MDP in Digital Health and Life Sciences

Thesis

The influence of maternal prenatal stress on the infant's gut microbiota development

Student: Anastasia Karavaeva

Supervisor: Leo Lahti

Date: 06/07/2020

Acknowledgements

First of all, I would like to thank my primary supervisor Leo Lahti, Associate Professor in Data Science at the University of Turku. His exemplary guidance, active participation and patience with me throughout the thesis made learning and applying of complex statistical methods possible.

I wish to also express my gratitude to the FinnBrain group and the group`s Principal Researcher and Adjunct Professor, Hasse Karlsson and Linnea Karlsson, respectively, for providing me with the infant gut microbiota cohort data, office space and further assistance.

In particular, the biological and bioinformatics knowledge provided by my secondary thesis supervisor, Anna Aatsinki, a PhD student in FinnBrain, was instrumental in me developing an interdisciplinary view on the thesis, from the perspective of bioinformatics, epidemiology and biology. I would also like to thank Paula Mustonen, a PhD student in FinnBrain, for giving me insight into the associations between hair cortisol concentrations and maternal variables.

Table of Contents

1. Introduction

1.1 Maternal psychological distress and its effect on the infant

Early life stress (ELS) is known to have adverse, long lasting consequences on an infant's development, capable of persisting into adulthood. ELS initially manifests as maternal prenatal psychological distress (PD) during the prenatal period. Maternal prenatal PD is defined as distress in an expectant mother that is caused by depression, anxiety, major life events or environmental hardships. Exposure to maternal prenatal PD may predispose infants to abnormal behavioural, emotional and cognitive development [1]. Other gestational stressors, such as infection, obesity, hypoxia and malnourishment, also may result in altered brain development. The child's behavioural and emotional development may be affected, for example, by their attention span and reaction to stress. In regard to cognitive development, spatial learning and hippocampal plasticity may be impaired [2]. Maternal prenatal PD is of particular interest to researchers because it could potentially account for the 17% of variance seen in childhood cognitive ability and for doubling the prevalence of child psychiatric disorders [3].

One of the main mechanisms by which maternal prenatal PD affects the infant's neurodevelopment is the altered maternal hypothalamus-pituitary-adrenal (HPA) axis functioning. The maternal HPA axis activates in response to stress and produces cortisol, a glucocorticoid hormone, as the end product [1]. Briefly, the hypothalamus reacts to stress by secreting two hormones - corticotropin-releasing hormone (CRH) and arginine vasopressin (AVP). Both hormones are then released into the anterior pituitary gland, where they trigger the release of the adrenocorticotropic hormone (ACTH). In turn, ACTH stimulates the synthesis and secretion of cortisol in the adrenal glands (Figure 1) [4]. Exposure to maternal prenatal PD ultimately increases the child's hypothalamic–pituitary–adrenal (HPA) axis sensitivity via the altered maternal HPA axis, therefore negatively impacting the infant's

reactivity to stress and making them more susceptible to psychiatric disorders [1, 5]. Other mechanisms mediating the effects of maternal prenatal PD on the infant include changes in the mother's and infant's autonomic nervous system, gut microbiota, immune system, and the infant's telomere length. The mother's diet, fitness and other lifestyle choices also influence maternal prenatal PD [6].

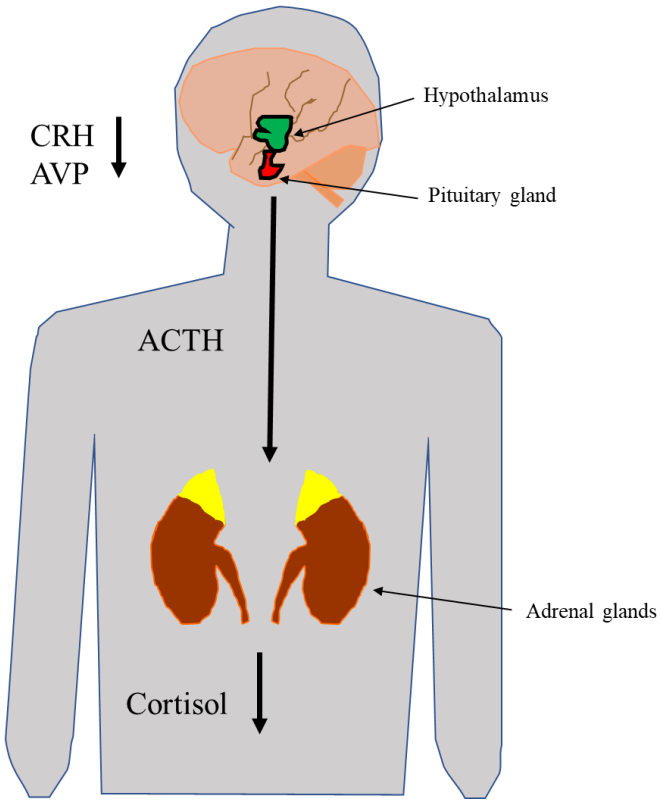


Figure 1. Components of the HPA axis.

The hypothalamus secretes CRH and AVP, which are then released into the anterior pituitary gland, where they trigger the release of ACTH. In turn, ACTH stimulates the synthesis and secretion of cortisol in the adrenal glands [4].

1.2 Role of maternal psychological distress and cortisol concentrations as biomarkers

Cortisol, an end product of the HPA stress response system, is perhaps responsible for imparting the effects of maternal prenatal PD on the fetus - via altered cortisol signalling patterns. High maternal prenatal cortisol concentration has been observed to be associated with impaired neurodevelopment in the infant, with motor development, cognitive development, regional brain volumes being affected. Maternal prenatal PD can be assessed by maternal prenatal cortisol concentrations and self-reported maternal prenatal PD. Maternal prenatal cortisol concentrations have traditionally been measured using the maternal saliva, blood or urine samples, all which have in studies been linked with altered neurodevelopment in the infant [3]. On the other hand, self-reported maternal prenatal PD is measured in expectant mothers using a variety of questionnaires that target distinct components of the stress response, such as depressive symptoms, overall anxiety, and pregnancy-specific anxiety [1,3].

Associations between maternal prenatal cortisol concentrations and self-reported maternal prenatal PD have been weak or inconsistently significant in prior studies [3]. For instance, maternal salivary cortisol measurements taken between 24 and 38 weeks gestation, or at earlier gestation points in other related studies, didn't correspond with self-reported maternal prenatal PD [7]; whereas in another study, maternal morning salivary cortisol in late pregnancy was significantly inversely associated with positive life events [8].

Assessing cortisol levels by short-term measurements, such as saliva, plasma or urine, is problematic for several reasons: daytime and seasonal fluctuations in the circadian clock and homeostatic regulatory mechanisms can cause high variability in baseline cortisol levels between and within subjects; multiple sampling is required for short-term measurements, which often isn't performed enough; and maternal HPA axis functioning changes during pregnancy [9,3]. To explain, maternal cortisol levels normally increase at the end of pregnancy since they're essential for the maturation of organs in the fetus and initiating childbirth, consequently making cortisol concentration readings at separate time points non-

generalizable to other trimesters. Assessment of cortisol levels by hair cortisol concentration (HCC) has gained popularity as an alternative sampling method, since it addresses some limitations of the previously mentioned methods. Cortisol accumulates into hair, with hair on average growing by one centimetre per month, hence cortisol in hair segments of a selected length represent the mean levels of cortisol during the corresponding months. A single HCC measure is enough to assess cumulative, long-term cortisol levels in a non-invasive manner [3].

1.3 Maternal and infant gut microbiota as stress-mediating mechanisms

One of the main issues is that the existing association between maternal prenatal PD and altered child neurodevelopment isn't completely understood. Composition of the gut microbiota could potentially serve as one of the underlying stress-mediating mechanisms [10].

Gut bacteria start to colonize the infant's gut as early as in the gestational period but expand greatly during delivery and the first months of life. Gut microbiota are involved in the maturation of the hypothalamic pituitary-adrenal system in infants, metabolism of host nutrients and drugs, immunomodulation, protecting against pathogens, production of vitamins and bioactive compounds, etc [2]. Gut microbiota metabolizes host nutrients via hydrolyzation and fermentation of complex, indigestible polysaccharides into simpler products. For instance, gut microbiota produces compounds such as short chain fatty acids and the neurotransmitters serotonin and gamma-aminobutyric acid (GABA). The microbial diversity of the gut has been linked to the health of the gut ecosystem, with high diversity indicating a healthy gut ecosystem and low diversity being associated with conditions such as obesity, inflammatory bowel disease. Therefore, gut microbiota plays an essential role in human health and disease, including in the development of the stress response in infants [2].

There are at least three pathways by which maternal prenatal cortisol concentrations could affect the infant gut microbiota. Cortisol has several functions such as controlling bile acid production, gut motility, cholesterol and bile acid homeostasis. Firstly, high maternal cortisol levels may cause increased bile acid production, which would hinder the normal development of the maternal gut microbiota during pregnancy, hence potentially impacting the transmission of maternal gut microbiota to the infant at birth. Secondly, maternal cortisol can pass through the placenta, if the placental 11beta-HSD2 is downregulated, and increase the fetal cortisol concentrations, which could alter development of the fetal HPA axis, causing higher base cortisol concentrations and cortisol sensitivity later in life. The higher infant cortisol concentrations would alter the permeability of the gut, disturb the gut barrier, and change the gut microbiota composition. Thirdly, mothers with high prenatal cortisol concentrations could potentially, through their breast milk, transfer cortisol to the infant in the postnatal period, therefore affecting the infant's HPA axis, gut permeability, and gut microbiota composition (Figure 2) [11].

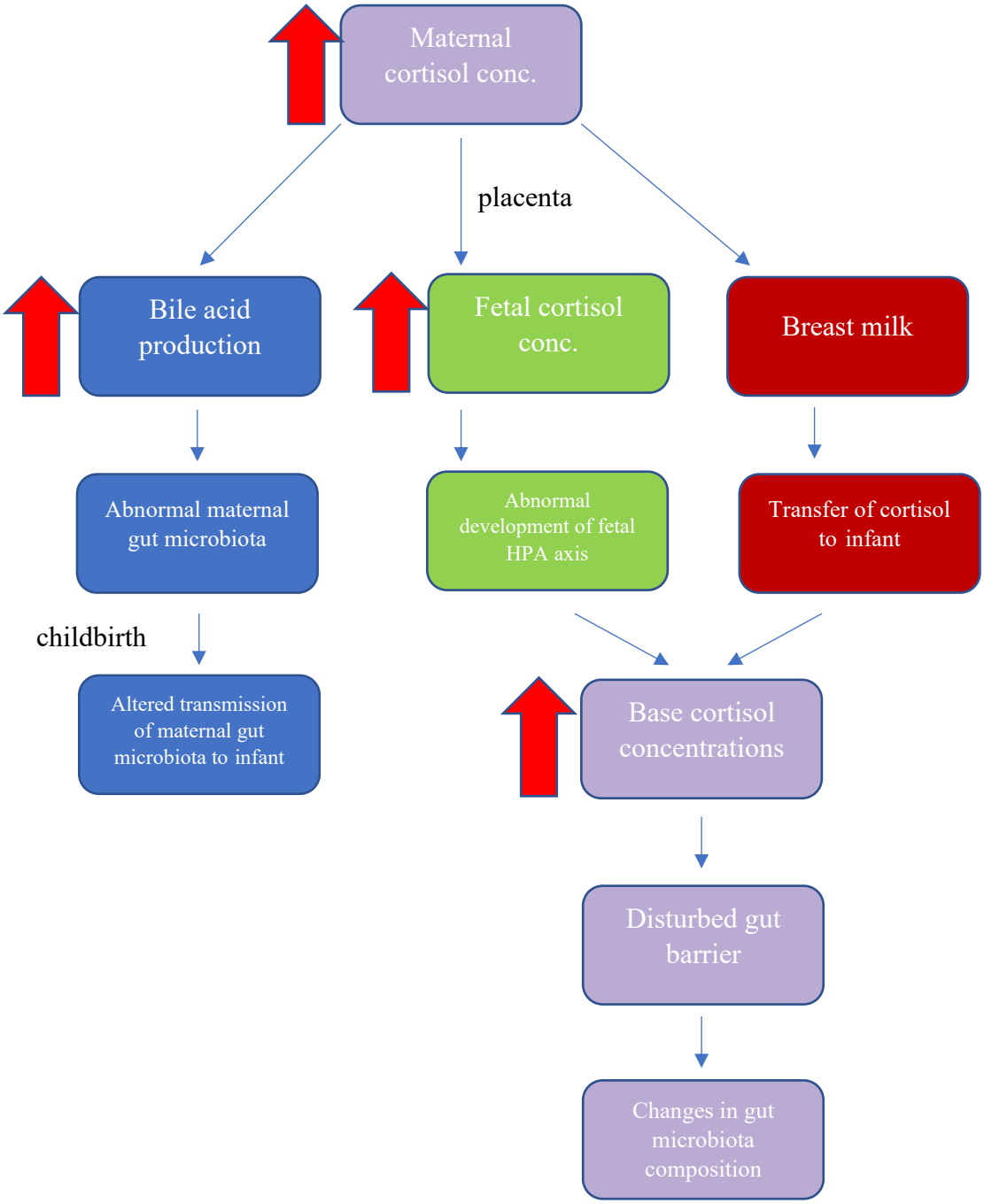


Figure 2. Three pathways by which maternal prenatal PD can affect the infant gut microbiota. High maternal cortisol concentrations may cause increased bile acid production that alters the mother's gut microbiota, can increase the fetal cortisol concentrations by passing through the placenta, or can transfer to the infant through breast milk during the postnatal period. As a result, development of the infant gut microbiota is altered [11].

1.4 Maternal prenatal PD and cortisol concentrations associate with infant gut microbiota

Previous animal and human studies have shown that either or both reported maternal prenatal PD and maternal salivary cortisol concentrations are associated with the infants' gut microbiota composition. Rhesus monkey infants of high-stress mothers had lower abundances of *Bifidobacteria* and *Lactobacilli* than infants whose mothers were not stressed during pregnancy [2,11]. In humans, infants whose mothers had both high reported prenatal PD and high cortisol concentrations during pregnancy had higher relative abundances of the phylum group Proteobacteria, including members such as *Escherichia*, *Serratia*, and *Enterobacter,* and lower relative abundances of lactic acid bacteria, such as *Lactobacillus, Lactoccus, Aerococcus,* and *Bifidobacteria*. Genus level groups like *Escherichia*, *Serratia*, and *Enterobacter* contain species that are potentially capable of causing infections and harbor Lipopolysaccharide (LPS) in their outer membrane, which is an inflammatory endotoxin that has been linked in inflammation in metabolic diseases and regulating stress responses; while *Bifidobacterium* and *Lactobacillus* have been found to be associated with healthy microbiota development in children [11].

1.5 Statistical methods in gut microbiota analysis

Microbiota abundance data is typically organized into a table containing operational taxonomic units – also referred to as OTU tables, where columns represent samples and the rows depict observed counts of clustered sequence reads [12]. Microbial abundances are either calculated using small-subunit ribosomal RNA (16SrRNA) gene sequences of each species, which serve as sufficient proxies for the microbes' full-length sequences, or the entire community DNA via shotgun sequencing. While the former can be used to recover whole genomes and to assess the functional potential of the microbial community, determining abundances with 16SrRNA sequences is much cheaper [13]. OTU tables are often normalized or transformed before conducting any downstream analyses [12].

Statistical considerations - microbiota samples have different library sizes, and hence can't be compared to each other without first normalizing. The two popular approaches are either to rarefy counts or to transform absolute abundancies into compositional data. Rarefaction of counts involves the selection of a minimum library sample, and then random subsampling without replacement of the remaining libraries so that each sample has an equal number of sequences. Transforming absolute abundances into compositional data results in relative abundances that are non-negative and sum to 1 within a sample [12]. After running multivariate analyses, some p-values will be less than the significance level entirely by chance and hence will be false positive, even though the null hypothesis is true. The null hypothesis is rejected and the alternative hypothesis is accepted if a p-value is less than the significance level, for instance 0.05. Correction for multiplicity can be done by methods such as Benjamini–Hochberg, which controls for the false discovery rate (FDR) [14].

Alpha-diversity (α-diversity) - is defined as the variation in species identity and abundances within a sample. Various alpha-diversity metrics hold different views on true diversity and perform differently [15]. These metrics can be either qualitative - also referred to as richness metrics - or quantitative, analysing presence-absence data or relative abundance data, respectively [13]. Furthermore, while traditional diversity metrics consider species to be equally different from one another, some diversity metrics have expanded to utilize extra information such as phylogenetic, functional, and other differences among species [16]. Chao1 index is an example of a qualitative metric that uses abundances to estimate species richness; it is based on the idea that rare species can deduce the most information about the number of missing species, consequently Chao1 index gives more weight to species with low abundances. Shannon index is an example of a quantitative metric that estimates both species richness and species evenness, in other words, how close in numbers are the different species' relative abundances within a sample [17].

Multiple linear regression models can test the association between the alpha-diversity values and a host trait, while also adjusting for host covariates [15]. The continuous dependent variable, in this case alpha-diversity, will be to some extent predicted by the independent variables: host trait and host covariates. The line of best fit in a regression models partially

explains the variance in a dependent variable as well as the relationship between a dependent variable and independent variables. The regression models can be evaluated by statistics such as R-squared, F-test, and t-test. R-squared is a measure of the percentage of variation in a dependent variable that is explained by the model. However, adding more independent variables to the model always increases the R-squared statistic. Its extension, adjusted R-squared, takes the number of independent variables into consideration and doesn't necessarily increase with the addition of a new independent variable [18]. The F-test evaluates multiple independent variables simultaneously to assess whether the regression model provides a better fit, therefore is significant, compared to a model with no independent variables [19]. On the other hand, the t-test tests the significance of individual independent variables [20].

Beta-diversity (β-diversity) - is defined as the variation in species identity and abundances between samples; it is measured by pairwise sample-to-sample distances based on either presence-absence data or relative abundance data and is organized into a square distance matrix (Figure 3). The distance matrix is constructed using beta-diversity metrics, such as the Jaccard and Bray-Curtis indices that have been adopted in microbial ecology. Jaccard index relies on the presence/absence of OTUs to check for similarity between the sample sets; it's calculated as the size of intersection, divided by the size of the union of the sample sets [21]. In contrast, the Bray-Curtis index takes OTU abundances into account and estimates the dissimilarity between the sample sets; it's calculated as the sum of the minimum counts for species in common between the samples sets, divided by the total amount of counts for all species present in both sample sets [22].

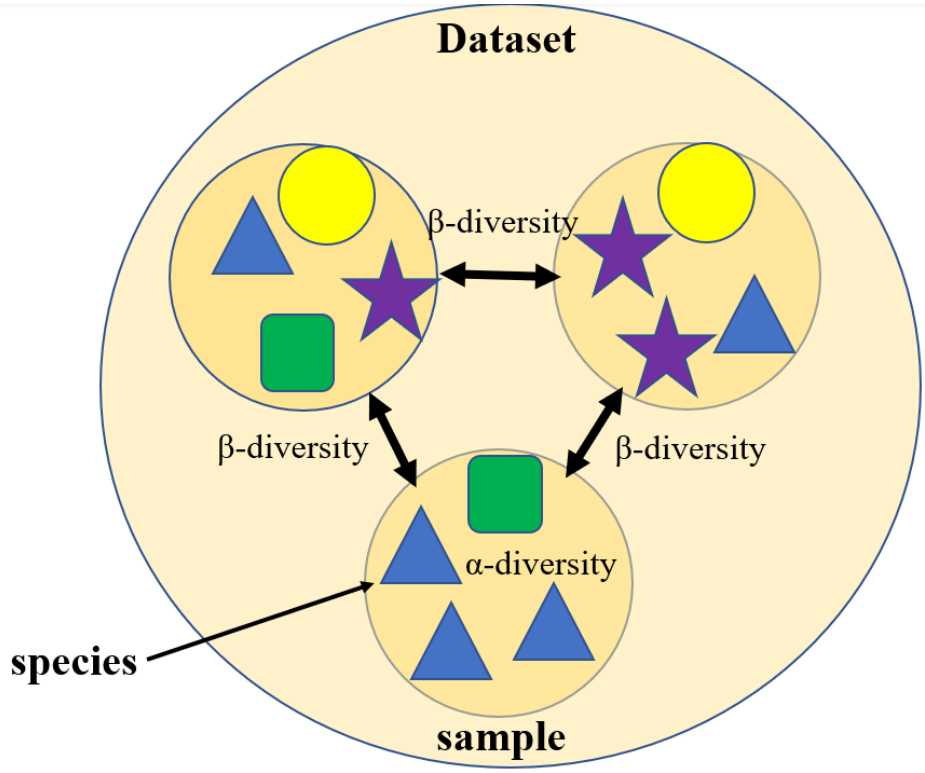


Figure 3. Measurements of biodiversity: α-diversity and β-diversity. Alpha-diversity (α-diversity) is defined as the variation in species identity and abundances within a sample [15], while beta-diversity (β-diversity) is defined as the variation in species identity and abundances between samples [21].

Multiple linear regression analysis and many other multivariate analyses aren't appropriate for testing the association between multiple dependent variables, in this case the beta-diversity distance matrix, and a host trait and its covariates. For example, Parametric multivariate analysis of variance (MANOVA) assumes multivariate normality and homogeneity of the distance matrix, but the presence of many zeros in the beta-diversity distance matrix, due to rare species, violates the normality assumption. Furthermore, MANOVA can't handle data sets containing more variables than replicates, yet it's common for ecological data to have more species than replicates [23]. As a solution, a nonparametric, distribution free multivariate analysis called permutational MANOVA (PERMANOVA) can be run instead. PERMANOVA compares the variances of between-sample and within-sample

sum of squares of distances. The significance of this ratio, called pseudo F-ratio, is calculated by shuffling, aka permutations. First, the order of species (rows) is randomly shuffled a certain number of times to generate empirical F distributions. Then, the significance between samples can be derived from the empirical F distribution. The underlying null hypothesis is that the samples aren't different, and hence species (rows) are exchangeable among the different samples [24].

Clustering into enterotypes – Samples can be clustered into enterotypes based on their abundances of key microbial taxa, where samples within the same cluster are similar to one another and dissimilar to samples in other clusters. Interpretation of enterotypes is subjective, since their selection is affected by distance metrics, clustering approaches, etc. used, and there's a lack of universal practices [25].

Cluster analysis is comprised of choosing a distance measure to depict the data's variability, the clustering method, and an appropriate number of clusters. Clustering methods fall into two categories: hierarchical and non-hierarchical. Hierarchical algorithms sort the data into clusters that are nested hierarchically within other clusters, while non-hierarchical algorithms partition the data into separate clusters. The standard agglomerative strategy for hierarchical clustering is to start with each observation in its own cluster, and merge pairs of clusters until all observations are in the same cluster. Hierarchical clustering also depends on the linkage criteria, such as average, median, centroid, which dictate how the distance between two clusters is defined [26]. Some of the most widely used non-hierarchical methods are k-means and k-medoids/ partitioning around medoids (PAM). k-means and PAM both partition the data into separate groups by trying to minimize the distance between the center point in a cluster and points inside that cluster. The main differences between them is that k-means assigns the average between the points in a cluster as center points of clusters, while PAM assigns input data points as center points [27]. Determining the optimal number of clusters (k-value) is required to perform partitioning clustering methods, such as k-means and PAM. k-value selection algorithms include the Gap Statistic, Elbow method, Silhouette coefficient, etc. The Gap Statistic computes performance scores for each k-value by comparing changes in within-cluster dispersion, aka variation, with those expected under an appropriate null reference distribution [28].

PAM is preferred over k-means when clustering ecological data. Compared to PAM, the means statistics used by k-means is a poorer indicator of centrality due to its higher sensitivity to outliers and noise (Figure 4) [26]; and the distance measure used by k-means, Euclidean distance, is insensitive to small changes in absolute species abundances and is incapable of distinguishing if a species is truly absent from two samples or just under sampled (referred to as the "double zero" problem) [29]. On the other hand, PAM can be run with any chosen distance metric, such as Bray-Curtis or Jaccard [27]. Unlike the Euclidean distance, Bray-Curtis and Jaccard indices don't suffer from the "double zero" problem, since they ignore double zeros [30]. Other clustering algorithms have previously been applied to microbiota data, such as the hierarchical methods Agnes, Hclust, and Dirichlet Multinomial Mixture, with various linkage criteria [31].

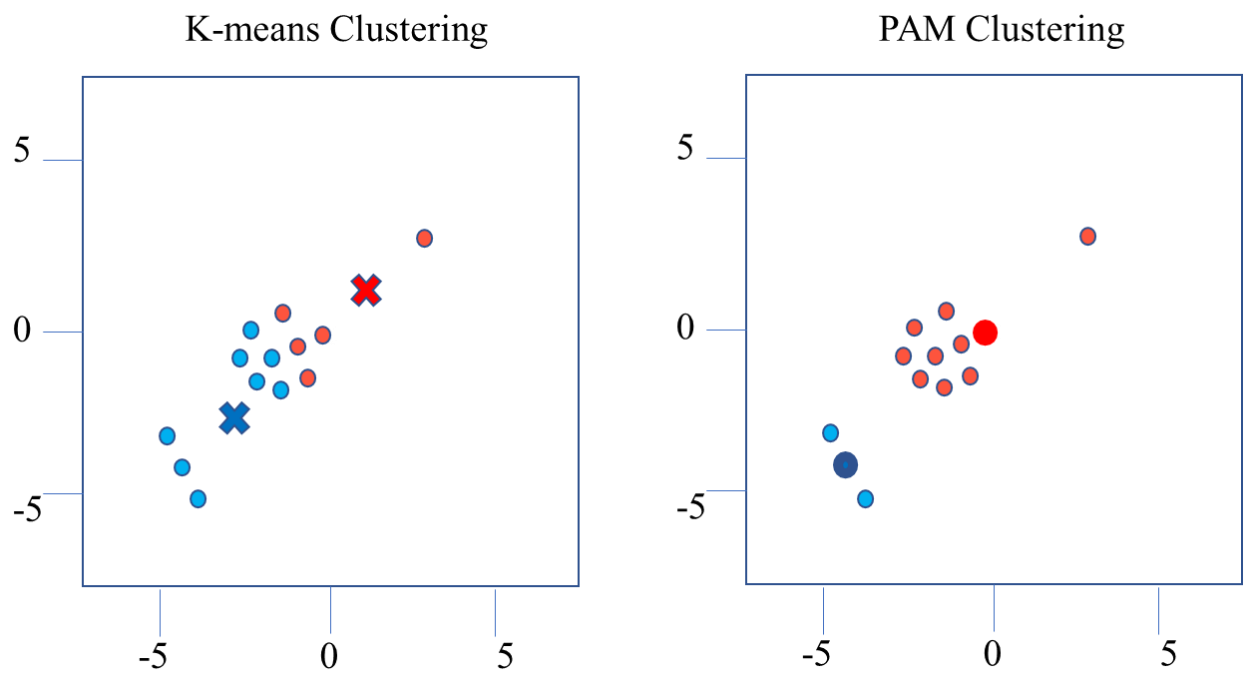


Figure 4. Differences between k-means and PAM clustering methods. k-means assigns the average between the points in a cluster as center points of clusters, while PAM assigns input data points as center points [27]. Compared to PAM, the means statistics used by k-means is a poorer indicator of centrality due to its higher sensitivity to outliers and noise [26].

Analysis of individual taxa – Microbial communities can also be profiled by testing which species, genera, or etc. are differentially abundant between two ecosystems or groups [12]. Differential abundance analyses were initially applied to transcript abundances, which are estimated from RNA-sequencing data to identify genes whose expression levels varies between different conditions, for example, cancer versus normal condition [32]. Although rarefying counts or transforming absolute abundances into relative abundances serve as common normalization approaches, their usage can result in a high rate of false positives in differential abundance analyses. For example, rarefying equalizes variances between samples, but it comes at the cost of underestimating the true variance due to the loss of information during subsampling. Instead, it's more data-efficient to model the noise and extra species. Differential abundance analysis packages such as DESeq2 and edgeR incorporate their own normalization algorithm and fit count data using negative binomial regression models. As count data are a discrete type of variable, they can't be modelled with a normal distribution. Poisson distribution is better suited for modelling the discrete count data; however, due to the high variance of taxon counts compared to their mean (aka overdispersion), an extension of the Poisson distribution called negative binomial distribution is ultimately used to account for this high variance [33].

DESeq2 first normalizes the count data by estimating the size factors in order to handle the differing sequencing depth between the libraries [24]. Specifically, DESeq2 implements the median of ratios as its normalization method; it's calculated as the ratio of each sample to the geometric mean of each taxon across all samples [34]. Dispersion parameters are then estimated to account for the within-group variability and variability between replicates; and finally, a negative binomial model is fitted, and the resulting log fold changes between the two conditions are checked by the Wald test for significant differentially abundant taxa [24].

1.6 Objectives

While only one past study has investigated the association between maternal saliva concentration and the infant gut microbiota, there have been no studies that have utilized HCC as a measure for maternal prenatal PD to examine its association with the infant gut microbiota [3]. This is mainly because HCC is a newer and not yet widely used method. The aim of this thesis was to examine if and how maternal prenatal HCC associates with the infant gut microbiota.

In greater detail, the aim was to study whether the maternal hair cortisol concentration (HCC) taken during the 24th week of pregnancy associates with the infant gut microbiota's operational taxonomic units (OTU's), genera, enterotype clusters, diversity, and richness at the age of 2.5 months. HCC taken during the 40th week of pregnancy and its association with the infant gut microbiota was also briefly investigated. In addition, important maternal and infant covariates were selected and adjusted for in the above analyses.

2. Methods

2.1 Study design

The study population originated from the FinnBrain Birth Cohort Study [1]. The following datasets were available for this study population - an OTU table storing the analysed infant stool samples collected at around 2.5 months of age, and a sample data table containing complete or missing maternal HCC observations taken during the 24th week of pregnancy and other information. The OTU table gave the number of reads per sample per OTU, while the HCC samples represented the average maternal cortisol concentration of the last 5 months [1]. Out of the original 445 observations and 274 variables in the sample data table, only 120 observations containing the complete HCC cases were used for majority of the analyses. Furthermore, only hair samples weighing 5-15 mg were included in the analyses to incorporate the hair weight covariate into the HCC variable, and HCC was log transformed to make the original HCC variable less skewed. The other 273 variables contained extensive information about the infant and mother, from which 11 covariates were deemed as of potential interest based on preliminary evidence from the FinnBrain group and its Birth Cohort Study. These covariates included the mother's age, level of education, freezer time of HCC sample, mother's BMI, breastfeeding at 2.5 months, infant sex, mode of delivery, stool sampling age in weeks, number of previous deliveries, selective serotonin reuptake inhibitor (SSRI) use by mother, and season of HCC_24 sampling (Table 1). Usage of antibiotics, another covariate, was discarded as a covariate of interest because there were too few observations (~ 50). Furthermore, HCC data collected at week 40 was also available, even though there were only 20 complete observations, and the HCC_24 variable was modified into a new variable divided by its quantiles 1 & 4 (Table 1). Therefore, there were three HCC measures in total that could be utilized in analyses.

Table 1. Overview of HCC variables and covariates

Variable	Type of variable	Values
HCC_24	continuous	NA
HCC_24 Q1/Q4 quantiles	categorical (binary)	1st quantile 4th quantile
HCC_40	continuous	NA
Mother's age	discrete	NA
Level of education	categorical (ordinal)	1: basic to upper secondary level 2: vocational school diploma 3: lower degree level tertiary education to doctorate or equivalent diploma
Freezer time of HCC_24 sample	categorical (ordinal)	0: 0 days 1: 1-2 days 2: 98-103 days 3: > 103 days
Mother's BMI	continuous	NA
Breastfeeding at 2.5 months	categorical (nominal)	0: never breastfed 1: no breastfeeding anymore 2: partial 3: exclusive
Infant sex	categorical (binary)	1: male 2: female
Mode of delivery	categorical (binary)	1: all vaginal 2: all caesarean section
Stool sampling age in weeks	continuous	NA
Number of previous deliveries	discrete	NA
SSRI use by mother	categorical (binary)	0: no 1: yes
Season of HCC_24 sampling	categorical (nominal)	1: winter 2: spring 3: summer 4: autumn

2.2 HCC & stool sample collection

HCC and stool samples were collected and analysed prior to the project's beginning as part of the FinnBrain Birth Cohort Study. Parents collected the stool samples at their homes, stored the samples at +4 C, and brought the samples to the laboratory within 24 hours after collection to have the DNA extracted from them [35, 36]. The V4 region of bacterial 16S ribosomal RNA was sequenced with Illumina MiSeq approach. The read quality was then checked using FastQC (v. 0.10.1) and downstream analyses were conducted using QIIME (v.1.9) [36]. Reads were quality filtered to at least 20 Phred quality, chimeric sequences were filtered out by the usearch tool, and Operational Taxonomic Units (OTUs) were selected using UCLUST with 97% sequence similarity and excluded if total sequence count was less than 0.05%. OTUs were annotated using the GreenGenes database [36].

The HCC measuring procedure consisted of two isopropanol washes of the collected hair strands, powdering of the clean & dried hair, 24-hour long methanol extraction, reconstitution of the dried extract in an assay buffer, and quantification of extracted cortisol using a specific enzyme immunoassay [37].

2.3 Covariate selection

To test the association between microbial alpha-diversity indices (Shannon index and Chao1 index) and HCC_24 via a linear regression model, 11 infant and maternal covariates deemed as of potential interest based on preliminary evidence from the FinnBrain group had to be first reduced in number. Several variable selection methods were employed – filter methods (correlations), and wrapper methods (backwards elimination). Initially, correlations and associations between HCC_24 and maternal and infant covariates of interest were assessed using Spearman's correlation, Kruskal–Wallis test, or Mann-whitney U test. Specifically,

associations between HCC_24 and nominal/ordinal variables - the level of education, freezer time of HCC sample, season of HCC_24 sampling, breastfeeding at 2.5 months - were assessed using the Kruskal–Wallis test. Associations between HCC_24 and binary variables - infant sex, mode of delivery, SSRI use by mother - were assessed using the Mann-whitney U test. Correlations between HCC_24 and continuous variables - mother's age, mother's BMI, number of previous deliveries, stool sampling age in weeks - were assessed using the Spearman's correlation. However, after p-value adjustment using the Benjamini & Hochberg method, all covariates were identified as insignificant [38]. As a solution, the final alpha-diversity linear regression models were obtained via the backward elimination method, which based on the Akaike information criterion (AIC) involved repeatedly removing insignificant covariates till only significant covariates remained [39]. Infant covariates considered crucial by the FinnBrain group- infant sex, stool sampling age in weeks, mode of delivery – were always included in the models. The selected covariates were permanently incorporated into most downstream analyses, such as analysis of beta-diversity, differential abundance analysis of individual genera, etc.

2.4 Statistical analyses

All statistical analyses were run with the R software. Alpha-diversity indices were calculated using phyloseq R package [40]. Shannon index represented species richness and evenness, while the Chao1 index represented estimated species richness. Regression analyses of alpha-diversity indices was performed in relation to the following independent variables – one of the three HCC measures (24,40 weeks, or quantiles), infant covariates, and maternal covariates. To evaluate the gut microbiota's enterotypes, subjects were clustered based on their core OTUs (where OTU's representing less than 0.1% abundance and with less than 5% prevalence were excluded) with the Bray-Curtis distance matrix via the Partitioning Around Medoids (PAM) method from the cluster R package [41]. The optimal number of clusters was calculated using the gap statistics on core OTUs with the Bray-Curtis distance matrix [41]. Correlations and associations between clusters and HCC_24, alpha-diversity, maternal

and infant covariates of interest were assessed using Kruskal–Wallis test, or chi-squared test ($\chi 2$). Differential expression analysis, performed using the DESeq2 R package, was run to identify infant gut genera that were differentially abundant for HCC_24 when adjusted for infant covariates [42]. Non-rarefied absolute data was used to generate results at genus and core genus level. All maternal covariates were left out because there were initially very few results at genus level. Individual effects of each categorical infant covariate on the associations between infant gut genera and HCC_24 were examined by subsetting the data. A permutational analysis of variance (PERMANOVA) was performed using adonis function from the vegan R package to test whether the overall microbial community, in other words beta-diversity, differs by the variable of interest and covariates. In addition to PERMANOVA, betadisper and anosim functions from the vegan R package were run to further confirm that the data divided by various covariates have equal beta dispersion, and therefore holds the assumptions of PERMANOVA [43]. The variable of interest was either HCC_24, Q1/Q4 quantiles of HCC_24, or HCC_40. Jaccard and Bray-Curtis distances were utilized to assess the beta richness and diversity, respectively, of the community. The PERMANOVA analyses were always run with 1000 permutations. Model coefficients were extracted for the top taxa separating HCC_24 Q1/Q4 quantiles' groups using the generic coefficients function from the stats R package [44]. p-values were adjusted after multiple testing in an analysis using the Benjamini & Hochberg method (R function p.adjust), where p-values less or equal to 0.05 were considered statistically significant [38,45].

3.Results

3.1 Maternal & infant covariates selection

Reduction of covariates was necessary for performing analyses. Otherwise, these analyses would have lost their statistical power due to the large number of covariates and the comparatively small number of available observations. Linear regression analysis of alpha-diversity was the first planned analysis for this thesis project.

Initially, the correlations and associations between HCC_24 and the 11 maternal and infant covariates of interest were checked for. However, after p-value adjustment all 11 covariates were identified as insignificant. Therefore, significant covariates were determined while constructing the alpha-diversity linear regression models. Both the Shannon and Chao1 index models fitted with HCC_24 were obtained via the backwards elimination method, which based on the AIC value involved repeatedly removing insignificant covariates till only significant covariates remained. Crucial infant covariates - infant sex, stool sampling age in weeks, mode of delivery – were always included in the models. After backwards elimination, the Shannon index model kept the mother's age, breastfeeding at 2.5 months, and season of HCC_24 sampling maternal covariates. On the other hand, the Chao1 index model kept the mother's age, season of HCC_24 sampling, breastfeeding at 2.5 months, and SSRI use by mother maternal covariates. Overall, the Shannon and Chao1 index models shared the mother's age, season of HCC_24 sampling, and breastfeeding at 2.5 months covariates. Along with the infant covariate - breastfeeding at 2.5 months, only mother's age was the sole shared maternal covariate picked for future analyses. To explain, season of HCC_24 sampling was left out midway because its role has been contradictory in preliminary studies from the FinnBrain group.

Generally, the mother's age was positively associated with both the Shannon and Chao1 indices, while partial (breastfeeding at 2.5 months group 2) and exclusive (breastfeeding at 2.5 months group 3) breastfeeding were negatively associated with these indices. The HCC_40 models had different associations, but they hadn't yielded any significant covariates. Finally, according to the F-test, which is a test that is incorporated into linear regression, nearly all the models had non-significant F-test results, except for the Chao1 index model with HCC_24 Q1/Q4 quantiles. A significant F-test would have indicated that the observed R-squared is reliable.

3.2 Linear regression analyses of alpha-diversity and maternal HCC

The relationship between alpha-diversity indices, hair cortisol concentration, and previously selected maternal/infant covariates - mother's age, breastfeeding at 2.5 months, infant sex, mode of delivery, stool sampling age in weeks, was assessed by building alpha-diversity linear regression models fitted with either HCC_24, HCC_24 Q1/Q4 quantiles, or HCC_40. Neither of the alpha-diversity indices, Shannon and Chao1, were associated with HCC_24 (Table 2-3), HCC_24 Q1/Q4 quantiles (Table 4-5), or HCC_40 before or after p-value adjustment. However, several covariates were significant before p-value adjustment.

For both Shannon and Chao1 index models fitted with HCC_24, the exclusive breastfeeding (breastfeeding at 2.5 months group 3) group was the only significant covariate, and mother's age was only significant for the Chao1 index model fitted with HCC_24 (Table 3). However, there were no significant covariates left after p-value adjustment. The exclusive breastfeeding (breastfeeding at 2.5 months group 3) covariate was the only significant covariate in both Shannon and Chao1 index models fitted with HCC_24 Q1/Q4 quantiles, and the mother's age and partial breastfeeding (breastfeeding at 2.5 months group 2) were only significant for the

Chao1 index model fitted with HCC_24 Q1/Q4 quantiles (Table 5). The results were quite similar to those of the HCC_24 models. However, again there were no significant covariates left after p-value adjustment. The Chao1 and Shannon index models fitted with HCC_40 didn't have any significant covariates before or after p-value adjustment.

Table 2. Shannon index model fitted with HCC_24.

	Estimate	**Standardized**	**Std.Error**	**T-value**	**Pr(>|t|)**
(Intercept)	1.551	0	0.536	2.893	0.005 **
HCC_24	0.003	0.006	0.043	0.067	0.947
Mother's age	0.019	0.179	0.010	1.937	0.055 .
Sample week	0.025	0.105	0.023	1.119	0.266
Breastfeeding 1	-0.724	-0.314	0.387	-1.868	0.064 .
Breastfeeding 2	-0.600	-0.475	0.346	-1.732	0.086 .
Breastfeeding 3	-0.767	-0.686	0.332	-2.312	0.023 *
Birth mode 2	-0.063	-0.052	0.114	-0.551	0.582
Gender 2	-0.025	-0.027	0.085	-0.288	0.774
Residual standard error: 0.456 on 111 degrees of freedom Multiple R-squared: 0.096 Adjusted R-squared: 0.031 F-statistic: 1.478 on 8 and 111 DF p-value: 0.173 **Signif. codes: 0** = *** **0.001** = ** **0.01** = * **0.05** = . **0.1** = " " 1 = " "					

Table 3. Chao1 index model fitted with HCC_24.

	Estimate	Standardized	Standard Error	T-value	p-value (Pr(>|t|))
(Intercept)	307.800	0	111.187	2.768	0.007 **
HCC_24	-2.365	-0.025	8.900	-0.266	0.791
Mother's age	4.025	0.189	1.984	2.029	0.045 *
Sample week	2.997	0.060	4.720	0.635	0.527
Breastfeeding 1	-138.860	-0.292	80.353	-1.728	0.087 .
Breastfeeding 2	-11.239	-0.505	71.808	-1.828	0.070 .
Breastfeeding 3	-150.396	-0.653	68.827	-2.185	0.031 *
Birth mode 2	-24.813	-0.099	23.557	-1.053	0.295
Gender 2	-6.977	-0.037	17.641	-0.395	0.693
Residual standard error: 94.470 on 111 degrees of freedom Multiple R-squared: 0.083 Adjusted R-squared: 0.017 F-statistic: 1.259 on 8 and 111 degrees of freedom p-value: 0.272 Signif. codes: 0 = *** 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "					

Table 4. Shannon index model fitted with HCC_24 Q1/Q4 quantiles.

	Estimate	Standardized	Standard Error	T-value	p-value (Pr(>|t|))
(Intercept)	1.910	0	0.691	2.763	0.008 **
HCC24_quantilesQ4	-0.003	-0.003	0.121	-0.026	0.980
Mother's age	0.021	0.211	0.013	1.548	0.128
Sample week	-0.019	-0.068	0.038	-0.499	0.620
Breastfeeding 1	-0.271	-0.107	0.471	-0.575	0.568
Breastfeeding 2	-0.691	-0.567	0.362	-1.908	0.062 .
Breastfeeding 3	-0.804	-0.749	0.335	-2.401	0.020 *
Birth mode 2	-0.011	-0.009	0.163	-0.066	0.948
Gender 2	-0.015	-0.016	0.122	-0.119	0.906
Residual standard error: 0.446 on 51 degrees of freedom Multiple R-squared: 0.180 Adjusted R-squared: 0.052 F-statistic: 1.403 on 8 and 51 degrees of freedom p-value: 0.218 Signif. codes: 0 = *** 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "					

Table 5. Chao1 index model fitted with HCC_24 Q1/Q4 quantiles.

	Estimate	Standardized	Standard Error	T-value	p-value (Pr(>|t|))
(Intercept)	324.191	0	141.084	2.298	0.026 *
HCC24_quantilesQ4	-12.167	-0.063	24.690	-0.493	0.624
Mother's age	6.270	0.298	2.727	2.3	0.026 *
Sample week	-6.574	-0.112	7.702	-0.854	0.397
Breastfeeding 1	-6.900	-0.013	96.005	-0.072	0.943
Breastfeeding 2	-169.700	-0.650	73.847	-2.298	0.026 *
Breastfeeding 3	-168.686	-0.734	68.304	-2.47	0.017 *
Birth mode 2	-20.335	-0.081	33.218	-0.612	0.543
Gender 2	-1.314	-0.007	24.985	-0.053	0.958

Residual standard error: 91 on 51 degrees of freedom
Multiple R-squared: 0.256
Adjusted R-squared: 0.139
F-statistic: 2.187 on 8 and 51 degrees of freedom
p-value: 0.044

Signif. codes: 0 = *** 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "

3.3 HCC and clusters in infant gut microbiota

The infant gut microbiota was clustered, via the PAM method, into enterotypes in order to assess its association with HCC_24. The association between enterotypes and alpha-diversity, or selected maternal/infant covariates was also checked. Prior to clustering, the optimal number of clusters had to be found. The gap statistics was calculated using dimensionally reduced core OTUs, where OTU's representing less than 0.1% abundance and with less than 5% prevalence were excluded, and the Bray-Curtis dissimilarity measure. The gap statistics showed that 3 was the optimal number of clusters (Figure 5). The number of clusters was also validated by using all the OTUs instead of core OTUs, and various distance matrices – Bray-Curtis, Jaccard. At first only 120 samples that had the HCC_24 measurements were looked at when testing the association between the clusters and HCC_24, but this reduced the statistical power of future analyses because the association of covariates with the clusters would have been limited to only those 120 samples. Therefore, clustering was repeated using all the 445 samples with all the available alpha-diversity, HCC_24, and 11 maternal/infant covariate measurements. Multidimensional scaling (MDS) was utilized using the Bray-Curtis dissimilarity measure to construct a plot showing the clusters (Figure 6).

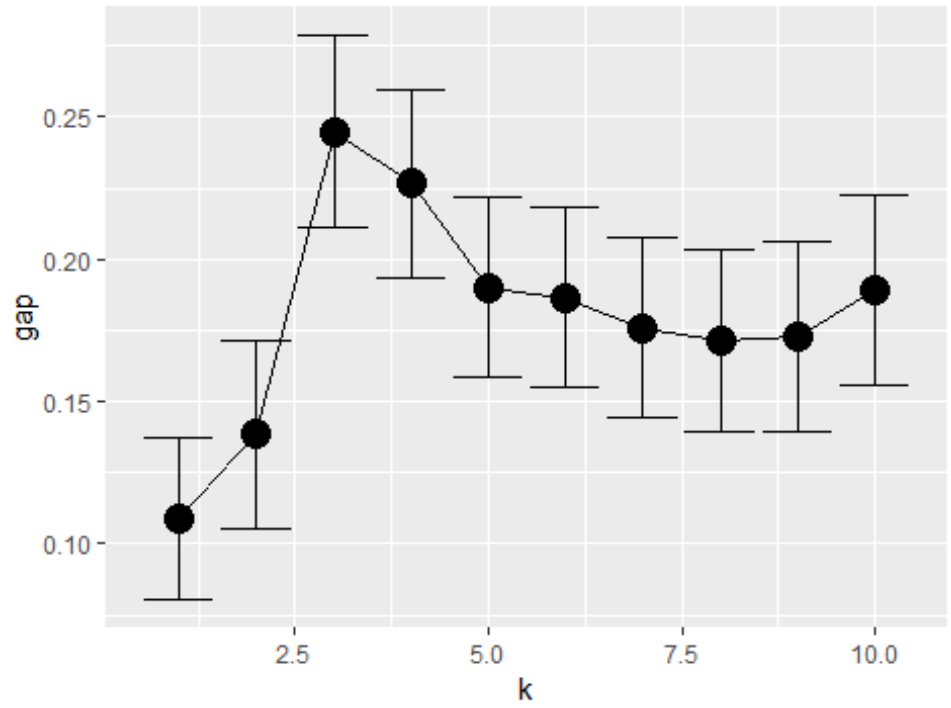


Figure 5. Calculated gap statistics using dimensionally reduced core OTUs & Bray-Curtis distances. The gap statistics showed that 3 was the optimal number of clusters.

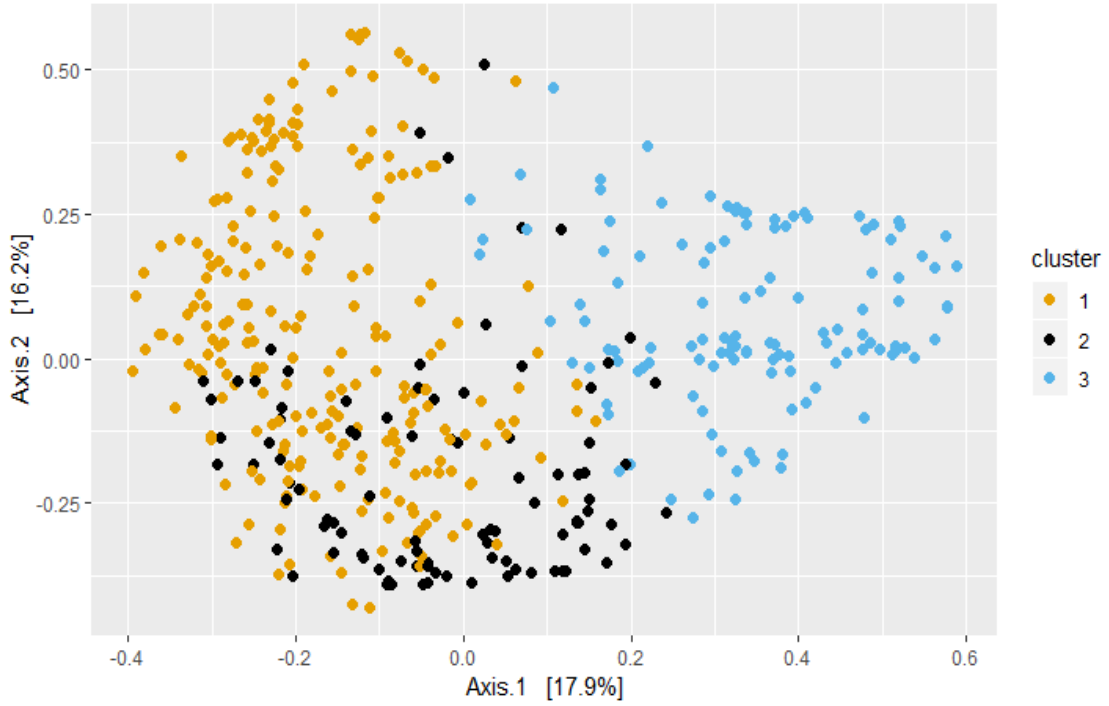


Figure 6. MDS plot of the 3 infant gut microbiota clusters.

Since the infant gut microbiota clusters are a categorical variable, associations between the clusters and HCC_24 & other covariates could be tested. Associations between the clusters and continuous variables – HCC_24, Shannon, Chao1, mother's age, mother's BMI, number of previous deliveries, stool sampling age in weeks - were assessed using the Kruskal–Wallis test. Associations between the clusters and categorical variables - the level of education, freezer time of HCC sample, season of HCC_24 sampling, breastfeeding at 2.5 months, infant sex, mode of delivery, SSRI use by mother - were assessed using the chi-squared test. However, HCC_24 wasn't significantly different among the clusters before or after p-value adjustment. Only number of previous deliveries, mode of delivery, and Chao1 seem to be different among the clusters after p-value adjustment out of the 14 listed variables (Table 6).

Table 6. Associations between infant gut microbiota clusters and HCC_24 & covariates

Names	Adjusted p-value
HCC_24	0.952
Mother's age	0.929
Previous deliveries	0.007
Mother's BMI	0.683
Season sampling	0.583
Freezertime	0.289
SSRI use	0.349
Education level	0.431
Sample week	0.683
Gender	0.349
Birth mode	2.6×10^{-6}
Breastfeeding	0.114
Shannon	0.106
Chao1	0.002

Re-running alpha-diversity linear regression analyses with clusters indicated that cluster 3 was a significant group in the Chao1 index regression model with HCC_24, although it wasn't significant after p-value adjustment (Table 7-8). Shannon and Chao1 index models formulated with HCC_24 Q1/Q4 quantiles and HCC_40 didn't yield any significant cluster groups.

