HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Influence of environmental and host factors on the temporal development of early life infant gut microbiome

Master's thesis

Xiaodong Wei

Genetics and Molecular Biosciences

Faculty of Biological and Environmental Sciences

University of Helsinki

2022

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Abstract

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildingsprogram – Degree Programme	
Faculty of Biological and Environmental Sciences	Master's Programme in Genetics and Molecular Biosciences	
Tekijä – Författare – Author		
Xiaodong Wei		
Työn nimi – Arbetets titel – Title		
Influence of environmental and host factors on the temporal development of early life infant gut microbiome		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track		
Genetics and Genomics		
Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's thesis	April 2022	45

Tiivistelmä – Referat – Abstract

The composition and dynamics of the early life gut microbiota plays a major role in establishing neonatal immunity and is suggested to have multiple impacts on the child's long-term health. Meanwhile, the composition of the infant gut microbiome has been shown to be affected by the birth mode, infant health and diet. However, the characterization of the infant gut microbiome and its impact on the host's health is still challenging as the contribution and importance of multiple co-factors on the early microbiome during infant growth is still poorly understood and characterized.

The Health and Early-life microbiota (HELMi) is a cohort of more than 1000 healthy Finnish infants currently followed from birth to 4-5 years old. By now, the HELMi dataset comprises more than 400 whole genome shotgun metagenomes obtained from stool samples from 80 infants and parents, but also an in-depth characterization of the families' lifestyle, environment, health and nutrition, allowing for a precise and cutting-edge characterization of the early gut microbiota. Based on the datasets from the HELMi, this project used Metaphlan3, Kraken and Braken to determine the best computational approach for the taxonomic profiling of the metagenomic reads. Then a PERMANOVA test was performed to evaluate and determine the factors significantly associated with the compositional microbiota variation within the infant gut metagenomes.

This study first identified technical factors introducing bias in taxonomic profiling (*e.g.*, DNA extraction batch), which served as confounders in the analysis of environmental and host variables. The investigation of these biological factors indicates that pre-natal and peri-natal variables such as the mode of delivery significantly impact the infant gut microbiota, while we did not identify any significant impact of breastfeeding habits and medication exposures in this study.

Avainsanat – Nyckelord – Keywords

Infant microbiome, intestinal microbiota development, taxonomic annotation, metagenomics

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Alise Ponsero, Anne Salonen, Roosa Jokela
Säilytyspaikka – Förvaringställe – Where deposited
HELDA – Digital Repository of the University of Helsinki
Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Contents

Abbreviations

BMI	Body mass index
BLAST	Basic Local Alignment and Search Tool
C-section; CS	Caesarean section
CSC	IT Center for science
DNA	Deoxyribonucleic acid
FDR	False discovery rate
FIMM	Institute for Molecular Medicine Finland
GBS	Group B *Streptococcus*
GI	Gastrointestinal
HELMi	Health and early life microbiota
KEGG	Kyoto encyclopedia of genes and genomes
NCBI	National center for biotechnology information
PCoA	Principal coordinate analysis
PERMANOVA	Permutational multivariate analysis
QC	Quality control
RefSeq	NCBI Reference Sequence database
TiB	Tebibyte
WGS	Whole genome shotgun sequencing

1. Introduction

1.1 Early life gut microbiome and infant health

The human microbiota is a complex community of microorganisms (prokaryotes, eukaryotes and viruses), living in the human body space. One of the most studied and well-characterized microbiotas is the gut microbiota, that have been shown to be central in human digestion and health. The acquisition of the gut microbiota occurs immediately at birth with the seeding of the infant from the mother's microbiota playing a major role. Recent studies have shown that the infant gut microbiota is obtained at birth from maternal faecal sources (Korpela et al., 2020; Wilson et al., 2021). Regardless of the exact maternal source, the early life microbiome is characterized by a low bacterial diversity and a microbiome dominated by bacterial genera such as *Bifidobacterium*, *Bacteroides, Escherichia* and *Veillonella,* also *Clostridium* and *Prevotella* (Bokulich et al., 2016; Korpela & de Vos, 2018). This initial microbial community will drastically change over the first years of life, becoming more and more diverse, until reaching a climax around 3 years old with some developments continuing still at school age (Derrien et al., 2019).

Extensive research has been devoted to describing the dynamics of the early life gut microbiome, and several studies have shown that the early life microbiota plays a central role in the infant's early and lifelong health. Indeed, the infant gut microbiota colonization has been suggested to be essential for the neonatal immunity establishment (Romano-Keeler et al., 2014; Stiemsma & Michels, 2018), to the infant's well-being (de Weerth et al., 2013; Pärtty et al., 2012), and has been also suggested to impact the brain development (Ihekweazu & Versalovic, 2018). Additionally, the infant microbiota development has been shown to have long-term health impacts, and disruptions in the early microbiota acquisitions has been linked to obesity risks (Stanislawski et al., n.d.), and to the development of allergic diseases and asthma (Abrahamsson et al., 2014).

During the gut colonization and during the early childhood, the infant gut microbiota is highly dynamic and can be affected by several host and environmental factors. Firstly, the birth mode is one of the clearest factors that has profound impacts on the infant gut microbiota development. Indeed, infants born by caesarean section (CS) have a microbial colonization drastically disrupted

in all body sites (Bokulich et al., 2016; Dominguez-Bello et al., 2010; Shao et al., 2019; Yassour et al., 2016). In the gut, CS born infants have a typical delayed acquisition of bacteria from the genus *Bacteroides*, still visible at 1 year of age (Stewart et al., 2018). Secondly, antibiotic exposures at birth or in early-life have also been shown to have drastic impacts on the gut microbiota development. Indeed, early antibiotic exposure can lead to a temporary decreased gut microbial diversity, and an enrichment in antibiotic resistance genes (Busi et al., 2021). Finally, infant diet, in particular breastfeeding, has been shown to have a large effect on microbiota development (Stewart et al., 2018). The use of formula feeding instead of breastfeeding leads to a higher gut microbial diversity and more instability in the microbial community dynamics (Forbes et al., 2018).

In total, a large number of exposures and host variables have been suggested to influence the gut microbiota development. However, the contribution and importance of these multiple factors as the microbiota matures is still poorly understood and requires large infant cohorts to be fully characterized.

1.2 Metagenomic sequencing approaches for the study of the gut microbiome

Early microbiota studies relied on a culture-based approach to study the faecal microbiota. However, these approaches allow the investigation of only a small subset of microbes, and do not allow to investigate the relative abundances of each microbe in their environment. The relatively recent advent of next-generation sequencing methods has allowed to leverage the genetic material directly extracted from the environmental sample to enable a rapid, untargeted and less biased taxa detection and by doing so allowed discovery of novel microbial species from a large array of ecosystems. The use of 16S rRNA gene sequencing methods, have been extensively used to support investigation of the taxonomic composition of the faecal microbiota (Almonacid et al., 2017; Cortes et al., 2019), and more recently, whole genome shotgun sequencing (WGS) has been applied to investigate the taxonomic but also the functional diversity of the human gut microbiota (Qin et al., 2010; Ranjan et al., 2016).

However, identifying the microbial genus or species present in complex environmental sample is still challenging and prone to biases. Importantly, modern metagenomic datasets are typically composed of several millions of short DNA sequences (ranging from 50 to 200 nucleotides). In this context, the main computational challenge is to be able to accurately classify such a large number of sequences in a reasonable computational time. Algorithms such as BLAST (basic local alignment and search tool), allowing sensitive sequence alignments, are too computationally intensive for these tasks (Altschul et al., 1990). Additionally, this challenge is made even more complex by the exponential growth of the number of sequenced genomes in recent years, requiring tools able to search a large database of reference sequences, but also to allow for a regular update of their databases.

A large number of tools have recently been developed which are focused on classifying large amounts of sequencing reads to known taxa with increasing speed. These tools are typically not as sensitive as BLAST but are designed to be much faster. We can divide these tools in two main groups: DNA-to-DNA classification and marker-based classification. On the one hand, DNA-to-DNA classifiers such as Kraken2 (Wood et al., 2019), Centrifuge (Kim et al., 2016) or Kaiju (Menzel et al., 2016) will compare each sequencing reads with a comprehensive DNA genome database. On the other hand, marker-based methods such as Metaphlan2 (Truong et al., 2015) and Metaphlan3 (Beghini et al., 2021), typically compare sequencing reads to a reference database containing marker gene sequences (specific gene families that allows to discriminate between species). The use of such a small subset of genes makes these methods particularly rapid; however, the marker sequences used can introduce a bias in the results when they are not evenly distributed among the microbial sequences of interest (D'Amore et al., 2016), and will not be able to classify sequencing reads that do not carry a marker gene. To ensure the accurate analysis and interpretation of metagenomic data, it is important to compare results obtained by these different classifiers, in order to determine the best approach for a given study.

1.3 The HELMi cohort

In order to have a better understanding of the interaction between early life gut microbiota, early exposures and infant health, the Health and Early Life Microbiota (HELMi) cohort was established to follow up healthy Finnish infants from birth to 2 years old, collecting both regular faecal samples and an extensive longitudinal metadata related to infant exposures, health and development (Korpela et al., 2019). The diverse data provided by the HELMi cohort allows for a precise and cutting-edge characterization of the impact of host and environmental variables on the early gut microbiota. Importantly, particular efforts have been made to control technical variables during the entire study in regards to faecal sample collection, storage and processing, thus providing an opportunity to assess the impact of these technical variables on the observed results.

2. Aims of the Thesis

This thesis project will leverage the WGS metagenome dataset generated from HELMi infant faecal samples. The goal of this thesis is to identify potential associations between environmental and host factors with the infant gut microbiota variations at different ages. To achieve this global goal, this thesis was divided into three aims:

- **Aim 1**: Assessing the impacts of the taxonomic annotation methods on the infant microbiota profile.
- **Aim 2**: Exploring the potential impact of technical variables such as sample collection, storage and sequencing technology. This will allow for the identification of confounders for the rest of the analysis.
- **Aim 3**: Leveraging the extensive metadata collected from the HELMi families, the potential source of variation due to peri- and post-natal variables will be investigated.

3. Materials and Methods

3.1 The HELMi cohort

This section describes work conducted prior to the beginning of the current master's project.
The Health and Early-Life Microbiota (HELMi) is a cohort study of 1,055 healthy term Finnish infants born in the Uusimaa region in 2016-2018 (Korpela et al., 2019). In total, more than 10,000 faecal samples from both infants from birth to 2 years of age and parents were collected. Additionally, families enrolled in the HELMi project answered regular online questionnaires, allowing for an in-depth description of the lifestyle, living environment, infant and maternal health, and infant nutrition. The study protocols have been approved by the ethical committee of The Hospital District of Helsinki and Uusimaa (263/13/03/03 2015 and HUS/2126/2020) and are performed in accordance with the principles of the Helsinki Declaration. The guardians have provided an informed and written consent. Participation is completely voluntary and the participants could withdraw from the study at any point.

The generated samples and questionnaire data are considered personal data. In this setting, any data processing (data collection, storage, protection, retention, and destruction) takes place according to the General Data Protection Directive and Finnish laws to ensure lawful, fair, and transparent data processing. All data is processed in pseudonymized format, as longitudinal samples and datasets cannot be fully anonymized.

3.2 HELMi questionnaire information

3.2.1 HELMi questionnaire collection

This section describes work conducted prior to the beginning of the current master's project.

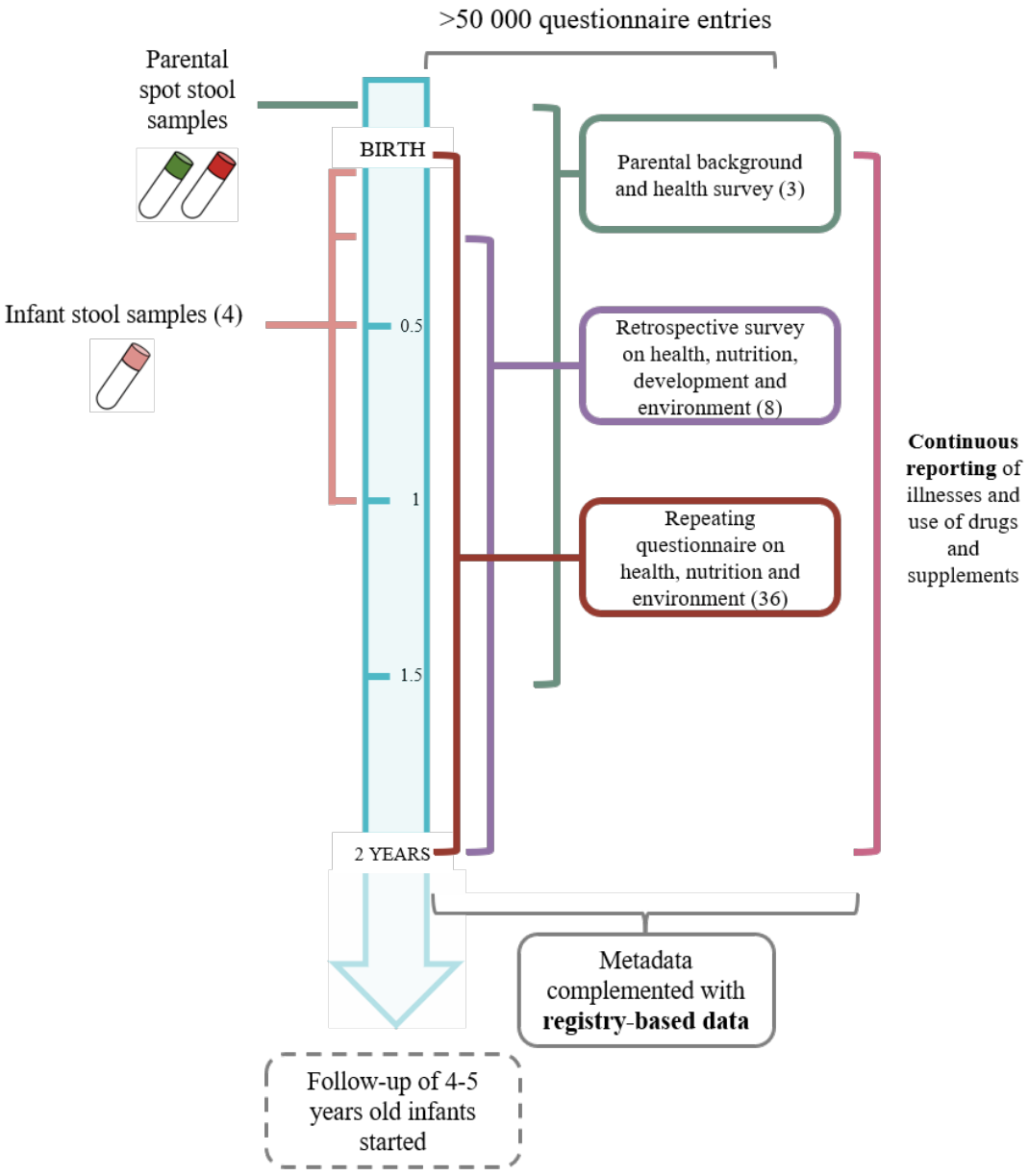


Figure 1. Overview schema of the questionnaires and stool samples involved in this project. *The parental samples were collected before the birth (only once). The infant stool samples used in this project were collect at 3 weeks, 12 weeks, 6 months, and 12 months of age. The parental survey (shown in green) tracks the basic background and health status of the parents, and the details of the childbirth, especially the mothers' medicine usage during gestation. The retrospective survey (shown in purple) and repeating questionnaire (shown in red) records the overall health status and nutrition of the infant at different frequencies. The image is modified from Korpela et al., 2019.*

In addition to stool sample collection, the families participating in the HELMi cohort were asked to answer a series of repeating questionnaires through an online platform (Figure 1). This allowed the collection of a large amount of data on the infant and their parents. Only the questionnaires relevant for this project are described below:

- Background questionnaire

A series of basic information concerning the pregnancy conditions (*e.g.*, gestational diabetes, probiotic or vitamin intake) and concerning the childbirth (*e.g.*, date of birth, mode of birth, place of birth, gestational age) were collected immediately after birth. Additionally, this questionnaire contains information concerning the maternal and paternal factors (*e.g.*, age, education, BMI, allergy, or genetic diseases) and living environment (*e.g.*, number of siblings, number of pets).

- Breastfeeding questionnaire as part of a repeating questionnaire

Breastfeeding and nutritional practices were monitored at weekly to monthly frequency in order to follow the evolution of the infant nutrition during early childhood. Questionnaires contained questions concerning the breastfeeding exclusivity and duration, exposure to infant formula, as well as introduction to solid food.

- Medication diary

Exposures to medication were recorded by the parents in a continuous diary-type of questionnaire. The parents described the medication name, date of start and end. Manual curation of this diary was conducted to classify the medication types and active compounds.

- Retrospective health and life-style questionnaire

The infant's global health and well-being was monitored through a questionnaire recurring every 3 months. In particular, information concerning the gastrointestinal function (*e.g.*, stool frequency, consistency and colour, frequency of symptoms such as regurgitation, stomach pain, flatulence) and allergy (*e.g.*, skin health, asthma) were monitored through these recurring questionnaires. Additionally, the child's sleeping habits, weight and height trajectories, living environment, and outdoor exposure were recorded in this questionnaire.

3.2.2 HELMi questionnaire aggregation and curation

First, each question was grouped according to the information they carried (maternal/paternal health, pregnancy, infant health, medication exposures, infant nutrition, environment, and lifestyle) and according to the data format (integer, boolean and categorical). Free text answers were excluded from the analysis. Then, the variables were mapped to their corresponding microbiota sample. For the background questionnaire variables, the answers were mapped to all samples collected for the family. However, for breastfeeding, health questionnaire and medication diary, each sample was mapped to the closest answer given in time.

The selected variables were then explored for their usability in this project. In particular, variables with a high number of missing answer (>60%) were excluded. Additionally, for categorical variables, highly unbalanced groups were excluded (one category accounting for more than 80% of the responses). Finally, for numeric variables, answers were manually checked for inconsistencies and impossible values that were curated out of the dataset. The complete list of variables selected, their definition and original questions are available in supplemental table 1. These analyses were conducted and visualized in R (version 1.4.1717), using the tidyverse R package (Wickham et al., 2019).

3.3 HELMi WGS metagenomes

3.3.1 Collection and sequencing of the metagenomes in the HELMi project

This section describes work conducted prior to the beginning of the current master's thesis project. In this project, we are leveraging a collection of stool samples from 80 infants, collected at 3 weeks, 12 weeks, 6 months, and 12 months of age, and of parental (maternal sample and father's/current partner's sample) stool samples collected in the two weeks prior delivery. In total the dataset consisted of 307 infant samples and 106 adult samples. Stool samples were collected by the parents at home using a provided kit, and were immediately stored at home at -20°C. Upon receival in the laboratory, the samples were immediately stored at -80°C until processing. After faecal DNA extraction in 96-plate format using a repeated bead-beating method (Korpela, Kallio, et al., 2021), WGS DNA libraries were prepared using Illumina Nextera Flex or iGenomX Riptide

High Throughput Rapid Library Prep Kit and sequenced using Illumina HiSeq and NovaSeq 6000 sequencing platforms at the sequencing unit of the Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland.

3.3.2 Taxonomic annotation of the metagenomes

During this project, a computational pipeline was assembled to perform the following computational tasks: (1) quality control (QC) of the raw metagenomic reads, (2) human read filtering and (3) taxonomic profiling of the reads after QC (Figure 2).

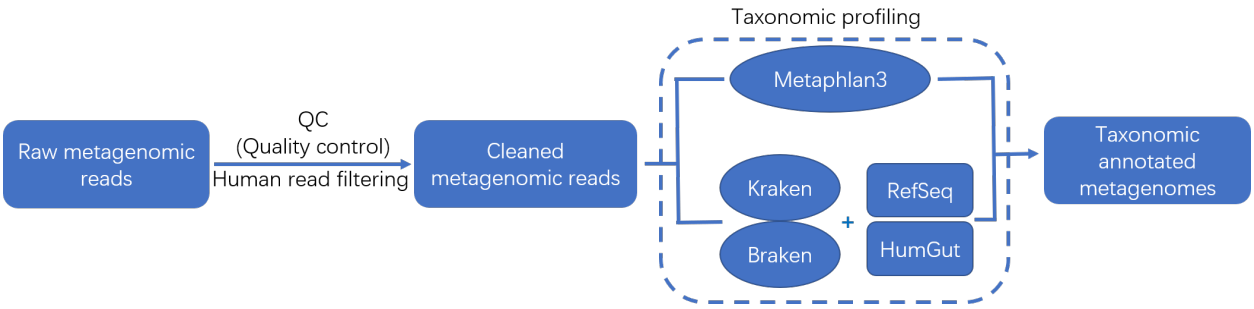


Figure 2. Taxonomic profiling of metagenomes. *After quality control (QC) and human read trimming, the metagenomic reads were annotated by using three taxonomic tools. Metaphlan3 annotated the metagenomes with its inbuilt reference database, while Kraken and Braken relies on an external database (RefSeq or HumGut). In total, five taxonomic profiling strategies were determined in this project: (1) Kraken with RefSeq database, (2) Kraken with HumGut database, (3) Kraken+Braken with RefSeq database, (4) Kraken+Braken with HumGut database and (5) Metaphlan3.*

After sequencing, the sequencing fastq files first underwent quality control. First, low-quality base calls (below 20 Phred score) were trimmed off from the 3' end of the reads before sequencing adapter removal using the tools FastQC v0.11.9 (Andrews, 2010) and TrimGalore v0.6.6 (Krueger, 2016/2022). Then, adapters and primers were removed using the adapter detection tool of Cutadapt v3.4 (Martin, 2011). After these two trimming steps, reads shorter than 200 bp were discarded, while taking into account the paired-end nature of the dataset. After QC, the metagenomes were screened to remove any contaminating human host sequences. The reads were mapped against a

non-redundant version of the Genome Reference Consortium Human Build 38 (available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40; patch release 14) using Bowtie2 v2.4.2 (Langmead & Salzberg, 2012). Reads with a significant match with the human genome were removed from the dataset.

Next, the remaining reads were annotated and the taxonomic profiles of the samples were determined. In order to determine the best computational approach for the taxonomic profiling, several tools were used. First, Metaphlan3 (Truong et al., 2015), a widely used taxonomic tool that relies on a gene database to annotate the reads from an unassembled metagenome, was run using the inbuilt reference database. We additionally used Kraken2 (Wood et al., 2019), a tool that uses *k*-mer profiles to efficiently map reads to a database of known genomes. Kraken2 is often used along with Braken (Lu et al., 2017), which uses a Bayesian inference to re-calculate and improve Kraken2 profiles. In order to assess the impact of the reference database in the taxonomic profile obtained, we annotated the metagenomes obtained by Kraken2 and Braken using the RefSeq database (downloaded from https://github.com/BenLangmead/aws-indexes on 21/05/2021) (Pruitt et al., 2005) and the HumGut database (downloaded from https://github.com/larssnip/HumGut on 20/09/2021) (Hiseni et al., 2021). While RefSeq is a more complete database composed of all non-redundant complete genomes from microbial species deposited in NCBI, the HumGut database is a database tailored for profiling human gut metagenomes and comprises complete genomes and metagenome assembled genomes from human gut.

3.3.3 Computational resources and environment

As this project involves a large dataset composed of 413 metagenomes (*ca.* 2 TiB), the cloud computing resources for the QC, human read removal, taxonomic and functional profiling as well as for the storing of the metagenomes were utilized from the Finland's IT Center for Science (CSC). Because the data used are considered as low-risk personal data, the metagenomes were stored encrypted in the "Allas" object storage system after human read removal. The supercomputer "Puhti", composed of several computational unit nodes, was used for the computational tasks.

