abundant in exclusive feeding and Bacteroidetes and Firmicutes are differentially abundant in formula feeders.

Appendix N – Ethics approval clearance certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Miss Claudine Nkera-Gutabara

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M180754

NAME: (Principal Investigator)	Miss Claudine Nkera-Gutabara
DEPARTMENT:	School of Pathology Division of Human Genetics
PROJECT TITLE:	Infant feeding practices and the gut microbiome: impact on infant growth and development in South Africa
DATE CONSIDERED:	27/07/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr Venesa Sahibdeen
APPROVED BY:	Prof C Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	07/08/2018

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed July and will therefore be due in the month of July each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES