Intestinal microbiota assembly and dynamics in health and disease

Intestinal microbiota assembly and dynamics in health and disease:

A focus on longitudinal data analysis

Gianluca Galazzo

ISBN: 9789464193282
Cover & Layout: Gianluca Galazzo
Printing: Gildeprint

The research presented in this thesis was conducted within NUTRIM School of Nutrition and Translational Research in Metabolism and CAPHRI School of Care and Public Health Research Institute of Maastricht University.

Intestinal microbiota assembly and dynamics in health and disease:

A focus on longitudinal data analysis

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,

on the authority of the Rector Magnificus,

Prof.dr. Rianne M. Letschert

in accordance with the decision of the Board of Deans,

to be defended in public

on Friday 5 November 2021, at 12:00 hours

by

Gianluca Galazzo

Supervisor:

Prof.dr. Paul H.M. Savelkoul

Co-supervisor:

Dr. John Penders

Assessment Committee :

Prof.dr. Frederik-Jan van Schooten (chair)

Dr. Hans Bogaards (Amsterdam UMC)

Prof. dr. Kristel Van Steen (University of Liège, Belgium)

Prof. dr. Koen Venema

Table of contents

Chapter 1


groups
metagenomics
nbmm
cup
however
main
state
approach
composition
bias
among
available
fitcase
taxa
research
results
one
red
study
step
time
also
analysis
gut
affect
used
data
microbiota
studies
sample
microbiome
methods
different
sequencing
species
bacterial
sequence
approaches
variation
structure
technical
outcome
based
sequences
related
may
now
size
otus
bacteria
factors
extraction
observed
section
community
identify
collection
crosssectional
evidence
another
many
model
samples
microbial
biological
coda
traditional
dna
balls
first
diet
far
per
thus
due
two
must
still
able
shifts
let
fill
otu
ibd
cells
set
adult
ratio
tests
birth
lead
will
real
side
vastlimited
host
relies
well
like
pcr
aldex
slot
clr
user
past
finally
mixed
often
work
tools
see
uses

General introduction

Introduction

The human intestinal tract is a unique habitat creating a nutrient-rich environment for its microbial inhabitants, while in return these microorganisms and their metabolites interact with the host. This mutual relationship provides the host with benefits such as metabolic balance[1-4], processing of nutrients, including fibres digestion, vitamin synthesis, colonization resistance against invading pathogens[5, 6] and maturation and homeostasis of the gastrointestinal lymphoid tissues[7].

The intestinal microbiota is a complex community counting up to 100 trillion of microbes, including bacteria, archaea, viruses and fungi that have co-evolved with the host[9]. Between 40–60% of the bacteria residing within the gut are reported to be unculturable[10]. Thus, current research relies on DNA-based, culture-independent methods for a comprehensive characterization of the intestinal microbiota[11]. Technological improvements have instigated microbiome research over the past years leading to a rapid expansion of knowledge on the ecological dynamics of gut microbiota and allowing for bigger and more complex, intervention and cohort studies [12].

Initial colonization of the human intestinal tract starts at birth with the rupture of the amniotic membranes and subsequent passage through the birth canal where the infant is seeded by maternal microbial strains, a process which is impeded in case of a caesarean section delivery[13, 14]. Subsequently, microbial populations evolve as the diet changes and the host matures. The infant gut microbiota is very unstable, showing big fluctuations in composition during the first 2,5 years of life[15, 16]. It has been argued that this time window is crucial for the maturation of the immune system.

Animal studies have highlighted the importance of the gut microbiota in educating the immune system. In germ-free animals the gastrointestinal lymphoid tissue shows less IgA+ B cells, but colonization with lactobacilli strains replenishes these IgA-producing plasma cells[17]. The early colonization appears to be of particular importance as *Bifidobacterium longum* subsp. *infantis* could restore Th1 responses in neonatal but not in adult ex germ-free mice[18].

At around school-age the microbiota stabilizes[19] and resembles the mature adult composition. Once matured, the gut microbiota has been shown to be stable and relatively resilient[20-23]. Nevertheless, it can also undergo dramatic compositional shifts, a condition known as dysbiosis, due to stressors like profound changes in diet, antibiotics use or diseases. During the past decade, many researchers investigated the association between the microbiota and health, providing evidence on the importance of our indigenous microbes for maintaining human health. Numerous studies have shown association between the gut microbiota and infant health[24-28], autoimmune disorders[29-31], obesity[32-36], diabetes[37-39], IBD[40-45], and longevity[46-50]. However, the causal mechanism behind this necessity is still far from being understood.

Aiming to fill this gap in knowledge, researchers made efforts to investigate the dynamics of gut microbiota over time. Longitudinal studies, in fact, provide more information than cross-sectional studies, providing richer information about the system under study especially because of the inherent, irreversible ordering of the samples.

The past years have been the golden age for microbiota research with an exponential increase in studies and publications. Nevertheless, the knowledge of the gut microbiota and its role in human health is still growing and in need of more fine-tuned studies. The

vast majority of microbiota research is still based upon cross-sectional studies that can only partially explain the role of the gut microbiota in health and disease. Additionally, many of these studies did not adequately adjust for confounding factors known to affect the gut microbiota composition. Differences in diet or medication used among subjects of different groups can significantly bias the results and lead to misleading conclusions. Finally, the heterogeneity in sampling methods and processing, the rapid evolution in NGS technologies and software algorithms and the wide variety of applied downstream methods for data analysis all contribute to a lack of consistency and reproducibility among studies.

Sample collection and storage

A first and crucial step in the analysis of the gut microbiota is the procedure of sample collection. It is known that processing and storage immediately affects the composition of the microbial community of faecal samples. Although such non-biological factors don't explain most of the observed inter-individual variation, several previous studies have demonstrated that it can introduce a sizable effect on the microbial community structure[51]. Different procedures are used to collect and preserve faecal samples for microbiome analysis.

Lowering the temperature, in general, prevents the proliferation of bacteria, thereby preserving the microbial community structure in faecal samples. The first consortium that sequenced the gut microbiota on a large scale opted to freeze the samples immediately after collection at -80 °C[52]. However, this approach requires special equipment and thus is not always suitable or logistically feasible in large-scale population-based studies. A more suitable option is to freeze the samples at -20 °C. This temperature is reached by the majority of the home freezers allowing the participant to collect and temporarily store the samples at home. One of the main drawbacks of cooling or freezing samples is that the cold chain must be preserved during transport and subsequent processing of the samples in the laboratory. Failure to do so may dramatically impact metagenomic DNA quality and introduce bias as freeze-thaw cycles are well known to cause DNA degradation.

A multitude of other options have been explored to increase the user-friendliness of sample collection and storage. One such approach is to collect the faecal samples into tubes containing a stabilizing buffer (e.g., OMNIgene®-GUT, RNALater, DNA/RNA Shield® or 95% ethanol) after which samples can be stored at room temperature. On one hand, storage at room temperature and the usage of stabilizing buffers makes the sample collection and transport process much easier and flexible. On the other hand, those methods induce cell lysis of the bacteria limiting the analysis that can be performed (e.g., excluding the possibility of culture-based methods).

Other methods meant to be user-friendly include swabs and stabilizing cards that retain the DNA among the fibers of the paper. These methods have the advantage that they can be used everywhere without training. The drawbacks are represented by the unknown initial weight/amount of the faecal matter and the possible proliferation of aerobic bacteria if no stabilizing buffer is used. Moreover, for both approaches in which stabilization buffers are being used as well as for methods based upon stabilizing cards, the initial stool consistency (dry weight percentage), an important confounding factor (see below), can no longer be determined.

All the methods described above either intend to preserve the microbial structure of the sample or to improve the user convenience. It is crucial to use a single standardized procedure that is both logistically feasible and minimizes bias within a study. In addition, however, pooling of data to detect disease associations with specific microbial taxa across multiple studies is also becoming more important to strengthen current evidence. If individual studies collect faecal samples using different methods, conducting meta-analyses may reveal extensive heterogeneity. For this reason, efforts have been made to propose an international standard for sample collection in metagenomics studies, but so far protocols are still widely variable between published studies.

In conclusion, the choice of the sample collection method as well as the storage methods are very important aspects of a study design that can affect the study budget, the user-friendliness as well as the reproducibility of the results. For this reason, much more focus should be given to the standards suggested by the International Microbiome Standards Consortium[53].

How many samples to collect?

An adequate number of subjects must be recruited to ensure that the expected effect from the exposure/intervention on the microbiome or from the microbiome on a disease outcome of interest can be detected. Large-scale studies have provided a wealth of insight into which variables have the strongest effects on the microbiome[16, 54, 55], but also small and "tailored" studies with a limited scope have a large potential to advance the field[56]. The number of samples required for a microbiome study depends on the effect size, i.e., a quantitative measure of the differences between two or more groups. Several tools, such as "Evident" (https://github.com/biocore/Evident) and R packages like "micropower"[57] are currently available to perform power and sample size calculations for microbiome studies in order to guide in the required sample size.

A question that often arises while designing a study comprises whether to collect multiple repeated measurements from the same individual or to allocate the same resources to sample more subjects at a single time-point. Once again, the answer depends on the goal of the study. For example, if the goal of the study is to identify microbial biomarkers that can differentiate diseased individuals from healthy controls, then it may be more advisable to opt for a cross-sectional design with one sample per subject since the within-subject variability of the gut microbiota is quite stable over time[58-61]. On the other hand, if the aim of the study is to investigate if certain microbial taxa or community shifts are associated with disease course or treatment response, then a longitudinal design is preferable. Moreover, time-series data can also reveal interesting characteristics of the microbiota that are not apparent from single time points, such as the volatility of the microbiota and its resilience[20, 62, 63].

Collection of metadata

For optimal experimental design, it is important to include all information related to a sample. This includes information of the patient or healthy individual before, during, and after sample collection, the sample itself and the experimental procedures. Many factors in these metadata can influence the gut microbial composition. Among others, currently the most acknowledged are dietary habits, medication use (including in particular the use of antibiotics, proton pump inhibitors, laxatives and antidiabetics), age,

gender, body mass, stool consistency and bowel habits[64, 65]. However, as research in this field is rapidly progressing, more and more environmental and host factors are found to affect the gut microbiota. For this reason, it might be unpractical to register each and every possible known confounding factor. Therefore, the choice of the metadata that should be recorded should at least contain the major confounding factors and furthermore be hypothesis-driven. The population under study is also, to a large extend, determining which metadata should be collected. For example, in studies among newborns and infants it is crucial to collect data on birth mode and sibship size as they are known to affect the infant microbiome, whereas in adult populations these factors have limited impact on the inter-individual microbiome variation[54].

Questionnaires should therefore be designed both according to the outcome of the study as well as to the characteristics of the population studied. Even in this case, efforts to standardize the questionnaires have been made. The integrative Human Microbiome Project (iHMP) had designed a series of data formats to record clinical metadata in a consistent manner[66]. It is recommended to think about the interplay between technical variation, biological variation, and the temporal distribution of sample collection.

Sample processing

Just like sample collection, also sample processing plays an important role and introduces variability in the analysis of faecal samples. The steps involved in sample processing can be summarized in sample homogenization, DNA extraction and isolation and library preparation.

It is well known that depending on the protocol used to extract metagenomic DNA from stool samples, results in the generated microbial profiles can vary dramatically[67-73]. Comparing results from studies is therefore significantly hampered as biological variation cannot be distinguished from the technical variation introduced when using different DNA isolation protocols[74]. This is exemplified in the study by Sunagawa and colleagues in which they tested their newly developed method to establish metagenomic operational taxonomic units (mOTUs) based upon single-copy phylogenetic marker genes[75]. Applying their mOTU-approach on datasets from the Human Microbiome Project (HMP) and a European IBD cohort, the authors found a lower species diversity in asymptomatic US individuals (HMP) compared to those collected from European IBD patients. This is in large contrast with the widely observed reduced microbial diversity in IBD patients when compared to healthy controls and is most likely the result of the different isolation protocols in the two studies from which the data of asymptomatic US individuals and European IBD patients were retrieved.

Several international initiatives have therefore been initiated to compare isolation protocols and provide recommendations on the most optimal extraction method. The International Human Microbiome Standards consortium for example, compared 21 representative DNA extraction protocols and applied them on the same faecal samples, which were subsequently profiled using whole metagenome shotgun sequencing[74]. The analyses revealed that DNA extraction has a much stronger impact on the observed community composition than differences due to library preparation and sample storage. Taking into account, DNA quality and quantity as well as biases in community diversity estimates and the ratio of Gram-positive to Gram-negative bacteria, the authors were able to recommend a standardized DNA extraction method that was further

benchmarked using a mock community and appeared to be transferable across labs. It is therefore advisable to follow the SOPs described by the IHMS for DNA extraction or alternative methods that have benchmarked against this IHMS protocol.

Amplicon library preparation is also known to introduce various PCR artifacts that can impact the perception of a community, including the formation of chimeras[76-78], misincorporation of nucleotides[79], preferential amplification of some populations over others, leading to bias, and accumulation of random amplification events[80-82]. All those technical biases can be circumvented when using whole metagenome sequencing (WMGS) which lacks a PCR amplification step. But even this solution is far from being perfect as WMGS relies on DNA yield and is much more expensive and therefore not always feasible for large numbers of samples. The sequencing of mock communities as a benchmark for the accuracy is a common practice to identify the number of artificial sequences.

Finally, the accuracy of 16S rRNA gene and whole shotgun metagenomic sequencing is limited in practice by several processes that introduce biological contaminants, i.e., bacteria or bacterial DNA from the environment, the individual handling the samples and in the laboratory reagents. Next to technical (PCR or sequencing artefacts) contamination, also biological contamination falsely inflates within-sample diversity[84, 85], but also obscures differences between samples[85, 86]. One common practice to identify biological contaminants is to process reagent-only[84] or blank sampling instrument[87] negative control samples alongside biological samples at the DNA extraction and PCR steps. Including such negative controls along with mock communities will guide the cleaning of the sequencing data during the data pre-processing step.

Data Pre-processing

Pre-processing of the data is a fundamental step before any kind of data analysis should be performed. This step is meant to remove low-quality data and increase the reliability of the results and avoid problems such as inflation of the observed microbial diversity [88-90]. As described in the previous section, technical bias in the data can result from PCR artifacts (e.g., chimeric sequences) and sequencing errors. In addition to the indispensable inclusion of positive and negative controls [91, 92], various in silico solutions, based on open-source software, are nowadays available to subsequently clean the sequencing data. Various in silico solutions exist to identify chimeric sequences and other PCR artefacts and remove them before the downstream analysis step. UCHIME[83] and Chimera Slayer[78] are two of the most commonly used packages for chimera removal. Decontam is, on the other hand, a popular R package that identifies and corrects for biological contaminants by comparing the frequencies and abundances of bacterial taxa within samples and negative controls [93].

But how to deal with sequencing errors? Or more specifically, how to distinguish them from real sequence variants? Up to now, one solution consists of annotating each sequence using a Bayesian approach. In this case, the algorithm computes the probability that a sequence belongs to an organism, overcoming the problem of sequence errors. One of the most used tools that uses this approach is the RDP classifier[94]. Another solution, and the most commonly used at the moment, is to cluster sequences that are identical to a certain percentage, usually 97% sequence similarity. This process is called Operational Taxonomic Unit (OTU) picking. More recently, alternative solutions to avoid

OTU picking have been proposed in order to identify the exact Sequence Variants.

DADA2 and Deblur[95, 96] use algorithms to model the error rate and denoise the sequence data so that these exact sequence variants (also referred to as sub-OTUs, Amplicon Sequence Variants (ASVs) or zero-radius OTUs (ZOTUS)) can be used.

There is still quite a debate about the use of ASVs as alternative to OTUs. On one side OTUs classifications are biologically useful to compare microbial diversity[97], while on the other side OTUs remain features that emerge from a specific data set (de novo OTUs) or reference database (closed-reference OTUs). This characteristic makes OTUs incomparably between different data sets or, at the very least, OTU-picking leads to the loss of the biological variation of the sequences. ASVs are claimed to overcome this problem by inferring finer sequence resolution. Moreover, ASVs have consistent labels, representing the DNA sequence of the organism, allowing the direct comparison of ASVs across different datasets. However, despite the clear advantages that ASVs offer, the question whether ASV-based approaches outperform OTU-clustering remains. It has been argued that biological trends might be obscured since existing sequencing technologies are often not sufficiently accurate to resolve exact sequences. Moreover, increasing (alpha-)diversity and inter-sample variation may actually complicate downstream statistical analyses. Many recent studies have, however, now been conducted focusing on this comparison and show that ASVs have a sensitivity and specificity as good or even better than OTUs[96, 98-101].

Data processing

Data generated with high throughput sequencing (HTS) show some features that must be taken into account when performing data analysis. In this section we will explore those features and describe how to account for them.

Microbiota data are constrained to a constant sum due to the maximum output of reads delivered by the sequencing machine, and it is also well known that, for each sample, the total count of reads (read depth) can differ drastically. The former feature causes the data to be compositional, while the latter is a major confounder known as library size effect. Finally, microbiota data tend to be enriched in zero counts making the data sparse.

If not addressed properly, those features can lead to biased or even incorrect results when traditional statistics is applied. For example, compositional data (CoDa) provide information only about the relative abundance of species in relation to each other and not about their absolute abundance[8], moreover the read counts of the taxa are not independent from each other. Thus, not accounting for compositionality can result in spurious correlations[102], while not accounting for sparseness can result in inflations of beta-diversity. Until recently, is has been common practice to convert the read counts into relative abundances using numerous techniques. However, none of those approaches corrects for compositionality, leading some researchers to conclude that many studies may suffer of many false positive inferences [103, 104].

During the past 30 years the pioneer work from Aitchison had an enormous influence on CoDa analysis, on the other hand his work was based on relatively simple datasets that are far away from resembling the complexity of microbiota data[8]. Luckily, metagenomics data analysis is starting to be examined by different research groups and now several tools are available to address various research questions.

The majority of microbiota studies have 3 major research goals:

- Identify clusters of samples linked to a certain phenotype.
- Identify taxonomic or structural differences among study groups.
- Find correlations (co-abundance) between microbial taxa.

Usually, studies achieve those objectives using analysis like Principal Coordinate Analysis (commonly based on UniFrac distance or Bray-Curtis dissimilarity), differential abundance testing (e.g., by using simple univariate non-parametric tests or more sophisticated alternatives such as Linear discriminant analysis Effect Size (LefSe)) and Spearman's correlation coefficient, respectively. Unfortunately, those methods often do not completely fulfil the conditions required for the analysis of compositional data[105, 106], such as:

- Permutation invariance: the order of the variables should not influence the results.
- Scale invariance: multiplication of a composition by a positive constant must not change the information in the composition.
- Subcompositional coherence: any subset of the data should have distances between samples and variances that are equal to or less than those found in the full composition.

Identify clusters of samples

Principal Coordinate Analysis (PCoA) is a typical analysis that is commonly performed to investigate the structure of the data for clusters of samples having common features. Due to its easiness and straightforward visual interpretation PCoA is a valuable tool especially as explorative analysis. A drawback related with PCoA is that the original relationship among the features, i.e., the effect of each individual taxon on the observed variance, is lost. Another limitation is related to the distance metric (or dissimilarity index) that is required. The most commonly used indices are the UniFrac, Bray-Curtis and the Jensen-Shannon divergence. Although some of those metrics (e.g., UniFrac) still capture important phylogenetic information, once again they do not account for the compositionality and can be affected by the sparseness.

The CoDa version of the PCoA is a principal component analysis (PCA) on centred log ratio (clr) transformed data. Because of the clr transformation, the underlying relationships between the components are maintained and because it is a PCA the analysis focuses on the variation in the ratio between the parts on an absolute scale. The main problem with this approach arises from the sparseness of the data. Due to the unfeasibility of taking the logarithm of zero, a common, and very questionable, solution is to add a pseudo-count to the data. Unfortunately, this solution has been proven to have an enormous impact on the results[107, 108]. Dealing with the sparseness of metagenomics CoDa can be very challenging. A zero can arise because an OTU is really absent in the sample (structural zero) or because of undersampling and therefore it didn't reach the detection threshold (sampling zero). A more righteous approach, but more computationally intense, is to model the proportions directly from the count data using the Dirichlet distribution[109].

An application of this approach is ALDEx2, a R package originally designed for RNA

-seq data and now used also for metagenomics[106].

Once the zeros in the data have been eliminated, it is possible to apply a log-ratio transformation to obtain independent components that are equivalent to an Euclidean vector[110, 111].

Identify differentially abundant taxa between groups

Another major goal in metagenomics research is to identify differentially abundant taxa between groups. Standard statistical non-parametric tests (e.g. Kruskal-Wallis, Mann-Whitney-Wilcoxon) are commonly used to identify differentially abundant taxa between two or more groups. An extended version of such class comparison methods that is commonly used is Linear discriminant analysis Effect Size (LEfSe)[112]. LEfSe first identifies features that are statistically different among biological classes using Kruskal-Wallis comparisons and subsequently uses the (unpaired) Wilcoxon rank-sum test to investigate if the observed associations are biologically consistent among subclasses. Finally, Linear Discriminant Analysis (LDA) is used to estimate the effect size of each differentially abundant feature which enables to rank the relevance of different biological aspects. In this way LefSe envisions to move beyond simply identifying potential biomarkers by elucidating biological consistency and revealing features with the largest effect size. However, since it uses proportions as input data this can result in distortion of the data. Within the CoDa framework there are two advisable approaches to investigate microbial community differences between groups. Probably the simplest approach is the ANCOM which assesses statistical significance after additive log-ratio transformation[113]. The second one is ALDEx2 that performs statistical testing after the Dirichlet transformation described previously[109]. Additionally, ALDEx2 reports the effect size estimates.

Find correlations between taxa

Finding correlated taxa is another aspect which is commonly part of microbiome studies. As mentioned before, CoDa may suffer of spurious correlations when traditional methods such as spearman's correlation coefficients or Kendall's τ are being used. Several approaches are available for CoDa analysis. The first one consist of computing the φ statistic[114]. Given two bacterial taxa X and Y this statistic has the advantage that it can be computed directly from relative abundance data after clr transformation. A second approach is to use a Bayesian-like approach such as the one used in SparCC. SparCC estimates the linear Pearson correlations between the log-transformed components and is based on the assumption that the number of different components is large and the true correlation network is sparse[115]. Both approaches account for the sparseness of metagenomics data.

Longitudinal data analysis

In the previous section we discussed techniques and methods used to process and analyse metagenomic data from a cross-sectional perspective, but the gut and its microbiota represent a complex and highly dynamic ecosystem. Even though many studies showed the resilience of the gut microbiota and its stability over time[116-118], the gut microbiota is subject to dramatic shifts due to interventions such as changes in diet or medication use. Those shifts might include a temporary bloom of certain species fol-

Compositionality

In statistics, compositional data are vectors of non-negative values that sum up to a constant. A classic example of compositional data are vectors of probabilities. Other forms of compositional data include proportions (summing up to 1), percentages (summing up to 100) and ppm (summing up to 106). The reason why metagenomic data are compositional relies on the presence of a "constant", represented by the maximum output of reads that a sequencing machine can give. As practical example, a standard illumina sequencer like the MiSeq can generate up to 20M reads per run, this maximum output is the constraint that makes microbiome data compositional. In order to better understand let's imagine having an urn filled with balls of 3 different colors: red, blue and white, and we want to describe how the colored balls are distributed inside the urn. Let's imagine now that we are going to sample our urn using a cup and express the result in the form S= {Red, Blue, White}. Our constraint, in this case, is the maximum number of balls that can fit in our cup, {Red+Blue+White} =10. In this example, the urn represents the gut, the colored balls represent the bacteria, and the capacity of the cup represents the maximum output of the sequencing machine. Once the sample is collected from the urn, we can count how many balls we get per each color, e.g., S= {3/10, 5/10, 2/10}. Now let's imagine that after this first sampling, someone re-arranges the composition of the balls inside the urn. This time the result of our sampling is S= {6/10 ,2/10, 2/10}. We can clearly see that the distribution of the colors inside the urn is changed but we are not able to state if someone added more red balls or removed some blue and white balls from the urn. The only information that we can retrieve from our results is that the red balls now are more prevalent. In other words the data that we are collecting provides information only about relative, and not absolute, values of the components[8]. The key feature driving the compositionality of the data relies on the fixed number of balls that fit in our cup. Having a fixed sampling space causes each element of the urn to compete to fill a slot in the empty cup; therefore, if the number of balls of one color increases, there will be more chance it will fill a slot in the cup. But it is also true that the same effect can be achieved reducing the number of balls of the other colors and thus reducing the chances for a blue or white ball to fill a slot in the cup. Analogously this happens with metagenomic studies in which the bacterial sequences present in the amplicon library compete for available slots in the sequencing machine.

lowed by stabilization of the microbiome into the original or an alternative state[119, 120]. A cross-sectional approach might not capture the temporal fluctuations of the microbial community or poorly describe the effect of a covariate. More importantly, cross-sectional studies are prone to selection bias, confounding factors (e.g., medication use or dietary differences in cases when compared to healthy controls) and are unable to reveal whether the observed microbial perturbations among cases actually preceded the disease onset or are merely a consequence of the disease (e.g., due to inflammatory processes). In prospective cohort studies, the outcome of interest (e.g., disease or disease exacerbation) has not yet manifested in any of the participants at baseline. This provides the advantage that temporal associations between exposures (microbial perturbations) and manifestation of the outcome can be explored. Prospective cohort studies have therefore the potential to provide the strongest scientific evidence of all types of observational study designs. In addition, as prospective studies examine changes in microbial composition over time in association to the manifestation of disease (exacerbation), each individual serves as its own control. This significantly reduces the number of potential confounding factors that could lead to either spurious or undetected associations.

The vast majority of statistical tools and methods available to study the microbiota are based upon cross-sectional study designs, however several approaches are nowadays also available to model the dynamics of bacterial species over time.

As mentioned previously, metagenomics data have characteristics such as noisiness, compositionality, and sparseness. Altogether those characteristics pose a big challenge on modelling the microbial structure. For this reason, when it comes to model the dynamics of the microbial community over time, things become even more complex. That is because time series data add another layer of complexity to the analysis. In time series analysis, data coming from the same subject correlate more than data from different subjects. Moreover time-series data may show cyclic patterns and lagged responses to stimuli[121-123]. While the latter case could be addressed when designing the study, e.g., increasing the frequency of sampling, the former can be more difficult to address, especially in association with the other features of metagenomics data.

In the past, many statistical tools have been proposed to model metagenomics data but, up to date, the majority of them still cannot account for all the problems related with the nature of the data. In this section we will screen the most popular tools available for longitudinal modelling of metagenomics data.

Negative binomial mixture models (NBMM)

The Generalized Linear Mixed Models (GLMM) family is the most common method used in the literature to model microbial data over time. As member of this family, the negative binomial mixture model (NBMM) was originally developed by Zhang et al.[124] and extends the principles of a negative binomial model by accounting for the dynamic time trend and within-subject correlation among repeated samples.

Like other negative binomial models, NBMM is well suited for over-dispersed count data. Moreover, as mixture model it can account for the random noise introduced by technical artifacts during the (pre)processing of the samples. The strengths of NBMM include the handling of count data, addressing for the sample depth and the possibility to account for sample variables such as repeated measurements from the same individ-

ual or other covariates. On the other hand, NBMM models suffer of some drawbacks. NBMM analyse each taxon separately, they cannot fully address the sparseness of less abundant species, and do not take into account the possible interactions among species[124, 125]. In conclusion the NBMM models are useful to analyse abundant bacterial taxa, to identify differentially abundant taxa, and to investigate the longitudinal effect of external factors on bacterial abundance but are not well suited to model low abundant taxa or when the focus lays on the identification of bacterial interactions.

Zero-inflated Beta regression

As pointed out in the previous paragraph, standard negative binomial models cannot perform well in presence of many zeros. Besides this, researchers often use the relative abundance of bacterial taxa. An available model developed by Chen et al.[126] is the Zero-inflated Beta regression (ZIBR). This two-part mixture model is a combination of a logistic regression, meant to model the presence/absence of a taxon, and of a Beta regression to model the non-zero relative abundance of the taxon. Additionally, this model allows covariates to affect both parts of the model i.e., a covariate might affect the presence/absence of a bacterial taxon, it might affect its relative abundance, or it might affect both. Finally, this model can take into account repeated measurement from the same subject and the random source of variability. ZIBR has been shown to outperform classic linear mixed models (LMM) when applied to real data set[126]. In addition, compared to NBMM, ZIBR allows two components to have different subject-specific random effects to allow for possible different dependency structures for the zero and non-zero parts of the data. Like with NBMM, a limitation of ZIBR is that it analyses each taxon separately, which assumes that bacteria do not influence each other. Another drawback, related to the use of the relative abundance data in ZIBR is that the unit sum constrain of the data may lead to dependency of the likelihood ratio statistics among the taxa, which may affect the performance of the FDR correction[126].

Generalized Lotka-Volterra equation

Another popular model that has been suggested is the generalized Lotka-Volterra equation. The original Lotka-Volterra equation (LV) was developed to describe 2-species predator-prey dynamics. Its generalized version (gLV) extends to any number of species and can model changes over time of microbial species modelling interaction with another species such as competition, mutualism, or parasitism. Stein et al.[127] proposed another version of gLV equation that also accounts for external perturbations. Since this model can well describe the interactions among bacterial species and how perturbations can affect the microbial dynamics, gLV and LV-based approaches (MC-TIMME)[128] have been increasingly applied in recent years. In spite of this it remains a complex model and therefore less easy to be used by inexperienced researchers and cannot account for within-subject correlation and the random noise typical of metagenomics longitudinal data.

Aims and objectives

The aim of the present thesis was to examine the role of the microbiota in health and disease with a special focus on methodological issues related to: i. the compositionality, and; ii. the longitudinal analysis of microbiome data.

Chapter 2 presents a study on the use of quantitative microbiome profiling to overcome the compositional structure of microbiome sequencing data by integrating absolute quantification of microbial abundances into the NGS data. Prior studies either used cell-based methods (e.g., flow cytometry) or molecular methods (qPCR) to determine the absolute microbial abundances. However, to what extent different quantification methods generate similar quantitative microbiome profiles had thus far not been explored. In **Chapter 2**, we compared relative microbiome profiling (without incorporation of microbial quantification) to three variations of quantitative microbiome profiling: (1) microbial cell counting using flow cytometry (QMP), (2) counting of microbial cells using flow cytometry combined with Propidium Monoazide pre-treatment of faecal samples before metagenomics DNA isolation in order to only profile the microbial composition of intact cells (QMP-PMA), and (3) molecular based quantification of the microbial load using qPCR targeting the 16S rRNA gene.

Chapter 3 presents how various ecological principles, including dispersal (limitation), neutral processes and environmental filtering contribute to the assembly of microbial communities during early infancy within the context of the LucKi Gut Study. While several previous birth cohort studies have already focused on this topic, these studies collected only one or few samples during a time window of extremely dynamic microbial maturation. As such, the impact of specific (dietary) determinants could not always be disentangled. For this purpose, we collected faecal samples from 98 infants repeatedly at 1-2, 4 and 8 weeks, as well as 4, 5, 6, 9, 11 and 14 months of age. The collection of maternal samples allowed us to examine the sharing of microbes between mother-infant dyads, whereas the detailed collection of metadata enabled us to examine the impact of lifestyle, perinatal factors, health status, medication use and diet on the developing infant gut microbiota.

Chapter 4 presents the microbiota maturation in another birth cohort, the German PAPS study. We collected 1453 stool samples, at 5, 13, 21, and 31 weeks postpartum (infants), and once at school age (6-11 years), from 440 children with a familial predisposition for atopic diseases. Next to studying how various determinants shaped the microbiota composition throughout infancy, the extensive clinical follow-up in this study also allowed us to link microbial maturation and composition to the development of atopic dermatitis and asthma. We applied various models, including *joint modelling*, to link longitudinal microbiota development to the onset of atopic manifestations.

In **Chapter 5**, we examined the composition and resilience of the microbiota in association to new-onset post-infectious Irritable Bowel Syndrome (PI-IBS). Many studies have linked the microbiota composition to (subphenotypes of) IBS, however no study to date has been examining the microbiota prior to the onset of IBS. As previous studies have reported an incidence of post-infectious IBS of approximately 5% among travellers that experienced an episode of traveller's diarrhoea (TD), we had the unique opportunity to examine the microbiota in new-onset IBS within the worldwide largest cohort of 2,001 intercontinental travellers (the COMBAT-study). Out of the total cohort, we select-

ed travellers that developed PI-IBS and age-, gender- and travel destination-matched controls. Microbial profiling of faecal samples collected prior to travel, immediately upon return and 1-month post-travel enabled us to study the baseline microbiota as well as the stability and resilience of the microbiota in association to PI-IBS development.

In **Chapter 6**, we examine the dynamics of the faecal microbiota in Crohn's Disease (CD) patients in relation to the disease course. Numerous studies have shown differences between the intestinal microbiota composition of Inflammatory Bowel Disease (IBD) patients and healthy controls, as well as between IBD patients in remission and patients with active flares. However, longitudinal studies in which the temporal (in)stability of the microbiota is associated to the disease course in these patients are largely lacking. Here we collected faecal samples at two time-points from 15 healthy control individuals, 35 CD patients who were in remission and who maintained remission, and 22 CD patients during remission and also during subsequent exacerbation.

Chapter 7 discusses the results of this thesis and brings us back to the question what is still to be discovered and which methodological challenges still need to be addressed to unravel the role of the microbiome in health and disease.

References

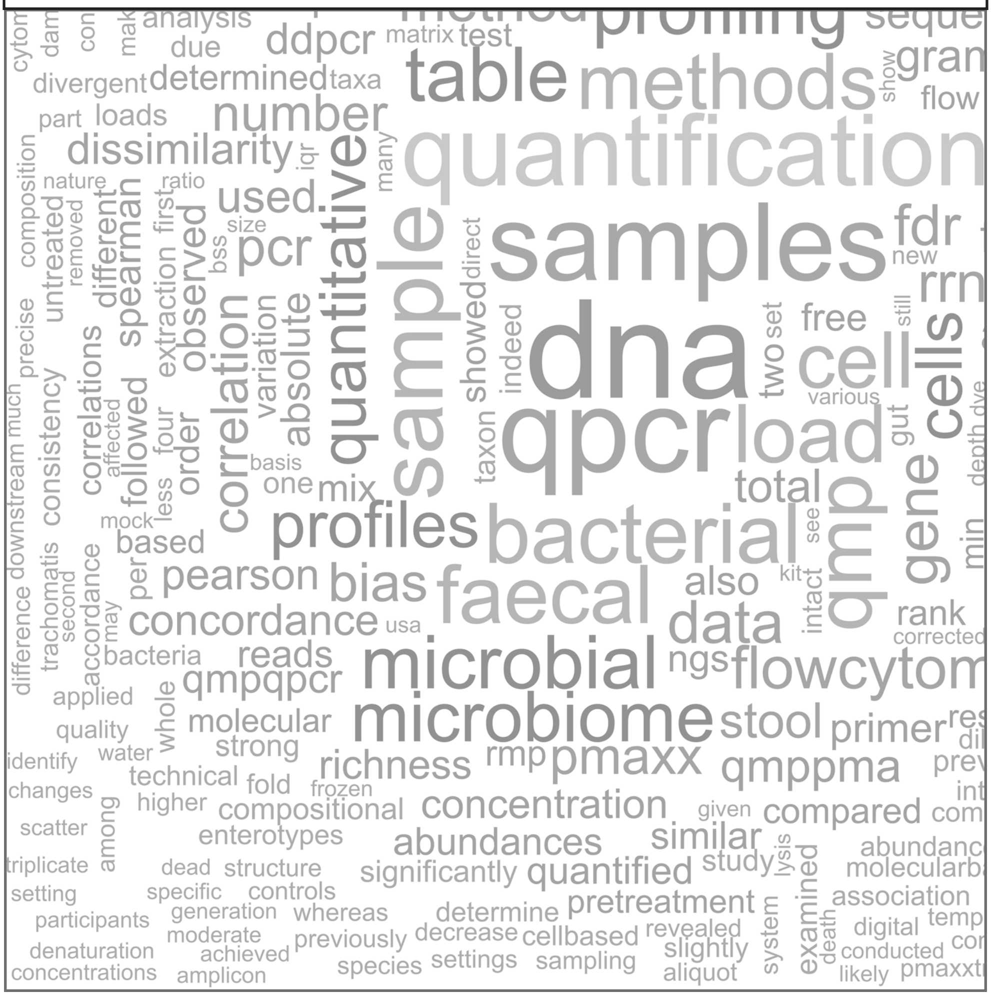
1. Backhed, F., et al., The gut microbiota as an environmental factor that regulates fat storage. Proceedings of the National Academy of Sciences of the United States of America, 2004. 101(44): p. 15718-15723.
2. Duncan, S.H., P. Louis, and H.J. Flint, Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. Applied and Environmental Microbiology, 2004. 70(10): p. 5810-5817.
3. Chen, F.D., W. Wu, and Y.Z. Cong, Short chain fatty acids regulation of neutrophil production of IL-10. Journal of Immunology, 2016. 196.
4. Sartor, R.B., Microbial influences in inflammatory bowel diseases. Gastroenterology, 2008. 134(2): p. 577-594.
5. Hooper, L.V., OPINION Do symbiotic bacteria subvert host immunity? Nature Reviews Microbiology, 2009. 7(5): p. 367-374.
6. Othman, M., R. Aguero, and H.C. Lin, Alterations in intestinal microbial flora and human disease. Current Opinion in Gastroenterology, 2008. 24(1): p. 11-16.
7. Takahashi, K., Interaction between the Intestinal Immune System and Commensal Bacteria and Its Effect on the Regulation of Allergic Reactions. Bioscience Biotechnology and Biochemistry, 2010. 74(4): p. 691-695.
8. Aitchison, J., The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological), 1982. 44(2): p. 139-160.
9. Rawls, J.F., et al., Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell, 2006. 127(2): p. 423-433.
10. 10. Carroll, I.M., et al., Characterization of the faecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. PLoS One, 2012. 7(10): p. e46953.
11. Turnbaugh, P.J. and J.I. Gordon, The core gut microbiome, energy balance and obesity. Journal of Physiology-London, 2009. 587(17): p. 4153-4158.
12. Gilbert, J.A., et al., Current understanding of the human microbiome. Nature Medicine, 2018. 24(4): p. 392-400.
13. Korpela, K., et al., Selective maternal seeding and environment shape the human gut microbiome. Genome Research, 2018. 28(4): p. 561-568.
14. Wampach, L., et al., Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. Nature Communications, 2018. 9.
15. Zhuang, L., et al., Intestinal Microbiota in Early Life and Its Implications on Childhood Health. Genomics Proteomics Bioinformatics, 2019. 17(1): p. 13-25.
16. Stewart, C.J., et al., Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature, 2018. 562(7728): p. 583-588.
17. Ibnou-Zekri, N., et al., Divergent patterns of colonization and immune response elicited from two intestinal Lactobacillus strains that display similar properties in vitro. Infection and Immunity, 2003. 71(1): p. 428-436.
18. Sudo, N., et al., The requirement of intestinal bacterial flora for the development of an IgE production system fully susceptible to oral tolerance induction. Journal of Immunology, 1997. 159(4): p. 1739-1745.
19. Zhong, H.Z., et al., Impact of early events and lifestyle on the gut microbiota and metabolic phenotypes in young school-age children. Microbiome, 2019. 7.
20. Costello, E.K., et al., The application of ecological theory toward an understanding of the human microbiome. Science, 2012. 336(6086): p. 1255-62.
21. Gilbert, J.A. and S.V. Lynch, Community ecology as a framework for human microbiome research. Nat Med, 2019. 25(6): p. 884-889.
22. Donaldson, G.P., S.M. Lee, and S.K. Mazmanian, Gut biogeography of the bacterial microbiota. Nat Rev Microbiol, 2016. 14(1): p. 20-32.
23. Faust, K., et al., Metagenomics meets time series analysis: unraveling microbial community dynamics. Current Opinion in Microbiology, 2015. 25: p. 56-66.
24. van Nimwegen, F.A., et al., Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. Journal of Allergy and Clinical Immunology, 2011. 128(5): p. 948-U371.
25. 25. Bunyavanich, S., et al., Early-life gut microbiome composition and milk allergy resolution. Journal of Allergy and Clinical Immunology, 2016. 138(4): p. 1122-1130.
26. Carlson, A.L., et al., Infant Gut Microbiome Associated With Cognitive Development. Biological Psychiatry, 2018. 83(2): p. 148-159.
27. Gonzalez, M.E., et al., Cutaneous microbiome effects of fluticasone propionate cream and adjunctive bleach baths in childhood atopic dermatitis. Journal of the American Academy of Dermatology, 2016. 75(3): p. 481-+.
28. Byrd, A.L., et al., Staphylococcus aureus and Staphylococcus epidermidis strain diversity underlying pediatric atopic dermatitis. Science Translational Medicine, 2017. 9(397).
29. Liu, C.H., et al., [Allergic airway response associated with the intestinal microflora disruption induced by antibiotic therapy]. Zhonghua Er Ke Za Zhi, 2007. 45(6): p. 450-4.
30. Penders, J., et al., The role of the intestinal microbiota in the development of atopic disorders. Allergy, 2007. 62(11): p. 1223-36.
31. Verhulst, S.L., et al., A Longitudinal Analysis on the Association Between Antibiotic Use, Intestinal Microflora, and Wheezing During the First Year of Life. Journal of Asthma, 2008. 45(9): p. 828-832.
32. 2013. 500(7464): p. 541-+.

33. Sonnenburg, J.L. and F. Backhed, Diet-microbiota interactions as moderators of human metabolism. Nature, 2016. 535(7610): p. 56-64.
34. Zheng, X.J., et al., The modulatory effect of nanocomplexes loaded with EGCG3 '' Me on intestinal microbiota of high fat diet-induced obesity mice model. Journal of Food Biochemistry, 2018. 42(3).
35. Dalby, M.J., et al., Dietary Uncoupling of Gut Microbiota and Energy Harvesting from Obesity and Glucose Tolerance in Mice. Cell Reports, 2017. 21(6): p. 1521-1533.
36. Liu, R.X., et al., Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. Nature Medicine, 2017. 23(7): p. 859-+.
37. Allin, K.H., et al., Aberrant intestinal microbiota in individuals with prediabetes. Diabetologia, 2018. 61(4): p. 810-820.
38. Hanninen, A., et al., Akkermansia muciniphila induces gut microbiota remodelling and controls islet autoimmunity in NOD mice. Gut, 2018. 67(8): p. 1445-1453.
39. Silverman, M., et al., Protective major histocompatibility complex allele prevents type 1 diabetes by shaping the intestinal microbiota early in ontogeny. Proceedings of the National Academy of Sciences of the United States of America, 2017. 114(36): p. 9671-9676.
40. Honda, K. and K. Takeda, Regulatory mechanisms of immune responses to intestinal bacteria. Mucosal Immunology, 2009. 2(3): p. 187-196.
41. Packey, C.D. and R.B. Sartor, Commensal bacteria, traditional and opportunistic pathogens, dysbiosis and bacterial killing in inflammatory bowel diseases. Current Opinion in Infectious Diseases, 2009. 22(3): p. 292-301.
42. Peterson, D.A., et al., Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. Cell Host & Microbe, 2008. 3(6): p. 417-427.
43. Sartor, R.B., Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. Nature Clinical Practice Gastroenterology & Hepatology, 2006. 3(7): p. 390-407.
44. Srikanth, C.V. and B.A. McCormick, Interactions of the intestinal epithelium with the pathogen and the indigenous microbiota: a three-way crosstalk. Interdiscip Perspect Infect Dis, 2008. 2008: p. 626827.
45. Xavier, R.J. and D.K. Podolsky, Unravelling the pathogenesis of inflammatory bowel disease. Nature, 2007. 448(7152): p. 427-34.
46. Biagi, E., et al., Gut Microbiota and Extreme Longevity. Current Biology, 2016. 26(11): p. 1480-1485.
47. Wang, F., et al., Gut Microbiota Community and Its Assembly Associated with Age and Diet in Chinese Centenarians. Journal of Microbiology and Biotechnology, 2015. 25(8): p. 1195-1204.
48. Kong, F.L., et al., Gut microbiota signatures of longevity. Current Biology, 2016. 26(18): p. R832-R833.
49. Yang, Z., et al., Metabolic shifts and structural changes in the gut microbiota upon branched-chain amino acid supplementation in middle-aged mice. Amino Acids, 2016. 48(12): p. 2731-2745.
50. Han, B., et al., Microbial Genetic Composition Tunes Host Longevity (vol 169, pg 1249, 2017). Cell, 2018. 173(4): p. 1058-1058.
51. Debelius, J., et al., Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome Biol, 2016. 17(1): p. 217.
52. Qin, J., et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010. 464(7285): p. 59-65.
53. Santiago, A., et al., Processing faecal samples: a step forward for standards in microbial community analysis. BMC Microbiol, 2014. 14: p. 112.
54. Falony, G., et al., Population-level analysis of gut microbiome variation. Science, 2016. 352(6285): p. 560-4.
55. Zhernakova, A., et al., Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science, 2016. 352(6285): p. 565-9.
56. David, L.A., et al., Diet rapidly and reproducibly alters the human gut microbiome. Nature, 2014. 505(7484): p. 559-63.
57. Kelly, B.J., et al., Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics, 2015. 31(15): p. 2461-8.
58. Osrin, D., et al., Ethical challenges in cluster randomized controlled trials: experiences from public health interventions in Africa and Asia. Bull World Health Organ, 2009. 87(10): p. 772-9.
59. Turnbaugh, P.J., et al., A core gut microbiome in obese and lean twins. Nature, 2009. 457(7228): p. 480-4.
60. Wu, G.D., et al., Linking long-term dietary patterns with gut microbial enterotypes. Science, 2011. 334(6052): p. 105-8.
61. Ley, R.E., et al., Microbial ecology: human gut microbes associated with obesity. Nature, 2006. 444(7122): p. 1022-3.
62. Lozupone, C.A., et al., Diversity, stability and resilience of the human gut microbiota. Nature, 2012. 489(7415): p. 220-30.
63. Carvalho, F.A., et al., Transient inability to manage proteobacteria promotes chronic gut inflammation in TLR5-deficient mice. Cell Host Microbe, 2012. 12(2): p. 139-52.
64. Wu, W.K., et al., Optimization of faecal sample processing for microbiome study - The journey from bathroom to bench. J Formos Med Assoc, 2019. 118(2): p. 545-555.
65. Forslund, K., et al., Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature, 2015. 528(7581): p. 262-266.
66. Integrative, H.M.P.R.N.C., The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe, 2014. 16(3): p. 276-89.
67. Salonen, A., et al., Comparative analysis of faecal DNA extraction methods with phylogenetic microarray:

Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. Journal of Microbiological Methods, 2010. 81(2): p. 127-134.
68. Maukonen, J., C. Simoes, and M. Saarela, The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human faecal samples. Fems Microbiology Ecology, 2012. 79(3): p. 697-708.
69. Kennedy, N.A., et al., The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. Plos One, 2014. 9(2).
70. Smith, B., et al., Optimising bacterial DNA extraction from faecal samples: comparison of three methods. Open Microbiol J, 2011. 5: p. 14-7.
71. Hang, J., et al., 16S rRNA gene pyrosequencing of reference and clinical samples and investigation of the temperature stability of microbiome profiles. Microbiome, 2014. 2.
72. Li, F., M.A.J. Hullar, and J.W. Lampe, Optimization of terminal restriction fragment polymorphism (TRFLP) analysis of human gut microbiota. Journal of Microbiological Methods, 2007. 68(2): p. 303-311.
73. Larsen, A.M., H.H. Mohammed, and C.R. Arias, Comparison of DNA extraction protocols for the analysis of gut microbiota in fishes. Fems Microbiology Letters, 2015. 362(5).
74. Costea, P.I., et al., Towards standards for human faecal sample processing in metagenomic studies. Nat Biotechnol, 2017. 35(11): p. 1069-1076.
75. Sunagawa, S., et al., Metagenomic species profiling using universal phylogenetic marker genes. Nature Methods, 2013. 10(12): p. 1196-+.
76. Kebschull, J.M. and A.M. Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Research, 2015. 43(21).
77. Wang, G.C.Y. and Y. Wang, The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. Microbiology-Sgm, 1996. 142: p. 1107-1114.
78. Haas, B.J., et al., Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research, 2011. 21(3): p. 494-504.
79. McInerney, P., P. Adams, and M.Z. Hadi, Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. Mol Biol Int, 2014. 2014: p. 287430.
80. Suzuki, M.T. and S.J. Giovannoni, Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Applied and Environmental Microbiology, 1996. 62(2): p. 625-630.
81. Brooks, J.P., et al., The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. Bmc Microbiology, 2015. 15.
82. Acinas, S.G., et al., PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Applied and Environmental Microbiology, 2005. 71(12): p. 8966-8969.
83. Edgar, R.C., et al., UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 2011. 27(16): p. 2194-2200.
84. Jousselin, E., et al., Assessment of a 16S rRNA amplicon Illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus. Molecular Ecology Resources, 2016. 16(3): p. 628-640.
85. Jervis-Bardy, J., et al., Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. Microbiome, 2015. 3.
86. Adams, R.I., et al., Microbiota of the indoor environment: a meta-analysis. Microbiome, 2015. 3.
87. Bittinger, K., et al., Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing. Genome Biology, 2014. 15(10).
88. Kunin, V., et al., Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental Microbiology, 2010. 12(1): p. 118-123.
89. Knight, R., et al., Best practices for analysing microbiomes. Nat Rev Microbiol, 2018. 16(7): p. 410-422.
90. Schloss, P.D., The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. Plos Computational Biology, 2010. 6(7).
91. Pollock, J., et al., The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. Applied and Environmental Microbiology, 2018. 84(7).
92. Hornung, B.V.H., R.D. Zwittink, and E.J. Kuijper, Issues and current standards of controls in microbiome research. Fems Microbiology Ecology, 2019. 95(5).
93. Davis, N.M., et al., Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome, 2018. 6.
94. Wang, Q., et al., Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, 2007. 73(16): p. 5261-5267.
95. Callahan, B.J., et al., DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 2016. 13(7): p. 581-+.
96. Amir, A., et al., Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. Msystems, 2017. 2(2).
97. Delgado-Baquerizo, M., et al., A global atlas of the dominant bacteria found in soil. Science, 2018. 359(6373): p. 320-+.
98. Callahan, B.J., P.J. McMurdie, and S.P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. Isme Journal, 2017. 11(12): p. 2639-2643.
99. Eren, A.M., et al., Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. Isme Journal, 2015. 9(4): p. 968-979.
100. Tikhonov, M., R.W. Leach, and N.S. Wingreen, Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. Isme Journal, 2015. 9(1): p. 68-80.
101. Edgar, R.C., UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv,

2016: p. 081257.
102. Pearson, K., Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 1897. 60(359-367): p. 489-498.
103. Jackson, D.A., Compositional data in community ecology: The paradigm or peril of proportions? Ecology, 1997. 78(3): p. 929-940.
104. Faust, K., et al., Microbial Co-occurrence Relationships in the Human Microbiome. Plos Computational Biology, 2012. 8(7).
105. Lozupone, C. and R. Knight, UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology, 2005. 71(12): p. 8228-8235.
106. Fernandes, A.D., et al., Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome, 2014. 2.
107. Feit, M.K., A fair comparison? Geriatric Nursing, 1996. 17(6): p. 260-260.
108. Walker, A., A fair comparison? Reply. Geriatric Nursing, 1996. 17(6): p. 260-260.
109. Fernandes, A.D., et al., ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. Plos One, 2013. 8(7).
110. Egozcue, J.J. and V. Pawlowsky-Glahn, Groups of parts and their balances in compositional data analysis. Mathematical Geology, 2005. 37(7): p. 795-828.
111. Egozcue, J.J., et al., Isometric logratio transformations for compositional data analysis. Mathematical Geology, 2003. 35(3): p. 279-300.
112. Segata, N., et al., Metagenomic biomarker discovery and explanation. Genome Biology, 2011. 12(6).
113. Mandal, S., et al., Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis, 2015. 26: p. 27663.
114. Lovell, D., et al., Proportionality: A Valid Alternative to Correlation for Relative Data. Plos Computational Biology, 2015. 11(3).
115. Friedman, J. and E.J. Alm, Inferring Correlation Networks from Genomic Survey Data. Plos Computational Biology, 2012. 8(9).
116. Relman, D.A., The human microbiome: ecosystem resilience and health. Nutrition Reviews, 2012. 70: p. S2-S9.
117. Caporaso, J.G., et al., Moving pictures of the human microbiome. Genome Biology, 2011. 12(5).
118. Lozupone, C.A., et al., Diversity, stability and resilience of the human gut microbiota. Nature, 2012. 489(7415): p. 220-230.
119. Walker, A.W., et al., Dominant and diet-responsive groups of bacteria within the human colonic microbiota. The ISME journal, 2011. 5(2): p. 220.
120. Dethlefsen, L. and D.A. Relman, Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. Proc Natl Acad Sci U S A, 2011. 108 Suppl 1: p. 4554-61.
121. Fuhrman, J.A., et al., Annually reoccurring bacterial communities are predictable from ocean conditions. Proceedings of the National Academy of Sciences of the United States of America, 2006. 103(35): p. 13104-13109.
122. Dakos, V., et al., Interannual variability in species composition explained as seasonally entrained chaos. Proceedings of the Royal Society B-Biological Sciences, 2009. 276(1669): p. 2871-2880.
123. Giovannoni, S.J. and K.L. Vergin, Seasonality in Ocean Microbial Communities. Science, 2012. 335(6069): p. 671-676.
124. Zhang, X.Y., et al., Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. Frontiers in Microbiology, 2018. 9.
125. Martin, T.G., et al., Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters, 2005. 8(11): p. 1235-1246.
126. Chen, E.Z. and H.Z. Li, A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics, 2016. 32(17): p. 2611-2617.
127. Stein, R.R., et al., Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. Plos Computational Biology, 2013. 9(12).
128. Gerber, G.K., A.B. Onderdonk, and L. Bry, Inferring Dynamic Signatures of Microbes in Complex Host Ecosystems. Plos Computational Biology, 2012. 8(8).

Chapter 2



analysis
due
ddpcr
matrix
test
profiling
divergent
determined
taxa
table
methods
gram
flow
part
loads
number
dissimilarity
quantification
nature
ratio
used
samples
fdr
new
pcr
size
dna
qpcr
sample
quantitative
free
cell
cells
load
various
total
bacterial
profiles
faecal
microbial
microbiome
stool
primer
pmaxx
qmppma
richness
concentration
compared
abundances
similar
significantly
quantified
pretreatment
association
molecular
strong
reads
qmpqpcr
concordance
pearson
bias
data
also
rank
corrected
ngs
flowcytom
technical
compositional
enterotypes
structure
controls
whereas
previously
species
settings
sampling
aliquot
cellbased
slightly
revealed
determine
decrease
moderate
achieved
amplicon
denaturation
concentrations
participants
setting
triplicate
scatter
changes
identify
quality
applied
water
higher
fold
frozen
dead
specific
generation
bacteria
based
mock
one
mix
basis
usa
digital
conducted
likely
study
given
correlation
absolute
variation
spearman
observed
extraction
different
correlations
consistency
followed
order
downstream
trachomatis
accordance
difference
composition
untreated
precise

How to count our microbes? The effect of different Quantitative Microbiome Profiling approaches

Gianluca Galazzo*, Niels van Best*, Birke J. Benedikter, Kevin Janssen, Liene Bervoets, Christel Driessen, Melissa Oomen, Mayk Lucchesi, Pascalle H. van Eijck, Heike Becker, Mathias W. Hornef, Paul H. Savelkoul, Frank R.M. Stassen, Petra F. Wolffs, John Penders

* Shared first authorship

Front. Cell. Infect. Microbiol. 2020 August; 10:403

Abstract

Next-generation sequencing (NGS) has instigated the research on the role of the microbiome in health and disease. The compositional nature of such microbiome datasets makes it however challenging to identify those microbial taxa that are truly associated with an intervention or health outcome. Quantitative microbiome profiling overcomes the compositional structure of microbiome sequencing data by integrating absolute quantification of microbial abundances into the NGS data. Both cell-based methods (e.g., flow cytometry) and molecular methods (qPCR) have been used to determine the absolute microbial abundances, but to what extent different quantification methods generate similar quantitative microbiome profiles has so far not been explored. Here we compared relative microbiome profiling (without incorporation of microbial quantification) to three variations of quantitative microbiome profiling: 1) microbial cell counting using flow cytometry (QMP); 2) counting of microbial cells using flow cytometry combined with Propidium Monoazide pre-treatment of faecal samples before metagenomics DNA isolation in order to only profile the microbial composition of intact cells (QMP-PMA), and; 3) molecular based quantification of the microbial load using qPCR targeting the 16*S* rRNA gene.

Although qPCR and flow cytometry both resulted in accurate and strongly correlated results when quantifying the bacterial abundance of a mock community of bacterial cells, the two methods resulted in highly divergent quantitative microbial profiles when analysing the microbial composition of faecal samples from 16 healthy volunteers. These differences could not be attributed to the presence of free extracellular prokaryotic DNA in the faecal samples as sample pre-treatment with Propidium Monoazide did not improve the concordance between qPCR-based and flow cytometry-based QMP. Also lack of precision of qPCR was ruled out as a major cause of the discordant findings, since quantification of the faecal microbial load by the highly sensitive digital droplet PCR correlated strongly with qPCR.

In conclusion, quantitative microbiome profiling is an elegant approach to bypass the compositional nature of microbiome NGS data, however it is important to realize that technical sources of variability may introduce substantial additional bias depending on the quantification method being used.

Introduction

Next-generation sequencing (NGS) has instigated microbiome research and resulted in many novel insights on the role of the microbiome in health and disease. One of the challenges of NGS however relates to the compositional nature of the generated data. As compositional data always sum up to a constant (e.g., 100%), an increase of a specific microbial taxon in response to a given condition will inevitably lead to a decrease in the relative abundance of other taxa. This mutual dependence between microbial taxa when expressed as relative abundances makes it particularly challenging to identify those microbial taxa that are truly affected by an intervention or a disease state.[1, 2]

Vandeputte *et al.* introduced the concept of Quantitative Microbiome Profiling (QMP) as a way to quantify absolute microbial abundances from NGS data to bypass many of the statistical and interpretative challenges that arise from the compositional structure of microbiome sequencing data.[3] In their work, QMP was achieved by determining the total bacterial load of stool samples by flow-cytometry and subsequently normalizing the 16S rRNA gene sequencing data for sampling depth taking the total bacterial cell counts into account. In contrast, Jian *et al.* used quantitative PCR (qPCR) as a simple and cost-effective alternative to determine the bacterial load and estimate the absolute taxon abundance from NGS data.[1]

Both cell-counting and qPCR come with their advantages and limitations which can impact the subsequent estimation of absolute taxon abundances. Flow-cytometry counts only intact microbial cells. Therefore, new bias could theoretically be introduced when samples contain a significant amount of free extracellular prokaryotic DNA. This free DNA is captured during sequencing but is excluded during flow-cytometry cell counting. In case the taxonomic composition of free circulating DNA differs from the composition of intact microbial cells (e.g., due to differences in the resistance of microbial cells to environmental stress), this might result in the introduction of a new source of bias in downstream analysis. Enumerating bacteria on the basis of qPCR would introduce biases through the extraction, purification, and amplification of DNA. Although, one could argue that this also applies to the NGS data and as such could be considered an advantage of qPCR-based quantification.[1] Advantages of qPCR-based quantification are the cost-effectiveness, simplicity and accessibility, whereas the sensitivity might be a limitation as qPCR has been reported to be only sensitive enough to detect twofold changes in gene concentration or microbial load.[4]

Although Vandeputte *et al.*[3] showed only a moderate correlation between quantification of microbial load by flow-cytometry and qPCR, a direct comparison between cell-based and molecular-based methods to estimate absolute taxon abundances from NGS data has not yet been conducted. As such the level of potential bias that could additionally be introduced when applying quantitative microbial profiling remains unknown.

Here we explored both cell-based and molecular-based methods for QMP and examined the potential effect of various sources of bias by analysing the faecal microbial profiles of 16 healthy volunteers.

First, we compared the estimation of absolute microbial taxon abundances by combining 16S rRNA gene amplicon profiling with respectively flow-cytometry and qPCR to determine the microbial load.

Second, we examined to what extend extracellular DNA derived from lysed bacteria

might introduce differences between cell-based and molecular-based QMP approaches by eliminating free DNA and non-viable cells from stool samples using Propidium Monoazide (PMAxx™, Biotium, Fremont, CA, USA) treatment. Last, we compared the (lack of) sensitivity of qPCR-based methods for microbial quantification to digital droplet PCR [5] as a more precise, discriminating and reproducible molecular quantification method [6, 7].

Materials and Methods

Study population

To assess the impact of different quantitative microbial profiling methods, we collected faecal samples from 16 healthy volunteers. Along with sample collection, a limited number of demographic data were retrieved, including date and time of faecal collection, age, sex, dietary lifestyle, and antibiotic consumption in the previous 3 months (Table 1).

Participants were instructed to collect a complete defecation in a FecesCatcher (Tag Hemi VOF, Zeijen, The Netherlands), transfer a maximum amount of feces in a labelled faeces tube (Sarstedt, Nümbrecht, Germany) and deposit the sample and accompanying questionnaire in a sealed plastic safety bag at the research department as soon as possible. All samples were aliquoted (200 mg aliquots) and stored at -80°C by the researchers.

Cell counts and stool moisture

For cell counting, 200 mg aliquots of the samples were processed and stained as described by Vandeputte *et al.*[3] followed by flow cytometric analysis using a BD FACSCanto II with FACS Diva V8.0.1 software (BD Biosciences). A side scatter of 2000 was set as acquisition threshold. All other instrument and gating settings were in accordance with the method described by Vandeputte *et al.* [3] and were kept constant for all samples. To obtain bacterial concentrations, the total number of events in the cell gate was divided by the sample volume, which was determined by weighing each tube before and after acquisition.

Stool moisture content was determined in duplicate on 200 mg homogenized faecal material as the percentage of mass loss upon vacuum concentration for 5 hrs. at 60°C in a Vacufuge plus (Eppendorf) using the 'AQ' setting.

PMAxx treatment

For the QMP-PMA approach, extracellular DNA and DNA from dead or membrane-compromised bacterial cells was removed by pre-treatment of faecal samples with the viability dye PMAxx™. PMAxx is a DNA-intercalating agent that forms photo-induced crosslinks making the bound DNA inaccessible for downstream molecular applications. PMAxx was added to tenfold diluted faecal specimens at a final concentration of 50 μM, followed by 10 min. shaded incubation at 4°C. Photoactivation was performed by using the PMA-Lite™ LED Photolysis Device (Biotium) with the exposure time set to 10 min. This procedure was repeated 3 times after which metagenomic DNA was isolated from the samples.

In order to assess the effectiveness of PMA-treatment, three faecal samples were

spiked with $3.7x10^7$ copies/gram feces of heat-killed Chlamydia trachomatis (CT). Subsequently, samples were split in two aliquots of which one aliquot was treated with PMAxx as described above and one aliquot remained untreated. Upon DNA isolation (see below), the CT load was quantified by subjecting the treated and untreated samples to a qPCR assay targeting the single-copy *ompA* gene, coding for the major outer membrane protein (MOMP) of Chlamydia trachomatis, on a 7900HT Real-Time PCR System (Applied Biosystems, Foster City, California) as described previously.[8]

DNA isolation and qPCR assessment of bacterial load

DNA was extracted from 200 mg of frozen aliquots of homogenized feces according to the recommended protocol Q of the International Human Microbiome Standards Consortium.[9]

Extracted DNA was quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific).

Enumeration of total bacterial load by qPCR was achieved by amplification of the 16S rRNA genes (primer pair 16S-341_F and 16S-805_R; CCTACGGGNGGCWGCAG and GACTACHVGGGTATCTAATCC, respectively) using a MyiQ Single-Color Real-Time PCR Detection System (BioRad) in 25 µl reactions containing 12.5 µl iQ SYBR Green Supermix (BioRad), 2 µl template DNA (1:1000 diluted), 300 nM of both primers 16S-341_F and 16S-805_R. The PCR amplification program consisted of an initial denaturation set at 95°C for 3 min. followed by 35 three-step cycles at 95°C for 15 s and at 55°C for 20 s and 72°C for 30 s. In each run, negative template controls (DNA replaced by nuclease-free water in qPCR), negative isolation controls (feces replaced by nuclease free water during DNA extraction) and positive controls (quantified recombinant plasmid construct containing the target sequence) were included. Melting curves were checked for each sample to confirm amplification of the correct product.

Digital droplet PCR for assessment of bacterial load

Next to molecular quantification by qPCR, all samples were also quantified by ddPCR by amplifying the 16S rRNA gene (primer pair 515F/806R [10]) using a QX200 Droplet Digital PCR system (Bio-rad). Reaction mixtures consisting of 11µl EvaGreen ddPCR Supermix (Bio-Rad), 2.2µl template DNA and 300 nM of both primers in 22µl reaction volumes were prepared and 20µl was transferred to the DG8 droplet generator cartridge. Upon the addition of 70 µl Droplet Generation Oil in the dedicated wells, the cartridge was placed in the QX200 droplet generator. After droplets have been generated, 40µl was transferred to a 96-wells PCR plate and the plate was sealed using a PX1 PCR plate sealer. The PCR amplification program consisted of an initial denaturation set at 95°C for 3 min. followed by 30 three-step cycles at 95°C for 30 s and at 50°C for 45 s and 72°C for 1 min and finally followed by post-cycling steps of 98 °C for 10 min (enzyme inactivation) and an infinite 12°C hold. The plate was subsequently placed in a QX200 droplet reader and results were analysed using the Quantasoft application.

Comparison of cell-based and molecular-based quantification of a standard microbial community

We used the Gut Microbiome Whole cell Mix (ATCC® MSA-2006™) containing an even mixture of whole bacterial cells (twelve different species) in order to assess whether cell-based or molecular-based quantification was more accurate. The lyo-

philized pellet was dissolved in 1 ml PBS according to the manufacturer's instructions and serial 2-fold dilutions, ranging from $3.3x10^6$ – $5.56x10^4$, were subsequently made. The dilutions were used for cell counting as well as for DNA-isolation followed by qPCR as described above. For qPCR, the number of copies/ml were converted into cells/ml by taking into account the copy numbers for each of the bacterial species in the mock community (average copy number 6.435/genome).

Microbiota profiling

Faecal microbiota profiling was performed in accordance with the paper by Vandeputte *et al.*[3]. Briefly, the V4 region of the 16S rRNA gene was PCR amplified from each DNA sample in triplicate using the 515F/806R primer pair described previously.[10] Pooled amplicons from the triplicate reactions were purified using AMPure XP purification (Agencourt) according to the manufacturer's instructions and eluted in 25 µl 1× low TE (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0). Quantification of amplicons was subsequently performed by the Quant-iT PicoGreen dsDNA reagent kit (Invitrogen) using a Victor3 Multilabel Counter (Perkin Elmer, Waltham, USA). Amplicons were mixed in equimolar concentrations to ensure equal representation of each sample and sequenced on an Illumina MiSeq instrument (MiSeq Reagent Kit v3, 2x250 cycles, 10% PhiX) to generate paired-end reads of 250 bases in length in both directions. After demultiplexing using MiSeq reporter software using default settings, fastq sequences were merged, quality and chimera filtered using FLASH [11], seqtk trimq (https://github.com/lh3/seqtk) and usearch[12], respectively, using the same settings as Vandeputte *et al.*[3].

Finally, between 153,527 and 282,297 reads per untreated sample and between 152,968 and 268,362 reads per PMAxx-treated samples remained for downstream analysis.

Relative Microbiome profiling (RMP)

Samples were downsized to 153,527 reads/sample by randomly selecting reads. Taxonomic assignment of reads was performed using RDP classifier 2.12[13].

Cell-based Quantitative Microbiome Profiling (QMP)

QMP was done in accordance with the method proposed by Vandeputte *et al.*, downsizing the samples to an even sampling depth, defined as the ratio between sample size (16S rRNA gene copy-number-corrected sequencing depth) and microbial load (average total cell count/gram frozen feces; Table S3).

PMA-based Quantitative Microbiome Profiling (QMP-PMA)

Quantitative microbiome profiling after removal of extracellular DNA and DNA from dead and damaged bacterial cells was conducted identical to the standard QMP method with the exception of the additional PMA pre-treatment prior to metagenomic DNA-isolation.

qPCR-based Quantitative Microbiome Profiling (QMP-qPCR)

The bacterial load was determined by qPCR targeting the 16S rRNA gene. Comparing cycle threshold values of each sample to a standard quantification curve (using quantified recombinant plasmid constructs) resulted in the total number of 16S rRNA gene copies/gram feces (Table S4). In order to use the qPCR-based determination of

bacterial load, total numbers of 16S rRNA gene copies/gram feces were converted into the total number of bacterial cells/gram feces. First, the average number of 16S rRNA gene copies per bacterium was calculated for each sample based upon the sequencing data (total number of sequencing reads for a given sample divided by the copy-number corrected number of reads for that respective sample). Next, the total number of 16S rRNA gene copies/gram feces as determined by qPCR was divided by the average 16S rRNA gene copy number of that respective sample. Subsequently, the same approach as for the standard QMP method was followed.

Statistical analyses

No sample size calculations were performed. Statistical analyses were performed in R using the packages *vegan* [14] and *DirichletMultinomial* [15]. Two sided statistical tests were used for all comparisons and corrected for multiple testing using the false discovery rate (FDR according to Benjamini-Hochberg method [16]) where appropriate. *Observed genus richness* was calculated using the R package vegan and *enterotyping* using the DMM approach was performed in R as described previously.[17] As the DMM-clustering was based on a limited number of samples, hence having potentially a limited accuracy, we examined whether the clustering was in concordance with the classification according to the reference-based enterotype classification model fitted on MetaHIT samples (enterotypes.org).[18] DMM-based clustering and reference-based classification was in accordance for all samples, with the exception of three samples that were classified into the Firmicutes enterotype according to the classification model. To calculate microbiome variation between replicates and methods, the *Bray-Curtis dissimilarity* based on the genus-abundance matrix was calculated and visualized by PCoA using the vegan package.

Pearson's or, where appropriate, non-parametric Spearman's correlations were calculated to determine the association between continuous variables (genus richness and abundances, bacterial load and/or metadata).

Paired Wilcoxon Signed Rank test was used to test for differences in observed genus richness between profiling methods and to test for differences in microbiome variation between replicates and profiling methods. The Mann-Whitney U-test was used to test for differences in bacterial loads between enterotypes.

To calculate the ordinal association between genera in the four different profiling methods, the Kendall rank correlation coefficient was used to test for concordance of ranking for the 15 most abundant genera (based upon RMP) between the methods.

Results and discussion

To examine the correlation between cell-counting and molecular quantification, we first compared the quantification of the serially diluted Gut Microbiome Whole cell Mix by means of flow cytometric and qPCR. Cell counting resulted in a concentration very similar to the expected concentration as provided by the manufacturer as quantified by a cellometer (Table S1). Although quantification of serial 2-fold dilutions of the cell mix by FACS and qPCR correlated very strongly (Pearson's $r = -0.967$, $P = 1.7 \times 10^{-8}$, Table S1), qPCR resulted in a much higher than expected concentration (1.0×10^8 cells/ml). Since qPCR also detects extracellular DNA, while cell-counting only quantifies intact mi-

crobial cells, we next removed extracellular DNA by pre-treatment of the suspended Gut Microbiome cell Mix with the viability dye PMAxx™ prior to metagenomic DNA isolation. After PMAxx™ pre-treatment, the number of bacterial cells in the mix as quantified by qPCR was 3.57 x 10^6 cells/ml, and thereby almost identical to the expected concentration. Also, after PMAxx™ pre-treatment, the correlation between cell counting and qPCR of serially diluted Gut Microbiome cell Mix remained very strong (Pearson's r = -0.966, P = 2.1 x 10^{-8}, Table S1). Altogether these results indicate that flow cytometry-based cell counting, and qPCR-based quantification correlated strongly, but absolute quantification might differ substantially in the presence of large quantities of extracellular DNA.

Next, we profiled the microbiota of faecal samples of the 16 healthy volunteers in duplicate for each of the four methods: i. Relative Microbial Profiling (RMP); ii. Quantitative Microbial Profiling using flow cytometry-based microbial load (QMP); iii. Quantitative Microbial profiling using flow-cytometry-based microbial load and PMAxx™ pre-treatment before metagenomics DNA isolation and 16S rRNA gene amplicon sequencing in order to only profile the microbial composition of intact cells (QMP-PMA), and; iv. Quantitative Microbial Profiling using qPCR to determine the microbial load (QMP-qPCR) (see Methods for details and Table 1 for study-specific data).

Stool moisture negatively correlated with observed richness (Spearman's ρ = -0.685, FDR = 9.0 x 10^{-3}, Table S2) confirming previous observations between stool consistency and microbial richness.[3, 19, 20] A similar correlation with microbial richness was not observed when using the Bristol Stool Scale (BSS) as a measure for stool consistency (Spearman's ρ = -0.15, FDR = 5.9 x 10^{-1}). Indeed, BSS scores only weakly and non-significantly correlated with stool moisture (Spearman's ρ = 0.27, FDR = 4.6 x 10^{-1}). This lack of correlation is likely the result of the potential bias introduced by the self-reporting of BSS scores by the study participants, advocating the standardized scoring of stool consistency by research staff or using more objective markers such as stool moisture[21] .

Microbial loads as assessed by flow-cytometry were shown to vary between 1.2 x 10^{10} and 5.3 x 10^{10} cell counts per gram of faecal material (median 2.3 x 10^{10} cell counts per gram; Table S3) and a comparison with qPCR enumeration revealed a moderate correlation (Pearson's r = -0.50, P = 4.7 x 10^{-2}, Table S3, Fig. S1) similar to what has been described by Vandeputte *et al.*[3]

Using DMM clustering on RMP profiles, we identified two enterotypes enriched in *Bacteroides* or *Prevotella* (Fig. S2). The microbial loads, as determined by flow-cytometry, significantly differed between the two enterotypes (median 1.91 x 10^{10} and 2.43 x 10^{10} cells/gram, respectively, P = 4.4 x 10^{-2}). A similar difference between enterotypes was, however, absent when the microbial loads were determined by qPCR (P = 6.0 x 10^{-1}; Table S4).

Prior to comparing the QMP- and QMP-PMA data, we examined the efficacy of PMAxx treatment in removing extracellular DNA in a faecal matrix. First, spiking of faecal samples with heat-killed *Chlamydia trachomatis* (CT) showed that PMAxx treatment effectively eliminated free DNA as indicated by a substantial reduction of qPCR detection of the CT-target DNA. (*i.e.,* average increase of 11.6 Ct-values (range: 10.2-12.7) in qPCR which is equivalent to a signal reduction of 99.96%). Second, enumeration of total bacterial load in faecal samples by qPCR revealed an average decrease in bacterial load of 1.5 x 10^{10} 16S rRNA gene copies/gram feces [IQR: 5.1 x 10^9 – 2.7 x 10^{10}, P = 5.2 x 10^{-4}, Fig. S3] upon PMAxx treatment corresponding to an average of 39.0% of metagenomic

DNA being extracellular or originating from non-viable cells however, the correlation between microbial loads as assessed by flow-cytometry and

qPCR appeared to be slightly weaker after PMAxx-treatment (Pearson's r = -0.41, P = 1.1 x 10^{-1}, Table S3, Fig. S1).

Generating quantitative microbiome profiles revealed that profiles obtained after PMAxx-treatment remained highly similar to the standard QMP profiles (Fig. 1b and c), although the observed genus richness slightly decreased upon PMAxx-treatment (median richness 66.0 and 64.0 for QMP and QMP-PMA, respectively, FDR = 4.0 x 10^{-3}, Table S2). Determination of bacterial load by qPCR, however, resulted in highly divergent profiles (Fig. 1d) and a strong decrease in the observed genus richness (median: 52.0, FDR = 1.2 x 10^{-3}) when compared to QMP and QMP-PMA.

We subsequently analysed the divergence in microbial community structure both between replicates of samples analysed by the same QMP method (within-method dissimilarity) as well as between aliquots of the same sample but profiled by different quantitative methods (between-methods dissimilarity). The within-method variation, as indicated by the average Bray Curtis (BC) dissimilarity, was similar for QMP with and without PMAxx-treatment (Fig. 2, Table S5, FDR = 5.62 x 10^{-1}), whereas the within-method variation was slightly higher for QMP-qPCR when compared to the standard QMP method (FDR = 9.66 x 10^{-4}). Although the between QMP and QMP-PMA method dissimilarity was significantly larger than the within QMP-method dissimilarity (FDR =1.44 x 10^{-3}), the dissimilarity in microbial community structure between both methods was still modest (median [IQR] BC dissimilarity: 0.082 [0.062-0.108]) and far lower than the dissimilarity between QMP-qPCR and QMP (median [IQR]: 0.260 [0.199-0.364], FDR = 1.83 x 10^{-4}). From these results it cannot yet be deduced whether the slightly yet significantly dissimilar QMP-PMA and QMP microbial profiles are due to the elimination of free extracellular DNA (bias in QMP) or merely due to the introduction of additional technical variation during sample handling (noise).

We therefore subsequently examined to what extent the sample rank order for each genus was conserved between the four profiling methods, similar to Vandeputte *et al.* When comparing RMP to QMP, sample rank order concordance within the 15 most abundant genera varied widely with the highest concordance observed for *Fusicatenibacter* and the lowest concordance for *Blautia* (Kendall's rank correlation test, τ, range = 0.47–0.95, Table S6). This confirmed the previous observation that absolute abundance profiles differ significantly from those generated by relative approaches.

When comparing the average sample rank concordance among the 15 most abundant genera between each of the four profiling methods, QMP and QMP-PMA showed the highest overall concordance (average τ among the 15 most abundant genera = 0.82, Table S6, Fig. S4). The overall concordance between RMP and either QMP or QMP-PMA did not differ significantly (average τ among 15 most abundant genera = 0.75 and 0.69, respectively, FDR = 3.8 x 10^{-1}, Table S6, Fig. S4), indicating that PMAxx-treatment did not appear to result in a higher overall concordance with RMP. For each of the 15 genera the lowest sample rank order concordance was observed between QMP-qPCR and the other three methods, confirming that qPCR-based absolute abundance profiles are highly divergent from both the other quantitative as well as the relative profiling methods.

Furthermore, we could clearly identify the strong trade-off between *Bacteroides* and *Prevotella* as commonly reported [22] in RMP-based analysis (Spearman's $\rho = -0.70$,

FDR = 3.2 x 10^{-5}) and confirmed that the association between these two genera became weaker in a quantitative context, although the association remained statistically significant in the QMP and QMP-PMA profiles (QMP: Spearman's $\rho = -0.64$, FDR = 1.7 x 10-4; QMP-PMA: Spearman's $\rho = -0.55$, FDR = 1.4 x 10-3; QMP-qPCR: Spearman's $\rho = -0.17$, FDR = 0.353; Table S7).

To explore the possibility that the deviant profiles generated by QMP-qPCR are the result of the lack of precision and sensitivity of qPCR-based quantification, we finally quantified the microbial load in all faecal samples by means of Droplet Digital PCR (ddPCR). As with qPCR, this more recently introduced technology uses Taq polymerase in a standard PCR reaction to amplify the target DNA. The ddPCR technology however partitions the PCR reaction into thousands of droplets (individual reaction vessels) prior to amplification and acquires the data at the reaction end point. This enables more precise and reproducible data and direct quantification without the need of standard curves.[6, 7] Quantification of microbial load based upon ddPCR however correlated strongly with qPCR-based quantification both for untreated (Pearson's r = 0.72, P = 2.0 x 10^{-3}) and PMAxx-treated faecal samples (Pearson's r = 0.90, P = 2.0 x 10^{-6}, Table S8). More importantly, correlations between ddPCR and FACS for untreated (Pearson's r = 0.50, P = 4.9 x 10^{-3}) and PMAxx-treated faecal samples (Pearson's r = 0.39, P = 1.4 x 10^{-1}, Table S8) were not stronger than correlations between qPCR and FACS (Table S3). Indeed, when quantifying serial 2-fold dilutions of 3 samples and mock mix (within the concentration range of ~10^2-10^5 copies/uL), we showed that qPCR and ddPCR results correlated strongly (Pearson's r = 0.988, P = 5.6 x 10^{-46} Table S9, Fig. S5).

Altogether these results indicate that the deviant QMP-qPCR based profiles when compared to the other profiling methods cannot be explained by a lack of precision or sensitivity of qPCR.

This indicates that extracellular DNA does not seem to introduce a new source of bias when combining 16S NGS with flow-cytometry cell counts. It should, however, be noted that a previous study did report markedly distinct faecal microbial profiles of extremely preterm infants upon PMA-treatment.[23] The rapid processing and storage of faecal samples in the present study might have contributed to the limited differences, underscoring the importance of careful sample handling.

The results of our analysis further demonstrate that quantification of bacterial load by qPCR results in highly divergent profiles, indicating that qPCR-based quantification might not be an adequate approach for quantitative microbiome profiling. Flow-cytometry quantification indicated that the difference in bacterial load varied less than 3 times between the vast majority of samples (14/16). Several studies have indicated that qPCR is only useful for determining dissimilarity between two samples if the true difference is at least 2-3 fold [4, 24], suggesting that qPCR-based enumeration is too imprecise to be an adequate alternative for flow-cytometry in quantitative microbiome profiling. However, we showed that using the highly precise and sensitive ddPCR for microbial quantification did not result in improved correlation with flow cytometry-based cell counting. The strong correlation between ddPCR and qPCR moreover makes PCR bias an unlikely cause of the divergent profiles as different primer pairs were used for the two molecular quantification methods. Indeed, *in silico* analyses showed that the primer pairs used for qPCR quantification are highly specific for the domains of archaea and bacteria. Less than 0.1% of eukaryotic sequences are detected while over 95% of

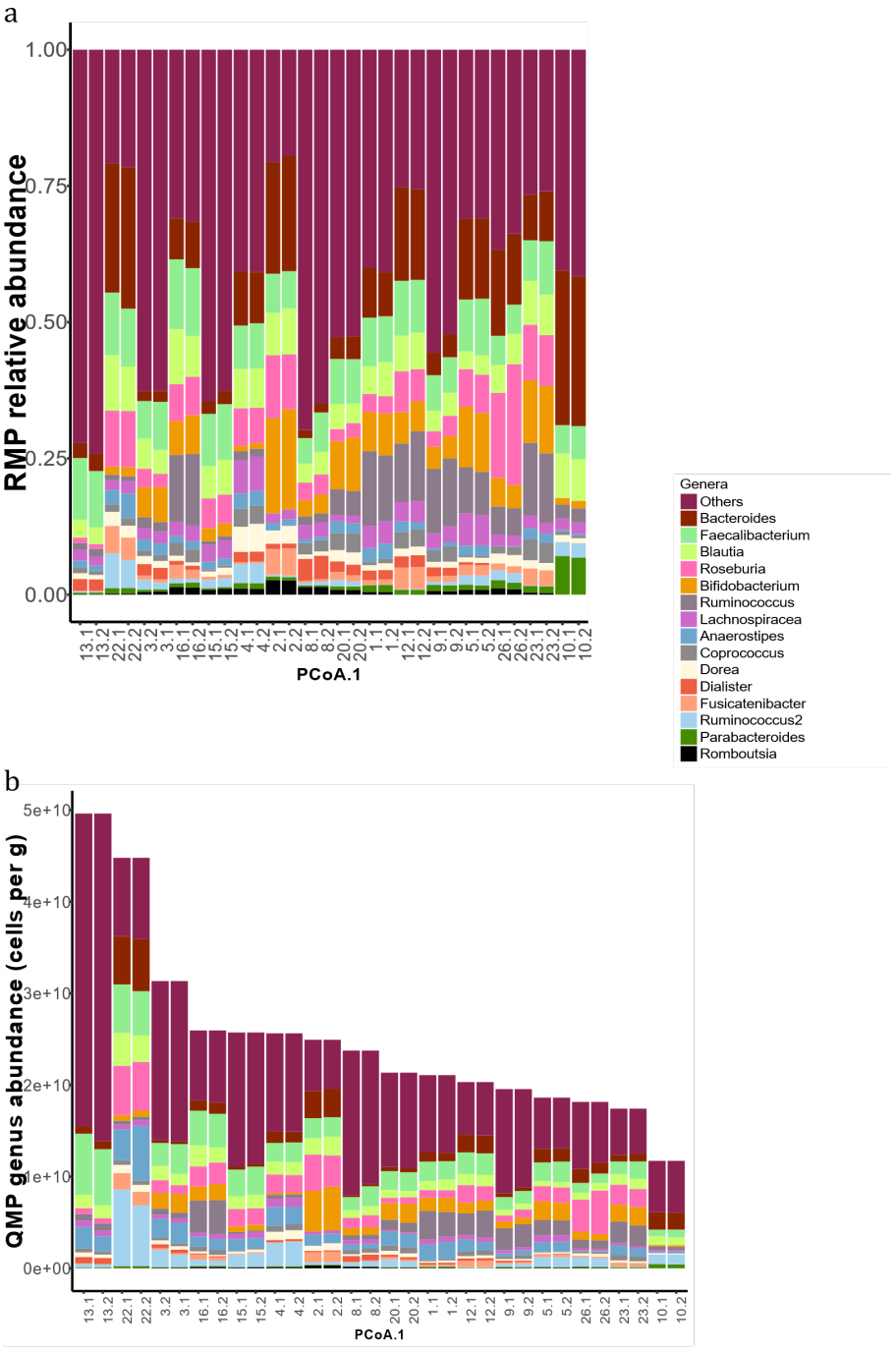
all bacterial sequences are being detected. Although only 65% of all archaeal sequences match our primer pair, this is mainly due to mismatches to many environmental archaea whereas the methanogenic archeal species commonly observed in the human intestinal tract are all covered by our primer pair.

When using Gut Microbiome Whole Cell Mix, we did show strong correlations between flow-cytometry and qPCR-based quantification. Moreover, ddPCR and qPCR quantification of 16*S* rRNA gene copies in faecal samples also correlated strongly despite the use of different primer pairs and amplification protocols. Together these results indicate that primer bias or other technical aspects related to qPCR-based quantification are an unlikely cause for the dramatic deviant QMP-qPCR profiles. It is much more likely that the bias is introduced during the process of extracting DNA from the complex faecal matrix. In contrast to cell counting, molecular quantification is a multi-step process on a small aliquot of the original faecal sample, which might result in increased intra-sample variation when performed on multiple aliquots. Indeed, the standard deviation between (some) replicates was substantially larger when using qPCR as compared to flow cytometry. This is in line with a recently published method to decompose spatiotemporal variance on microbial communities, which confirmed substantial heterogeneity between spatial sampling locations of faecal samples.[25]. Also, incomplete lysis and DNA fragmentation can bias results during DNA extraction, however the protocol used in the present study has been comprehensively optimized to maximize DNA quality and quantity and benchmarked to limit bias in community diversity and Gram-positive to Gram-negative ratio. [9] Moreover, the DNA extraction might also become saturated which even further hampers direct correlation between DNA yield and microbial load in the original sample. These limitations may also impact the use of alternative methods for quantitative profiling such as spiking in reference DNA as an internal standard to extrapolate the amount of starting nucleic material.[26, 27]

A previous study did report near perfect correlations between QMP-qPCR and absolute abundances as determined by various taxon-specific qPCRs [1]. However as both methods were applied on the same DNA sample this further suggests that the bias is not due to the qPCR-based approach itself but rather the lack of correlation between yield upon DNA extraction and the microbial load in the original faecal sample.

Flow-cytometry, being executed on the original sample, performed better in terms of intra-sample variations and showed stronger correlations with RMP and stool consistency. This suggests that flow-cytometry would be a more preferable method to quantify bacterial load in feces, however also flow-cytometry comes with several limitations. One such limitation is cell aggregation which can result in underestimation of cell counts.[28, 29] Moreover, this method is more laborious, expensive and requires technical expertise (e.g. quality control and monitoring size-related resolution, setting-up reproducible scatter detection, measuring in accurate concentration range), which makes it less suitable as a standard method that can be applied by all labs on a high-throughput basis. This calls for more high-throughput and user-friendly cell-counting methods.

Alternatively, computational solutions are now becoming available to make stable inferences of changes in abundances in compositional data such as the application of "reference frames"[26]. Such computational solutions should however always be accompanied by careful controlling for important confounding factors, in particular stool consistency.


a
RMP relative abundance
1.00
0.75
0.50
0.25
0.00
13.1
13.2
22.1
22.2
3.2
3.1
16.1
16.2
15.1
15.2
4.1
4.2
2.1
2.2
8.1
8.2
20.1
20.2
1.1
1.2
12.1
12.2
9.1
9.2
5.1
5.2
26.1
26.2
23.1
23.2
10.1
10.2
PCoA.1
Genera
Others
Bacteroides
Faecalibacterium
Blautia
Roseburia
Bifidobacterium
Ruminococcus
Lachnospiracea
Anaerostipes
Coprococcus
Dorea
Dialister
Fusicatenibacter
Ruminococcus2
Parabacteroides
Romboutsia
b
QMP genus abundance (cells per g)
5e+10
4e+10
3e+10
2e+10
1e+10
0e+00
13.1
13.2
22.1
22.2
3.2
3.1
16.1
16.2
15.1
15.2
4.1
4.2
2.1
2.2
8.1
8.2
20.1
20.2
1.1
1.2
12.1
12.2
9.1
9.2
5.1
5.2
26.1
26.2
23.1
23.2
10.1
10.2
PCoA.1

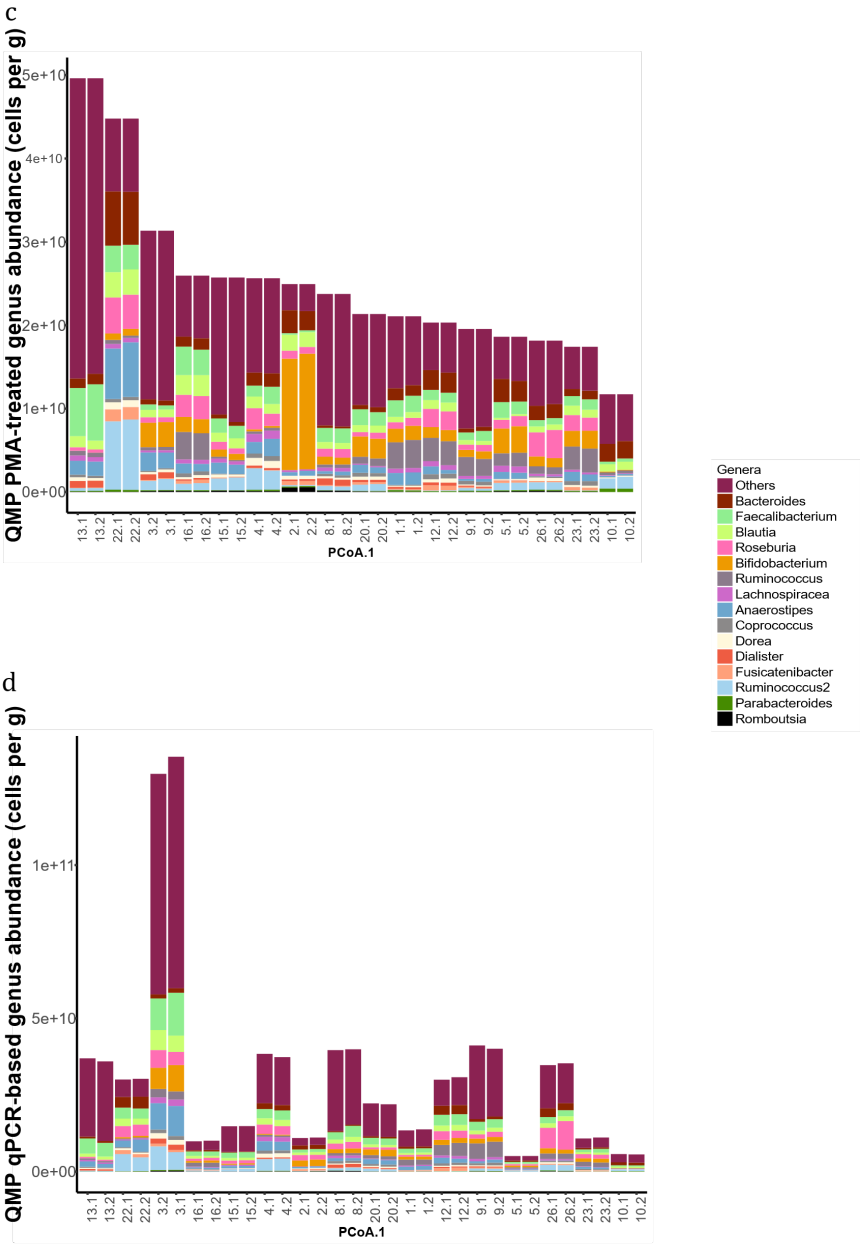


Figure 1 Microbiome profile comparisons. Genus-level faecal microbial composition of both replicates of all 16 healthy study subjects (n = 32 samples) based upon a) relative microbiome profiling (RMP), b) quantitative microbiome profiling (QMP, cells per gram feces), c) QMP after PMAxx-treatment of faecal samples (QMP-PMA, cells per gram feces), d) QMP using qPCR for quantification of bacterial load (QMP-qPCR, cells per gram feces).

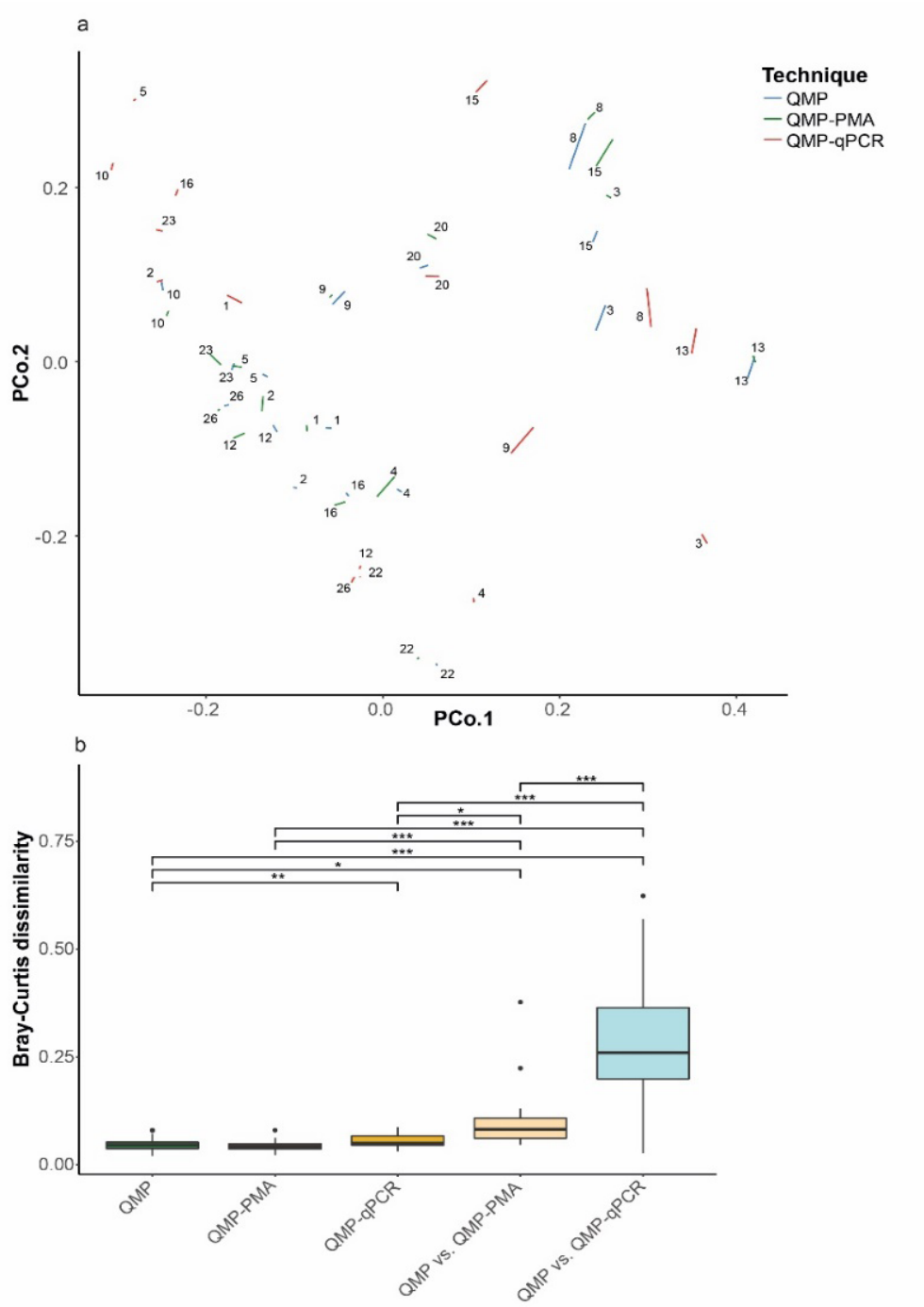

a
Technique
QMP
QMP-PMA
QMP-qPCR
PCo.2
PCo.1
0.2
0.0
-0.2
-0.2
0.0
0.2
0.4
b
Bray-Curtis dissimilarity
0.75
0.50
0.25
0.00
QMP
QMP-PMA
QMP-qPCR
QMP vs. QMP-PMA
QMP vs. QMP-qPCR

Figure 2 Within method dissimilarity of sample replicates and between methods dissimilarity of samples. Faecal microbial community structure variation based upon Bray-Curtis (BC) dissimilarity between samples and sample replicates. a) Principal coordinates analysis of the study cohort based upon BC dissimilarity. Each segment connects the two replicates of the same sample as profiled by QMP (blue), QMP-PMA (green) and QMP-qPCR (red), b) Box-plot of BC distance between sample replicates for all quantitative profiling methods (within-method variability) and BC distance in microbial community structure from the same sample profiled with different quantitative methods (between-method variability). The significance was checked pairwise using the Wilcoxon test and then adjusted for multiple comparisons using the FDR correction. The significance coding is indicated as *** for $p<0.005$, ** for $p<0.01$, * for $p<0.05$ and N.S. for $p > 0.05$. For clarity only significance of the comparisons between within QMP-method dissimilarity and all other within- and between-method dissimilarities are indicated (all FDR-corrected p-values are presented in Supplementary Table 5).

In conclusion, quantitative microbiome profiling is an elegant approach to bypass the compositional nature of microbiome NGS data, however it is important to realize that technical sources of variability may introduce substantial additional bias depending on the quantification method being used.

Table 1 Characteristics of the (faecal samples of) healthy subjects included in the present study.

Subject	Age	Sex	Alternative dietary lifestyle	Antibiotic use in past 3 months	Bristol Stool Score	Average % Dry Weight
1	25	Female	Vegetarian	No	3	26.99
2	24	Male	No	No	3	18.70
3	28	Female	Vegetarian	No	6	23.43
4	23	Female	No	No	4	28.08
5	31	Female	No	No	4	23.34
8	29	Female	No	No	6	18.52
9	27	Female	No	No	3	36.29
10	49	Female	No	No	3	14.73
12	30	Male	No	No	4	26.82
13	27	Male	No	No	4	22.47
15	29	Female	No	No	4	13.00
16	26	Female	No	No	6	16.37
20	26	Male	No	No	3	24.49
22	26	Male	No	No	3	24.58
23	31	Male	No	No	4	29.89
26	31	Male	No	No	4	19.56

Data Availability Statement

The datasets generated for this study can be found in the EBI archive under accession number ERP108719.

Ethics statement

The study design and the experimental procedures were approved by the Medical Ethics Committee of the Maastricht University Medical Center and align with the revised declaration of Helsinki. All subjects gave their written informed consent.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Jian, C., et al., Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. PLoS One, 2020. 15(1): p. e0227285.
2. Knight, R., et al., Best practices for analysing microbiomes. Nat Rev Microbiol, 2018. 16(7): p. 410-422.
3. Vandeputte, D., et al., Quantitative microbiome profiling links gut community variation to microbial load. Nature, 2017. 551(7681): p. 507-511.
4. Smith, C.J. and A.M. Osborn, Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. FEMS Microbiol Ecol, 2009. 67(1): p. 6-20.
5. Hindson, C.M., et al., Absolute quantification by droplet digital PCR versus analog real-time PCR. Nat Methods, 2013. 10(10): p. 1003-5.
6. Gobert, G., et al., Droplet digital PCR improves absolute quantification of viable lactic acid bacteria in faecal samples. J Microbiol Methods, 2018. 148: p. 64-73.
7. Kim, T.G., S.Y. Jeong, and K.S. Cho, Comparison of droplet digital PCR and quantitative real-time PCR for examining population dynamics of bacteria in soil. Appl Microbiol Biotechnol, 2014. 98(13): p. 6105-13.
8. Janssen, K.J., et al., Viability-PCR Shows That NAAT Detects a High Proportion of DNA from Non-Viable Chlamydia trachomatis. PLoS One, 2016. 11(11): p. e0165920.
9. Costea, P.I., et al., Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol, 2017. 35(11): p. 1069-1076.
10. Caporaso, J.G., et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J, 2012. 6(8): p. 1621-4.
11. Magoc, T. and S.L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 2011. 27(21): p. 2957-63.
12. Edgar, R.C., et al., UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 2011. 27(16): p. 2194-200.
13. Wang, Q., et al., Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol, 2007. 73(16): p. 5261-7.
14. Oksanen, J., et al., Vegan: Community Ecology Package. R Package Version. 2.5-3. CRAN. 2013.
15. Morgan, M., DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning for Microbiome Data. R Package Version. 1.18-0. https://cran-r-project.org/package=dirmult. 2017.
16. Benjamini, Y. and Y. Hochberg, Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. J Roy Stat Soc B Met, 1995. 57: p. 289-300.
17. Holmes, I., K. Harris, and C. Quince, Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One, 2012. 7(2): p. e30126.
18. Costea, P.I., et al., Enterotypes in the landscape of gut microbial community composition. Nat Microbiol, 2018. 3(1): p. 8-16.
19. Tigchelaar, E.F., et al., Gut microbiota composition associated with stool consistency. Gut, 2016. 65(3): p. 540-2.
20. Vandeputte, D., et al., Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. Gut, 2016. 65(1): p. 57-62.
21. Vork, L., et al., Stool Consistency: Looking Beyond the Bristol Stool Form Scale. J Neurogastroenterol Motil, 2019. 25(4): p. 625.
22. Lozupone, C.A., et al., Diversity, stability and resilience of the human gut microbiota. Nature, 2012. 489(7415): p. 220-30.
23. Young, G.R., et al., Response: Commentary: Reducing Viability Bias in Analysis of Gut Microbiota in Preterm Infants at Risk of NEC and Sepsis. Front Cell Infect Microbiol, 2018. 8: p. 374.
24. Hospodsky, D., N. Yamamoto, and J. Peccia, Accuracy, precision, and method detection limits of quantitative PCR for airborne bacteria and fungi. Appl Environ Microbiol, 2010. 76(21): p. 7004-12.
25. Ji, B.W., et al., Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. Nat Methods, 2019. 16(8): p. 731-736.
26. Morton, J.T., et al., Establishing microbial composition measurement standards with reference frames. Nat Commun, 2019. 10(1): p. 2719.
27. Tkacz, A., M. Hortala, and P.S. Poole, Absolute quantitation of microbiota abundance in environmental samples. Microbiome, 2018. 6(1): p. 110.
28. Gunasekera, T.S., P.V. Attfield, and D.A. Veal, A flow cytometry method for rapid detection and enumeration of total bacteria in milk. Appl Environ Microbiol, 2000. 66(3): p. 1228-32.
29. Ou, F., et al., Absolute bacterial cell enumeration using flow cytometry. J Appl Microbiol, 2017. 123(2): p. 464-477.

Supplementary tables

Table S1 Quantification of Mock Gut Microbiome Community by FACS and qPCR

Microbial load as determined by FACS

Expected concentration*	FACS cell count/ml replicate #1	FACS cell count/ml replicate #2	Average FACS cell count/ml	S.D.
3.30E+06	2.22E+06	1.56E+06	1.89E+06	4.68E+05
1.65E+06	5.52E+05	5.69E+05	5.61E+05	1.17E+04
8.25E+05	2.25E+05	2.29E+05	2.27E+05	2.83E+03
4.13E+05	1.27E+05	1.23E+05	1.25E+05	2.76E+03
2.06E+05	5.56E+04	5.69E+04	5.63E+04	9.19E+02
1.03E+05	3.10E+04	3.04E+04	3.07E+04	4.24E+02
5.56E+04	1.47E+04	1.78E+04	1.63E+04	2.19E+03

*Based upon information provided by manufacturer

Microbial load as determined by qPCR

Expected concentration*	qPCR Ct replicate #1	qPCR Ct replicate #2	Average Ct	S.D. Ct	cells/ml replicate #1	cells/ml replicate #2	Average cells/ml	S.D. cells/ml
							1.03E+08	4.13E+06
3.30E+06	19.15	19.23	19.19	0.06	1.06E+08	1.00E+08		
1.65E+06	21.13	20.83	20.98	0.21				
8.25E+05	22.11	21.95	22.03	0.11				
4.13E+05	23.53	24.10	23.82	0.40				
2.06E+05	23.04	23.74	23.39	0.49				
1.03E+05	25.65	24.28	24.97	0.97				
5.56E+04	25.70	26.34	26.02	0.45				

*Based upon information provided by manufacturer

Table S1 (cont'd)

Microbial load as determined by qPCR with PMA treatment prior to DNA isolation

Expected concentration*	qPCR-PMA Ct replicate #1	qPCR-PMA Ct replicate #2	Average Ct	S.D.	cells/ml replicate #1	cells/ml replicate #2	Average cells/ml	S.D. cells/ml
3.30E+06	24.38	23.60	23.99	0.55	2.61E+06	4.53E+06	3.57E+06	1.36E+06
1.65E+06	25.80	25.09	25.45	0.50				
8.25E+05	28.11	28.84	28.48	0.52				
4.13E+05	29.04	27.68	28.36	0.96				
2.06E+05	29.62	29.73	29.68	0.08				
1.03E+05	30.71	31.40	31.06	0.49				
5.56E+04	31.35	31.08	31.22	0.19				

*Based upon information provided by manufacturer

Table S1 (cont'd)

Variable 1	Variable 2	N	Test	Effect size	p-value
FACS cell counts	qPCR ct-values untreated faeces	7	Pearson	-0.967	1.65E-08
FACS cell counts	qPCR ct-values PMA-treated faeces	7	Pearson	-0.966	2.05E-08

Table S2 Pairwise correlations between stool moisture content, Bristol Stool Scale scores and observed species richness. Spearman test effect size, p-value, and FDR by Benjamini-Hochberg.

Variable 1	Variable 2	N	Test	Effect size	p-value	FDR
Bristol Stool Score	Stool moisture (%)	16	Spearman	0.274	0.304	0.456
Bristol Stool Score	Observed genus richness in RMP	16	Spearman	-0.147	0.587	0.587
Stool moisture (%)	Observed genus richness in RMP	16	Spearman	-0.685	0.003	0.009

Variable 1	Variable 2	N	Test	Median Variable 1	IQR Variable 1	Median Variable 2	IQR Variable 2	p-value	FDR
Observed genus richness RMP	Observed genus richness QMP	16	Paired Wilcoxon	73.5	[71.25 - 80.75]	66	[62.25 - 71.75]	0.001	0.0012
Observed genus richness RMP	Observed genus richness QMP-PMA	16	Paired Wilcoxon	73.5	[71.25 - 80.75]	64	[55.75 - 69.50]	0.001	0.0012
Observed genus richness RMP	Observed genus richness QMP-qPCR	16	Paired Wilcoxon	73.5	[71.25 - 80.75]	52	[40.50 - 58.50]	0.001	0.0012
Observed genus richness in QMP	Observed genus richness in QMP-PMA	16	Paired Wilcoxon	66	[62.25 - 71.75]	64	[55.75 - 69.50]	0.004	0.004
Observed genus richness in QMP	Observed genus richness in QMP-qPCR	16	Paired Wilcoxon	66	[62.25 - 71.75]	52	[40.50 - 58.50]	0.001	0.0012
Observed genus richness in QMP-PMA	Observed genus richness in QMP-qPCR	16	paired Wilcoxon	64	[55.75 - 69.50]	52	[40.50 - 58.50]	0.001	0.0012

Table S3 Microbial load as determined by FACS analysis and as determined by qPCR of untreated and PMA-treated faecal samples.
Pairwise correlation between microbial load as determined by FACS and qPCR using Pearson test effect size, p-value.

Microbial load as determined by FACS

Sample ID	FACS cell count/g faeces replicate #1	FACS cell count/g faeces replicate #2	Average FACS cell count/g faeces	S.D.
1	2.46E+10	2.52E+10	2.49E+10	4.39E+(
2	9.75E+09	1.37E+10	1.17E+10	2.80E+(
3	1.72E+10	3.47E+10	2.60E+10	1.24E+1
4	2.17E+10	2.10E+10	2.13E+10	5.23E+(
5	3.76E+10	5.20E+10	4.48E+10	1.01E+1
8	1.48E+10	2.00E+10	1,74E+10	3,65E+(
9	1.88E+10	1.76E+10	1,82E+10	8,49E+(
10	2.22E+10	2.00E+10	2,11E+10	1,53E+(
12	3.72E+10	2.55E+10	3,13E+10	8,30E+(
13	2.17E+10	2.96E+10	2,56E+10	5,61E+(
15	2.22E+10	1.51E+10	1,86E+10	5,01E+(
16	2.39E+10	2.36E+10	2,38E+10	2,60E+(
20	2.24E+10	1.67E+10	1,96E+10	4,00E+(
22	2.23E+10	1.84E+10	2,03E+10	2,77E+(
23	5.57E+10	4.96E+10	5,27E+10	4,30E+(
26	2.17E+10	2.98E+10	2,57E+10	5,70E+(
		Median: 2.26E+10 (range 1.17E+10 - 5.27E+10)		

Table S3 (cont'd) Microbial load as determined by FACS analysis and as determined by qPCR of untreated and PMA-treated faecal samples.

Microbial load as determined by qPCR

Sample ID	Ct-value replicate 1	Ct-value replicate 2	Average Ct-value	S.D.	log Copies per gram/faeces	Copies per gram/faeces
1	23.74	24.42	24.08	0.48	10.32	2.07E+10
1-PMA	24.12	24.49	24.31	0.26	10.25	1.77E+10
2	24.89	24.86	24.88	0.02	10.07	1.18E+10
2-PMA	26.59	27.37	26.98	0.55	9.42	2.66E+09
3	24.12	24.31	24.22	0.13	10.28	1.88E+10
3-PMA	26.08	26.73	26.41	0.46	9.60	3.99E+09
4	23.17	23.78	23.48	0.43	10.50	3.18E+10
4-PMA	23.99	24.76	24.38	0.54	10.23	1.68E+10
5	22.70	22.71	22.71	0.01	10.74	5.49E+10
5-PMA	23.90	23.60	23.75	0.21	10.42	2.62E+10
8	23.87	23.79	23.83	0.06	10.39	2.48E+10
8-PMA	23.59	23.73	23.66	0.10	10.45	2.79E+10
9	22.19	22.90	22.55	0.50	10.79	6.15E+10
9-PMA	22.75	22.83	22.79	0.06	10.71	5.17E+10
10	23.79	23.82	23.81	0.02	10.40	2.52E+10
10-PMA	23.92	24.15	24.04	0.16	10.33	2.14E+10
12	22.42	18.62	20.52	2.69	11.41	2.58E+11
12-PMA	23.12	23.37	23.25	0.18	10.57	3.75E+10
13	22.09	22.31	22.20	0.16	10.90	7.86E+10
13-PMA	23.09	23.31	23.20	0.16	10.59	3.87E+10
15	26.88	23.36	25.12	2.49	10.00	9.96E+09
15-PMA	26.30	24.81	25.55	1.05	9.86	7.31E+09
16	22.19	22.26	22.23	0.05	10.89	7.72E+10
16-PMA	22.76	23.26	23.01	0.35	10.65	4.43E+10
20	22.10	22.39	22.25	0.21	10.88	7.61E+10
20-PMA	22.48	22.73	22.61	0.18	10.77	5.90E+10
22	22.97	23.12	23.05	0.11	10.64	4.32E+10
22-PMA	23.85	24.16	24.01	0.22	10.34	2.19E+10

Table S3 (cont'd)

23	22.38	22.40	22.39	0.01	10.84	6.87E+10
23-PMA	22.81	23.10	22.96	0.21	10.66	4.60E+10
26	24.76	22.69	23.73	1.46	10.43	2.67E+10
26-PMA	25.14	23.54	24.34	1.13	10.24	1.73E+10
Positive plasmid control 10^6 copies/µl	17.93	17.82	17.88	0.08	NA	NA
Positive plasmid control 10^4 copies/µl	24.78	24.55	24.67	0.16	NA	NA

Variable 1	Variable 2	N	Test	Effect size	p-value
FACS cell counts (log10)/g faeces	qPCR ct-values untreated faeces	16	Pearson	-0.5	0.047
FACS cell counts (log10)/g faeces	qPCR ct-values PMA-treated faeces	16	Pearson	-0.41	0.11

Table S4 Top 5 genera driving the DMM-based enterotype clustering

Top 5 genera driving DMM-based enterotypes and median microbial load in enterotypes

Genus	Main relative abundance in Enterotype 1 (n =10)	Main relative abundance in Enterotype 2 (n = 6)
Prevotella	0.003976499	0.012462066
Bacteroides	0.145341645	0.083816963
Bifidobacterium	0.053733343	0.041333001
Blautia	0.073070461	0.065577759
Roseburia	0.050321123	0.045548263

Bacterial load	Median bacterial load Enterotype 1	Median bacterial load Enterotype 2	Test	p-value
FACS (cell counts/g faeces)	2.43E+10	1.91E+10	unpaired Wilcoxon	0.044
qPCR (cell counts/g faeces) *	1.84E+10	1.62E+10	unpaired Wilcoxon	0.604
qPCR PMA-treated samples (cell counts/g faeces) *	9.21E+09	1.09E+10	unpaired Wilcoxon	0.526

*cell counts calculated by dividing copy number/g faeces by the average 16S copy number per sample as calculated from 16S NGS data

Table S5 Within and between-method variation in microbial community structure as indicated by Bray-Curtis dissimilarity.

	Median [IQR] BC dissimilarity
Within QMP method dissimilarity	0.04621 [0.03824-0.05242]
Within QMP-PMA method dissimilarity	0.04254 [0.03720-0.04910]
Within QMP-qPCR method dissimilarity	0.05027 [0.04539-0.06720]
Between QMP and QMP-PMA method dissimilarity	0.08274 [0.06203-0.10828]
Between QMP and QMP-qPCR method dissimilarity	0.26020 [0.19934-0.36433]

Variable 1	Variable 2	Test	FDR
Within QMP method dissimilarity	Within QMP-PMA method dissimilarity	paired Wilcoxon signed rank	5.62E-01
Within QMP method dissimilarity	Within QMP-qPCR method dissimilarity	paired Wilcoxon signed rank	9.66E-04
Within QMP-PMA method dissimilarity	Within QMP-qPCR method dissimilarity	paired Wilcoxon signed rank	2.34E-01
Within QMP method dissimilarity	Between QMP and QMP-PMA method dissimilarity	paired Wilcoxon signed rank	1.44E-03
Within QMP method dissimilarity	Between QMP and QMP-qPCR method dissimilarity	paired Wilcoxon signed rank	1.83E-04
Within QMP-PMA method dissimilarity	Between QMP and QMP-PMA method dissimilarity	paired Wilcoxon signed rank	1.83E-04

Table S5 (cont'd)

Within QMP-PMA method dissimilarity	Between QMP and QMP-qPCR method dissimilarity	paired Wilcoxon signed rank	1.83E-04
Within QMP-qPCR method dissimilarity	Between QMP and QMP-PMA method dissimilarity	paired Wilcoxon signed rank	2.67E-03
Within QMP-qPCR method dissimilarity	Between QMP and QMP-qPCR method dissimilarity	paired Wilcoxon signed rank	1.83E-04
Between QMP and QMP-PMA method dissimilarity	Between QMP and QMP-qPCR method dissimilarity	paired Wilcoxon signed rank	1.83E-04

Table S6 Sample rank concordance (Kendall's tau) among the 15 most abundant genera between each of the four profiling methods

Genus	RMP vs. QMP	RMP vs. QMP-PMA	RMP vs. QMP-qPCR	QMP vs. QMP-PMA	QMP vs. QMP-qPCR	QMP-PMA vs. QMP-qPCR
Bacteroides	0.833333333	0.75	0.45	0.883333333	0.45	0.466666667
Faecalibacterium	0.666666667	0.533333333	0.216666667	0.466666667	0.383333333	0.116666667
Blautia	0.466666667	0.55	0.016666667	0.85	0.216666667	0.2
Roseburia	0.783333333	0.716666667	0.433333333	0.766666667	0.45	0.483333333
Bifidobacterium	0.883333333	0.833333333	0.533333333	0.95	0.55	0.566666667
Ruminococcus	0.733333333	0.75	0.583333333	0.916666667	0.583333333	0.533333333
Lachnospiracea incertae sedis	0.533333333	0.55	0.266666667	0.783333333	0.266666667	0.35
Anaerostipes	0.65	0.616666667	0.55	0.833333333	0.666666667	0.566666667
Coprococcus	0.833333333	0.75	0.516666667	0.816666667	0.516666667	0.5
Dorea	0.583333333	0.466666667	0.466666667	0.816666667	0.616666667	0.5
Dialister	0.869267957	0.833333333	0.852980514	0.835509978	0.863846946	0.800757217
Fusicatenibacter	0.95	0.883333333	0.566666667	0.833333333	0.55	0.583333333
Ruminococcus2	0.816666667	0.833333333	0.583333333	0.916666667	0.633333333	0.683333333
Parabacteroides	0.733333333	0.55	0.383333333	0.716666667	0.383333333	0.2
Romboutsia	0.866666667	0.75	0.6	0.85	0.633333333	0.683333333

Table S6 (cont'd)

*paired Wilcoxon signed rank

Variable 1	Variable 2	N	Median [IQR] Variable 1	Median [IQR] Variable 2	Test	FDR3
RMP vs. QMP	RMP vs. QMP-PMA	16	0.747 [0.658-0.85]	0.691 [0.55-0.792]	PWSR*	3.84E-01
RMP vs. QMP	RMP vs. QMP-qPCR	16	0.747 [0.658-0.85]	0.468 [0.408-0.575]	PWSR*	1.34E-03
RMP vs. QMP	QMP vs. QMP-PMA	16	0.747 [0.658-0.85]	0.816 [0.8-0.867]	PWSR*	2.32E-01
RMP vs. QMP	QMP vs. QMP-qPCR	16	0.747 [0.658-0.85]	0.518 [0.417-0.625]	PWSR*	2.24E-03
RMP vs. QMP	QMP-PMA vs. QMP-qPCR	16	0.747 [0.658-0.85]	0.482 [0.408-0.575]	PWSR*	1.84E-03
RMP vs. QMP-PMA	RMP vs. QMP-qPCR	16	0.691 [0.55-0.792]	0.468 [0.408-0.575]	PWSR*	6.29E-03
RMP vs. QMP-PMA	QMP vs. QMP-PMA	16	0.691 [0.55-0.792]	0.816 [0.8-0.867]	PWSR*	1.20E-02
RMP vs. QMP-PMA	QMP vs. QMP-qPCR	16	0.691 [0.55-0.792]	0.518 [0.417-0.625]	PWSR*	1.56E-02
RMP vs. QMP-PMA	QMP-PMA vs. QMP-qPCR	16	0.691 [0.55-0.792]	0.482 [0.408-0.575]	PWSR*	7.70E-03
RMP vs. QMP-qPCR	QMP vs. QMP-PMA	16	0.468 [0.408-0.575]	0.816 [0.8-0.867]	PWSR*	5.65E-04
RMP vs. QMP-qPCR	QMP vs. QMP-qPCR	16	0.468 [0.408-0.575]	0.518 [0.417-0.625]	PWSR*	5.10E-01
RMP vs. QMP-qPCR	QMP-PMA vs. QMP-qPCR	16	0.468 [0.408-0.575]	0.482 [0.408-0.575]	PWSR*	8.68E-01
QMP vs. QMP-PMA	QMP vs. QMP-qPCR	16	0.816 [0.8-0.867]	0.518 [0.417-0.625]	PWSR*	5.65E-04
QMP vs. QMP-PMA	QMP-PMA vs. QMP-qPCR	16	0.816 [0.8-0.867]	0.482 [0.408-0.575]	PWSR*	5.65E-04
QMP vs. QMP-qPCR	QMP-PMA vs. QMP-qPCR	16	0.518 [0.417-0.625]	0.482 [0.408-0.575]	PWSR*	7.43E-01

Table S7 Spearman rank correlations between *Bacteroides* and *Prevotella* in each of the four profiling methods

Variable 1	Variable 2	N	Test	Effect size	p-value	FDR
Bacteroides abundance RMP	Prevotella abundance RMP	32	Spearman	-0.701	0.000008	0.000032
Bacteroides abundance QMP	Prevotella abundance QMP	32	Spearman	-0.639	0.000084	0.000168
Bacteroides abundance QMP-PMA	Prevotella abundance QMP-PMA	32	Spearman	-0.553	0.001038	0.001384
Bacteroides abundance QMP-qPCR	Prevotella abundance QMP-qPCR	32	Spearman	-0.17	0.352664	0.352664

Table S8 Microbial load as determined by ddPCR of untreated and PMA-treated faecal samples.

Pairwise correlation between microbial load as determined by ddPCR and qPCR using Pearson test effect size, p-value.

Microbial load as determined by ddPCR

Sample ID	ddPCR copies per uL DNA replicate #1	ddPCR copies per uL DNA replicate #2	ddPCR average copies per uL DNA	ddPCR SD copies per uL DNA	qPCR copies per uL DNA	Fold-differences qPCR vs ddPCR
1	1.14E+07	1.01E+07	1.08E+07	9.19E+05	2,07E+07	1,93
1-PMA	1.35E+07	1.00E+07	1.18E+07	2.47E+06	1,77E+07	1,50
2	5.12E+06	4.75E+06	4.94E+06	2.62E+05	1,18E+07	2,39
2-PMA	2.25E+06	2.06E+06	2.16E+06	1.34E+05	2,66E+06	1,23
3	1.17E+07	1.02E+07	1.10E+07	1.06E+06	1,88E+07	1,72
3-PMA	5.73E+06	5.18E+06	5.46E+06	3.89E+05	3,99E+06	0,73
4	1.09E+07	9.70E+06	1.03E+07	8.49E+05	3,18E+07	3,09
4-PMA	7.50E+06	7.40E+06	7.45E+06	7.07E+04	1,68E+07	2,26
5	1.63E+07	1.85E+07	1.74E+07	1.56E+06	5,49E+07	3,16
5-PMA	1.21E+07	1.17E+07	1.19E+07	2.83E+05	2,62E+07	2,20
8	1.01E+07	1.02E+07	1.02E+07	7.07E+04	2,48E+07	2,44
8-PMA	1.57E+07	1.60E+07	1.59E+07	2.12E+05	2,79E+07	1,76
9	2.41E+07	2.21E+07	2.31E+07	1.41E+06	6,15E+07	2,66
9-PMA	1.97E+07	1.83E+07	1.90E+07	9.90E+05	5,17E+07	2,72
10	8.89E+06	8.98E+06	8.94E+06	6.36E+04	2,52E+07	2,82
10-PMA	8.80E+06	8.45E+06	8.63E+06	2.47E+05	2,14E+07	2,48

Table S8 (cont'd)

12	2.23E+07	2.34E+07	2.29E+07	7.78E+05	2,58E+08	11,31
12-PMA	1.27E+07	1.30E+07	1.29E+07	2.12E+05	3,75E+07	2,92
13	2.66E+07	2.74E+07	2.70E+07	5.66E+05	7,86E+07	2,91
13-PMA	1.96E+07	2.09E+07	2.03E+07	9.19E+05	3,87E+07	1,91
15	1.58E+07	1.75E+07	1.67E+07	1.20E+06	9,96E+06	0,60
15-PMA	1.03E+07	9.60E+06	9.95E+06	4.95E+05	7,31E+06	0,73
16	2.32E+07	2.27E+07	2.30E+07	3.54E+05	7,72E+07	3,36
16-PMA	1.67E+07	1.78E+07	1.73E+07	7.78E+05	4,43E+07	2,57
20	2.72E+07	2.28E+07	2.50E+07	3.11E+06	7,61E+07	3,04
20-PMA	2.63E+07	2.39E+07	2.51E+07	1.70E+06	5,90E+07	2,35
22	1.12E+07	1.11E+07	1.12E+07	7.07E+04	4,32E+07	3,87
22-PMA	8.30E+06	8.30E+06	8.30E+06	0.00E+00	2,19E+07	2,64
23	2.10E+07	1.91E+07	2.01E+07	1.34E+06	6,87E+07	3,43
23-PMA	1.61E+07	1.50E+07	1.56E+07	7.78E+05	4,60E+07	2,96
26	1.71E+07	1.65E+07	1.68E+07	4.24E+05	2,67E+07	1,59
26-PMA	1.19E+07	1.18E+07	1.19E+07	7.07E+04	1,73E+07	1,46

MEDIAN FOLD-DIFFERENCE 2.46 [IQR 1.72 - 2.91]

Variable 1	Variable 2	N	Test	Effect size	p-value
qPCR ct-values untreated faeces	ddPCR copies per uL DNA	16	Pearson	0.72	2.00E-03

Table S8 (cont'd)

qPCR ct-value PMA-treated faeces	ddPCR copies per uL DNA from PMA-treated faeces	16	Pearson	0.90	2.00E-06
FACS cell counts (log10)/g faeces	ddPCR copies per uL DNA	16	Pearson	0.5	4.90E-02
FACS cell counts (log10)/g faeces	ddPCR copies per uL DNA from PMA-treated faeces	16	Pearson	0.39	1.40E-01

Table S9 Microbial load of serial 2-fold diluted mock and samples as determined by ddPCR and qPCR.

Pairwise correlation between microbial load as determined by ddPCR and qPCR using Pearson test effect size, p-value.

Sample	Dilution factor	ddPCR 1 log10 copies/uL	ddPCR 2 log10 copies/uL	qPCR 1 log10 copies/uL	qPCR 2 log10 copies/uL
Mock	1	4.687528961	4.673941999	4.790915246	4.932479842
Mock	2	4.425371166	4.378397901	4.569335877	4.633963193
Mock	4	4.146128036	4.107888025	4.313904105	4.387763895
Mock	8	3.859138297	3.810904281	3.944605158	3.981535053
Mock	16	3.489958479	3.462397998	3.609158614	3.615313596
Mock	32	3.243038049	3.184691431	3.356804333	3.230627193
Mock	64	2.890421019	2.888179494	2.987505386	2.876715701
sample A	1	4.179551791	4.250663919	4.707822983	4.643195667
sample A	2	3.911157609	3.892651034	4.24927679	4.356988983
sample A	4	3.604226053	3.670245853	3.969225088	4.046162368

Table S9 (cont'd)

sample A	8	3.267171728	3.352182518	3.587616175	3.636856035
sample A	16	3	3.096910013	3.187542316	3.316796947
sample A	32	2.685741739	2.665580991	2.929033052	2.676678771
sample A	64	2.340444115	2.369215857	2.402782052	2.519726719
sample B	1	3.292256071	3.334453751	3.036745245	3.169077368
sample B	2	3.025305865	3.025305865	2.876715701	2.882870684
sample B	4	2.64246452	2.792391689	2.538191666	2.682833754
sample B	8	2.369215857	2.451786436	2.156582754	2.208900105
sample B	16	2.201397124	2.08278537	2.073490491	2.036560596
sample B	32	1.698970004	1.838849091	1.642641718	1.818058719
sample B	64	1.707570176	1.653212514	1.44568228	1.648796701
sample C	1	3.809559715	3.84135947	3.932295193	4.040007386
sample C	2	3.503790683	3.5132176	3.661475965	3.627623561
sample C	4	3.178976947	3.281033367	3.258324614	3.470671509
sample C	8	2.912753304	2.943494516	2.925955561	2.981350403
sample C	16	2.562292864	2.670245853	2.578199052	2.670523789
sample C	32	2.201397124	2.264817823	2.055025543	2.091955438
sample C	64	1.886490725	1.963787827	1.851911122	1.836523666

Table S9 (cont'd)

Pearson correlation serial dilutions	ddPCR	qPCR
Mock	0.865852181	0.847761759
Sample A	0.864580823	0.859651092
Sample B	0.888031531	0.892300398
Sample C	0.862322445	0.853170557

Pearson correlation ddPCR vs qPCR (Fig S5)	**p-value**
0.988483687	5.60E-46

Supplementary Figures

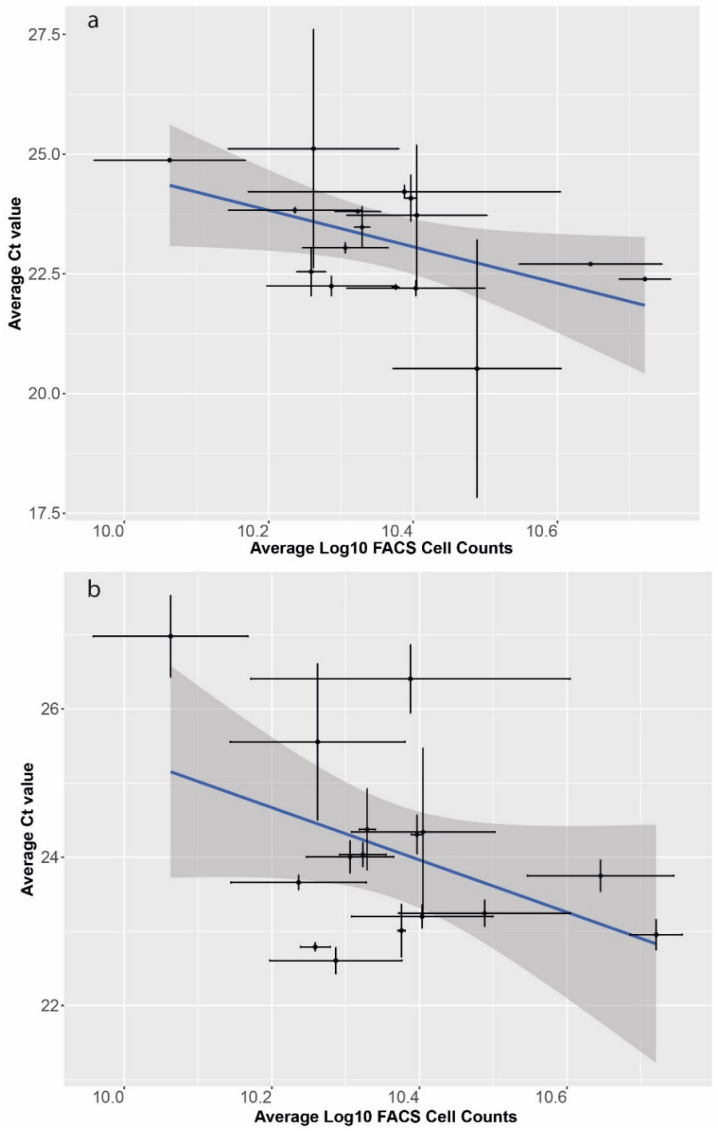


Figure S1 Quantification of microbial load by flow-cytometry and qPCR. (a) Correlation between microbial loads as assessed by flow cytometry (log10 cell counts/gram faeces) and bacterial abundance as assessed by qPCR (n = 16 healthy subjects, Pearson's r = -0.50, P = 4.7 x 10-2), (b) Correlation between microbial loads as assessed by flow cytometry (log10 cell counts/gram faeces) and bacterial abundance as assessed by qPCR (n = 16 healthy subjects) after treatment of faecal samples with PMAxx to remove non-viable cells and extracellular DNA (Pearson's r = -0.41, P = 1.1 x 10-1). Data points represent median values of replicate samples, error bars represent the standard deviation.

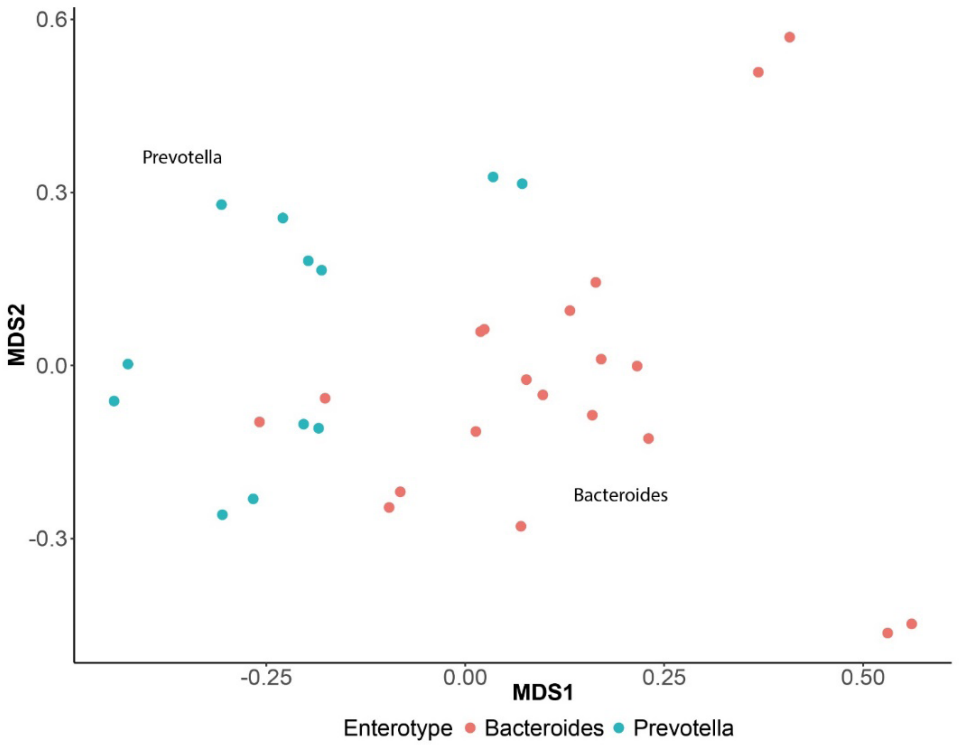


Figure S2 Genus-level faecal microbial community variation based upon Bray-Curtis dissimilarity and represented by non-metric Multidimensional Scaling. Both replicates from all 16 healthy individuals were included (n = 32), enterotyped based upon Dirichlet Multinomial Mixtures (DMM) and coloured accordingly.

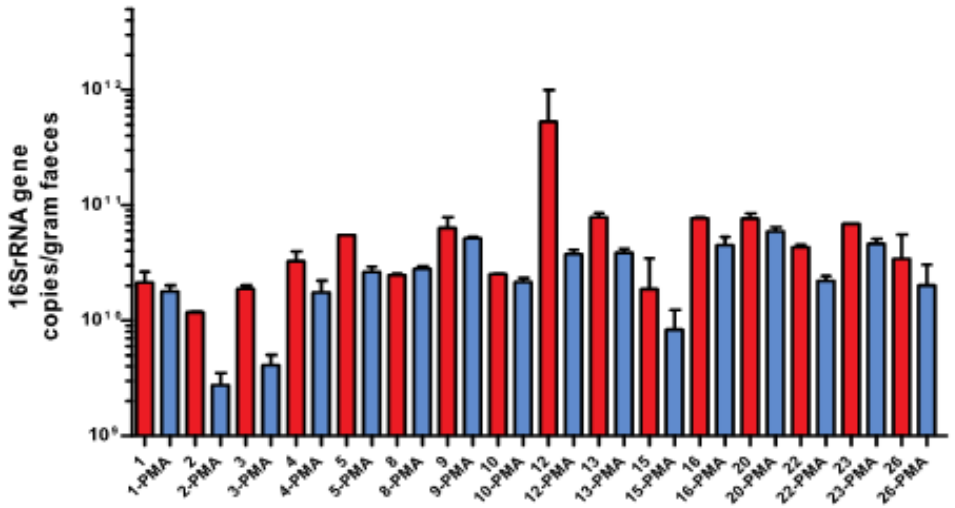


Figure S3 Average number of 16S rRNA gene copies per gram faeces for each sample before (red bars) and after (blue bars) treatment with PMAxx. Error bars indicate standard deviation based upon sample replicates.

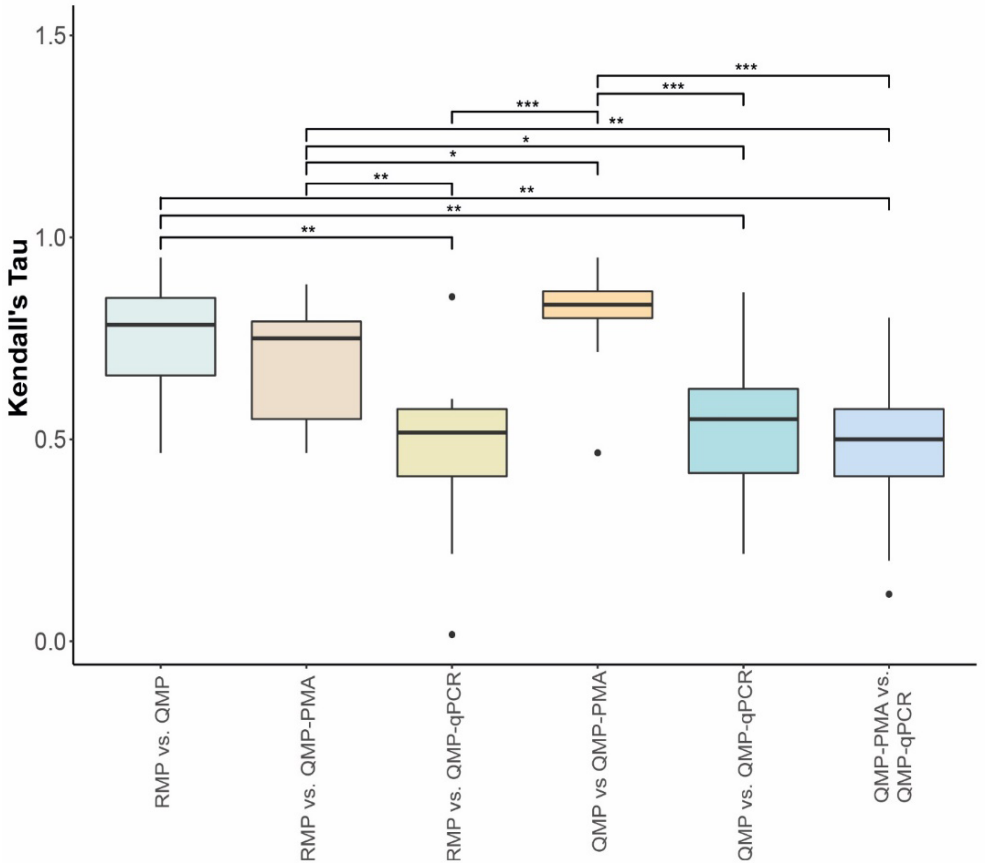


Figure S4 Boxplot based on the average Kendall's tau concordance between profiling methods among the ranked abundance of the 15 most abundant genera. The significance was checked pairwise using the Wilcoxon test and then adjusted for multiple comparisons using the FDR correction. The significance coding is indicated as *** for p<0.005, ** for p<0.01, and * for p<0. 05.

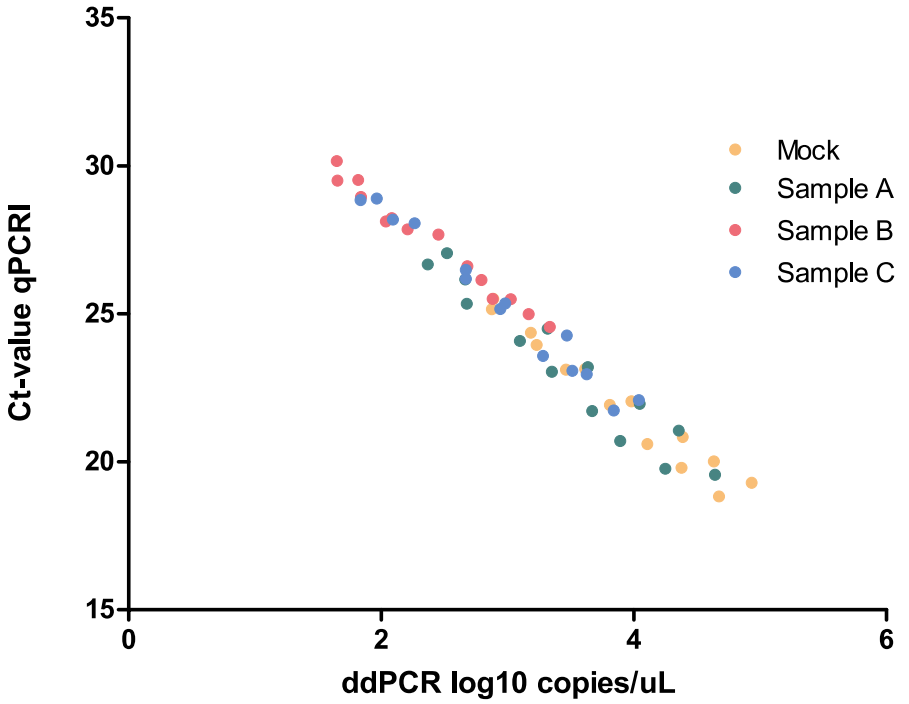


Figure S5 Scatter plot based on the quantification of serial 2-fold diluted mock and faecal samples by ddPCR as compared to qPCR. (Pearson's r = -0.988, P = 5.6 x 10-46)

Chapter 3


months
age
infants
microbiota
community
microbial
structure
dietary
diversity
samples
postpartum
maternal
associated
mothers
impact
cluster
neutral
composition
processes
asvs
infant
delivered
also
diet
gut
still
bacteroides
delivery
maturation
studies
birth
infancy
clusters
patterns
analysis
metacommunity
revealed
supplementary
determinants
feeding
vaginal
collected
assembly
first
food
dmm
high
used
foods
within
breastfeeding
data
adults
children
sprockett
relative
early
items
specific
similar
medication
effects
local
total
ages
largely

Assembly, structure, and dynamics of the infant gut microbiota

Gianluca Galazzo[*], Niels van Best[*], Liene Bervoets[*], Paul H. Savelkoul,

Monique Mommers[§], John Penders[§]

** Shared first authorship*

§ Shared last authorship

Manuscript in preparation

Abstract

Studies on the assembly and development of the infant microbiota have often used limited time-points to profile the faecal microbiota during this critical window and mainly focused on deterministic processes over neutral (i.e., random) processes. The present study aimed to identify processes and dynamics involved in the assembly of microbial communities during infancy considering the impact of host, diet, environment, and their consecutive, dynamic interactions.

We collected 806 faecal samples of 98 mother-infant dyads participating in the LucKi Gut study. At 1-2 weeks post-partum both infant and maternal faecal samples were collected, whereas additional infant samples were collected at age 4-, 8-weeks and at 4,5,6,9,11 and 14 months. Microbial composition was determined by 16SrRNA V3-4 region amplicon sequencing. Alpha- and beta-diversity indices were calculated and microbial clusters were identified based on Dirichlet Multinomial Modelling. Information on current health status, medication use, diet and lifestyle of both mother and child was obtained by administering repeated questionnaires accompanying faecal sample collection.

Microbial diversity increased throughout the first 14 months of life with largest changes in both diversity and community structure occurring between the age of 6- and 9-months post-partum. When comparing infant to maternal samples it appeared that at the age of 14 months the microbiota is still far from mature both in terms of diversity and community structure. To identify ecological processes that are involved in the microbial community maturation, we first clustered the genus-level data using DMM-modelling. At 1-2 weeks of age infants had a microbial community that was either characterized by a high abundance of *Bifidobacterium* (cluster 1), *Escherichia* (cluster 4) or *Bacteroides-Parabacteroides* (cluster 6). The latter cluster was only observed in infants born by vaginal delivery. When examining shared amplicon sequence variants (ASVs) between mother-infant dyads, we confirmed that within the genera of *Bacteroides* and *Parabacteroides* maternal ASVs were significantly more frequently shared in vaginally delivered as compared to Caesarean section delivered infants. Subsequent maturation of the microbial community structure was mainly driven by diet. Although breastfeeding had the strongest impact on microbiota maturation, the age of introduction of solid food, composition and diversity of complementary foods also had a profound impact, as indicated by an increased microbial diversity and altered community structure. In particular, a more "mature omnivore diet" characterized by consumption of meat, fish, pasta and rice was associated with a microbial community dominated by amongst others *Faecalibacterium* and *Blautia* (cluster 5). Next to dispersal from the maternal microbiota and environmental selection by dietary substrates, stochastic processes also appeared to have a profound impact on the microbiota. Over 80% of ASVs were under neutral selection as indicated by the application of a neutral community model. Finally, from the age of 9 months onwards, the interaction with peers in the form of older siblings was significantly associated with the infant microbial community structure.

In conclusion, microbiota assembly and maturation are dynamic processesthat are influenced by birth mode, diet and dispersal from household members. However, also neutral processes, that are far from being completed at 14 months of age have a profound impact. Additional follow-up and application of ecological theory could shed fur-

ther light on the microbial assembly and maturation during this critical time window.

Introduction

The indigenous microbiota of the human gut has long been recognized to contribute to health and disease by influencing gut maturation, host nutrition and pathogen resistance [1]. In line with this, perturbations in gastrointestinal (GI) microbiota composition have been associated with development of obesity [2], allergies [3], inflammatory bowel diseases [4], and colorectal cancer [5]. Our understanding of intestinal microbial ecology, therefore, has a direct impact on our ability to manage and maintain human health [6, 7]. The unexpected variation in the composition of the microbiota of healthy individuals [8-11] highlights the importance of identifying the processes that could give rise to such variation. Large studies on genetic determinants of the gut microbiome indicated that, although a small part of the microbial taxa (<2%, mainly within the phylum Firmicutes) appear to be heritable, host genetics only play a small role in determining the microbiota composition compared to environmental factors [12-14]. Two recent large-scale studies on the impact of dietary, lifestyle, medication and health-related factors revealed that together these specific factors could only explain a limited amount of the variation (<20%) in microbiota composition between individual adults [15, 16]. Whereas human adults have highly differentiated bacterial communities in the GI tract, in infants these communities appear to be largely undifferentiated [17]. The initial pioneer bacteria colonizing the neonatal gut, are strongly affected by birth mode [17-20]. However, the factors driving this first inoculum to differentiate into a highly complex and more stable 'adult-like' microbiota, as established after the first years of life, are largely unknown. Insight into the ecological factors that shape the microbiome during the first year of life is crucial, not only because it sets the stage for microbial maturation, but also because it is the critical time-window during which the microbiota provides a stimulus for the adequate development of the gut and immune system with persistent local as well as systemic effects [21].

It is therefore increasingly being recognized that in order to refine our understanding on the largely unexplained (processes driving) inter-individual variation in intestinal microbial composition and maturation we need to apply the principles of ecological theory to explain and predict community characteristics in order to develop successful strategies to reshape the microbiota where needed.

Environmental filtering (or niche-based interactions), dispersal (limitation), historical contingency and random sampling likely all contribute to the assembly of microbial communities and give rise to the compositional variations in the human microbiota [6].

Next to the (absence of) exposure to maternal faecal and vaginal microbes during delivery, other factors that have been indicated to influence dispersal of microbial species through host-host and environment-host contact include place of delivery, hospitalization, family size, day care attendance and pet exposure [22-25]. The impact of environmental filtering (or niche-based interactions) by diet has been described for both infants and adults. Next to the well-established large geographical differences in the composition of the human gut microbiome that can largely be attributed to dietary habits [26, 27], type of infant feeding (breast vs. formula-based milk) and the introduction of solid food have a profound impact on the microbiota development [28]. Besides diet,

medication use (in particular antibiotics) and bowel habits (i.e., transit time, stool moisture) are among the abiotic factors that may drive environmental filtering [15, 16, 25]. Moreover, the founder species themselves can initiate processes that influence the establishment of the more complex and stable adult ecosystems (historical contingency). This is clearly demonstrated by the initial facultative anaerobes which help to reduce the redox potential, thereby paving the way for the strict anaerobic species. A study by Faith and colleagues, applying a targeted approach focusing on *Methanobrevibacter smithii* and *Bacteroides thetaiotaomicron* strains suggested that early gut colonizers such as those acquired from parents and siblings, have the potential to exert their physiological, metabolic, and immunologic effects for most, and perhaps, all of our lives [29]. Characteristics of microbial communities as well as their environment fuel processes like diversification and ecological drift. Growth rate, size of microbial population and the capacity to adapt through recombination or mutation can all lead to survival or alternatively extinction of species with no competitive advantage. In addition, the order and timing of species arrival (priority effects) potentially determines consecutive colonization patterns [30].

Studies performed to assess the relationships between determinants and compositional differences so far, focused mainly on singular time-points in adult populations or singular determinants. Little is still known on how the microbiota composition develops over time during early infancy and on how the timing of exposure to determinants influences inter-individual variations in microbial composition. Moreover, studies that explore the relative contribution of deterministic over neutral (i.e., random) processes, such as dispersal and stochasticity, are largely lacking [31].

The assessment of temporal dynamics might lead to insight into when processes affect the microbiota the most. This knowledge is essential to be able to specifically address deterministic factors that increase the risk for microbiota related disease and determines success of intervention studies.

The present study aimed to identify processes involved in the assembly of microbial communities in the gut of 98 infants during the first 14 months of life, considering host, microbiota, environment, and their consecutive, dynamic interactions in their entirety.

Materials and Methods

Study population and design

The LucKi Gut study is an ongoing birth cohort study that aims to monitor gut microbiota development throughout infancy and early childhood and is embedded within the larger Lucki Birth Cohort Study [32]. Pregnant women from the South Limburg area in the Netherlands are recruited via professionals involved in mother and childcare and through the internet (study website and Facebook). Women are eligible to participate in the LucKi Gut study if they give birth at >37 weeks of completed gestation. Study questionnaires and faecal samples of the infant are collected at ages 1-2, 4, 8 weeks, 4, 5, 6, 9, 11 and 14 months postpartum.

Parents were instructed to collect infant feces from the diaper and freeze immediately at -20°C in their home freezer in a cool transport container (Sarstedt, Hilden, Germany). Next to infant samples, a maternal sample was self-collected at 1-2 weeks

post-partum using a FecesCatcher (Tag Hemi VOF, Zeijen, the Netherlands). Samples were transported to the laboratory maintaining the cold chain.

Together with the collection of the maternal and first infant faecal samples at the age of 1-2 weeks postpartum, parents were asked to fill in a questionnaire on the pregnancy period as well as two short questionnaires on the current health status, medication use and diet of both mother and child. Each subsequent sampling time-point was also accompanied with a questionnaire to collect information on lifestyle, health status, medication use, development, and diet.

Identification of dietary patterns in infancy

Next to the duration of (exclusive) breastfeeding, formula feeding and the age at introduction of solid foods, we assessed the dietary patterns at 9 months of age. The current dietary intake at a given age was obtained by asking parents whether their child had ever been fed a specific food item. Dietary patterns were identified using dimension reduction by means of categorical PCA (IBM SPSS Statistics Version 25.0). Food items that had been consumed by less than 5 infants at a certain age were excluded from the analysis. Moreover, food items with a VAF below 0.25 were excluded from the PCA. The number of dietary patterns extracted was based on a break (elbow) in the Scree plot and the interpretability of the patterns [33].

Next to dietary patterns, the dietary diversity was calculated at 4, 5, 6, 9 and 11 months of age as the total number of food items consumed multiplied by the number of the following food categories: fruits, vegetables, bread, potato, cereal products, flesh foods (fish, meat, poultry), dairy and water/juices.

Microbial Profiling of Faecal Samples

In total 806 faecal samples from 98 infants and 90 mothers were available for sequencing.

Metagenomic DNA was extracted with a custom extraction protocol involving mechanical and enzymatic lysis as described previously [34] with some modifications. First, 300 µl of sample was added to a tube containing 2.8mm ceramic beads and 0.1mm glass beads (MoBio Laboratories Inc., Carlsbad, CA, USA) along with 800 µl of 200 mm sodium phosphate monobasic (pH 8) and 100 µl guanidinium thiocyanate EDTA N-lauroylsarkosine buffer (50.8 mM guanidine thiocyanate, 100 mM ethylenediaminetetraacetic acid and 34 mm N-lauroylsarcosine). Samples were bead-beated in a PowerLyzer 24 Benchtop Homogenizer (MoBio Laboratories Inc.) for 3 min at 3000 revolutions per minute and centrifuged. The supernatant was further processed using the MagMAX Express 96-Deep Well Magnetic Particle Processor from Applied Biosystems with the Multi-Sample kit (Life Technologies#4413022). Purified DNA was used to amplify the V3-V4 region of the 16S rRNA gene by PCR. 50 ng of DNA was used as template with 1U of Taq, 1x buffer, 1.5 mM MgCl2, 0.4 mg/mL BSA, 0.2 mM dNTPs, and 5 pmoles each of 341F (CCTACGGGNGGCWGCAG) and 806R (GGACTACNVGGGTWTCTAAT) Illumina adapted primers, as described in Bartram et al. [35]. The reaction was carried out at 94C for 5 minutes, 5 cycles of 94C for 30 seconds, 47C for 30 seconds and 72C for 40 seconds, followed by 25 cycles of 94C for 30 seconds, 50C for 30 seconds and 72C for 40 seconds, with a final extension of 72C for 10 minutes. Resulting PCR products were visualized on a 1.5% agarose gel. Positive amplicons were normalized using the SequalPrep normalization kit (ThermoFisher#A1051001) and sequenced on the Illu-

mina MiSeq platform at the McMaster Genomics Facility.

Raw 16S-rRNA sequence reads were demultiplexed and technical sequences were trimmed off. We subsequently used the DADA2 pipeline for sequence quality control and feature table construction using default settings for trimming and filtering. The high-quality reads resulting from denoising and chimera-filtering steps were clustered into a table of Amplicon Sequence Variant (ASV) and taxonomy was assigned both using SILVA version 138 [36] and GreenGenes version 13.8 [37].

Samples with low sequencing depth (<10,000 reads/sample) were excluded, resulting in a total of 775 samples from 98 infants and 86 mothers that were retained for downstream analyses. Moreover, rare ASVs (present in less than 5 samples) and very low abundant ASVs (average relative abundance across all samples of <0.001%) were filtered out.

Data transformation and descriptive analysis

To make the data scale-invariant and control for the compositional nature of sequencing data, we modelled the probability of the observed count data using the Dirichlet distribution (using 128 Monte Carlo iterations) as implemented in the ALDEX2 package. Subsequently the probabilities were centred log ratio (*clr*) transformed [38].

The Shannon index, as a metric of microbial diversity within samples (alpha diversity), was computed using the R package *vegan 2.5.3* [39]. The Skillings-Mack test was used to test for significant differences in Shannon between all timepoints, followed by Dunn's test for post hoc pairwise comparisons between individual time points. The *P* values were finally false discovery rate (FDR) adjusted for multiple comparisons.

The Aitchison distance, reflecting the Euclidean distance between two *clr*-transformed compositions, was computed using the *vegan* package and was used as a metric of between sample dissimilarity (beta-diversity). Ordination of samples based on the Aitchison distance was obtained by principal component analysis using the *vegan* package and visualized using *ggplot2*.

Dispersal of microbial taxa between mother-child pairs and neutral assembly

To identify the proportion of ASVs shared between mother-infant dyads and between infants and unrelated adults, the Jaccard index was calculated. The Mann-Whitney (Wilcoxon rank sum) test was used to examine whether infants shared more ASVs with their mothers than with unrelated adults. We subsequently examined whether the proportion of shared ASVs in mother-infant dyads differed between infants born by Caesarean (C-) section and vaginally delivery.

The neutral assembly theory assumes that all microbes in a meta-community (all microbes in a regional species pool) have an equal ability to disperse to a local area, e.g., an infant's gut, and once established, all have equal fitness, growth, and death rates. If there is a higher abundance of a specific taxon in the meta-community then the prevalence of this taxon in the local communities is higher, but each single microorganism has an equal chance of establishing.

We have used the neutral assembly model as recently applied by Sprockett et al. [31]. This model is based upon a nonlinear least-squares algorithm (R package *minpack.lm* version 1.2-1) to predict the prevalence of a microbial taxon in a local community

based upon its average relative abundance in the meta-community. Neutral dispersal holds true for those taxa for which the microbial distributions that are consistent with the model's predictions, while taxa deviating from the model are under either positive or negative selection [31].

Age-dependent transition of enterotypes

Dirichlet Multinomial Mixture (DMM) clustering, an unsupervised clustering method that in which Laplace approximation was used to identify groups of communities (enterotypes) with similar composition, was performed as previously described [40]. Next, the transition of infants through these DMM clusters with age, was analysed [41].

Impact of environmental and dietary factors on microbial richness, diversity, and community structure

Multivariable linear regression models were run to identify environmental and dietary factors associated with microbial richness (Chao1) and diversity (Shannon). Separate models were run for each of the different ages at which faecal samples were collected.

We examined which perinatal, environmental, and dietary factors were associated with the microbiota community structure throughout infancy. The effect size and significance of each of the covariates on the microbial community structure was determined using the *envfit* function in *vegan*. All *P* values derived from *envfit* were adjusted for multiple comparisons using FDR adjustment (Benjamini–Hochberg procedure). To understand which of the covariates had the strongest impact on the overall microbial community structure, we performed a permutational analysis of variance based on the Aitchison distance. Only covariates that were statistically significant in the *envfit* analyses were included in the permutational analysis of variance.

Results

Study population

A total of 775 samples from 98 infants and 86 mothers were available after quality processing of the sequencing data. Of the infants, 44 (44.9%) were girls and 54 (55.1%) were boys with a mean birth weight of 3418 (SD 467.4) gram. The majority of children (84/98, 85.7%) were born by vaginal delivery of which 12 were delivered at home. About half of the infants (48/98, 49.0%) had older siblings, while furry pets were present in 37.8% (37/98) of the families. By the age of 6 months, almost two-thirds (54/85, 62.4%) of the infants attended day care or a guest family.

Breastfeeding was initiated in 86 out of the 98 infants and the median duration of breastfeeding was 5 months (IQR 1 - ≥14 months). Solid foods were introduced at a median age of 18 weeks [IQR 17 - 18 weeks].

Infant microbiome development

The microbial diversity, as measured by the Shannon index, gradually increased from a median of 1.77 [IQR 1.39 - 2.07] at 1-2 weeks postpartum to a median of 3.18 [IQR 2.93 - 3.57] at 14 months of age. Accompanying the increasing diversity in dietary substrates, the largest increase in microbial diversity was observed from 6 months on-

wards (Figure 1A). The significantly higher diversity in mothers (median 4.04, IQR 3.63 - 4.23, P<0.001) as compared to infants aged 14 months, indicated that microbial maturation is still not completed by that age.

Ordination of the microbial community structure, as assessed by the Aitchison distance, also showed a gradual shift throughout infancy along the first component (Figure 1B) with the largest shift in within-subject microbial structure observed between 6 and 9 months of age (Figure 1C). The ordination was driven by a high abundance of amplicon sequence variants representing *Staphylococcus epidermidis* (ASV50) and *Enterococcus spp.* (ASV22) in the early-life samples. Moreover, a high abundance of amongst others *Faecalibacterium prausnitzii, Blautia spp., Coprococcus spp., SMB53 spp.* (family Clostridiaceae) and unknown Lachnospiraceae (ASV29, 8, 40, 46 and 39 respectively) in the samples from 9 months onwards was detected. When including maternal samples (Supplementary Figure 1), it is apparent that at the age of 14, months the microbial community structure is still far from mature.

Ecological processes in microbial immunity maturation

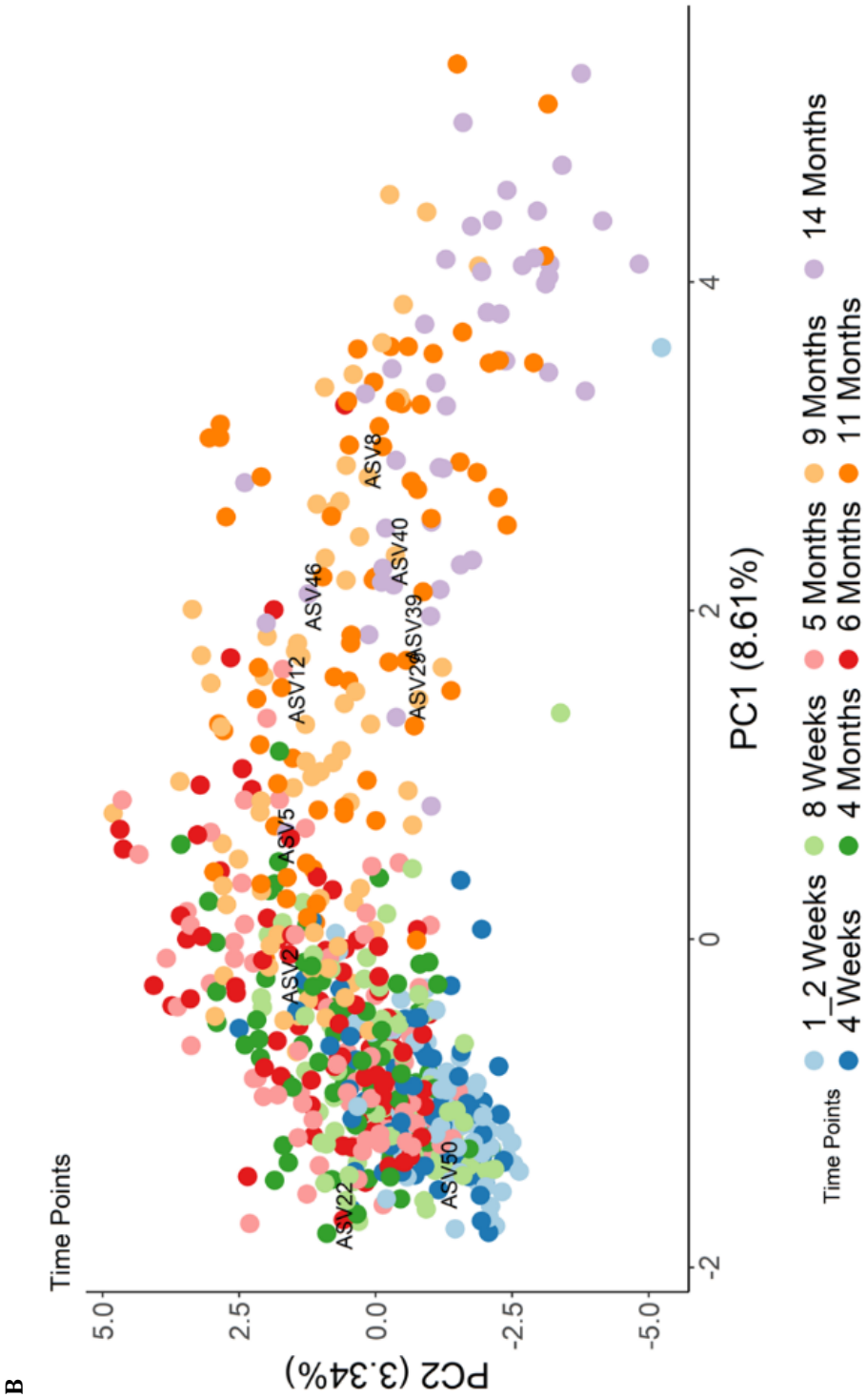
To identify ecological processes that are involved in the microbial community maturation, we first clustered the genus-level data using DMM-modelling. Based on the lowest Laplace approximation, 6 clusters were identified that strongly associated with age (Figure 2A). Cluster 1, 4 and 6 mainly included samples collected during the first 6 months of life. These clusters were characterized by a high abundance of *Bifidobacterium* (cluster 1), *Escherichia* (cluster 4) and *Bacteroides-Parabacteroides* (cluster 6) (Figure 2B). Clusters 3 and 5 were almost exclusively populated by samples collected at the ages of 9 months and beyond and were differentiated by the much higher relative abundance of *Faecalibacterium* and *Blautia* in cluster 5.

Transition modelling of the microbiota throughout infancy showed that in the neonatal period the microbiota of infants either belonged to cluster 1, 4 or 6 (Figure 2C). Interestingly, the developmental trajectories differed substantially depending on the initial microbiota cluster at 1-2 weeks of age. Many infants in cluster 1 and 4 shifted to cluster 2 between the age of 8 weeks and 4 months, and further transitioned to clusters 3 and 5. Infants that started off in cluster 6 on the other hand mostly remained in this cluster up to the age of 6 months after which they directly transitioned to cluster 3. As expected by the high abundance of *Bacteroides* and *Parabacteroides,* all children in cluster 6 were born by vaginal delivery suggesting that exposure to maternal feces during delivery is mainly driving this cluster.

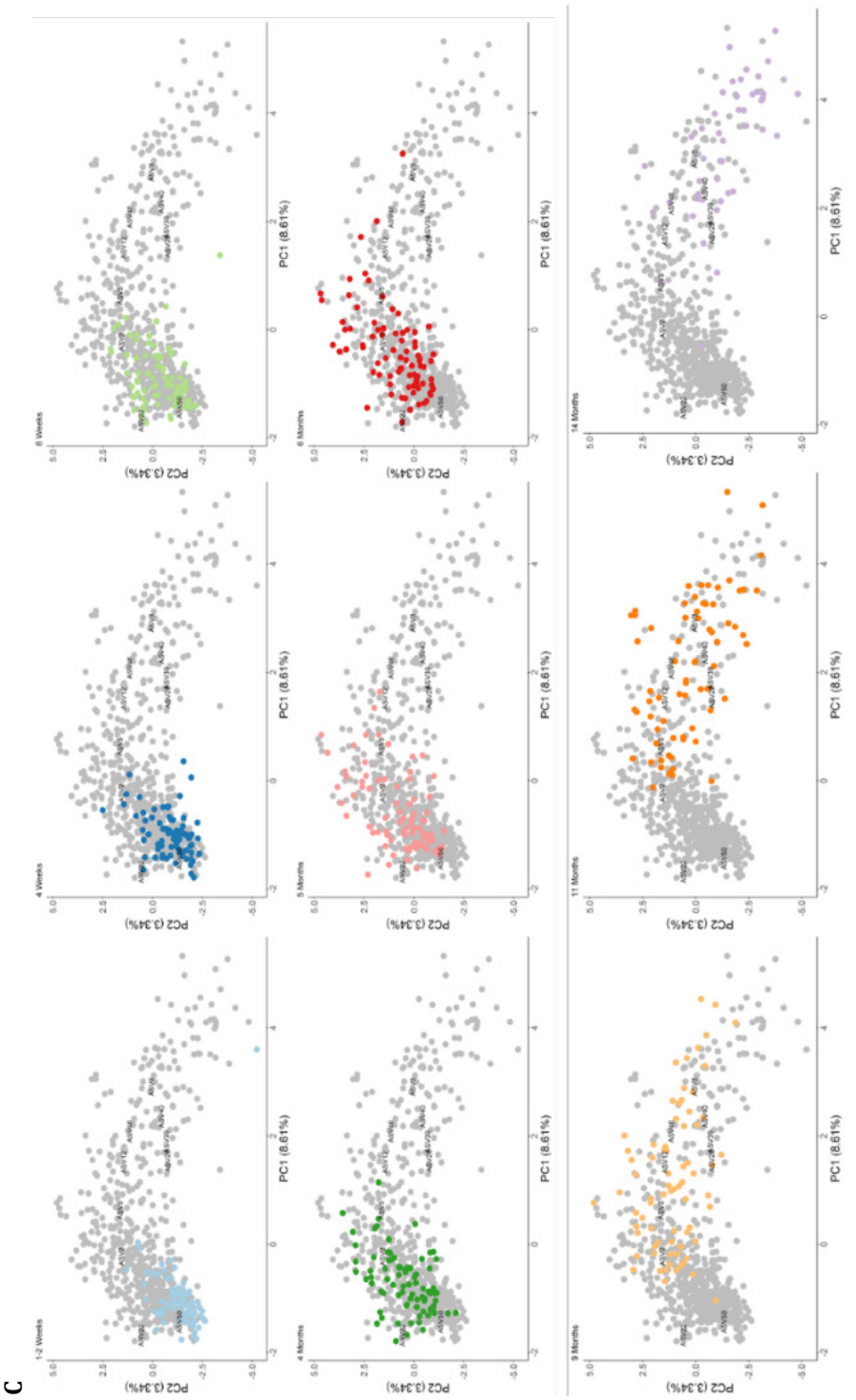
To further explore the impact of birth mode on the dispersal of maternal microbes we next examined the proportion of ASVs shared between mother-infant dyads as compared to the proportion of shared ASVs between infants and unrelated adults (Figure 3). When examining the entire study population infants did not share significantly more ASVs with their own mothers than with unrelated adult individuals (Figure 3A). However, when stratifying according to birth mode it became apparent that infants born by vaginal delivery did share significantly more ASVs with their own mothers than with unrelated adults throughout the first 8 weeks of life (Figure 3B). When examining the proportion of mother-infant dyads that shared ASVs at 1-2 weeks postpartum within each genus separately, it was confirmed that within the genera of *Bacteroides* and *Parabacteroides* ASVs were statistically significantly more frequently shared in vaginally delivered

A

	4 Weeks	8 Weeks	4 Months	5 Months	6 Months	9 Months	11 Months	14 Months
1_2 Weeks	4.455e-01	-	-	-	-	-	-	-
4 Weeks	1.078e-01	4.941e-01	-	-	-	-	-	-
8 Weeks	1.401e-02	2.035e-01	2.437e-01	-	-	-	-	-
4 Months	2.005e-04	1.383e-02	4.527e-03	1.619e-02	-	-	-	-
5 Months	1.878e-04	1.320e-03	7.313e-04	1.056e-02	5.192e-01	-	-	-
6 Months	7.495e-10	2.870e-10	2.358e-11	2.121e-11	2.397e-09	1.654e-10	-	-
9 Months	1.026e-10	5.037e-11	1.705e-11	2.121e-11	2.669e-11	5.037e-11	1.311e-03	-
11 Months	1.705e-11	5.037e-11	4.093e-12	4.093e-12	9.550e-12	1.705e-11	4.333e-07	3.603e-04


Time Points
ASV22
ASV50
ASV2
ASV5
ASV12
ASV46
ASV8
ASV40
ASV39
ASV28
PC2 (3.34%)
5.0
2.5
0.0
-2.5
-5.0
-2
0
2
4
PC1 (8.61%)
Time Points
1_2 Weeks
4 Weeks
8 Weeks
4 Months
5 Months
6 Months
9 Months
11 Months
14 Months

B


1-2 Weeks
4 Weeks
6 Weeks
4 Months
5 Months
6 Months
9 Months
11 Months
14 Months
PC1 (8.61%)
PC2 (3.34%)

c

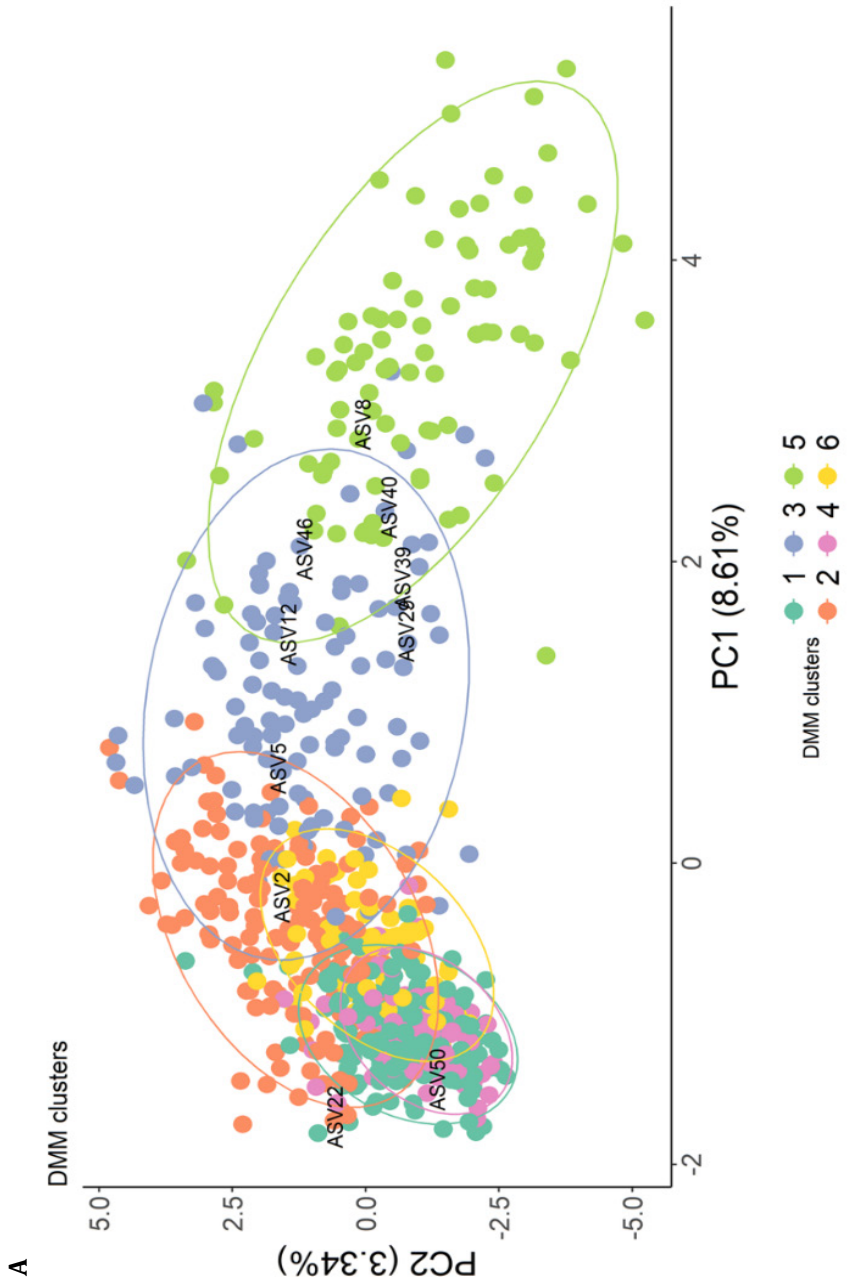
	1_2 Weeks vs 4 Weeks	4 Weeks vs 8 Weeks	8 Weeks vs 4 Months	4 Months vs 5 Months	5 Months vs 6 Months	6 Months vs 9 Months	9 Months vs 11 Months
4 Weeks vs 8 Weeks	1.429e-02	-	-	-	-	-	-
8 Weeks vs 4 Months	3.001e-07	1.694e-05	-	-	-	-	-
4 Months vs 5 Months	2.524e-02	1.229e-01	1.091e-03	-	-	-	-
5 Months vs 6 Months	1.591e-01	7.712e-01	1.003e-04	4.854e-01	-	-	-
6 Months vs 9 Months	4.977e-09	1.115e-09	4.977e-09	1.115e-09	9.759e-11	-	-
9 Months vs 11 Months	4.977e-09	4.977e-09	1.970e-07	5.702e-09	4.977e-09	1.880e-01	-
11 Months vs 14 Months	5.526e-07	4.217e-07	1.970e-07	1.156e-07	4.624e-08	1.879e-02	1.267e-04

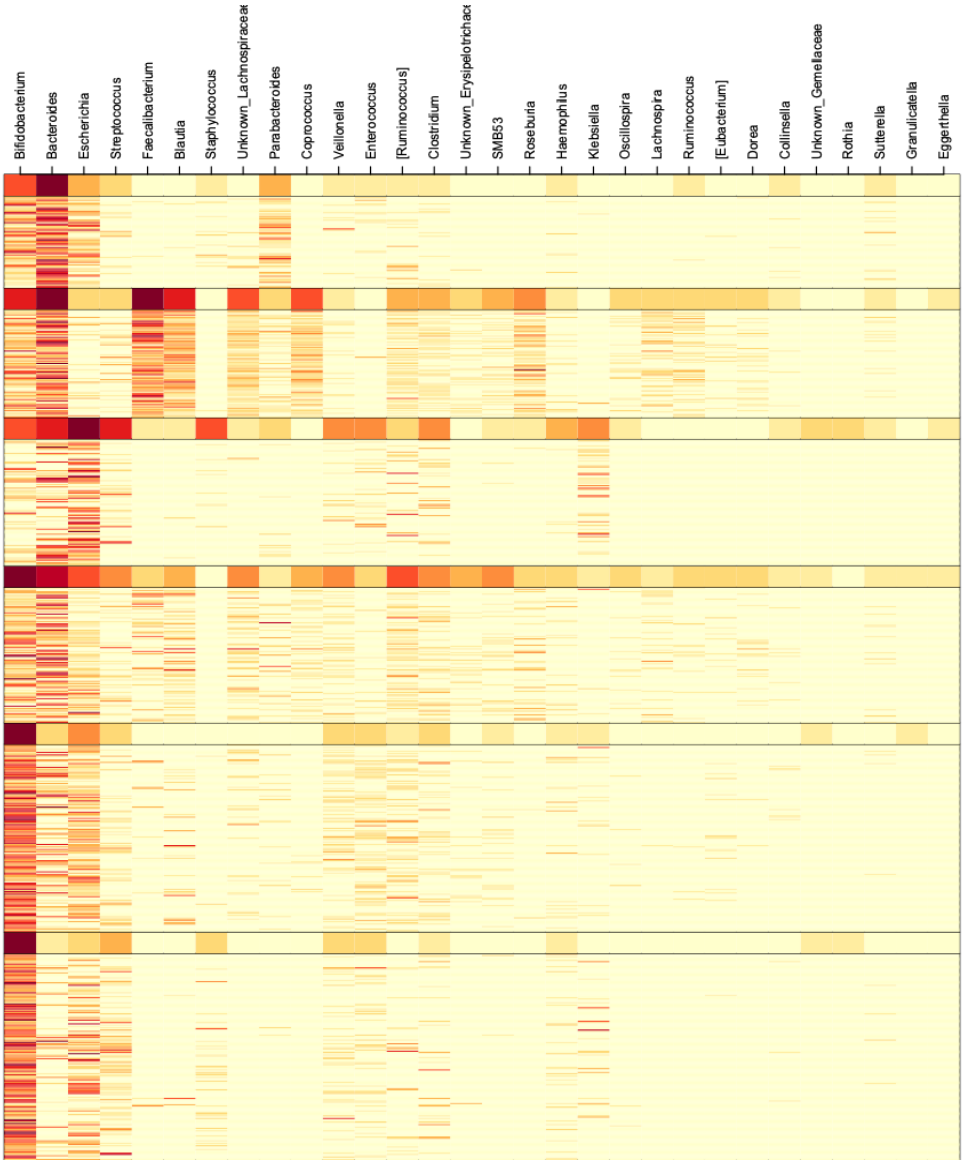
D

Figure 1 General development of the microbial diversity and community structure throughout infancy. **(*A*)** Microbial diversity (Shannon index) in faecal samples throughout infancy. The overall significance was tested using the Skillings-Mack test, the table depicts FDR-adjusted p-values based upon post-hoc test performed using the Wilcoxon signed-rank test to examine significance of within-subject changes in microbial diversity. **(*B*)** Principal Component Analyses visualizing the ordination of the microbial community structure (based upon the Aitchison distance) of infant samples coloured according to age. Depicted ASV's are the most external ASVs based on their coordinates. The smaller plots **(C)** are identical PCAs with only one of the time-points coloured and all remaining samples depicted in grey **(*D*)** Within-subject change in microbial community structure between subsequent time-points. Significance was tested using the Skilling-Mack test, the table depicts FDR-adjusted p-values based upon post-hoc test performed using the Wilcoxon signed-rank test to examine significance of within-subject changes in microbial community structure

as compared to C-section delivered infants (Figure 3C and 3D, both P<0.0001). In fact, none of the mothers that delivered by C-section shared any of the *Parabacteroides* ASVs with their infants. Further exploration of the dynamics of the five most abundant *Bacteroides/Parabacteroides* ASVs, revealed that while some ASVs became shared between mothers and their C-section delivered infants at later time-points (delayed sharing) for other ASVs sharing between mother-infant dyads in the C-section delivered infants remained sporadic (Supplementary Figure 2).

Infant diet was mainly driving the transition to DMM cluster 2. At 4 months, only 34.2% (13/38) of children in DMM cluster 2 were still receiving breastfeeding, while 86.7% (39/45) of infants in the other DMM clusters were still breastfed (Chi-square $P = 8.5618*10E^{-7}$). Formula feeding on the other hand was common among children in DMM cluster 2 at the age of 4 months (83.3% (30/36), while much less frequent in children in the other DMM clusters (27.3% 12/44, $P = 5.8736*10E^{-7}$). Also, at 5 months of age a similar pattern was observed with a minority of infants in DMM cluster 2 still receiving breastfeeding and a vast majority receiving formula feeding.


A
DMM clusters
PC2 (3.34%)
5.0
2.5
0.0
-2.5
-5.0
PC1 (8.61%)
-2
0
2
4
DMM clusters
1
2
3
4
5
6
ASV2
ASV5
ASV8
ASV12
ASV22
ASV23
ASV39
ASV40
ASV46
ASV50


Bifidobacterium
Bacteroides
Escherichia
Streptococcus
Faecalibacterium
Blautia
Staphylococcus
Unknown_Lachnospiraceae
Parabacteroides
Coprococcus
Veillonella
Enterococcus
[Ruminococcus]
Clostridium
Unknown_Erysipelotrichace
SMB53
Roseburia
Haemophilus
Klebsiella
Oscillospira
Lachnospira
Ruminococcus
[Eubacterium]
Dorea
Collinsella
Unknown_Gemellaceae
Rothia
Sutterella
Granulicatella
Eggerthella

B

C

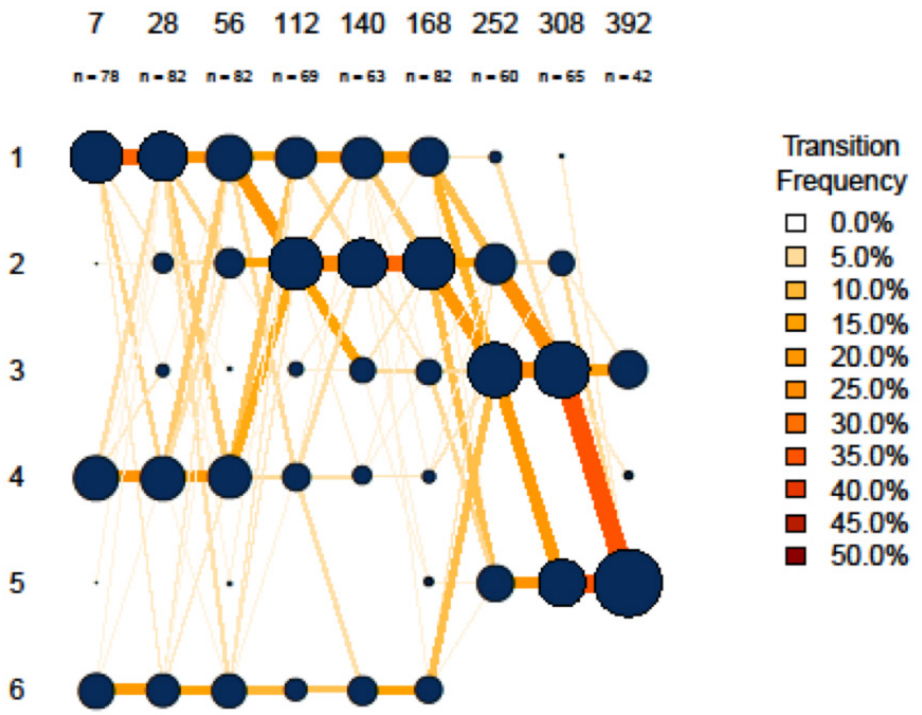
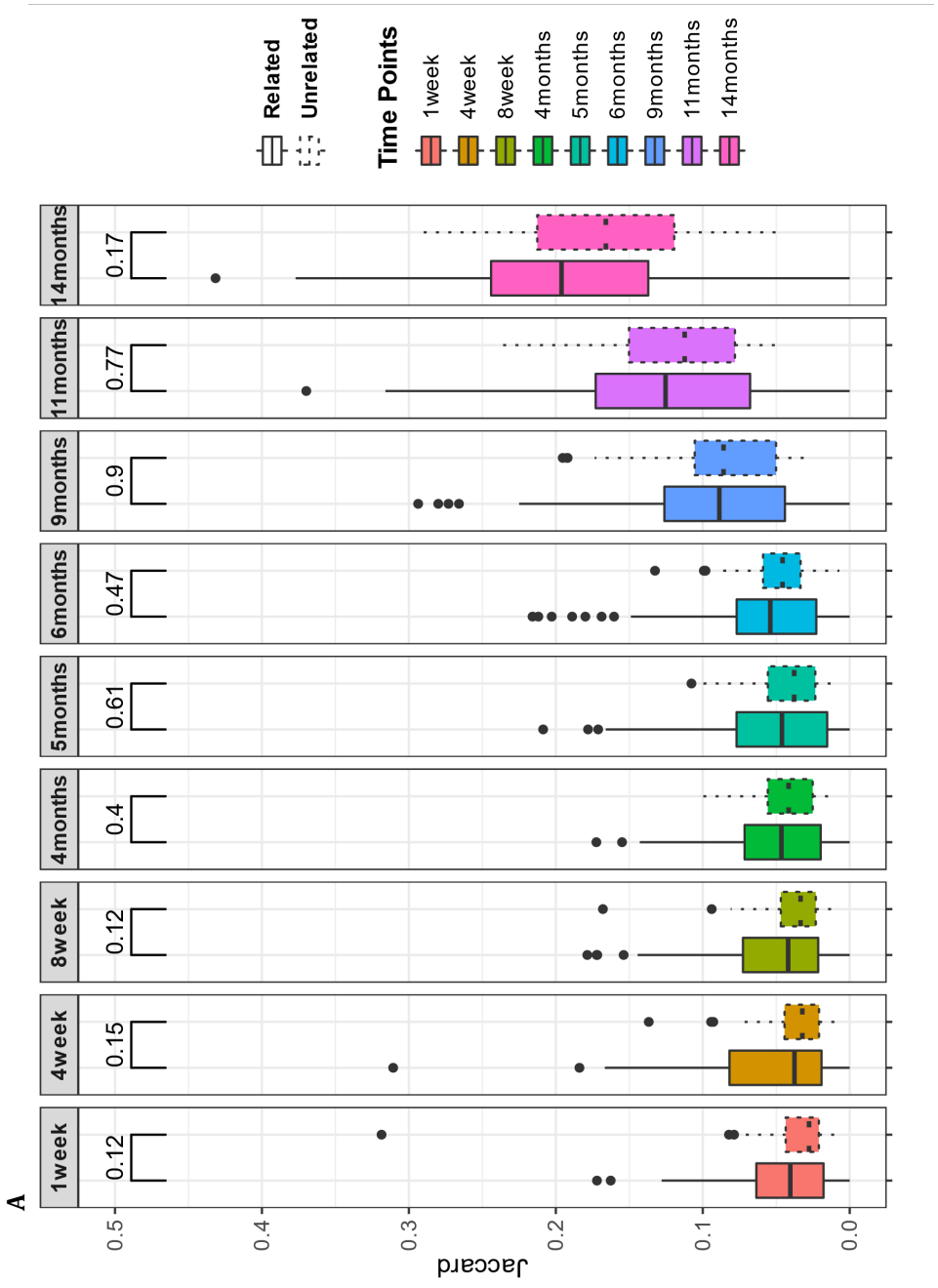
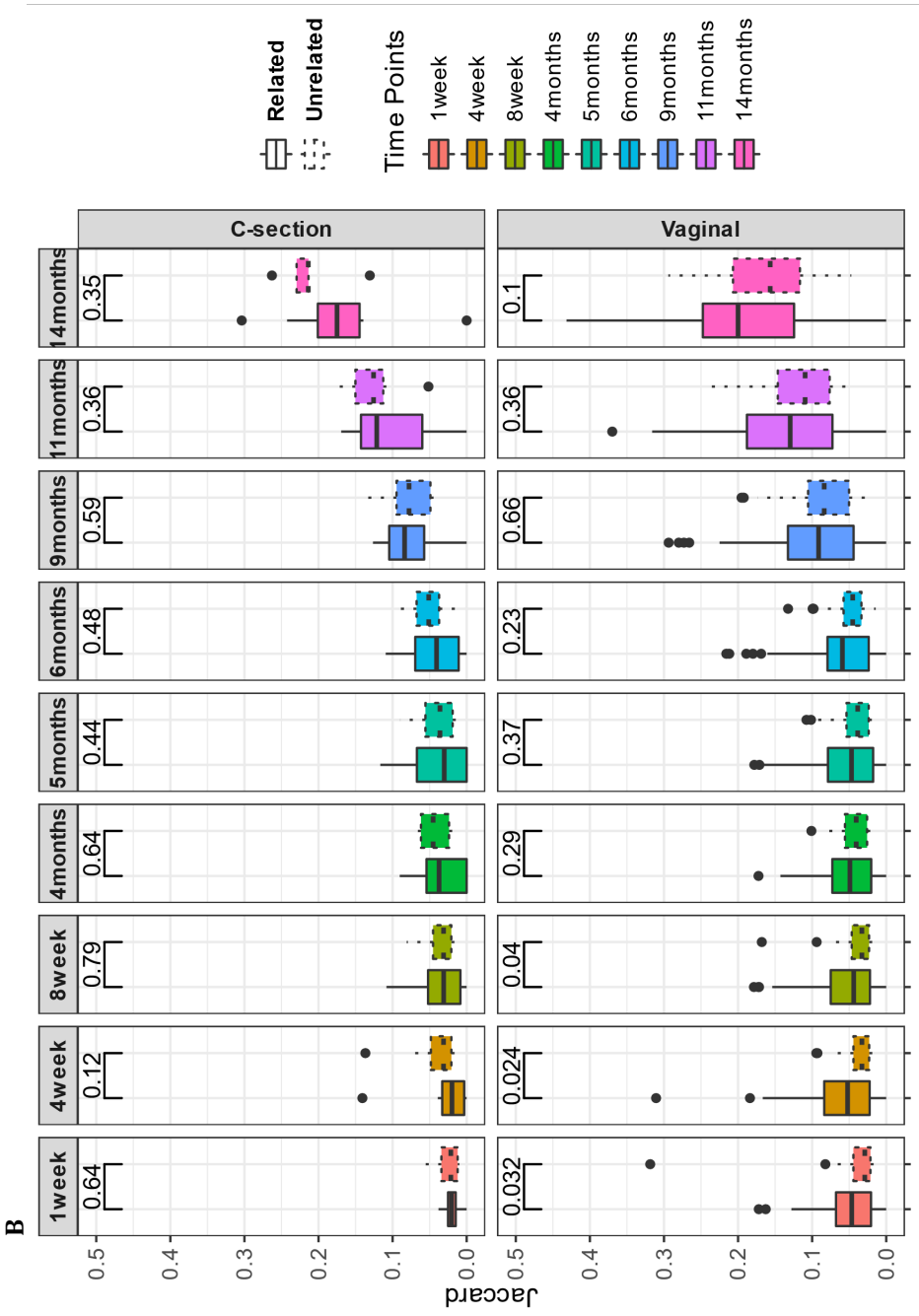
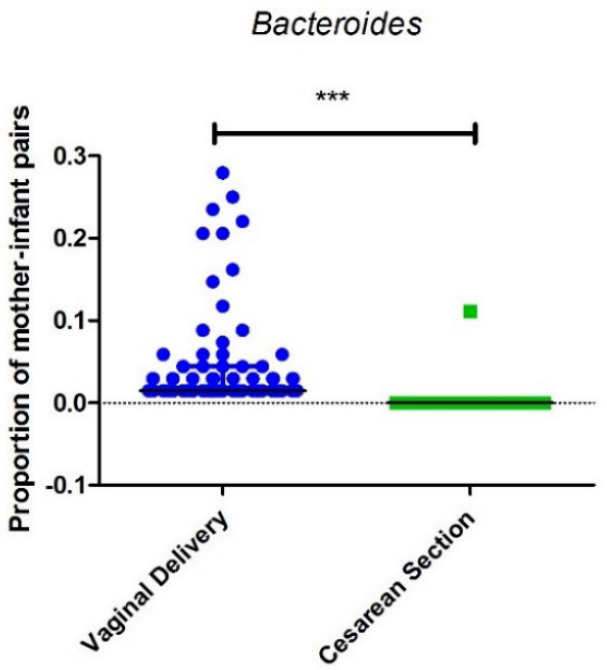


Figure 2 Community typing by DMMs of infant samples revealed 6 clusters (n = 689 stool samples from 98 children). **(A)** Principal component analysis on ASV-level data with samples coloured according to DMM cluster. Ellipses indicate the 95% confidence interval. **(B)** Heat map showing the relative abundance of the 30 most important/dominant taxa per DMM cluster. **(C)** Transition model showing the progression of samples through the 6 DMM clusters from one sampling time point to the next time point. The thickness and colour of the connecting lines represent the transition frequency, the size of the circles represent the number of infants in that particular cluster at each time-point.


A
Related
Unrelated
Time Points
1week
4week
8week
4months
5months
6months
9months
11months
14months
1week
4week
8week
4months
5months
6months
9months
11months
14months
0.12
0.15
0.12
0.4
0.61
0.47
0.9
0.77
0.17
0.5
0.4
0.3
0.2
0.1
0.0
Jaccard


B
Related
Unrelated
Time Points
1week
4week
8week
4months
5months
6months
9months
11months
14months
C-section
Vaginal
1week
4week
8week
4months
5months
6months
9months
11months
14months
0.64
0.12
0.79
0.64
0.44
0.48
0.59
0.36
0.35
0.032
0.024
0.04
0.29
0.37
0.23
0.66
0.36
0.1
Jaccard
0.5
0.4
0.3
0.2
0.1
0.0

C


Bacteroides
***
Proportion of mother-infant pairs
0.3
0.2
0.1
0.0
-0.1
Vaginal Delivery
Cesarean Section

D

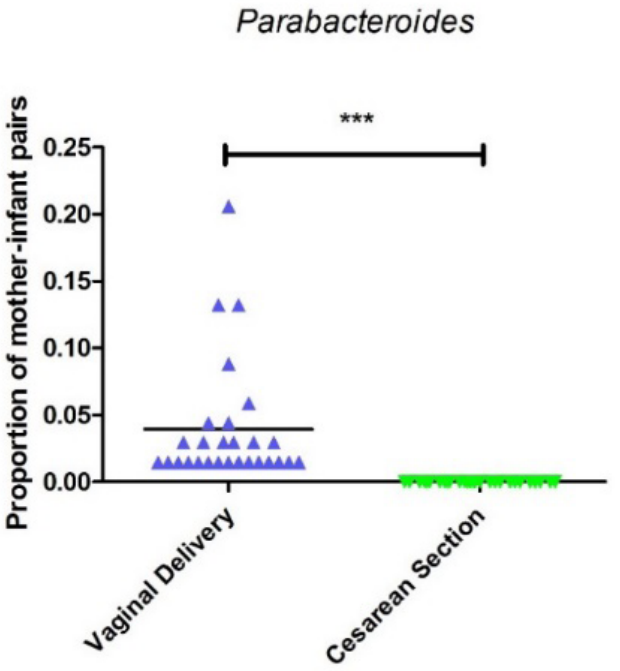

Parabacteroides
***
Proportion of mother-infant pairs
0.25
0.20
0.15
0.10
0.05
0.00
Vaginal Delivery
Cesarean Section

Figure 3 Impact of dispersal from maternal faecal microbiota. (A, B) Microbiota similarity (Jaccard similarity index) on ASV-level in mother-infant dyads as compared to the microbiota similarity of infants and unrelated mothers in the entire population (A) and stratified according to mode of delivery (B). The significance was tested using the Mann-Whitney test. (C, D) Proportion of mother-infant dyads sharing Bacteroides (C) and Parabacteroides (D) ASVs at 1-2 weeks post-partum according to mode of delivery.

We next quantified the contribution of stochastic processes using a neutral community model (NCM) that has recently been applied to study microbial assembly in the Tsimane horticulturalists of the Bolivian Amazon [31]. This model predicts the prevalence of each microbe given its average relative abundance in the meta-community. Microbes are considered to have assembled neutrally if the true prevalence fits the prediction. Microbes that have a higher or a lower prevalence than predicted are considered to have been under local positive or negative ecological selection, respectively. The meta-community was estimated by summing all the individual infant microbial communities. In line with the study by Sprockett *et al.* [31], we randomly selected one sample per infant with 1000 permutations to calculate a bootstrapped estimate of model fit. The NCM shows that the majority of ASVs in infant stool samples, 82.5%, were neutrally distributed in at least 80% of permutations, while only 5% were consistently under positive and 0.2% under negative selection. The remaining 12.3% of ASVs were variable in their predicted fit to the NCM across permutations (Figure 4A).

It should be noted that, due to the many simplified assumptions, the NCM cannot be considered as a complete description of community assembly and that neutrally distributed taxa can still experience selection, if the selection was of similar direction and magnitude as the selection present in the meta-community.

The model is however particularly useful to help identify factors that lead to divergence from neutral dynamics. In this respect, *Bacteroides fragilis* ASV6 which was commonly shared among mothers and their vaginally born infants was under negative selection in line with the dispersal limitation in case of C-section. In addition, a *Bifidobacterium longum* ASV was also under negative selection whereas most ASVs that did not fit the neutral distribution were under positive selection, meaning that the prevalence of these ASVs in the infant study population was higher than predicted based upon the average abundance in the meta-community. In line with previous infant populations several *Veillonella* ASVs were under positive selection, however the positive selection of *Bacteroides* species as previously reported in Finnish and Bolivian infants [31] was absent in our study (Figure 4B).

Deterministic factors driving microbial community structure

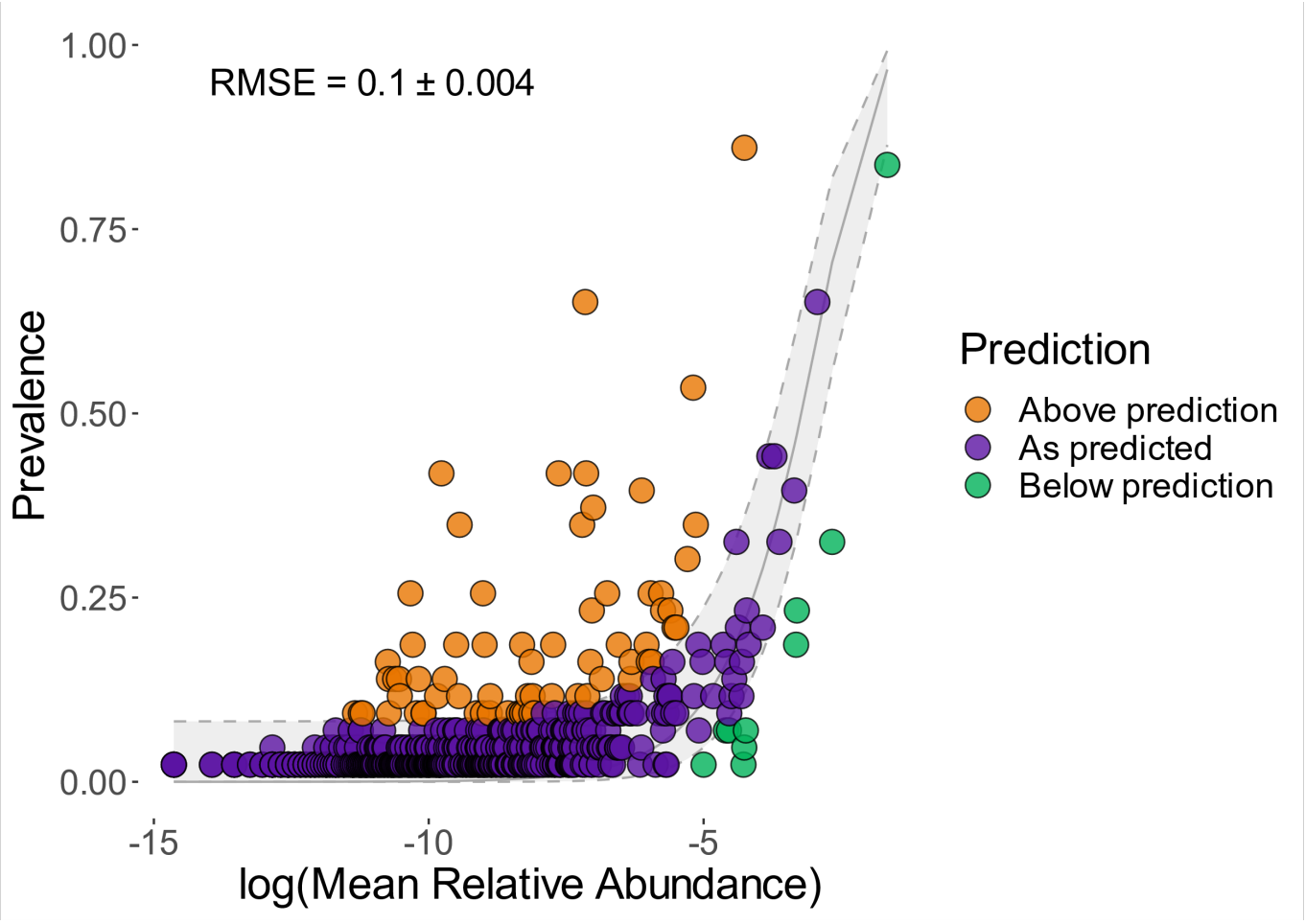
We next examined the impact of perinatal, lifestyle and dietary factors on the overall microbial community structure. These analyses revealed that in early infancy both mode and place of delivery had a strong impact on the microbial community structure. The impact of place of delivery (home vs. hospital) and subsequent hospitalization only had a short-term impact as shown by their significant association with the microbial community structure at 1-2 weeks and 4 weeks post-partum respectively (Figure 5A-J). In addition, multivariable linear regression analysis, showed that hospital delivery was also associated with a significantly lower microbial richness at 1-2 weeks (beta = -21.876 [95%CI = -38.385 to -5.367], p = 0.01) and 4 weeks post-partum (beta = -8.78 [-17.139 to -0.421], p = 0.04) when compared to home-delivery (Supplementary Table S1). Our previous analyses showed that vaginal delivery not only resulted in sharing of ASVs between mothers and infants, but also was associated with the DMM clusters in early life. Our permutational multivariate analyses of variance confirmed the significant impact of birth mode on the overall microbial community structure. It was shown that its

effect remained until the age of 6 months and was thus far more persistent than the impact of birthplace. However, as from the age of 4 weeks post-partum, dietary factors had a stronger influence on the infant microbial community structure than birth mode. Microbial richness and diversity did not appear to differ between infants born vaginally or by C-section (Supplementary Table S1 and S2). When examining infant feeding, the microbial community structure was both impacted by breastfeeding and formula feeding as indicated by *envfit* analyses (Figure 5 A-I). After, performing multivariate analyses, it however became apparent that breastfeeding was more strongly affected and more persistently associated with the microbial community structure up to the age of 11 months (Figure 5J). Formula feeding on the other hand was significantly associated with an increased microbial richness and diversity up to the age of 4 months and 8 weeks post-partum, respectively (Supplementary Table S1 and S2).

Children that already received solid foods by the age of 4 months had a significantly different microbial community structure as compared to children that were still fully breast- and or formula-fed. This impact was no longer apparent by the age of 5 months post-partum, likely because by that age the vast majority of children already were introduced to some form of solid foods (Figure 5J). However, the dietary diversity index as a marker of the complexity of foods introduced by that age did significantly impact the microbial community structure by the age of 5 months post-partum. In accordance, the dietary diversity at the age of 5 months was also associated with a significantly higher microbial diversity as indicated by the Shannon index (Supplementary Table S2, beta = 0.259 [0.013 – 0.5050], p = 0.04). A similar association between dietary and microbial diversity was also observed at the ages of 9 (beta = 0.09 [-0.001 – 0.18], p = 0.051) and 11 months (beta = 0.109 [0.002 – 0.216], p = 0.046).

Stool consistency as measured by the Bristol Stool Scale (BSS) was associated with the microbial community structure at 5 months of age (Figure 5J). A higher BSS (looser stools) was furthermore associated with decreased microbial richness and diversity at the ages of 5 and 6 months (Supplementary Table S1 and S2). Finally, from the age of 9 months onwards the presence of older siblings significantly affected the microbial community structure and also increased microbial richness and diversity. However, this was only observed to be significant at the age of 11 months (Supplementary Table S1 and S2).

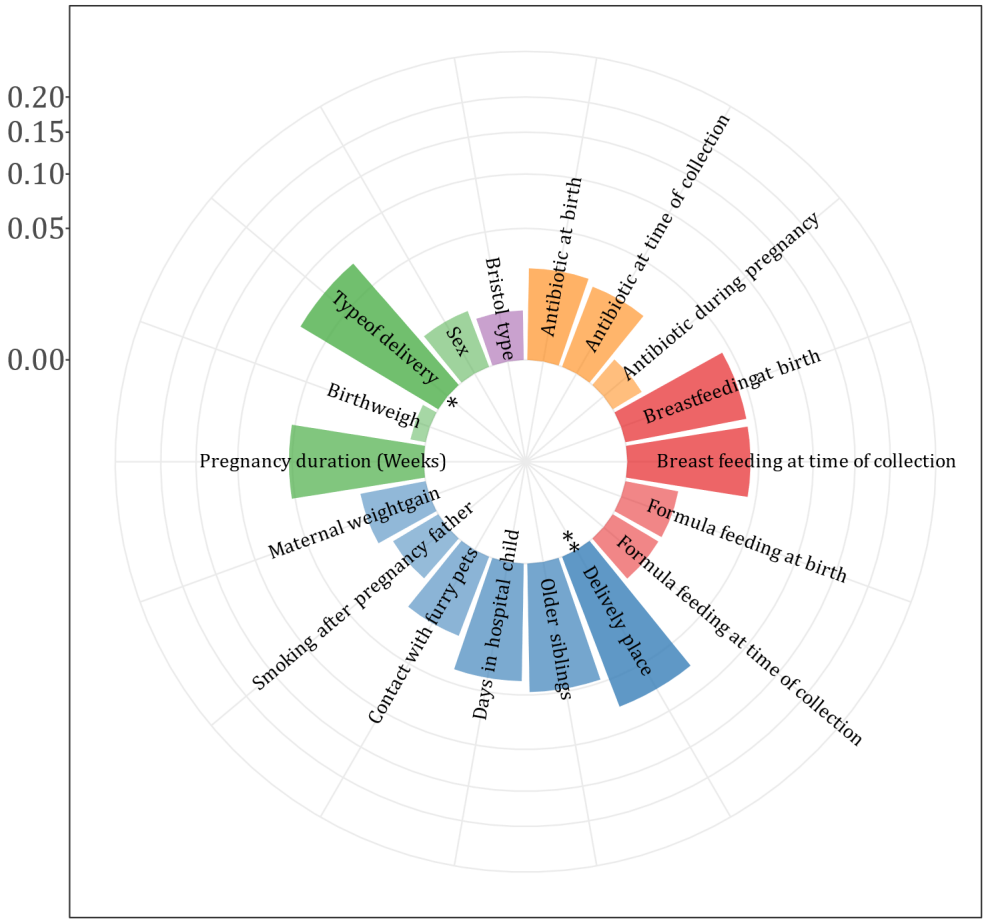
A


RMSE = 0.1 ± 0.004
Prevalence
1.00
0.75
0.50
0.25
0.00
-15
-10
-5
log(Mean Relative Abundance)
Prediction
Above prediction
As predicted
Below prediction

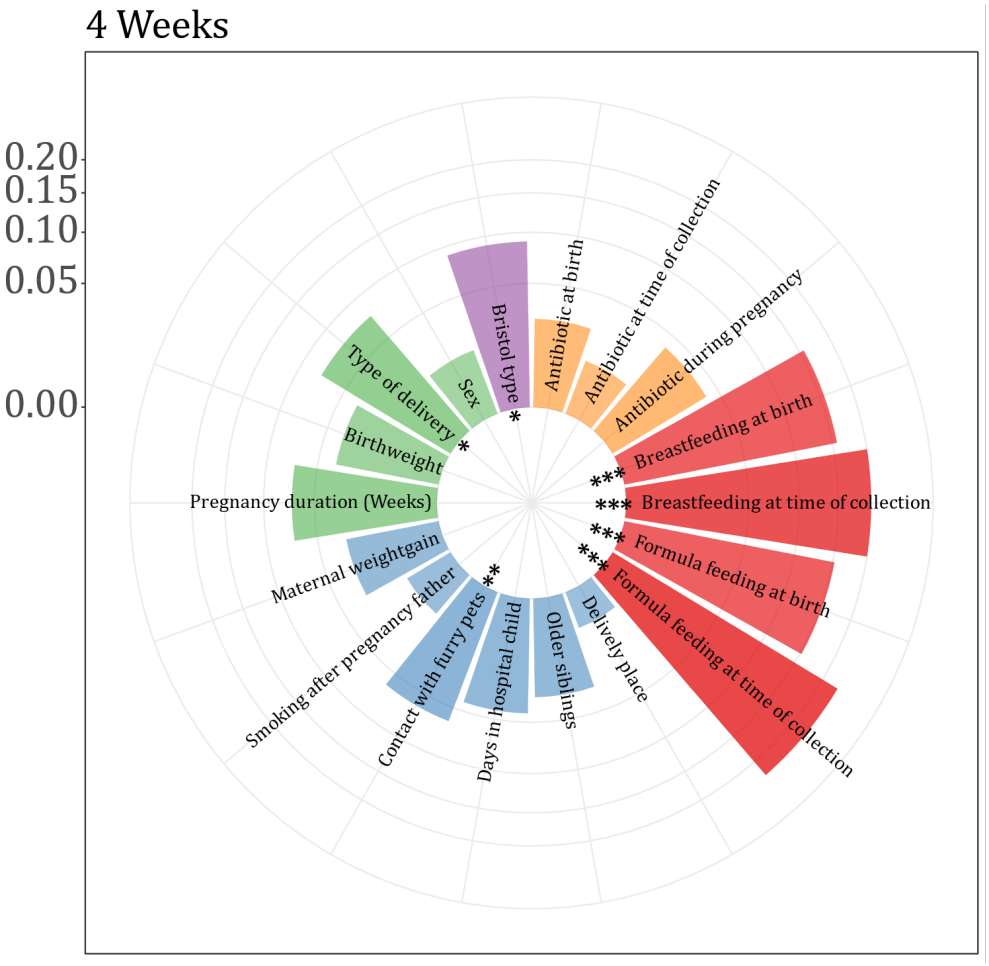
B
Age (Days)
ASV 16 Streptococcus salivarius
ASV 14 Streptococcus infantis
ASV 22 Enterococcus faecalis
ASV 56 Haemophilus parainfluenzae
ASV 46 Intestinibacter bartlettii
ASV 10 Veillonella dispar
ASV 81 Collinsella aerofaciens
ASV 50 Staphylococcus epidermidis
ASV 502 Gemella haemolysans
ASV 310 Streptococcus GT732452
ASV 838 Bifidobacterium breve
ASV 62 Staphylococcus aureus
ASV 57 Clostridium ramosum
ASV 26 Veillonella dispar
ASV 177 Clostridium neonatale
ASV 349 Caproiciproducens PAC001597
ASV 355 Finegoldia magna
ASV 283 Granulicatella adiacens
ASV 850 Escherichia coli
ASV 90 Eubacterium dolichum
ASV 246 Streptococcus parasanguinis
ASV 29 Clostridium citroniae
ASV 252 Bifidobacterium pseudocatenulatum
ASV 61 Clostridium celatum
ASV 122 Clostridium paraputrificum
ASV 324 Anaerococcus vaginalis
ASV 837 Corynebacterium pseudodiphtheriticum
ASV 487 Streptococcus AFUC
ASV 160 Haemophilus parainfluenzae
ASV 1 Flavonifractor plautii
ASV 190 Veillonella dispar
ASV 135 Enterococcus avium
ASV 768 Rothia mucilaginosa
ASV 343 Bifidobacterium adolescentis
ASV 262 Eggerthella lenta
ASV 147 Clostridioides difficile
ASV 127 Veillonella dispar
ASV 159 Enterococcus gilvus
ASV 563 Mixta theicola
ASV 919 Lancefieldella JRND
ASV 313 Veillonella dispar
ASV 514 Peptoniphilus grossensis
ASV 864 Enterococcus faecalis
ASV 3800 Blautia wadsworthia
ASV 395 Eggerthella lenta
ASV 244 Veillonella dispar
ASV 225 Clostridium perfringens
ASV 2 Bifidobacterium longum
ASV 6 Bacteroides fragilis
Age (Days)
350
50
Prediction
Above prediction
Below prediction
Prevalence
70
20
4
3
2
1
0
Prevalence
Prediction

Figure 4 Impact of neutral processes in infant stool microbiota assembly **(A)** Neutral community model fitted to each ASV observed in the infant stool samples within the LucKi Gut cohort. Points are coloured according to whether the taxon prevalence in infant samples was above (yellow), at (purple), or below (green) the predicted prevalence according to the Neutral community model. Average RMSE (±standard deviation) was calculated from 1000 bootstrap resamplings. **(B)** Heatmap of the log10 normalized abundances of the ASVs from infant stool samples that were observed either to be consistently above (yellow) or below (green) their predicted prevalence in the Neutral community model. Rows are sorted by taxa prevalence; columns are sorted by the subject's age (months)

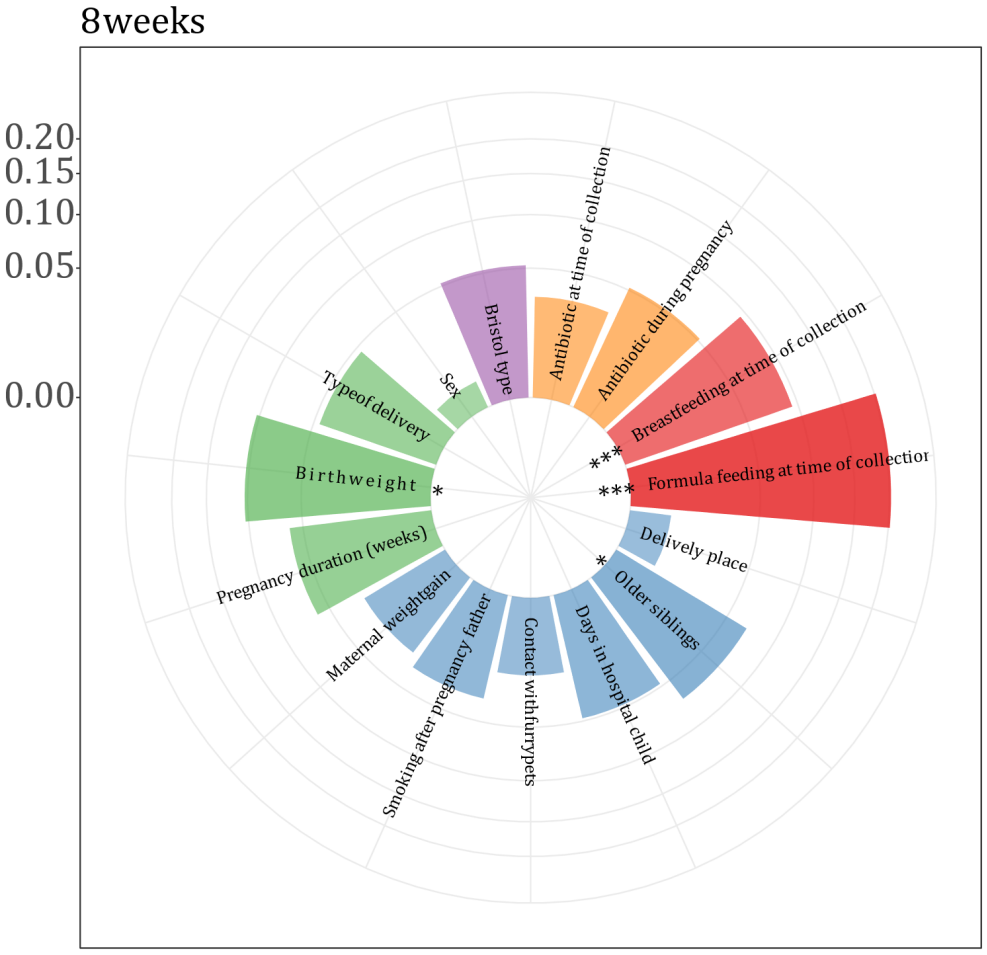
A


1-2 Weeks
0.20
0.15
0.10
0.05
0.00
Bristol type
Antibiotic at birth
Antibiotic at time of collection
Antibiotic during pregnancy
Breastfeedingat birth
Breast feeding at time of collection
Formula feeding at birth
Formula feeding at time of collection
Delively place
Older siblings
Days in hospital child
Contact with furry pets
Smoking after pregnancy father
Maternal weightgain
Pregnancy duration (Weeks)
Birthweigh
Typeof delivery
Sex
*
**

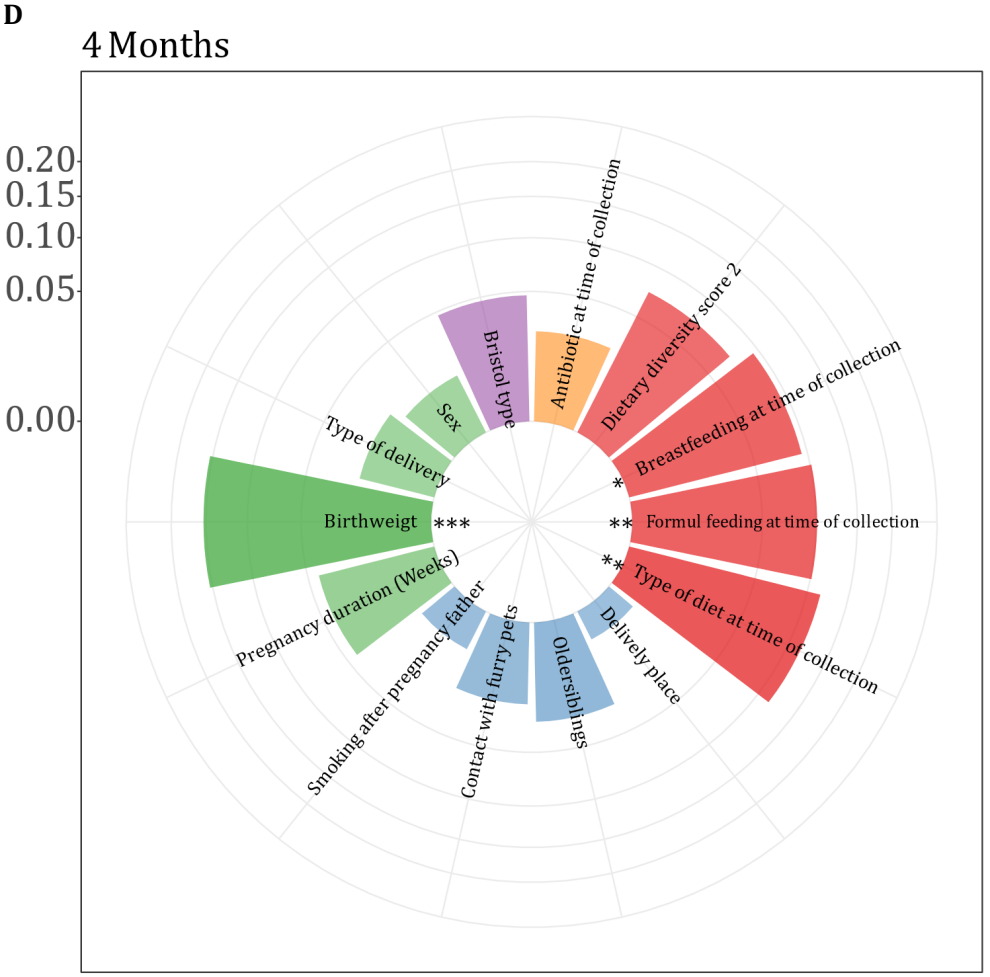
B


4 Weeks
0.20
0.15
0.10
0.05
0.00
Bristol type
*
Antibiotic at birth
Antibiotic at time of collection
Antibiotic during pregnancy
Breastfeeding at birth
***
Breastfeeding at time of collection
***
Formula feeding at birth
***
Formula feeding at time of collection
Delively place
Older siblings
Days in hospital child
Contact with furry pets
**
Smoking after pregnancy father
Maternal weightgain
Pregnancy duration (Weeks)
Birthweight
*
Type of delivery
Sex

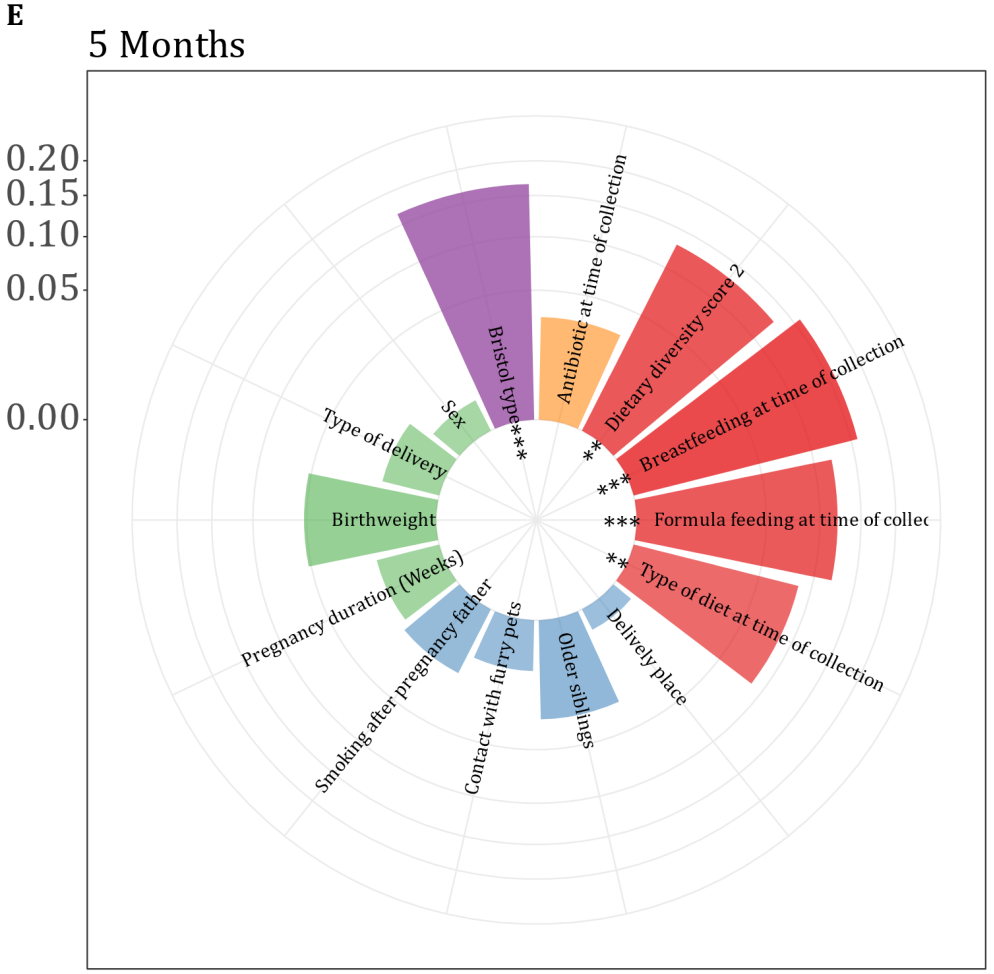
c


8weeks
0.20
0.15
0.10
0.05
0.00
Bristol type
Antibiotic at time of collection
Antibiotic during pregnancy
Breastfeeding at time of collection
Formula feeding at time of collectior
Delively place
Older siblings
Days in hospital child
Contact withfurrypets
Smoking after pregnancy father
Maternal weightgain
Pregnancy duration (weeks)
Birthweight
Typeofdelivery
Sex
***
***
*
*
*

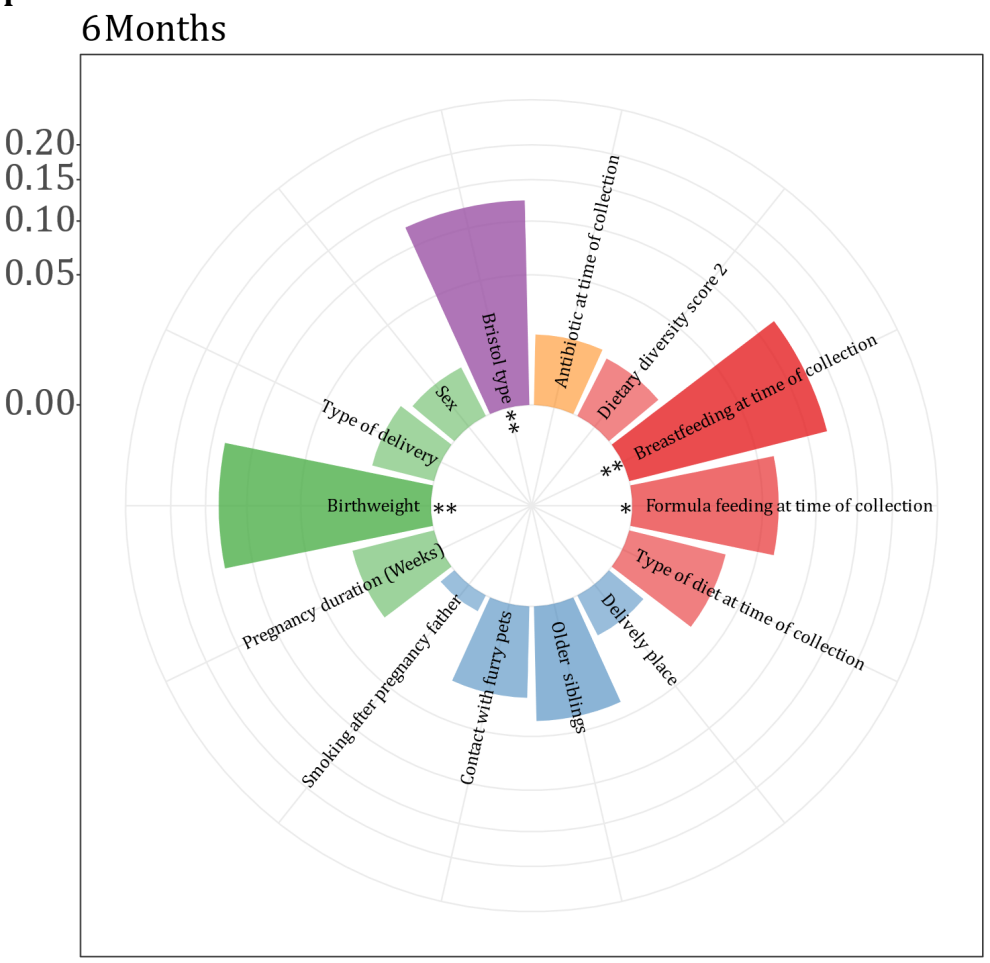
D


4 Months
0.20
0.15
0.10
0.05
0.00
Antibiotic at time of collection
Dietary diversity score 2
Breastfeeding at time of collection
Formul feeding at time of collection
Type of diet at time of collection
Delively place
Oldersiblings
Contact with furry pets
Smoking after pregnancy father
Pregnancy duration (Weeks)
Birthweigt
Type of delivery
Sex
Bristol type
*
**
**
***

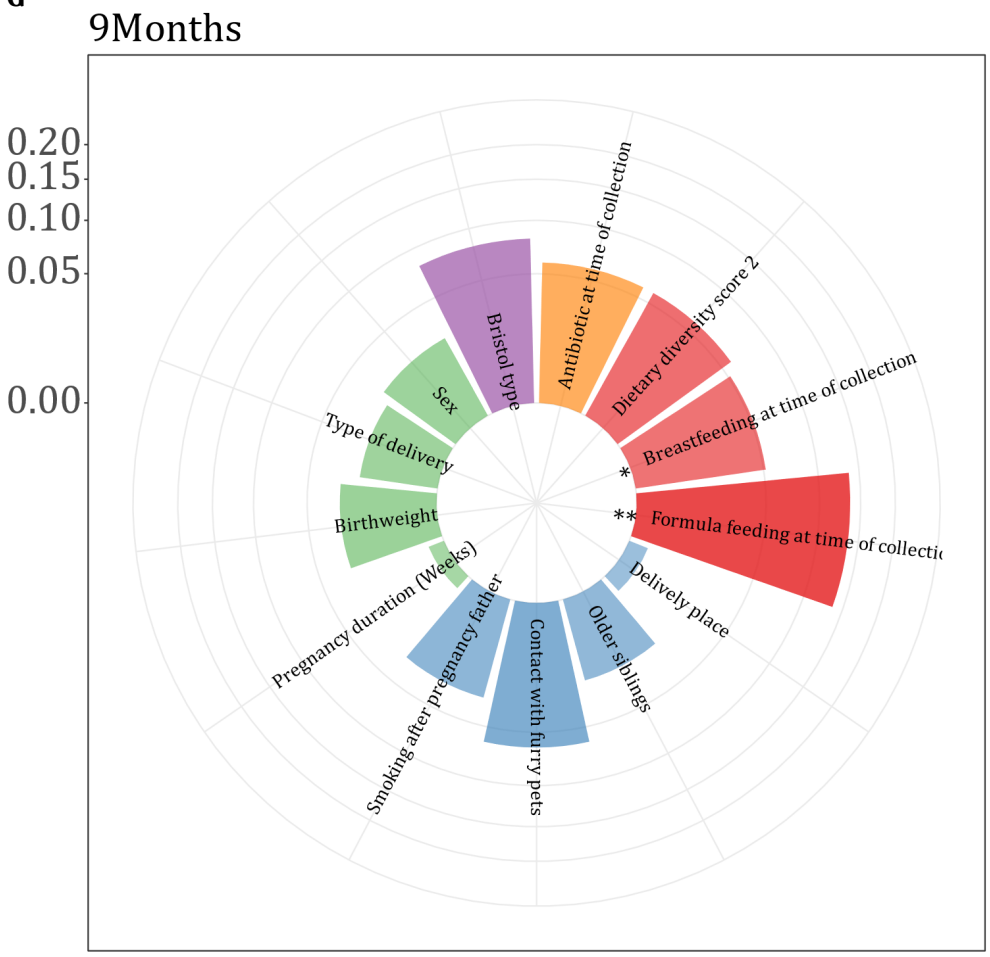
E


5 Months
0.20
0.15
0.10
0.05
0.00
Bristol type***
Antibiotic at time of collection
Dietary diversity score 2
**
Breastfeeding at time of collection
***
Formula feeding at time of collec
***
Type of diet at time of collection
**
Delively place
Older siblings
Contact with furry pets
Smoking after pregnancy father
Pregnancy duration (Weeks)
Birthweight
Type of delivery
Sex

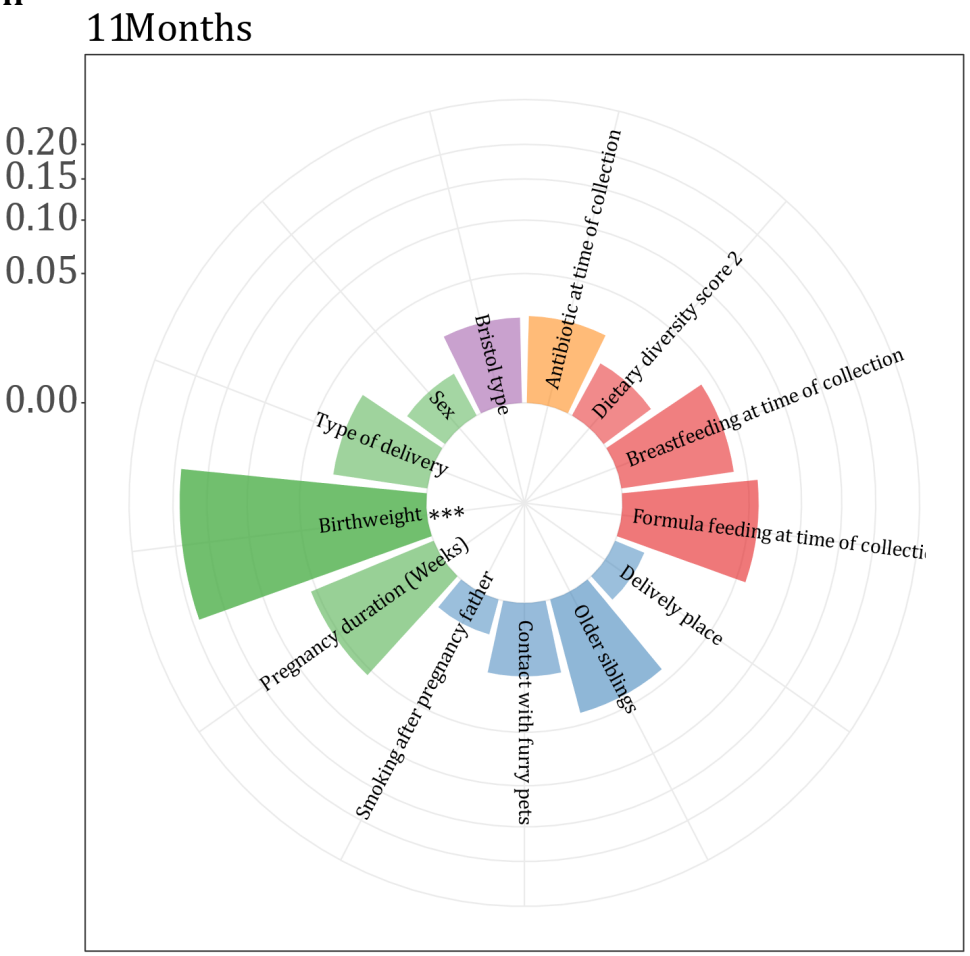
F


6 Months
0.20
0.15
0.10
0.05
0.00
Bristol type
Antibiotic at time of collection
Dietary diversity score 2
Breastfeeding at time of collection
Formula feeding at time of collection
Type of diet at time of collection
Delively place
Older siblings
Contact with furry pets
Smoking after pregnancy father
Pregnancy duration (Weeks)
Birthweight
Type of delivery
Sex
**
**
*
**

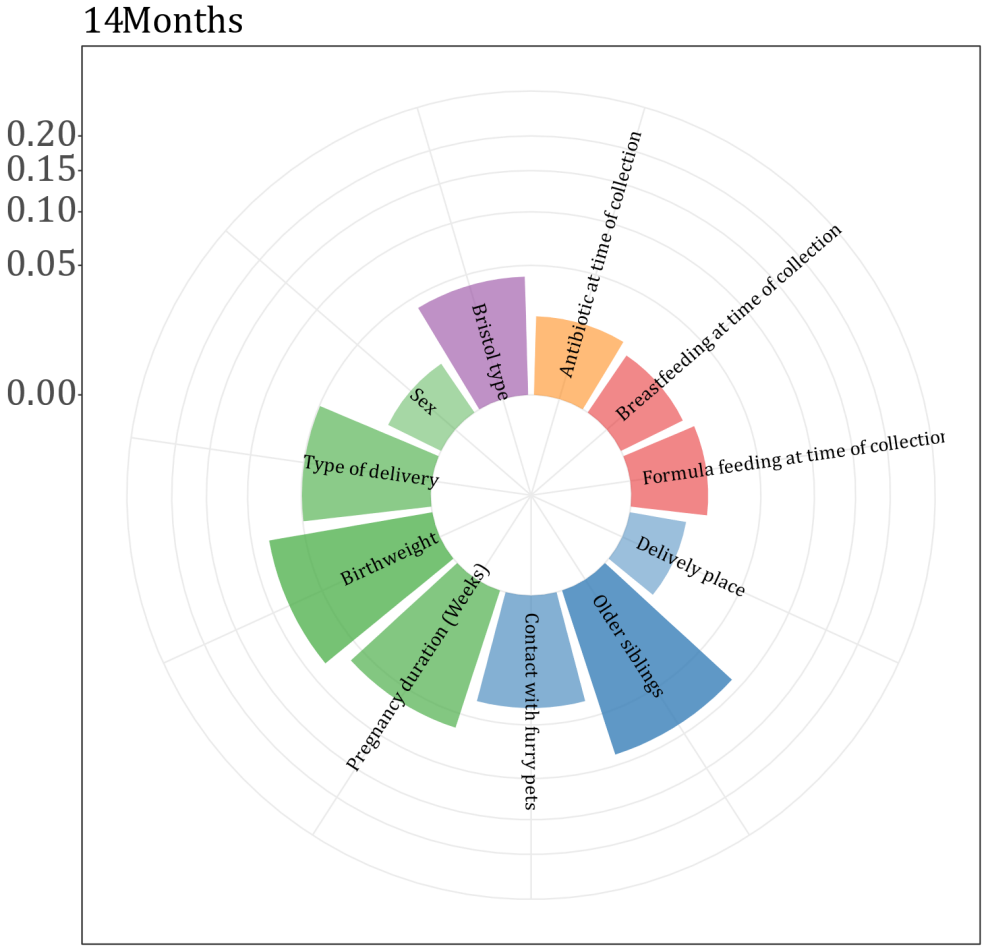
G


9Months
0.20
0.15
0.10
0.05
0.00
Bristol type
Antibiotic at time of collection
Dietary diversity score 2
Breastfeeding at time of collection
*
**
Formula feeding at time of collecti
Delively place
Older siblings
Contact with furry pets
Smoking after pregnancy father
Pregnancy duration (Weeks)
Birthweight
Type of delivery
Sex

H


11Months
0.20
0.15
0.10
0.05
0.00
Antibiotic at time of collection
Dietary diversity score 2
Breastfeeding at time of collection
Formula feeding at time of collecti
Delively place
Older siblings
Contact with furry pets
Smoking after pregnancy father
Pregnancy duration (Weeks)
Birthweight ***
Type of delivery
Sex
Bristol type

I


14Months
0.20
0.15
0.10
0.05
0.00
Antibiotic at time of collection
Breastfeeding at time of collection
Formula feeding at time of collectior
Delively place
Older siblings
Contact with furry pets
Pregnancy duration (Weeks)
Birthweight
Type of delivery
Sex
Bristol type

J

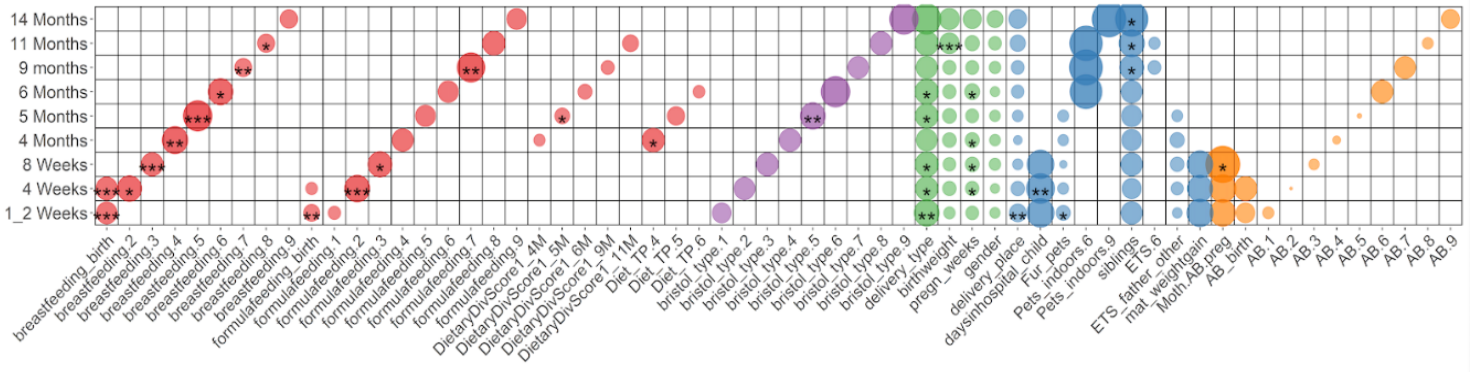


Figure 5 Deterministic influence of perinatal, lifestyle and dietary factors on microbiota community structure.
A-I, Polar plots visualizing the amount of variance of microbial communities at 1-2 weeks **(*A*)**, 4 weeks **(*B*)**, 8 weeks **(*C*)**, 4 months **(*D*)**, 5 months **(*E*)**, 6 months **(*F*)**, 9 months **(*G*)**, 11 months **(*H*)**, and 14 months **(*I*)** postpartum that could be explained by covariates as analysed using *Envfit*. The height of the bars reflects the amount of variance (r2) explained by each covariate. Covariates are coloured to highlight antibiotic use (orange), environment (blue), diet (red), stool consistency (purple), environmental (blue), and perinatal covariates (green). Asterisks indicate significant covariates (false discovery rate (FDR) P< 0.05) at each time point. **(*J*)**, Permutational Multivariate Analysis of Variance (PERMANOVA) combining all covariates that were significantly associated with microbial community variation at any given time point in the *Envfit* analyses. The size of the dots reflects the R2. Only samples without missing data on the included covariates were included in PERMANOVA. Asterisks indicate statistical significance with *p<0.05, **p<0.01, ***p<0.001.

Impact of dietary patterns on the microbial community structure

As our previous analyses revealed a substantial proportion of infants at the age of 9 months transitioning to DMM clusters 3 and 5, we next examined in more detail to what extent this was driven by differences in dietary patterns focusing on solid foods. We first performed PCA to reduce the high dimensional dietary data and identify dietary patterns and identified 3 components that together explained 35% of the variance in dietary patterns (15.1%, 11.3%, 8.8% for components 1, 2 and 3 respectively). The first component was characterized by meat (beef, pork, and chicken), fish and pasta and rice and points towards a more "mature omnivore diet" (Figure 6B). Components 2 and 3, both characterized by health-conscious foods, including a mixture of fruits and vegetables (Figure 6C-D).

Next, we used the scores of each infant on these first three components as explanatory variables in a redundancy analysis to examine the impact of dietary patterns on the microbial community structure. Dietary component 1 ("mature omnivores diet") clearly separated the microbial community structure along the first axis in the RDA (Figure 6A) with a high score on dietary component 1 being associated with increased levels of *F. prausnitzii* and reduced levels of *Enterococcus spp.* and *Stapylococcus spp.* Dietary component 3 was mainly explaining variation along the second RDA axis. *Envfit* analysis revealed that dietary component 1 was significantly associated with the microbial community structure (data not shown) and also with an increase in microbial diversity (beta = 0.142 [0.027-0.258], p = 0.0017).

Strikingly, the separation along the first RDA axis, driven by dietary component 1, almost perfectly matched the separation into the various DMM clusters. Together this indicates that the transition from DMM Cluster 3 to 5 is largely driven by maturation in the infant's diet.

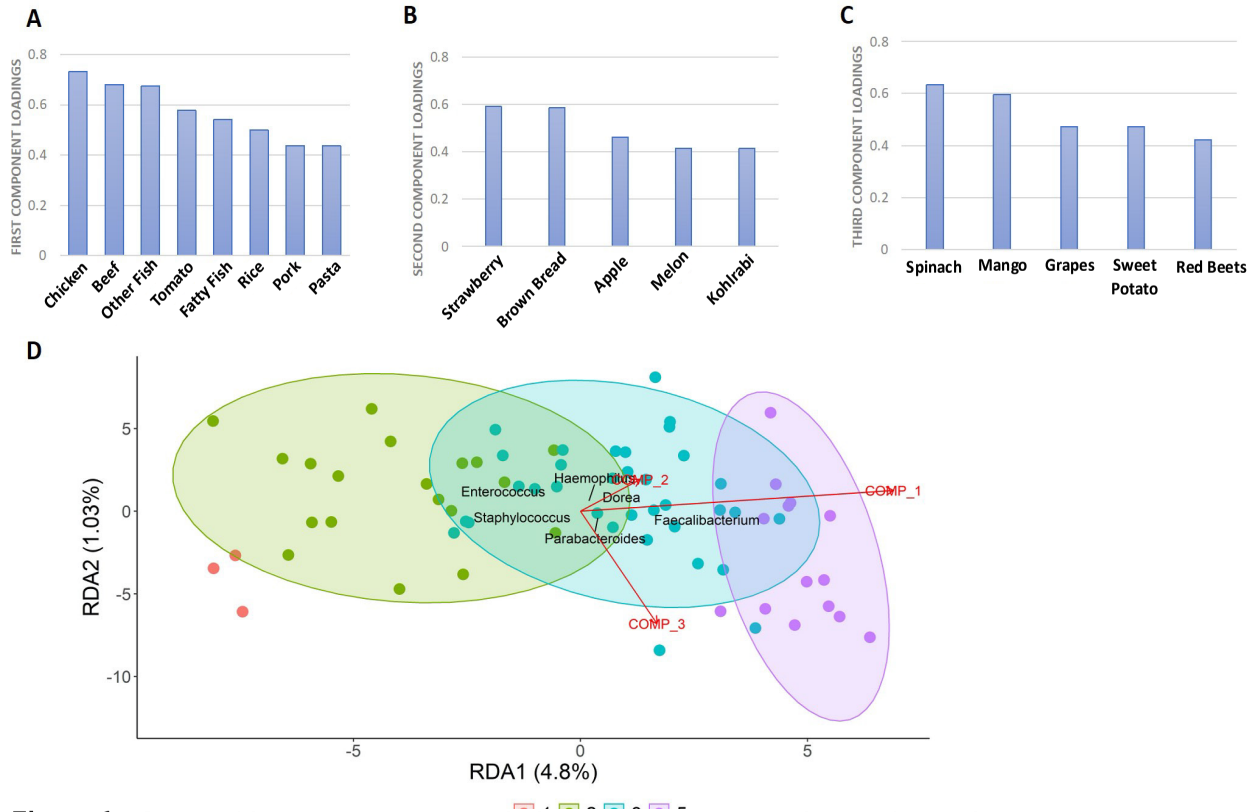


Figure 6 Influence of dietary patterns on microbial community structure at the age of 9 months.
A-C Loadings of food items on the first (A), second (B) and third (C) component in a Categorical Principal Components Analyses on the dietary data of infants at the age of 9 months post-partum. Redundancy analysis plot based upon *clr*-transformed genus abundance data, including the 30 genera that contributed most to the DMM clustering (Fig. 2B), as response variables and the scores on the dietary PCA components as explanatory variables (D). Samples are coloured according to DMM cluster, ellipses indicate 95% confidence intervals

Discussion

Within the present study, we aimed to identify the processes and dynamics involved in the assembly of intestinal microbial communities throughout infancy. We observed a general increase in microbial diversity and shift in community structure throughout the first 14 months of life. The largest changes were observed between the age of 6- and 9-months postpartum. However, at the age of 14 months, the microbiota was still far from mature both in terms of diversity and composition as compared to the maternal microbiota.

At the earliest time-point of 1-2 weeks postpartum, mode of delivery had the strongest impact on the microbial community structure. In line with previous studies [42-44] reduced levels of Bacteroides in C-section delivered infants suggested a disrupted transmission of maternal Bacteroides strains. Indeed, the microbiota of infants, as compared to that of their mothers, revealed a higher similarity in the microbial communities between infants and their own (as compared to unrelated) mothers, but only when born via vaginal delivery. Focusing on the ASVs within the Bacteroides and Parabacteroides genera showed an even more striking difference with a complete lack of shared ASVs between mothers and C-section delivered infants at the age of 1-2 weeks postpartum. While some ASVs became shared between C-section delivered infants and their mothers at a later stage during infancy, for other ASVs sharing between mother-infant dyads in the C-section delivered infants remained sporadic. Altogether, these results point towards a role of maternal faecal rather than vaginal bacteria as an important inoculum during vaginal delivery. This is indeed consistent with previous studies [42-44] and might explain the lack of persistent effects of vaginal seeding, a procedure in which C-section delivered infants are being exposed to maternal vaginal microbes [45]. A recent proof-of-principle study indeed showed that maternal faecal microbiota transplantation to restore the microbiota in C-section delivered infants [46] by engraftment of maternal faecal Bacteroides strains might hold more promise. Of note, our DMM-clustering analysis showed that part of the vaginally delivered infants had a microbiota more similar to that of C-section delivered newborns. This calls for further studies to unravel what determines successful engraftment of Bacteroides and other maternal strains during natural delivery.

Previous studies have suggested that cessation of breastfeeding is more important for microbiota maturation than the introduction of solid foods [41, 47], however it is questionable whether these studies could properly disentangle these effects given the limited number of samples collected during the weaning period.

At the age of 4 months postpartum, we found that both breastfeeding as well as the introduction of solid foods both had a significant impact on the microbial community structure. Moreover, the diversity of introduced food items had a significant impact on the microbial community structure at the age of 5 months post-partum and persisted to have an effect on the diversity of the microbiota up to the age of 11 months post-partum. Together our results demonstrate that while the presence of human milk oligosaccharides has a strong impact on environmental filtering by supporting the growth of Bifidobacterium spp., and repressing the growth of many other microorganisms, it is the diversity and complexity of solid food items being introduced which drives further microbiota maturation. The impact of the weaning process as a main driver of the microbi-

al maturation was further underscored by the large shift in the microbiota between the age of 6- and 9-months post-partum, although it should be noted that this could also be partly explained by the larger time-interval between subsequent samples as compared to earlier time-points. Many bacterial genera within the family Lachnospiraceae (e.g., Blautia, Coprococcus) as well as Faecalibacterium prausnitzii increase in relative abundance at the expense of facultative anaerobic bacteria such as staphylococci and enterococci. The transition analysis based on the DMM-clustering confirmed these findings and showed that in particular cluster 5 was characterized by high levels of Faecalibacterium and Lachnospiraceae. Members of the Lachnospiraceae family are well-known to be main short-chain fatty acid-producers and genomic analysis has shown a considerable capacity to metabolise diet-derived polysaccharides (e.g., starch, inulin) with substantial variability among different species and strains [48, 49]. Faecalibacterium prausnitzii, a major butyrate-producer, has also been shown to thrive on complex indigestible oligosaccharides such as resistant maltodextrin, inulin and polydextrose [50].

To further confirm the impact of dietary patterns, we performed factor analysis on the food items that were being consumed by the infants at the age of 9 months and demonstrated that infants receiving a diet rich in meat, fish, rice, pasta had the most mature microbiota composition with high levels of amongst others Faecalibacterium prausnitizii. Our results are in strong agreement with the study of Laursen et. al., in which a PCA was also performed on the dietary data collected at the age of 9 months [51]. In line with our study, the authors reported that the first component was driven by the consumption of family-foods with high loadings of meat, animal fat, but also pasta/rice and fish. Strikingly, similar negative and positive correlations were found with Enterococcaceae and Lachnospiraceae, respectively. Together these results suggest that the progression from early infant food to a more mature diet with higher protein and fibre contents is a major driver of gut microbial maturation during late infancy.

Many deeply phenotyped population-based (birth) cohort studies, including ours, have revealed important determinants of microbiota composition yet can only explain a limited portion of the interindividual microbiota variation [15, 16, 52]. What is often ignored in many studies is that not all variation among host-associated microbial communities needs to be caused by differences among hosts or their microorganisms [53]. Neutral community assembling theory assumes that there is an equal growth, death, and dispersal of species, i.e., species have equal ecological fitness. Under such circumstances the assembly of microbial communities is the result of stochastic processes of dispersal and drift. Although the current neutral assembly models are oversimplistic, they have been applied to successfully predict the structures of various microbial ecosystems [31, 54, 55]. In line with the study by Sprockett and colleagues [31], we showed that the vast majority of ASVs were neutrally distributed in the NCM indicating that neutral dispersal is an important force shaping microbial community structure. It should however be noted that we randomly selected a single sample per infant for the NCM, while recent studies in zebrafish showed that the importance of neutral processes declines as hosts mature [53]. Sprockett et. al., also showed that the role of neutral processes was significantly decreased in adults, suggesting that non-neutral processes (e.g., microbe-microbe interactions, active dispersal or selection by the host) become more important with changing lifestyle factors [31]. It would therefore be of interest to examine how the influence of neutral processes changes throughout infancy and

childhood.

A limitation of the present study is the lack of details on macro- and micronutrient intake as well as information on intake of specific dietary fibres. We collected information on specific food items being introduced and aimed to identify different dietary profiles, but in the future these data could be used to gain more insight into which specific dietary components might drive microbiota development. Despite the high frequency of sample collection within each individual child, the number of children included in our study is still relatively limited and hampered the identification of other perinatal determinants of microbiota development. In particular, dispersal of microbial species from the metacommunity deserves further exploration. Especially since we showed that having older siblings impacted the microbial community structure of infants from the age of 9 months onwards. But also, because recent studies have identified connections to natural environments and green spaces [56, 57] but also the built environment [58] as important sources of microbial dispersion. Larger birth cohort studies with a similar narrow sampling time-intervals are therefore needed to further explore the relative contribution of neutral processes, such as dispersal and stochasticity. Such studies should also aim to sample this so-called metacommunity, for example by profiling the microbiome of family members, pets, day care centres and the environment, in order to directly track the transmission of microbial strains.

Altogether, within the present study we have shown that microbiota assembly is a dynamic process, influenced by birth mode, diet and dispersal from household members but also to a large extent by neutral processes, that is far from being completed at 14 months of age.

Increasing our knowledge on the processes that shape the microbiota during the critical infant time-window is pivotal to understand the mechanisms underlying inter-individual heterogeneity in microbiota composition. Finally, future inshight will be crucial to find new leads for microbiota-based interventions in the primary and secondary prevention of non-communicable diseases.

References

1. Dethlefsen, L., et al., Assembly of the human intestinal microbiota. Trends Ecol Evol, 2006. 21(9): p. 517-23.
2. Turnbaugh, P.J. and J.I. Gordon, The core gut microbiome, energy balance and obesity. J Physiol, 2009. 587(Pt 17): p. 4153-8.
3. Penders, J., et al., The role of the intestinal microbiota in the development of atopic disorders. Allergy, 2007. 62(11): p. 1223-36.
4. Manichanh, C., et al., The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol, 2012. 9(10): p. 599-608.
5. Konstantinov, S.R., Diet, microbiome, and colorectal cancer. Best Pract Res Clin Gastroenterol, 2017. 31(6): p. 675-681.
6. Costello, E.K., et al., The application of ecological theory toward an understanding of the human microbiome. Science, 2012. 336(6086): p. 1255-62.
7. Martinez, I., C.E. Muller, and J. Walter, Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. PLoS One, 2013. 8(7): p. e69621.
8. Huttenhower, C., et al., Structure, function and diversity of the healthy human microbiome. Nature, 2012. 486(7402): p. 207-214.
9. Palmer, C., et al., Development of the human infant intestinal microbiota. PLoS Biol, 2007. 5(7): p. e177.
10. Wu, G.D., et al., Linking long-term dietary patterns with gut microbial enterotypes. Science, 2011. 334(6052): p. 105-8.
11. Yatsunenko, T., et al., Human gut microbiome viewed across age and geography. Nature, 2012. 486(7402): p. 222-7.
12. Davenport, E.R., Elucidating the role of the host genome in shaping microbiome composition. Gut Microbes, 2016. 7(2): p. 178-84.
13. Goodrich, J.K., et al., Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe, 2016. 19(5): p. 731-43.
14. Rothschild, D., et al., Environment dominates over host genetics in shaping human gut microbiota. Nature, 2018. 555(7695): p. 210-215.
15. Falony, G., et al., Population-level analysis of gut microbiome variation. Science, 2016. 352(6285): p. 560-4.
16. Zhernakova, A., et al., Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science, 2016. 352(6285): p. 565-9.
17. Dominguez-Bello, M.G., et al., Development of the human gastrointestinal microbiota and insights from high-throughput sequencing. Gastroenterology, 2011. 140(6): p. 1713-9.
18. Azad, M.B., et al., Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: a prospective cohort study. BJOG, 2016. 123(6): p. 983-93.
19. Penders, J., et al., New insights into the hygiene hypothesis in allergic diseases: mediation of sibling and birth mode effects by the gut microbiota. Gut microbes, 2014. 5(2): p. 239-244.
20. Penders, J., et al., Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics, 2006. 118(2): p. 511-21.
21. Renz, H., P. Brandtzaeg, and M. Hornef, The impact of perinatal immune development on mucosal homeostasis and chronic inflammation. Nat Rev Immunol, 2011. 12(1): p. 9-23.
22. Azad, M.B., et al., Infant gut microbiota and the hygiene hypothesis of allergic disease: impact of household pets and siblings on microbiota composition and diversity. Allergy Asthma Clin Immunol, 2013. 9(1): p. 15.
23. Penders, J., et al., Establishment of the intestinal microbiota and its role for atopic dermatitis in early childhood. J Allergy Clin Immunol, 2013. 132(3): p. 601-607 e8.
24. Tun, H.M., et al., Exposure to household furry pets influences the gut microbiota of infant at 3-4 months following various birth scenarios. Microbiome, 2017. 5(1): p. 40-40.
25. van Best, N., et al., On the origin of species: Factors shaping the establishment of infant's gut microbiota. Birth Defects Res C Embryo Today, 2015. 105(4): p. 240-51.
26. De Filippo, C., et al., Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A, 2010. 107(33): p. 14691-6.
27. Schnorr, S.L., et al., Gut microbiome of the Hadza hunter-gatherers. Nat Commun, 2014. 5: p. 3654.
28. Backhed, F., et al., Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe, 2015. 17(5): p. 690-703.
29. Faith, J.J., et al., The long-term stability of the human gut microbiota. Science, 2013. 341(6141): p. 1237439.
30. Sprockett, D., T. Fukami, and D.A. Relman, Role of priority effects in the early-life assembly of the gut microbiota. Nat Rev Gastroenterol Hepatol, 2018. 15(4): p. 197-205.
31. Sprockett, D.D., et al., Microbiota assembly, structure, and dynamics among Tsimane horticulturalists of the Bolivian Amazon. Nat Commun, 2020. 11(1): p. 3772.
32. de Korte-de Boer, D., et al., LucKi Birth Cohort Study: rationale and design. BMC Public Health, 2015. 15: p. 934.
33. Smithers, L.G., et al., Dietary patterns of infants and toddlers are associated with nutrient intakes. Nutrients, 2012. 4(8): p. 935-48.
34. Stearns, J.C., et al., Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age. ISME J, 2015. 9(5): p. 1246-59.

35. Bartram, A.K., et al., Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. Appl Environ Microbiol, 2011. 77(11): p. 3846-52.
36. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res, 2013. 41(Database issue): p. D590-6.
37. DeSantis, T.Z., et al., Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol, 2006. 72(7): p. 5069-72.
38. Fernandes, A.D., et al., Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome, 2014. 2: p. 15.
39. Jari Oksanen , F.G.B., Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner, vegan: Community Ecology Package. 2019.
40. Holmes, I., K. Harris, and C. Quince, Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One, 2012. 7(2): p. e30126.
41. Stewart, C.J., et al., Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature, 2018. 562(7728): p. 583-588.
42. Mitchell, C.M., et al., Delivery Mode Affects Stability of Early Infant Gut Microbiota. Cell Rep Med, 2020. 1(9): p. 100156.
43. Shao, Y., et al., Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature, 2019. 574(7776): p. 117-121.
44. Yassour, M., et al., Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med, 2016. 8(343): p. 343ra81.
45. Dominguez-Bello, M.G., et al., Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nat Med, 2016. 22(3): p. 250-3.
46. Korpela, K., et al., Maternal Fecal Microbiota Transplantation in Cesarean-Born Infants Rapidly Restores Normal Gut Microbial Development: A Proof-of-Concept Study. Cell, 2020. 183(2): p. 324-334 e5.
47. Heath, G.W. and J.S. Kendrick, Outrunning the risks: a behavioral risk profile of runners. Am J Prev Med, 1989. 5(6): p. 347-52.
48. Vacca, M., et al., The Controversial Role of Human Gut Lachnospiraceae. Microorganisms, 2020. 8(4).
49. P, O.S., et al., Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes. Microb Genom, 2016. 2(2): p. e000043.
50. Ganesan, K., et al., Causal Relationship between Diet-Induced Gut Microbiota Changes and Diabetes: A Novel Strategy to Transplant Faecalibacterium prausnitzii in Preventing Diabetes. Int J Mol Sci, 2018. 19(12).
51. Laursen, M.F., et al., Infant Gut Microbiota Development Is Driven by Transition to Family Foods Independent of Maternal Obesity. mSphere, 2016. 1(1).
52. Galazzo, G., et al., Development of the Microbiota and Associations With Birth Mode, Diet, and Atopic Disorders in a Longitudinal Analysis of Stool Samples, Collected From Infancy Through Early Childhood. Gastroenterology, 2020. 158(6): p. 1584-1596.
53. Burns, A.R., et al., Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. ISME J, 2016. 10(3): p. 655-64.
54. Venkataraman, A., et al., Application of a neutral community model to assess structuring of the human lung microbiome. mBio, 2015. 6(1).
55. Ofiteru, I.D., et al., Combined niche and neutral effects in a microbial wastewater treatment community. Proc Natl Acad Sci U S A, 2010. 107(35): p. 15345-50.
56. Thorsen, J., et al., Evaluating the Effects of Farm Exposure on Infant Gut Microbiome. Journal of Allergy and Clinical Immunology, 2019. 143(2): p. AB299.
57. Nielsen, C.C., et al., Natural environments in the urban context and gut microbiota in infants. Environ Int, 2020. 142: p. 105881.
58. Sbihi, H., et al., Thinking bigger: How early-life environmental exposures shape the gut microbiome and influence the development of asthma and allergic disease. Allergy, 2019. 74(11): p. 2103-2115.

Supplementary Tables

Supplementary Table 1 Multivariable linear regression models on the association between dietary, lifestyle and other perinatal determinants and microbial richness (Chao1 index).

Age*	Model	Unstandardized Coef. B	Std. Error	95% Confidence Interval for B	Sig.
1-2 weeks	(Constant)	105.144	96.91	(-88.089, 298.376)	0.282
	Breastfed at timepoint	-9.402	10.172	(-29.684, 10.881)	0.358
	Formula fed at timepoint	3.563	6.894	(-10.183, 17.309)	0.607
	Antibiotic use prior to sampling	-9.363	14.697	(-38.668, 19.943)	0.526
	Older siblings (Yes/no)	1.718	6.006	(-10.257, 13.694)	0.776
	Birth mode (Vaginal/C-section)	-0.934	9.703	(-20.281, 18.412)	0.924
	Birthplace (Home/Hospital	-21.876	8.28	(-38.385, -5.367)	**0.01**
	Bristol Stool score	0.544	1.887	(-3.218, 4.306)	0.774
	Gestational age (weeks)	-1.43	2.815	(-7.044, 4.184)	0.613
	Birth weight (grams)	0	0.009	(-0.017, 0.018)	0.977
	Sex (Male/Female)	8.177	5.819	(-3.426, 19.78)	0.164
	Hospitalization after birth (days)	-0.361	1.596	(-3.543, 2.822)	0.822
4 weeks	(Constant)	-46.149	46.46	(-138.882, 46.585)	0.324
	Breastfed at timepoint	7.891	5.374	(-2.835, 18.617)	0.147
	Formula fed at timepoint	16.615	4.783	(7.069, 26.161)	**0.001**
	Antibiotic use prior to sampling	-33.791	12.968	(-59.675, -7.908)	**0.011**
	Older siblings (Yes/no)	4.852	2.859	(-0.854, 10.558)	0.094
	Birth mode (Vaginal/C-section)	0.556	4.559	(-8.544, 9.656)	0.903
	Birthplace (Home/Hospital	-8.78	4.188	(-17.139, -0.421)	**0.04**

*Separate regression models were run for each of the timepoints
**Infant feeding not included as all children were weaned

Supplementary Table 1 (cont'd)

	Bristol Stool score	-0.916	1.257	(-3.425, 1.592)	0.468
	Gestational age (weeks)	2.181	1.323	(-0.46, 4.822)	0.104
	Birth weight (grams)	-0.002	0.004	(-0.01, 0.006)	0.588
	Sex (Male/Female)	-1.544	2.885	(-7.303, 4.214)	0.594
	Hospitalization after birth (days)	0.063	0.878	(-1.69, 1.817)	0.943
8 weeks	(Constant)	43.134	61.171	(-78.867, 165.135)	0.483
	Breastfed at timepoint	-1.618	5.628	(-12.843, 9.607)	0.775
	Formula fed at timepoint	9.606	4.7	(0.233, 18.979)	**0.045**
	Antibiotic use prior to sampling	-12.378	15.931	(-44.151, 19.395)	0.44
	Older siblings (Yes/no)	4.402	3.436	(-2.451, 11.254)	0.204
	Birth mode (Vaginal/C-section)	0.191	4.969	(-9.72, 10.102)	0.969
	Birthplace (Home/Hospital	-4.169	5.26	(-14.659, 6.322)	0.431
	Bristol Stool score	-2.459	1.575	(-5.6, 0.681)	0.123
	Gestational age (weeks)	0.524	1.75	(-2.967, 4.015)	0.766
	Birth weight (grams)	-0.004	0.005	(-0.014, 0.006)	0.398
	Sex (Male/Female)	-1.368	3.569	(-8.486, 5.75)	0.703
	Hospitalization after birth (days)	-0.548	0.895	(-2.332, 1.237)	0.542
4 Months	(Constant)	51.113	35.047	(-18.84, 121.067)	0.149
	Breastfed at timepoint	-8.45	4.213	(-16.859, -0.042)	**0.049**
	Formula fed at timepoint	9.956	4.319	(1.335, 18.577)	**0.024**
	Antibiotic use prior to sampling	-11.287	10.064	(-31.375, 8.801)	0.266

Supplementary Table 1 (cont'd)

	Older siblings (Yes/no)	6.684	3.23	(0.238, 13.131)	0.042
	Birth mode (Vaginal/C-section)	5.307	4.76	(-4.194, 14.809)	0.269
	Birthplace (Home/Hospital	-0.316	4.467	(-9.232, 8.6)	0.944
	Bristol Stool score	-1.616	1.41	(-4.43, 1.198)	0.256
	Gestational age (weeks)	0.232	0.969	(-1.703, 2.166)	0.812
	Birth weight (grams)	-0.005	0.004	(-0.013, 0.004)	0.277
	Sex (Male/Female)	-0.916	3.151	(-7.205, 5.373)	0.772
	Hospitalization after birth (days)	0.925	0.876	(-0.823, 2.673)	0.295
	Solids introduced (Yes/no)	4.89	3.458	(-2.012, 11.792)	0.162
5 Months	(Constant)	-31.648	58.477	(-148.434, 85.138)	0.59
	Breastfed at timepoint	-10.067	5.45	(-20.951, 0.817)	0.069
	Formula fed at timepoint	8.07	5.426	(-2.766, 18.906)	0.142
	Antibiotic use prior to sampling	-31.242	9.93	(-51.073, -11.411)	**0.002**
	Older siblings (Yes/no)	1.444	3.82	(-6.184, 9.073)	0.707
	Birth mode (Vaginal/C-section)	6.035	5.61	(-5.169, 17.239)	0.286
	Birthplace (Home/Hospital	-2.944	5.424	(-13.776, 7.888)	0.589
	Bristol Stool score	-4.267	1.884	(-8.03, -0.505)	**0.027**
	Gestational age (weeks)	2.819	1.71	(-0.596, 6.234)	0.104
	Birth weight (grams)	-0.004	0.006	(-0.015, 0.007)	0.506
	Sex (Male/Female)	1.811	3.765	(-5.709, 9.331)	0.632
	Hospitalization after birth (days)	1.366	1.107	(-0.845, 3.576)	0.222

Supplementary Table 1 (cont'd)

	DietaryDivScore1	4.144	3.566	(-2.977, 11.265)	0.249
6 Months	(Constant)	10.124	72.326	(-134.408, 154.656)	0.889
	Breastfed at timepoint	-7.467	6.824	(-21.104, 6.169)	0.278
	Formula fed at timepoint	2.817	6.762	(-10.696, 16.33)	0.678
	Antibiotic use prior to sampling	-21.309	11.93	(-45.149, 2.531)	0.079
	Older siblings (Yes/no)	2.52	5.089	(-7.65, 12.689)	0.622
	Birth mode (Vaginal/C-section)	8.727	6.774	(-4.81, 22.264)	0.202
	Birthplace (Home/Hospital	4.116	7.344	(-10.559, 18.791)	0.577
	Bristol Stool score	-5.287	2.149	(-9.582, -0.991)	**0.017**
	Gestational age (weeks)	2.631	2.124	(-1.613, 6.875)	0.22
	Birth weight (grams)	-0.012	0.007	(-0.026, 0.002)	0.094
	Sex (Male/Female)	-3.445	4.916	(-13.269, 6.379)	0.486
	Hospitalization after birth (days)	0.373	1.216	(-2.056, 2.802)	0.76
	DietaryDivScore1	-0.002	1.98	(-3.959, 3.955)	0.999
9 Months	(Constant)	135.686	87.007	(-38.752, 310.124)	0.125
	Breastfed at timepoint	-12.596	6.744	(-26.118, 0.925)	0.067
	Formula fed at timepoint	9.381	7.801	(-6.259, 25.022)	0.234
	Antibiotic use prior to sampling	-9.68	8.053	(-25.824, 6.464)	0.235
	Older siblings (Yes/no)	5.449	6.299	(-7.178, 18.077)	0.391
	Birth mode (Vaginal/C-section)	3.341	8.683	(-14.068, 20.75)	0.702
	Birthplace (Home/Hospital	-11.403	8.011	(-27.464, 4.659)	0.16

Supplementary Table 1 (cont'd)

	Bristol Stool score	-2.858	2.171	(-7.211, 1.494)	0.193
	Gestational age (weeks)	-1.202	2.542	(-6.298, 3.895)	0.638
	Birth weight (grams)	-0.003	0.009	(-0.021, 0.015)	0.749
	Sex (Male/Female)	-6.537	5.778	(-18.121, 5.047)	0.263
	Hospitalization after birth (days)	-2.063	1.664	(-5.398, 1.273)	0.22
	DietaryDivScore1	2.934	1.994	(-1.064, 6.932)	0.147
11 Months	(Constant)	67.132	103.157	(-139.515, 273.78)	0.518
	Breastfed at timepoint	-14.125	8.032	(-30.216, 1.965)	0.084
	Formula fed at timepoint	-2.822	9.666	(-22.185, 16.542)	0.771
	Antibiotic use prior to sampling	-8.548	10.818	(-30.219, 13.124)	0.433
	Older siblings (Yes/no)	17.161	6.889	(3.362, 30.961)	**0.016**
	Birth mode (Vaginal/C-section)	-1.567	9.632	(-20.862, 17.728)	0.871
	Birthplace (Home/Hospital	-7.438	9.126	(-25.721, 10.844)	0.419
	Bristol Stool score	2.204	2.527	(-2.858, 7.266)	0.387
	Gestational age (weeks)	0.016	3.039	(-6.071, 6.103)	0.996
	Birth weight (grams)	-2.00E-03	0.01	(-0.022, 0.018)	0.841
	Sex (Male/Female)	-13.085	6.604	(-26.315, 0.145)	0.052
	Hospitalization after birth (days)	-0.735	1.759	(-4.259, 2.79)	0.678
	DietaryDivScore1	5.059	2.783	(-0.516, 10.634)	0.074
14 Months**	(Constant)	145.616	170.482	(-200.845, 492.077)	0.399
	Antibiotic use prior to sampling	28.141	19.756	(-12.008, 68.291)	0.163

Supplementary Table 1 (cont'd)

Older siblings (Yes/no)	3.61	11.803	(-20.378, 27.597)	0.762
Birth mode (Vaginal/C-section)	8.297	13.683	(-19.509, 36.103)	0.548
Birthplace (Home/Hospital	2.926	14.586	(-26.716, 32.569)	0.842
Bristol Stool score	4.648	4.128	(-3.742, 13.037)	0.268
Gestational age (weeks)	-3.471	4.928	(-13.485, 6.543)	0.486
Birth weight (grams)	1.60E-02	0.016	(-0.017, 0.048)	0.335
Sex (Male/Female)	-7.59	9.971	(-27.854, 12.673)	0.452
Hospitalization after birth (days)	-3.341	3.089	(-9.618, 2.936)	0.287

*Separate regression models were run for each of the timepoints
**Infant feeding not included as all children were weaned

Supplementary Table 2 Multivariable linear regression models on the association between dietary, lifestyle and other perinatal determinants and microbial diversity (Shannon index).

Age*	Model	Unstandardized Coef. B	Std. Error	95% Confidence Interval for B	Sig.
1-2 weeks	(Constant)	1.461	2.348	(-3.22, 6.142)	0.536
	Breastfed at timepoint	-0.301	0.246	(-0.793, 0.19)	0.226
	Formula fed at timepoint	0.119	0.167	(-0.214, 0.452)	0.479
	Antibiotic use prior to sampling	-0.452	0.356	(-1.162, 0.258)	0.208
	Older siblings (Yes/no)	-0.011	0.146	(-0.301, 0.279)	0.941
	Birth mode (Vaginal/C-section)	-0.219	0.235	(-0.687, 0.25)	0.355
	Birthplace (Home/Hospital	-0.266	0.201	(-0.665, 0.134)	0.19
	Bristol Stool score	-0.02	0.046	(-0.111, 0.071)	0.663
	Gestational age (weeks)	0.038	0.068	(-0.098, 0.174)	0.584
	Birth weight (grams)	0	0	(-0.001, 0)	0.42
	Sex (Male/Female)	0.073	0.141	(-0.208, 0.355)	0.604
	Hospitalization after birth (days)	-0.014	0.039	(-0.091, 0.063)	0.715
4 weeks	(Constant)	-0.844	2.079	(-4.993, 3.306)	0.686
	Breastfed at timepoint	0.293	0.24	(-0.187, 0.773)	0.227
	Formula fed at timepoint	0.537	0.214	(0.11, 0.965)	**0.014**
	Antibiotic use prior to sampling	-1.207	0.58	(-2.365, -0.049)	**0.041**
	Older siblings (Yes/no)	0.078	0.128	(-0.178, 0.333)	0.546
	Birth mode (Vaginal/C-section)	-0.311	0.204	(-0.718, 0.097)	0.133
	Birthplace (Home/Hospital	-0.212	0.187	(-0.586, 0.162)	0.262

Supplementary Table 2 (cont'd)

	Bristol Stool score	-0.083	0.056	(-0.196, 0.029)	0.143
	Gestational age (weeks)	0.091	0.059	(-0.027, 0.209)	0.13
	Birth weight (grams)	0	0	(-0.001, 0)	0.314
	Sex (Male/Female)	0.002	0.129	(-0.256, 0.26)	0.989
	Hospitalization after birth (days)	-0.01	0.039	(-0.088, 0.068)	0.8
8 weeks	(Constant)	0.579	2.295	(-3.998, 5.156)	0.802
	Breastfed at timepoint	-0.004	0.211	(-0.425, 0.417)	0.985
	Formula fed at timepoint	0.404	0.176	(0.053, 0.756)	**0.025**
	Antibiotic use prior to sampling	-0.211	0.598	(-1.403, 0.981)	0.725
	Older siblings (Yes/no)	0.114	0.129	(-0.143, 0.371)	0.378
	Birth mode (Vaginal/C-section)	0.045	0.186	(-0.327, 0.417)	0.811
	Birthplace (Home/Hospital	-0.136	0.197	(-0.529, 0.258)	0.494
	Bristol Stool score	-0.069	0.059	(-0.187, 0.049)	0.247
	Gestational age (weeks)	0.06	0.066	(-0.071, 0.191)	0.364
	Birth weight (grams)	0	0	(-0.001, 0)	0.219
	Sex (Male/Female)	-0.101	0.134	(-0.368, 0.166)	0.453
	Hospitalization after birth (days)	-0.001	0.034	(-0.068, 0.066)	0.969
4 Months	(Constant)	1.056	1.517	(-1.972, 4.084)	0.489
	Breastfed at timepoint	-0.165	0.182	(-0.529, 0.199)	0.367
	Formula fed at timepoint	0.105	0.187	(-0.268, 0.478)	0.576
	Antibiotic use prior to sampling	-0.277	0.436	(-1.147, 0.592)	0.527

Supplementary Table 2 (cont'd)

	Older siblings (Yes/no)	0.171	0.14	(-0.108, 0.45)	0.225
	Birth mode (Vaginal/C-section)	0.141	0.206	(-0.27, 0.552)	0.497
	Birthplace (Home/Hospital	0.043	0.193	(-0.343, 0.429)	0.823
	Bristol Stool score	-0.029	0.061	(-0.15, 0.093)	0.641
	Gestational age (weeks)	0.033	0.042	(-0.051, 0.117)	0.437
	Birth weight (grams)	0	0	(-0.001, 0)	0.407
	Sex (Male/Female)	0.044	0.136	(-0.228, 0.317)	0.746
	Hospitalization after birth (days)	-0.009	0.038	(-0.085, 0.067)	0.813
	Solids introduced (Yes/no)	0.274	0.15	(-0.025, 0.573)	0.072
5 Months	(Constant)	-3.175	2.023	(-7.215, 0.866)	0.121
	Breastfed at timepoint	-0.138	0.189	(-0.514, 0.239)	0.468
	Formula fed at timepoint	0.306	0.188	(-0.069, 0.681)	0.107
	Antibiotic use prior to sampling	-1.107	0.344	(-1.793, -0.42)	**0.002**
	Older siblings (Yes/no)	-0.042	0.132	(-0.306, 0.222)	0.751
	Birth mode (Vaginal/C-section)	0.222	0.194	(-0.166, 0.61)	0.257
	Birthplace (Home/Hospital	-0.154	0.188	(-0.529, 0.221)	0.414
	Bristol Stool score	-0.127	0.065	(-0.257, 0.004)	0.056
	Gestational age (weeks)	0.167	0.059	(0.049, 0.285)	**0.006**
	Birth weight (grams)	0	0	(-0.001, 0)	0.179
	Sex (Male/Female)	0.114	0.13	(-0.147, 0.374)	0.387
	Hospitalization after birth (days)	0.023	0.038	(-0.054, 0.099)	0.556

Supplementary Table 2 (cont'd)

	DietaryDivScore1	0.259	0.123	(0.013, 0.505)	**0.04**
6 Months	(Constant)	-1.642	1.925	(-5.489, 2.204)	0.397
	Breastfed at timepoint	-0.128	0.182	(-0.491, 0.235)	0.482
	Formula fed at timepoint	0.247	0.18	(-0.113, 0.607)	0.175
	Antibiotic use prior to sampling	-0.652	0.317	(-1.286, -0.017)	**0.044**
	Older siblings (Yes/no)	0.308	0.135	(0.038, 0.579)	**0.026**
	Birth mode (Vaginal/C-section)	0.243	0.18	(-0.117, 0.604)	0.182
	Birthplace (Home/Hospital	0.145	0.195	(-0.246, 0.535)	0.461
	Bristol Stool score	-0.127	0.057	(-0.241, -0.012)	**0.03**
	Gestational age (weeks)	0.135	0.057	(0.022, 0.248)	**0.02**
	Birth weight (grams)	0	0	(-0.001, 0)	**0.032**
	Sex (Male/Female)	-0.093	0.131	(-0.354, 0.168)	0.48
	Hospitalization after birth (days)	0.01	0.032	(-0.054, 0.075)	0.752
	DietaryDivScore1	0.065	0.053	(-0.04, 0.171)	0.22
9 Months	(Constant)	4.775	1.96	(0.845, 8.705)	0.018
	Breastfed at timepoint	-0.193	0.152	(-0.498, 0.111)	0.209
	Formula fed at timepoint	0.281	0.176	(-0.072, 0.633)	0.116
	Antibiotic use prior to sampling	-0.311	0.181	(-0.675, 0.053)	0.092
	Older siblings (Yes/no)	0.111	0.142	(-0.174, 0.396)	0.438
	Birth mode (Vaginal/C-section)	0.079	0.196	(-0.313, 0.472)	0.686
	Birthplace (Home/Hospital	-0.187	0.181	(-0.549, 0.175)	0.305

Supplementary Table 2 (cont'd)

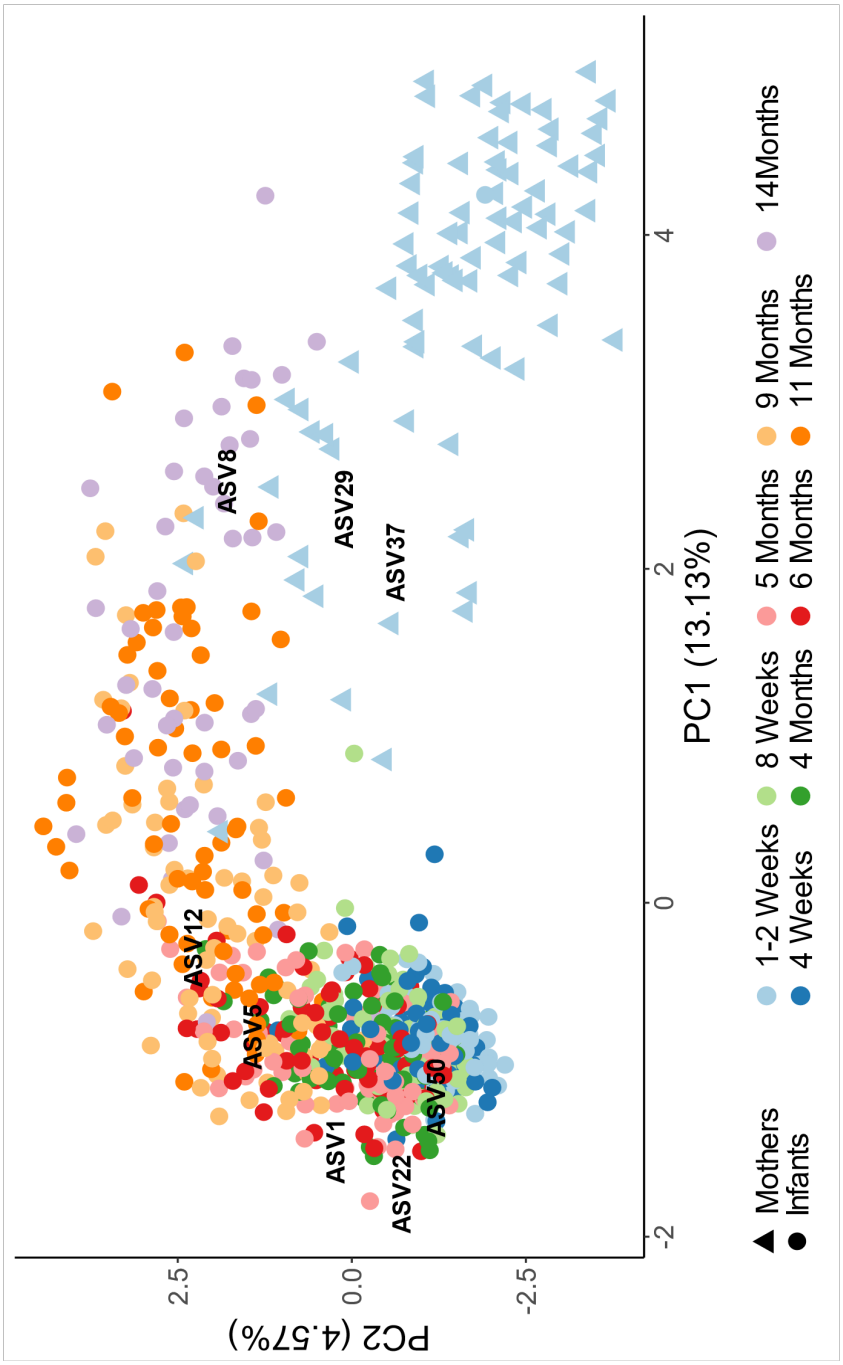
	Bristol Stool score	-0.058	0.049	(-0.156, 0.04)	0.239
	Gestational age (weeks)	-0.068	0.057	(-0.183, 0.047)	0.243
	Birth weight (grams)	0	0	(0, 0.001)	0.505
	Sex (Male/Female)	-0.087	0.13	(-0.348, 0.174)	0.507
	Hospitalization after birth (days)	-0.027	0.037	(-0.102, 0.048)	0.471
	DietaryDivScore1	0.09	0.045	(-0.001, 0.18)	0.051
11 Months	(Constant)	4.069	1.984	(0.094, 8.043)	0.045
	Breastfed at timepoint	-0.202	0.154	(-0.512, 0.107)	0.195
	Formula fed at timepoint	0.065	0.186	(-0.308, 0.437)	0.729
	Antibiotic use prior to sampling	-0.25	0.208	(-0.667, 0.166)	0.234
	Older siblings (Yes/no)	0.304	0.132	(0.038, 0.569)	**0.026**
	Birth mode (Vaginal/C-section)	-0.058	0.185	(-0.429, 0.313)	0.757
	Birthplace (Home/Hospital	-0.203	0.176	(-0.555, 0.148)	0.251
	Bristol Stool score	0.059	0.049	(-0.038, 0.157)	0.227
	Gestational age (weeks)	-0.049	0.058	(-0.166, 0.068)	0.406
	Birth weight (grams)	1.38E-05	0	(0, 0)	0.942
	Sex (Male/Female)	-0.121	0.127	(-0.376, 0.133)	0.344
	Hospitalization after birth (days)	0.028	0.034	(-0.04, 0.096)	0.411
	DietaryDivScore1	0.109	0.054	(0.002, 0.216)	**0.046**
14 Months**	(Constant)	3.299	2.592	(-1.958, 8.557)	0.211
	Antibiotic use prior to sampling	0.249	0.288	(-0.335, 0.833)	0.393

Supplementary Table 2 (cont'd)

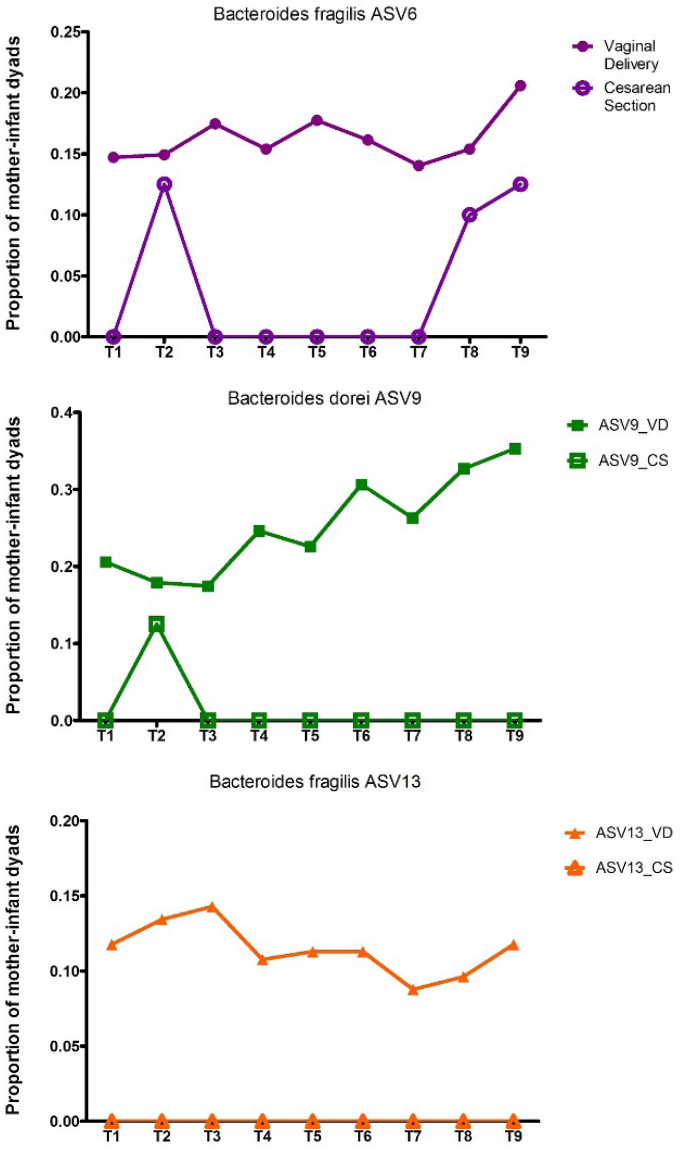
Older siblings (Yes/no)	0.066	0.174	(-0.287, 0.419)	0.709
Birth mode (Vaginal/C-section)	0.115	0.214	(-0.319, 0.55)	0.594
Birthplace (Home/Hospital	0.094	0.223	(-0.358, 0.547)	0.675
Bristol Stool score	0.058	0.063	(-0.071, 0.186)	0.367
Gestational age (weeks)	-0.014	0.076	(-0.167, 0.139)	0.854
Birth weight (grams)	3.30E-05	0	(0, 0.001)	0.89
Sex (Male/Female)	-0.066	0.156	(-0.382, 0.251)	0.677
Hospitalization after birth (days)	-0.014	0.048	(-0.111, 0.083)	0.767

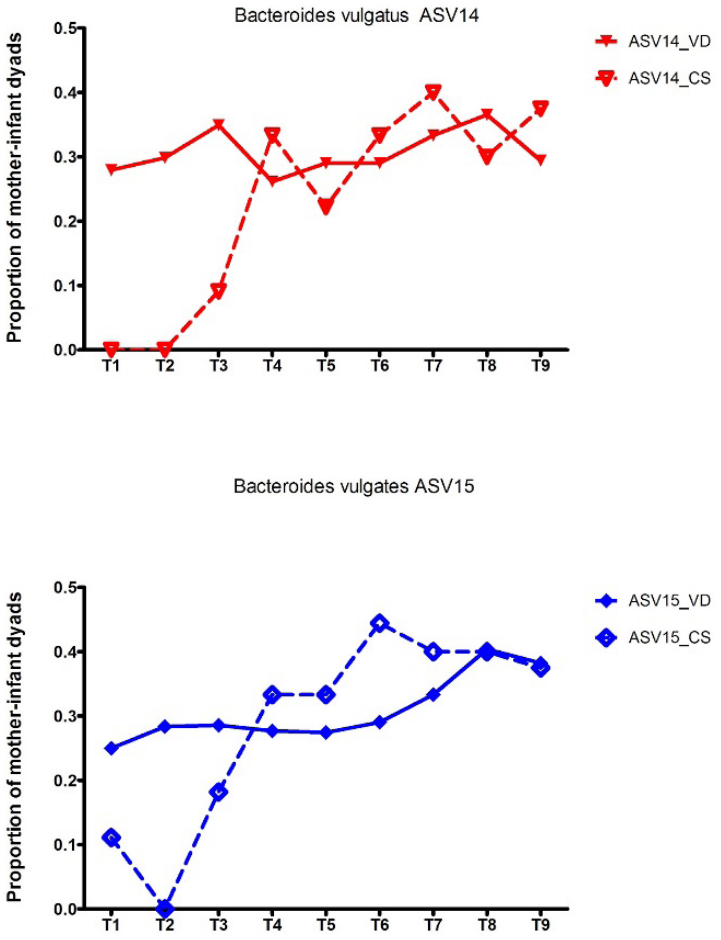
*Separate regression models were run for each of the timepoints
**Infant feeding not included as all children were weaned

Supplementary Figures


PC2 (4.57%)
2.5
0.0
-2.5
PC1 (13.13%)
-2
0
2
4
ASV8
ASV29
ASV37
ASV12
ASV5
ASV1
ASV22
ASV50
Mothers
Infants
1-2 Weeks
4 Weeks
8 Weeks
4 Months
5 Months
6 Months
9 Months
11 Months
14Months

Supplementary figure 1 Principal Component Analyses visualizing the ordination of the microbial community structure (based upon the Aitchison distance) of infant samples coloured according to age in comparison to maternal samples. Depicted ASV's are the most external ASVs based on their coordinates.