3.4 Variance analysis

In order to assess each variable's impact on the gut microbiota composition, we used a series of permutational multivariate analysis (PERMANOVA) tests using the adonis2 command in the vegan package version (Oksanen et al., 2022). First, the metagenomic taxonomic read counts were aggregated at the species level, and the taxonomic counts were transformed into relative abundances. Then, a Bray-Curtis distance was computed between the taxonomic profiles and a PERMANOVA test was applied for each variables using 9999 permutations. For each of these PERMANOVA test, a beta dispersion test was performed as well. The result of beta dispersion test indicates whether or not the significant result in PERMANOVA could be due to a difference in dispersion rather than in composition (Weiss et al., 2017). Finally, the *p*-values obtained from the multiple PERMANOVA tests were corrected for false positive rate using a false discovery rate (FDR) correction.

Importantly, this project first assessed the influence of technical variables on the observed microbiota composition. From the PERMANOVA tests, we determined that the DNA extraction plate ID and reads number had a significant impact on the observed variance and were chosen as cofounders for the biological variables.

4. Results

In this project, we first evaluated the impact of different taxonomic annotation methods on the observed gut microbiota profiles (section 4.1), allowing us to select one annotation approach. We then used this profiling method to evaluate the impact of technical variables on the observed microbiota variance between samples (section 4.2), allowing us to identify confounding factors. Using this knowledge, we explored the impact of biological variables on the observed taxonomic variance in our samples (section 4.3, 4.4 and 4.5).

4.1 Assessing the impact of annotation tools on taxonomic profiling results

To explore the potential impact of the choice of taxonomic annotation tools on the taxonomic profile in WGS metagenomes, we annotated the HELMi WGS metagenomes using various bioinformatic tools and databases.

To assess the impact of the choice of database, we first compared the annotations obtained for Kraken using RefSeq and HumGut databases. We observed that use of the HumGut database allowed us to radically reduce the proportion of reads left unannotated in particular for adult samples (Figure 5). We then compared the results obtained using Kraken+Braken against the same databases. The abundance profiles obtained at the phylum and family levels using Kraken+Braken against the RefSeq or HumGut were highly similar (Figure 6A and B). The median Bray-Curtis distance computed for the same sample using the two databases was therefore low (0.16 IQR: 0.17), and was significantly lower than the Bray-Curtis distance obtained between different samples at the same time point (0.60 IQR: 0.37) (Unpaired Wilcoxon test, p-value < 2.2e-16).

These results show that using Kraken+Braken along with the HumGut database allowed a higher proportion of reads to be annotated compared to Kraken or Kraken+Braken using the RefSeq database.

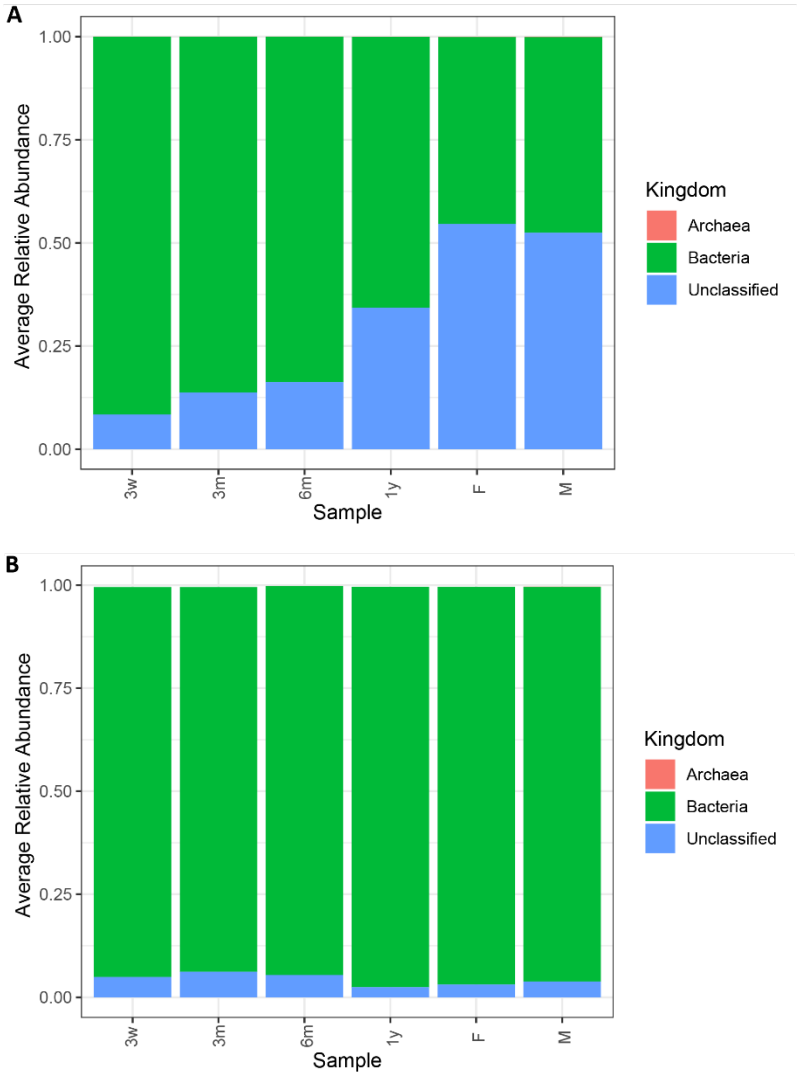


Figure 5. Proportions of annotated and unannotated reads at the phylum level using Kraken with the (A) RefSeq database or the (B) HumGut database. *The taxonomic profiles obtained by Kraken using each database were aggregated at the kingdom level and averaged by infant age (3w= 3 weeks; 3m=3 months; 6m=6 months and 1y=1 year) and by type of parental sample (F=Father's sample, M=Mother's sample).*

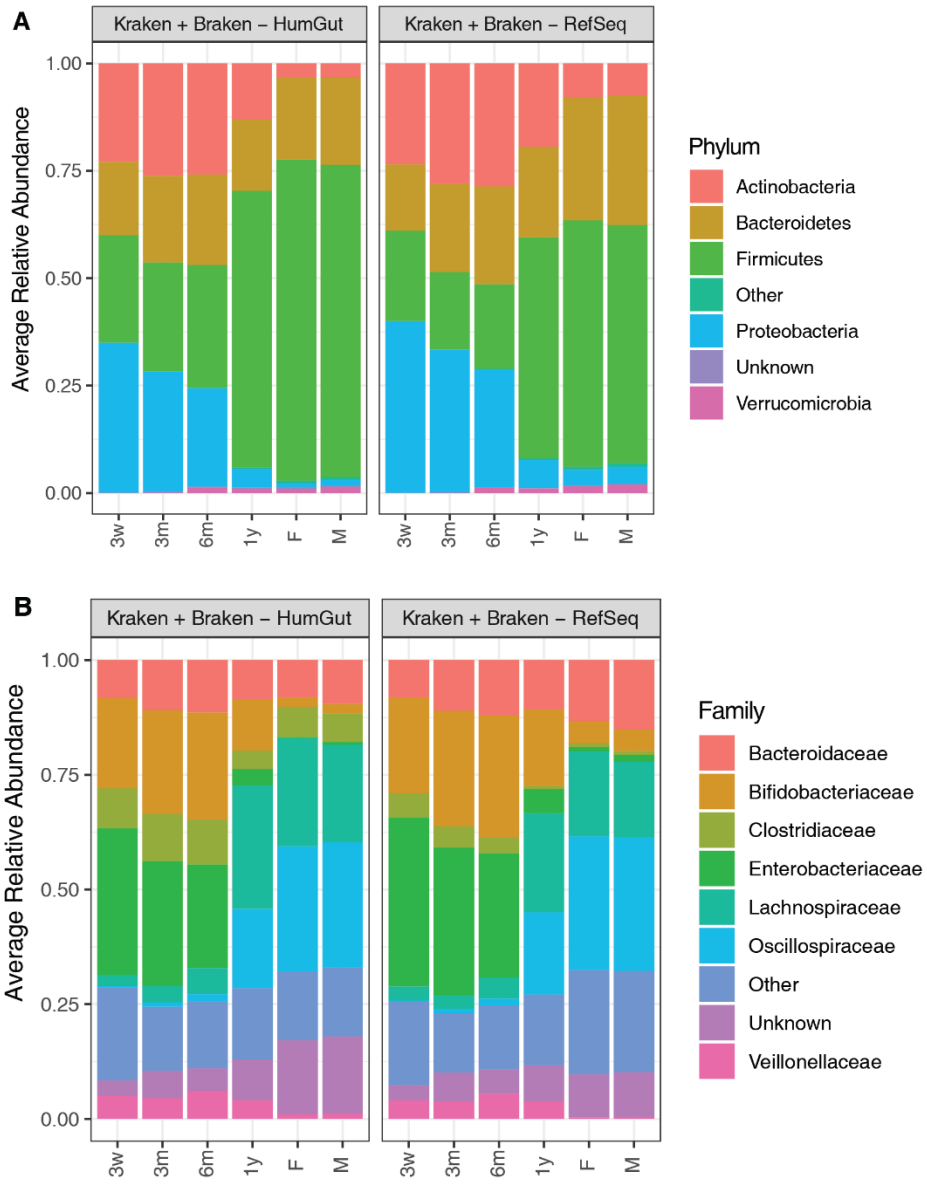
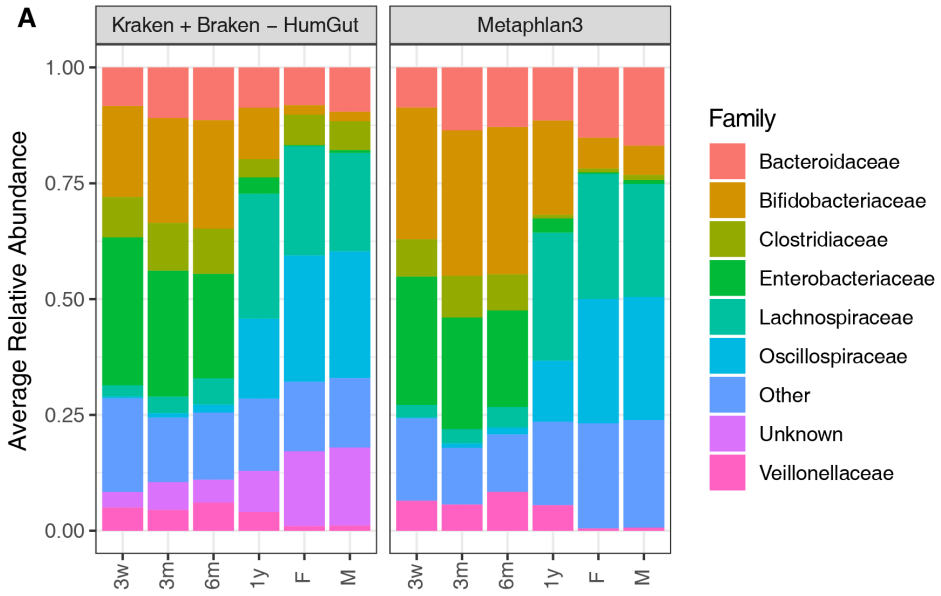


Figure 6. Taxonomic profiles of the annotated reads obtained using Kraken+Braken with the RefSeq database or the HumGut database at the (A) phylum or (B) family level. *The taxonomic profiles obtained by Kraken+Braken using each database were aggregated at the phylum or family level and averaged by infant age (3w= 3 weeks; 3m=3 months; 6m=6 months and 1y=1 year) and by type of parental sample (F=Father's sample, M=Mother's sample). Low abundance taxa (<10% relative abundance) were grouped as "Other".*

We next compared Kraken+Braken using the HumGut database to Metaphlan3. The taxonomic profiles obtained by the two methods at the family level were highly similar (Figure 7A). However, Metaphlan3 did not report the proportion of sequences left unannotated, whereas *k*-mer-based approaches such as Kraken+Braken reported them as "unknown". Importantly, all tools reported the same evolution of richness (Chao1 index) and alpha-diversity (Shannon index) during the infant growth, with a significant increase in richness and alpha-diversity during the first year of life, and no differences between the two parental samples (Figure 7B). However, Kraken+Braken allowed the detection of a higher number of different species than Metaphlan3, hence the species richness was consistently higher for the Kraken and Kraken+Braken using either database than using Metaphlan3.