Table 7. Chao1 index model with infant gut microbiota clusters fitted with HCC_24

	Estimate	**Standardized**	**Standard Error**	**T-value**	**p-value** (Pr(>|t|))
(Intercept)	310.881	0	98.063	3.170	0.002 **
HCC_24	0.561	0.007	6.903	0.081	0.935
Mother's age	3.831	0.185	1.711	2.240	0.027 *
Sample week	4.145	0.092	3.765	1.101	0.273
Breastfeeding 1	-143.268	-0.284	76.734	-1.867	0.064 .
Breastfeeding 2	-109.743	-0.456	67.458	-1.627	0.106
Breastfeeding 3	-147.886	-0.671	65.693	-2.251	0.026 *
Birth mode 2	-14.162	-0.060	20.970	-0.675	0.501
Gender 2	-8.932	-0.048	15.868	-0.563	0.575
Cluster 2	-37.865	-0.172	20.068	-1.887	0.061 .
Cluster 3	-39.521	-0.192	19.065	-2.073	0.040 *
Residual standard error: 90.180 on 130 degrees of freedom Multiple R-squared: 0.139 Adjusted R-squared: 0.073 F-statistic: 2.095 on 10 and 130 degrees of freedom p-value: 0.029 Signif. codes: 0 = *** 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "					

Table 8. Shannon index model with infant gut microbiota clusters fitted with HCC_24

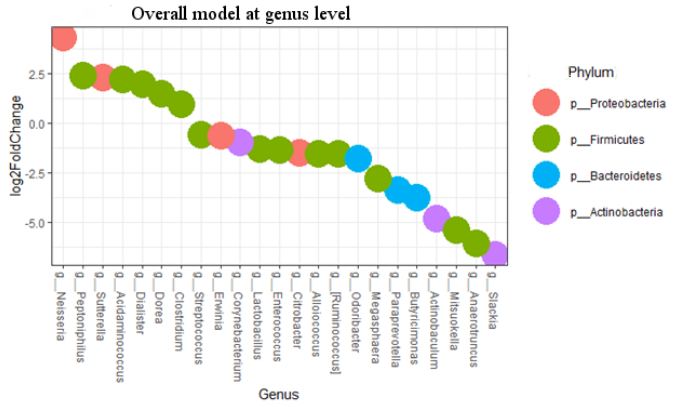
	Estimate	Standardized	Standard Error	T-value	p-value (Pr(>|t|))
(Intercept)	1.489	0	0.493	3.021	0.003 **
HCC_24	-0.002	-0.005	0.035	-0.057	0.954
Mother's age	0.022	0.211	0.009	2.573	0.011 *
Sample week	0.027	0.118	0.019	1.425	0.157
Breastfeeding 1	-0.717	-0.280	0.386	-1.859	0.065 .
Breastfeeding 2	-0.509	-0.417	0.339	-1.502	0.135
Breastfeeding 3	-0.738	-0.660	0.330	-2.236	0.027 *
Birth mode 2	-0.098	-0.081	0.105	-0.927	0.356
Gender 2	0.008	0.008	0.080	0.100	0.921
Cluster 2	-0.153	-0.137	0.101	-1.515	0.132
Cluster 3	-0.174	-0.167	0.096	-1.817	0.071 .
Residual standard error: 0.453 on 130 degrees of freedom Multiple R-squared: 0.155 Adjusted R-squared: 0.090 F-statistic: 2.387 on 10 and 130 degrees of freedom p-value: 0.012 Signif. codes: 0 = *** 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "					

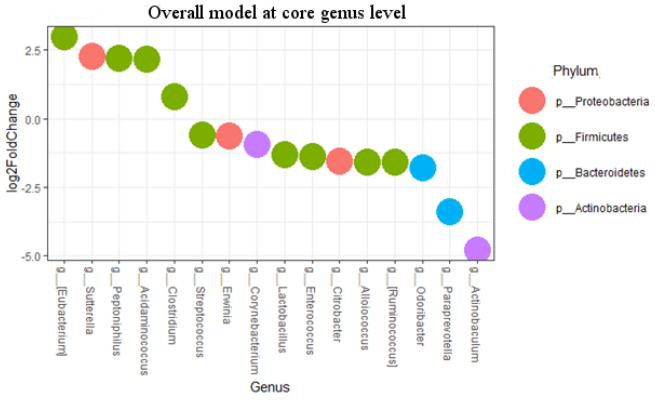
3.4 Gut microbiota composition and HCC

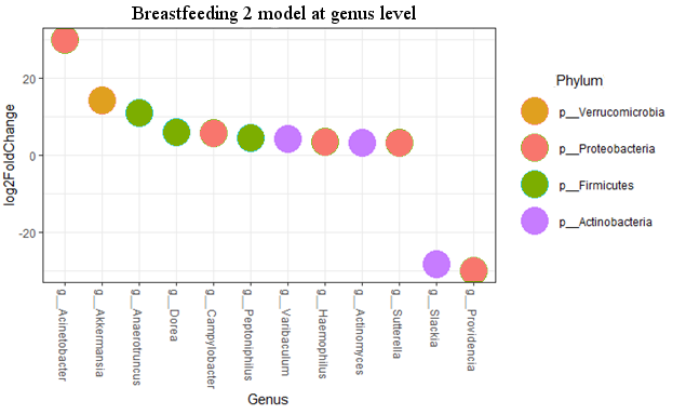
DESeq2 analyses were run to identify which genera in infant gut microbiota were differentially abundant for HCC_24, when adjusted for the covariates. Non-rarefied data was used to generate results at genus and core genus levels. All maternal covariates, except HCC_24, were left out, as suggested by the FinnBrain group, because there were initially very few results at genus and core genus levels. Therefore, DESeq2 analyses were run only with the infant covariates - breastfeeding at 2.5 months, infant sex, mode of delivery, and stool sampling age in weeks. Individual effects of each categorical infant covariate - breastfeeding at 2.5 months, infant sex, and mode of delivery – on the association between the gut microbiota composition and HCC_24 was examined by subsetting the data. HCC_24

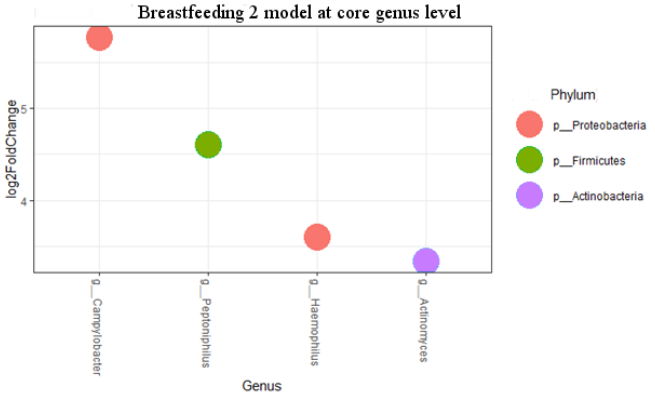
in the overall model, containing all the infant covariates, had negative associations with the phyla Bacteroidetes (e.g. genera *Paraprevotella, Odoribacter*) and Actinobacteria (e.g. genera *Actinobaculum, Corynebacterium*), while Proteobacteria (e.g. genera *Sutterella, Erwinia, Citrobacter*) and Firmicutes (e.g. genera *Clostridium, Streptococcus, Lactobacillus*) showed both positive and negative associations (Figure 7).

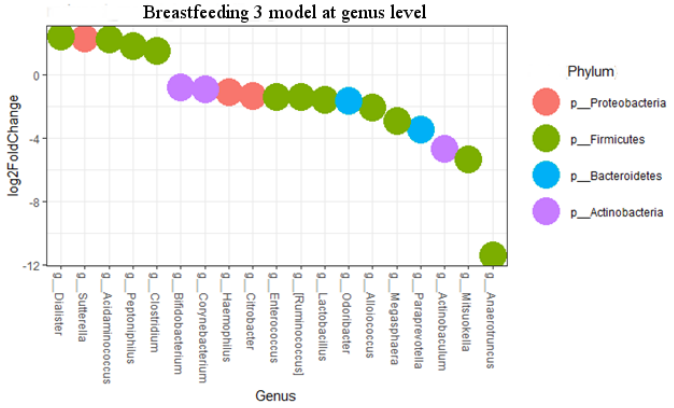
In the breastfeeding at 2.5 months subset, partial breastfeeding (breastfeeding at 2.5 months group 2) had mainly positive associations, while exclusive breastfeeding (breastfeeding at 2.5 months group 3) had slightly more negative associations than positive. Bifidobacteria, which is high in mother's milk, was negatively associated in exclusive breastfeeding at genus and core genus level [2]. Firmicutes members were the most abundant phylum in both the partial and exclusive breastfeeding groups. Additionally, the exclusive breastfeeding group had more associations than the partial breastfeeding group, which was mainly due to the partial and exclusive breastfeeding groups possessing different number of observations - 19 and 94, respectively. In the mode of delivery subset, cesarean section (mode of delivery group 2) was associated with more Actinobacteria and oral, skin, and placental species (*Propionibacterium*) than vaginal delivery (mode of delivery group 1). Specifically, the association with *Propionibacterium* was negative in the cesarean section group. On the other hand, vaginal delivery was associated with more Firmicutes and Bacteroidetes members (Figure 7). The vaginal delivery group had more associations than the caesarean section group, which was mainly due to them having different number of observations – 99 and 21, respectively. Finally, in the infant sex subset, no prominent differences were observed between the two genders.

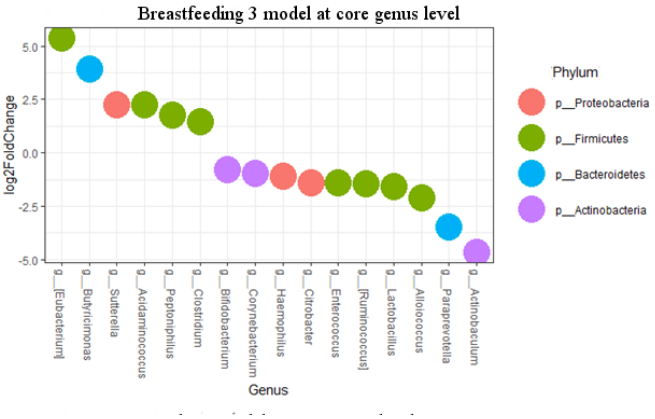

Overall model at genus level
log2FoldChange
Genus
g__Neisseria
g__Peptoniphilus
g__Sutterella
g__Acidaminococcus
g__Dialister
g__Dorea
g__Clostridium
g__Streptococcus
g__Erwinia
g__Corynebacterium
g__Lactobacillus
g__Enterococcus
g__Citrobacter
g__Alloiococcus
g__[Ruminococcus]
g__Odoribacter
g__Megasphaera
g__Parapprevotella
g__Butyricimonas
g__Actinobaculum
g__Mitsuokella
g__Anaerotruncus
g__Slackia
Phylum
p__Proteobacteria
p__Firmicutes
p__Bacteroidetes
p__Actinobacteria

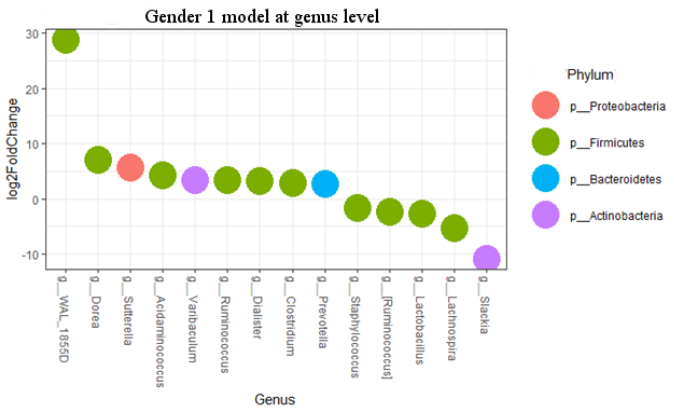

Overall model at core genus level
log2FoldChange
Genus
g__[Eubacterium]
g__Sutterella
g__Peptoniphilus
g__Acidaminococcus
g__Clostridium
g__Streptococcus
g__Erwinia
g__Corynebacterium
g__Lactobacillus
g__Enterococcus
g__Citrobacter
g__Alloiococcus
g__[Ruminococcus]
g__Odoribacter
g__Parapprevotella
g__Actinobaculum
Phylum
p__Proteobacteria
p__Firmicutes
p__Bacteroidetes
p__Actinobacteria

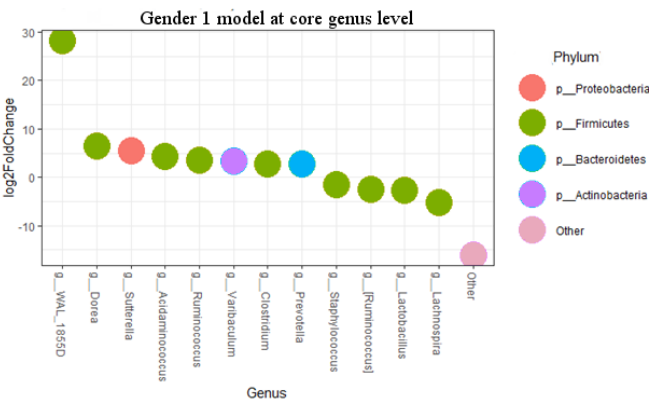

Breastfeeding 2 model at genus level
log2FoldChange
Genus
g__Acinetobacter
g__Akkermansia
g__Anaerotruncus
g__Dorea
g__Campylobacter
g__Peptoniphilus
g__Varibaculum
g__Haemophilus
g__Actinomyces
g__Sutterella
g__Slackia
g__Providencia
Phylum
p__Verrucomicrobia
p__Proteobacteria
p__Firmicutes
p__Actinobacteria


Breastfeeding 2 model at core genus level
log2FoldChange
Genus
g__Campylobacter
g__Peptoniphilus
g__Haemophilus
g__Actinomyces
Phylum
p__Proteobacteria
p__Firmicutes
p__Actinobacteria


Breastfeeding 3 model at genus level
log2FoldChange
Genus
g__Dialister
g__Sutterella
g__Acidaminococcus
g__Peptoniphilus
g__Clostridium
g__Bifidobacterium
g__Corynebacterium
g__Haemophilus
g__Citrobacter
g__Enterococcus
g__[Ruminococcus]
g__Lactobacillus
g__Odoribacter
g__Alloiococcus
g__Megasphaera
g__Parapprevotella
g__Actinobaculum
g__Mitsuokella
g__Anaerotruncus
Phylum
p__Proteobacteria
p__Firmicutes
p__Bacteroidetes
p__Actinobacteria


Breastfeeding 3 model at core genus level
log2FoldChange
Genus
g__[Eubacterium]
g__Butyricimonas
g__Sutterella
g__Acidaminococcus
g__Peptoniphilus
g__Clostridium
g__Bifidobacterium
g__Corynebacterium
g__Haemophilus
g__Citrobacter
g__Enterococcus
g__[Ruminococcus]
g__Lactobacillus
g__Alloiococcus
g__Parapprevotella
g__Actinobaculum
Phylum
p__Proteobacteria
p__Firmicutes
p__Bacteroidetes
p__Actinobacteria