Bacteroides fragilis ASV6
Proportion of mother-infant dyads
0.25
0.20
0.15
0.10
0.05
0.00
T1
T2
T3
T4
T5
T6
T7
T8
T9
Vaginal Delivery
Cesarean Section
Bacteroides dorei ASV9
Proportion of mother-infant dyads
0.4
0.3
0.2
0.1
0.0
T1
T2
T3
T4
T5
T6
T7
T8
T9
ASV9_VD
ASV9_CS
Bacteroides fragilis ASV13
Proportion of mother-infant dyads
0.20
0.15
0.10
0.05
0.00
T1
T2
T3
T4
T5
T6
T7
T8
T9
ASV13_VD
ASV13_CS



Supplementary figure 2 Proportion of mother-infant dyads sharing the most abundant Bacteroides ASVs throughout infancy according to mode of delivery.

Chapter 4


potential
included
allergies
identify
escherichia
developed
previous
higher
maturation
lachnospira
decreased
data
infants
previously
microbiome
year
observed
obacteria
one
first
significantly
another
equencing
became
also
maturity
diseases
longitudinal
disease
many
half
tract
role
infancy
asthma
lower
adjusted
results
genera
diversity
schoolage
infant
later
otus
otu
yet
associat
structure
test
allergic
age
born
life
analyses
identified
result
table
prick
risk
lack
taxa
used
line
fdr
chao
weeks
children
associated
sitization
affects
often
oral
birth
far
study
microbiota
development
breastfeeding
mong
sters
microbial
total
acterial
onset
dividual
supplementary
dings
maz
opic
composition
alterations
diet
faecal
new
samples
estinal

Development of the microbiota and associations with birth mode, diet, and atopic disorders in a longitudinal analysis of stool samples, collected from infancy through early childhood

Gianluca Galazzo*, Niels van Best*, Liene Bervoets, Isaac Oteng Dapaah, Paul H. Savelkoul, Mathias W. Hornef, the GI-MDH consortium, Susanne Lau§, Eckard Hamelmann§, and John Penders§

** Shared first authorship*

§ Shared last authorship

Gastroenterology 2020 May;158(6):1584-1596

Abstract

Establishment of the gastrointestinal microbiota during infancy affects immune system development and oral tolerance induction. Perturbations in the microbiome during this period can contribute to development of immune-mediated diseases. We monitored microbiota maturation and associations with subsequent development of allergies in infants and children.

We collected 1453 stool samples, at 5, 13, 21, and 31 weeks post-partum (infants), and once at school-age (6–11 years), from 440 children (49.3% girls, 24.8% born by caesarean section; all children except for 6 were breastfed for varying durations; median 40 weeks; interquartile range, 30–53 weeks). Microbiota were analysed by amplicon sequencing. Children were followed through 3 years of age for development of atopic dermatitis; data on allergic sensitization and asthma were collected when children were school age.

Diversity of faecal microbiota, assessed by Shannon index, did not differ significantly among children from 5 through 13 weeks after birth, but thereafter gradually increased to 21 and 31 weeks. Most bacteria within the Bacteroidetes and Proteobacteria phyla were already present at 5 weeks after birth, whereas many bacteria of the Firmicutes phylum were acquired at later times in infancy. At school age, many new Actinobacteria, Firmicutes, and Bacteroidetes bacterial taxa emerged. The largest increase in microbial diversity occurred after 31 weeks. Vaginal, compared with cesarean section delivery, was most strongly associated with an enrichment of Bacteroides species at 5 weeks through 31 weeks. From 13 weeks onwards, diet became most the important determinant of the microbiota composition—cessation of breastfeeding, rather than solid food introduction, was associated with changes. For example, bifidobacteria, staphylococci, and streptococci significantly decreased upon cessation of breastfeeding, whereas bacteria within the Lachnospiraceae family (*Pseudobutyrivibrio*, *Lachnobacterium*, *Roseburia*, and *Blautia*) increased. When we adjusted for confounding factors, we found the faecal microbiota composition to be associated with development of atopic dermatitis, allergic sensitization, and asthma. Members of the Lachnospiraceae family, as well as the genera *Faecalibacterium* and *Dialister*, were associated with a reduced risk of atopy.

In a longitudinal study of faecal microbiota of children from 5 weeks through 6–11 years, we tracked changes in diversity and composition associated with the development of allergies and asthma.

Introduction

Colonization of the intestinal tract during the neonatal period is of crucial importance for the maturation of the mucosal immune system and the induction of oral tolerance [1-3]. Animal studies have provided compelling evidence to support a causal role of the intestinal microbiota and its metabolites, especially in early life, in the etiology of allergic diseases [3-6].

Numerous epidemiological studies [7-12] also suggest that the infant intestinal microbiota plays an important role in the manifestation of allergic diseases and asthma, although actual results vary considerably between studies. About half of the studies that examined intestinal microbial diversity in infancy and childhood reported a lower diversity among children with (subsequent) allergies, whereas the remaining studies found no evidence for such an association [13]. Moreover, despite many specific microbial taxa have been linked to allergies and asthma, it remains unclear which bacterial taxa prevent or promote disease onset [14].

Lack of early stool sampling and different ages of stool sample collection, different microbiological profiling methods, and an inadequate control for potential confounders have been suggested to contribute to the heterogeneity between study results [9, 13]. Additionally, cross-sectional studies are prone to reverse causality, i.e., changes in the microbiota composition as a result of the disease manifestation, and only very few studies have sufficient clinical follow-up to link infant microbiota maturation to the subsequent development of asthma [11].

Initial microbial colonization starts upon rupture of the amniotic membranes and subsequent passage through the birth canal when the infant is seeded by maternal microbial strains, a process which is impeded in case of a caesarean section delivery [15, 16]. Subsequently, microbial populations evolve as the diet changes and the host develops. Given the highly dynamic and complex process of microbial assembly, succession and maturation, repeated sampling is important to allow analysis of the overall development of the indigenous infant microbial ecology [17]. Moreover, many of the antenatal and postnatal factors that influence microbial community assembly during infancy, such as birth mode and the presence of older siblings and furry pets in the household, have also been associated with the development of allergic diseases and asthma [9, 13, 18, 19], highlighting the importance to account for potential confounding factors.

In the present study, we aimed to monitor microbial assembly, succession, and maturation during the first year of life and identify hereditary, perinatal, environmental, lifestyle and dietary factors that drive microbiota development. Through the application of various multivariable longitudinal models, including joint modelling, we next examined the dynamics in microbial diversity, composition, and community structure in association with the subsequent risk of developing atopic dermatitis and asthma until school-age.

Our findings indicate that alterations in microbial diversity and composition precede the onset of allergic manifestations, while emphasizing the importance of possible confounders.

Materials and Methods

Design and clinical outcome measurements

Originally, this study was designed as a randomized placebo-controlled clinical trial to examine the impact of a bacterial lysate, containing heat-killed *Escherichia coli* and *Enterococcus faecalis,* on the primary prevention of atopic dermatitis (AD) (registration no. ISRCTN60475069, ISRCTN registry) [20]. However, we did not find any evidence that the intervention affected the microbiota composition and therefore pooled both treatment arms in the downstream statistical analyses. Infants were clinically examined by a paediatrician on signs of AD at the ages of 1, 21 and 31 weeks and again at 1, 2 and 3 years of age as described previously [20]. The school-age follow-up of the study population (at 6-11 years) took place in 2013 including clinical examination, lung function testing, skin prick tests, and serum analyses of specific IgE to the most common aero-allergens (house dust mite, dog, cat, mold (*Alternaria*, *Cladosporium*), birch and grass pollen). Children were classified as having current asthma in case of a doctor's diagnosis in combination of any indicative symptoms in the last 12 months (wheezing, shortness of breath, nocturnal awakening due to shortness of breath and/or wheezing). Allergic sensitization was assessed by Skin Prick Test and serum sensitization for the above-mentioned allergens. The study and follow-up were approved by the hospitals local review board Charité Ethics Committee in 2002 and 2012. Parents and participants gave written informed consent.

Microbial profiling of faecal samples

Faecal DNA was isolated by a combination of bead beating and column-based purification as described in detail previously [21]. Barcoded universal primers adapted from Bartram and collegues [22] were used to amplify the variable 3 region of the 16S rRNA gene. Amplicons were sequenced using the Illumina MiSeq platform using 2x250 paired-end reads. The resulting sequencing data were processed using the short-read library 16S rRNA gene sequencing pipeline (sl1p) [23] (for description see supplementary methods). This resulted in a total of 93,475,612 reads from 1,468 samples that were clustered into 7,961 OTUs. Removal of OTUs that were observed in only a single sample and discarding OTUs with a fraction of the total number of sequences below 0.001, retained the majority of sequences (92,997,277) while significantly reducing the number of OTUs to 873. Finally, we eliminated 15 samples with a low coverage (<15,000 reads) and normalized the data using *Rhea* [24]. In order not to discard informations, normalization in *Rhea* is performed by dividing OTU counts per sample for their total count (sample depth) and then multiplying the obtained relative abundance for the lowest sample depth (15,540 reads).

Statistical Analysis

All the statistical analyses were performed two-sided using R, version 3.4.3.

GI microbiota richness and maturation

The Chao1 index, as measure for the estimated microbial richness, and the Shannon index, as microbial diversity metric, were computed using the R package *vegan 2.5.3* [25]. In order to compare the microbial community structure of samples, we used the unweighted UniFrac which incorporates phylogenetic distances between observed or-

ganisms. The R package *GUniFrac 1.1* [26] was used to compute the unweighted UniFrac metric. Ordination of samples based upon the unweighted UniFrac dissimilarity was visualized using Principal Coordinate Analysis (PCoA) using the *capscale* function of the *vegan* package. The Friedman test was used to test for significant differences in Chao1, Shannon and loadings scores on the principal coordinates between all time-points, followed by Dunn's test for post hoc pairwise comparisons between individual time-points. The p-values were finally FDR adjusted (Benjamini–Hochberg procedure) for multiple comparisons. In order to track the dynamics of individual OTUs within the four main bacterial phyla, we created Sankey plots using Sankeymatic. For readability of the Sankey plots, the OTU table was further filtered to OTUs that were present in ≥ 10% of the samples during one or more time-points. We next investigated the maturity of the GI microbiota by computing the microbial-by-age z-score (MAZ) of the sample as described previously [27]. Shortly, we started feeding a random forest with a training set made of the microbial community composition of healthy children after filtering out all OTUs with a prevalence below 5%. Once the model was trained, we used it to predict the age of all the samples. Finally, the z-score was computed using the following formula:

$$MAZ = \frac{Microbial\ age - \ median\ of\ microbial\ age\ of\ healthy\ children\ of\ same\ chronologic\ age}{(standard\ deviation\ of microbial\ age\ of\ healthy\ children\ of\ the\ same\ chronologic\ age)}$$

Dirichlet Multinomial Mixture (DMM) clustering, an unsupervised clustering method that uses Laplace approximation to identify groups of communities (enterotypes) with similar composition, was performed as previously described [28]. We then analysed the transition of infants through these DMM-clusters with age [29].

Analysis of factors shaping the GI microbiota

We examined which hereditary, perinatal, environmental, lifestyle and dietary factors were associated with the establishment of the microbiome during infancy (see supplementary methods for detailed description). In order to examine which of these factors were associated with the DMM clusters at baseline and/or with the transition of DMM clusters between the age of 5 and 31 weeks, multinomial logistic regression analysis was used. Only factors that were significantly associated with the (transition of) DMM clusters in the univariable analyses, were included in the final multivariable model. We next used multivariate association with linear models (MaAsLin) [30] to examine the association between these factors and individual microbial taxa. The effect size and significance of each of the covariates on the microbial community structure was determined using the *envfit* function in *vegan* [25]. Ordination was performed using the PCoA based on unweighted UniFrac metric obtained as described above. The significance value was determined based on 999 permutations. All *P* values derived from *envfit* were adjusted for multiple comparisons using FDR adjustment. In order to understand which of the covariates had the strongest impact on the overall microbial community structure, we performed a Permutational Analysis of the Variance (PERMANOVA) based on unweighted UniFrac. Only covariates that were statistically significant in the *envfit* analyses were included in the PERMANOVA.

Analysis on association between microbiota and allergic manifestations.

To examine how the longitudinal variation of the microbial diversity (Shannon index) and maturity (MAZ) of the GI microbiota affects the time to development of AD, we applied a joint model [31] using the JM function of the *JM*package [32] (for details see supplementary methods). To examine the impact of microbial diversity and maturity on asthma and allergic sensitization at school-age, a Generalized Linear Model (GLM) was built using *lme41.1.19.* The same covariates as included in the JM were incorporated as potential confounding factors. Since both asthma and sensitization were binary outcomes, a binomial distribution was chosen for the GLM. To identify if specific bacterial genera were differentially abundant in children with and without allergic manifestations, we used the MetaLonDa package [33]. To ensure meaningful p-values we performed 999 permutations. To select only the significant associations, we choose a threshold of 0.05 for the p-values after FDR adjustment.

Results

Study population characteristics

The study, initially designed as a randomized, placebo-controlled trial on the primary prevention of AD by an orally applied lysate of heat-killed *Escherichia coli* and *Enterococcus faecalis*, consisted of healthy newborns (n = 606) with a single or double heredity for atopy [20]. During the first 3 years of life, children were deeply phenotyped by physical examination and the collection of detailed questionnaires at 7 clinical visits. At school-age, children were contacted again to determine the establishment of asthma and allergic sensitization.

For the present study, only children with at least three faecal samples collected during the first year and/or faeces collected at school-age were included (n = 440). Of these children, 217 (49.3%) were girls, 187 (42.5%) had older siblings, 109 (24.8%) were born by caesarean section and 29 (6.6%) were reportedly treated with antibiotics in the first 31 weeks of life. All except 6 children received breastfeeding, although the duration of breastfeeding varied considerably with a median duration of 40 weeks [IQR: 30-53]. Solid food was introduced at a median age of 25 weeks [IQR: 22 – 27] (Supplementary Table S1).

Development of the microbiota between early infancy and school-age

We first examined the compositional changes in the microbiota during infancy and compared this to the school-age microbiota composition. Samples collected at the age of 5 (n = 306), 13 (n = 287), 21 (n = 268) and 31 (n = 307) weeks post-partum and again at school-age (n = 300) were profiled by amplicon sequencing of the 16S rRNA hypervariable V3 gene region. Upon quality filtering and removal of samples with low sequencing depth (n = 15), 1,453 samples with a median sequencing depth of 62,420 reads/sample (range 15,540 – 168,848) were retained for downstream analysis and clustered into 873 operational taxonomic units (OTUs).

Microbial diversity, assessed by the Shannon index, was not significantly different

between ages 5 and 13 weeks but thereafter gradually increased from 13 to 21 and 31 weeks after birth (Fig. 1A). The largest increase in microbial diversity occurred after the age of 31 weeks as indicated by the steep increase in the Shannon index at school-age ($p = 7.99*10^{-28}$). Similar findings were observed for the microbial richness as assessed by the Chao1 (Supplementary Table S2).

Principal coordinate analyses indicated that the microbial community structure, as assessed by the unweighted UniFrac metric, also gradually shifted during infancy with the most prominent shift between the age of 21 and 31 weeks (Fig 1B, $p = 1.58*10^{-27}$, Supplementary Table S3). The school-age samples, however, clustered separately and showed less inter-individual variation as compared to the infant samples.

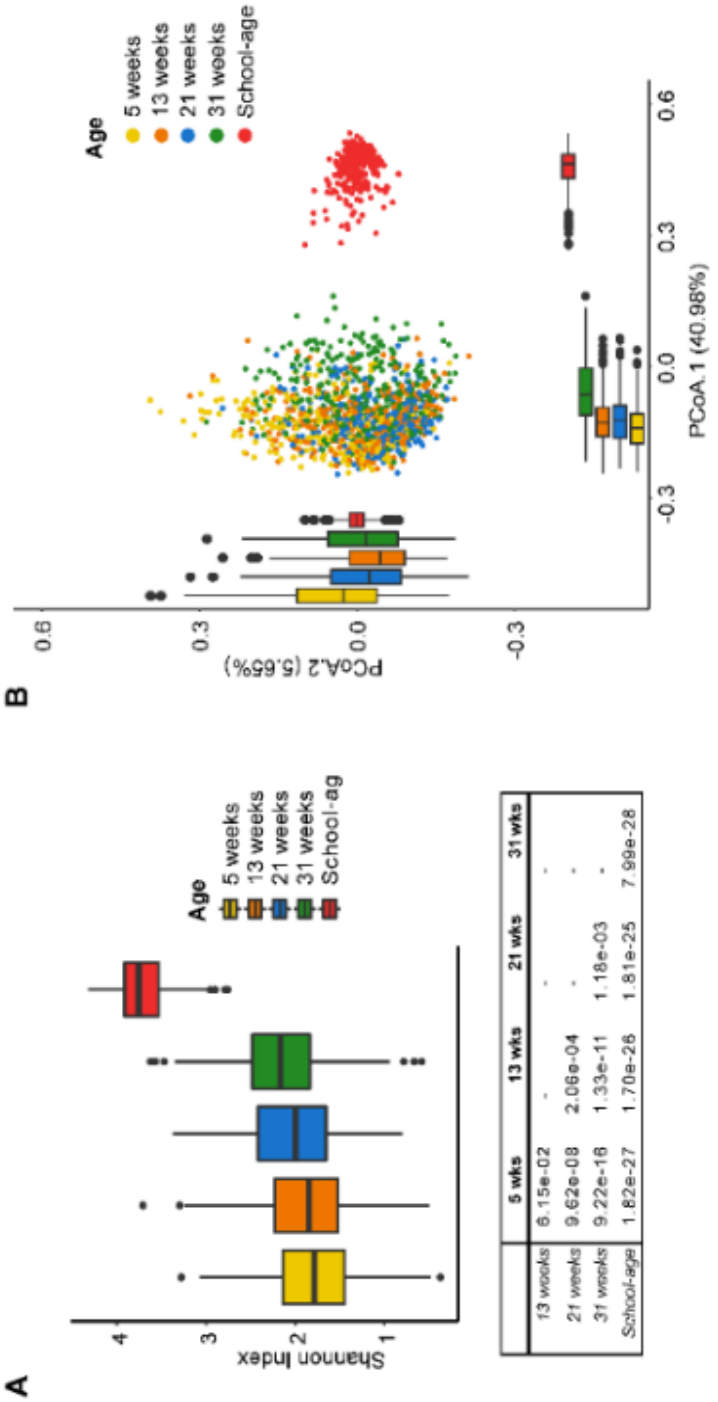
Tracking individual OTUs based on their presence or absence revealed different dynamics within the dominant phyla (Supplementary Fig 1). The majority of OTUs within the phyla of Actinobacteria, Bacteroidetes and Proteobacteria found during infancy were already present at 5 weeks after birth, whereas almost half of the OTUs within the phylum of Firmicutes were only acquired at later infant time-points. At school-age, many new Actinobacteria, Firmicutes, and Bacteroidetes OTUs emerged on top of the OTUs already present during infancy. In contrast, only few new OTUs emerged within the phylum Proteobacteria at school-age, while a significant portion of the infant OTUs were lost thereafter.

We next examined the bacterial genera that contribute most to the temporal dynamics in microbial diversity and community structure. Towards school-age the prevalence in many of the genera within the phylum of Proteobacteria dramatically decreased, whereas the prevalence of genera within the phylum of Firmicutes, and in particular within the Lachnospiraceae and Ruminococcaceae families strongly increased (Fig 1C). Moreover, with the exception of *Bifidobacterium*, the relative abundance of all of the major bacterial genera changed significantly (Friedman test, all P-values <0.001, Supplementary Table 4) throughout infancy and childhood (Supplementary Fig 2A-B). *Escherichia* was the most abundant genus at 5 weeks of age followed by *Bifidobacterium* and *Streptococcus. Escherichia* still remained the most abundant genus at 31 weeks of age but was now followed by *Bacteroides* and *Veillonella.* At school-age the most abundant genera were *Blautia*, *Faecalibacterium* and *Ruminococcus*.

The length of breastfeeding represents the main driver of the infant's microbiota composition

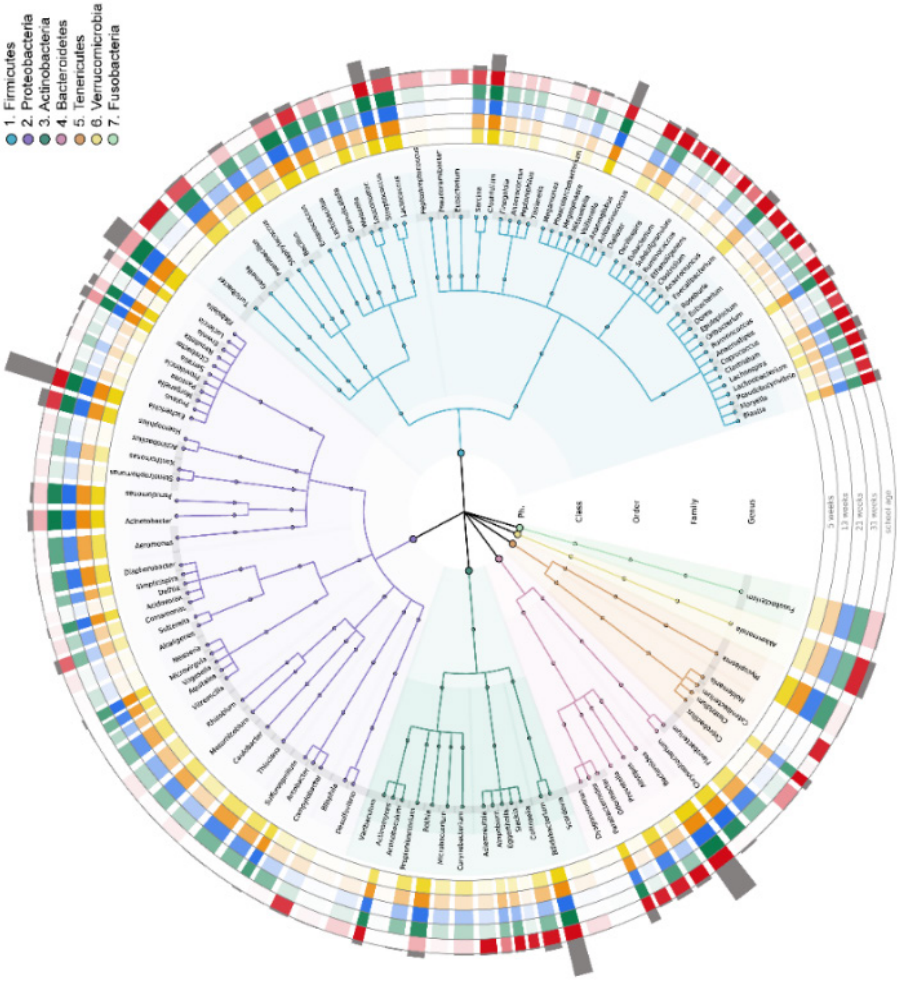
To identify covariates associated with the microbiota dynamics during infancy, we continued our analyses focusing on the infant samples. Dirichlet multinomial mixtures (DMM) modelling on OTU-level data formed six clusters (based on lowest Laplace approximation) (Fig. 2A-B).

To illustrate the progression of samples through each DMM cluster with age, we applied a transition model as described previously [29]. Clusters 1 and 3 were the most dominant at the age of 5 weeks and thereafter transitions were chaotic, consistent with the previously identified developmental phase of the microbiome during the first 14 months of life [29]. Although cluster 1 remained dominant until the age of 31 weeks, cluster 3 gradually disappeared in favour of clusters 4 and 5 (Fig 2C). Multinomial logistic regression analyses indicated that the initial microbiota cluster at 5 weeks of age was mainly determined by birth mode. The chance that a newborn's microbiota belonged to



	5 wks	13 wks	21 wks	31 wks
13 weeks	6.15e-02	-	-	-
21 weeks	9.62e-08	2.06e-04	-	-
31 weeks	9.22e-16	1.33e-11	1.18e-03	-
School-age	1.82e-27	1.70e-26	1.81e-25	7.99e-28

Figure 1 Microbiota maturation throughout infancy and childhood (N= 1,453 stool samples from 440 children). A) Microbial diversity (Shannon index) gradually increased throughout infancy and has markedly risen at school-age ($p = 7.72*10^{-53}$, Friedman, p-values for post-hoc analyses using Dunn's test are depicted in the table). B) Principal Coordinate Analysis (PCoA) based on unweighted UniFrac dissimilarity indicates a gradual shift in microbial community structure along PC1 during infancy and a completely distinct structure at school-age ($p = 6.0*10-51$, Friedman, p-values for post-hoc analyses using Dunn's test are depicted in Supplementary Table 3). C) Cladogram depicting the bacterial genera detected in the children's faecal microbiota. Background and branch colours reflect the different phyla. The height of the outer ring reflects the average relative abundance of a genus across all infant time points, whereas the colour density of the five inner rings reflects the prevalence of the genus at the individual time points (with opaque colour indicating a prevalence of 100 percent and fully transparent indicating a prevalence of 0 percent).



c

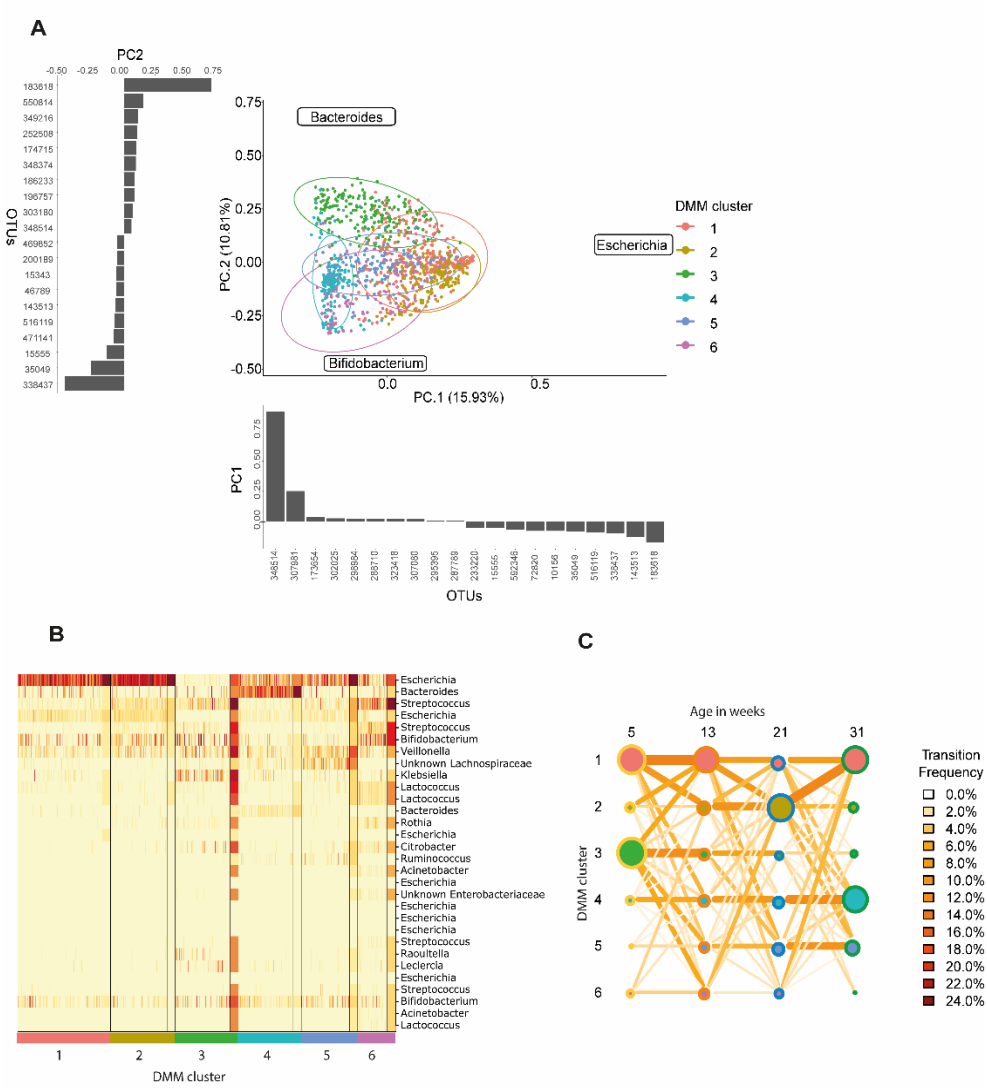


Figure 2 Community typing by Dirichlet Multinomial Mixtures of infant samples revealed six clusters (N=1,154 stool samples from 312 children). A) Heat map showing the relative abundance of the 30 most important/dominant OTUs per DMM cluster. B) Principal Component Analysis (PCA) on OTU-level data with samples colored according to DMM cluster. Ellipses indicate the 95% confidence interval. Vertical and horizontal bar charts depict OTUs with the highest loadings on PC1 and PC2, respectively. The OTUs with the highest positive and negative loadings on PC1 and PC2 are plotted in the PCA. C) Transition model showing the progression of samples through the six DMM clusters from one sampling time-point to the next time-point.

cluster 3 was strongly increased among infants born by caesarean section [Supplementary Table S5]. This cluster was remarkably different with respect to the abundance of several of the driving OTUs. In particular a *Klebsiella* OTU exhibited a high abundance at the expense of an *Escherichia* OTU that dominated many of the other clusters. Additionally, *Citrobacter*, *Leclercia* and *Raoultella* OTUs were characteristic for cluster 3 (Fig 2B).

Analysis of the most common transition trajectories, revealed that for both children starting in cluster 1 as well as for children starting in cluster 3, transition towards clusters 4 and 5 significantly increased when breastfeeding was ceased [Supplementary Table S5]. These results were further supported by the overall bacterial profiles throughout infancy. At the age of 5 weeks, the largest amount of variance was explained by birth mode (Fig 3A). At the genus level, vaginal as compared to caesarean section delivery was most strongly associated with an enrichment of *Bacteroides spp.,* at 5 weeks and until the age of 31 weeks (Supplementary Table S6).

At the age of 13, 21 and 31 weeks, breastfeeding explained by far the greatest variance in bacterial community profiles (Fig 3B-D). Permutational multivariate analyses of variance confirmed that the duration of breastfeeding had a stronger impact than the introduction of solid foods (Fig 3E, Supplementary Table S7). Bifidobacteria, staphylococci and streptococci amongst others significantly decreased upon cessation of breastfeeding, whereas many bacteria within the Lachnospiraceae family (e.g., *Pseudobutyrivibrio*, *Lachnobacterium*, *Roseburia*, *Blautia*) increased (Supplementary Table S6).

A longer duration of breastfeeding was also associated with a lower microbial diversity (Supplementary Table S8) as well as with a lower microbial-by-age z-score (MAZ) (Supplementary Table S9). The MAZ is calculated by training a machine-learning algorithm on the microbiota composition of a dataset with known biological age, thereafter the age of samples is predicted based on its microbiota composition. A lower MAZ is thus indicative for a delayed microbial maturation.

Furthermore, the exposure to older siblings was associated with an increase in several genera within the phylum of Actinobacteria (*Bifidobacterium* and *Corynebacterium* at 5 weeks and *Egghertella* at 21 weeks, Supplementary Table S6) and a higher microbial diversity at 31 weeks of age (Supplementary Table S9). Finally, besides dietary factors, the microbial community structure was most strongly associated with the presence of AD at time of sample collection.

Alterations in microbial composition, diversity and maturity precede manifestations of atopy

To further investigate whether differences in microbiota development precede the onset of atopic disease, we applied several longitudinal analyses while controlling for potential confounding factors by adjusting for other covariates.

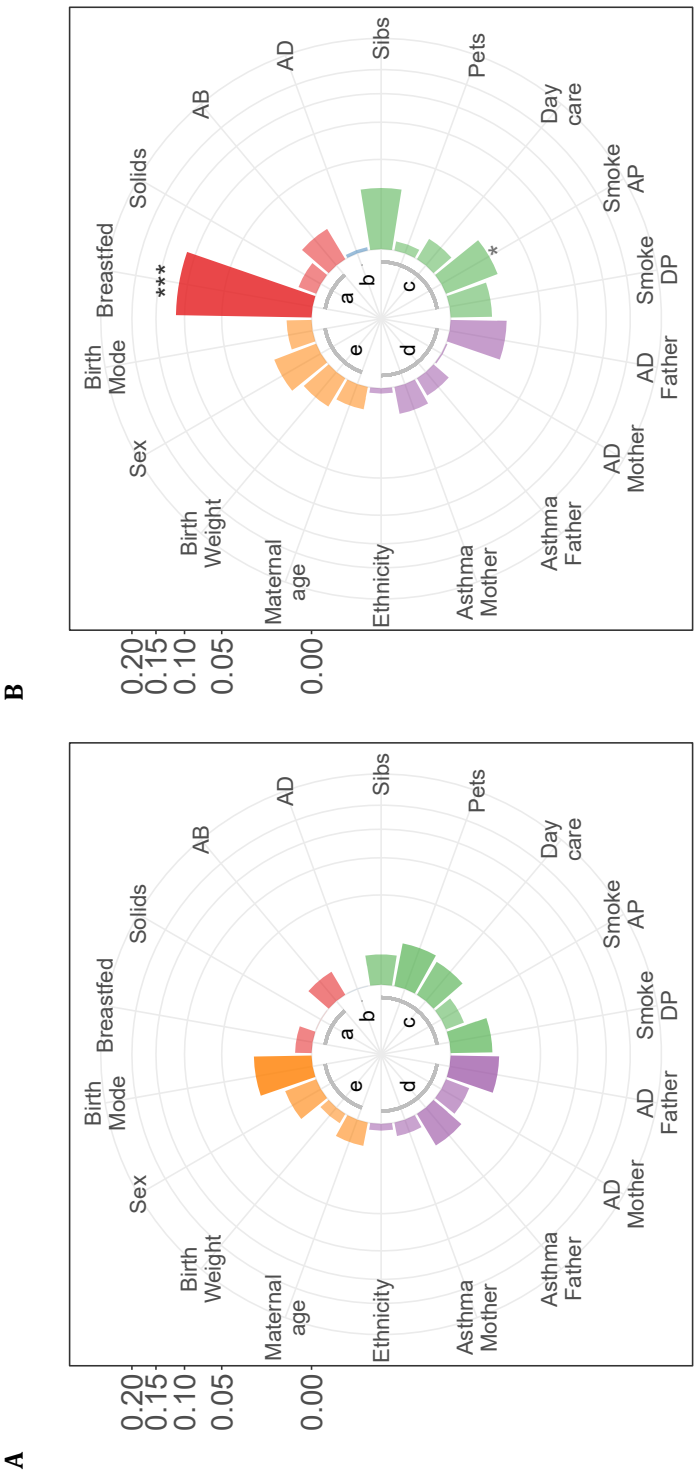
We first applied multivariate joint models on the microbial diversity and maturity in association to AD. Joint models have become increasingly popular as a statistical framework to concurrently analyse longitudinal data (e.g., biomarker evolution) and survival data (e.g., time-to-disease onset) [31]. To our knowledge, they have not been applied in the microbiome research field so far. While accounting for known risk factors for AD, we found that the temporal pattern of microbial diversity was independently and inversely associated with AD (Hazard ratio (HR) = 0.21; $p = 1.15*10^{-4}$, Fig 4A-B, Supplementary Table S10), indicating that a lower microbial diversity throughout infancy is associated

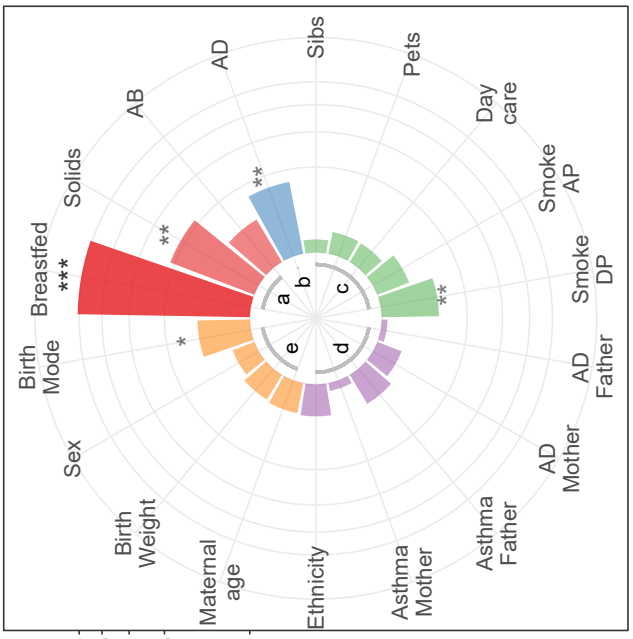
with an increased risk of AD. For the temporal pattern of microbial maturity, expressed as microbial age z-scores, we found a statistically significant positive association with AD (HR = 1.14, p = $1.94*10^{-5}$, Fig 4C-D, Supplementary Table S11). Next, we used the recently introduced metagenomics longitudinal differential abundance (MetaLonDa) method [33] to identify time intervals of differentially abundant bacterial genera between infants that did or did not develop AD.

Among children who did not develop AD during follow-up, the relative abundance of *Atopobium* (days 25.6 - 79.4, $FDR_{adjusted}$ p = $7.65*10^{-3}$), *Corynebacterium* (days 126.1 - 151.2, $FDR_{adjusted}$ p = $9.68*10^{-3}$), both members of the phylum Actinobacteria, and *Prevotella* (days 104.6 - 133.3, $FDR_{adjusted}$ p <0.001) were temporarily enriched when compared to children who developed AD. Most pronounced were, however, the associations of *Lachnobacterium* and *Faecalibacterium,* which were significantly enriched during the entire period of faecal sampling among children who remained free from AD (Figure 4E-G, Supplementary Table 12).

Next, we examined whether the infant microbiota composition was also associated with allergic manifestations at school-age, including allergic sensitization and asthma. Blood samples for the determination of allergic sensitization at school age were available for 292 out of the 440 children included in the present study. Like for AD, we found a higher diversity of the infant microbiota to be associated with decreased risk of allergic sensitization at school-age (Shannon index at 31 weeks $OR_{adjusted}$ = 0.19, p= $7.33*10^{-3}$, Supplementary Figure 3A, Supplementary Table 13). A higher microbial maturity very early in life was associated with an increased risk of allergic sensitization (MAZ at 5 weeks $OR_{adjusted}$ = 1.46, p = $5.01*10^{-3}$, Supplementary Figure 3B, Supplementary Table 14), again in line with findings for AD. We could, however, not identify individual bacterial genera with differential abundance over a significant period of time between children who did or did not develop allergic sensitization.

A clear association between microbial diversity and asthma could not be detected. Yet, in line with allergic sensitization and AD, a higher microbial maturity at the age of 5 weeks was also associated with an increased risk for asthma (MAZ at 5 weeks $OR_{adjusted}$ = 1.43, p = $7.78*10^{-3}$, Supplementary Figure 3C-D, Supplementary Tables 15 and 16). Multiple bacterial genera were differentially abundant over time in children who did or did not develop asthma. The genera that were differentially abundant across the entire time-period during which the microbiota composition was monitored included *Lachnobacterium,* Lachnospira (both members of the Lachnospiraceae family) and *Dialister* (Veillonellaceae), which were all significantly enriched in healthy as compared to asthmatic children (Supplementary Figure 3E-G, Supplementary Table 17).


A
B
0.20
0.15
0.10
0.05
0.00
Sibs
Pets
Day care
Smoke AP
Smoke DP
AD Father
AD Mother
Asthma Father
Asthma Mother
Ethnicity
Maternal age
Birth Weight
Sex
Birth Mode
Breastfed
Solids
AB
AD
a
b
c
d
e
***
*


D
0.20
0.15
0.10
0.05
0.00
Sibs
Pets
Day care
Smoke AP
Smoke DP
AD Father
AD Mother
Asthma Father
Asthma Mother
Ethnicity
Maternal age
Birth Weight
Sex
Birth Mode
Breastfed
Solids
AB
AD
a
b
c
d
e
**
**
**
*
***
*

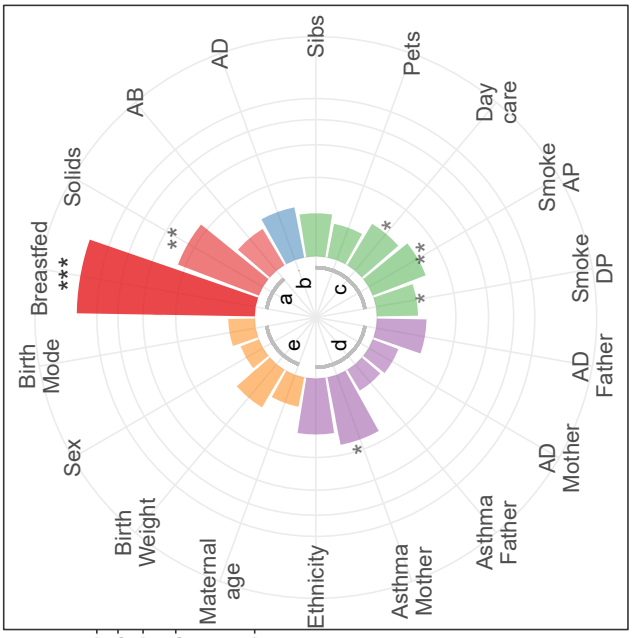

C
0.20
0.15
0.10
0.05
0.00
Sibs
Pets
Day care
Smoke AP
Smoke DP
AD Father
AD Mother
Asthma Father
Asthma Mother
Ethnicity
Maternal age
Birth Weight
Sex
Birth Mode
Breastfed
Solids
AB
AD
a
b
c
d
e
**
*
**
*
***
*

Figure 3 Microbiota community structure is most strongly influenced by breastfeeding (N= 961 stool samples from 312 children). A-D, Polar plots visualizing the amount of variance of microbial communities at 5 weeks A), 13 weeks B), 21 weeks C) and 31 weeks D) of age that could be explained by 18 covariates as analyzed using EnvFit. The height of the bars reflects the amount of variance (r2) explained by each covariate. Covariates are colored to highlight parental health status and ethnicity (purple), perinatal covariates (orange), diet and medication (red) and environmental exposures (green). Asterisks indicate significant covariates (false discovery rate (FDR) P< 0.05) at each time point. E) Permutational Multivariate Analysis of Variance (PERMANOVA) combining all covariates that were significantly associated with microbial community variation at any given time point in the EnvFit analyses. The size of the dots reflects the R2. Only samples without missing data on the included covariates were included in PERMANOVA (5 weeks: N= 238, 13 weeks: N = 233, 21 weeks: N = 231, 31 weeks: N= 259). Asterisks indicate statistical significance with *p<0.05, **p<0.01, ***p<0.001.

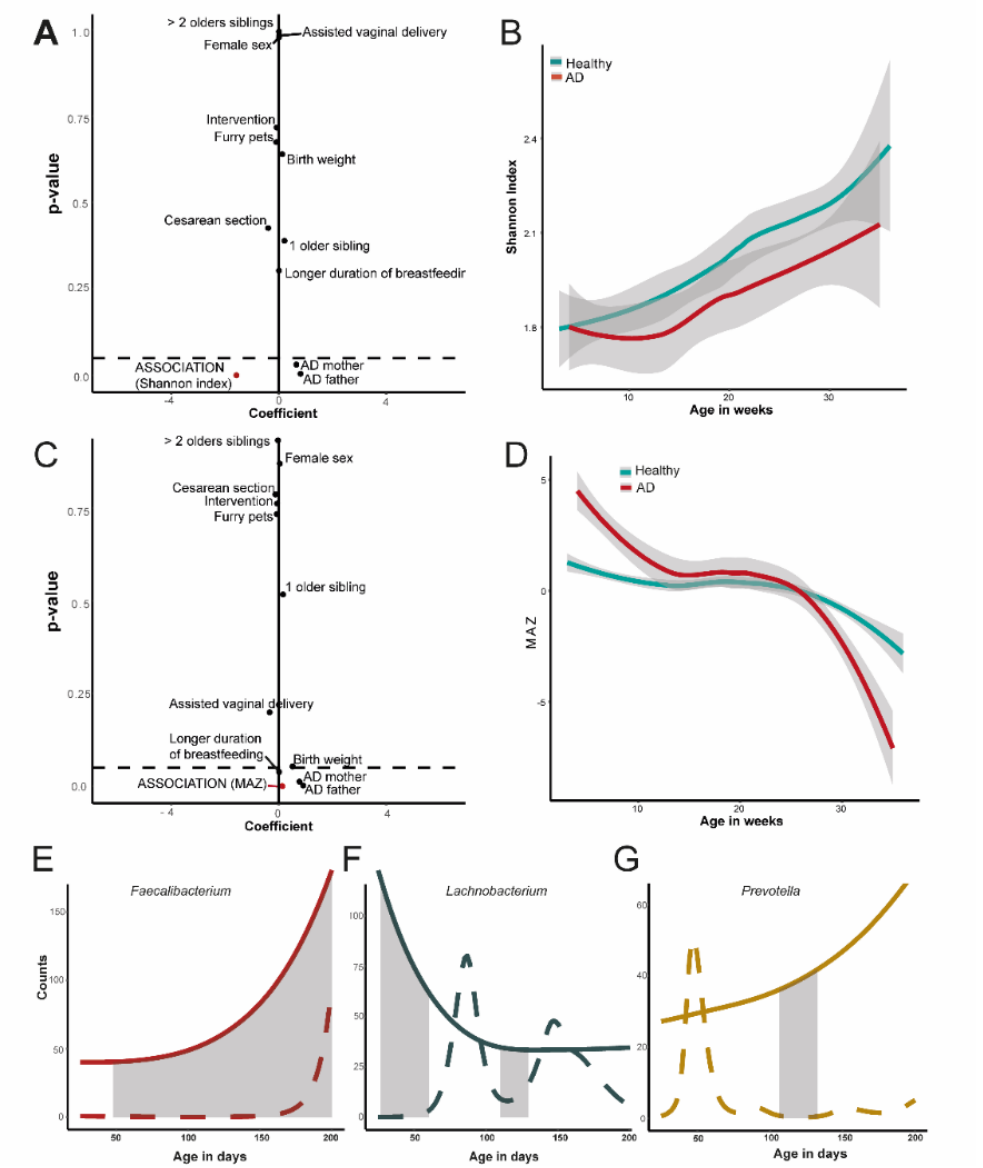

A
> 2 olders siblings
Female sex
Assisted vaginal delivery
Intervention
Furry pets
Birth weight
Cesarean section
1 older sibling
Longer duration of breastfeedir
ASSOCIATION (Shannon index)
AD mother
AD father
p-value
Coefficient
B
Healthy
AD
Shannon Index
Age in weeks
C
> 2 olders siblings
Female sex
Cesarean section
Intervention
Furry pets
1 older sibling
Assisted vaginal delivery
Longer duration of breastfeeding
Birth weight
ASSOCIATION (MAZ)
AD mother
AD father
p-value
Coefficient
D
Healthy
AD
MAZ
Age in weeks
E
Faecalibacterium
Counts
Age in days
F
Lachnobacterium
Age in days
G
Prevotella
Age in days

Figure 4 Microbiota composition, diversity and maturity is linked to the subsequent development of atopic dermatitis (N= 961 stool samples from 312 children). A) Volcano plot depicting the regression coefficients from the joint model on the association between the Shannon index and atopic dermatitis (AD). The dashed line depicts the threshold for statistical significance at $p<0.05$. Variables depicted below the dashed line were statistically significantly associated with AD in the final model. Positive coefficients (variables to the right of the vertical line) were associated with an increased AD risk. Negative coefficients (variables to the left of the vertical line) were associated with a decreased AD risk. The Hazard Ratio is given by the exponent of the coefficient (e.g. for Shannon index: e^-1.57 results in a HR of 0.21). B) Development of microbial diversity (Shannon index) throughout infancy among children that did (red line) or did not (green line) develop AD as modelled by Loess regression. Grey areas represent the 95% confidence intervals. C) Volcano plot depicting the regression coefficients from the joint model on the association between the microbial age z-score (MAZ) and the development of atopic dermatitis (AD). The dashed line depicts the threshold for statistical significance at $p<0.05$. D, Development of microbial maturity (MAZ-score) throughout infancy among children that did (red line) or did not (green line) develop AD as modelled by Loess regression. Grey areas represent the 95% confidence intervals. E-G, Time intervals of differential abundance in Faecalibacterium, E) Lachnobacteriumn, F) and Prevotella, G) between infants that did or did (dashed line) not develop AD (solid line) as identified from MetaLonda analyses. Significantly different time-intervals (FDR-adjusted $p<0.05$) are depicted by gray shading.

Discussion

This study aimed to longitudinally analyse the process of gastrointestinal microbial community assembly, succession, and maturation throughout the most critical time-window of immune development and linked microbiota maturation during this time to the development of clinical signs of allergic disease, while carefully controlling for potential confounding factors.

Our results indicate a dynamic microbiota during infancy which is far from completely matured at 31 weeks of age. In early infancy the microbial composition was most strongly affected by birth mode, while from 13 weeks onwards diet became the most important factor. Our data support previous reports, showing that Bacteroides are most strongly affected by birth mode [29, 34-36]. The difference in microbial community structure and lower abundance of Bacteroides in caesarean section as compared to vaginally delivered infants persisted up to the age of 31 weeks of age and withstood mutual adjustment for other determinants, including breastfeeding. This suggests that the impact of caesarean section delivery could not be compensated by breastfeeding. Given the increased risk of future diseases, including allergies and asthma [37, 38], among children born by caesarean section, more research is warranted to elucidate the need for and efficacy of restoring the natural microbial colonization process upon caesarean delivery.

We furthermore we showed that cessation of breastfeeding was more strongly associated with microbial composition and maturity than solid food introduction. In line with previous results [11, 35, 39], these results suggest that the introduction of solid food does not appear to result in a profound shift in microbial community structure as long as breastfeeding is continued. Only when breastfeeding is ceased, maturation of the microbiota is accelerated with a decrease in degraders of human milk oligosaccharides and an increase in microbial diversity and compositional changes towards bacterial genera specialized in degrading complex dietary carbohydrates. The generally observed lower microbial diversity in infants during breastfeeding [40] seems at first contradictory with the concept that a “healthy” and resilient microbiome is highly diverse [41]. However, in line with most prospective studies [42], we did not find a direct association between breastfeeding duration and the risk of AD. The fact that breastfeeding reduces the risk of several other diseases, including metabolic diseases, which on the other hand are also associated with a lower microbial diversity, suggests that the context is of crucial importance when considering microbial diversity. For example, loss of microbiota diversity generally opens up niches for opportunistic invaders [41], while the plethora of bioactive components transferred by breastfeeding protects against colonization by such opportunistic pathogens [43]. This further underscores the need for a meticulous adjustment for diet as a confounding factor in the association between microbiota and disease outcomes.

Using various multivariable longitudinal analysis, we furthermore demonstrated that the microbial community structure, diversity, and maturity as well as the relative abundance of several individual genera were associated with the subsequent development of allergic manifestation. We know from previous animal studies and large longitudinal human cohorts that intestinal microbial dysbiosis in allergic diseases is mainly observed within a critical window in early life [44]. The comparability between studies,

however, hampered by the highly dynamic microbial communities within this early time window, which likely results in different associations at different sampling time points.

The main strengths of the present study are its prospective design, repeated sample collection and the deep clinical phenotyping. The regular physical examinations of the children throughout the first 3 years of life in combination with the collection of detailed questionnaire data allowed not only deep clinical phenotyping, but also an accurate assessment of the time of disease onset. The follow-up into school-age further facilitated a reliable classification of children who developed allergic asthma as it is well-known that wheezing symptoms at an earlier age are often transient and caused by episodic viral infections [45].

We observed a lower microbial diversity to be associated with AD development and allergic sensitization, but not with asthma. This is consistent with previous studies that also reported a reduced microbial diversity in association to AD [46-49] and sensitization [7, 50]. In contrast, a link between microbial diversity and wheeze or asthma could often not be observed [51-53]. Although atopic manifestations are common comorbidities, these results support previous conclusions from COPSAC that extrapolation of risk factors between different atopic disorders may not always be justified [54].

The increased risk of AD, sensitization and asthma among children with a higher microbial maturity might at first seem in contrast with the above-mentioned results for microbial diversity and with findings of previous studies. Indeed, recent results from COPSAC2010 linked a low microbial maturity with later onset of asthma in children born to asthmatic mothers [11]. In our study, the microbial maturity was however only significantly increased at 5 weeks of age in children who developed sensitization (as determined by Skin Prick Tests and serum IgE levels to the most common aero-allergens) and asthma. Also, for children with AD, we observed a microbial maturity (MAZ) that was significantly higher at the earliest recorded time-point but gradually decreased and became even lower at the age of 31 weeks when compared to the MAZ of children that did not develop AD. This temporally higher MAZ in very young infants might therefore suggest a dysregulated colonization process, e.g., with some bacterial taxa arriving (too) early, rather than a more mature overall microbial community structure.

Next to differences in microbial diversity and maturity, we were able to identify microbial taxa that were differentially abundant among infants who did or did not develop allergic disease manifestations. Lachnobacterium and Faecalibacterium were significantly decreased throughout infancy among children who developed AD. Also, Lachnospira and Dialister, next to Lachnobacterium, were significantly decreased among children who developed asthma.

The fact that these bacterial taxa were not only differentially abundant at a single time-point but throughout infancy strengthens the likelihood of a causal role in the protection against allergic disease. Altogether our results indicate that microbial perturbations in early life are also associated with asthma at school age, although perturbations are not identical to those observed in children that developed AD. In line with our findings, analysis on the microbiota composition at 3 months of age within the Canadian Healthy Infant Longitudinal Development (CHILD) Study revealed Lachnospira and *Faecalibacterium* to be significantly decreased among children at risk for allergic wheeze at the age of 1 year [52]. Moreover, a lower relative abundance of amongst others *Lachnospiraraeceae incertae sedis, Faecalibacterium* and *Dialister* at the age of

1 year in children from COPSAC2010 was associated with an increased risk of asthma at 5 years [11]. Fermentation products of these bacteria are a possible explanation for the protective effect of these bacteria. Acetate is one of the fermentation products of *Lachnospira* and to a lesser account *Lachnobacterium*. Animal studies have previously shown that acetate-feeding leads to a marked suppression of allergic airway disease in a mouse-model for human asthma. The underlying cellular mechanism was related to the effect of acetate on regulatory T (T-reg) cells, particularly through epigenetic modification of the Foxp3 promotor [5].

Faecalibacterium (*prausnitzii*) is well-known for its anti-inflammatory effects, amongst others through the production of butyrate [55] and a microbial anti-inflammatory molecule (MAM) that inhibits the pro-inflammatory NF-κB pathway [56]. Two recent studies have identified another lactate-consuming butyrate-producing genus, Anaerostipes, associated with a decreased risk of food allergy [57] and eczema [58]. The very low abundance of this genus in our population could potentially explain the lack of association in our study.

The application of several types of longitudinal data-analysis, including the joint modelling of longitudinal and survival data which had previously not been used for microbiota data analyses, enabled us to demonstrate that alterations in microbial diversity, maturity and composition preceded the clinical manifestations of atopic diseases. Although this statistical framework reveals the temporality of associations thereby suggesting causal relationships, causality can never be proven in an observational study. For example, microbial perturbations could be an epiphenomenon of exposure to yet another unknown risk factor for allergy. Also, it cannot be ruled out that early preclinical manifestation of allergies or genetic predisposition for allergy might impact the microbiota composition. Moreover, our findings on faecal collections might not fully reflect alterations in the microbiome on allergy development at the level of (small) intestinal mucosa.

It is therefore of importance not only to replicate findings in similar cohorts, but also to conduct future experimental studies building upon these findings in order to reveal the underlying biological mechanisms and prove causality.

In conclusion, our results demonstrate the importance of birth mode and diet on the early maturation of the infant microbiota and demonstrate that, upon careful adjustment of important confounding factors, alterations in the microbial colonization process of the infant intestinal tract precede the development of AD, sensitization, and asthma. In particular, members of the Lachnospiraceae family, as well as the genera *Faecalibacterium* and *Dialister* appear to protect against allergies. These findings further support the future development of evidence-based intervention strategies targeting the microbiota to prevent or treat allergic diseases in early life.

Acknowledgments

We thank Christel Driessen for lab analyses and Mayk Lucchesi for designing the graphical abstract.

References

1. Bennek, E., et al., Subcellular antigen localization in commensal E. coli is critical for T cell activation and induction of specific tolerance. Mucosal Immunol, 2019. 12(1): p. 97-107.
2. Cahenzli, J., et al., Intestinal microbial diversity during early-life colonization shapes long-term IgE levels. Cell Host Microbe, 2013. 14(5): p. 559-70.
3. Stefka, A.T., et al., Commensal bacteria protect against food allergen sensitization. Proc Natl Acad Sci U S A, 2014. 111(36): p. 13145-50.
4. Olszak, T., et al., Microbial exposure during early life has persistent effects on natural killer T cell function. Science, 2012. 336(6080): p. 489-93.
5. Thorburn, A.N., et al., Evidence that asthma is a developmental origin disease influenced by maternal diet and bacterial metabolites. Nat Commun, 2015. 6: p. 7320.
6. Trompette, A., et al., Gut microbiota metabolism of dietary fiber influences allergic airway disease and hematopoiesis. Nat Med, 2014. 20(2): p. 159-66.
7. Azad, M.B., et al., Infant gut microbiota and food sensitization: associations in the first year of life. Clin Exp Allergy, 2015. 45(3): p. 632-43.
8. Dzidic, M., et al., Aberrant IgA responses to the gut microbiota during infancy precede asthma and allergy development. J Allergy Clin Immunol, 2017. 139(3): p. 1017-1025 e14.
9. Penders, J., et al., The role of the intestinal microbiota in the development of atopic disorders. Allergy, 2007. 62(11): p. 1223-36.
10. Simonyte Sjodin, K., et al., Temporal and long-term gut microbiota variation in allergic disease: A prospective study from infancy to school age. Allergy, 2019. 74(1): p. 176-185.
11. Stokholm, J., et al., Maturation of the gut microbiome and risk of asthma in childhood. Nat Commun, 2018. 9(1): p. 141.
12. Fujimura, K.E., et al., Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. Nat Med, 2016. 22(10): p. 1187-1191.
13. Zimmermann, P., et al., Association between the intestinal microbiota and allergic sensitization, eczema, and asthma: A systematic review. J Allergy Clin Immunol, 2019. 143(2): p. 467-485.
14. Simonyte Sjodin, K., et al., Emerging evidence of the role of gut microbiota in the development of allergic diseases. Curr Opin Allergy Clin Immunol, 2016. 16(4): p. 390-5.
15. Korpela, K., et al., Selective maternal seeding and environment shape the human gut microbiome. Genome Res, 2018. 28(4): p. 561-568.
16. Wampach, L., et al., Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. Nat Commun, 2018. 9(1): p. 5091.
17. Gilbert, J.A. and S.V. Lynch, Community ecology as a framework for human microbiome research. Nat Med, 2019. 25(6): p. 884-889.
18. Davidson, R., et al., Influence of maternal and perinatal factors on subsequent hospitalisation for asthma in children: evidence from the Oxford record linkage study. BMC Pulm Med, 2010. 10: p. 14.
19. van Nimwegen, F.A., et al., Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. J Allergy Clin Immunol, 2011. 128(5): p. 948-55 e1-3.
20. Lau, S., et al., Oral application of bacterial lysate in infancy decreases the risk of atopic dermatitis in children with 1 atopic parent in a randomized, placebo-controlled trial. J Allergy Clin Immunol, 2012. 129(4): p. 1040-7.
21. Penders, J., et al., Establishment of the intestinal microbiota and its role for atopic dermatitis in early childhood. J Allergy Clin Immunol, 2013. 132(3): p. 601-607 e8.
22. Bartram, A.K., et al., Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. Appl Environ Microbiol, 2011. 77(11): p. 3846-52.
23. Whelan, F.J. and M.G. Surette, A comprehensive evaluation of the sl1p pipeline for 16S rRNA gene sequencing analysis. Microbiome, 2017. 5(1): p. 100.
24. Lagkouvardos, I., et al., Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ, 2017. 5: p. e2836.
25. Oksanen, J., et al., Vegan: Community Ecology Package. R Package Version. 2.5-3. CRAN. 2013.
26. Chen, J., GUniFrac: Generalized UniFrac Distances. R Package Version. 1.1. CRAN. 2018.
27. Subramanian, S., et al., Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature, 2014. 510(7505): p. 417-21.
28. Holmes, I., K. Harris, and C. Quince, Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One, 2012. 7(2): p. e30126.
29. Stewart, C.J., et al., Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature, 2018. 562(7728): p. 583-588.
30. Morgan, X.C., et al., Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol, 2012. 13(9): p. R79.
31. Cadarso Suarez, C., et al., Editorial "Joint modeling of longitudinal and time-to-event data and beyond". Biom J, 2017. 59(6): p. 1101-1103.
32. Rizopoulos, D., JM: An R Package for the Joint Modelling of Longitudinal and Time-to-event Data. Journal of Statistical Software, 2010. 35(9).
33. Metwally, A.A., et al., MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. Microbiome, 2018. 6(1): p. 32.
34. Azad, M.B., et al., Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet

at 4 months. CMAJ, 2013. 185(5): p. 385-94.
35. Backhed, F., et al., Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe, 2015. 17(5): p. 690-703.
36. Penders, J., et al., Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics, 2006. 118(2): p. 511-21.
37. Huang, L., et al., Is elective cesarean section associated with a higher risk of asthma? A meta-analysis. J Asthma, 2015. 52(1): p. 16-25.
38. Thavagnanam, S., et al., A meta-analysis of the association between Caesarean section and childhood asthma. Clin Exp Allergy, 2008. 38(4): p. 629-33.
39. Pannaraj, P.S., et al., Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. JAMA Pediatr, 2017. 171(7): p. 647-654.
40. Ho, N.T., et al., Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations. Nat Commun, 2018. 9(1): p. 4169.
41. Bello, M.G.D., et al., Preserving microbial diversity. Science, 2018. 362(6410): p. 33-34.
42. Lin, B., et al., Breastfeeding and Atopic Dermatitis Risk: A Systematic Review and Meta-Analysis of Prospective Cohort Studies. Dermatology, 2019: p. 1-16.
43. van den Elsen, L.W.J., et al., Shaping the Gut Microbiota by Breastfeeding: The Gateway to Allergy Prevention? Front Pediatr, 2019. 7: p. 47.
44. Stiemsma, L.T. and S.E. Turvey, Asthma and the microbiome: defining the critical window in early life. Allergy Asthma Clin Immunol, 2017. 13: p. 3.
45. Townshend, J., S. Hails, and M. McKean, Diagnosis of asthma in children. BMJ, 2007. 335(7612): p. 198-202.
46. Abrahamsson, T.R., et al., Low diversity of the gut microbiota in infants with atopic eczema. J Allergy Clin Immunol, 2012. 129(2): p. 434-40, 440 e1-2.
47. Forno, E., et al., Diversity of the gut microbiota and eczema in early life. Clin Mol Allergy, 2008. 6: p. 11.
48. Ismail, I.H., et al., Reduced gut microbial diversity in early life is associated with later development of eczema but not atopy in high-risk infants. Pediatr Allergy Immunol, 2012. 23(7): p. 674-81.
49. Wang, M., et al., Reduced diversity in the early fecal microbiota of infants with atopic eczema. J Allergy Clin Immunol, 2008. 121(1): p. 129-34.
50. Chen, C.C., et al., Alterations in the gut microbiotas of children with food sensitization in early life. Pediatr Allergy Immunol, 2016. 27(3): p. 254-62.
51. Arrieta, M.C., et al., Associations between infant fungal and bacterial dysbiosis and childhood atopic wheeze in a nonindustrialized setting. J Allergy Clin Immunol, 2018. 142(2): p. 424-434 e10.
52. Arrieta, M.C., et al., Early infancy microbial and metabolic alterations affect risk of childhood asthma. Sci Transl Med, 2015. 7(307): p. 307ra152.
53. Bisgaard, H., et al., Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. J Allergy Clin Immunol, 2011. 128(3): p. 646-52 e1-5.
54. Bisgaard, H., et al., Risk analysis of early childhood eczema. J Allergy Clin Immunol, 2009. 123(6): p. 1355-60 e5.
55. Canani, R.B., et al., Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol, 2011. 17(12): p. 1519-28.
56. Quevrain, E., et al., Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn's disease. Gut, 2016. 65(3): p. 415-425.
57. Feehley, T., et al., Healthy infants harbor intestinal bacteria that protect against food allergy. Nat Med, 2019. 25(3): p. 448-453.
58. Wopereis, H., et al., Intestinal microbiota in infants at high risk for allergy: Effects of prebiotics and role in eczema development. J Allergy Clin Immunol, 2018. 141(4): p. 1334-1342 e5.