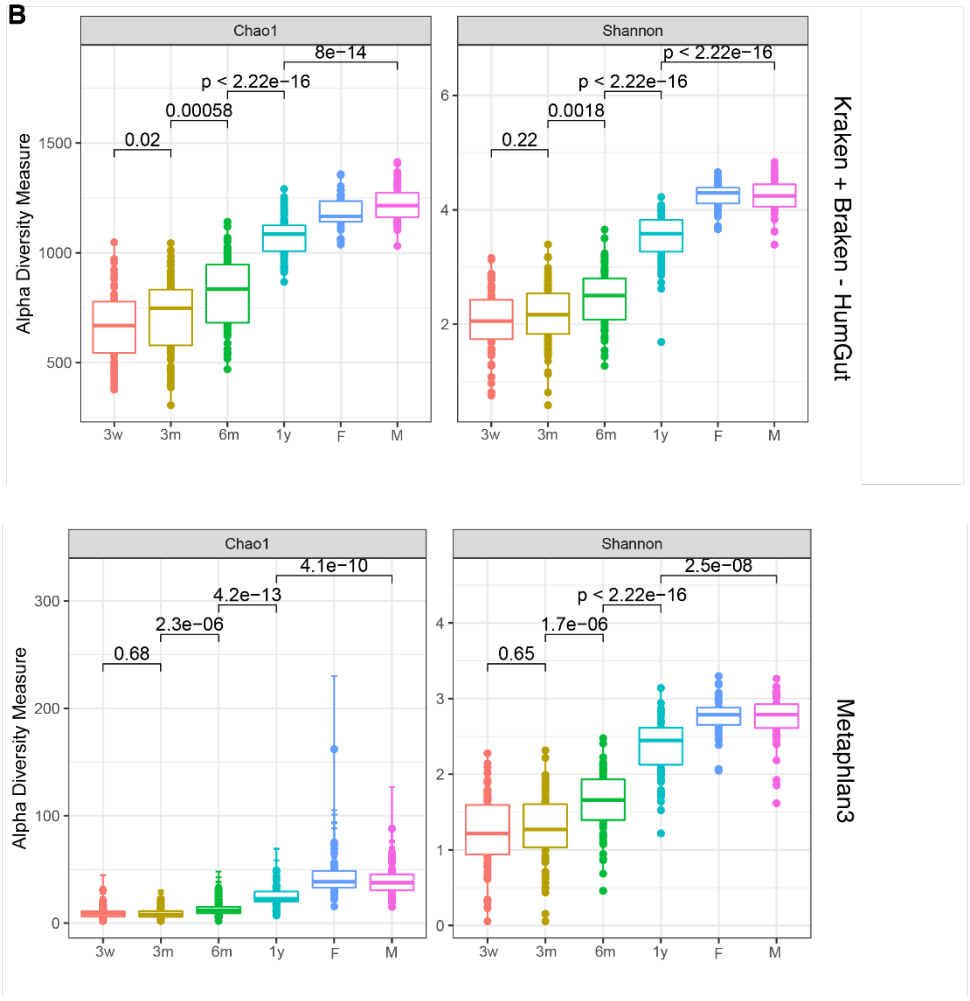


Figure 7. Microbiota composition, species richness and alpha-diversity obtained with Kraken+Braken using the HumGut database and Metaphlan3.

(A) Relative abundance in taxa for infant and parental samples using Kraken+Braken and Metaphlan3. For each tool, the relative abundance profiles obtained for each metagenome were aggregated to the family level, and rare families (below 10% relative abundance and 10% prevalence) were grouped as "Other". The relative abundances were then averaged by infant age (3w= 3 weeks; 3m=3 months; 6m=6 months and 1y=1 year) and by type of parental sample (F=Father's sample, M=Mother's sample).

(B) The species richness (Chao1 index) and alpha-diversity (Shannon index) for infant and parental samples. Richness and alpha-diversity were calculated after rarefying the raw counts to the same sequencing depth. The indices are grouped by infant age (3w= 3 weeks; 3m=3 months; 6m=6 months and 1y=1 year) and by type of parental sample (F=Father sample, M=mother sample). Comparisons between samples groups were performed using unpaired Wilcoxon test.

All in all, these results favoured the selection of the Kraken+Braken using the HumGut database annotation method, as it achieved the lowest relative abundance of unannotated sequences as well as captured the highest taxonomic richness and diversity.

Using this method, the maturation of the infant microbiota can be observed during the first year of life, with infant microbiota composition converging toward a more adult-like composition, while being still distinct from the adult samples at 1 year (Figure 8).

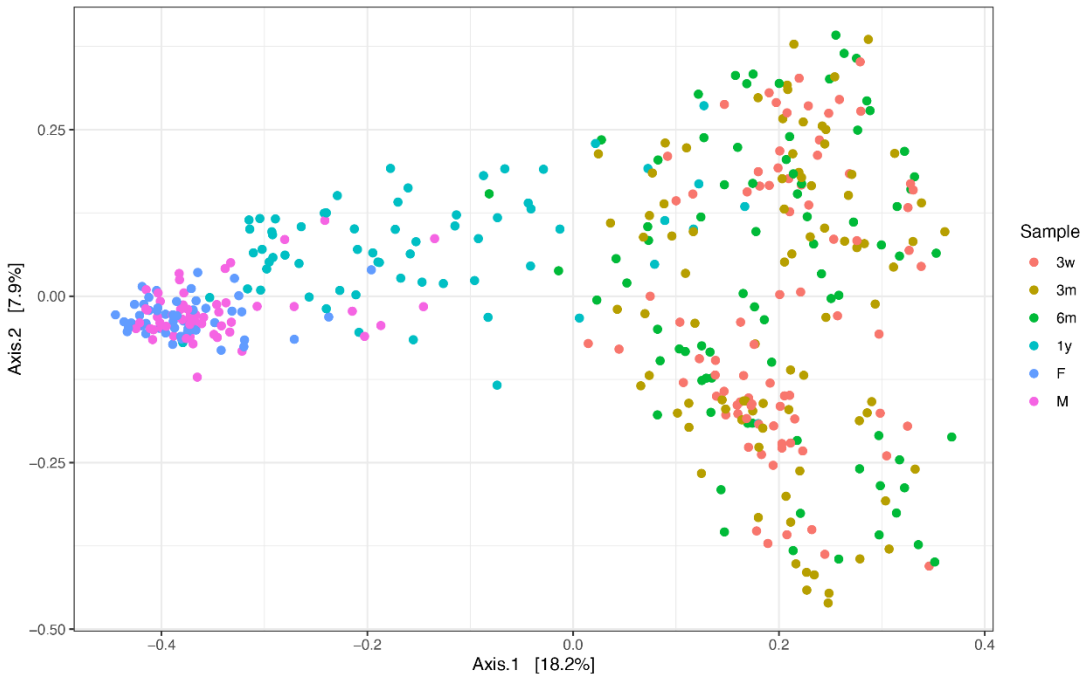


Figure 8. Principal coordinate analysis (PCoA) plot of the HELMi metagenomes using Bray-Curtis distance. *The samples were annotated using Kraken with Braken using the HumGut database, the counts were aggregated at the species level and a Bray-Curtis distance computed between samples. The samples are coloured by infant age (3w= 3 weeks; 3m=3 months; 6m=6 months and 1y=1 year) and by type of parental sample (F=Father sample, M=mother sample).*

4.2 Effect of technical variables on the taxonomic composition

First, we aimed to address the impact of technical variables in the taxonomic variations observed in the metagenomes. In total, 14 technical variables were taken into account in this project. Among them, 4 categorical variables were excluded because the categories were highly unbalanced (one category covering more than 80% of the answers) and one variable was removed because of a high overlap with another variable. In total, 10 technical variables were kept for the variance analysis. These variables described the sample collection, in particular the storage time in days; the sample consistency (Bristol score) as well as the DNA extraction batch (extraction Plate ID); DNA yield; and the sequencing run (Figure 9).

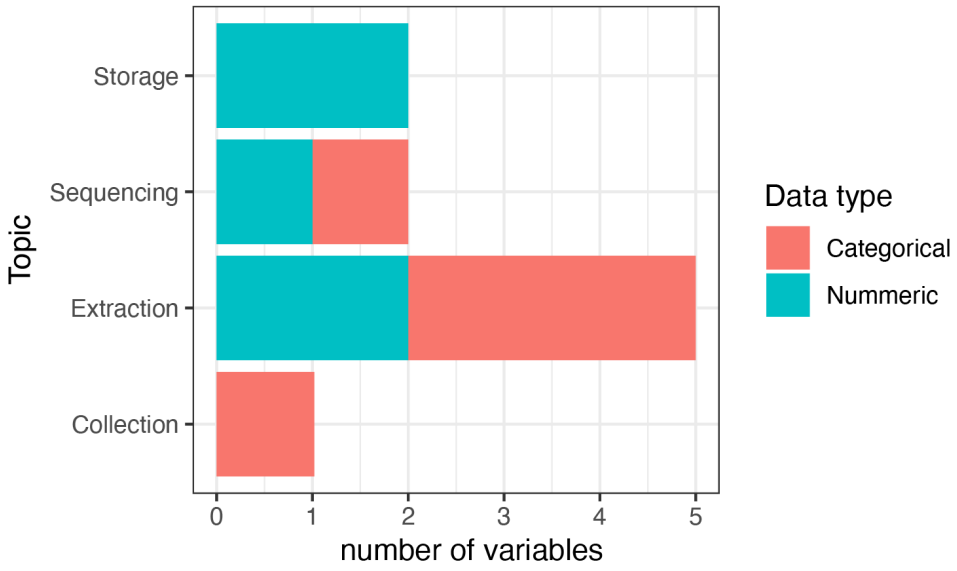


Figure 9. Distribution of selected technical variables for stool samples, coloured by data type.

Just a few technical variables were shown to have a significant impact on the taxonomic variation (Figure 10). In particular, none of the variables concerning sample collection and storage conditions were found to have a significant impact. However, the person performing the DNA extraction was found to have a significant impact on the gut microbiota variation, which suggests that the experimenters' habits as well as other variations among different batches of extractions can cause a significant bias in the final data. Additionally, the sequencing depth (total number of

sequencing reads/samples) and the DNA concentration after extraction were found to have a significant impact on the taxonomic variation.

From these results, we chose the DNA extraction batch (extraction Plate ID) as a confounder in further analysis as it takes into account the effect of the person performing the DNA extractions as well as possible variations due to the date of extraction. The total number of reads were also selected as a confounder for the further variance analysis.

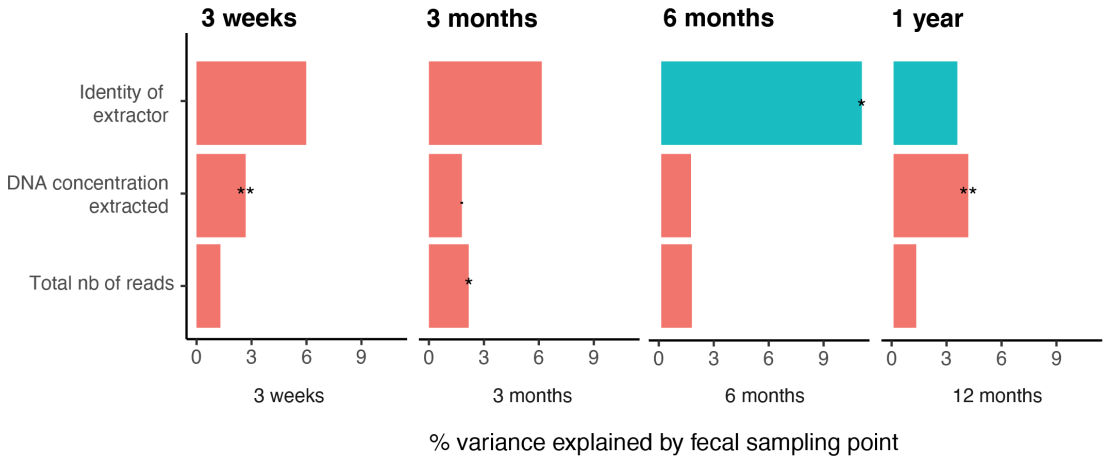


Figure 10. Significance and explained variance of the gut microbiota by technical covariates modelled with PERMANOVA test*. Results for variables with homogenous dispersion (Red) were deemed reliable. ***; $p < 0.0001$, **: $p < 0.001$, *: $p < 0.01$, .: $p < 0.1$ from adonis2, (permutation=999). Only variables with an $p \leq 0.05$ in at least one time point were plotted.*

4.3 Effect of background variables on the taxonomic composition

In total, 132 background variables for the study subjects were taken into account in this project. These variables describe the mother's and partner's health prior and during pregnancy, life style such as cleaning habits, living surrounding, and variables describing the birth conditions (mode of delivery, antibiotic exposures at birth, etc.) (Figure 11). From these 132 variables, 56 were excluded because of they had a highly imbalanced distribution (one category covering more than

80% of the answers), and 37 variables were removed because of a high overlap with one or more variables.