Gender 1 model at genus level
log2FoldChange
Genus
g__WAL_1855D
g__Dorea
g__Sutterella
g__Acidaminococcus
g__Varibaculum
g__Ruminococcus
g__Dialister
g__Clostridium
g__Prevotella
g__Staphylococcus
g__[Ruminococcus]
g__Lactobacillus
g__Lachnospira
g__Slackia
Phylum
p__Proteobacteria
p__Firmicutes
p__Bacteroidetes
p__Actinobacteria


Gender 1 model at core genus level
log2FoldChange
Genus
g__WAL_1855D
g__Dorea
g__Sutterella
g__Acidaminococcus
g__Ruminococcus
g__Varibaculum
g__Clostridium
g__Prevotella
g__Staphylococcus
g__[Ruminococcus]
g__Lactobacillus
g__Lachnospira
Other
Phylum
p__Proteobacteria
p__Firmicutes
p__Bacteroidetes
p__Actinobacteria
Other

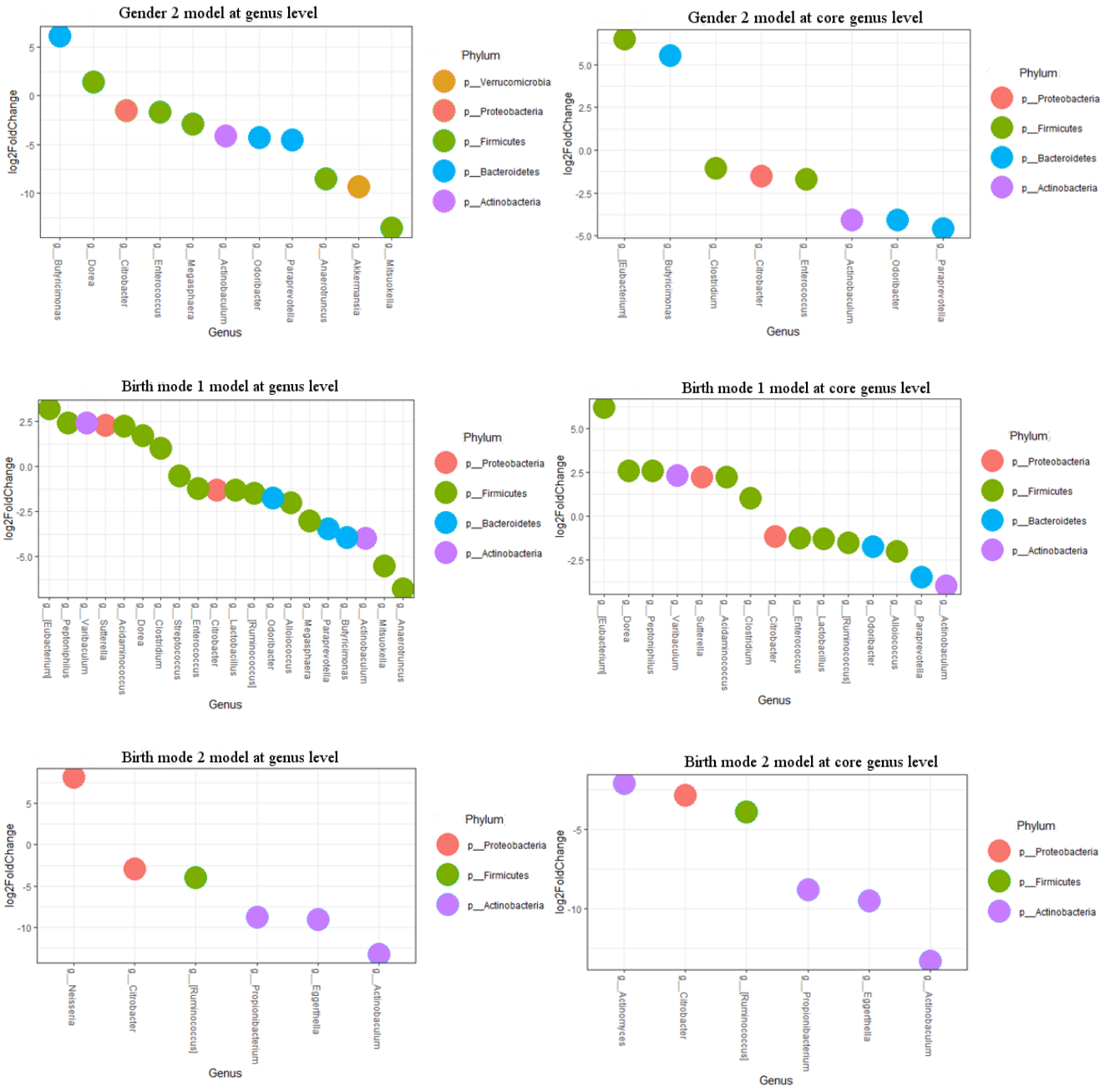


Figure 7. Plots of associations between HCC_24 and infant gut microbiota genera/core genera. DESeq2 analyses were run without the maternal covariates to construct a reduced overall model, and models for each of the infant covariates - breastfeeding at 2.5 months, infant sex, and mode of delivery. The breastfeeding at 2.5 months and mode of delivery subset groups were visually different.

3.5 Linear regression analyses of beta-diversity

A permutational analysis of variance (PERMANOVA) was run to test whether the overall microbial community differs by the variable of interest and covariates. The variable of interest was either HCC_24, HCC_24 Q1/Q4 quantiles, or HCC_40. Jaccard and Bray-Curtis distances were utilized to check the community's beta-diversity qualitatively and quantitatively, respectively.

In addition to PERMANOVA, Multivariate homogeneity of groups dispersions (betadisper) and Analysis of similarities (ANOSIM) tests were run to further confirm that HCC and each covariate holds the assumptions of PERMANOVA. To explain, while PERMANOVA does not assume normality, it does assume equal beta dispersion between the variable's groups. Therefore, betadisper, which calculates the average distance of group members to the group centroid in multivariate space, along with a permutation test for homogeneity of multivariate dispersions (permutest), were run to check whether the variable's variance is homogenous. Since betadisper only accepts categorical data, only the HCC_24 Q1/Q4 quantiles variable and categorical covariates - breastfeeding at 2.5 months, infant sex, and mode of delivery- were tested. In cases where the betadisper showed that the beta dispersion between groups was significantly different, ANOSIM was run in addition to PERMANOVA for each categorical variable to further confirm this. Only the breastfeeding at 2.5 months covariate had non-equal beta dispersions, so ANOSIM was especially run for it (Figure 8). There were no significant results for breastfeeding with ANOSIM, which meant that its significance in the PERMANOVA analyses would be dubious.

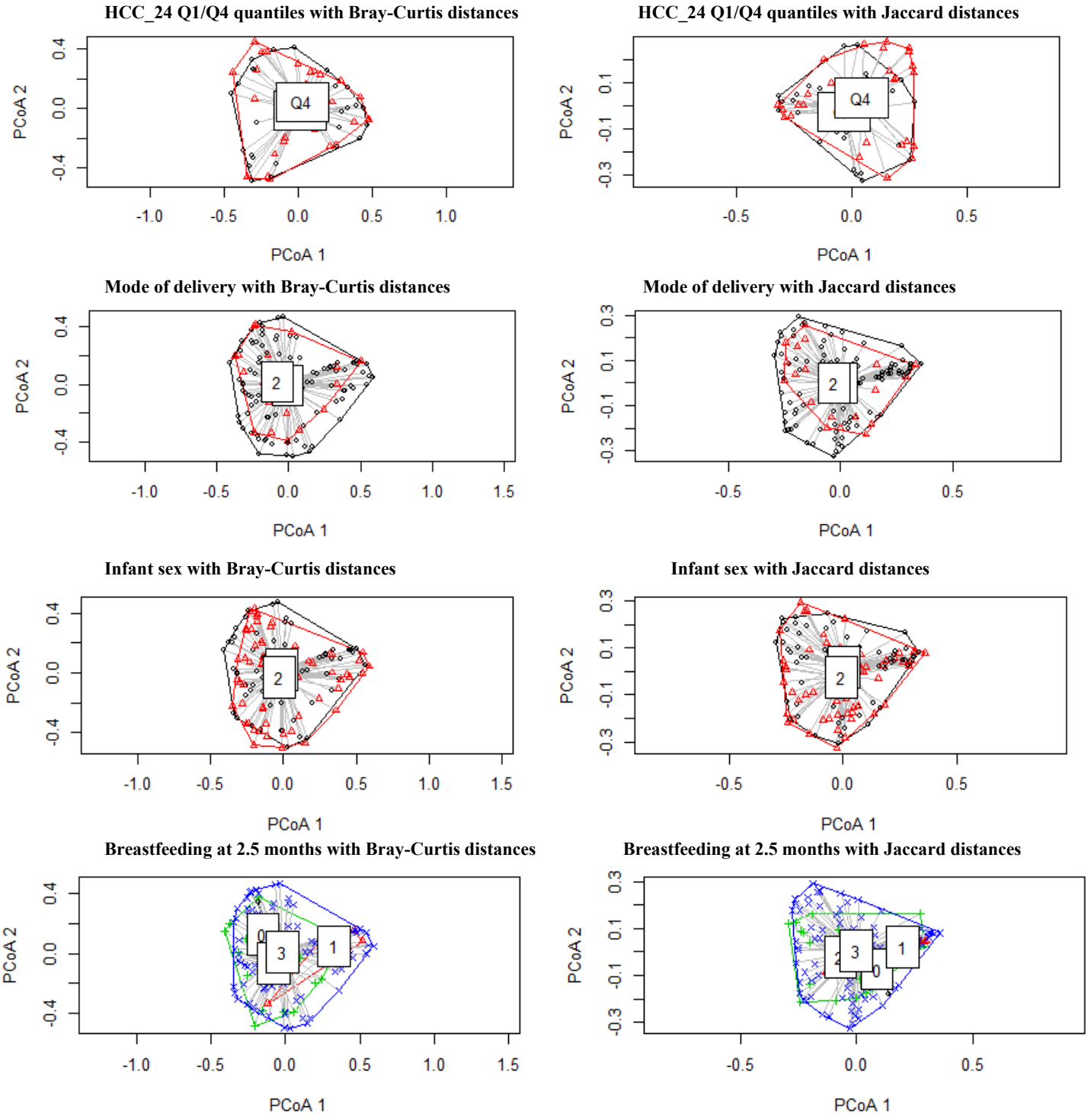


Figure 8. Plots of beta dispersion between variables' groups. Centroids of groups of all categorical variables - HCC_24 Q1/Q4 quantiles, breastfeeding at 2.5 months, infant sex, and mode of delivery, were in similar positions in the ordination space. However, the dispersions of breastfeeding at 2.5 months covariate groups differed more than for other covariates or HCC_24 Q1/Q4 quantiles.

Neither HCC_24 or the infant/maternal covariates, excluding the breastfeeding covariate, were significant for both Jaccard and Bray-Curtis distances in the PERMANOVA analysis run with HCC_24 (Table 9-10). The breastfeeding covariate was significant for both Jaccard and Bray-Curtis distances in the PERMANOVA analysis run with HCC_24, but because breastfeeding at 2.5 months has non-equal beta dispersions, the significance wasn't conclusive.

HCC_24 Q1/Q4 quantiles were significant for the PERMANOVA analysis run with the Jaccard distance (Table 11), and insignificant with the Bray-Curtis distance (p=0.087) (Table 12). ANOSIM run on the HCC_24 Q1/Q4 quantiles confirmed that HCC_24 Q1/ Q4 quantiles are statistically significant in the PERMANOVA analysis. Although, adjusted p-values were not significant for HCC_24 Q1/Q4 quantiles with either Bray-Curtis or Jaccard distances; it was around 0.07 for the HCC_24 Q1/Q4 quantiles variable with the Jaccard distance. The PERMANOVA analysis run with HCC_40 didn't yield any significant results.