Supplementary Tables

Supplementary Table S1 Demographics of the study population

Variable	n (%)*
Sex	
Boys	223 (50.7)
Girls	217 (49.3)
Birth mode	
Natural vaginal delivery	304 (69.1)
Assisted vaginal delivery	26 (5.9)
Cesarean section	109 (24.8)
Number of older siblings	
No older siblings	253 (57.5)
1 older sibling	150 (34.1)
≥ 2 older siblings	37 (8.4)
Eczema mother	
Yes	161 (36.2)
No	279 (63.4)
Eczema father	
Yes	92 (20.9)
No	346 (78.6)
Asthma mother	
Yes	142 (32.3)
No	298 (67.6)
Asthma father	
Yes	102 (23.2)
No	337 (76.6)
Maternal smoking in pregnancy	
No	348 (79.1)
Yes	92 (20.9)
Maternal smoking after pregnancy	
No	341 (77.5)
Yes	99 (22.5)
Households with furry pets	
No	138 (31.4)
Yes	277 (63.0)
Ethnicity	
Caucasian	429 (97.5)
Other	11 (2.5)
Antibiotics before age 31 weeks	
No	411 (93.4)
Yes	29 (6.6%)
**Numbers may not add up to 440 (100%) due to missing data*	

Supplementary Table S1 (cont'd)

Variable	Median (IQR)
Maternal age at delivery (in years)	33 (29-36)
Birth Weight (in grams)	3497 (3202-3800)
Breast feeding in weeks	40 (30-53)
Age at introduction of solids (in weeks)	25 (22-27)
Gestational age (in weeks)	40 (38-40)
Start at daycare (in months)	13 (11-22)

Supplementary Table S2 Microbial richness as assessed by Chao1 index across time-points

Age	Median Chao1	Interquartile range
5 weeks	147.23	41.6 - 188.64
13 weeks	170.8	65.13 - 214.3
21 weeks	182.71	66.11 - 220.88
31 weeks	189.83	74.13 - 239.52
School-age	379.83	216.52 - 432.17

Test	P-value
Friedman test	7.724E-53

Variable 1	Variable 2	Test	P-value
Chao1 at 5 weeks	Chao1 at 13 weeks	Dunn's test	**1.37E-02**
Chao1 at 5 weeks	Chao1 at 21 weeks	Dunn's test	**1.52E-05**
Chao1 at 5 weeks	Chao1 at 31 weeks	Dunn's test	**1.65E-07**
Chao1 at 5 weeks	Chao1 at school-age	Dunn's test	**7.55E-21**
Chao1 at 13 weeks	Chao1 at 21 weeks	Dunn's test	**4.09E-03**
Chao1 at 13 weeks	Chao1 at 31 weeks	Dunn's test	**4.09E-03**
Chao1 at 13 weeks	Chao1 at school-age	Dunn's test	**7.68E-21**
Chao1 at 21 weeks	Chao1 at 31 weeks	Dunn's test	9.09E-01
Chao1 at 21 weeks	Chao1 at school-age	Dunn's test	**7.55E-21**
Chao1 at 31 weeks	Chao1 at school-age	Dunn's test	**8.56E-21**

**Related samples statistics based only on children with samples collected at every single time-point (N = 606 samples of 121 children)*

Supplementary Table S3 Testing for statistically significant associations between age of sample collection and principal coordinates of PCo1 as depicted in Figure 1

Test	P-value
Friedman	6.002E-51

Variable 1	Variable 2	Test	P-value
PCo1 at 5 weeks	PCo1 at 13 weeks	Dunn's test	**2.32E-04**
PCo1 at 5 weeks	PCo1 at 21 weeks	Dunn's test	**2.12E-03**
PCo1 at 5 weeks	PCo1 at 31 weeks	Dunn's test	**5.65E-34**
PCo1 at 5 weeks	PCo1 at school-age	Dunn's test	**1.21E-27**
PCo1 at 13 weeks	PCo1 at 21 weeks	Dunn's test	3.27E-01
PCo1 at 13 weeks	PCo1 at 31 weeks	Dunn's test	**1.21E-27**
PCo1 at 13 weeks	PCo1 at school-age	Dunn's test	**8.53E-27**
PCo1 at 21 weeks	PCo1 at 31 weeks	Dunn's test	**1.58E-27**
PCo1 at 21 weeks	PCo1 at school-age	Dunn's test	**1.03E-25**
PCo1 at 31 weeks	PCo1 at school-age	Dunn's test	**3.99E-28**

Related samples statistics based only on children with samples collected at every single time-point (N = 606 samples of 121 children)

Supplementary Table S4 Testing for statistically significant associations between age of sample collection and relative abundance of the main bacterial genera as depicted in Supplementary Figures 2A-B

		Akkermansia	*Prevotella*	*Oscillospira*	*Subdoligranulum*	*Haemophilus*
	†	**7.69E-37**	**4.60E-05**	**3.76E-45**	**5.91E-75**	**1.48E-11**
	§					
Time-point 1	**Time-point 2**					
5 weeks	13 weeks	NS	NS	NS	NS	NS
5 weeks	21 weeks	<0.001	NS	NS	NS	NS
5 weeks	31 weeks	0.002	NS	<0.001	NS	<0.001
5 weeks	school-age	<0.001	NS	<0.001	<0.001	0.012
13 weeks	21 weeks	0.012	NS	NS	NS	NS
13 weeks	31 weeks	0.039	NS	NS	NS	0.012
13 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001
21 weeks	31 weeks	NS	NS	NS	NS	0.042
21 weeks	school-age	<0.001	0.037	<0.001	<0.001	<0.001
31 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001

		Lachnospira	*Citrobacter*	*Pseudobutyrivibrio*	*Ruminococcus*	*Lactococc*
	†	**8.79E-57**	**4.37E-39**	**4.96E-64**	**6.30E-55**	**7.75E-42**
	§					
Time-point 1	**Time-point 2**					
5 weeks	13 weeks	NS	NS	NS	NS	NS
5 weeks	21 weeks	NS	0.05	NS	0.005	NS
5 weeks	31 weeks	<0.001	NS	0.028	<0.001	NS
5 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001
13 weeks	21 weeks	NS	NS	NS	NS	NS
13 weeks	31 weeks	<0.001	NS	NS	0.023	NS
13 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001
21 weeks	31 weeks	<0.001	NS	NS	NS	NS
21 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001
31 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001

		Parabacteroides	*Ruminococcus*	*Faecalibacterium*	*Blautia*	*Clostridium*
	†	**1.54E-11**	**3.75E-78**	**8.95E-59**	**2.16E-59**	**2.55E-10**
	§					
Time-point 1	**Time-point 2**					
5 weeks	13 weeks	NS	NS	NS	NS	NS
5 weeks	21 weeks	NS	0.005	NS	0.012	NS
5 weeks	31 weeks	NS	<0.001	0.002	<0.001	<0.001
5 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001
13 weeks	21 weeks	NS	NS	NS	NS	NS
13 weeks	31 weeks	NS	0.023	<0.001	0.003	0.005
13 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001
21 weeks	31 weeks	NS	NS	<0.001	NS	0.007

Supplementary Table S4 (cont'd)

21 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001
31 weeks	school-age	<0.001	<0.001	<0.001	<0.001	<0.001

		Klebsiella	*Streptococcus*	*Veillonella*	*Bacteroides*	*Bifidobacterium*	*Escherichia*
	†	**3.42E-39**	**1.61E-14**	**6.00E-53**	**0.000067**	**3.00E-01**	**2.56E-40**
	§						
Time-point 1	**Time-point 2**						
5 weeks	13 weeks	NS	NS	NS	NS	-	NS
5 weeks	21 weeks	NS	NS	NS	NS	-	NS
5 weeks	31 weeks	NS	<0.001	<0.001	<0.001	-	NS
5 weeks	school-age	<0.001	<0.001	<0.001	NS	-	<0.001
13 weeks	21 weeks	NS	NS	NS	NS	-	NS
13 weeks	31 weeks	NS	<0.001	<0.001	0.011	-	0.003
13 weeks	school-age	<0.001	<0.001	<0.001	NS	-	<0.001
21 weeks	31 weeks	NS	<0.001	0.007	<0.001	-	0.02
21 weeks	school-age	<0.001	<0.001	<0.001	NS	-	<0.001
31 weeks	school-age	<0.001	NS	<0.001	NS	-	<0.001

Related samples statistics based only on children with samples collected at every single time-point (N = 606 samples of 121 children)

†*Friedman-test p-values*

§ *Post-hoc analyses Dunn's test Bonferroni adjusted p-values*

Supplementary Table S5 Multinomial logistic regression analyses on the association between demographic, lifestyle and medical factors in association with DMM clusters and trajectories

	FINAL MULTIVARIABLE MULTINOMIAL LOGISTIC REGRESSION MODEL ON COVARIATES IN ASSOCIATION TO DMM CLUSTERS AT 5 WEEKS OF AGE.			
	Variable	**Exp(B)**	**95% Confidence Interval for Exp (B)**	**P-value**
DMM Cluster 1 (n = 75) - Reference category	-	-	-	-
DMM Cluster 2 (n = 29)	**Breastfed (in weeks)**	0.998	0.983-1.014	8.27E-01
	C-section delivery	1.610	0.520 - 4.974	4.10E-01
	Eczema mother	1.574	0.476 - 5.204	4.57E-01
	Asthma father	0.685	0.188 - 2.497	5.67E-01
	1 older sibling	2.097	0.815 - 5.398	1.25E-01
	≥ 2 older siblings	2.931	0.645 - 13.315	1.64E-01
DMM Cluster 3 (n = 76)	**Breastfed (in weeks)**	0.997	0.985 - 1.010	6.81E-01
	C-section delivery	**3.346**	**1.488 - 7.528**	**3.00E-03**
	Eczema mother	2.238	0.895 - 5.596	8.50E-02
	Asthma father	0.700	0.274 - 1.787	4.56E-01
	1 older sibling	1.019	0.487 - 2.133	9.60E-01
	≥ 2 older siblings	0.765	0.180 - 3.254	7.17E-01
DMM Cluster 4 (n = 20)	**Breastfed (in weeks)**	**0.965**	**0.938 - 0.992**	**1.20E-02**
	C-section delivery	0.242	0.028 - 2.109	1.99E-01
	Eczema mother	2.564	0.713 - 9,220	1.49E-01
	Asthma father	**3.338**	**1.006 - 11.075**	**4.90E-02**
	1 older sibling	2.527	0.802 - 7.956	1.13E-01
	≥ 2 older siblings	2.815	0.443 - 17.903	2.73E-01
DMM Cluster 5 (n = 11)	**Breastfed (in weeks)**	0.990	0.964 - 1.017	4.57E-01
	C-section delivery	**6.347**	**1.467 - 27.456**	**1.30E-02**
	Eczema mother	0,.808	0.105 - 6.212	8.37E-01
	Asthma father	0.931	0.140 - 6.175	9.41E-01
	1 older sibling	1.635	0.323 - 8.277	5.52E-01
	≥ 2 older siblings	**16.669**	**2.535 - 109.587**	**3.00E-03**
DMM Cluster 6 (n = 14)	**Breastfed (in weeks)**	1.013	0.993 - 1.034	1.98E-01
	C-section delivery	1.371	0.310 - 6.065	6.78E-01
	Eczema mother	2.233	0.514 - 9.703	2.84E-01
	Asthma father	3.455	0.937 - 12.748	6.30E-02
	1 older sibling	0.343	0.066 - 1.783	2.03E-01
	≥ 2 older siblings	0.502	0.045 - 5.555	5.74E-01

Supplementary Table S5 (cont'd)

FINAL MULTIVARIABLE MULTINOMIAL LOGISTIC REGRESSION MODEL ON COVARIATES IN ASSOCIATION TO MOST COMMON DMM CLUSTER TRAJECTORIES

	Variable	Exp(B)	95% Confidence Interval for Exp (B)	P-value
DMM Cluster trajectory 1 -> 1 (n =26) - Reference category	-	-	-	-
DMM Cluster trajectory 1 -> 4 (n = 22)	**Breastfed (in weeks)**	**0.959**	**0.928 - 0.992**	**1.40E-02**
	Age introduction solids (weeks)	0.943	0.774-1.148	5.57E-01
	Age at start day care (months)	0.958	0.876 - 1.048	3.53E-01
DMM Cluster trajectory 1 -> 5 (n = 14)	**Breastfed (in weeks)**	**0.948**	**0.909 - 0.989**	**1.20E-02**
	Age introduction solids (weeks)	0.818	0.660 - 1.014	6.70E-02
	Age at start day care (months)	1.004	0.909 - 1.108	0.944
DMM Cluster trajectory 3 -> 1 (n = 16)	**Breastfed (in weeks)**	0.980	0.950 - 1.010	1.92E-01
	Age introduction solids (weeks)	1.074	0.876 - 1.315	4.93E-01
	Age at start day care (months)	0.981	0.890 - 1.081	6.95E-01
DMM Cluster trajectory 3 -> 3 (n = 11)	**Breastfed (in weeks)**	0.997	0.958 - 1.037	8.69E-01
	Age introduction solids (weeks)	0.836	0.636 - 1.100	2.01E-01
	Age at start day care (months)	**1.130**	**1.016 - 1.257**	**2.50E-02**
DMM Cluster trajectory 3 -> 4 (n = 13)	**Breastfed (in weeks)**	**0.939**	**0.896 - 0.985**	**1.00E-02**
	Age introduction solids (weeks)	**0.776**	**0.618 - 0.974**	**2.80E-02**
	Age at start day care (months)	0.970	0.865 - 1.087	5.97E-01
DMM Cluster trajectory 3 -> 5 (n = 25)	**Breastfed (in weeks)**	**0.949**	**0.916 - 0.983**	**4.00E-03**
	Age introduction solids (weeks)	0.905	0.744 - 1.102	3.21E-01
	Age at start day care (months)	1.009	0.925 - 1.099	8.44E-01

Supplementary Table S6 Multivariate Analysis by Linear Models (MaAsLin) on demographic, lifestyle, and medical factors in association with genus level taxa

Multivariate Analysis by Linear Models (MaAsLin) on demographic, lifestyle and medical factors associated with genus level taxa (only results with statistically significant associations upon FDR-correction are presented)								
Age	**Variable**	**Feature (genus)**	**Value**	**Coefficient**	**N**	**N not 0**	**P-value**	**FDR-adjusted P-value (Q)**
5 weeks	Birth Mode	Bacteroides	Vaginal delivery	0.175837645	297	297	1.38E-05	8.04E-03
5 weeks	Older siblings	Bifidobacterium	Older siblings in households	0.166120378	297	297	1.55E-05	8.04E-03
5 weeks	Older siblings	Corynebacterium	Older siblings in households	0.002526116	297	138	2.06E-05	8.04E-03
13 weeks	Cessation breastfeeding prior to sample collection	Pantoea	Breastfeeding ceased	0.00131711	286	79	2.96E-05	4.16E-02
21 weeks	Birth Mode	Bacteroides	Vaginal delivery	0.14714048	267	267	7.16E-04	6.28E-02
21 weeks	Birth Mode	Campylobacter	Vaginal delivery	-0.001529738	267	77	1.11E-03	8.25E-02
21 weeks	Cessation breastfeeding prior to sample collection	Other	Breastfeeding ceased	0.139999237	267	267	1.51E-09	1.72E-06
21 weeks	Cessation breastfeeding prior to sample collection	Lachnobacterium	Breastfeeding ceased	0.003857002	267	145	2.31E-09	1.72E-06
21 weeks	Cessation breastfeeding prior to sample collection	Pseudobutyrivibrio	Breastfeeding ceased	0.002832593	267	95	8.08E-08	4.02E-05
21 weeks	Cessation breastfeeding prior to sample collection	Bifidobacterium	Breastfeeding ceased	-0.197143826	267	267	7.22E-06	1.79E-03

Supplementary Table S6 (cont'd)

21 weeks	Cessation breastfeeding prior to sample collection	Clostridium_C	Breastfeeding ceased	0.009197433	267	154	1.47E-05	3.14E-03
21 weeks	Cessation breastfeeding prior to sample collection	Oscillospira	Breastfeeding ceased	0.002240703	267	155	2.20E-05	4.10E-03
21 weeks	Cessation breastfeeding prior to sample collection	Eubacterium_A	Breastfeeding ceased	0.002266881	267	81	5.18E-05	8.59E-03
21 weeks	Cessation breastfeeding prior to sample collection	Ruminococcus	Breastfeeding ceased	0.007854753	267	194	6.28E-05	9.37E-03
21 weeks	Cessation breastfeeding prior to sample collection	Blautia	Breastfeeding ceased	0.002264146	267	203	2.04E-04	2.76E-02
21 weeks	Cessation breastfeeding prior to sample collection	Staphylococcus	Breastfeeding ceased	-0.004207764	267	195	2.93E-04	3.36E-02
21 weeks	Cessation breastfeeding prior to sample collection	Clostridium	Breastfeeding ceased	0.040367347	267	265	6.21E-04	6.28E-02
21 weeks	Cessation breastfeeding prior to sample collection	Megasphaera	Breastfeeding ceased	0.001435393	267	85	7.13E-04	6.28E-02
21 weeks	Cessation breastfeeding prior to sample collection	Enterococcus	Breastfeeding ceased	-0.015797733	267	263	1.11E-03	8.25E-02
21 weeks	Ethnicity	Proteus	non Caucasian	0.033164297	267	49	8.31E-04	6.88E-02

Supplementary Table S5 (cont'd)

21 weeks	Older siblings	Eggerthella	Older siblings in households	0.006674972	267	124	4.98E-06	1.49E-03
21 weeks	Smoking_In_Pregnancy	Veillonella	Smoking_In_Pregnancy2	-0.088597495	267	267	2.60E-04	3.23E-02
21 weeks	Solids introduced prior to sample collection	Campylobacter	Solids introduced	0.002285898	267	77	2.94E-06	1.10E-03
21 weeks	Solids introduced prior to sample collection	Clostridium_C	Solids introduced	0.006088278	267	154	6.57E-04	6.28E-02
31 weeks	AD status at time of sample collection	Citrobacter	AD present	-0.0083787	304	288	1.19E-03	8.39E-02
31 weeks	Birth Mode	Bacteroides	Vaginal delivery	0.1860148	304	304	4.30E-05	5.14E-03
31 weeks	Birth Mode	Raoultella	Vaginal delivery	-0.0064118	304	294	1.05E-03	7.75E-02
31 weeks	Birth weight (in grams)	Corynebacterium	Birth weight (in grams)	1.22E-06	304	93	4.53E-04	4.14E-02
31 weeks	Cessation breastfeeding prior to sample collection	Oscillospira	Breastfeeding ceased	0.04809638	304	209	1.18E-17	1.83E-14
31 weeks	Cessation breastfeeding prior to sample collection	Ruminococcus	Breastfeeding ceased	0.06942166	304	259	1.04E-15	8.05E-13
31 weeks	Cessation breastfeeding prior to sample collection	Other	Breastfeeding ceased	0.1276393	304	304	1.13E-11	5.86E-09
31 weeks	Cessation breastfeeding prior to sample collection	Pseudobutyrivibrio	Breastfeeding ceased	0.00440686	304	147	1.53E-10	5.95E-08

Supplementary Table S6 (cont'd)

31 weeks	Cessation breastfeeding prior to sample collection	Roseburia	Breastfeeding ceased	0.00412616	304	60	6.19E-10	1.92E-07
31 weeks	Cessation breastfeeding prior to sample collection	Lachnobacterium	Breastfeeding ceased	0.00304302	304	182	2.20E-09	5.71E-07
31 weeks	Cessation breastfeeding prior to sample collection	Coprobacillus	Breastfeeding ceased	0.00542691	304	31	4.50E-09	9.99E-07
31 weeks	Cessation breastfeeding prior to sample collection	Eubacterium_A	Breastfeeding ceased	0.00411655	304	156	6.38E-09	1.24E-06
31 weeks	Cessation breastfeeding prior to sample collection	Bifidobacterium	Breastfeeding ceased	-0.1679934	304	304	1.06E-07	1.82E-05
31 weeks	Cessation breastfeeding prior to sample collection	Bacteroides	Breastfeeding ceased	0.20501827	304	304	6.47E-06	1.00E-03
31 weeks	Cessation breastfeeding prior to sample collection	Clostridium_C	Breastfeeding ceased	0.01510415	304	223	8.10E-06	1.14E-03
31 weeks	Cessation breastfeeding prior to sample collection	Staphylococcus	Breastfeeding ceased	-0.0034376	304	216	3.92E-05	5.07E-03
31 weeks	Cessation breastfeeding prior to sample collection	Actinomyces	Breastfeeding ceased	-0.0055536	304	276	1.61E-04	1.67E-02

Supplementary Table S6 (cont'd)

31 weeks	Cessation breastfeeding prior to sample collection	Streptococcus	Breastfeeding ceased	-0.043623	304	304	3.06E-04	2.97E-02
31 weeks	Cessation breastfeeding prior to sample collection	Blautia	Breastfeeding ceased	0.00395886	304	231	8.53E-04	6.98E-02
31 weeks	Cessation breastfeeding prior to sample collection	Escherichia	Breastfeeding ceased	-0.135552	304	304	9.71E-04	7.54E-02
31 weeks	Cessation breastfeeding prior to sample collection	Veillonella	Breastfeeding ceased	-0.0786019	304	304	1.41E-03	9.50E-02
31 weeks	Smoking during pregnancy	Bilophila	Yes	-0.0073827	304	49	5.82E-04	5.02E-02
31 weeks	Solids introduced prior to sample collection	Veillonella	Solids introduced	0.14597588	304	304	4.96E-05	5.50E-03

Supplementary Table S7 Permutational analyses of variance on demographic, dietary, life-style and medical factors associated with microbial community structure

Permutational analysis of the variance at age 5 weeks (N = 238)					
Variable	**Df**	**SumOfSqs**	**R2**	**F**	**P-value**
Asthma Mother	1	0.049	0.002	0.556	9.59E-01
Ethnicity	1	0.063	0.003	0.716	8.13E-01
Birth Mode	2	0.291	0.014	1.654	**1.70E-02**
Day care	1	0.095	0.005	1.079	3.22E-01
Smoke DP	1	0.112	0.005	1.276	1.69E-01
Smoke AP	1	0.102	0.005	1.157	2.56E-01
Breastfed	1	0.094	0.004	1.066	3.90E-01
Residual	229	20.169	0.962		
Total	237	20.976	1.000		

Permutational analysis of the variance at age 13 weeks (N = 233)					
Variable	**Df**	**SumOfSqs**	**R2**	**F**	**P-value**
Asthma Mother	1	0.082	0.004	0.997	4.47E-01
Ethnicity	1	0.074	0.004	0.903	5.69E-01
Birth Mode	2	0.169	0.009	1.025	4.07E-01
Day care	1	0.099	0.005	1.202	2.25E-01
Smoke DP	1	0.107	0.005	1.299	1.74E-01
Smoke AP	1	0.122	0.006	1.486	7.39E-02
Breastfed	1	0.305	0.016	3.717	**9.99E-04**
Solids	1	0.119	0.006	1.448	9.19E-02
AD	1	0.085	0.004	1.038	3.81E-01
Residual	222	18.241	0.940		
Total	232	19.403	1.000		

Permutational analysis of the variance at age 21 weeks (N = 231)					
Variable	**Df**	**SumOfSqs**	**R2**	**F**	**P-value**
Asthma Mother	1	0.103	0.006	1.360	1.25E-01
Ethnicity	1	0.113	0.006	1.498	8.19E-02
Birth Mode	2	0.171	0.009	1.132	2.36E-01
Day care	1	0.099	0.006	1.319	1.33E-01
Smoke DP	1	0.111	0.006	1.466	7.79E-02
Smoke AP	1	0.099	0.005	1.312	1.27E-01
Breastfed	1	0.468	0.026	6.212	**9.99E-04**
Solids	1	0.103	0.006	1.370	1.15E-01
AD	1	0.135	0.007	1.785	**2.00E-02**
Residual	220	16.589	0.922		
Total	230	17.991	1.000		

Permutational analysis of the variance at age 31 weeks (N = 259)					
Variable	**Df**	**SumOfSqs**	**R2**	**F**	**P-value**
Asthma Mother	1	0.046	0.002	0.516	9.84E-01
Ethnicity	1	0.148	0.006	1.647	**4.80E-02**
Birth Mode	2	0.244	0.010	1.359	**4.90E-02**
Day care	1	0.098	0.004	1.093	2.93E-01
Smoke DP	1	0.152	0.006	1.694	3.70E-02
Smoke AP	1	0.078	0.003	0.871	6.28E-01
Breastfed	1	0.663	0.027	7.377	**9.99E-04**
Solids	1	0.180	0.007	1.999	**2.00E-03**
AD	1	0.241	0.010	2.679	**9.99E-04**
Residual	248	22.287	0.923		
Total	258	24.137	1.000		

Supplementary Table S8 Multivariable linear regression analyses on demographic, dietary, lifestyle and medical factors associated with microbial maturity (Microbial Age z-scores)

MAZ index at age 5 weeks (N = 235)							
	Unstandardized Coefficients Beta	Std. Error	Standardized Coefficients Beta	t	P-value	95,0% Confidence Interval for B	
						Lower Bound	Upper Bound
(Constant)	-2.378	5.265		-0.452	6,52E+02	-12.757	8.000
Birthmode assisted vaginal	-0.476	0.680	-0.040	-0.701	4,84E+02	-1.816	0.863
Birthmode caesarean section	0.711	0.367	0.111	1.937	5,40E+01	-0.013	1.434
1 older sibling	-0.362	0.371	-0.060	-0.976	3,30E+02	-1.092	0.369
≥ 2 older siblings	0.763	0.620	0.078	1.230	2,20E+02	-0.459	1.984
Sex	-0.287	0.320	-0.051	-0.897	3,71E+02	-0.918	0.344
Birth Weight	-0.001	0.000	-0.077	-1.220	2,24E+02	-0.001	0.000
Gestational age	0.175	0.136	0.082	1.284	2,01E+02	-0.094	0.444
Breastfeeding duration	-0.010	0.007	-0.097	-1.507	1,33E+02	-0.024	0.003
Ethnicity	0.104	1.022	0.006	0.102	9,19E+02	-1.910	2.119
Maternal age delivery	-0.035	0.035	-0.060	-1.003	3,17E+02	-0.104	0.034
AD father	0.279	0.405	0.043	0.689	4,92E+02	-0.519	1.077

Supplementary Table S8 (cont'd)

AD mother	-0.694	0.436	-0.103	-1.592	1,13E+02	-1.553	0.165
Asthma father	0.417	0.407	0.060	1.023	3,08E+02	-0.386	1.220
Asthma mother	**0.894**	**0.366**	**0.152**	**2.444**	**1,50E+01**	**0.173**	**1.615**
Smoking in pregnancy	0.364	0.393	0.056	0.928	3,54E+02	-0.410	1.138
Smoking after pregnancy	-0.168	0.401	-0.026	-0.418	6,76E+02	-0.959	0.623
AD	**3.203**	**0.356**	**0.536**	**8.986**	**0.000E+00**	**2.500**	**3.905**
Household with furry pets	0.179	0.339	0.030	0.528	5,98E+02	-0.490	0.848
Start of daycare in months	-0.014	0.021	-0.039	-0.647	5,18E+02	-0.055	0.028
Age of introduction solids (months)	-0.020	0.043	-0.030	-0.477	6,34E+02	-0.105	0.064
antibiotic use prior to sample collection	-0.116	1.717	-0.004	-0.068	9,46E+02	-3.500	3.268

Supplementary Table S8 (cont'd)

	MAZ index at age 13 weeks (N = 231)						
	Unstandardized Coefficients Beta	Std. Error	Standardized Coefficients Beta	t	P-value	95,0% Confidence Interval for B	
						Lower Bound	Upper Bound
(Constant)	5.589	5.543		1.008	3,14E+02	-5.338	16.516
Birthmode assisted vaginal	0.262	0.734	0.025	0.357	7,22E+02	-1.185	1.709
Birthmode caesarean section	0.337	0.384	0.060	0.878	3,81E+02	-0.420	1.094
1 older sibling	0.172	0.382	0.033	0.449	6,54E+02	-0.581	0.925
≥ 2 older siblings	0.598	0.674	0.071	0.887	3,76E+02	-0.731	1.927
Sex	-0.138	0.336	-0.028	-0.410	6,82E+02	-0.799	0.524
Birth Weight	-0.001	0.000	-0.147	-1.937	5,40E+01	-0.002	0.000
Gestational age	-0.059	0.143	-0.031	-0.409	6,83E+02	-0.341	0.224
Breastfeeding duration	-0.008	0.007	-0.087	-1.111	2,68E+02	-0.022	0.006
Ethnicity	-0.764	1.009	-0.054	-0.757	4,50E+02	-2.753	1.225
Maternal age delivery	0.019	0.037	0.038	0.523	6,01E+02	-0.054	0.092
AD father	-0.043	0.410	-0.008	-0.105	9,16E+02	-0.852	0.766
AD mother	0.211	0.469	0.036	0.451	6,53E+02	-0.714	1.137
Asthma father	0.724	0.450	0.115	1.610	1,09E+02	-0.162	1.610
Asthma mother	0.293	0.379	0.057	0.774	4,40E+02	-0.453	1.039

Supplementary Table S8 (cont'd)

Smoking in pregnancy	0.296	0.400	0.054	0.740	4,60E+02	-0.492	1.084
Smoking after pregnancy	-0.349	0.403	-0.064	-0.867	3,87E+02	-1.144	0.445
AD	0.052	0.366	0.010	0.142	8,87E+02	-0.669	0.773
Household with furry pets	-0.136	0.347	-0.027	-0.391	6,96E+02	-0.820	0.549
Start of daycare in months	0.023	0.018	0.085	1.229	2,21E+02	-0.014	0.059
Age of introduction solids (months)	-0.005	0.045	-0.009	-0.113	9,10E+02	-0.093	0.083
antibiotic use prior to sample collection	-0.021	0.054	-0.026	-0.388	6,99E+02	-0.128	0.086

MAZ index at age 21 weeks (N - 228)							
	Unstandardized Coefficients Beta	Std. Error	**Standardized Coefficients Beta**	**t**	**P-value**	**95,0% Confidence Interval for B**	
						Lower Bound	**Upper Bound**
(Constant)	8.762	5.692		1.539	1,25E+02	-2.463	19.986
Birthmode assisted vaginal	-0.628	0.734	-0.056	-0.856	3,93E+02	-2.076	0.820

Supplementary Table S8 (cont'd)

Birthmode caesarean section	-0.652	0.395	-0.110	-1.651	1,00E+02	-1.430	0.126
1 older sibling	-0.134	0.384	-0.025	-0.349	7,27E+02	-0.891	0.623
≥ 2 older siblings	0.985	0.632	0.113	1.558	1,21E+02	-0.262	2.232
Sex	**0.727**	**0.335**	**0.143**	**2.171**	**3,10E+01**	**0.067**	**1.386**
Birth Weight	0.000	0.000	0.004	0.055	9,56E+02	-0.001	0.001
Gestational age	-0.105	0.150	-0.052	-0.700	4,85E+02	-0.400	0.190
Breastfeeding duration	-0.004	0.007	-0.040	-0.542	5,88E+02	-0.018	0.010
Ethnicity	**-4.363**	**1.117**	**-0.255**	**-3.907**	**0.000E+00**	**-6.565**	**-2.161**
Maternal age delivery	-0.047	0.036	-0.091	-1.289	1,99E+02	-0.118	0.025
AD father	-0.437	0.417	-0.076	-1.049	2,96E+02	-1.260	0.385
AD mother	-0.243	0.465	-0.039	-0.522	6,02E+02	-1.160	0.674
Asthma father	0.072	0.429	0.012	0.167	8,68E+02	-0.775	0.918
Asthma mother	**0.838**	**0.382**	**0.156**	**2.191**	**3,00E+01**	**0.084**	**1.592**
Smoking in pregnancy	-0.504	0.421	-0.083	-1.198	2,32E+02	-1.333	0.325
Smoking after pregnancy	**1.106**	**0.419**	**0.189**	**2.637**	**9,00E+00**	**0.279**	**1.933**
AD**	**0.917**	**0.358**	**0.172**	**2.559**	**1,10E+01**	**0.210**	**1.623**
Household with furry pets	0.215	0.362	0.039	0.593	5,54E+02	-0.499	0.929
Start of daycare in months	0.002	0.019	0.008	0.114	9,09E+02	-0.035	0.039

Supplementary Table S8 (cont'd)

Age of introduction solids (months)	-0.018	0.044	-0.029	-0.411	6,82E+02	-0.104	0.068
antibiotic use prior to sample collection	-1.055	0.676	-0.107	-1.561	1,20E+02	-2.388	0.278

MAZ index at age 31 weeks (N = 253)							
	Unstandardized Coefficients Beta	Std. Error	**Standardized Coefficients Beta**	**t**	**P-value**	**95,0% Confidence Interval for B**	
						Lower Bound	**Upper Bound**
(Constant)	-2.944	4.998		-0.589	5,56E+02	-12.792	6.904
Birthmode assisted vaginal	1.245	0.686	0.113	1.814	7,10E+01	-0.107	2.597
Birthmode caesarean section	0.254	0.354	0.045	0.719	4,73E+02	-0.443	0.952
1 older sibling	-0.361	0.353	-0.069	-1.024	3,07E+02	-1.056	0.334
≥ 2 older siblings	-0.273	0.581	-0.033	-0.470	6,39E+02	-1.418	0.871
Sex	-0.072	0.306	-0.015	-0.236	8,13E+02	-0.676	0.531
Birth Weight	-0.001	0.000	-0.097	-1.408	1,60E+02	-0.001	0.000
Gestational age	0.077	0.130	0.041	0.592	5,54E+02	-0.179	0.332
Breastfeeding duration**	**-0.018**	**0.007**	**-0.192**	**-2.724**	**7,00E+00**	**-0.031**	**-0.005**
Ethnicity	1.271	0.947	0.086	1.342	1,81E+02	-0.594	3.136

Supplementary Table S8 (cont'd)

Maternal age delivery*	**0.072**	**0.034**	**0.141**	**2.130**	**3,40E+01**	**0.005**	**0.139**
AD father	-0.551	0.384	-0.099	-1.434	1,53E+02	-1.309	0.206
AD mother	-0.384	0.421	-0.064	-0.912	3,63E+02	-1.214	0.446
Asthma father	-0.329	0.398	-0.054	-0.827	4,09E+02	-1.114	0.455
Asthma mother	-0.526	0.348	-0.101	-1.511	1,32E+02	-1.211	0.160
Smoking in pregnancy	-0.195	0.373	-0.034	-0.523	6,02E+02	-0.930	0.540
Smoking after pregnancy	0.131	0.373	0.023	0.351	7,26E+02	-0.605	0.866
AD**	**-0.977**	**0.332**	**-0.188**	**-2.945**	**4,00E+00**	**-1.630**	**-0.323**
Household with furry pets	0.287	0.322	0.055	0.890	3,74E+02	-0.348	0.921
Start of daycare in months	-0.030	0.017	-0.113	-1.798	7,40E+01	-0.063	0.003
Age of introduction solids (months)	-0.012	0.040	-0.021	-0.308	7,59E+02	-0.092	0.067
antibiotic use prior to sample collection	0.431	0.502	0.055	0.860	3,91E+02	-0.557	1.420

Supplementary Table S9 Multivariable linear regression analyses on demographic, dietary, lifestyle and medical factors associated with microbial diversity (Shannon)

	Shannon index at age 5 weeks (N = 235)						
	Unstandardized Coefficients Beta	Std. Error	Standardized Coefficients Beta	t	P-value	95,0% Confidence Interval for B	
						Lower Bound	Upper Bound
(Constant)	0.984	1.142		0.861	3,90E+02	-1.267	3.234
Birthmode assisted vaginal	0.075	0.153	0.032	0.487	6,27E+02	-0.227	0.376
Birthmode caesarean section	-0.025	0.081	-0.020	-0.314	7,54E+02	-0.184	0.134
1 older sibling	0.057	0.080	0.050	0.722	4,71E+02	-0.099	0.214
≥ 2 older siblings	0.006	0.134	0.003	0.047	9,62E+02	-0.257	0.270
Sex	0.045	0.070	0.042	0.636	5,25E+02	-0.093	0.183
Birth Weight	0.000	0.000	0.080	1.110	2,68E+02	0.000	0.000
Gestational age	0.023	0.030	0.056	0.765	4,45E+02	-0.036	0.082
Breastfeeding duration	-0.001	0.002	-0.065	-0.875	3,82E+02	-0.004	0.002
Ethnicity	0.017	0.220	0.005	0.078	9,38E+02	-0.417	0.451
Maternal age delivery	0.003	0.008	0.030	0.441	6,59E+02	-0.012	0.018
AD father	0.019	0.088	0.015	0.212	8,32E+02	-0.154	0.191
AD mother	-0.183	0.097	-0.140	-1.888	6,00E+01	-0.374	0.008
Asthma father	0.083	0.091	0.062	0.917	3,60E+02	-0.095	0.261
Asthma mother	-0.068	0.079	-0.060	-0.860	3,90E+02	-0.224	0.088

Supplementary Table S9 (cont'd)

Smoking in pregnancy	-0.029	0.087	-0.023	-0.331	7,41E+02	-0.199	0.142
Smoking after pregnancy	0.014	0.084	0.012	0.166	8,68E+02	-0.152	0.180
AD	-0.004	0.076	-0.004	-0.055	9,56E+02	-0.154	0.145
Household with furry pets	-0.146	0.075	-0.127	-1.953	5,20E+01	-0.293	0.001
Start of daycare in months	0.002	0.004	0.027	0.413	6,80E+02	-0.006	0.010
Age of introduction solids (months)	-0.010	0.009	-0.078	-1.084	2,80E+02	-0.028	0.008
antibiotic use prior to sample collection	-0.025	0.281	-0.006	-0.090	9,28E+02	-0.580	0.529

Supplementary Table S9 (cont'd)

	Shannon index at age 13 weeks (N = 231)						
	Unstandardized Coefficients Beta	Std. Error	Standardized Coefficients Beta	t	P-value	95,0% Confidence Interval for B	
						Lower Bound	Upper Bound
(Constant)	4.139	1.268		3.264	1,00E+00	1.639	6.639
Birthmode assisted vaginal	0.184	0.168	0.076	1.098	2,73E+02	-0.147	0.515
Birthmode caesarean section	0.031	0.088	0.025	0.358	7,21E+02	-0.142	0.205
1 older sibling	0.005	0.087	0.004	0.054	9,57E+02	-0.168	0.177
≥ 2 older siblings	0.157	0.154	0.081	1.019	3,09E+02	-0.147	0.461
Sex	0.018	0.077	0.016	0.236	8,14E+02	-0.133	0.170
Birth Weight	0.000	0.000	-0.119	-1.573	1,17E+02	0.000	0.000
Gestational age	-0.034	0.033	-0.079	-1.035	3,02E+02	-0.099	0.031
Breastfeeding duration	-0.001	0.002	-0.046	-0.587	5,58E+02	-0.004	0.002
Ethnicity	0.100	0.231	0.031	0.431	6,67E+02	-0.356	0.555
Maternal age delivery	-0.004	0.008	-0.036	-0.493	6,23E+02	-0.021	0.013
AD father	0.020	0.094	0.016	0.208	8,35E+02	-0.166	0.205
AD mother	0.124	0.107	0.092	1.154	2,50E+02	-0.088	0.336

Supplementary Table S9 (cont'd)

Asthma father	-0.012	0.103	-0.008	-0.113	9,10E+02	-0.214	0.191
Asthma mother	-0.006	0.087	-0.005	-0.064	9,49E+02	-0.176	0.165
Smoking in pregnancy	-0.053	0.091	-0.042	-0.583	5,61E+02	-0.234	0.127
Smoking after pregnancy	-0.035	0.092	-0.028	-0.380	7,04E+02	-0.217	0.147
AD	-0.074	0.084	-0.062	-0.888	3,76E+02	-0.239	0.091
Household with furry pets	-0.034	0.079	-0.029	-0.429	6,68E+02	-0.191	0.123
Start of daycare in months	-0.006	0.004	-0.101	-1.452	1,48E+02	-0.014	0.002
Age of introduction solids (months)	-0.006	0.010	-0.047	-0.605	5,46E+02	-0.026	0.014
antibiotic use prior to sample collection	0.001	0.012	0.006	0.094	9,25E+02	-0.023	0.026

Supplementary Table S9 (cont'd)

Shannon index at age 21 weeks (N = 228)							
	Unstandardized Coefficients Beta	Std. Error	Standardized Coefficients Beta	t	P-value	95,0% Confidence Interval for B	
						Lower Bound	Upper Bound
(Constant)	3.346	1.192		2.807	5,00E+00	0.996	5.696
Birthmode assisted vaginal	-0.156	0.154	-0.067	-1.009	3,14E+02	-0.460	0.148
Birthmode caesarean section	-0.074	0.082	-0.061	-0.902	3,68E+02	-0.236	0.088
1 older sibling	0.042	0.081	0.038	0.526	6,00E+02	-0.117	0.201
≥ 2 older siblings	0.005	0.133	0.003	0.037	9,71E+02	-0.257	0.267
Sex	0.083	0.070	0.079	1.183	2,38E+02	-0.055	0.222
Birth Weight	0.000	0.000	-0.056	-0.757	4,50E+02	0.000	0.000
Gestational age	0.004	0.031	0.009	0.122	9,03E+02	-0.058	0.065
Breastfeeding duration**	**-0.004**	**0.001**	**-0.199**	**-2.685**	**8,00E+00**	**-0.007**	**-0.001**
Ethnicity	-0.207	0.235	-0.058	-0.882	3,79E+02	-0.669	0.255
Maternal age delivery	0.000	0.008	-0.002	-0.024	9,81E+02	-0.015	0.015
AD father	-0.111	0.088	-0.094	-1.265	2,07E+02	-0.284	0.062
AD mother	-0.069	0.097	-0.054	-0.712	4,78E+02	-0.260	0.122
Asthma father	-0.012	0.089	-0.010	-0.137	8,91E+02	-0.188	0.164
Asthma mother	0.059	0.080	0.053	0.729	4,67E+02	-0.100	0.217

Supplementary Table S9 (cont'd)

Smoking in pregnancy	-0.058	0.087	-0.046	-0.661	5,09E+02	-0.230	0.115
Smoking after pregnancy	-0.060	0.088	-0.050	-0.686	4,93E+02	-0.233	0.113
AD	-0.100	0.075	-0.091	-1.334	1,84E+02	-0.248	0.048
Household with furry pets	0.036	0.076	0.032	0.477	6,34E+02	-0.113	0.186
Start of daycare in months	0.003	0.004	0.060	0.882	3,79E+02	-0.004	0.011
Age of introduction solids (months)**	**-0.028**	**0.009**	**-0.224**	**-3.086**	**2,00E+00**	**-0.046**	**-0.010**
antibiotic use prior to sample collection	0.109	0.138	0.055	0.793	4,29E+02	-0.162	0.380

Shannon index at age 31 weeks (N = 253)

	Unstandardized Coefficients Beta	**Std. Error**	**Standardized Coefficients Beta**	**t**	**P-value**	**95,0% Confidence Interval for B**	
						Lower Bound	**Upper Bound**
(Constant)	2.551	1.064		2.399	1,70E+01	0.456	4.647
Birthmode assisted vaginal	-0.012	0.142	-0.005	-0.082	9,34E+02	-0.292	0.269
Birthmode cesarean section	0.009	0.076	0.008	0.122	9,03E+02	-0.140	0.158

Supplementary Table S9 (cont'd)

1 older sibling	0.104	0.075	0.091	1.386	1,67E+02	-0.044	0.251
≥ 2 older siblings*	**0.246**	**0.124**	**0.136**	**1.978**	**4,90E+01**	**0.001**	**0.491**
Sex	-0.029	0.065	-0.028	-0.448	6,55E+02	-0.157	0.099
Birth Weight	0.000	0.000	0.017	0.248	8,04E+02	0.000	0.000
Gestational age	0.021	0.028	0.051	0.757	4,50E+02	-0.033	0.075
Breastfeeding duration**	**-0.005**	**0.001**	**-0.238**	**-3.449**	**1,00E+00**	**-0.008**	**-0.002**
Ethnicity	-0.202	0.203	-0.063	-0.996	3,20E+02	-0.602	0.198
Maternal age delivery	-0.007	0.007	-0.060	-0.915	3,61E+02	-0.021	0.008
AD father	-0.073	0.082	-0.060	-0.885	3,77E+02	-0.235	0.089
AD mother	0.072	0.089	0.056	0.810	4,19E+02	-0.104	0.249
Asthma father	0.092	0.084	0.071	1.106	2,70E+02	-0.072	0.257
Asthma mother	0.068	0.074	0.061	0.918	3,60E+02	-0.078	0.215
Smoking in pregnancy	0.031	0.079	0.025	0.391	6,96E+02	-0.125	0.187
Smoking after pregnancy	0.002	0.079	0.002	0.025	9,80E+02	-0.155	0.158
AD	-0.178	0.070	-0.158	-2.525	1,20E+01	-0.317	-0.039
Household with furry pets	-0.010	0.069	-0.009	-0.148	8,82E+02	-0.145	0.125
Start of daycare in months	0.000	0.004	0.002	0.028	9,78E+02	-0.007	0.007
Age of introduction solids (months)**	**-0.029**	**0.008**	**-0.232**	**-3.459**	**1,00E+00**	**-0.046**	**-0.013**

Supplementary Table S9 (cont'd)

antibiotic use prior to sample collection	-0.001	0.106	-0.001	-0.010	9,92E+02	-0.209	0.207

Supplementary Table S10 Joint modelling on the association between longitudinal microbial diversity (Shannon index) and the time to development of atopic dermatitis (AD)

Joint modelling on MAZ scores and atopic dermatitis (N = 312)				
Effect	**Value**	**Std. Err.**	**z-value**	**p-value**
Treatment (reference = placebo)	-0.06407407	0.211695794	-0.30267049	7.62E-01
Gender (reference = boy)	0.034598927	0.211291403	0.16374981	8.70E-01
AD father (reference = no)	0.906944883	0.290099079	3.126328035	1.77E-03
Breastfeeding duration (weeks)	0.007557587	0.00364736	2.072070356	3.83E-02
1 older sibling (reference = no sibs)	0.154546605	0.238565208	0.647817031	5.17E-01
≥ 2 older siblings	-0.034512157	0.375489343	-0.091912482	9.27E-01
Furry pets in household (reference = no)	-0.078684706	0.23136654	-0.340086798	7.34E-01
Caesarean section birth mode (reference = natural vaginal delivery)	-0.12821753	0.474555339	-0.27018457	7.87E-01
Assisted vaginal delivery	-0.343478873	0.267486101	-1.284099886	1.99E-01
Birth weight (in grams)	0.510627325	0.264140586	1.933164954	5.32E-02
AD mother (reference = no)	0.766246518	0.305654581	2.506903427	1.22E-02
ASSOCIATION (MAZ scores)	**0.127799617**	**0.029913897**	**4.272249057**	**1.94E-05**
log(xi.1)	-4.595666187	1.097884853	-4.185927309	2.84E-05
log(xi.2)	-4.671636806	1.098344847	-4.253342489	2.11E-05
log(xi.3)	-2.512362383	1.302212452	-1.929302995	5.37E-02
log(xi.4)	-7.065394817	8240.266544	-0.000857423	9.99E-01

Supplementary Table S11 Joint modelling on the association between longitudinal microbial maturity (MAZ score) and the time to development of atopic dermatitis (AD)

Joint modelling on MAZ scores and atopic dermatitis (N = 312)				
Effect	**Value**	**Std. Err.**	**z-value**	**p-value**
Treatment (reference = placebo)	-0.06407407	0.211695794	-0.30267049	7.62E-01
Gender (reference = boy)	0.034598927	0.211291403	0.16374981	8.70E-01
AD father (reference = no)	0.906944883	0.290099079	3.126328035	1.77E-03
Breastfeeding duration (weeks)	0.007557587	0.00364736	2.072070356	3.83E-02
1 older sibling (reference = no sibs)	0.154546605	0.238565208	0.647817031	5.17E-01
≥ 2 older siblings	-0.034512157	0.375489343	-0.091912482	9.27E-01
Furry pets in household (reference = no)	-0.078684706	0.23136654	-0.340086798	7.34E-01
Caesarean section birth mode (reference = natural vaginal delivery)	-0.12821753	0.474555339	-0.27018457	7.87E-01
Assisted vaginal delivery	-0.343478873	0.267486101	-1.284099886	1.99E-01
Birth weight (in grams)	0.510627325	0.264140586	1.933164954	5.32E-02
AD mother (reference = no)	0.766246518	0.305654581	2.506903427	1.22E-02
ASSOCIATION (MAZ scores)	**0.127799617**	**0.029913897**	**4.272249057**	**1.94E-05**
log(xi.1)	-4.595666187	1.097884853	-4.185927309	2.84E-05
log(xi.2)	-4.671636806	1.098344847	-4.253342489	2.11E-05
log(xi.3)	-2.512362383	1.302212452	-1.929302995	5.37E-02
log(xi.4)	-7.065394817	8240.266544	-0.000857423	9.99E-01

Supplementary Table S12 METALONDA analyses on microbial genera in association to atopic dermatitis

MetaLonDa Analysis Output in association to Atopic Dermatitis

Bacterial Genera	start (days)	end (days)	dominant	FDR adjusted p-value
Corynebacterium	126.110	151.240	HC	9.68E-03
Atopobium	25.590	79.440	HC	7.65E-03
Prevotella	104.570	133.290	HC	0.00E+00
Lachnobacterium	25.590	61.490	HC	1.18E-02
Lachnobacterium	108.160	129.700	HC	6.49E-03
Lachnobacterium	208.680	269.710	HC	5.00E-03
Faecalibacterium	47.130	201.500	HC	5.08E-03

Supplementary Table S13 Logistic regression model of microbial diversity (Shannon index) in association to allergic sensitization at school-age

Logistic regression model of microbial diversity (Shannon index) in association to allergic sensitization at school-age				
Coefficients	**Estimate**	**Std. Error**	**z value**	**P-value**
(Intercept)	5.627791598	3.674612075	1.531533529	1.26E-01
Shannon at 5 weeks	-0.410285084	0.454296315	-0.903122193	3.66E-01
Shannon at 13 weeks	-0.208920785	0.527504064	-0.396055309	6.92E-01
Shannon at 21 weeks	0.279310503	0.601239738	0.464557623	6.42E-01
Shannon at 31 weeks	-1.620866258	0.604467106	-2.681479675	7.33E-03
Breastfeeding duration (weeks)	-3.030233351	0.873149718	-3.470462497	5.20E-04
Solid food introduction (week)	6.294484319	3.381347875	1.861531126	6.27E-02
Caesarean section	-0.238619825	1.081714761	-0.220594036	8.25E-01
Assisted vaginal delivery	-0.94531093	0.626875226	-1.507973023	1.32E-01
Birth Weight (g)	-0.000366139	0.00065891	-0.555673949	5.78E-01
Sex Female (reference = male)	-0.527185935	0.500811149	-1.052664135	2.92E-01
Treatment Control (reference = yes)	0.742100392	0.503761986	1.473117091	1.41E-01
No AD Father (reference=Yes)	-1.635728509	0.645015214	-2.535953375	1.12E-02
No AD Mother (reference=Yes)	-0.655147341	0.663647229	-0.987192159	3.24E-01
1 older sibling (reference = no sibs)	0.45591163	0.568647631	0.801747172	4.23E-01
≥ 2 older siblings	-6.44E-05	0.999108354	-6.44E-05	1.00E+00
Furry pets in household (reference = no)	-0.72774705	0.548034524	-1.327921908	1.84E-01

Supplementary Table S14 Logistic regression model of microbial maturity (MAZ score) in association to allergic sensitization at school-age

Logistic regression model of microbial maturity (MAZ-score) in association to allergic sensitization at school-age				
Coefficients	**Estimate**	**Std. Error**	**z value**	**P-value**
(Intercept)	-0.087464637	2.640818493	-0.033120276	9.74E-01
MAZ at 5 weeks	0.382210056	0.136202238	2.806195127	5.01E-03
MAZ at 13 weeks	-0.164301672	0.140623631	-1.16837882	2.43E-01
MAZ at 21 weeks	0.050549523	0.182113301	0.277571833	7.81E-01
MAZ at 31 weeks	-0.016145267	0.133545263	-0.120897342	9.04E-01
Breastfeeding duration (weeks)	-3.026775739	0.884920175	-3.420394091	6.25E-04
Solid food introduction (week)	8.847895913	3.301639509	2.679849175	7.37E-03
Birth Weight (g)	-0.00039128	0.000649369	-0.602553926	5.47E-01
Treatment Control (reference = yes)	0.957059287	0.534603349	1.790223143	7.34E-02
Sex Female (reference = male)	-0.402875364	0.508750997	-0.791891054	4.28E-01
Caesarean section	-0.236701605	1.019008403	-0.232286215	8.16E-01
Assisted vaginal delivery	-0.842345125	0.656241797	-1.283589569	1.99E-01
No AD Father (reference=Yes)	-1.43835435	0.64026914	-2.246483953	2.47E-02
No AD Mother (reference=Yes)	-0.52709629	0.640499938	-0.822945108	4.11E-01
1 older sibling (reference = no sibs)	0.119864913	0.552340617	0.217012672	8.28E-01

Supplementary Table S14 (cont'd)

≥ 2 older siblings	-0.745760306	1.048383644	-0.711342942	4.77E-01
Furry pets in household (reference = no)	-0.840188796	0.565073526	-1.486866323	1.37E-01

Supplementary Table S15 Logistic regression model of microbial diversity (Shannon index) in association to asthma at school-age

Logistic regression model of microbial diversity (Shannon index) in association to asthma at school-age				
Coefficients	**Estimate**	**Std. Error**	**z value**	**P-value**
(Intercept)	-9.519296394	5.319877537	-1.789382618	7.36E-02
Shannon at 5 weeks	-0.431558541	0.692669049	-0.623037137	5.33E-01
Shannon at 13 weeks	0.065089728	0.751687765	0.086591443	9.31E-01
Shannon at 21 weeks	-0.02811964	0.661640302	-0.04249989	9.66E-01
Shannon at 31 weeks	0.164692463	0.801999013	0.205352451	8.37E-01
Breastfeeding duration (weeks)	-0.254586729	0.764690582	-0.332927768	7.39E-01
Solid food introduction (week)	6.194863693	5.019759796	1.234095643	2.17E-01
Caesarean section	0.344889489	1.456595944	0.236777735	8.13E-01
Birth Weight (g)	0.001270255	0.00090951	1.396637011	1.63E-01
Assisted vaginal delivery	-0.825554535	0.948850266	-0.870057758	3.84E-01
Treatment Control (reference = yes)	-0.166071147	0.657007741	-0.252768936	8.00E-01
Sex Female (reference = male)	-2.051878298	0.859550668	-2.387152234	1.70E-02
No AD Father (reference=Yes)	0.085116272	0.823378904	0.103374365	9.18E-01
No AD Mother (reference=Yes)	-0.313860614	1.003653848	-0.31271799	7.54E-01
1 older sibling (reference = no sibs)	0.643148333	0.754705643	0.852184343	3.94E-01
≥ 2 older siblings	-15.68707851	1926.712049	-0.00814189	9.94E-01
Furry pets in household (reference = no)	1.483880523	0.934270057	1.588277941	1.12E-01

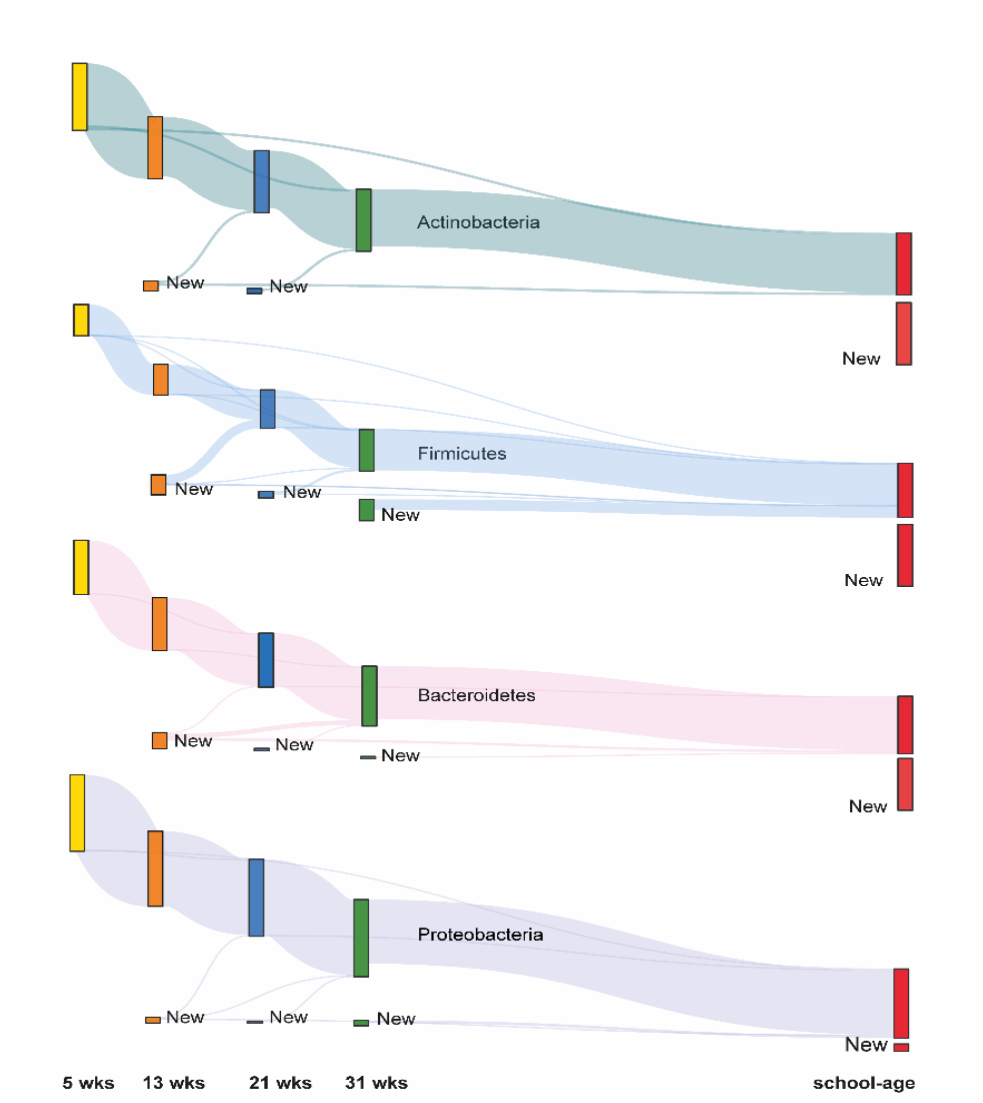
Supplementary Table S16 Logistic regression model of microbial maturity (MAZ score) in association to asthma at school-age

Logistic regression model of microbial maturity (MAZ-score) in association to asthma at school-age				
Coefficients	**Estimate**	**Std. Error**	**z value**	**P-value**
(Intercept)	-12.40821269	5.470392851	-2.268248923	2.33E-02
MAZ at 5 weeks	0.363442479	0.136558324	2.661445086	7.78E-03
MAZ at 13 weeks	-0.415956399	0.187305674	-2.220735705	2.64E-02
MAZ at 21 weeks	0.254280302	0.228921442	1.110775385	2.67E-01
MAZ at 31 weeks	-0.104036691	0.157184716	-0.661875362	5.08E-01
Breastfeeding duration (weeks)	-0.486254724	0.839838341	-0.578986099	5.63E-01
Solid food introduction (week)	10.08291775	5.955334706	1.69309002	9.04E-02
Caesarean section	1.480313854	1.561995355	0.94770695	3.43E-01
Assisted vaginal delivery	-1.1528046	1.0910663	-1.056585287	2.91E-01
Birth Weight (g)	0.001070665	0.000970668	1.103018401	2.70E-01
Treatment Control (reference = yes)	0.126012384	0.750777261	0.167842569	8.67E-01
Sex Female (reference = male)	-3.129532598	1.16397764	-2.688653536	7.17E-03
No AD Father (reference=Yes)	0.731989105	0.882583537	0.829370903	4.07E-01
No AD Mother (reference=Yes)	-1.282273155	1.392172865	-0.921058862	3.57E-01
1 older sibling (reference = no sibs)	0.961383693	0.768278099	1.25134856	2.11E-01
≥ 2 older siblings	-14.55804809	1838.156717	-0.007919917	9.94E-01
Furry pets in household (reference = no)	1.829645625	1.082872207	1.68962285	9.11E-02

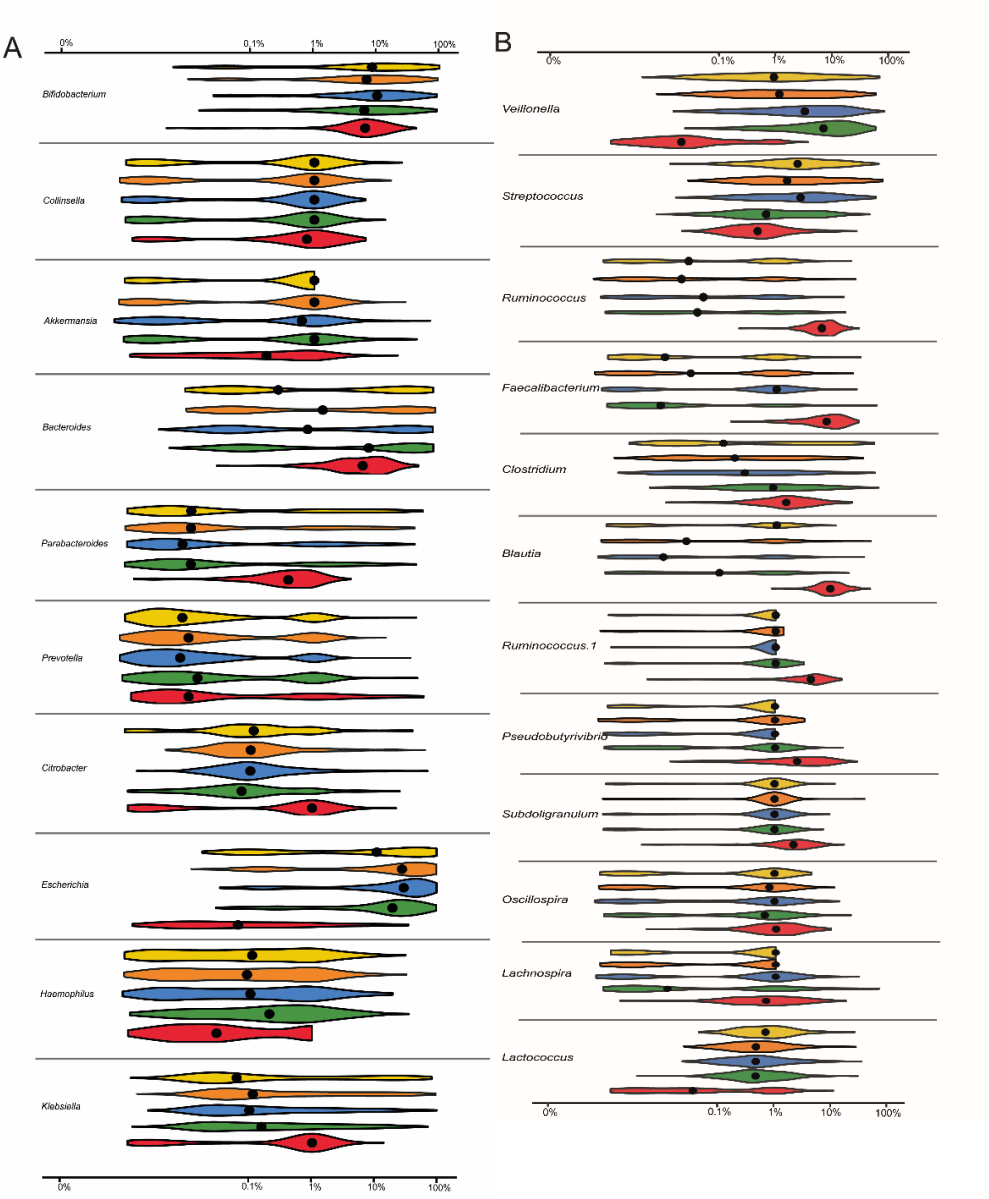
Supplementary Table S17 METALONDA analyses on microbial genera in association to asthma

MetaLonDa Analysis Output in association to Asthma				
Bacterial Genera	**start (days)**	**end (days)**	**dominant**	**FDR adjusted p-value**
Actinomyces	183.040	192.000	HC	1.809E-02
Enterococcus	167.360	194.240	HC	3.307E-02
Streptococcus	44.160	129.280	Asthma	1.741E-02
Streptococcus	183.040	236.800	HC	3.156E-03
Sarcina	28.480	57.600	HC	8.062E-03
Sarcina	84.480	86.720	HC	4.844E-02
Sarcina	88.960	93.440	HC	4.521E-02
Lachnobacterium	28.480	100.160	HC	1.365E-02
Lachnobacterium	138.240	176.320	HC	4.618E-03
Lachnobacterium	221.120	236.800	HC	2.783E-02
Lachnospira	48.640	80.000	HC	1.717E-02
Lachnospira	91.200	200.960	HC	2.878E-03
Lachnospira	218.880	236.800	HC	1.040E-03
Dialister	28.480	122.560	HC	6.209E-04
Dialister	131.520	144.960	Asthma	1.378E-02
Dialister	151.680	176.320	HC	1.200E-02
Dialister	203.200	207.680	HC	4.823E-02
Dialister	216.640	236.800	HC	1.935E-02
Klebsiella	28.480	102.400	HC	6.922E-03
Leclercia	28.480	187.520	HC	2.746E-03
Haemophilus	115.840	236.800	HC	1.038E-02

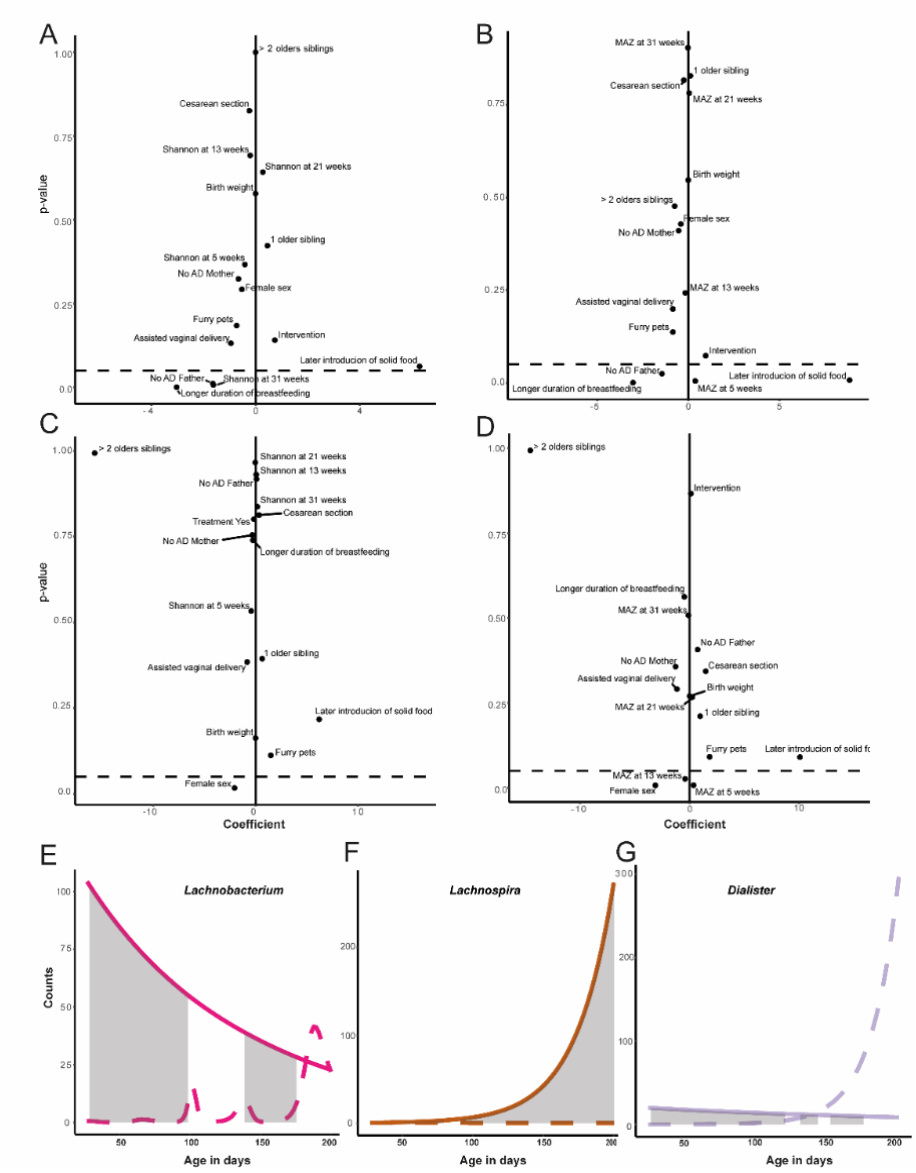
Supplementary Figures



Supplementary Figure 1 Tracking the flow of OTUs within the four main phyla throughout infancy and childhood (N= 1,453 stool samples from 440 children). OTUs that were shared by at least 10 percent of the population during one or more time points were tracked using Sankey plots in the four major phyla. The rectangle height indicates the relative number of OTUs and the rectangle color reflects the children's age. The lines represent the transfer of OTUs between time points. At school-age many new OTUs within the Actinobacteria, Firmicutes and Bacteroidetes were gained, whereas a major loss of OTUs in Proteobacteria was observed.