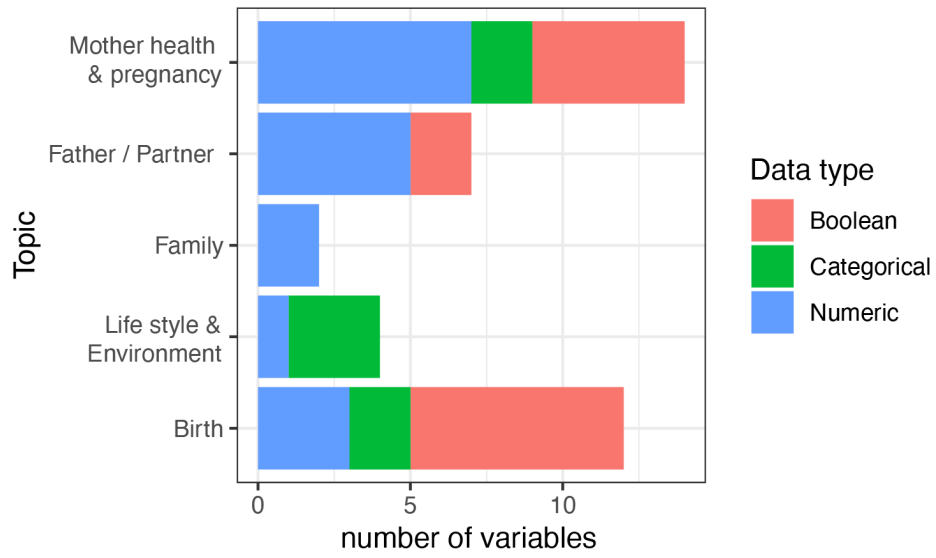


Figure 11. Distribution of background variables, coloured by data type.

The PERMANOVA and beta-dispersion test of the usable variables showed that there were 21 variables with significant impact (FDR q-value ≤ 0.25) on the variation observed in the infant gut microbiota on at least one time point. The infant delivery mode and intrapartum antibiotic had the most long-lasting effect on the infant gut microbiota, with an effect visible from 3 weeks to 12 months after birth (FDR q-value < 0.01, with non-significant beta-dispersion at all time points for birth mode and after 6 months for intrapartum antibiotics) (Figure 12). Other variables on the birth conditions, such as stay in intensive care and the gestational age were mildly significant at only one time point, suggesting a potential effect that needs to be further confirmed.

Paternal and maternal education level were found to have a significant impact in the early time points, although significant beta-dispersion in the maternal education level does not allow to exclude an effect of imbalanced categories. Interestingly, the maternal and paternal status in heritable diseases (including allergies and asthma), were found to have a significant impact on the infant's gut microbiota variation at one year (Figure 12).

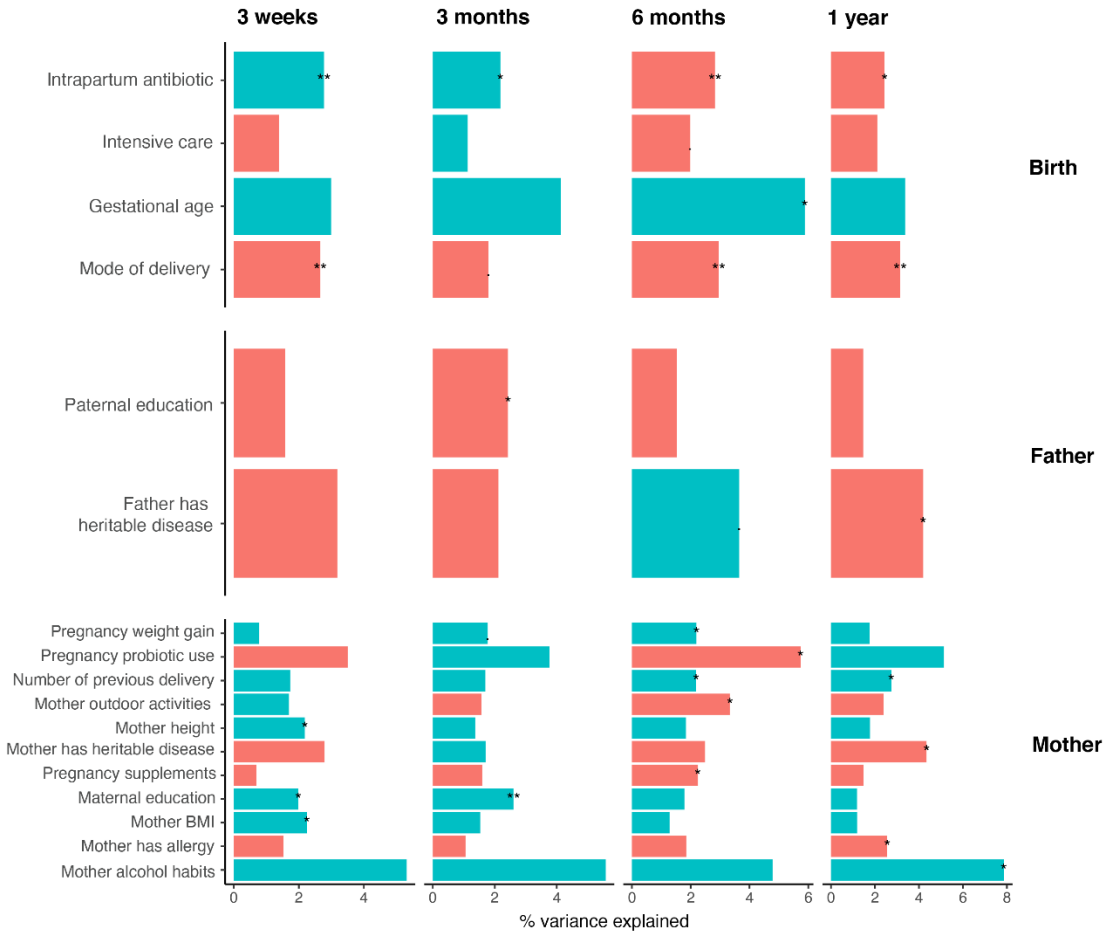


Figure 12. Significance and explained variance of the microbiota by background variables on birth, mother and father modelled with PERMANOVA test. *Results for variables with homogenous dispersion (Red) were deemed reliable. ***: $p < 0.0001$, **: $p < 0.001$, *: $p < 0.01$, .: $p < 0.1$ from adonis2 (PERMANOVA permutation=9999, confounders: plate ID and total number of reads). Only variables with an FDR corrected q-values ≤ 0.25 were plotted.*

We next explored the impact of variables describing the infant's living environment and family variables (Figure 13). As previously observed for the maternal and paternal education level, the maximum education level of the couple had a significant impact on the infant gut microbiota variations at 3 weeks and 3 months. Additionally, the number of siblings and pets, as well as the frequency of cleaning was found to be significant, however their positive beta-dispersion does not allow to exclude an effect of imbalanced group sizes.

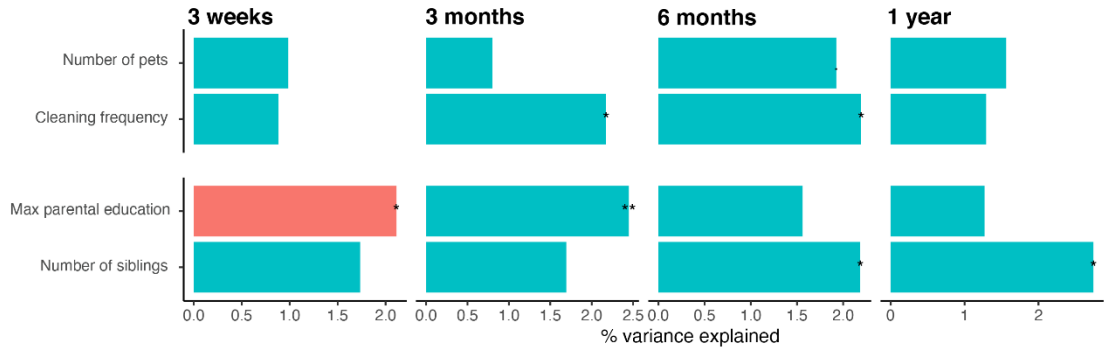


Figure 13. ***Significance and explained variance of the infant gut microbiota by life-style and family variables modelled with PERMANOVA test.*** *Results for variables with homogenous dispersion (Red) were deemed reliable. **: $p < 0.001$, *: $p < 0.01$, .: $p < 0.1$ from adonis2 (PERMANOVA permutation=9999, cofounders: plate ID and total number of reads). Only variables with an FDR corrected q-values ≤ 0.25 were plotted.*

4.4 Effect of breastfeeding variables on the taxonomic composition

In total, we surveyed the breastfeeding habits using 20 variables. These variables described the infant early life diet, such as duration of breastfeeding, formula exposure, solid food introduction. Altogether 11 variables were excluded for a highly imbalanced distribution, and 2 variables were removed because they overlapped with other variables. The distribution of data types for the selected variables can be found in Figure 14. Interestingly, no breastfeeding and diet variables were shown to have a significant impact on the infant gut microbiota variation at any age.

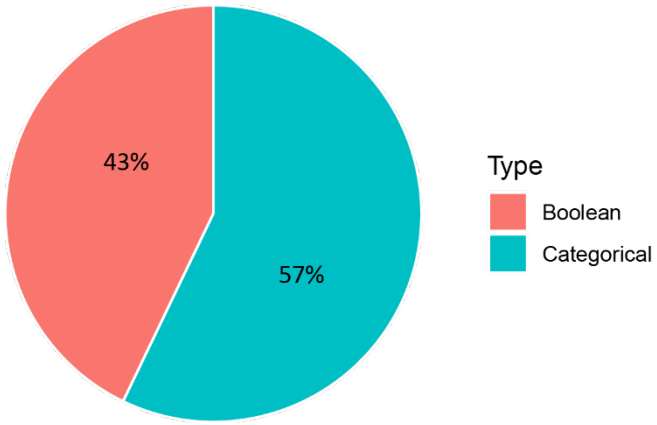


Figure 14. Distribution of data types for breastfeeding variables.

4.5 Effect of health and medication treatments variables on the metagenome composition

In total, we collected 12 variables describing medication exposures for infants. These variables described the frequency and types of exposures for the infant, in particular to antibiotic treatments. However, 11 of these variables were highly unbalanced and had to be excluded. We tested the presence or absence of any medication treatment course in the 3 weeks prior sampling; however, this variable was not found to have a significant impact on the taxonomic composition of the gut microbiota at any age.

Additionally, we collected 61 health-related variables in this project. These variables explore the infant health status, including the infant's development of allergic disease, general health status, gastrointestinal health and well-being. Altogether 48 variables were excluded because of highly imbalanced group or overlap with other variables. The distribution and data type of the selected variables is represented in Figure 15.