Table 9. Jaccard distance matrix model fitted with HCC_24

	Degrees of freedom (Df)	**Sequential sums of squares** (SumsOfSqs)	**Mean squares** (MeanSqs)	**F statistic** (F.Model)	**R-squared** (R2)	**p-value** (Pr(>F))
HCC_24	1	0.425	0.425	1.108	0.009	0.201
Mother's age	1	0.375	0.375	0.977	0.008	0.488
Breastfeeding	3	1.349	0.450	1.172	0.029	0.032 *
Gender	1	0.367	0.367	0.957	0.008	0.542
Sample week	1	0.398	0.398	1.038	0.009	0.326
Birth mode	1	0.354	0.354	0.922	0.008	0.650
Residuals	111	42.581	0.384		-0.929	
Total	119	45.849			1	
Signif. codes: 0 = * 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "**						

Table 10. Bray-Curtis distance matrix model fitted with HCC_24

	Degrees of freedom (Df)	**Sequential sums of squares** (SumsOfSqs)	**Mean squares** (MeanSqs)	**F statistic** (F.Model)	**R-squared** (R2)	**p-value** (Pr(>F))
HCC_24	1	0.411	0.411	1.185	0.010	0.280
Mother's age	1	0.221	0.221	0.637	0.005	0.788
Breastfeeding	3	1.475	0.492	1.419	0.035	0.058 .
Gender	1	0.310	0.310	0.893	0.007	0.513
Sample week	1	0.427	0.427	1.231	0.010	0.255
Birth mode	1	0.356	0.356	1.026	0.009	0.403
Residuals	111	38.459	0.346		-0.923	
Total	119	41.658			1	
Signif. codes: 0 = * 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "**						

Table 11. Jaccard distance matrix model fitted with HCC_24 Q1/Q4 quantiles

	Degrees of freedom (Df)	**Sequential sums of squares** (SumsOfSqs)	**Mean squares** (MeanSqs)	**F statistic** (F.Model)	**R-squared** (R2)	**p-value** (Pr(>F))
HCC24_quantiles	1	0.595	0.595	1.552	0.026	0.006 **
Mother's age	1	0.414	0.414	1.082	0.018	0.247
Breastfeeding	3	1.073	0.358	0.933	0.047	0.796
Gender	1	0.341	0.341	0.890	0.015	0.756
Sample week	1	0.352	0.351	0.917	0.015	0.682
Birth mode	1	0.395	0.395	1.030	0.017	0.352
Residuals	51	19.542	0.383		-0.860	
Total	59	22.712			1	
Signif. codes: 0 = * 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "**						

Table 12. Bray-Curtis distance matrix model fitted with HCC_24 Q1/Q4 quantiles

	Degrees of freedom (Df)	**Sequential sums of squares** (SumsOfSqs)	**Mean squares** (MeanSqs)	**F statistic** (F.Model)	**R-squared** (R2)	**p-value** (Pr(>F))
HCC24_quantiles	1	0.572	0.571	1.669	0.028	0.091 .
Mother's age	1	0.414	0.414	1.208	0.020	0.275
Breastfeeding	3	0.943	0.314	0.918	0.047	0.609
Gender	1	0.252	0.251	0.734	0.012	0.685
Sample week	1	0.188	0.188	0.549	0.009	0.862
Birth mode	1	0.369	0.369	1.078	0.018	0.366
Residuals	51	17.463	0.342		-0.865	
Total	59	20.200			1	
Signif. codes: 0 = *** 0.001 = ** 0.01 = * 0.05 = . 0.1 = " " 1 = " "						

Investigation into which top taxa in HCC_24 Q1/Q4 quantiles contributed most to the community differences showed that *Bifidobacteriaceae*, *Lactobacillaceae, Clostridiaceae* taxa are dominant in the 1st quantile for both distances, while *Bacteroidaceae* is dominant in the 4th quantile for both distances. *Porphyromonadaceae* is dominant in the 4th quantile for only the Jaccard distance (Figure 9).

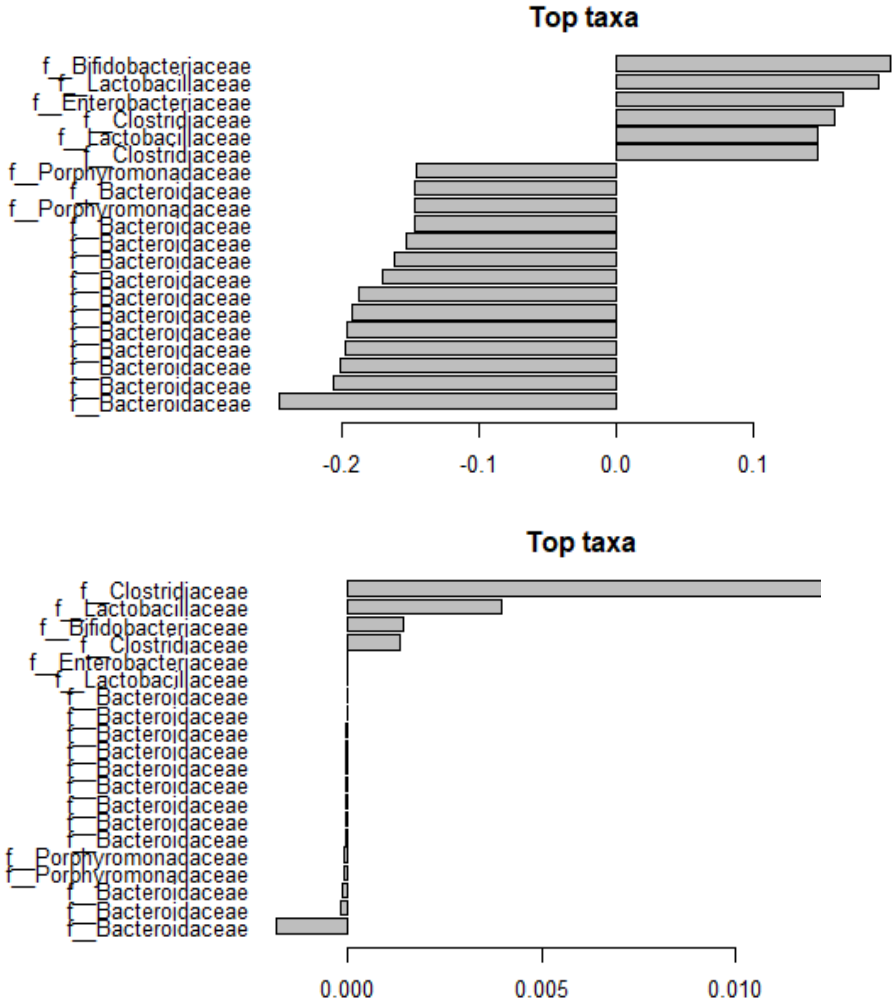


Figure 9. Top taxa in HCC_24 Q1/Q4 quantiles that contributed most to the community differences. Top taxa calculated by either Bray-Curtis or Jaccard for the 1st quantile are represented by the top and bottom diagrams, respectively. The top taxa members and their positive or negative association with the 1st quantile were identical for both distances.

4.Discussion

4.1 Covariate selection & linear regression analyses

Linear regression analyses revealed that HCC isn't associated with alpha-diversity. There were several issues as well regarding the covariate selection. Shannon and Chao1 models fitted with HCC_24 kept slightly different covariates in the final selected models, which made it harder to pick the optimal covariates. For example, the Chao1 model fitted with HCC_24 also kept the SSRI use by mother covariate, unlike the Shannon model fitted with HCC_24. Secondly, at least one significant maternal covariate had to be left out because the role of several maternal covariates was found to be contradictory in preliminary studies by the FinnBrain group. For example, season of HCC_24 sampling covariate was left out midway, even though it was present in the final models selected by AIC.

Thirdly, the final models selected during backward elimination may differ depending on the criterion used and its definition of goodness of fit, among which are AIC, Bayesian information criterion (BIC), and adjusted R-squared criteria. Both AIC and BIC are information criteria that aim for the simplest model with the greatest explanatory power, but usage of BIC would have led to simpler models because it applies a larger penalty for complex models than AIC; whereas adjusted R-squared aims for better models according to their predictive power rather than explanatory power [46]. For example, the last eliminated maternal covariate in the Shannon index model fitted with HCC_24 was the number of previous deliveries covariate, and it interestingly, contributed a lot to the Shannon index model's explained variance, aka adjusted R-squared, even though it wasn't picked during model selection by the AIC criterion and was never a statistically significant covariate in any of the linear regression models. Lastly, covariate selection depends on the method used, whose performance may be less or more conservative than that of other methods. Variable

selection methods can be divided into stepwise selection methods (forward selection, backward elimination), penalized regression methods (lasso, elastic net), Bayesian model averaging (BMA), etc. Penalized regression methods would have given more conservative estimates of coefficients, standard errors and number of variables than stepwise selection methods. BMA utilizes prior knowledge about the variables during the estimation procedure and would have produced more robust results than stepwise selection methods. Despite the limitations of stepwise selection methods, they remain the standard in epidemiology and other disciplines [47].

Past studies have reported active breastfeeding as being associated with lower gut microbiota diversity, and therefore the findings confirmed that partial (breastfeeding at 2.5 months group 2) and exclusive (breastfeeding at 2.5 months group 3) breastfeeding groups have negative associations with the alpha-diversity indices [36]. It was hard to assess the impact of mother's age on the infant gut microbiota. The HCC_24, HCC_40, HCC_24 Q1/Q4 quantiles linear regression models had mostly insignificant p-values for the F-test, suggesting that the adjusted R-squared values weren't optimal. Furthermore, the adjusted R-squared values were quite low. Only the Chao1 index model fitted with HCC_24 Q1/Q4 quantiles had a significant p-value for the F-test and a high adjusted R-squared value. These poor statistical values, in the case of the F-test and adjusted R-squared values, may have been obtained due to the low number of available samples. There were 120 HCC_24 samples, and only 20 HCC_40 samples. HCC_40 variable's very low sample size prevented its usage during covariate selection, in the DESeq2 analysis, and most likely affected the results' interpretation. A larger study may be required in order to improve the statistical power and to find an association between HCC and the alpha-diversity. Alternatively, HCC may not at all be associated with alpha-diversity.

4.2 Associations between the infant gut microbiota clusters and the covariates

Clustering of the infant gut microbiota into enterotypes revealed that HCC isn't associated with the clusters. Nevertheless, several infant covariates were significantly associated with the clusters after p-value adjustment, and most infant covariates had significant associations before p-value adjustment. Since the clusters represented the infant gut microbiota, associations between the clusters and infant covariates were expected. However, Chao1 index was significantly associated with the infant gut microbiota clusters while Shannon index wasn't, even though Chao1 and Shannon indices are both estimates of microbial diversity. Mode of delivery was significantly associated with the infant gut microbiota clusters, but for some reason hadn't been significant in the linear regression analyses of alpha-diversity. Interestingly, the number of previous deliveries covariate was significantly associated with the infant gut microbiota clusters, even though this covariate was a maternal covariate. This may explain why the number of previous deliveries covariate, absent in the final models selected by AIC and a non-significant covariate in the linear regression models, raised the Shannon index model's adjusted R-squared value.

4.3 Gut microbiota composition and HCC

When controlled for the following infant covariates - breastfeeding at 2.5 months, infant sex, and mode of delivery, The DESeq2 analysis revealed several associations between the infant gut microbiota and HCC_24. Nonetheless, sequencing wasn't conducted at the strain level due to its lack of reliability and wasn't always possible at the species level, which limited the resolution of the infant gut microbiota to the genera level and reduced the number of identified associations. The association patterns found in the overall model could only be compared with results from past studies examining the association of saliva, blood, urine cortisol concentrations or reported stress with the infant gut microbiota, since few studies have utilized HCC as a maternal prenatal PD marker. Past studies examining how each infant

covariate affects the gut microbiota diversity will be used below to roughly assess the observed association patterns in the subset models and the differences between two or more subset groups. For the most part, the subsetting approach has been problematic because it can only point out potential interactions by the selected covariate; interaction analyses in DESeq2 would have to be conducted to accurately assess the interaction between an infant covariate term and HCC_24.

As previously mentioned in the introduction part, infants of mothers with high cumulative stress, meaning mothers who had high levels of both reported stress and cortisol concentrations, had higher relative abundances of Proteobacteria, such as Escherichia, Serratia, and Enterobacter, and lower relative abundances of Bifidobacteria and lactic acid bacteria, such as *Lactoccus, Aerococcus*, and *Lactobacillus*. Furthermore, infants of mothers with high cumulative stress had a decreased abundance of Actinobacteria. The overall models' association patterns partially matched those seen in past studies. In detail, Actinobacteria and *Lactobacillus* were present and had negative associations with HCC_24 in the overall model [11].

The two mode of delivery subset groups followed very few of the patterns observed in past studies examining how mode of delivery affects the gut microbiota diversity. For example, the differences between important phyla such as Bacteroidetes, Firmicutes and Actinobacteria were occasionally opposite to findings from past studies. It was also hard to evaluate whether the composition of the vaginal delivered and caesarean section subset groups differed greatly, as there were a lot more associations for the vaginal delivered group than the caesarean section group. Most likely it was due to the different sizes of the subsets, with the vaginal delivery group being much bigger. According to one past study, cesarean section delivery was supposed to be associated with a higher abundance of Firmicutes, and lower abundance of Actinobacteria and Bacteroides in the first 3 months of life. Contrary to these findings, the vaginal delivery group had more associations with Firmicutes members than the cesarean section group. Furthermore, vaginally delivered infants are known to harbour more Bifidobacteria than cesarean section delivered infants, yet no Bifidobacteria were present in neither of the mode of delivery subset groups [48]. On the other hand, Bacteroidetes were present in the vaginal delivery group and completely absent in the caesarean section group.

The breastfeeding at 2.5 months subset groups followed several of the patterns observed in past studies examining how breastfeeding at 2.5 months affects the gut microbiota diversity. It was hard to evaluate whether the composition or diversity of the partial and exclusive breastfeeding subset groups differed greatly, as there were a lot more associations for the exclusive breastfeeding subset than the partial breastfeeding subset. Most likely it was due to the different sizes of the subsets, with the exclusive breastfeeding subset being much bigger. According to one past study, exclusive breastfeeding was associated with a higher relative abundance of *Bifidobacteriaceae* and Enterobacteriaceae and with less *Lachnospiraceae, Veillonellaceae*, and *Ruminococcaceae* [49]. Another past study demonstrated that non breastfed infants had a higher relative abundance of Peptostreptococcaceae and *Verrucomicrobiaceae* [50]. Both subset groups had associations with *Enterobacteriaceae, Lachnospiraceae*, and *Ruminococcaceae*. Interestingly, the exclusive breastfeeding group had associations with *Bifidobacteriaceae*, while the partial breastfeeding group had associations with *Verrucomicrobiaceae*. The differences between the two subsets should have been more noticeable because even a small degree of formula milk supplementation to breastfeeding at 2.5 months infants, which presumably happened in the case of the partial breastfeeding group, can change the gut microbiota pattern [49].

Past studies have demonstrated that the gut microbiota differs between male and female preterm infants. Female infants tend to have a more diverse gut microbiota, a higher abundance of *Clostridiales* and lower abundance of *Enterobacteriales* than male infants during early life [51]. According to another study, after 3 months of age males have a lower relative abundance of *Bacteroides* species than females [51, 52]. Results of the infant sex subset groups were contrary to findings from past studies, as few differences had been observed between the two genders in the infant sex subset, or they were opposite of those in past studies. For example, *Clostridiales* members were present in both subset groups, and they had both negative and positive associations with HCC_24. *Enterobacteriales* (*Citrobacter*) was present in the female group but not in the male group, while *Bacteroides* species were absent in both the female and male groups.