Supplementary Figure 2 Relative abundance of the main bacterial genera at different time-points during infancy and at school-age (N= 1,453 stool samples from 440 children). Comparison were made among the 22 most abundant genera and stratified according to age. The black dots represent the median values and the violins are colored according to age. A, relative abundance of the main genera within the phyla of Proteobacteria, Actinobacteria, Bacteroidetes and Verrucomicrobia. B, relative abundances of the main genera within the Firmicutes phylum. To test for significant changes in relative abundances with age, the Friedman test was used followed by the Dunn's test for post-hoc analyses. FDR-adjusted p-values are presented in Supplementary Table 4.



Supplementary Figure 3 (N= 961 stool samples from 312 children). A-B, Volcano plots depicting the regression coefficients from the logistic regression analyses on the association between microbial diversity (Shannon index) and maturity (microbial age z-scores, MAZ) at the ages of 5, 13, 21 and 31 weeks and the development of allergic sensitization at school-age. C-D, Volcano plots depicting the regression coefficients from the logistic regression analyses on the association between microbial diversity (Shannon index) and maturity (microbial age z-scores, MAZ) at the ages of 5, 13, 21 and 31 weeks and the development of asthma at school-age. The dashed lines depict the threshold for statistical significance at p<0.05. E-G, Time intervals of differential abundance in Lachnobacteriumn (E), Lachnospira (F) and Dialister (G) between infants that did (dashed line) or did not develop asthma (solid line) as identified from MetaLonda analyses. Significantly different time-intervals (fdr-adjusted p<0.05) are depicted by gray shading.

Chapter 5

pibs
cases
travel
controls
microbiota
ibs
study
symptoms
diversity
travellers
microbial
community
significantly
among
also
onset
prior
subjects
identified
structure
faecal studies
control
abundance
episode
samples
alterations
differences
gastrointestinal
experienced
sequencing
patients
differentially
criteria
previous
significant
collected
longitudinal
immediately
posttravel
according
time
however

Gut microbiota perturbations precede the onset of Post-Infectious IBS in intercontinental travellers

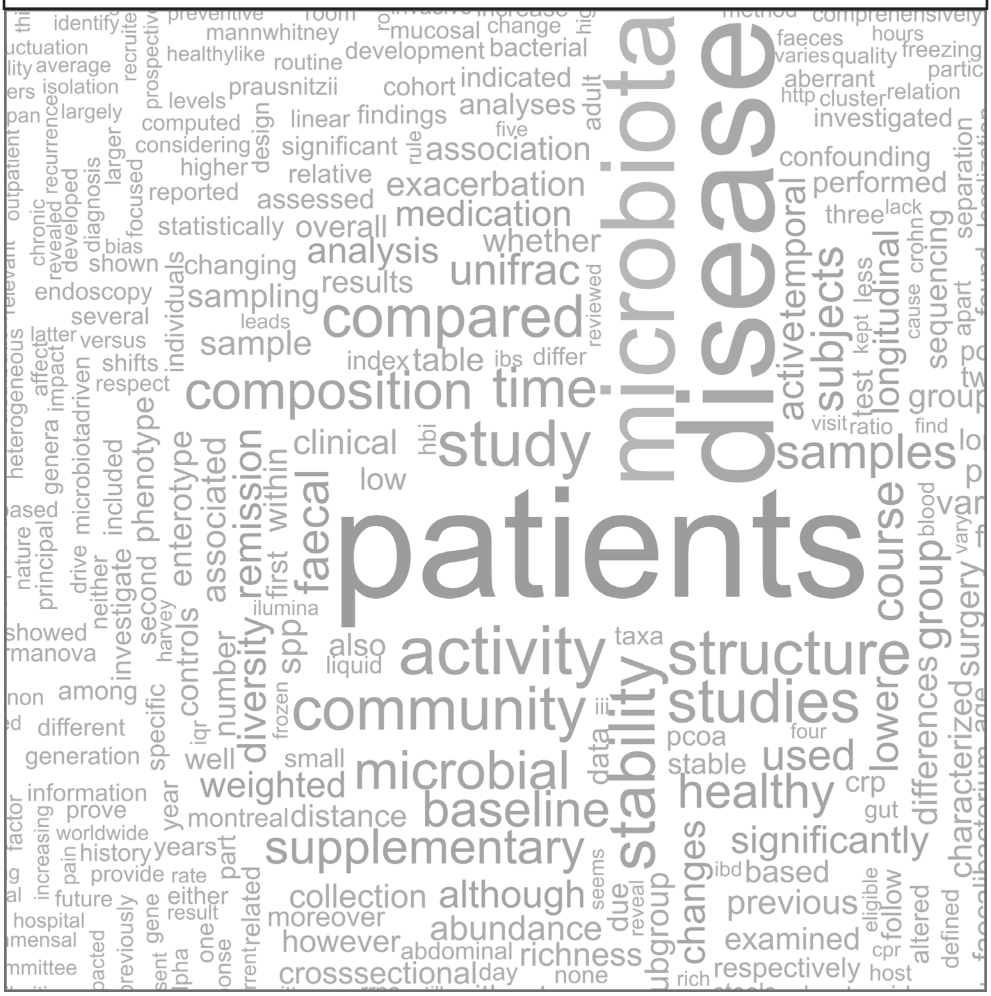
Gianluca Galazzo[*], Jiyang Chan[*], Markia Ward, Maris Arcilla, Perry van Genderen[§] and John Penders[§] on behalf of the COMBAT-consortium

** Authors contributed equally*

§ Co-authors contributed equally

Manuscript in preparation

Chapter 6


patients
disease
microbiota
stability
activity
structure
community
studies
composition
time
study
compared
samples
microbial
baseline
supplementary
healthy
unifrac
analysis
faecal
remission
course
subjects
longitudinal
healthy
used
significantly
although
collection
abundance
richness
previous
examined
respectively
crosssectional
however
weighted
distance
clinical
sample
sampling
medication
exacerbation
association
changes
differences
group
diversity

Faecal microbiota dynamics and its relation with disease course in Crohn's disease

Gianluca Galazzo[*], Danyta I. Tedjo[*], Dion S.J. Wintjens, Paul H.M. Savelkoul, Ad A.M. Masclee, Alexander G. L. Bodelier, Marie J. Pierik, John Penders[§], Daisy M.A.E. Jonkers[§]

** Shared first authorship*

§ Shared last authorship

J. Crohns Colitis 2019 Sep ;13(10):1273-1282

Abstract

Microbial shifts have been associated with disease activity in Crohn's disease (CD), but findings on specific taxa are inconsistent. This may be due to differences in applied methods and cross-sectional study designs. We prospectively examined the faecal microbiota in adult CD patients with changing or stable disease course over time.

Faeces was collected at two time-points from 15 healthy individuals (HC), 35 CD patients that maintained remission (RR) and 22 during remission and subsequent exacerbation (RA). The microbial composition was assessed by 16S rRNA (V4) gene sequencing.

Compared to HC, CD patients had a lower microbial richness (p=0.0002) and diversity (p=0.005). Moreover, the microbial community structure of a subset of patients clustered apart from HC, characterized by low microbial diversity and *Faecalibacterium* abundance. Patients within this cluster did not differ with respect to long-term disease course compared to patients with a "healthy-like" microbiota.

Over time, microbial richness and diversity did not change in RR versus RA patients. Although the microbial community structure of both RR and RA patients was less stable over time compared to HC, no differences were observed between the patient groups (p=0.17), nor was the stability impacted by Montreal classification, medication use or surgery.

This study shows that altered microbiota composition and stability in CD was neither associated with disease activity nor long-term disease course, questioning its involvement in the development of an exacerbation. The aberrant microbiota composition in a subset of CD patients, warrants further exploration of a more microbiota-driven etiology in this group.

Introduction

Crohn's disease (CD) is a chronic gastrointestinal inflammatory disease of which the incidence is increasing worldwide [1]. It is a relapsing disease characterized by periods of active inflammation with symptoms as abdominal pain and (bloody) diarrhoea, alternated by periods of remission. The disease course varies between patients and has a poor predictability [2], hindering clinical decision making. CD has a significant impact on the patient's quality of life and health-related costs, especially during active disease [3, 4]. Further insight in factors contributing to disease activity, may provide leads for preventive strategies and improve disease outcome.

Although the exact cause is unclear, the generally accepted hypothesis is that CD results from an aberrant immune response against commensal bacteria in genetically susceptible hosts. Previous studies reported microbiota perturbations, characterized by a decreased diversity and changes in the abundance of specific taxa (e.g., reduction of *Faecalibacterium prausnitzii*, increase of Enterobacteriaceae) in CD when compared to healthy individuals [5-10]. Moreover, several studies have reported microbial shifts in relation to disease activity. When compared to inactive patients, the microbiota of CD patients during an exacerbation is characterized by increased members of Enterobacteriaceae [11, 12] and Bacteroides spp. [11, 13], and a reduction of *F. prausnitzii* [14-16] and *Clostridium coccoides* group [14, 17], although these associations vary between studies. These inconsistencies may in part be due to differences in assessing disease activity, applied molecular methods, and study populations, but also to potential confounding factors such as medication use. Many studies are based on a cross-sectional design comparing active with inactive patients. Considering the inter-individual variation in microbiota composition and heterogeneous nature of CD, longitudinal studies are particularly relevant.

To our knowledge, nine studies have investigated the microbiota in adult CD patients in relation to changing disease course over time [8, 13, 17-23]. Four focused on the predictive value of the microbiota on treatment response [13, 20] or post-surgery recurrence [18, 19]. The other five focused on remission patients subsequently developing an exacerbation [8, 17, 21-23]. With the exception of two [8, 23], these studies only included few subjects and only three studies comprehensively assessed the microbiota with next-generation sequencing techniques [8, 22, 23]. Although other techniques used in previous studies result in valuable information, they do not provide the same resolution as next-generation sequencing. Moreover, to what extend microbiota composition and stability is related to long-term disease course is largely unknown.

Within this study we therefore aimed to i) compare the faecal microbiota stability of CD patients and healthy individuals, ii) compare the stability of the faecal microbiota of CD patients with either changing or stable disease activity over time, and iii) explore the association between microbiota composition and stability in association to long-term disease course, by means of next-generation sequencing.

Materials and Methods

Study population

A total of 57 CD patients and 15 healthy subjects were included in this study[24]. The CD patients participated in a prospective follow-up study [25] of the deeply phenotyped IBDSL cohort. Clinical data, blood and faeces were collected at each outpatient visit and during an exacerbation during follow-up. As the current standard endoscopy is too invasive for disease monitoring over time and clinical indices do not correlate well with mucosal inflammation [26], disease activity was defined by the combination of faecal calprotectin (FC), serum CRP and the Harvey Bradshaw Index (HBI): *i.e*. FC >250 µg/g or FC >100 µg/g with at least a fivefold increase from baseline. Remission was defined by FC <100 µg/g and CRP <5 mg/l or FC <100 µg/g, CRP <10 mg/l and HBI ≤4. Patients being in remission at baseline were eligible for further analyses. Healthy subjects (HC), all without any GI disease, GI symptoms or comorbidities, were recruited among the controls that participated in the Maastricht IBS cohort as a reference group [27].

Faecal samples were collected from all CD and HC subjects at two time-points. The CD group comprised 22 patients with baseline sampling at time of remission and subsequent sampling during an exacerbation (*i.e.,* RA group), and 35 patients with two subsequent samples while maintaining remission (*i.e*., RR group, without any flares in between subsequent samples). Complete defecations were collected at home, kept at room temperature, and brought to the hospital within 12 hours after defecation. Part of the faecal sample of CD patients was sent to the laboratory of Clinical Chemistry for routine analysis of FC. The remaining part was aliquoted and frozen at -80°C for microbiota analysis. We have previously shown that this sample collection procedure does not significantly alter the microbiota composition when compared to immediate freezing of samples upon defecation at -80°C [28]. Blood was collected for routine analysis of CRP.

The standardized computer registration of the IBDSL and Maastricht IBS cohort (for HC) were used to retrieve demographics, data on disease phenotype by the Montreal classification, surgery (including (hemi)colectomy and ileocecal resection), medication use and clinical activity scores (HBI).

All study subjects gave written informed consent prior to participation. Both studies have been approved by the Medical Ethics Committee of Maastricht University Medical Centre+ and have been registered in the US National Library of Medicine (http://www.clinicaltrials.gov, NCT02130349 and NCT00775060, respectively).

Microbiota analysis and statistics

The faecal microbiota composition was assessed by Ilumina Miseq sequencing of the V4-region of the 16S rRNA gene. A detailed description on metagenomic DNA isolation, sequencing and quality control is provided in the supplemental information.

Statistical analysis were performed in R Studio 1.0.143 (R 3.4.1) using *vegan, Rhea, stats,igraph,ggraph, GUniFrac and DirichletMultinomial* packages. Alpha diversity estimates (observed species, Chao1 and Shannon index) were computed using Rhea standard script and settings [29]. Those indices were computed, per individual and the average differences were subsequently compared between the three study groups (RR, RA and HC). Significance was tested using the Kruskal-Wallis test and Mann-Whitney U test for *post-hoc* analysis.

Figure 1 Graphic schematization of the study design and demographics of the study population.
Age legend: A1 (≤16 year); A2 (17-40 year); A3 (>40 year). Phenotype legend: B1 (non-structuring/non-penetrating); B2 (structuring); B3 (penetrating). Disease localization legend: L1 (ileal); L2 (colonic); L3 (ileocolonic)

Bray-Curtis and (un)weighted UniFrac dissimilarities within subjects were used to investigate both the changes in the microbiota community structure between subjects at baseline and within subjects over time. Enterotype analysis was performed at baseline using the Dirichlet multinomial mixture method as described previously [30]. Principal Coordinate Analysis (PCoA) was used as an unconstrained ordination technique. To investigate whether the microbiota was more stable in healthy subjects as compared to CD patients, the Kruskal-Wallis and Mann-Whitney U tests were used to check for

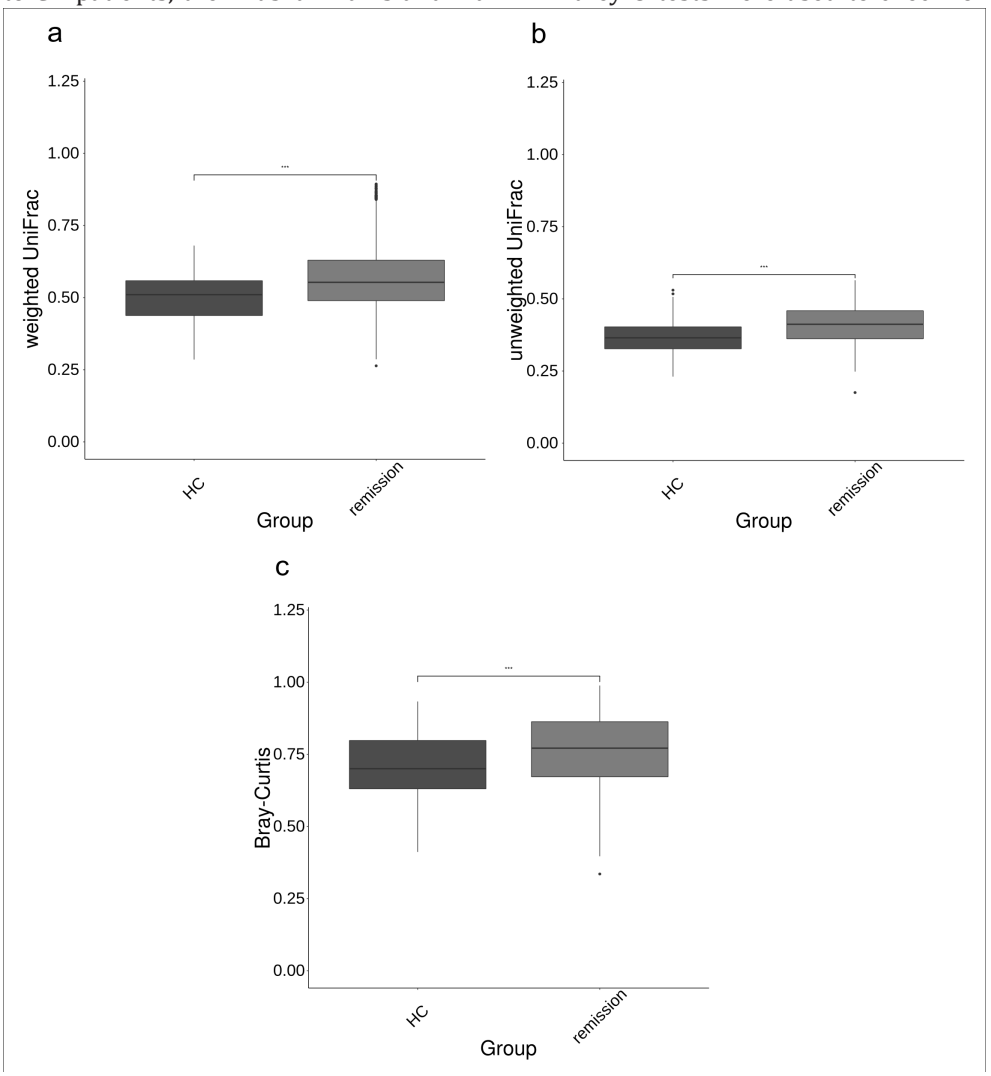


Figure 2 Within-group dissimilarity in the microbial community structure based upon **(a)** weighted UniFrac, **(b)** unweighted UniFrac and **(c)** Bray-Curtis for healthy controls (HC) and Crohn's disease patients (remission) at baseline (T1). All three beta-diversity indices indicate that the microbial community structure is significantly more heterogenous between CD patients than between healthy controls.
Significance was tested using Wilcoxson Signed-Ranks Test; *** indicates $p<0.001$.

statistically significant differences between the groups with respect to the variation in relative abundance of bacterial genera as well as to the within-subject beta-diversity (distance between first and second sample). A linear model was used to test if the time-span between collection of the first and second sample was affecting the within-subject weighted UniFrac distance, as well as to investigate the correlation between the within-subject weighted UniFrac and the variation overtime of the Calprotectin and CPR levels.

To examine the variation in microbial community structure, we first performed a PERMANOVA using Montreal classification factors (age at diagnosis, disease localization and disease behaviour)[31], medications used, number of liquid stools per day, surgery, and smoking habits as explanatory variables for the microbial community structure. We then performed a distance-based redundancy analysis (dbRDA) [32] to test if CD patients clustered according to the disease activity or the medications used, with and without removing the effect of age, gender and *Bacteroides/Prevotella* ratio.

Finally, we examined whether the history of disease activity and/or disease activity in the years following sample collection were associated with microbial community structure at baseline. To this purpose, the disease course of each individual patient was reviewed from the year before until 5 years after inclusion. Each yearly quarter was assessed for disease activity, defined by either i) active disease on endoscopy or imaging, ii) hospitalization due to an exacerbation, iii) surgery for active IBD or iv) treatment adjustment for increased symptoms. The number of active quarters before inclusion was used as marker for 'disease course before sample collection' and the number of active quarters after inclusion was used as marker for 'disease course after baseline sample collection'.

Results

Study population

A total of 144 faecal samples of 57 CD patients (35 RR, 22 RA) and 15 HC were available for analysis (1 and Supplementary Table 1). The median time between baseline and follow-up samples was 14 (IQR 11-21), 20 (8-36) and 13 (12-16) weeks for RR, RA and HC, respectively. Neither substantial differences were found in overall medication use between the different CD patient populations, nor within each patient group over time (Supplementary Table 2).

Baseline microbial richness, diversity, and community structure

At baseline, CD patients had a significantly lower faecal microbial richness and diversity when compared to HC as indicated by the number of observed species (median (IQR): 170 (97-233) and 209 (135-251), respectively; p=0.0002), Chao1 index (173 (107-236) and 209 (135-251), respectively; p=0.0006) and Shannon index (3.5 (1.8-4.1) and 3.8 (2.7-4.4), respectively; p=0.005) (Supplementary Figure 1).

Differences in the faecal microbial community structure between samples at baseline were assessed using the Bray-Curtis and (un)weighted UniFrac. Microbial community structure was more heterogeneous among CD as indicated by the significantly higher distances in CD when compared to HC for weighted UniFrac as well as for the other beta-diversity indices (Figure 2).

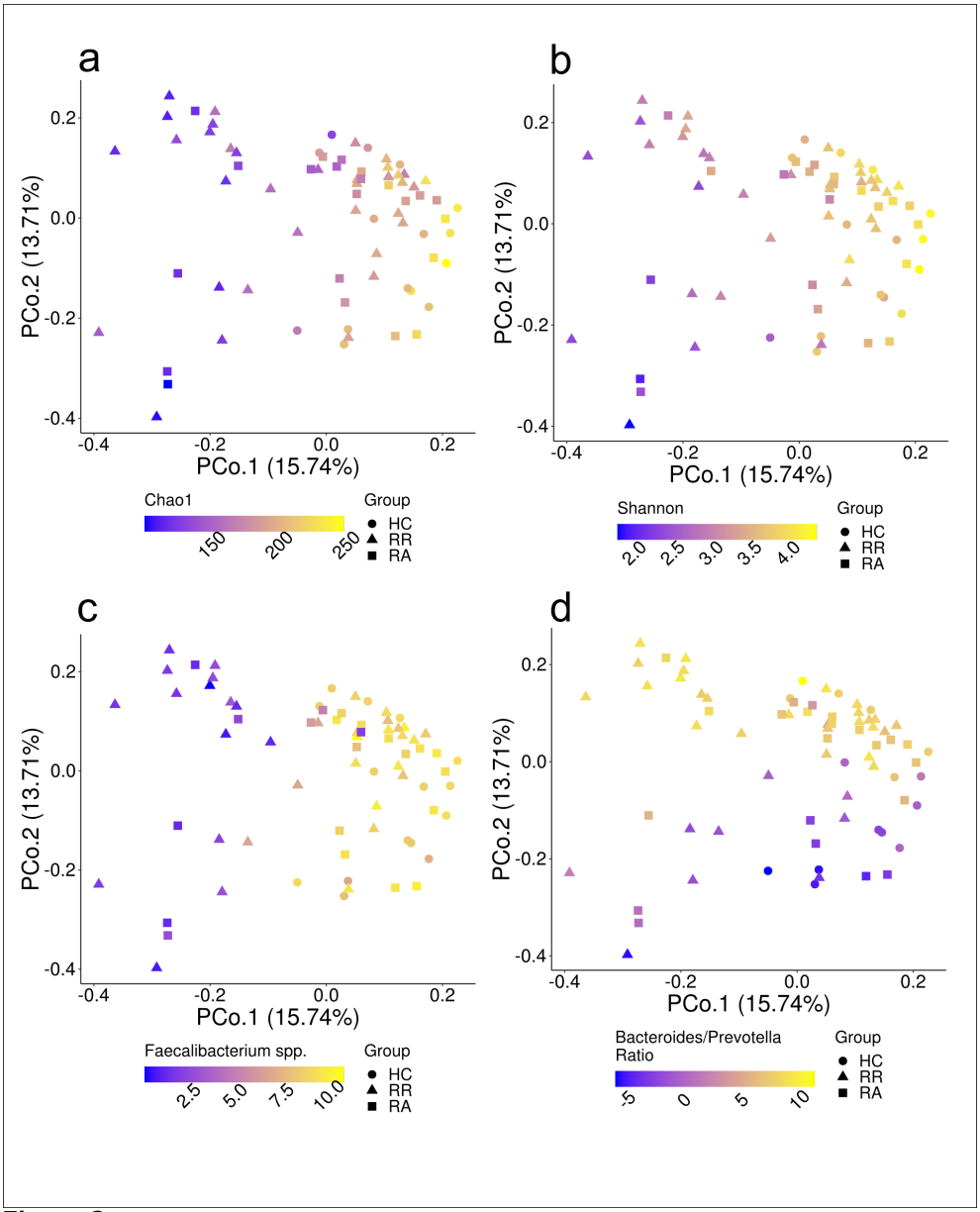


Figure 3 PCoA based on weighted UniFrac distance metric of baseline (T1) faecal microbial community structure in healthy controls and Crohn's disease patients. Samples are coloured based on (a) Chao1 index, (b) Shannon Index, (c) relative abundance of Faecalibacterium spp. and (d) log2 ratio of the relative abundance of Bacteroides spp. and Prevotella spp. Alpha diversity and abundance of Faecalibacterium spp. drive separation along the first principal coordinate, whereas the Bacteroides to Prevotella ratio drives separation along the second coordinate.

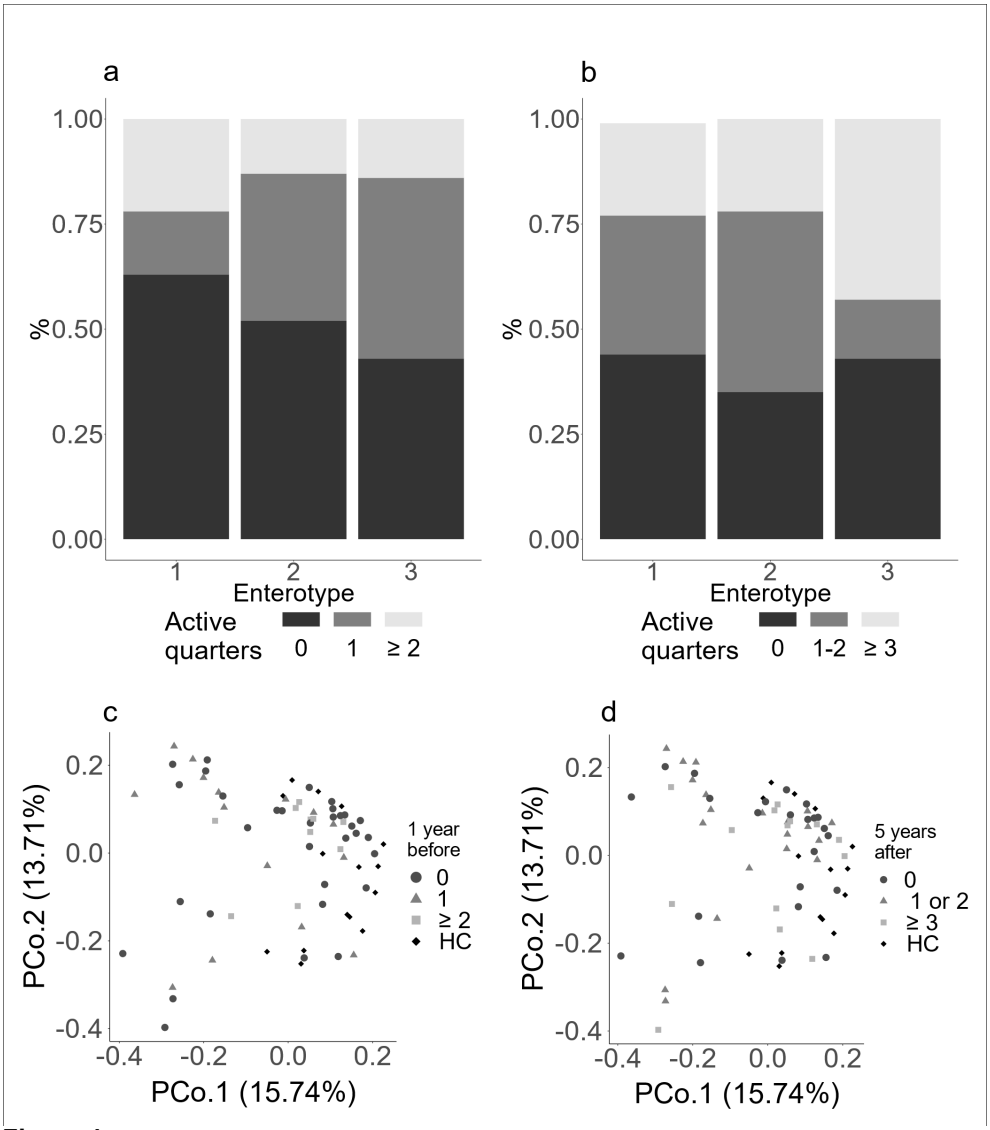


Figure 4 The number of active disease quarters is not associated with the microbial community structure. (a) bar plot representing the proportion of CD patients among each enterotype with 0, 1 or ≥2 active quarters within the year prior to inclusion, (b) bar plot depicting the proportion of CD patients among each of the enterotypes with 0, 1-2 or ≥3 active quarters in the 5 years after inclusion, c+d) PCoA based on weighted UniFrac distance metric of baseline faecal microbiota samples of healthy controls and Crohn's disease patients. Samples are coloured based on the number of active quarters during (c) 1 year before inclusion, and (d) 5 years after inclusion.

Enterotype analysis revealed the presence of three enterotypes driven by high abundances of Bacteroides spp. (E1) and Prevotella spp. (E2) and a low abundance of Faecalibacterium spp. (E3) but none of the enterotype was significantly associated with one of the subject groups.

Next, we performed a PCoA based on the weighted UniFrac distance of the faecal microbiota of CD patients and HC at baseline, aiming to visualize differences among sample groups and to identify factors driving the separation of samples. The results highlighted a subgroup of CD patients that clustered apart from HC along the first principal coordinate.

This subgroup was characterized by a lower relative abundance of Faecalibacterium spp., as well as a lower microbial richness and diversity (Figure 3a-c). In addition, the PCoA showed that the ratio of the relative abundance between Bacteroides spp. and Prevotella spp seems to drive the separation along the second principal coordinate (Figure 3d).

Baseline microbial community structure in association to preceding and subsequent disease course

Finally, we examined whether disease activity in the year preceding the baseline sampling was predictive for the microbial community structure or whether the microbial community structure was predictive for the disease activity up to 5 years after sampling.

Although enterotype 1 comprised slightly more patients without active quarters within the year prior to baseline when compared to the other enterotypes, the distribution of the number of active quarters in the year prior to sampling was not statistically significantly different among the three enterotypes as examined by a generalized linear model using a logistic regression (Figure 4a, Supplementary Table 3). PCoA and PERMANOVA based on weighted UniFrac distance of the overall community structure also did not show any separation according to the clinical history (Figure 4c, Supplementary Table 5). Also the disease activity in the 5 years following baseline sampling did neither show any association with baseline enterotype (Figure 4b) (Supplementary Table 4) nor with the microbial community structure based upon the weighted UniFrac (Figure 4d, Supplementary Tables 6). This indicates that the overall microbial community structure appears not to be predictive for future disease course.

Temporal dynamics of the microbial richness, diversity, and community structure

Changes in alpha and beta diversity indices within study subjects over time were compared between RR and RA patients and HC. Although the microbial richness and diversity of CD patients was lower than HC at baseline (Supplementary Figure 1), the temporal dynamics of these parameters did not differ significantly between healthy controls and CD patients that either maintained remission (RR) or developed an exacerbation during follow-up (RA) (Figure 5).

We next examined the fluctuation over time of the individual bacterial genera in association with the disease groups. Although some bacterial genera seem to increase over time in the RA group and decrease in the RR group (Supplementary Figure 2), the Kruskall-Wallis test prove that, after false-discovery rate adjustment for multiple com-

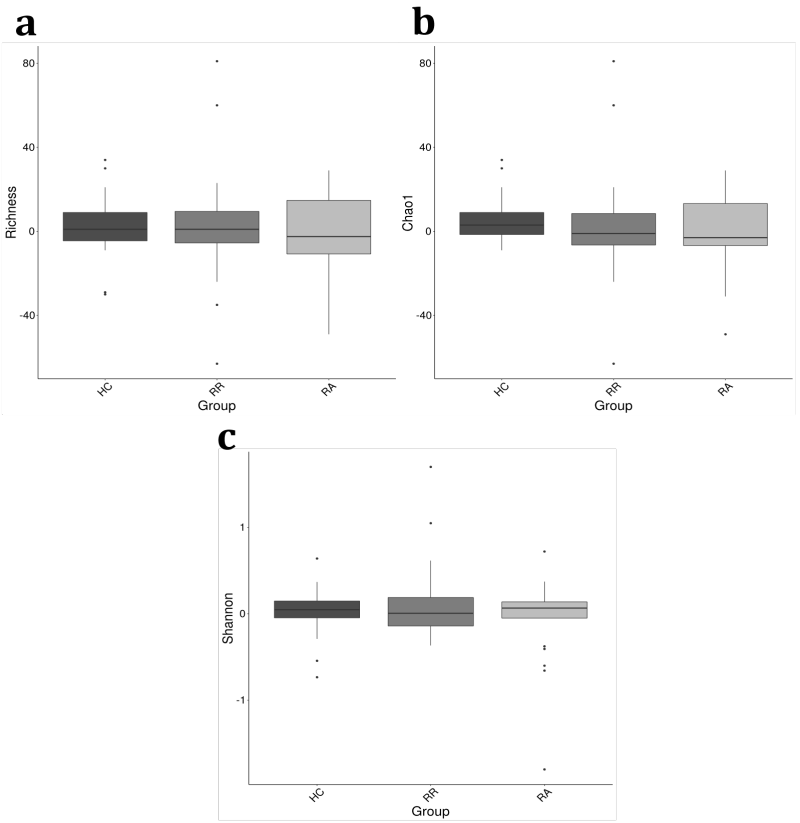


Figure 5 Changes in alpha-diversity indices in samples collected at first and second time-point (T1-T2) from healthy controls (HC), CD patients staying in remission (RR) and CD patients in remission followed by an exacerbation (RA). Panels depict: (a) observed species, (b) Chao1 index and (c) Shannon index between. Changes in alpha-diversity were not significantly different between HC, RR and RA groups. Significance was tested using Wilcoxson Signed-Ranks Test.

parisons, this trend is not significant (Supplementary Table 7).

We then investigated in the temporal dynamics in the microbial community structure as indicated by the within-subject beta-diversity. First, a PCoA was performed based on the weighted UniFrac distance of faecal samples from healthy subjects and CD patients at baseline and the second time-point (Figure 6a). As for the baseline data, no discrete separation between RR and RA samples could be observed. However, the temporal stability in microbial community structure between two subsequent samples, as indicated by the weighted UniFrac distance, was significantly higher in healthy controls as compared to the RA patients (Fig. 6b). When performing similar analyses based upon the unweighted UniFrac, the microbial community structure of healthy subjects appeared to be more stable than that of both RR and RA patients (Supplementary Figure 3). We next examined whether subjects switched enterotypes over time. None of the healthy controls changed enterotype during the sampling period whereas 9 CD patients switched from one enterotype to another, again indicating a lower stability in (some) CD patients as compared to HC. The proportion of patients that changed enterotype (6/35 for RR and 3/22 for RA) was however not associated with disease course.

We then investigated if the temporal stability of the microbiota composition of CD patients was related with the variation over time of Calprotectin or CRP. As expected, the RA group show higher levels of faecal Calprotectin over time as compared to RR group. Nonetheless our results prove that the variation over time of both Calprotectin and CRP does not affect the temporal stability of the microbiota composition (Supplementary Figure 4)

To examine the possibility that the temporal stability was confounded by variation in the time-period (in weeks) between the collection of the two subsequent samples, we constructed a linear model between time and the weighted UniFrac distance per study group (Figure 6c). These analyses did not reveal any evidence that the difference in collection time acted as a confounding factor (p-values HC p=0.64; RR p=0.16; RA p=0.16).

We subsequently examined whether the stability of the gut microbiota might also differ for CD patients according to disease localizations (ileal, colonic or ileocolonic CD) or abdominal surgery. We found that only the subgroup of colonic CD patients had a gut microbiota composition that was statistically significantly less stable when compared to the healthy subjects (Supplementary Figure 5a), whereas abdominal surgery did not affect microbiota stability (Supplementary Figure 5b).

To better understand the covariates that drive the microbial variation between samples, we next used distance-based Redundancy-Analysis (dbRDA) as an additional ordination method. The results show that, when looking at the CD patients only, disease activity only creates a minor shift on the spatial distribution of the data, which is insufficient to create separate clusters (Supplementary figures 6 and 7).

Using permutational multivariate analyses of variance (PERMANOVA), we ruled out that our study results were not confounded by Montreal classification factors (age at diagnosis, disease localization and disease phenotype), medication use (mesalazines, thiopurines, biologicals, corticosteroids, or proton pump inhibitors) prior to or during the study period, age, gender, number of liquid stools/day, surgery or original sequencing depth. Only disease phenotype was statistically significantly associated with the microbial community structure of CD patients (supplementary Table 8). This association was mainly related to a significantly different microbial community structure in

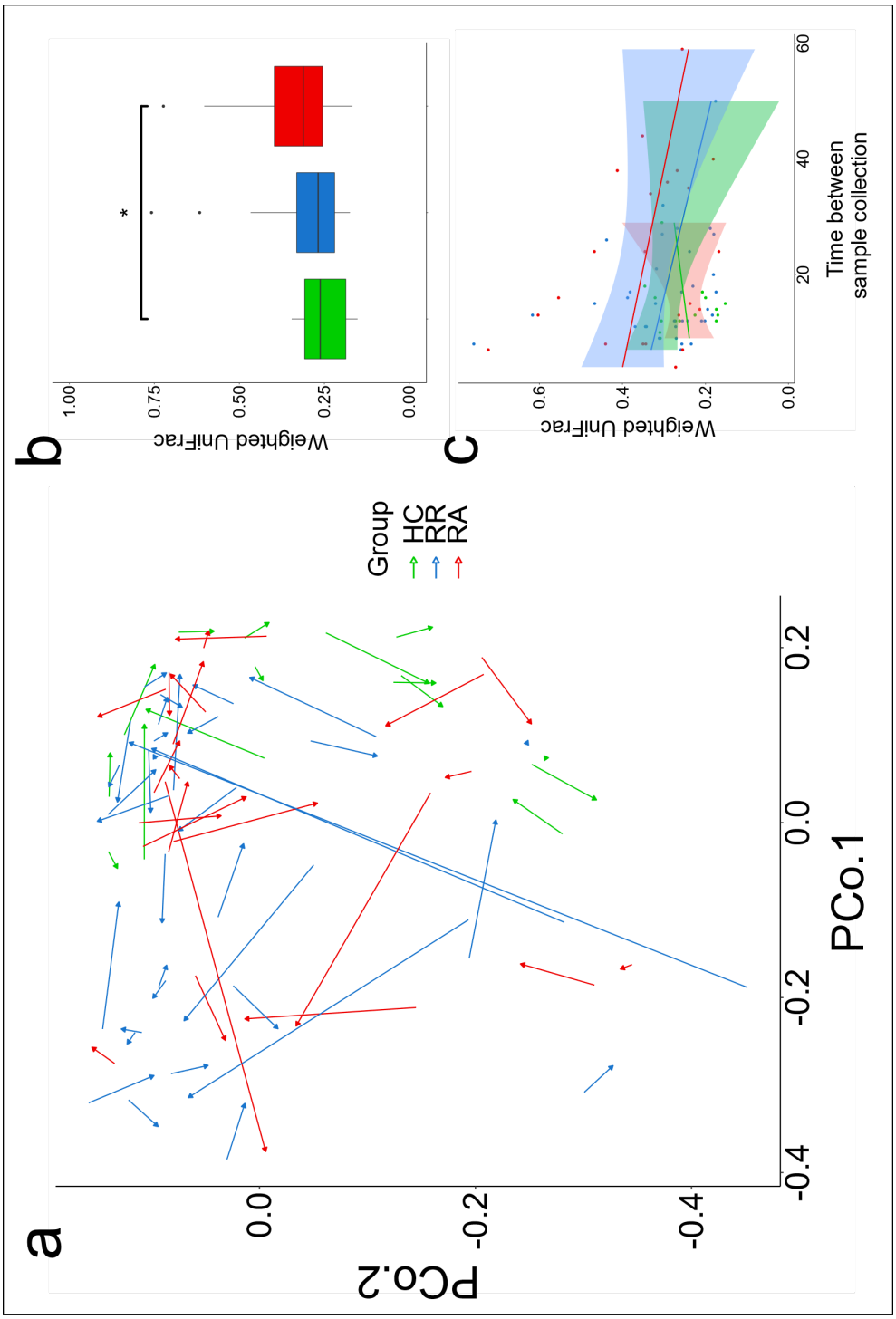

a
PCo.2
0.0
-0.2
-0.4
PCo.1
-0.4
-0.2
0.0
0.2
Group
HC
RR
RA
b
Weighted UniFrac
1.00
0.75
0.50
0.25
0.00
*
c
Weighted UniFrac
0.0
0.2
0.4
0.6
Time between sample collection
0
20
40
60

Figure 6 (a) PcoA based on within-subject weighted UniFrac distance of faecal microbiota at baseline (T1) and second sampling time-point (T2). Samples of healthy controls (HC) are indicated in green, whereas samples of CD patients that remain in remission (RR) are indicated in blue and patients that develop an exacerbation (RA) in red. The arrows connect two samples from the same individual. The direction goes from T1 to T2. (b) Healthy controls show a statistically significantly smaller within-subject UniFrac distance between the two subsequent time-points, when compared to patients that develop an exacerbation (RA) whereas no difference is observed when compared to patients that remain in remission. Significance was tested using Wilcoxson Signed-Ranks Test; * indicates $p<0.05$.
(c) There is no association between the within-subject distance and the actual time (in weeks) between subsequent sampling time-points as assessed by a linear model.

CD patients with a penetrating (B3) phenotype when compared to the non-structuring/ non-penetrating (B1) phenotype (Supplementary Table 9). This association is partially in contrast with previous findings in which the authors argue that gut bacterial infections do not play a major role on maintaining the fistulas phenotype in CD patients[24]. Further studies are auspicable to further characterize the nature of this association.

Altogether, our findings suggest that the microbiota community structure only marginally differs between RA and RR patients.

Discussion

To the best of our knowledge, together with the work from Pascal, Pozuelo [8] and Halfvarson, Brislawn [23], this is one of the largest longitudinal studies that comprehensively investigated the stability of the faecal microbiota of adult CD patients during their disease course. First, CD patients showed a lower microbial richness and diversity when compared to HC. Second, a subset of CD patients clustered separate from healthy controls and were characterized by a low microbial diversity and a relatively low abundance of *Faecalibacterium* spp. Third, the temporal stability of the microbial community structure was lower in CD patients when compared to healthy controls, but the microbial stability was not affected by changes in disease activity. And finally, the overall microbial community structure was not associated with disease history or subsequent disease course.

By collecting multiple samples of healthy individuals and CD patients both with and without a changing disease activity over time, we were able to assess the microbial stability and to investigate the microbial changes during remission and active disease, thereby limiting potential confounding associated with cross-sectional studies. The present study confirms previous observations that the faecal microbiota of CD patients is less diverse as compared to HC [33]. This lower stability in CD was for some indices (weighted UniFrac) only reached statistical significance for the RA patients, whereas for other indices of microbial stability (unweighted UniFrac) both RR and RA patients has a lower temporal stability when compared to HC. However, for none of the indices of microbial stability we found a significant difference between the RA and RR patient groups. Altogether these results prove that HC have a stronger temporal stability of the microbial community structure when compared to CD patients, regardless of whether these patients maintained remission or developed an exacerbation. Our results hereby can be used to confirm that CD patients have a less rich microbiota with larger intra-individual variations. Moreover, we observed that especially patients with colonic disease had a lower temporal microbiota stability.

Although the absence of an altered microbial composition and stability in patients developing active disease as compared to patients maintaining remission is in agreement with previous longitudinal studies [8, 23], it contrasts with several previous cross-sectional studies [11, 16]. As indicated by the study of Halfvarson [23] changes in medication use are more strongly linked to the dynamics in the microbial community structure than changes in disease activity. This might explain why previous cross-sectional studies, with large variations in medication use between subjects, have reported stronger associations between disease activity and microbiota composition. Altogether, this pleas for longitudinal studies with repeated sampling to rule out confounding, also

questions the involvement of the overall microbiota in the development of exacerbations. Still, small shifts in (a combination of) specific taxa may be present in CD patients with changing disease activity over time. It is also plausible that patient specific changes are present, but due to our focus on the overall microbiota composition, these changes remained undetected. Therefore, further analyses focusing on (small) changes in individual taxa are warranted in large groups of patients with longitudinal follow-up, considering disease phenotype, medication use, surgery as well as dietary habits. The latter factor was not included in the present study as dietary information was not available yet could be a potential reason for the lack of consistent changes in the microbiota community structure between CD patients.

We found a subgroup of CD patients of whom the microbiota composition, characterized by a low microbial richness and diversity and a low relative abundance of *Faecalibacterium* spp., deviated from the microbiota of HC. Subgroups of CD patients that clustered apart from other CD patients and healthy controls have also been demonstrated in previous studies [6, 23, 34, 35]. Consistent with the recent study of Halfvarson *et al.*[23], we found this subgroup of CD patients to be characterized by lower *F. prausnitzii* abundance and low microbial richness, although in our study this subcluster was not restricted to ileal CD. The existence of a subgroup of CD patients with a deviating microbiota composition might (in part) be explained by disease-related factors. We could however not find clear differences in age at onset, disease localization, disease behaviour, disease activity before and after inclusion in the study, number of liquid stools per day, surgery or medication use when comparing the CD patients with a more deviant microbiota profile versus those with a more 'healthy' microbiota profile. It should however be noted that numbers were relatively small. Larger studies are needed to further characterize the subgroup of CD patients that do not cluster with HC, and to investigate whether this altered microbiota might be related to a more microbiota-driven disease etiology or certain host or environmental factors. Incorporation of host genetics, metabolomics and/or transcriptomics data in future large-scale studies could potentially explain the reason for this subgroup of CD patients.

A potential limitation of our study is the collection and transport of faecal samples at room temperature. However, although immediate freezing of samples is considered the gold standard, we already proved that the sample processing as applied in the present study does not significantly alter the microbiota composition. On the other hand, our study has several strengths, in particular the longitudinal study design. Although cross-sectional microbiota studies are restricted by the large inter-individual variation of the microbiota, most microbiota studies on disease activity in adult CD patients are based on a cross-sectional design. Longitudinal studies are able to circumvent this limitation. Furthermore, only small numbers of patients in each group (RR and RA) did have a change in medication use in between the consecutive samples, further limiting potential bias due to confounding. Moreover, repeating our analyses without these patients did not impact our findings (data not shown).

Another strength of our study is the use of a composite score, including both clinical and inflammation markers, to determine disease activity. Repeated endoscopy, which is the golden standard for disease activity assessment, is too invasive in a longitudinal patient cohort. Most previous studies, including the longitudinal study of Pascal and colleagues ([8, 36, 37], have used clinical activity indices to assess disease activity [11,

13, 17, 20, 21], which have however shown to correlate only moderately with mucosal inflammation [26]. A combination of inflammatory and clinical markers as used in the present study, provides a more reliable and accepted surrogate for mucosal inflammation[26]. Finally, due to the deeply phenotyped patients included, we were able to explore the association of the microbiota composition and stability with long-term disease course.

In conclusion, our prospective longitudinal study showed that the altered microbiota composition and stability in CD was not associated with disease activity or with long-term disease course. This questions the involvement of the overall microbiota structure in the development of exacerbations. The aberrant microbiota composition in a subset of CD patients warrants further exploration of a more microbiota-driven etiology in this group.

Acknowledgements

This study was financially supported by a grant from the Academic Fund of Maastricht UMC+.

References

1. Molodecky, N.A., et al., Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. Gastroenterology, 2012. 142(1): p. 46-54.e42.
2. Baumgart, D.C. and W.J. Sandborn, Crohn's disease. The Lancet, 2012. 380(9853): p. 1590-1605.
3. Burisch, J., et al., The burden of inflammatory bowel disease in Europe. Journal of Crohn's and Colitis, 2013. 7(4): p. 322-337.
4. Romberg-Camps, M.J.L., et al., Fatigue and health-related quality of life in inflammatory bowel disease. Inflammatory Bowel Diseases, 2010. 16(12): p. 2137-2147.
5. Fujimoto, T., et al., Decreased abundance ofFaecalibacterium prausnitziiin the gut microbiota of Crohn's disease. Journal of Gastroenterology and Hepatology, 2013. 28(4): p. 613-619.
6. Gevers, D., et al., The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host & Microbe, 2014. 15(3): p. 382-392.
7. Morgan, X.C., et al., Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology, 2012. 13(9): p. R79.
8. Pascal, V., et al., A microbial signature for Crohn's disease. Gut, 2017. 66(5): p. 813-822.
9. Walker, A.W., et al., High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. BMC Microbiology, 2011. 11(1): p. 7.
10. Willing, B.P., et al., A Pyrosequencing Study in Twins Shows That Gastrointestinal Microbial Profiles Vary With Inflammatory Bowel Disease Phenotypes. Gastroenterology, 2010. 139(6): p. 1844-1854.e1.
11. Kolho, K.-L., et al., Faecal Microbiota in Pediatric Inflammatory Bowel Disease and Its Relation to Inflammation. The American Journal of Gastroenterology, 2015. 110(6): p. 921-930.
12. Papa, E., et al., Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease. PLoS ONE, 2012. 7(6): p. e39242.
13. Andoh, A., et al., Characterization of gut microbiota profiles by disease activity in patients with Crohn's disease using data mining analysis of terminal restriction fragment length polymorphisms. Biomedical Reports, 2014.
14. Sokol, H., et al., Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflammatory Bowel Diseases, 2009. 15(8): p. 1183-1189.
15. Swidsinski, A., et al., Active Crohn's disease and ulcerative colitis can be specifically diagnosed and monitored based on the biostructure of the faecal flora. Inflammatory Bowel Diseases, 2008. 14(2): p. 147-161.
16. Wang, W., et al., Increased Proportions of Bifidobacterium and the Lactobacillus Group and Loss of Butyrate-Producing Bacteria in Inflammatory Bowel Disease. Journal of Clinical Microbiology, 2013. 52(2): p. 398-406.
17. Seksik, P., Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. Gut, 2003. 52(2): p. 237-242.
18. De Cruz, P., et al., Association between specific mucosa-associated microbiota in Crohn's disease at the time of resection and subsequent disease recurrence: A pilot study. Journal of Gastroenterology and Hepatology, 2015. 30(2): p. 268-278.
19. Dey, N., et al., Association of gut microbiota with post-operative clinical course in Crohn's disease. BMC Gastroenterology, 2013. 13(1).
20. Rajca, S., et al., Alterations in the Intestinal Microbiome (Dysbiosis) as a Predictor of Relapse After Infliximab Withdrawal in Crohn's Disease. Inflammatory Bowel Diseases, 2014: p. 1.
21. Scanlan, P.D., et al., Culture-Independent Analyses of Temporal Variation of the Dominant Faecal Microbiota and Targeted Bacterial Subgroups in Crohn's Disease. Journal of Clinical Microbiology, 2006. 44(11): p. 3980-3988.
22. Wills, E.S., et al., Faecal Microbial Composition of Ulcerative Colitis and Crohn's Disease Patients in Remission and Subsequent Exacerbation. PLoS ONE, 2014. 9(3): p. e90981.
23. Halfvarson, J., et al., Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol, 2017. 2: p. 17004.
24. van den Heuvel, T.R.A., et al., Cohort Profile: The Inflammatory Bowel Disease South Limburg Cohort (IBDSL). International Journal of Epidemiology, 2015. 46(2): p. e7-e7.
25. Bodelier, A.G.L., et al., Pancreatitis-associated protein has no additional value as a marker of disease activity in a real-life cohort of IBD patients. European Journal of Gastroenterology & Hepatology, 2014. 26(8): p. 902-909.
26. Falvey, J.D., et al., Disease Activity Assessment in IBD. Inflammatory Bowel Diseases, 2015. 21(4): p. 824-831.
27. Mujagic, Z., et al., Small intestinal permeability is increased in diarrhoea predominant IBS, while alterations in gastroduodenal permeability in all IBS subtypes are largely attributable to confounders. Alimentary Pharmacology & Therapeutics, 2014. 40(3): p. 288-297.
28. Tedjo, D.I., et al., The effect of sampling and storage on the faecal microbiota composition in healthy and diseased subjects. PLoS One, 2015. 10(5): p. e0126685.
29. Lagkouvardos, I., et al., Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ, 2017. 5: p. e2836.
30. Morgan, M., DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning for Microbiome Data. 2017.
31. Satsangi, J., et al., The Montreal classification of inflammatory bowel disease: controversies, consensus,

and implications. Gut, 2006. 55(6): p. 749-53.
32. Legendre, P. and M.J. Anderson, DISTANCE-BASED REDUNDANCY ANALYSIS: TESTING MULTISPECIES RESPONSES IN MULTIFACTORIAL ECOLOGICAL EXPERIMENTS. Ecological Monographs, 1999. 69(1): p. 1-24.
33. Manichanh, C., et al., The gut microbiota in IBD. Nature Reviews Gastroenterology & Hepatology, 2012. 9(10): p. 599-608.
34. Frank, D.N., et al., Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences, 2007. 104(34): p. 13780-13785.
35. Lewis, James D., et al., Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host & Microbe, 2015. 18(4): p. 489-500.
36. af Björkesten, C.-G., et al., Surrogate markers and clinical indices, alone or combined, as indicators for endoscopic remission in anti-TNF-treated luminal Crohn's disease. Scandinavian Journal of Gastroenterology, 2012. 47(5): p. 528-537.
37. D'Haens, G., et al., Faecal calprotectin is a surrogate marker for endoscopic lesions in inflammatory bowel disease. Inflammatory Bowel Diseases, 2012. 18(12): p. 2218-2224.

Supplementary methods

DNA isolation and sequencing

DNA isolation of faecal samples was performed in batches of 11 or 23 by repeated bead beating in combination with the PSP spin stool kit (Stratec Molecular, Berlin, Germany) as described previously [1]. For each DNA isolation batch, one additional isolation was performed on PCR-grade water as a negative control.

Amplicon library preparation and sequencing was performed according to a previous published protocol [2]. The 515f/806r primer pair was used to amplify the V4 region of the 16S rRNA gene. PCR reactions were performed using 25 µL NEB Phusion High-Fidelity Master Mix (New England Biolabs, Ipswich, USA), 4 µl 515f/806r primer mix and 30 ng metagenomics DNA under the following conditions: denaturation at 98oC for 3 minutes, followed by 30 cycles of denaturation at 98oC for 45 seconds, annealing at 55oC for 45 seconds and extension at 72oC for 45 seconds. The final elongation step was at 72oC for 7 minutes. Amplicons were purified using the AMPure XP purification (Agencourt, Massachusetts, USA) according to the manufacturer's instructions. Amplicons were mixed in equimolar concentrations and sequenced on an Illumina MiSeq instrument.

Sequence Analysis

Quality control of the sequencing data were performed using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) using default settings. Data demultiplexing, length and quality filtering and clustering of reads into Operational Taxonomic Units (OTUs) at 97% sequence identity was done using the online Integrated Microbial Next Generation Sequencing (IMNGS) platform [3] using default settings except for minimum and maximum length for amplicons, which were set at 100 and 500 bp, respectively.

After quality filtering and binning and removing unassigned reads, a total of 2,829,437 sequences with an average of 19,649 paired sequences per sample (range: 11,744-26,641 sequences/sample) remained for downstream analysis and were clustered in 473 OTUs.

Data normalization, alpha indices, taxonomical binning, and unsupervised clustering were performed using Rhea [4].

In order not to discard informative data, normalization in Rhea is performed by dividing OTU counts per sample for their total count (sample depth) and then multiplying the obtained relative abundance for the lowest sample depth (11744 reads/sample).

References

1. Tedjo DI, Jonkers DMAE, Savelkoul PH, et al. The Effect of Sampling and Storage on the Faecal Microbiota Composition in Healthy and Diseased Subjects. PLOS ONE 2015;10:e0126685.
2. Caporaso JG, Lauber CL, Walters WA, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 2012;6:1621-1624.
3. Lagkouvardos I, Joseph D, Kapfhammer M, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci Rep 2016;6:33721.
4. Lagkouvardos I, Fischer S, Kumar N, et al. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ 2017;5:e2836.

Supplementary Tables

Supplementary Table 1 CD patients characteristics

	Remission-remission (n=35)	**Remission-active (n=22)**	**Healthy controls (n=15)**
Age at inclusion (median, IQR)	43(33-53)	38(26-59)	25 (23-30)
Male (%)	10(28.6)	10(45.5)	8 (53.3)
Smoking (%)	8(22.9)	2(9.1)	0 (0)
Age at diagnosis[1] (%)			
A1 (<16 year)	1(2.9)	1(4.5)	Na
A2 (17-40 year)	30(85.7)	13(59.1)	
A3 (>40 year)	4(11.4)	8(36.4)	
Disease localization[1] (%)			
L1 (ileal)	12(34.3)	7(31.8)	Na
L2 (colonic)	8(22.9)	7(31.8)	
L3 (ileocolonic)	15(42.9)	87(36.4)	
Phenotype at inclusion[1] (%)			
B1 (nonstricturing, non penetrating)	26(74.3)	12(54.5)	Na
B2 (stricturing)	6(17.1)	5(22.7)	
B3 (penetrating)	3 (8.6)	5 (22.7)	
Abdominal Surgery[2]	8 (22.9)	4 (18.2)	0 (0)

[1] According to Montreal Classification

[2] includes (hemi)colectomy and ileocecal resection

Supplementary Table 2 Medication use and time between sampling moments for active and remission samples.

	RR		RA	
	Remission (n=35)	**Remission (n=35)**	**Remission (n=22)**	**Active (n=22)**
Medication (%)‡				
Mesalazine	5(14.3)	5(14.3)	4(18.2)	5(22.7)
Thiopurines	11 (31.4)	11(31.4)	9(40.9)	7(31.8)
Biologicals	18 (51.4)	19(54.3)	13 (59.1)	15(68.2)
Corticosteroids	1(2.9)	0(0)	1(4.5)	1(4.5)
Proton Pump Inhibitors	7(20)	7(20)	8(36.4)	8(36.4)
Antibiotics#	1(2.9)	0(0)	1(4.5)	0(0)
Time between sampling moments (week, median, IQR)	-	14(11-21)[1]	-	20(8-36)[2]

‡Six RR and five RA patients had a medication change between consecutive samples during the study period. In the RR group, mesalazine was stopped by 1 patients, prednisone by 1 patient, and biologicals in 2 patients, while 1 patient started mesalazine and 1 patient with started biologicals. In the RA group, 2 patients started with biologicals, 2 patients stopped with thiopurines and 1 patient started with mesalazine.

[1]Time between first remission and second remission samples

[2]Time between first remission and first active samples

#Ciprofloxacin and cotrimoxazol were used two and one month prior to sample collection, respectively.

Supplementary Table 3 z-statistics and p-value resulting from Generalized Linear Model on clinical history 1 year before the study period and enterotype clusters.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.53063	0.398527	-1.33147	0.183033
E2	0.443617	0.577119	0.768675	0.442086
E3	0.81831	0.861485	0.949883	0.342172

Supplementary Table 4 z-statistics and p-value resulting from Generalized Linear Model on clinical course during the5 years following the study period and enterotype clusters.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.223144	0.387298	0.576154	0.564511
E2	0.405465	0.584523	0.693669	0.48789
E3	0.064539	0.856349	0.075365	0.939925

Supplementary Table 5 F-statistics and p-value resulting from PERMANOVA on microbiota composition using the clinical history 1 year before the study period as explanatory variable

	Df	SumOfSqs	R2	F	Pr(>F)
1 year before	2	0.393114439	0.036689	1.028328	0.399
Residual	54	10.32169678	0.963311	NA	NA
Total	56	10.71481121	1	NA	NA

Supplementary Table 6 F-statistics and p-value resulting from PERMANOVA on microbiota composition using the clinical course during the 5 year following the study period as explanatory variable.

	Df	SumOfSqs	R2	F	Pr(>F)
rec.5.Y.after	2	0.361603	0.033748	0.943019	0.5063
Residual	54	10.35321	0.966252	NA	NA
Total	56	10.71481	1	NA	NA

Supplementary Table 7 Kruskall-wallis test on the delta of log10 relative abundance of bacterial genera. The p-value is adjusted for multiple tests using FDR.

Genera	χ^2	p-value	FDR adj. p-value
Acidaminococcus	1.251437188	0.534876931	0.926024793
Actinomyces	0.70285608	0.703682485	0.926024793
Akkermansia	1.821321123	0.40225842	0.926024793
Alistipes	0.817114392	0.664608458	0.926024793
Allisonella	1.183107435	0.553466686	0.926024793
Alloprevotella	0.060219168	0.970339194	0.995745087
Anaerococcus	4.609475032	0.099784989	0.822426135
Anaeroglobus	0.548996798	0.759953221	0.926024793
Anaerostipes	1.229807863	0.540692836	0.926024793
Bacteroides	3.411117041	0.181670891	0.910635452
Barnesiella	0.476244836	0.788106207	0.926024793
Bifidobacterium	3.961706102	0.137951507	0.833112966
Bilophila	2.087301888	0.352166592	0.926024793
Blautia	1.187724604	0.552190435	0.926024793
Butyricicoccus	1.182671643	0.553587297	0.926024793
Butyricimonas	2.268014244	0.321741413	0.926024793
Campylobacter	7.733674009	0.020924449	0.655632729
Catenibacterium	0.607964904	0.737873815	0.926024793
Clostridium.IV	1.064213053	0.587366363	0.926024793
Clostridium.sensu.stricto	0.860878891	0.650223294	0.926024793
Clostridium.XI	0.31736955	0.853265287	0.947201524
Clostridium.XIVa	0.309773864	0.856512016	0.947201524
Clostridium.XIVb	0.369975997	0.831114258	0.941261931
Clostridium.XVIII	0.516850045	0.772266931	0.926024793
Collinsella	3.162352675	0.205732944	0.926024793
Coprobacter	2.41230142	0.299347339	0.926024793
Coprococcus	0.136336139	0.934103466	0.995745087
Desulfovibrio	1.49868898	0.472676295	0.926024793
Dialister	1.776196406	0.41143748	0.926024793
Dorea	0.569466682	0.752214812	0.926024793
Eikenella	0	1	1

Supplementary Table 7 (cont'd)

Elusimicrobium	1.302682403	0.521346077	0.926024793
Enterococcus	0.584954628	0.746412178	0.926024793
Erysipelotrichaceae_incertae_sedis	0.431879371	0.805783901	0.935107243
Escherichia.Shigella	2.16398723	0.338919177	0.926024793
Faecalibacterium	0.029918559	0.985152054	0.995745087
Finegoldia	0.608632941	0.737627392	0.926024793
Flavonifractor	2.140619952	0.342902209	0.926024793
Fusicatenibacter	0.036410427	0.9819595	0.995745087
Fusobacterium	1.953382623	0.376554943	0.926024793
Lachnospiracea_incertae_sedis	8.308279468	0.015699291	0.655632729
Lactobacillus	5.379217212	0.067907513	0.709256244
Mannheimia	0.624124288	0.731936041	0.926024793
Megamonas	5.789934664	0.055300832	0.709256244
Megasphaera	2.144329061	0.342266868	0.926024793
Methanobrevibacter	0.785070014	0.675342704	0.926024793
Mitsuokella	2.539643681	0.280881659	0.926024793
Morganella	1.926950373	0.381564571	0.926024793
Odoribacter	1.226367887	0.541623622	0.926024793
Olsenella	4.18549336	0.123347873	0.828192861
Oscillibacter	0.048919524	0.975836954	0.995745087
Parabacteroides	0.752890945	0.686296539	0.926024793
Paraprevotella	2.725283058	0.255983695	0.926024793
Parasutterella	6.008377315	0.049578964	0.709256244
Parvimonas	3.384936855	0.184064613	0.910635452
Pediococcus	4.89458601	0.086527499	0.813358493
Peptostreptococcus	1.203788293	0.54777309	0.926024793
Phascolarctobacterium	4.203202191	0.122260521	0.828192861
Prevotella	1.309533693	0.519563186	0.926024793
Proteus	0.657206222	0.719928692	0.926024793
Roseburia	2.012389057	0.365607645	0.926024793
Ruminococcus	0.529921524	0.767236054	0.926024793
Ruminococcus2	1.747224693	0.417440881	0.926024793
Slackia	10.34485425	0.005670788	0.533054109

Supplementary Table 7 (cont'd)

Streptococcus	4.507769476	0.10499057	0.822426135
Streptophyta	1.297010039	0.522826809	0.926024793
Succiniclasticum	0.390118573	0.822785877	0.941261931
Succinivibrio	2.983064024	0.225027646	0.926024793
Sutterella	5.56874024	0.061767983	0.709256244
Turicibacter	1.026805455	0.598455728	0.926024793
unknown Acidaminococcaceae	0.828487883	0.660839725	0.926024793
unknown Alphaproteobacteria	0.748436606	0.687826741	0.926024793
unknown Bacteria	2.474150501	0.290231833	0.926024793
unknown Bacteroidales	3.785334941	0.150669366	0.833112966
unknown Bacteroidetes	0.529467999	0.767410054	0.926024793
unknown Betaproteobacteria	1.655698003	0.436988237	0.926024793
unknown Burkholderiales	1.014929828	0.602019818	0.926024793
unknown Clostridia	2.132906103	0.344227311	0.926024793
unknown Clostridiales	0.100480342	0.951000994	0.995745087
unknown Coriobacteriaceae	1.099580371	0.577070875	0.926024793
unknown Desulfovibrionales	0.066157092	0.967462566	0.995745087
unknown Enterobacteriaceae	2.80881446	0.245512546	0.926024793
unknown Erysipelotrichaceae	0.947508364	0.622660295	0.926024793
unknown Firmicutes	1.804594195	0.405636802	0.926024793
unknown Lachnospiraceae	0.493679456	0.781265899	0.926024793
unknown Pasteurellaceae	1.210027412	0.546066942	0.926024793
unknown Peptostreptococcaceae	3.871963584	0.144282544	0.833112966
unknown Porphyromonadaceae	6.567552414	0.037486433	0.709256244
unknown Prevotellaceae	1.572437563	0.45556413	0.926024793
unknown Proteobacteria	5.8813429	0.052830244	0.709256244
unknown Ruminococcaceae	1.561176343	0.458136469	0.926024793
unknown Veillonellaceae	0.238858104	0.887426967	0.969978312
Veillonella	0.520185069	0.77098024	0.926024793
Victivallis	2.599967052	0.272536283	0.926024793

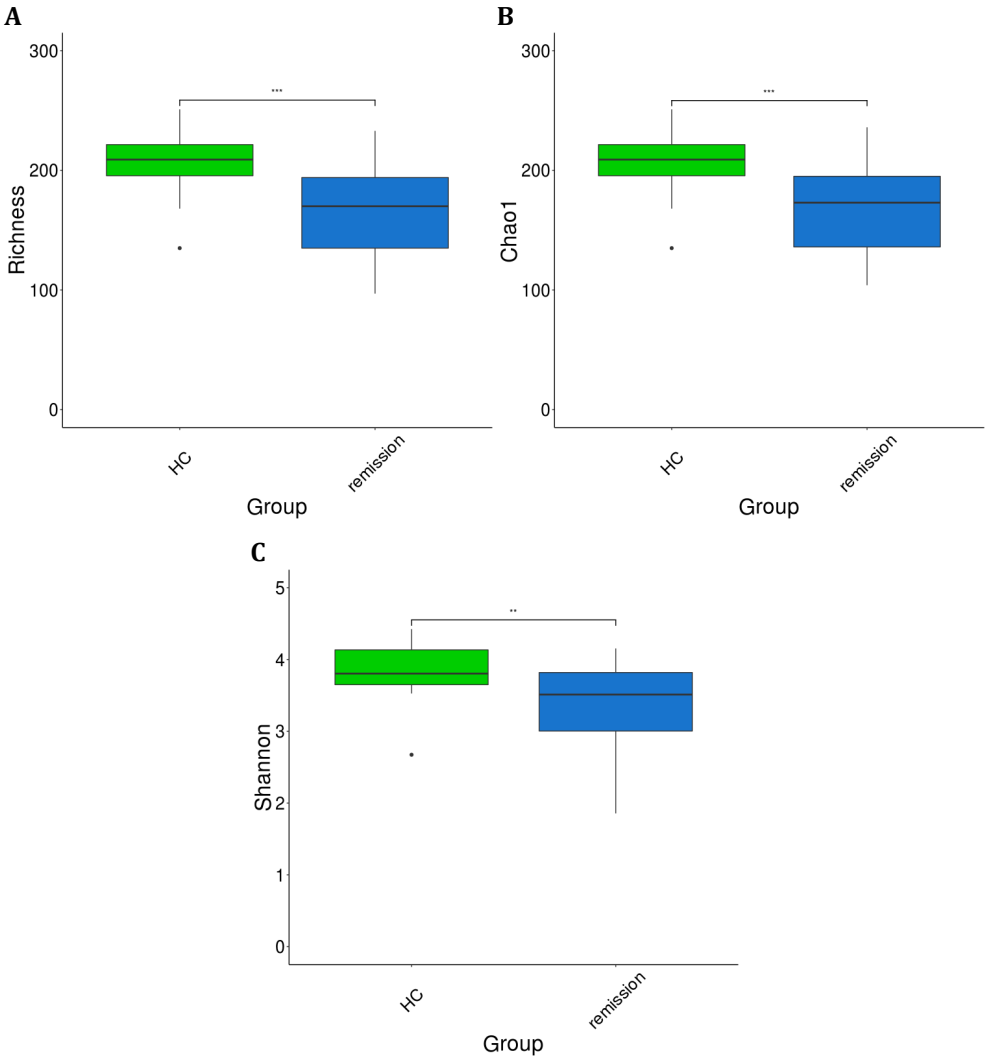
Supplementary Table 8 F-statistics and p-value resulting from PERMANOVA on microbiota composition using the Montreal classification factors, medication usage, smoking, surgery, and sample depth.