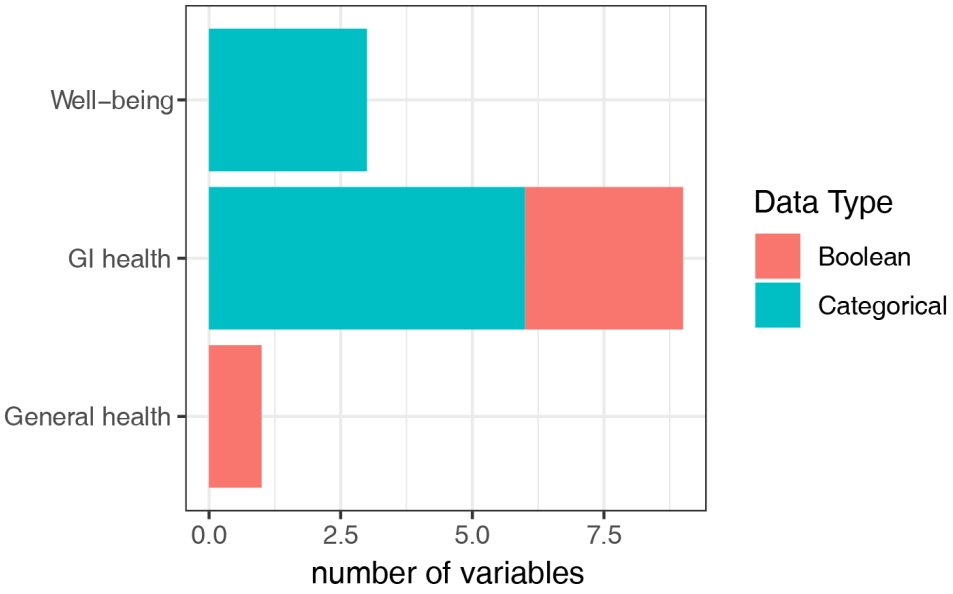


Figure 15. Distribution of health-related variables, coloured by data type.
GI health = Gastrointestinal Health

The PERMANOVA test identified 7 health variables that had significant impact on the infant gut microbiota variation at least in one time point (Figure 16). First, the general health variable describing if the infant was sick during the past three months was found to be significant. However, the beta-dispersion was also significant for this variable. Several gastrointestinal (GI) health variables were found to be significantly impacting the infant gut microbiota composition, in particular the number of defecations, and flatulence regularity.

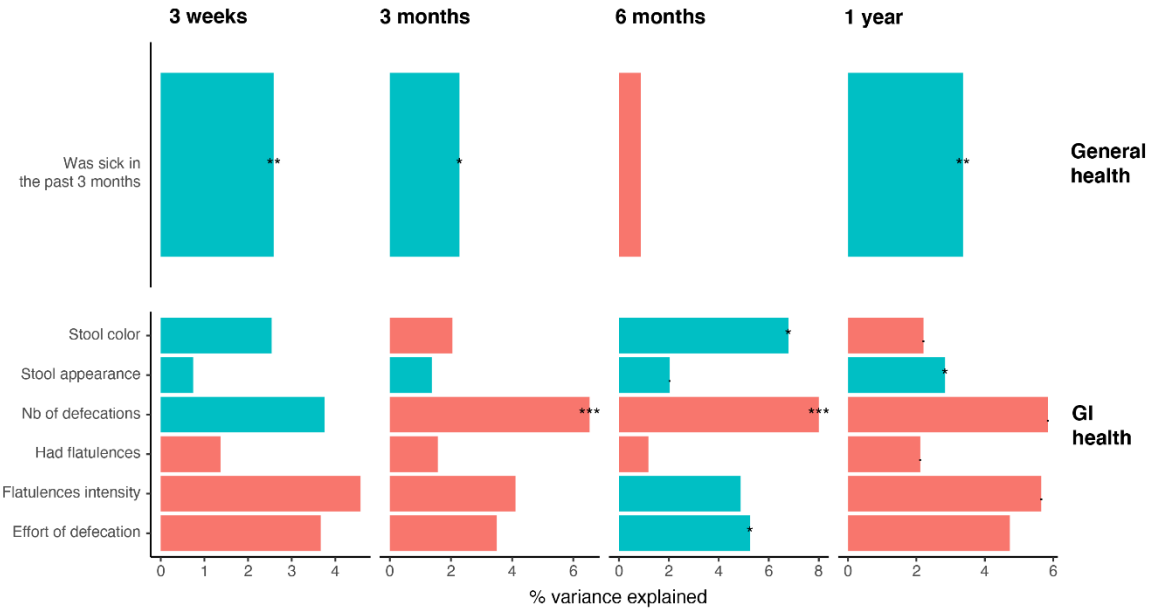


Figure 16. Significance and explained variance of the infant gut microbiota by health-related variables modelled with PERMANOVA test. *Results for variables with homogenous dispersion (Red) were deemed reliable. ***: $p < 0.0001$, **: $p < 0.001$, *: $p < 0.01$, .: $p < 0.1$ from adonis2 (PERMANOVA permutation=9999, cofounders: plate ID and total number of reads). Only variables with an FDR corrected q-values ≤ 0.25 were plotted.*

5. Discussion

5.1 Taxonomic annotation profiling methods

This study aimed to identify the technical and biological variables explaining the taxonomic variation observed in infant and parental faecal microbiota during the first year of life. We first explored the impact of the choice of taxonomic profiling methods, as they have been shown to have profound impacts on the profiling results (Mavromatis et al., 2007; Meyer et al., 2019; Ye et al., 2019). Here, we compared the annotation obtained using three computational tools: Metaphlan3, Kraken and Kraken along with Braken. Metaphlan3 is a taxonomic annotation tool widely used in metagenome research (published in 2021, cited 144 times as of date) (Beghini et al., 2021). On the other hand, Kraken is a taxonomic annotation tool using the sequence composition in *k*-mers to annotate the sequencing reads (published in 2019, cited 1148 times to date) (Wood et al., 2019). One particularity of Kraken is that reads may be classified at different taxonomic level (species, genus, family etc.), depending on if the sequence of the considered read is conserved between different taxa. Braken then uses a Bayesian inference to recalculate Kraken annotation and provide a more precise species-level profile (Lu et al., 2017). When comparing the different approaches, we noted that the gene-based approach used by Metaphlan3 masks the number of reads left unannotated by the tool (Beghini et al., 2021). In other words, this tool reports the relative shares of annotated taxa in relation to the total number of annotated reads, not the total reads. In our context, when comparing parent and infant metagenome profiles at different ages, it is important to assess how the fraction of sequences annotated by the tool is varying across sample types. Therefore, we determined that using Kraken along with Braken was most appropriate for this study, as it allowed us to (1) report the number of sequencing reads left un-annotated for each sample type and (2) easily adapt the reference database used by the tool.

Indeed, the reference database used by each annotation tools can have a major impact on the quality of annotation. The development of databases specifically tailored to an environment of interest allows for a better precision of annotation, as it limits the false positive hits. In our study we compared the annotation obtained by the general database RefSeq and the human gut specific HumGut database (Hiseni et al., 2021) using Kraken and Kraken+Braken. We observed that using

the HumGut database drastically reduces the number of reads left un-annotated by the tool. All in all, these methods and database comparison allowed us to determine that using Kraken+Braken along with the HumGut database was most appropriate for our study.

5.2 Technical factors

We next explored the impact of technical variables such as the sample collection, storage conditions, DNA extraction and sequencing methods on the overall microbiota variation. Indeed, in large-scale cohort such as the HELMi project, these variations may introduce significantly bias in microbiota composition that needs to be accounted for when studying biological variations (Voigt et al., 2015; Wesolowska-Andersen et al., 2014). In our dataset, few technical variables were found to have a significant impact on the observed microbiota composition. In particular, while storage time and conditions can introduce a significant bias on the microbiota profile (Cardona et al., 2012), we did not observe any significant impact of these variables on our dataset. However, we observed a significant impact of the identity of the person performing the DNA extraction, DNA extraction concentration and of the number of sequencing reads obtained. The samples were processed over a period of around 2 years by 3 different persons. This suggested that small differences in protocols may have significant impacts on the obtained microbiota composition. The DNA extraction concentration typically remains unaltered in repeated extraction for each sample, indicating that DNA yield is more a biological than technical variable of the samples. Sequencing depth is a known factor affecting richness and compositional outputs of the microbiota samples and hence typically normalized for samples by rarefaction or more recently with other methods available (McMurdie & Holmes, 2014). This exploration of the impacts of technical variations on the obtained microbiota profiles lead us to choose two variables as confounders for the rest of the analysis: (A) the DNA extraction plate ID, which accounts for variations in the extractor and the buffer batches and (B) the total number of reads obtained after sequencing, allowing to adjust for the DNA concentration and sequencing depth.

5.3 Biological factors from both environment and host

We explored the impact of biological variables on the infant faecal microbiota variations. As extensively reported in several other cohort studies, the mode of delivery has been observed to significantly alter the infant faecal microbiota composition during early infancy (Bokulich et al., 2016; Busi et al., 2021; Dominguez-Bello et al., 2010; Guittar et al., 2019; Reyman et al., 2019; Wampach et al., 2018). In our study, we observed a significant impact of the mode of delivery from 3 weeks to 1 year of age. This long-lasting impact on the early gut microbiota is thought to have impacts on the infant's future health. Indeed, children delivered by Caesarean section have an increased risk of developing asthma and obesity during childhood (Keag et al., 2018; Kuhle et al., 2015; Li et al., 2013).

We additionally observed a significant impact of the exposure to intrapartum antibiotics in our cohort. Mothers may be exposed to antibiotic treatment during labour and delivery to prepare for CS delivery, but also in cases of vaginal delivery due to carriage of neonatal pathogen Group B *Streptococcus* B (GBS) or suspected intrauterine infection. Previous studies have shown that exposures to several classes of intra-partum antibiotics can lead to a significantly less diverse early gut microbiota in the infant, and an altered acquisition of the gut microbiota (M. Azad et al., 2016; Coker et al., 2020; Korpela, Jokela, et al., 2021).

Aside from birth-mode related variables, we observed that several background and environmental factors such a as the parental education, number of pets and siblings as well as cleaning habits were found to have a significant impact on the infant microbiota composition. Impact of education level is a general proxy for socio-economic status, which can have a profound impact on lifestyle, living environment and parenting habits. In our cohort, the effect of number of pets and siblings are still inconclusive, as groups were also found to have a significant beta-dispersion. However, this result is of particular interest since several other studies also report these environmental variables to alter the infant microbiota composition and infant health (M. B. Azad et al., 2013; Laursen et al., 2015; Penders et al., 2014).

Surprisingly, we did not find any significant impact of the infant diet and breastfeeding variables on our cohort. This is surprising as several previous studies have shown that infant feeding pattern, in particular the length of breastfeeding and the age of introduction of solid foods, can impact the trajectory of gut microbiota acquisition (Bäckhed et al., 2015; Stewart et al., 2018). However, our dataset comprised of very few formula-fed infants, and 90% of the infants included in this study were breastfed until 6 months of age. This uniform group of infants in terms of feeding may explain why no significant association between feeding and infant microbiota variation was detected.

Finally, we explored the impacts of infant well-being and health on the gut microbiota composition. Recent sicknesses, as well as several variables regarding the GI health were found to have a significant association with variation in the infant microbiota composition. Importantly, some well-being variables have been linked to altered gut microbiota composition, such as colic (de Weerth et al., 2013), crying (Pärtty et al., 2012) and other GI symptoms (Korpela, Jokela, et al., 2021). In adults, the gut transit time is known to affect the stool microbiota composition (Asnicar et al., 2021), and the same phenomenon could be hypothesized for infants, explaining the impact of variables such as defecation rate and stool consistency. However, for these health-related variables, especially GI health and well-being variables, it can be difficult to assess what is the causal relationship between health and the microbiota composition.

5.4 Limitations and future directions

The present study explored the technical and biological variables impacting the composition of the early life microbiota during the first year of life. However, it is important to highlight that our study was limited by the uniformity in life-style and habits of the included families. Indeed, the HELMi cohort involves Finnish families living mostly in the Uusimaa region, globally Caucasian, highly educated and of a high socio-economic background. This lack of diversity may mask the effect of some variables that may be better addressed in other cohorts, using in particular a cross-sectional approach, although cohort heterogeneity is typically difficult to achieve, with some populations more difficult to enrol than others (Svensson et al., 2012). As previously noted, this

may explain why no association between breastfeeding habits and infant gut microbiota was found in this study.

The HELMi cohort variables used in this study have been obtained through the extensive questionnaires filled in by the parents. Importantly, some questions may have been misunderstood by some families, or base fully on subjective assessment (*e.g.*, parents' estimation of child well-being during the past months). We tried to reduce effects of subjective questions by creating categorical variables, and excluding questions with a high proportion of mistakes or impossible responses. Future studies focusing on a smaller subset of variables of interest may allow for a more curated approach which would increase the precision of the observed associations.

Finally, this study has focused on variance partitioning (*i.e.*, on detecting impacts of selected variables on the infant faecal microbiota). However, we did not investigate the impacts on gut microbiota richness and diversity nor on the microbiome functional composition and gene diversity, which will be investigated in future studies.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Acknowledgments

I am very grateful to Anne for giving me the opportunity to join this lab, and for the support and guidance for my study. I am so glad I can explore such a fascinating topic. It is a real pity that I couldn't communicate face-to-face with group members. But even if working remotely, I can feel that this is a group with a convivial atmosphere, and I am very thankful for the support they gave me.

I thank Alise for all the support she gave. Only under her guidance can I complete this thesis and improve it a lot. She has always been there to answer my question, to provide any help I needed and to encourage me when I was blue. She is more than a supervisor, but a concerned friend and a patient teacher to me. She is one of the best mentors I have ever met, and I will be grateful to her forever.

I thank Roosa for her constant support. I could not finish those complex data analyses without her technical support. She has taught me a lot and made it easier for me to understand the analysis process. She also gave me numerous valuable suggestions and comments for the entire thesis, especially the thesis writing. I am glad to have her as my friend through this project.

I also want to thank everyone in the lab who offered support and comments during this thesis. A special thank you to Emilia, who gave me a lot of advice at my lowest point.

Last but not least, I want to thank my roommate for pushing me to finish the painful paperwork. And I thank my family for giving me unconditional support. Their love gives me the courage to overcome all difficulties I have met.

References

Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., & Jenmalm, M. C. (2014). Low gut microbiota diversity in early infancy precedes asthma at school age. *Clinical & Experimental Allergy*, *44*(6), 842–850. https://doi.org/10.1111/cea.12253

Almonacid, D. E., Kraal, L., Ossandon, F. J., Budovskaya, Y. V., Cardenas, J. P., Bik, E. M., Goddard, A. D., Richman, J., & Apte, Z. S. (2017). 16S rRNA gene sequencing and healthy reference ranges for 28 clinically relevant microbial taxa from the human gut microbiome. *PLoS ONE*, *12*(5), e0176555. https://doi.org/10.1371/journal.pone.0176555

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrews, S. (2010). *FastQC: A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Asnicar, F., Leeming, E. R., Dimidi, E., Mazidi, M., Franks, P. W., Khatib, H. A., Valdes, A. M., Davies, R., Bakker, E., Francis, L., Chan, A., Gibson, R., Hadjigeorgiou, G., Wolf, J., Spector, T. D., Segata, N., & Berry, S. E. (2021). Blue poo: Impact of gut transit time on the gut microbiome using a novel marker. *Gut*, *70*(9), 1665–1674. https://doi.org/10.1136/gutjnl-2020-323877

Azad, M. B., Konya, T., Maughan, H., Guttman, D. S., Field, C. J., Sears, M. R., Becker, A. B., Scott, J. A., Kozyrskyj, A. L., & CHILD Study Investigators. (2013). Infant gut microbiota and the hygiene hypothesis of allergic disease: Impact of household pets and siblings on microbiota composition and diversity. *Allergy, Asthma & Clinical Immunology*, *9*(1), 15. https://doi.org/10.1186/1710-1492-9-15

Azad, M., Konya, T., Persaud, R., Guttman, D., Chari, R., Field, C., Sears, M., Mandhane, P., Turvey, S., Subbarao, P., Becker, A., Scott, J., Kozyrskyj, A., & Investigators, the C. S. (2016). Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: A prospective cohort study. *BJOG: An International Journal of Obstetrics & Gynaecology*, *123*(6), 983–993. https://doi.org/10.1111/1471-0528.13601

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M. T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y. S., Kotowska, D., Colding, C., … Wang, J. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe*, *17*(5), 690–703. https://doi.org/10.1016/j.chom.2015.04.004

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *ELife*, *10*, e65088. https://doi.org/10.7554/eLife.65088

Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., D. Lieber, A., Wu, F., Perez-Perez, G. I., Chen, Y., Schweizer, W., Zheng, X., Contreras, M., Dominguez-Bello, M. G., & Blaser, M. J. (2016).

Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine*, *8*(343), 343ra82-343ra82. https://doi.org/10.1126/scitranslmed.aad7121

Busi, S. B., de Nies, L., Habier, J., Wampach, L., Fritz, J. V., Heintz-Buschart, A., May, P., Halder, R., de Beaufort, C., & Wilmes, P. (2021). Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life. *ISME Communications*, *1*(1), 1–12. https://doi.org/10.1038/s43705-021-00003-5

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., Azpiroz, F., Roca, J., Guarner, F., & Manichanh, C. (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiology*, *12*(1), 158. https://doi.org/10.1186/1471-2180-12-158

Coker, M., Hoen, A., Dade, E., Lundgren, S., Li, Z., Wong, A., Zens, M., Palys, T., Morrison, H., Sogin, M., Baker, E., Karagas, M., & Madan, J. (2020). Specific class of intrapartum antibiotics relates to maturation of the infant gut microbiota: A prospective cohort study. *BJOG: An International Journal of Obstetrics & Gynaecology*, *127*(2), 217–227. https://doi.org/10.1111/1471-0528.15799

Cortes, L., Wopereis, H., Tartiere, A., Piquenot, J., Gouw, J. W., Tims, S., Knol, J., & Chelsky, D. (2019). Metaproteomic and 16S rRNA Gene Sequencing Analysis of the Infant Fecal Microbiome. *International Journal of Molecular Sciences*, *20*(6), 1430. https://doi.org/10.3390/ijms20061430

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., Shakya, M., Podar, M., Quince, C., & Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*, *17*(1), 55. https://doi.org/10.1186/s12864-015-2194-9

de Weerth, C., Fuentes, S., Puylaert, P., & de Vos, W. M. (2013). Intestinal Microbiota of Infants With Colic: Development and Specific Signatures. *Pediatrics*, *131*(2), e550–e558. https://doi.org/10.1542/peds.2012-1449

Derrien, M., Alvarez, A.-S., & Vos, W. M. de. (2019). The Gut Microbiota in the First Decade of Life. *Trends in Microbiology*, *27*(12), 997–1010. https://doi.org/10.1016/j.tim.2019.08.001

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, *107*(26), 11971–11975. https://doi.org/10.1073/pnas.1002601107

Forbes, J. D., Azad, M. B., Vehling, L., Tun, H. M., Konya, T. B., Guttman, D. S., Field, C. J., Lefebvre, D., Sears, M. R., Becker, A. B., Mandhane, P. J., Turvey, S. E., Moraes, T. J., Subbarao, P., Scott, J. A., Kozyrskyj, A. L., & for the Canadian Healthy Infant Longitudinal Development (CHILD) Study Investigators. (2018). Association of Exposure to Formula in the Hospital and Subsequent Infant Feeding Practices With Gut Microbiota and Risk of Overweight in the First Year of Life. *JAMA Pediatrics*, *172*(7), e181161. https://doi.org/10.1001/jamapediatrics.2018.1161

Guittar, J., Shade, A., & Litchman, E. (2019). Trait-based community assembly and succession of the infant gut microbiome. *Nature Communications*, *10*(1), 512. https://doi.org/10.1038/s41467-019-08377-w

Hiseni, P., Rudi, K., Wilson, R. C., Hegge, F. T., & Snipen, L. (2021). HumGut: A comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome*, *9*(1), 165. https://doi.org/10.1186/s40168-021-01114-w

Ihekweazu, F. D., & Versalovic, J. (2018). Development of the Pediatric Gut Microbiome: Impact on Health and Disease. *The American Journal of the Medical Sciences*, *356*(5), 413–423. https://doi.org/10.1016/j.amjms.2018.08.005

Keag, O. E., Norman, J. E., & Stock, S. J. (2018). Long-term risks and benefits associated with cesarean delivery for mother, baby, and subsequent pregnancies: Systematic review and meta-analysis. *PLOS Medicine*, *15*(1), e1002494. https://doi.org/10.1371/journal.pmed.1002494

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Korpela, K., & de Vos, W. M. (2018). Early life colonization of the human gut: Microbes matter everywhere. *Current Opinion in Microbiology*, *44*, 70–78. https://doi.org/10.1016/j.mib.2018.06.003

Korpela, K., Dikareva, E., Hanski, E., Kolho, K.-L., Vos, W. M. de, & Salonen, A. (2019). Cohort profile: Finnish Health and Early Life Microbiota (HELMi) longitudinal birth cohort. *BMJ Open*, *9*(6), e028500. https://doi.org/10.1136/bmjopen-2018-028500

Korpela, K., Helve, O., Kolho, K.-L., Saisto, T., Skogberg, K., Dikareva, E., Stefanovic, V., Salonen, A., Andersson, S., & Vos, W. M. de. (2020). Maternal Fecal Microbiota Transplantation in Cesarean-Born Infants Rapidly Restores Normal Gut Microbial Development: A Proof-of-Concept Study. *Cell*, *183*(2), 324-334.e5. https://doi.org/10.1016/j.cell.2020.08.047

Korpela, K., Jokela, R., Jian, C., Dikareva, E., Nikkonen, A., Saisto, T., Skogberg, K., Vos, W. M. de, Kolho, K.-L., & Salonen, A. (2021). *Quantitative insights into effects of intrapartum antibiotics and birth mode on infant gut microbiota in relation to well-being during the first year of life* (p. 2021.11.01.21265735). medRxiv. https://doi.org/10.1101/2021.11.01.21265735

Korpela, K., Kallio, S., Salonen, A., Hero, M., Kukkonen, A. K., Miettinen, P. J., Savilahti, E., Kohva, E., Kariola, L., Suutela, M., Tarkkanen, A., de Vos, W. M., Raivio, T., & Kuitunen, M. (2021). Gut microbiota develop towards an adult profile in a sex-specific manner during puberty. *Scientific Reports*, *11*(1), 23297. https://doi.org/10.1038/s41598-021-02375-z

Krueger, F. (2022). *Trim Galore* [Perl]. https://github.com/FelixKrueger/TrimGalore (Original work published 2016)

Kuhle, S., Tong, O. S., & Woolcott, C. G. (2015). Association between caesarean section and childhood obesity: A systematic review and meta-analysis. *Obesity Reviews*, *16*(4), 295–303. https://doi.org/10.1111/obr.12267

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Laursen, M. F., Zachariassen, G., Bahl, M. I., Bergström, A., Høst, A., Michaelsen, K. F., & Licht, T. R. (2015). Having older siblings is associated with gut microbiota development during early childhood. *BMC Microbiology*, *15*(1), 154. https://doi.org/10.1186/s12866-015-0477-6

Li, H. -t, Zhou, Y. -b, & Liu, J. -m. (2013). The impact of cesarean section on offspring overweight and obesity: A systematic review and meta-analysis. *International Journal of Obesity*, *37*(7), 893–899. https://doi.org/10.1038/ijo.2012.195

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, *3*, e104. https://doi.org/10.7717/peerj-cs.104

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., & Kyrpides, N. C. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, *4*(6), 495–500. https://doi.org/10.1038/nmeth1043

McMurdie, P. J., & Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*(1), 11257. https://doi.org/10.1038/ncomms11257

Meyer, F., Bremges, A., Belmann, P., Janssen, S., McHardy, A. C., & Koslicki, D. (2019). Assessing taxonomic metagenome profilers with OPAL. *Genome Biology*, *20*(1), 51. https://doi.org/10.1186/s13059-019-1646-y