	F	Pr(>F)
Disease localization	0.82734	0.655
Age at diagnosis	0.843010	0.558
Phenotype	2.084877	0.016
Thiopurines	0.408099	0.971
Mesalazines	0.655223	0.773
Biological	0.581659	0.828
Corticosteroids	1.044867	0.293
Proton Pump Inhibitors	0.781195	0.61
Surgery	1.554117	0.123
Smoking	0.879233	0.591
Number of liquid stools/day	1.1946	0.247
Sequencing depth	1.709101	0.089

Supplementary Table 9 F-statistics and p-value resulting from PERMANOVA on microbiota composition using the disease phenotype: B1(non stricturing, non penetrating); B2(structuring); B3(penetrating).

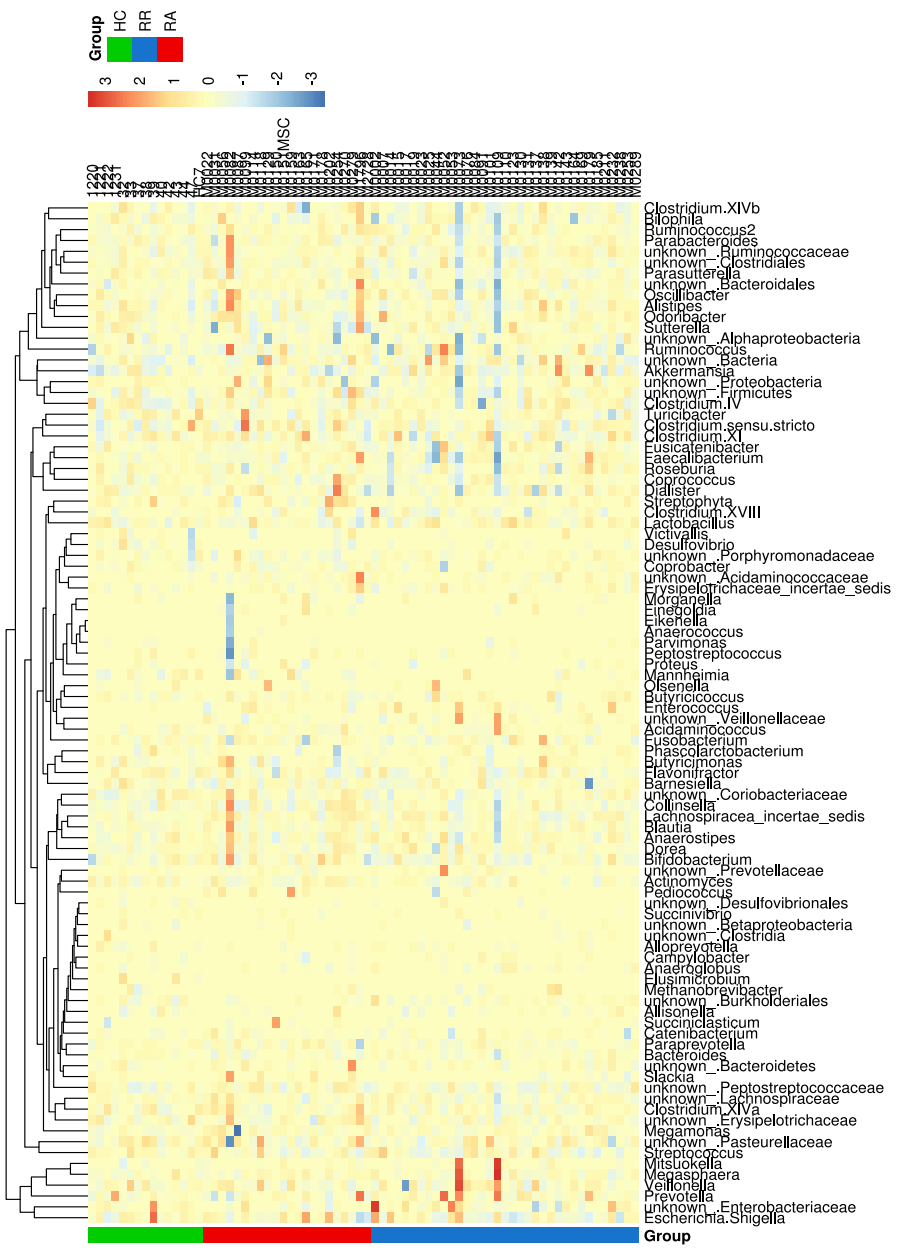
	F	Pr(>F)
B1 vs. B2	0.703237	0.691
B1 vs. B3	2.839477	0.013
B2 vs. B3	2.082749	0.039

Supplementary Figures

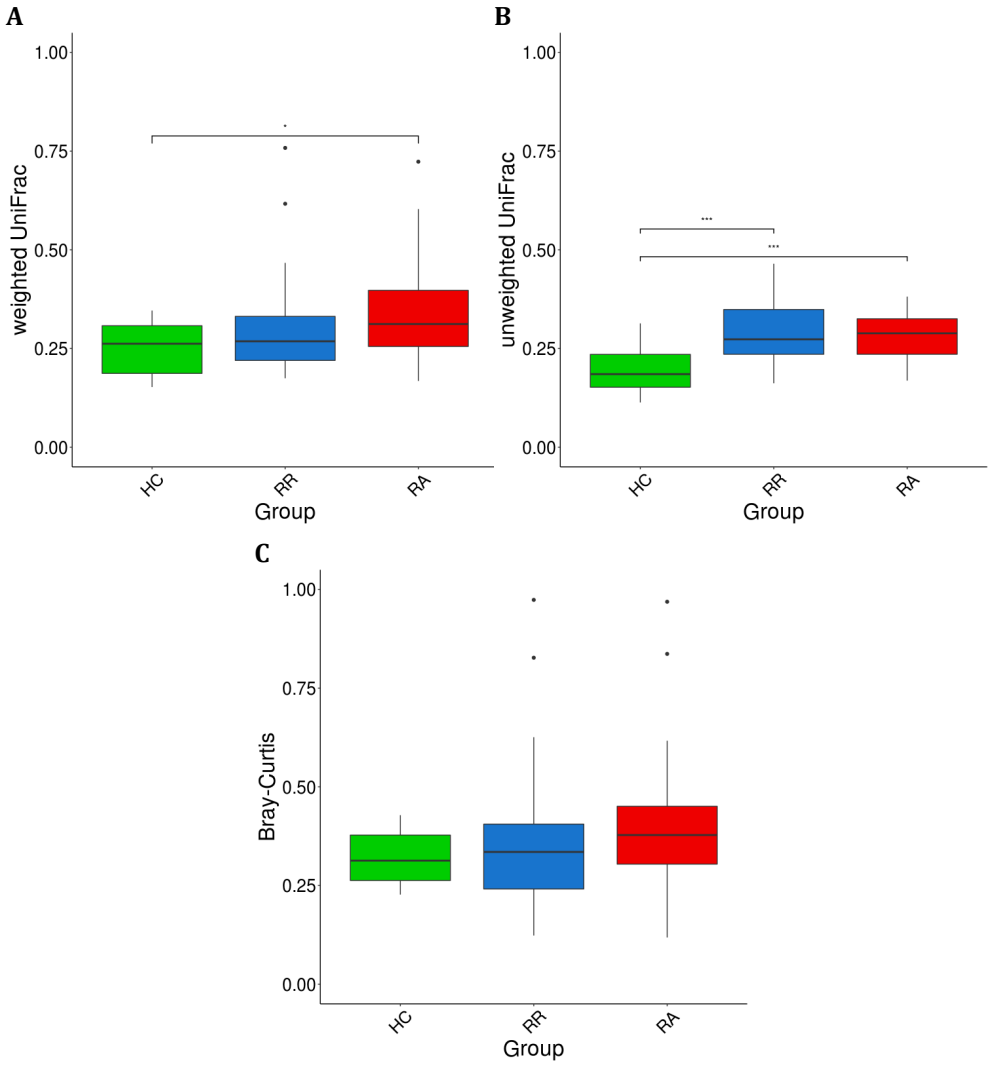


Supplementary Figure 1 Baseline alpha diversity indices: (a)Observed species (richness), (b) Shannon, and (c) Chao1 index within healthy controls (HC) and Crohn's disease patients at baseline. All patients were in remission state at baseline.

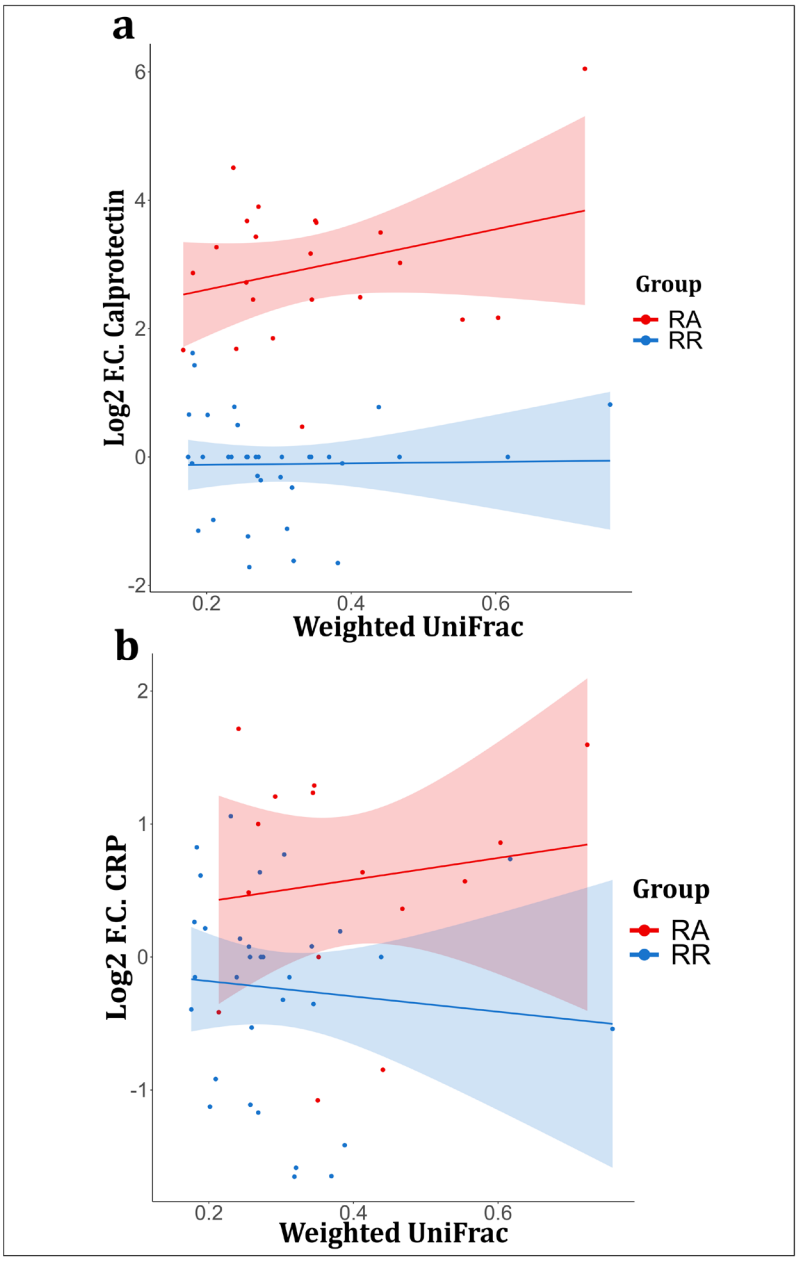
Significance was tested using Wilcoxson Signed-Ranks Test; *** indicates $p<0.01$.



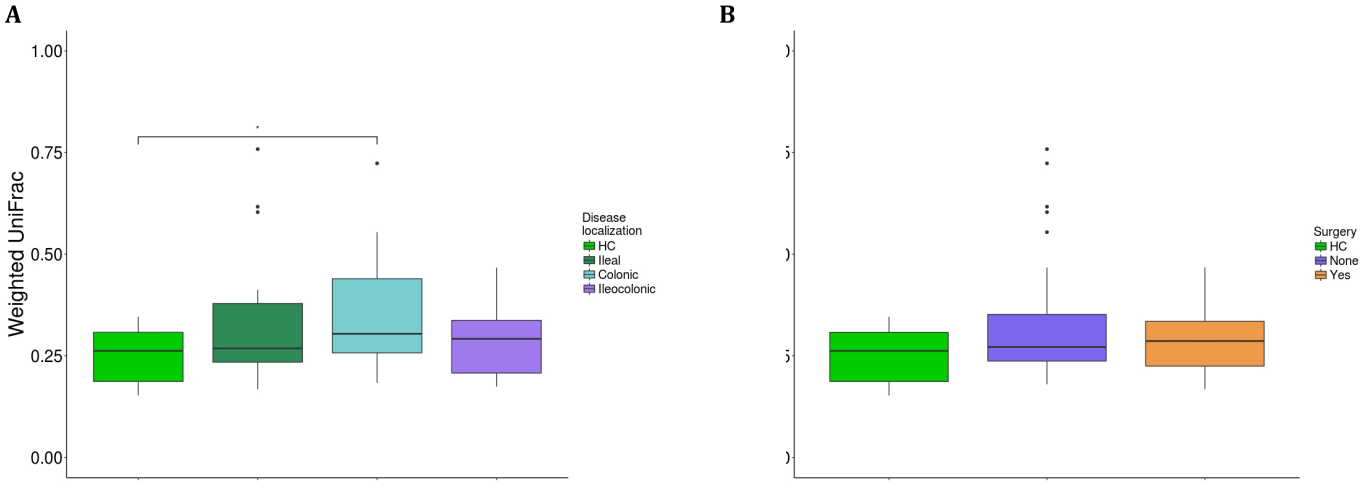
Supplementary Figure 2 Change in relative abundance over time of the bacterial genera. The cell colours represent the delta of log10 relative abundance of the two subsequent sampling time-points.



Supplementary Figure 3 Within-subject beta-diversity between two subsequent sampling time-points (T1-T2): (a)Weighted UniFrac, (b) Unweighted UniFrac and (c) Bray-Curtis distance, in healthy individuals (HC), CD patients maintaining in remission (RR) and CD patients in remission and subsequent exacerbation (RA).

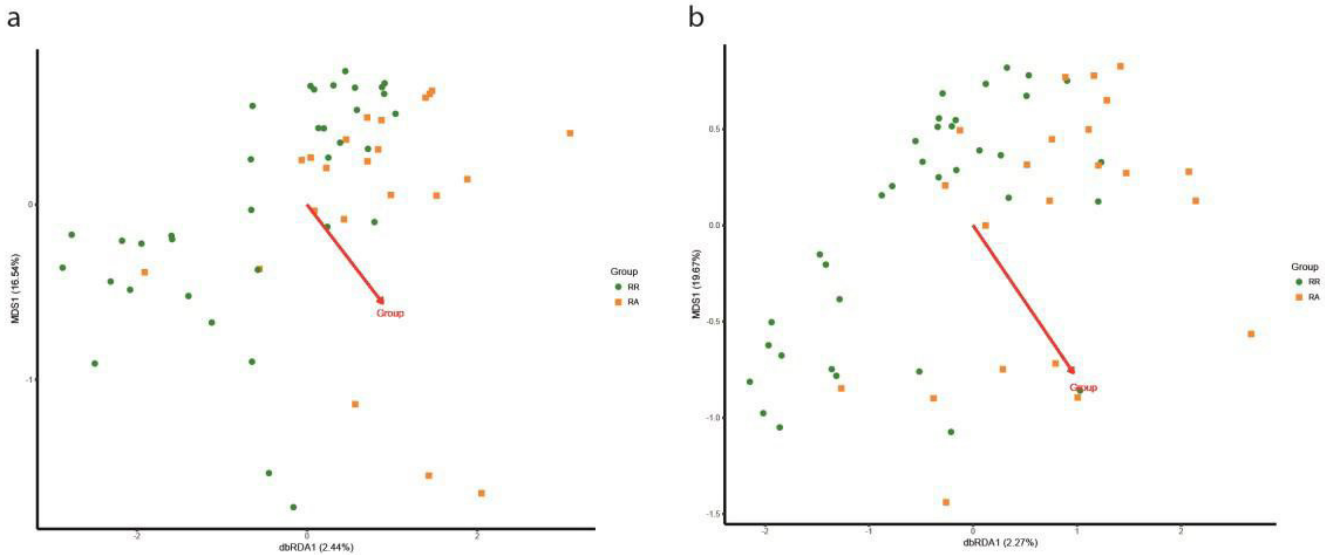


Supplementary Figure 4 Linear model between within-subject weighted UniFrac between two subsequent sampling time-points (T1-T2) and Log2 fold change of (a) faecal calprotectin; (b) C-Reactive Protein (CRP) in CD patients groups.



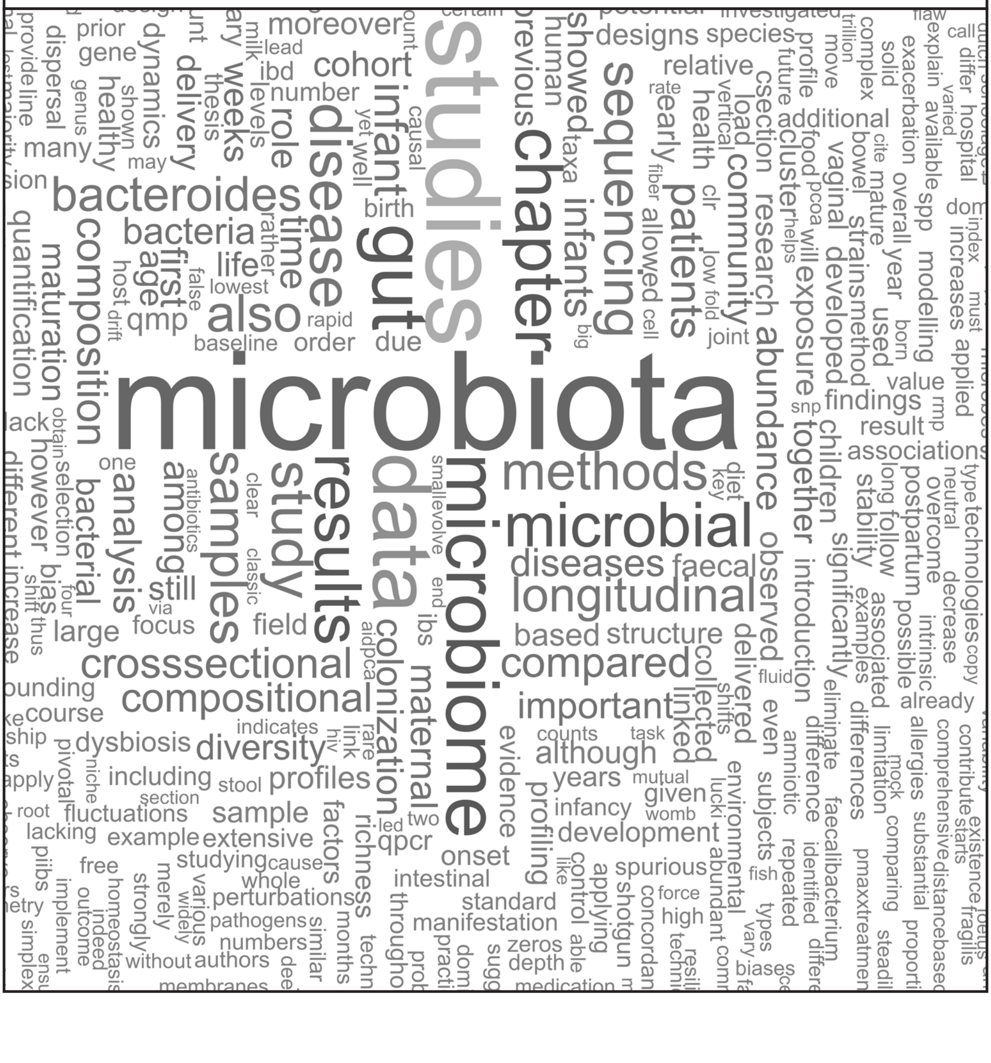
Supplementary Figure 5 Within-subject weighted UniFrac distance between two subsequent sampling time-points (T1-T2) among: (a) CD patients with different disease localization and healthy subjects, and (b) CD patients that did or did not received abdominal surgery.

Supplementary Figure 6 Distance-based redundancy analysis (dbRDA) based on weighted UniFrac distances using patient group as explanatory variable at (a) baseline (T1) and (b) at second time point (T2). Significance was tested using PERMANOVA resulting in p=0.12 and p=0.15 for T1 and T2 respectively.



Supplementary Figure 7 Distance-based redundancy analysis (dbRDA) based on weighted UniFrac distances using patient group as explanatory variable and partially filtering out Bacteroides:Prevotella ratio, age, gender, and medications use at (a) baseline (T1) and (b) at second time point (T2). Significance was tested using PERMANOVA resulting in p=0.12 and p=0.17 for T1 and T2 respectivel

Chapter 7


microbiota
studies
microbiome
data
results
gut
chapter
microbial
methods
disease
sequencing
patients
samples
study
bacteroides
longitudinal
crosssectional
compositional
composition
colonization
maternal
infant
diseases
compared
important
analysis
bacteria
diversity
development
although
profiles
also
cohort
moreover
designs
species
infants
community
research
abundance
exposure
together
vaginal
children
faecal
observed
delivered
introduction
structure
based
findings
associations
dysbiosis
fluctuations
including
sample
extensive
profiling
intestinal
onset
evidence

General Discussion & Summary

General discussion & Summary

The human intestinal tract is a unique setting that supplies a nutrient-rich environment for its complex microbial community that counts up to 100 trillion of microbes. All these microbes, including bacteria, archaea, viruses and eukarya, have co-evolved with the host [1]. This mutual relationship provides the host with benefits such as metabolic balance[2-5], processing of nutrients (including fiber digestion), vitamin synthesis, colonization resistance against invading pathogens[6, 7] and maturation and homeostasis of the gastrointestinal lymphoid tissues[8]. The initial colonization of the human intestinal tract starts at birth with the rupture of the amniotic membranes and subsequent passage through the birth canal [9, 10]. Subsequently, microbial populations evolve as the host matures and the diet changes. The infant gut microbiota is very unstable, showing big fluctuations in composition during the first 2.5 years of life[11, 12]. Given the strong dynamics in the microbiota during early infancy and its strong effect on the maturation of the host's immune system, a comprehensive insight into the processes that shape the infant microbiota is of particular importance. At around school-age the microbiota stabilizes [13] and resembles the mature adult composition. Once matured, the gut microbiota has been shown to be relatively resilient [14-17]. Nevertheless, it can still undergo dramatic compositional shifts, a condition known as dysbiosis, due to stressors like profound changes in diet, antibiotics use or diseases.

The past years have been the golden age for microbiota research with technological improvements leading to a rapid expansion of knowledge on the ecological dynamics of gut microbiota and an exponential increase in large-scale cohort studies and publications [18]. Nevertheless, the investigation of the gut microbiota and its role in human health is still young and in need of more fine-tuned studies. The vast majority of microbiota research is still based upon cross-sectional studies that can only partially explain the role of the gut microbiota in health and disease.

Studying the microbiome using a longitudinal design is pivotal in order to detect fluctuations of the microbial community but more important to find the relationships between bacteria and external factors and move from cross-sectional to temporal associations.

Moreover, to improve microbiome studies it is also key to better account for (or eliminate) the intrinsic compositional structure of next-generation sequencing data.

To this end, my thesis aimed to implement proper mathematical approaches prior to the data analysis of longitudinal data, to ensure more reliable results that can provide a deeper understanding on the assembly and maturation of the complex gut microbiota in humans and on the association between microbial perturbations with the development or progression of diseases.

The importance of longitudinal studies on the human microbiome

Since the introduction of next-generation sequencing techniques, the microbiome research field has revolutionised, and the microbiome has been studied in association to a plethora of diseases and determinants. However, for a long time the majority of these studies were cross-sectional in nature, comparing the microbiota composition of individuals exposed to a certain determinant or suffering from a disease to that of individuals without such an exposure or disease.

Such cross-sectional studies, however, cannot capture the dynamics in microbial

ecosystems. In particular, shifts that might occur as a result of the introduction of a specific environmental exposure, dietary changes or development or progression of diseases might easily be overlooked. Shifts might include temporary blooms of specific species followed by stabilization of the microbiome into the original or an alternative state [19, 20]. More importantly, cross-sectional studies are prone to selection bias, confounding and reverse causation (i.e., microbiota perturbations are a consequence rather than a cause of a disease). Many examples exist of bias in cross-sectional microbiome studies with the confounding effect of antidiabetic medication as a classic example[21].

In order to advance microbiome research, we should therefore move towards longitudinal study designs in which the outcome of interest (e.g., disease or disease exacerbation) has not yet manifested in any of the participants at baseline. Prospective cohort studies have the potential to provide the strongest scientific evidence of all types of observational study designs. In addition, as prospective studies examine changes in microbial composition over time in association to a certain exposure or the manifestation of disease (exacerbation), each individual serve as its own control. This significantly reduces the number of potential confounding factors that could lead to either spurious or undetected associations.

For those reasons all the studies presented in this thesis have a longitudinal design. In **chapter 3**. we collected faecal samples from 98 infants repeatedly at 1-2, 4 and 8 weeks, as well as 4, 5, 6, 9, 11 and 14 months of age. In **chapter 4**, 1453 stool samples were collected from 440 children, at 5, 13, 21 and 31 weeks of age and once again at school-age (6–11 years). In **chapter 5**, we studied the dynamics and resilience of the microbiota among 106 travellers that experienced a bout of diarrhoea and subsequently did or did not develop post-infectious Irritable Bowel Syndrome. In **chapter 6**, we examined the stability of the microbiota composition in 15 healthy control subjects and 57 patients with Crohn's disease (CD). As 22 of the CD patients developed an exacerbation during the course of the follow-up, we could examine whether the microbiota stability was associated with the disease course.

In all studies multiple stool samples along with extensive metadata were collected enabling us to move from mere cross-sectional to temporal associations.

Removing the compositionality of microbiome sequencing data using quantitative methods

Next to failing to acknowledge limitations of cross-sectional designs, another common flaw observed in microbiome studies is to not take into adequate consideration the compositional nature of microbiome sequencing data [22]. One way to overcome the compositional nature of microbiome sequencing data is to make the data quantitative again. In **chapter 2**, we investigated the use of various approaches to quantitatively profile the microbiota, as opposed to the traditional relative microbiome profiling (RMP), to overcome the compositional structure of sequencing data. Next to the Quantitative Microbial Profiling using flow cytometry-based microbial load (QMP) as introduced by Vandeputte *et al.* 2017 [23], we additionally combined Propidium Monoazide pre-treatment with flow cytometry-based cell counting in order to profile only intact cells (QMP-PMA), and also performed Quantitative Microbial Profiling using qPCR to determine the microbial load (QMP-qPCR). Overall, our results suggested that QMP could be a promising and elegant approach to overcome the compositional structure of microbiome data

but is still far from perfect as the use of cell flow cytometry can introduce additional biases. Moreover, this technique is still too laborious to apply in large cohort studies. We therefore investigated the possibility to use different, more high-throughput, methods to quantify the microbial load of the samples such as qPCR.

Our findings confirmed the previous observation [23] that absolute abundance profiles differ significantly from those generated by relative approaches. When comparing RMP to QMP, sample rank order concordance within the 15 most abundant genera varied widely with the highest concordance observed for *Fusicatenibacter* and the lowest concordance for *Blautia*. The differences among relative and quantitative methods extended also to enterotypes. When using DMM clustering, we identified two enterotypes enriched in *Bacteroides* or *Prevotella* with a significant difference in microbial load between these enterotypes. However, this difference was absent when the microbial loads were determined using qPCR.

When moving the focus on the different quantitative profile methods, we found that technical sources of variability may introduce additional bias depending on the quantification method being used. Generating quantitative microbiome profiles revealed that profiles obtained after PMAxx-treatment remained highly similar to the standard QMP profiles, although the observed genus richness slightly decreased upon PMAxx-treatment. This indicates that free extracellular DNA does not significantly bias the traditional flow-cytometry-based QMP method, although it cannot be deduced whether the existing dissimilarities between QMP-PMA and QMP microbial profiles are due to the elimination of free extracellular DNA or merely due to the introduction of additional technical variation during sample handling.

Previous studies have advocated qPCR as a more suitable and accessible alternative for microbial quantification compared to Flow-cytometry based quantification, although direct comparisons between both methods were lacking [24]. Our analysis demonstrated that quantification of bacterial load by qPCR results in highly divergent microbiome profiles as well as a strong decrease in the observed genus richness when compared to standard QMP or QMP-PMA methods. We ruled out that these deviant QMP-qPCR based profiles were the result of a lack of precision or sensitivity of qPCR as we proved that quantification of microbial load based upon Droplet Digital PCR (ddPCR) correlated strongly with qPCR-based quantification. Together these results prove that qPCR-based quantification might not be an adequate approach for quantitative microbiome profiling compared to flow-cytometry based quantification. Although 16S rRNA gene copy-number correction to account for variable in copy-numbers between bacterial taxa was applied in our study, the added value of this approach has recently been questioned as gene-copy number normalization even failed to improve the classification of 16S rRNA sequenced simple mock communities [25]. Indeed, low predictive accuracy and substantial disagreement has been observed between gene-copy number prediction tools [26]. As gene copy-number normalization is also applied in the traditional flow-cytometry-based QMP, a more comprehensive catalogue of copy numbers or other methods to account for variance are urgently needed.

This implies that there is currently no high-throughput and accessible laboratory method available to eliminate the compositional problem of microbiome sequencing data at its root. Therefore, to overcome the problem while still being able to investigate the role of the intestinal microbiota in human health, we decided to apply a mathemat-

ical approach to remove the compositional component from the data and thus allowing us to apply standard analysis techniques.

Removing the compositionality and sparseness using mathematical methods to ensure more reliable results and allow the implementation of new methods

As mentioned before**,** compositional data belongs to the mathematical space called the Simplex. As a consequence, when an element in a vector of compositional data increases, all the other elements together must decrease as all elements sum up to a constrained value. A mathematical solution to this problem was presented by Aitchison[27] who demonstrated how to project the compositional data from the Simplex to the Euclidean space by taking the logarithm of the ratio between each value and the geometric mean of the values. Applying this approach, called centred log-ratio (*clr*) transformation, to microbiome data is challenging because of the intrinsic sparseness of the data. Therefore, it is important to remove the zeros in the data without introducing biases. For example, a common practice is to add a small pseudo counts to the zeros in order to make the computation of the *clr* possible. But as shown in many studies [28-31] , the choice of the exact value of the pseudo count can have a large impact on the results. We therefore decided to model the zeros using the Dirichlet distribution[32], prior the *clr* transformation.

In **chapter 3**, the implementation of the Aitchison transformation, allowed us to obtain more reliable results on the effect that various ecological principles, including dispersal (limitation), neutral processes and environmental filtering contribute to the assembly of microbial communities during early infancy. Frequently used distance-based ordination methods such as PCoA and dbRDA rely on metrics such as the Bray-Curtis distance or the Jensen-Shannon divergence and are therefore more affected by the sequencing depth than by the actual microbial composition of samples. This was extensively illustrated by McMurdie and Holmes [33] by benchmarking the performance of commonly used distances and dissimilarities. The authors simulated publicly available data to introduce variations in sequencing depth and subsequently tested the performance of clustering methods and differential abundance testing. The results showed that a decrease in sequencing depth led to a poorer performance of the clustering methods and an increase in the false discovery rate for the differential abundance testing [33]. Moreover, distance-based ordination methods based on relative abundance data will mostly discriminate samples based on the most dominant bacteria rather than the most variable ones. A clear demonstration of this mechanism was given by Gorvitovskaia *et al.* [34] who demonstrated that the relative abundance of *Bacteroides* and *Prevotella,* rather than the underlying microbial community structures, is driving the separation of samples in ordination plots. To circumvent these types of bias, we applied *clr* transformation on microbial count data which allowed us to use PCA rather than PCoA. Besides adequately addressing the compositionality of the data, PCA on Euclidean distances also allows to identify and visually represent the contribution of individual taxa in the variations in overall microbial community structure between samples.

When this approach could not be used, we applied analysis methods that were developed specifically for compositional data.

The approaches to overcome the compositional nature of microbiome data applied

in this thesis have been carefully selected to address the research questions. It is important to stress that different transformations may be preferred depending on the analytic modelling tool of choice. The application of the Aitchison transformation is recommended when performing a PCA, but different data preprocessing steps might be preferred when performing regression analysis, network analysis or machine learning techniques (e.g., Random Forest).

The establishment and maturation of the gut microbiota: factors shaping the process

During the past few years, a lively academic controversy has emerged on the existence of prenatal microbial communities and *in utero* colonization of the foetus[35]. Although several studies showed molecular microbial profiles and limited numbers of viable microbes in placental tissue, amniotic fluid and foetal meconium[36, 37], current evidence favours contamination during sample collection and handling as the most likely source of these microbial signals[35, 38, 39]. The existence of interacting microbial communities in the womb that facilitate in utero colonization is thus highly unlikely, which makes rupture of the amniotic membranes the starting point of microbial colonization. The richness and diversity of the gut right after birth is extremely low when compared to adults, but as demonstrated in **chapters 3 and 4** gradually increases throughout and over the first year of life. In **chapter 3** the median Shannon index of children aged 1-2 weeks was 1.77 compared to a median diversity of 4.04 in the maternal faecal samples collected at the same time. This difference *per se* might not seems particularly large but as those values are logarithmic an increase of one unit is associated with substantial differences in microbial diversity. Similar results were reported in **chapter 4** in which children were sampled starting from 5 weeks of age. The assembly of the gut microbiota then steadily proceeded with gradual increases in microbial diversity up to the age of 6-8 months postpartum where a drastic increase in richness and diversity occurred [**Chapter 3 and 4**]. Unsurprisingly, this time window coincides with the start of weaning, pointing towards the introduction of solid foods or the cessation of breast feeding as a driving force of significant increases in the infant microbial richness and diversity.

Our results in **chapter 3** showed that longer breastfeeding duration was most strongly linked to a delayed increase in microbial diversity. This indicates that, regardless of the introduction of other food substrates, the availability of human milk maintains a simple microbial community dominated by only a few genera.

When examining the overall microbial community structure in both **chapter 3 and 4**, we identified 6 microbial clusters. In **chapter 3**, the majority of samples collected at 1-2 weeks of age grouped into 3 clusters, dominated by *Bifidobacterium* (Actinobacteria) for cluster 1, *Escherichia (*Proteobacteri*a)* and *Streptococcus (*Firmicutes) for cluster 4 and *Bacteroides* (Bacteroidetes) for cluster 6. In **chapter 4**, for samples collected at 5 weeks postpartum, only 2 cluster were prevailing and were characterized by *Escherichia* for cluster 1, while cluster 3 was characterized by *Streptococcus* and *Veillonella.* The early dominance of Bacteroides in some of the infants in the LucKi cohort [**Chapter 3**] could be linked to a vaginal delivery as has also been shown in several previous studies [40, 41]. Interestingly Yassour *et al.* [40] reported that children born via C-section also harboured *Bacteroides* strains during the first week of life but lost these bacteria

upon which streptococci became dominant. Nayfach *et al.* [42] further substantiated these findings by tracking the vertical transmission of bacterial strains from mother to infant based upon the identification of rare Single Nucleotide Polymorphism (SNP) characteristic of bacterial strains in shotgun metagenomic data. In line with our results in **chapter 3**, the authors show that *Bacteroides* and *Parabacteroides* are among the most vertically transmitted bacteria in case of vaginal delivery while transmission of these bacteria among C-section delivered infants was lacking. Further evidence of vertical transmission of maternal gut microbiota comes from the study of Korpela *et al.* [43]. Using publicly available shotgun metagenomic sequencing the authors used rare SNPs that were not shared with samples from any non-family members as marker to track the transmission of bacterial strains. Their results not only showed that in the 87% of vaginal delivered infant the vertical transmission of gut microbiota involved mainly bacteria of the classes Actinobacteria and Bacteroidia, but also that the colonization of maternal strains was more persistent. Altogether, our findings that vaginal delivered infants shared a significantly higher proportion of *Bacteroides* ASVs with faecal microbiota of their mothers as compared to C-section delivered infants [**Chapter 3**] are in strong agreement with these previous studies. Together these results suggest that the neonatal microbial composition is more strongly influenced by exposure to the maternal gut microbiota than by the passage through the birth canal.

Indeed, causal evidence for the relationship between exposure to maternal faeces and persistence of *Bacteroides* has recently been demonstrated in a proof-of-concept study on maternal faecal microbiota transplantation in Caesarean-section delivered infants [44]. The relative Bacteroidales abundance observed in the first week of life decreased more than 100-fold by the age of 3 weeks postpartum in C-section delivered infants. In contrast, the abundance of Bacteroidales (mainly *Bacteroides*) increased over the first weeks of life in C-section born infants that received maternal FMT. This again confirms engraftment of maternal faecal *Bacteroides* strains. Moreover, together with our findings, these results further question the benefits of bacterial baptism or vaginal seeding approaches [45].It remains however unclear why we could observe a dominance of *Bacteroides* in some of the children in the Dutch Lucki cohort as early as 1-2 weeks postpartum, whereas a similar *Bacteroides* dominated group of neonates was lacking in the German PAPS study. This is even more surprising as the proportion of C-section delivered infants was even lower in the latter cohort (6.6 vs. 14.3%). Although technical variation due to differences in DNA isolation protocols and sample collection methods could be partly responsible for this difference, national differences in childbirth practices likely also play a role. In contrast to Germany, where inpatient hospital delivery is standard practice and mothers and their newborns generally remain hospitalized for 3-4 days after delivery, homebirth and outpatient hospital delivery are common practice in the Netherlands. Further research on the potential impact of delivery practices, bowel movements during labor and in- vs. outpatient childbirth on the persistent colonization of *Bacteroides* is currently needed.

More research on priority effects, a well-known ecological concept, might also shed additional light on the dynamics of the infant gut colonization. The dispersion of bacteria from the “regional” pool is known to be affected by exposure to maternal gut microbiota, environment of delivery and contact with other persons (e.g., hospital staff). Altogether those elements can affect the pioneer species colonizing the infant gut.

Moreover, those early colonizers might change the environmental condition of the infant gut, affecting the chances for other bacteria to colonize the infant gut. An example of this process is (nontoxigenic) *Bacteroides fragilis* that upon colonization the niche is resistant to colonization by the same, but not different, species [46].

Altogether these results suggest that the early colonization of the gut is a chaotic but not stochastic phase characterized by a low diversity and a high interindividual variability, likely driven by dispersal limitation and environmental selection. In particular, vaginal delivery sets up a unique initial microbial community by exposing the child to maternal faeces during delivery. This was further supported by the results of our neutral community modelling [**Chapter 3**], in which *Bacteroides fragilis* resulted under negative selection in case of caesarean section.

Once the gut is colonized the maturation of the gut microbiota proceeds gradually and steadily towards adulthood with a major influence of dietary factors [47]. For example, the origin and trajectory of the *Bifidobacterium*-dominated cluster 1 (**Chapter 3**) is likely the result of a combination of different dispersal mechanisms (e.g., seeding by the maternal milk microbiome), as compared to the other clusters, and subsequently selection and drift driven by the ability of *Bifidobacterium* species to degrade HMOs derived from breast milk. As early as of 4 weeks postpartum and onwards, dietary factors showed to have a greater impact on the microbial community structure than perinatal determinants. In particular, results from **chapter 3 and 4** show that cessation of breastmilk relate more with changes in microbial composition than introduction of solid food. However, in contrast to previous studies with limited number of sampling time-points and lack of detailed dietary data[48], our data in **chapter 3** do demonstrate that also the type and complexity of solid foods is important for maturation of the infant gut microbiota. Infants with a more adult-like omnivore dietary pattern, characterized by the consumption of rice, pasta, fish, and meat products, at the age of 9 months had the most mature gut microbiota composition with highest levels of *Faecalibacterium spp.* and lowest levels of *Enterococcus spp*. and *Staphylococcus* spp. This elegantly demonstrates the importance of longitudinal study designs and sufficient numbers of repeated samples when studying the effect of diet and other determinants on the highly dynamic infant microbiota.

Next to the impact of birth mode and diet, also dispersal from other individuals and companion animals appeared to affect the infant gut microbiota [**Chapter 3 and 4**]. This is in line with several previous studies [49-52], although the specific effects of sibship size and pet exposure differ between studies. On this regard, future studies should focus on more refined microbial profiling on a strain level (e.g., by whole metagenome shotgun sequencing) to track the microbial dispersal from siblings and pets to newborns. Nonetheless, the goal to fully explain the inter-individual microbiota variations could be unachievable as indicated by the increasing evidence on the central role of stochastic events on the assembly and maturation of the gut microbiota.

Longitudinal methods to link disturbances in the maturation of the gut microbiota to the onset of allergic manifestations

Many epidemiological studies [53-58] suggest that the infant gut microbiota plays an important role in manifestation of allergic diseases and asthma, although the results vary considerably between studies. The lack or early samples and different ages of sam-

ple collection, different microbial profile methods and insufficient control for potential confounders might contribute to the heterogeneity between study results[55-59]. Moreover, cross-sectional case-control studies cannot discriminate whether the change in microbiota composition is the cause or the result of the allergic manifestation. In this context and given the complex dynamics of gut microbiota assembly and maturation, longitudinal studies are important to allow analysis on the overall development of the infant gut microbiota. In the study presented in **chapter 4**, the collection of an adequate number of repeated samples during the first year of life and the follow-up up to school age together with recording of extensive metadata allowed us to implement, for the first time in a microbiome study, a Joint Modelling while correcting for potential confounders. The joint modelling combines how an exposure variable changes over time (longitudinal modelling of a risk factor) with the time at which an outcome event occurs (survival analysis). The results of the joint modelling showed that throughout the first year of life a higher microbial richness was associated with a lower risk and delayed onset of atopic dermatitis.

The collection of repeated samples during such an important time window also allowed us to define the stage of maturity of the infant gut microbiota using a machine learning approach. We subsequently examined to what extent microbiota maturation was linked to atopic dermatitis. Our results [**Chapter 4**} showed that children that developed atopic dermatitis and allergic manifestation have a more mature microbiota in the earliest time points when compared to infants that remain free from allergies. After the first half year of life this trend appears to be reverted with a less mature microbiota among infants who developed allergic symptoms.

Moreover, the longitudinal design of the study in **chapter 4** allowed us to identify microbial taxa that were differentially abundant throughout the infancy among infants who did or did not developed allergic manifestations. In line with previous studies [57, 60], the abundance of *Lachnobacterium* and *Faecalibacterium* was decreased among children who subsequently developed atopic dermatitis. The extensive duration of this decreased abundance throughout infancy suggests a protective role of those bacterial genera in preventing the development of atopic dermatitis. Altogether, we showed that applying various multivariable longitudinal models can provide additional insight into the temporal associations between the dynamic infant gut microbiota and the onset of non-communicable diseases, such as allergies.

By applying such methods, we can separate bacterial taxa likely involved in the pathophysiology of allergies from bacteria that merely shift as a consequence from the disease, its treatment or disease-related dietary restrictions. This greatly helps to unveil those bacteria, such as *Faecalibacterium*, that are suitable candidates for next-generation probiotics in the primary prevention of allergic diseases.

Gut microbiota in adult human health: the role dysbiosis in the onset and disease course of gastrointestinal disorders

The last decades have been characterized by numerous studies that underscored the role of the human gut microbiota in health and homeostasis. The microbiome is involved in metabolism regulation, immune maturation and response and protection against pathogens [61-64] just to cite a few examples. Given the involvement of the microbiome in all these pivotal functions, it is not surprising that microbial perturbations

(dysbiosis) have been linked to the onset and course of numerous diseases. Obesity[65], type II diabetes[66], activation of HIV[67], IBS[68], IBD[69], and atopy[53-55, 57, 59] are few examples of a long and growing list of diseases and disorders that have been linked to microbial perturbations. The studies presented in **Chapter 5 and 6** investigated the temporal associations between dysbiosis and Inflammatory Bowel Disease and Irritable Bowel Syndrome, respectively.

Studying the microbiota prior to disease onset is often impossible for diseases that manifest in adulthood as it would require extremely large sample sizes and extensive follow-up. Therefore, no study had previously been able to study the microbiota composition among individuals that subsequently developed IBS. In **Chapter 5** we describe the first study that has examined the baseline microbiota as well as its stability and resilience upon a bout of diarrhoea among intercontinental travellers that subsequently did or did not develop post-infectious IBS. We demonstrated that, as compared to control subjects, the microbial diversity and community structure were already significantly different among PI-IBS cases prior to disease onset. Longitudinal analyses moreover revealed differentially abundant genera, including increased *Bacteroides* levels, in cases as compared to controls from baseline onwards. These results clearly show for the first time that the microbiota diversity and composition can predispose to the development of PI-IBS after an episode of gastroenteritis. We can therefore eliminate the possibility that the microbiota alterations are merely the result of avoidance of specific food items that IBS patients link to worsening of symptoms or that microbiota alterations are epiphenomena linked to an unknown trigger of IBS. As these alternative explanations could not be completely ruled out in all previous cross-sectional studies, our study greatly enhances the evidence for a role of the microbiota in the pathogenesis of (PI-) IBS.

Despite the large literature on the role of the microbiota in IBD, the results on the role of the microbiota in disease flares are often inconsistent and sometimes even contradictory. Many cross-sectional studies compared patients the microbiota of patients with active Crohn 's Disease (CD) to that of patients in remission. The microbiota of CD patients with an active disease flare has been linked to increased levels of Enterobacteriaceae [70, 71] and *Bacteroides spp.* [70, 72], a reduction of *F. prausnitzii* [73-75] and *Clostridium coccoides* group [73, 76] in some but far from all studies. These inconsistencies may in large part be due the inter-individual variation in microbiota composition, confounding factors, and the heterogeneous nature of CD, that can only be partly accounted for in cross-sectional studies. Longitudinal studies are therefore important to shed further light on the causal relationship between dysbiosis and disease onset and course. The longitudinal design of the study in **Chapter 6** allowed us to compare the stability of the intestinal microbiota in healthy subjects as compared to CD patients. We showed that CD patients have a less stable faecal microbiota as compared to healthy subjects. This is in line with a study in which faecal samples were sequentially collected from patients with ulcerative colitis (UC) remaining in remission and with stable medication during a year of follow-up. Only one-third of the dominant taxa were persistently detected among UC patients in this study, while healthy individuals showed a remarkable microbiota stability [77] .

However, contrasting previous cross-sectional studies [70, 75], by profiling multiple samples from CD patients with changing or stable disease activity we did not observe

the stability of the gut microbiota to be associated with disease course. Two other longitudinal studies also did not observe a correlation between microbial composition and active inflammation [78, 79]. Halfvarson and colleagues profiled the microbiota in a cohort of 128 IBD patients, including 49 CD patients, and observed a distinct microbiota as compared to healthy controls but no association with calprotectin levels, as inflammatory marker, in the patient group [78, 79]. In a large cohort of over 2,000 a cohort of non-IBD and IBD faecal samples from four countries also observed a clear dysbiosis in CD patients but no link with disease activity [78, 79]. Together this sharp contrast between case-control studies and longitudinal cohort studies when examining the link between the microbiota and IBD disease activity highlights the risk of identifying spurious findings and bias in cross-sectional studies.

Future perspectives

The results presented in this thesis show how much the microbiome field can benefit from longitudinal study designs and appropriate statistical frameworks to answer biological questions. Despite the improvements the microbiome field witnessed during the past years, there is still a profound lack of analysis methods tailored specifically to handle microbiome data. Many tools have been borrowed from the ecology field and adjusted for microbiome studies, however those tools are not meant to handle the high dimensionality, sparseness and compositionality typical for the data generated in microbiome studies [33].

This lack of analysis methods results in the inability to answer some important questions on the microbe-microbe and microbe-host interactions and the mechanisms linking dysbiosis to diseases. Statistical challenges related to sparseness, high dimensionality and complexity are becoming even more evident when applying methods with higher taxonomic resolution such as whole metagenome sequencing or when integrating multiple -omics data. A possible focus of future efforts is the microbial profiling technology itself. High-throughput and 16S rRNA gene amplicon sequencing technologies have set the path to revolutionize the microbiome field to what we know it today. Nonetheless as the knowledge and experience in this field grow, the awareness of its intrinsic limitation grows as well. It is time to re-think sequencing technologies to create data that are not intrinsically compositional. Many efforts have been made already in this direction, one also presented in this thesis, but this task should ultimately be accomplished together with the gene sequencing industry.

On the subject of data analysis, a promising method that is becoming more and more present in microbiome studies is represented by network analysis. This method holds the potential to discover keystone species and key players in the food chain allowing a deeper understanding of what defines a healthy microbiota. Unfortunately, the majority of the studies do not account for the possible spurious correlations generated by correlation indexes, i.e., Spearman's correlation coefficient, commonly used to generate the interaction networks, which can lead to visually impressive interaction networks that are hampered by a limited validity.

Next to studying bacterial interactions, the integration of metabolomic data with metagenomic data such as microbial abundance and gene pathway abundance, as well as extensive (clinical) metadata, is also pivotal to reveal causal pathways and leads for disease prevention. Further development of statistical frameworks for multi-omics

data integration should therefore be an important research focus. It is also important to stress how much the microbiome field is in need of more fundamental research on the gut microbiota. To define what a healthy microbiota is and what the inter-bacterial dynamics are. Because only answering to these questions it will be possible to make successful interventions to improve patient's health.

With respect to whole metagenome shotgun sequencing, technologies are improving at an unprecedented pace, in particular with respect to the improved quality of long-read sequencing technologies such as nanopore sequencing[80]. Longer reads will significantly simplify metagenomic data processing and assembly, yet almost all bioinformatic pipelines are still focused on short-read sequencing data. Given the pace at which long-read sequencing is developing, time is pressing to also shift the focus from a bioinformatics point of view.

Last, but not least, establishing gold standards or guidelines for microbiome data analysis and reporting are crucially important to enable comparison of results between studies and facilitate more rapid sample analysis and throughput. This call was already made more than 10 years ago [81] and is to date still largely unaddressed.

Together with tailored analytical methods and appropriate standards, future microbiome studies will benefit of bigger cohorts, more frequent sampling, and longer follow up to unravel the long- and short-term fluctuations of the gut microbial communities and how those fluctuations can impact human health later in life.

In conclusion the studies presented in this thesis demonstrate the importance of longitudinal study designs and appropriate analytical methods to study the microbiota in health and disease. The application of such analytical methods and longitudinal study designs substantially decreases the possibility of false positive findings due to spurious correlations, confounding factors, and reverse causalities. Together these results aid in a stronger foundation for effective microbiota-based intervention strategies to prevent or reduce the burden of non-communicable diseases

References

1. Rawls, J.F., et al., Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell, 2006. 127(2): p. 423-433.
2. Backhed, F., et al., The gut microbiota as an environmental factor that regulates fat storage. Proceedings of the National Academy of Sciences of the United States of America, 2004. 101(44): p. 15718-15723.
3. Duncan, S.H., P. Louis, and H.J. Flint, Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. Applied and Environmental Microbiology, 2004. 70(10): p. 5810-5817.
4. Chen, F.D., W. Wu, and Y.Z. Cong, Short chain fatty acids regulation of neutrophil production of IL-10. Journal of Immunology, 2016. 196.
5. Sartor, R.B., Microbial influences in inflammatory bowel diseases. Gastroenterology, 2008. 134(2): p. 577-594.
6. Hooper, L.V., OPINION Do symbiotic bacteria subvert host immunity? Nature Reviews Microbiology, 2009. 7(5): p. 367-374.
7. Othman, M., R. Aguero, and H.C. Lin, Alterations in intestinal microbial flora and human disease. Current Opinion in Gastroenterology, 2008. 24(1): p. 11-16.
8. Takahashi, K., Interaction between the Intestinal Immune System and Commensal Bacteria and Its Effect on the Regulation of Allergic Reactions. Bioscience Biotechnology and Biochemistry, 2010. 74(4): p. 691-695.
9. Korpela, K., et al., Selective maternal seeding and environment shape the human gut microbiome. Genome Research, 2018. 28(4): p. 561-568.
10. Wampach, L., et al., Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. Nature Communications, 2018. 9.
11. Zhuang, L., et al., Intestinal Microbiota in Early Life and Its Implications on Childhood Health. Genomics Proteomics Bioinformatics, 2019. 17(1): p. 13-25.

12. Stewart, C.J., et al., Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature, 2018. 562(7728): p. 583-588.
13. Zhong, H.Z., et al., Impact of early events and lifestyle on the gut microbiota and metabolic phenotypes in young school-age children. Microbiome, 2019. 7.
14. Costello, E.K., et al., The application of ecological theory toward an understanding of the human microbiome. Science, 2012. 336(6086): p. 1255-62.
15. Gilbert, J.A. and S.V. Lynch, Community ecology as a framework for human microbiome research. Nat Med, 2019. 25(6): p. 884-889.
16. Donaldson, G.P., S.M. Lee, and S.K. Mazmanian, Gut biogeography of the bacterial microbiota. Nat Rev Microbiol, 2016. 14(1): p. 20-32.
17. Faust, K., et al., Metagenomics meets time series analysis: unraveling microbial community dynamics. Current Opinion in Microbiology, 2015. 25: p. 56-66.
18. Gilbert, J.A., et al., Current understanding of the human microbiome. Nature Medicine, 2018. 24(4): p. 392-400.
19. Walker, A.W., et al., Dominant and diet-responsive groups of bacteria within the human colonic microbiota. The ISME journal, 2011. 5(2): p. 220.
20. Dethlefsen, L. and D.A. Relman, Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. Proc Natl Acad Sci U S A, 2011. 108 Suppl 1: p. 4554-61.
21. Forslund, K., et al., Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature, 2015. 528(7581): p. 262-266.
22. Lin, H. and S.D. Peddada, Analysis of microbial compositions: a review of normalization and differential abundance analysis. NPJ Biofilms Microbiomes, 2020. 6(1): p. 60.
23. Vandeputte, D., et al., Quantitative microbiome profiling links gut community variation to microbial load. Nature, 2017. 551(7681): p. 507-511.
24. Jian, C., et al., Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. PLoS One, 2020. 15(1): p. e0227285.
25. Starke, R., V.S. Pylro, and D.K. Morais, 16S rRNA Gene Copy Number Normalization Does Not Provide More Reliable Conclusions in Metataxonomic Surveys. Microb Ecol, 2021. 81(2): p. 535-539.
26. Louca, S., M. Doebeli, and L.W. Parfrey, Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome, 2018. 6(1): p. 41.
27. Aitchison, J., The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological), 1982. 44(2): p. 139-160.
28. Egozcue, J.J., et al., Isometric logratio transformations for compositional data analysis. Mathematical Geology, 2003. 35(3): p. 279-300.
29. Greenacre, M., Measuring Subcompositional Incoherence. Mathematical Geosciences, 2011. 43(6): p. 681-693.
30. Costea, P.I., et al., A fair comparison. Nat Methods, 2014. 11(4): p. 359.
31. Paulson, J.N., H.C. Bravo, and M. Pop, Reply to: "a fair comparison". Nat Methods, 2014. 11(4): p. 359-60.
32. Fernandes, A.D., et al., ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. Plos One, 2013. 8(7).
33. McMurdie, P.J. and S. Holmes, Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol, 2014. 10(4): p. e1003531.
34. Gorvitovskaia, A., S.P. Holmes, and S.M. Huse, Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome, 2016. 4: p. 15.
35. Blaser, M.J., et al., Lessons learned from the prenatal microbiome controversy. Microbiome, 2021. 9(1): p. 8.
36. Aagaard, K., et al., The placenta harbors a unique microbiome. Sci Transl Med, 2014. 6(237): p. 237ra65.
37. Rackaityte, E., et al., Viable bacterial colonization is highly limited in the human intestine in utero. Nat Med, 2020. 26(4): p. 599-607.
38. de Goffau, M.C., et al., Batch effects account for the main findings of an in utero human intestinal bacterial colonization study. Microbiome, 2021. 9(1): p. 6.
39. Hornef, M. and J. Penders, Does a prenatal bacterial microbiota exist? Mucosal Immunol, 2017. 10(3): p. 598-601.
40. Mitchell, C.M., et al., Delivery Mode Affects Stability of Early Infant Gut Microbiota. Cell Rep Med, 2020. 1(9): p. 100156.
41. Yassour, M., et al., Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med, 2016. 8(343): p. 343ra81.
42. Nayfach, S., et al., An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res, 2016. 26(11): p. 1612-1625.
43. Korpela, K., et al., Selective maternal seeding and environment shape the human gut microbiome. Genome Res, 2018. 28(4): p. 561-568.
44. Korpela, K., et al., Maternal Fecal Microbiota Transplantation in Cesarean-Born Infants Rapidly Restores Normal Gut Microbial Development: A Proof-of-Concept Study. Cell, 2020. 183(2): p. 324-334 e5.
45. Mueller, N.T., et al., Bacterial Baptism: Scientific, Medical, and Regulatory Issues Raised by Vaginal Seeding of C-Section-Born Babies. J Law Med Ethics, 2019. 47(4): p. 568-578.
46. Lee, S.M., et al., Bacterial colonization factors control specificity and stability of the gut microbiota. Nature, 2013. 501(7467): p. 426-9.
47. Yatsunenko, T., et al., Human gut microbiome viewed across age and geography. Nature, 2012. 486(7402):

p. 222-7.
48. Backhed, F., et al., Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe, 2015. 17(5): p. 690-703.
49. Azad, M.B., et al., Infant gut microbiota and the hygiene hypothesis of allergic disease: impact of household pets and siblings on microbiota composition and diversity. Allergy Asthma Clin Immunol, 2013. 9(1): p. 15.
50. Hasegawa, K., et al., Household siblings and nasal and fecal microbiota in infants. Pediatr Int, 2017. 59(4): p. 473-481.
51. Lane, A.A., et al., Household composition and the infant fecal microbiome: The INSPIRE study. Am J Phys Anthropol, 2019. 169(3): p. 526-539.
52. Martin, R., et al., Early-Life Events, Including Mode of Delivery and Type of Feeding, Siblings and Gender, Shape the Developing Gut Microbiota. PLoS One, 2016. 11(6): p. e0158498.
53. Azad, M.B., et al., Infant gut microbiota and food sensitization: associations in the first year of life. Clin Exp Allergy, 2015. 45(3): p. 632-43.
54. Dzidic, M., et al., Aberrant IgA responses to the gut microbiota during infancy precede asthma and allergy development. J Allergy Clin Immunol, 2017. 139(3): p. 1017-1025 e14.
55. Penders, J., et al., The role of the intestinal microbiota in the development of atopic disorders. Allergy, 2007. 62(11): p. 1223-36.
56. Simonyte Sjodin, K., et al., Temporal and long-term gut microbiota variation in allergic disease: A prospective study from infancy to school age. Allergy, 2019. 74(1): p. 176-185.
57. Stokholm, J., et al., Maturation of the gut microbiome and risk of asthma in childhood. Nat Commun, 2018. 9(1): p. 141.
58. Fujimura, K.E., et al., Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. Nat Med, 2016. 22(10): p. 1187-1191.
59. Zimmermann, P., et al., Association between the intestinal microbiota and allergic sensitization, eczema, and asthma: A systematic review. J Allergy Clin Immunol, 2019. 143(2): p. 467-485.
60. Arrieta, M.C., et al., Associations between infant fungal and bacterial dysbiosis and childhood atopic wheeze in a nonindustrialized setting. J Allergy Clin Immunol, 2018. 142(2): p. 424-434 e10.
61. O'Hara, A.M. and F. Shanahan, The gut flora as a forgotten organ. EMBO Rep, 2006. 7(7): p. 688-93.
62. Sekirov, I., et al., Gut microbiota in health and disease. Physiol Rev, 2010. 90(3): p. 859-904.
63. Vyas, U. and N. Ranganathan, Probiotics, prebiotics, and synbiotics: gut and beyond. Gastroenterol Res Pract, 2012. 2012: p. 872716.
64. Guarner, F. and J.R. Malagelada, Gut flora in health and disease. Lancet, 2003. 361(9356): p. 512-9.
65. Clarke, S.F., et al., The gut microbiota and its relationship to diet and obesity: new insights. Gut Microbes, 2012. 3(3): p. 186-202.
66. Qin, J., et al., A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 2012. 490(7418): p. 55-60.
67. Hofer, U. and R.F. Speck, Disturbance of the gut-associated lymphoid tissue is associated with disease progression in chronic HIV infection. Semin Immunopathol, 2009. 31(2): p. 257-66.
68. Ghoshal, U.C., et al., The gut microbiota and irritable bowel syndrome: friend or foe? Int J Inflam, 2012. 2012: p. 151085.
69. Li, Q., et al., Molecular-phylogenetic characterization of the microbiota in ulcerated and non-ulcerated regions in the patients with Crohn's disease. PLoS One, 2012. 7(4): p. e34939.
70. Kolho, K.L., et al., Fecal Microbiota in Pediatric Inflammatory Bowel Disease and Its Relation to Inflammation. Am J Gastroenterol, 2015. 110(6): p. 921-30.
71. Papa, E., et al., Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. PLoS One, 2012. 7(6): p. e39242.
72. Andoh, A., et al., Characterization of gut microbiota profiles by disease activity in patients with Crohn's disease using data mining analysis of terminal restriction fragment length polymorphisms. Biomed Rep, 2014. 2(3): p. 370-373.
73. Sokol, H., et al., Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflamm Bowel Dis, 2009. 15(8): p. 1183-9.
74. Swidsinski, A., et al., Active Crohn's disease and ulcerative colitis can be specifically diagnosed and monitored based on the biostructure of the fecal flora. Inflamm Bowel Dis, 2008. 14(2): p. 147-61.
75. Wang, W., et al., Increased proportions of Bifidobacterium and the Lactobacillus group and loss of butyrate-producing bacteria in inflammatory bowel disease. J Clin Microbiol, 2014. 52(2): p. 398-406.
76. Seksik, P., et al., Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. Gut, 2003. 52(2): p. 237-42.
77. Martinez, C., et al., Unstable composition of the fecal microbiota in ulcerative colitis during clinical remission. Am J Gastroenterol, 2008. 103(3): p. 643-8.
78. Halfvarson, J., et al., Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol, 2017. 2: p. 17004.
79. Pascal, V., et al., A microbial signature for Crohn's disease. Gut, 2017. 66(5): p. 813-822.
80. Maghini, D.G., et al., Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. Nat Protoc, 2021. 16(1): p. 458-471.
81. Chistoserdova, L., Recent progress and new challenges in metagenomics for biotechnology. Biotechnol Lett, 2010. 32(10): p. 1351-9.

Addendum

Impact paragraph
Curriculum vitae
Acknowledgements

Impact paragraph

In the last 10 years the human gut microbiota has drawn a lot of attention from the scientific community that through many research projects has discovered its role in human health such as in metabolism regulation, immune maturation, and protection against pathogen colonisation. Along with a deeper understanding of its physiological functions, scientists have linked changes in gut microbiota to many diseases. IBD, IBS, Celiac Disease, Type-1/-2 diabetes, atopic diseases, obesity and colorectal cancer are only few examples of a long list of diseases that together affect millions of people around the world and have all been associated to perturbations in the gut microbiota. Because of the vast societal impact of these diseases, the gut microbiota is the epicentre of a pharmaceutical "gold rush" worth 284.5M USD up to 2019. Beside its potential, the microbiome field is still very young and therefore tools specifically tailored for the unique challenges that its data analysis poses are only now becoming available. The shortage of analytical methods able to handle the compositionality, the sparseness and the high dimensionality of the data has resulted in the wide application of analysis techniques. These techniques are still inadequate for the intrinsic characteristics and dynamics of microbiome data. One example of this inadequacy is the current standard of using correlation indexes such as the Spearman's correlation. This coefficient is prone to spurious correlations if applied to compositional data. This issue, already addressed by Aitchison in 1982, remains essentially ignored. The use of inadequate analysis techniques is likely one of the reasons why the causal relationships as well as the mechanisms by which the gut microbiota affect human health remain largely unknown. In addition, many of the research papers published in this field to date are based upon cross-sectional study designs which can only partially explain the role of gut microbiota in health and disease.

The overall aim of this thesis was to provide more evidence on the benefit of the application of longitudinal designs and adequate analysis methods to elucidate the putative relationship between gut microbiota and health during early life as well as in children and adults.

In the first part of this thesis, we compared different methods to quantify the bacterial load in faecal samples as a way to deal with the compositionality of sequencing data. We however showed that the available quantification methods are either too laborious for high-throughput application in large population-based studies or too imprecise. These results are important as it shows the need to either develop other quantification methods or bioinformatic solutions to handle the compositionality of sequencing data. Better data analysis methods will therefore also have large societal impact on the results of microbiota studies and their role in health and disease.

In the second part of this thesis, we studied the impact of environmental and dietary factors, as well as stochastic factors, on the assembly and maturation of the gut microbiota in early life and how disruptions in these processes can contribute to the development of immune-mediated diseases. This time window in infancy is crucial, not only because it sets the conditions for microbial maturation, but also because the microbiota provides a stimulus for the adequate development of the gut and immune system. Our findings suggest that caesarean section delivery profoundly affects the colonisation pattern of the infant gut, mainly by limiting the transfer and expansion of maternal gut rather than vaginal microbiota. This might explain the lack of persistent effects of vag-

inal seeding, a procedure that is acquiring more and more popularity among mothers that deliver via caesarean section. Informing the general public and medical society on the detrimental impact of procedures such as caesarean section without a medical need might therefore have more impact than vaginal seeding.

Introducing a novel analytical method, joint modelling, to combine longitudinal microbiome data with survival analysis to study the time to disease onset, we demonstrated that alterations in the infant gut microbiota preceded the manifestation of atopic symptoms. Not only do these results add a new analysis technique to the toolbox of bioinformaticians to study longitudinal microbiome data, but they also provide insight into potential protective microorganisms in the prevention of atopic diseases. Faecalibacterium and Lachnobacterium were amongst the bacterial genera that were reduced among infants who subsequently developed allergic diseases and might therefore serve as candidates for potential next-generation probiotics to prevent these diseases. This is potentially very important for the health and wellbeing of future generations by creating an optimal starting condition of the gut microbiota leading to long term health effects.

In the third part of this thesis, we investigated the role of the gut microbiota in adults and its relation to the onset and course of two gastrointestinal diseases, Crohn's disease and Irritable Bowel Syndrome (IBS). Our findings show that while the microbiota of Crohn's disease patients differs from the microbiota of healthy individuals, the microbial community structure doesn't seem to play a role in disease exacerbations. We did however observe clear alterations in the gut microbiota of subjects who subsequently developed post-infectious IBS. To our knowledge this is the first time that gut microbiota alterations have been observed prior to the onset of active symptoms in IBS patients. Altogether, these findings substantially contribute to the causal role of the microbiota in the pathophysiology of IBS and provide further evidence that analysis of repeated samples over time can provide valuable knowledge to the field. The observed Bacteroides dominance in subjects susceptible to IBS development, moreover, provides strong leads for dietary intervention strategies in general to prevent functional bowel disorders.

To conclude, we demonstrate that robust analytical methods and adequate longitudinal study designs are essential to understand the role of microbiome alterations in health and disease and to subsequently develop strategies to modify or shape the gut microbiome to prevent or alleviate the disease burden caused by the global rise in non-communicable diseases.

About the author

Gianluca Galazzo was born in Siracusa (Italy) on the 24th of October 1987. After finishing high school at the Istituto Polivalente M.F.Quintiliano in 2006, he started his bachelor studies in molecular biotechnologies at the University of Bologna. In 2010 he graduated with a thesis on the role of the polymorphism at codon 72 of P53. He continued his studies obtaining a master's degree in bioinformatics in 2016 with a thesis about the role of RNA editing in Humans. The same year he started the PhD at the department of Medical Microbiology at Maastricht university. His research, under the supervision of Dr. John Penders and Paul Savelkoul, was focused on longitudinal analysis approaches for microbiome data.

Acknowledgements

Let's start with this: getting a PhD is never easy!

These four years have been full of both satisfactions and disappointments. I have learned to accept the both of them thanks, also, to the support of all the people who have been with me on this journey.

Therefore, I would like to thank everyone, co-workers and friends, who stood by, supported me, and made this piece of my life more interesting, colourful, and fun.

I would like to start by thanking John.

I could not have hoped for a better supervisor. Your infinite patience, your ability to never judge the point of view of your students, your dedication, your humour and sarcasm have taught me so much and have made the trip really fun.

A big thank you goes to my promoter Paul, who gave me the opportunity to embark on this adventure within the MMB and who has always found the time to listen, advise and make me reflect on the direction of my academic path.

Thanks to my family, whose sacrifices in the past have allowed me to get to where I am now.

Thank you, Jorge, my partner in crime and in life. I lived so long without you but now that you are beside me, I can't imagine how it could be without you.

A warm thanks and a hug to Heike, my paranymph and pole dance instructor. Your kindness and constant availability have carved a special place in my memory. Your unique point of view on problems has inspired me and helped me a lot. Thank you and best wishes for the newcomer.

A thanks to my former roommates Brian, Liene, Matt and Niels.

Matt, you came as a colleague and left as a friend. With you, Ian, Simone and Alex I shared weekends full of Downton Abbey, D&D and board games, Sometimes it wasn't easy between us but I couldn't imagine these four years without you and the guys.

Liene and Brian, who over time have become more than just colleagues, they have become friends. Thanks to both of you, for the long chats on data analysis and R, for the laughter even when we had a lot of work to do, thank you for the coffee breaks, walks and holidays together. These four years would have been very different (and a lot more boring) without you.

Special thanks to Niels, with whom I shared the office, the trip to Ireland, many long chats, and the Gastroenterology paper!

Thanks to Mayk and Kevin for your humour, sometimes nonsensical, sometimes dark but always able to make me laugh like few others can. Thanks also for creating two characters that will stay with me for a long time: Banana man and General Pomegranate.

Thanks to Jiyang and Danyta with whom I was lucky enough to collaborate and with whom I share an article.

Thanks to Casper, Nader, Petra (passionate gamer who I would like to challenge to Mario Kart), Melissa, Giang Li, David, Birke, Chris, Erik, Christel and all the colleagues of the department, with whom I shared many coffee breaks and lunches full of laughter.

Last but not least, a deep thank you to all the other co-authors and all the technicians, without whom I wouldn't have had any data to analyse and therefore, I wouldn't have been able to complete my PhD.

Thank you!