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M. D., Durand, S., … Weedon, J. (2022). *vegan: Community Ecology Package* (2.6-2) [Computer software]. https://CRAN.R-project.org/package=vegan

Pärtty, A., Kalliomäki, M., Endo, A., Salminen, S., & Isolauri, E. (2012). Compositional Development of Bifidobacterium and Lactobacillus Microbiota Is Linked with Crying and Fussing in Early Infancy. *PLOS ONE*, *7*(3), e32495. https://doi.org/10.1371/journal.pone.0032495

Penders, J., Gerhold, K., Thijs, C., Zimmermann, K., Wahn, U., Lau, S., & Hamelmann, E. (2014). New insights into the hygiene hypothesis in allergic diseases. *Gut Microbes*, *5*(2), 239–244. https://doi.org/10.4161/gmic.27905

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(suppl_1), D501–D504. https://doi.org/10.1093/nar/gki025

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., … Wang, J. (2010). A human gut microbial gene catalog established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. https://doi.org/10.1038/nature08821

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967–977. https://doi.org/10.1016/j.bbrc.2015.12.083

Reyman, M., van Houten, M. A., van Baarle, D., Bosch, A. A. T. M., Man, W. H., Chu, M. L. J. N., Arp, K., Watson, R. L., Sanders, E. A. M., Fuentes, S., & Bogaert, D. (2019). Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life. *Nature Communications*, *10*(1), 4997. https://doi.org/10.1038/s41467-019-13014-7

Romano-Keeler, J., Moore, D. J., Wang, C., Brucker, R. M., Fonnesbeck, C., Slaughter, J. C., Li, H., Curran, D. P., Meng, S., Correa, H., Lovvorn III, H. N., Tang, Y.-W., Bordenstein, S., George Jr, A. L., & Weitkamp, J.-H. (2014). Early life establishment of site-specific microbial communities in the gut. *Gut Microbes*, *5*(2), 192–201. https://doi.org/10.4161/gmic.28442

Shao, Y., Forster, S. C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., Kumar, N., Stares, M. D., Rodger, A., Brocklehurst, P., Field, N., & Lawley, T. D. (2019). Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature*, *574*(7776), 117–121. https://doi.org/10.1038/s41586-019-1560-1

Stanislawski, M. A., Dabelea, D., Wagner, B. D., Iszatt, N., Dahl, C., Sontag, M. K., Knight, R., Lozupone, C. A., & Eggesbø, M. (n.d.). Gut Microbiota in the First 2 Years of Life and the Association with Body Mass Index at Age 12 in a Norwegian Birth Cohort. *MBio*, *9*(5), e01751-18. https://doi.org/10.1128/mBio.01751-18

Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., Ross, M. C., Lloyd, R. E., Doddapaneni, H., Metcalf, G. A., Muzny, D., Gibbs, R. A., Vatanen, T., Huttenhower, C., Xavier, R. J., Rewers, M., Hagopian, W., Toppari, J., Ziegler, A.-G., … Petrosino, J. F. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*, *562*(7728), 583–588. https://doi.org/10.1038/s41586-018-0617-x

Stiemsma, L. T., & Michels, K. B. (2018). The Role of the Microbiome in the Developmental Origins of Health and Disease. *Pediatrics*, *141*(4), e20172437. https://doi.org/10.1542/peds.2017-2437

Svensson, K., Ramírez, O. F., Peres, F., Barnett, M., & Claudio, L. (2012). Socioeconomic determinants associated with willingness to participate in medical research among a diverse population. *Contemporary Clinical Trials*, *33*(6), 1197–1205. https://doi.org/10.1016/j.cct.2012.07.014

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *12*(10), 902–903. https://doi.org/10.1038/nmeth.3589

Voigt, A. Y., Costea, P. I., Kultima, J. R., Li, S. S., Zeller, G., Sunagawa, S., & Bork, P. (2015). Temporal and technical variability of human gut metagenomes. *Genome Biology*, *16*(1), 73. https://doi.org/10.1186/s13059-015-0639-8

Wampach, L., Heintz-Buschart, A., Fritz, J. V., Ramiro-Garcia, J., Habier, J., Herold, M., Narayanasamy, S., Kaysen, A., Hogan, A. H., Bindl, L., Bottu, J., Halder, R., Sjöqvist, C., May, P., Andersson, A. F., de Beaufort, C., & Wilmes, P. (2018). Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nature Communications*, *9*(1), 5091. https://doi.org/10.1038/s41467-018-07631-x

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. https://doi.org/10.1186/s40168-017-0237-y

Wesolowska-Andersen, A., Bahl, M. I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R., & Licht, T. R. (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome*, *2*(1), 19. https://doi.org/10.1186/2049-2618-2-19

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

Wilson, B. C., Butler, É. M., Grigg, C. P., Derraik, J. G. B., Chiavaroli, V., Walker, N., Thampi, S., Creagh, C., Reynolds, A. J., Vatanen, T., O'Sullivan, J. M., & Cutfield, W. S. (2021). Oral administration of maternal vaginal microbes at birth to restore gut microbiome development in infants born by caesarean section: A pilot randomised placebo-controlled trial. *EBioMedicine*, *69*. https://doi.org/10.1016/j.ebiom.2021.103443

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A.-M., Härkönen, T., Ryhänen, S. J., Franzosa, E. A., Vlamakis, H., Huttenhower, C., Gevers, D., Lander, E. S., Knip, M., on behalf of the DIABIMMUNE Study Group, & Xavier, R. J. (2016). Natural history of the infant gut microbiome and impact of antibiotic

treatment on bacterial strain diversity and stability. *Science Translational Medicine*, *8*(343), 343ra81-343ra81. https://doi.org/10.1126/scitranslmed.aad0917

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Supplementary Material

Table 1. Overview of the variables used for PERMANOVA test.

Variable short name	Variable description	Section	Type
inf_AgeHome	Age in days of the child when they got home after delivery	Background	Numeric
inf_BirthHeight	Infant height at birth	Background	Numeric
inf_BirthWeight	Infant weight at birth	Background	Numeric
inf_DeliveryMode	Birth mode (Vaginal or C-Section)	Background	Categorical
inf_HadDepartureDelayed	The departure from the maternity hospital was delayed for health reasons	Background	Boolean
inf_HadIntensiveCare	The infant was placed in intensive care after birth	Background	Boolean
inf_HospitalExtraMilk	The infant received extra milk at the hospital	Background	Boolean
inf_Sex	Infant sex at birth	Background	Categorical
reg_RecievingIAP	Intrapartum antibiotic was given at delivery	Background	Boolean
inf_Gestational_term	Infant gestational age	Background	Categorical
reg_WaterBreakH	Time between water break and delivery	Background	Numeric
birth_season	Season of Birth	Background	Categorical
env_CleanScore_sum	Global score computed from house cleaning habits	Background	Numeric
env_HabitationType	Type of habitation	Background	Categorical
env_NbPets	Number of pets living with the family	Background	Numeric
env_FurOrFeathers	The family has a furry/feather pet, other or no pet	Background	Categorical
familly_NbSiblingsFt	Number of full-time siblings	Background	Numeric
family_MaxEducation	Maximum education level in the family	Background	Numeric
f_AllergyDiseaseType_Allergy	The biological father had a diagnosed allergic disease	Background	Boolean
f_HasHeritableDisease	The biological father has a diagnosed heritable disease	Background	Boolean
f_height	Biological father height at delivery time	Background	Numeric
f_weight	Biological father weight at delivery time	Background	Numeric
f_BMI	Biological father BMI at delivery time	Background	Numeric
p_AgeDelivery	Partner's age at delivery	Background	Numeric
p_educationScore	Partner's education level	Background	Categorical
m_ActivityLeisure	Mother's typical physical activity during pregnancy	Background	Categorical
m_AgeDelivery	Mother's age at delivery	Background	Numeric
m_AlcoholPriorFreq	Mother's alcohol doses prior pregnancy	Background	Categorical
m_AllergyDisease-Type_Allergy	The mother has a diagnosed allergic disease	Background	Boolean
m_FattyAcid	Fatty acid and oil fish supplements used during pregnancy by the mother	Background	Boolean

Variable short name	Variable description	Section	Type
m_FolicAcid	Folic acid supplements used during pregnancy by the mother	Background	Boolean
m_HasHeritable-Disease	The mother has a diagnosed heritable disease	Background	Boolean
m_height	Mother's height at delivery time	Background	Numeric
m_OutdoorActivities-Score	Global score for the mother's outdoor activity hobbies	Background	Numeric
m_PreviousDeliveries	Mother's number of previous deliveries	Background	Numeric
m_ProbioticsUse	Usage of lactic acid bacteria and other probiotic during last trimester of pregnancy	Background	Boolean
m_BMI	Mother BMI before pregnancy	Background	Numeric
m_weightGain	Mother's weight gain during pregnancy	Background	Numeric
m_educationScore	Mother education level score	Background	Categorical
all_reads	Number of reads obtained after sequencing	Technical	Numeric
Coll_Antibiotic-Treatment	For parents, when was the last antibiotic treatment course before sample collection	Technical	Categorical
Ext_FecalWeight	Faecal weight in gram used for DNA extraction	Technical	Numeric
Ext_BristolScore	Bristol score of the faecal sample during DNA extraction	Technical	Categorical
Ext_DNAconc	DNA concentration obtained by DNA extraction	Technical	Numeric
Ext_Extractor	Person who did the DNA extraction	Technical	Categorical
Ext_Plate	Batch ID of DNA extraction	Technical	Categorical
Run_ID	Sequencing run	Technical	Categorical
Storage_AtHome_ Days	Number of days of sample storage at home (-20°C)	Technical	Numeric
Storage_Lab_Days	Numbers of days of sample storage in the lab before extraction (-80°C)	Technical	Numeric
ExclusiveBF_Retro	The infant was exclusively breastfed at least until the sample collection	Breastfeeding	Categorical
inf_SolidQuant	Quantity of solid food the infant received during the 2 weeks prior sample collection	Breastfeeding	Categorical
inf_AgeFirstSolids	Infant age when received solid food for the first time	Breastfeeding	Categorical
inf_MainFood	Main food received by the infant during the month prior sample collection	Breastfeeding	Categorical
inf_TypeMilk_BreastMilk	The infant received breast milk during the month prior sample collection	Breastfeeding	Boolean
inf_TypeMilk_ FormulaMilk	The infant received formula milk during the month prior sample collection	Breastfeeding	Boolean
inf_TypeMilk_ CowMilk	The infant received cow milk during the month prior sample collection	Breastfeeding	Boolean
Had_treatment	Has received any treatment course before the collection of the sample	Medication	Boolean
inf_WasSick	The baby has been sick at the time of the sample collection	Health	Boolean

Variable short name	Variable description	Section	Type
inf_StoolAppearance	Typical appearance of the infant stool in the weeks before sample collection	Health	Categorical
inf_StoolColor	Typical colour of the stool in the weeks before sample collection	Health	Categorical
inf_HadGI	The child presented signs of GI symptoms in the weeks before sample collection	Health	Boolean
inf_HadPain	The child presented signs of stomach pain in the weeks before sample collection	Health	Boolean
inf_HadFlat	The child presented signs of flatulence in the weeks before sample collection	Health	Boolean
inf_Crying_cat	Estimate of the total amount of baby crying per day in the weeks before sample collection	Health	Categorical
inf_IntensityCrying_cat	Intensity of crying episodes in the weeks before sample collection	Health	Categorical
inf_Soothing_cat	The effect of soothing on the baby's crying episodes in the weeks before sample collection	Health	Categorical
inf_NbDef_cat	Infant defecation frequency in the weeks before sample collection	Health	Categorical
inf_DefEffort_cat	Effort necessary for the infant to defecate in the weeks before sample collection	Health	Categorical
inf_PainIntensity_cat	Intensity of the stomach pain in the weeks before sample collection	Health	Categorical
inf_FlatIntensity_cat	Intensity of the flatulence pain in the weeks before sample collection	Health	Categorical