The lack of similarities between past study findings and any of the infant covariate subset results is mainly due to the problematic nature of the subsetting approach and different aims of the DESeq2 analysis. Instead of directly testing the association between some infant covariate and the infant gut microbiota, the infant covariate subset groups tested their respective gut microbiota for associations with HCC_24. The different associations with HCC_24 in subset groups indicate that mode of delivery and breastfeeding at 2.5 months may interact with HCC and future studies should consider them as interaction terms in interaction analyses. Sequencing using the shotgun sequencing method, which permits the sequencing of entire organisms, would have offered better resolution of the infant gut microbiota at species and strain level, and hence would've revealed more associations. The next step would be to determine whether there's any biologically meaningful associations, which could be tested for experimentally or by replication in independent future studies.

4.4 Association between HCC and beta-diversity of the infant gut microbiota

PERMANOVA was the only analysis in the entire thesis that clearly showed an association between HCC and the infant gut microbiota - in this case beta-diversity. The other analyses had at most revealed associations or correlations between infant or maternal covariates and the infant gut microbiota.

However, the inclusion of the breastfeeding at 2.5 months covariate in the PERMANOVA analyses may have affected the reliability of the results and they may be an artifact of heterogeneous dispersions, since the breastfeeding at 2.5 months covariate didn't have equal beta dispersion between its groups. In other words, the results could have been influenced by differences in composition within groups and not by the difference in composition between groups. In addition, the p-values of the covariate in the PERMANOVA analysis varied depending on the run and number of permutations, affecting the HCC_24 Q1/Q4 quantiles' significance in the process; although the HCC_24 Q1/Q4 quantiles always remained insignificant after p-value adjustment. The PERMANOVA analysis was also limited to

covariates selected during the backwards elimination of alpha-diversity linear regression models, therefore significant maternal covariates could have been left out from the PERMANOVA analyses.

The top taxa present in HCC_24 Q1/Q4 quantiles for both the Jaccard and Bray-Curtis distances partially matched the previously mentioned findings from past studies regarding the infant gut microbiota [11]. However, the top taxa, *Bacteroidaceae* and *Porphyromonadaceae*, were entirely absent in past studies and instead were found to have positive associations with other conditions such as malnourishment and Chrohn's disease. A past study, concerning the gut microbiota dysbiosis in children due to malnutrition, reported that malnourished children had an increase in abundance of *Bacteroidaceae* and *Porphyromonadaceae* bacteria [53].

5.Conclusion

This thesis aimed to investigate whether HCC is associated with the infant gut microbiota. Based on linear regression analyses of alpha-diversity and beta-diversity, analysis of individual genera, and cluster analysis of the gut microbiota, it can be concluded that HCC is associated with individual infant gut genera and perhaps with the infant gut microbiota's beta-diversity. Association patterns of the DESeq2 overall models and the top taxa that contributed most to the community differences in HCC_24 Q1/Q4 quantiles partially matched those seen in a past study examining the association of saliva cortisol concentrations with the infant gut microbiota. A larger, independent cohort study and better sequencing resolution, obtained using the shotgun sequencing method, are required to confirm the association between HCC and beta-diversity.

6.References

1. Karlsson, L., Tolvanen, M., Scheinin, N.M., Uusitupa, H.M., Korja, R., Ekholm, E., Tuulari, J.J., Pajulo, M., Huotilainen, M., Paunio, T. and Karlsson, H., 2018. Cohort profile: the FinnBrain birth cohort study (FinnBrain). *International Journal of Epidemiology*, 47(1), pp.15-16j.

2. O'Mahony, S.M., Clarke, G., Dinan, T.G. and Cryan, J.F., 2017. Early-life adversity and brain development: Is the microbiome a missing piece of the puzzle?. *Neuroscience*, 342, pp.37-54.

3. Mustonen, P., Karlsson, L., Scheinin, N.M., Kortesluoma, S., Coimbra, B., Rodrigues, A.J. and Karlsson, H., 2018. Hair cortisol concentration (HCC) as a measure for prenatal psychological distress—A systematic review. *Psychoneuroendocrinology*, 92, pp.21-28.

4. Papadimitriou, A. and Priftis, K.N., 2009. Regulation of the hypothalamic-pituitary-adrenal axis. *Neuroimmunomodulation,* 16(5), pp.265-271.

5. Golubeva, A.V., Crampton, S., Desbonnet, L., Edge, D., O'Sullivan, O., Lomasney, K.W., Zhdanov, A.V., Crispie, F., Moloney, R.D., Borre, Y.E. and Cotter, P.D., 2015. Prenatal stress-induced alterations in major physiological systems correlate with gut microbiota composition in adulthood. *Psychoneuroendocrinology*, 60, pp.58-74.

6. Van den Bergh, B.R., van den Heuvel, M.I., Lahti, M., Braeken, M., de Rooij, S.R., Entringer, S., Hoyer, D., Roseboom, T., Räikkönen, K., King, S. and Schwab, M., 2017. Prenatal developmental origins of behavior and mental health: The influence of maternal stress in pregnancy. *Neuroscience & Biobehavioral Reviews.*

7. Voegtline, K.M., Costigan, K.A., Kivlighan, K.T., Laudenslager, M.L., Henderson, J.L. and DiPietro, J.A., 2013. Concurrent levels of maternal salivary cortisol are unrelated to self-reported psychological measures in low-risk pregnant women. *Archives of Women's Mental Health,* 16(2), pp.101-108.

8. Pluess, M., Wurmser, H., Buske-Kirschbaum, A., Papousek, M., Pirke, K.M., Hellhammer, D. and Bolten, M., 2012. Positive life events predict salivary cortisol in pregnant women. *Psychoneuroendocrinology,* 37(8), pp.1336-1340.

9. D'Anna-Hernandez, K.L., Ross, R.G., Natvig, C.L. and Laudenslager, M.L., 2011. Hair cortisol levels as a retrospective marker of hypothalamic–pituitary axis activity throughout pregnancy: comparison to salivary cortisol. *Physiology & Behavior,* 104(2), pp.348-353.

10. Aatsinki, A.K., Keskitalo, A., Laitinen, V., Munukka, E., Uusitupa, H.M., Lahti, L., Kortesluoma, S., Mustonen, P., Rodrigues, A.J., Coimbra, B. and Huovinen, P., 2020. Maternal prenatal psychological distress and hair cortisol levels associate with infant fecal microbiota composition at 2.5 months of age. *Psychoneuroendocrinology*, p.104754.

11. Zijlmans, M.A., Korpela, K., Riksen-Walraven, J.M., de Vos, W.M. and de Weerth, C., 2015. Maternal prenatal stress is associated with the infant intestinal microbiota. *Psychoneuroendocrinology*, 53, pp.233-245.

12. Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A. And Hyde, E.R., 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), p.27.

13. Hamady, M. and Knight, R., 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*, 19(7), pp.1141-1152.

14. Waite, T.A. and Campbell, L.G., 2006. Controlling the false discovery rate and increasing statistical power in ecological studies. *Ecoscience*, 13(4), pp.439-442.

15. Koh, H., 2018. An adaptive microbiome α-diversity-based association analysis method. *Scientific Reports*, 8(1), pp.1-12.

16. Pellens, R. and Grandcolas, P., 2016. Biodiversity conservation and phylogenetic systematics: preserving our evolutionary heritage in an extinction crisis (p. 390). *Springer Nature.*

17. Kim, B.R., Shin, J., Guevarra, R., Lee, J.H., Kim, D.W., Seol, K.H., Lee, J.H., Kim, H.B. and Isaacson, R.E., 2017. Deciphering diversity indices for a better understanding of microbial communities. *Journal of Microbiology and Biotechnology*, 27(12), pp.2089-2093.

18. Thorpe, R. and Holt, R. eds., 2007. *The SAGE Dictionary of Qualitative Management Research.* Sage.

19. Allen, M.P., 1997. Testing hypotheses in nested regression models. *Understanding Regression Analysis*, pp.113-117.

20. Turner, P., 2006. Response surfaces for an F-test for cointegration. *Applied Economics Letters*, 13(8), pp.479-482.

21. Kelly, B.J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J.D., Collman, R.G., Bushman, F.D. and Li, H., 2015. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics*, 31(15), pp.2461-2468.

22. Gloor, G.B., Wu, J.R., Pawlowsky-Glahn, V. and Egozcue, J.J., 2016. It's All Relative: Analyzing Microbiome Data as Compositions. *Annals of Epidemiology*, 26(5), pp.322-329.

23. Legendre, P. and Anderson, M.J., 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69(1), pp.1-24.

24. Xia, Y., Sun, J. and Chen, D.G., 2018. *Statistical Analysis of Microbiome Data with R*. Singapore:: Springer.

25. Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C. and Ley, R.E., 2013. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLOS Computational Biology*, 9(1), p.e1002863.

26. Aho, K., 2006. Multivariate clustering for objective classification of vegetation data. *Proceedings America Society of Mining and Reclamation*, pp.1-23.

27. Rodríguez-Casado, C.I., Monleón-Getino, T., Cubedo, M. and Ríos-Alcolea, M., 2017. A priori groups based on Bhattacharyya distance and partitioning around medoids algorithm (PAM) with applications to metagenomics. *Journal of Mathematics* (Accepted).

28. Tibshirani, R., Walther, G. and Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), pp.411-423.

29. de Cárcer, D.A., Denman, S.E., McSweeney, C. and Morrison, M., 2011. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. *Applied and Environmental Microbiology*, 77(24), pp.8795-8798.

30. Buttigieg, P.L. and Ramette, A., 2014. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiology Ecology*, 90(3), pp.543-550.

31. García-Jiménez, B. and Wilkinson, M.D., 2019. Robust and automatic definition of microbiome states. *PeerJ*, 7, p.e6657.

32. Luo, H., Li, J., Chia, B.K.H., Robson, P. and Nagarajan, N., 2014. The importance of study design for detecting differentially abundant features in high-throughput experiments. *Genome Biology*, 15(12), p.527.

33. McMurdie, P.J. and Holmes, S., 2014. Waste not, want not: why rarefying microbiome data is inadmissible. *PLOS Computational Biology*, 10(4), p.e1003531.

34. Anders, S. and Huber, W., 2010. Differential expression analysis for sequence count data. *Nature Precedings*, pp.1-1.

35. Rintala, A., Riikonen, I., Toivonen, A., Pietilä, S., Munukka, E., Pursiheimo, J.-P., Elo, L.L.,Arikoski, P., Luopajärvi, K., Schwab, U., Uusitupa, M., Heinonen, S., Savilahti, E., Eerola, E., Ilonen, J., 2018. Early fecal microbiota composition in children who later develop celiac disease and associated autoimmunity. *Scandinavian Journal of Gastroenterology*, 53, pp.403–409.

36. Aatsinki, A.K., Lahti, L., Uusitupa, H.M., Munukka, E., Keskitalo, A., Nolvi, S., O'Mahony, S., Pietilä, S., Elo, L.L., Eerola, E. And Karlsson, H., 2019. Gut microbiota composition is associated with temperament traits in infants. *Brain, Behavior, and Immunity*, 80, pp.849-858.

37. Davenport, M.D., Tiefenbacher, S., Lutz, C.K., Novak, M.A. and Meyer, J.S., 2006. Analysis of endogenous cortisol concentrations in the hair of rhesus macaques. *General and Comparative Endocrinology*, 147(3), pp.255-261.

38. Benjamini, Y. and Hochberg, Y., 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics,* 25(1)*,* pp.60-83.

39. Lindsey, C. and Sheather, S., 2010. Variable selection in linear regression. *The Stata Journal,* 10(4), pp.650-669.

40. McMurdie, P.J. and Holmes, S., 2013. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One*, 8(4).

41. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K., 2012. Cluster: cluster analysis basics and extensions. *R Package Version*, 1(2), p.56.

42. Love, M.I., Huber, W. and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), p.550.

43. Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H. and Oksanen, M.J., 2013. Package 'vegan'. *Community Ecology Package*, version, 2(9), pp.1-295.

44. Chambers, J. M. and Hastie, T. J., 1992. *Statistical Models in S.* Wadsworth & Brooks/Cole.

45. Jafari, M. and Ansari-Pour, N., 2019. Why, when and how to adjust your P values?. *Cell Journal (Yakhteh),* 20(4), p.604.

46. Perez-Alvarez, S., Gómez, G. and Brander, C., 2015. FARMS: a new algorithm for variable selection. *BioMed Research International.*

47. Morozova, O., Levina, O., Uusküla, A. and Heimer, R., 2015. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. *BMC Medical Research Methodology,* 15(1), p.71.

48. Rutayisire, E., Huang, K., Liu, Y. and Tao, F., 2016. The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life: a systematic review. *BMC Gastroenterology*, 16(1), p.86.

49. 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. and Mandhane, P.J., 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), pp.e181161-e181161.

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

51. Cong, X., Xu, W., Janton, S., Henderson, W.A., Matson, A., McGrath, J.M., Maas, K. and Graf, J., 2016. Gut microbiome developmental patterns in early life of preterm infants: impacts of feeding and gender. *PloS One*, 11(4).

52. Kozyrskyj, A.L., Kalu, R., Koleva, P.T. and Bridgman, S.L., 2016. Fetal programming of overweight through the microbiome: boys are disproportionately affected. *Journal of Developmental Origins of Health and Disease*, 7(1), pp.25-34.

53. Gupta, S.S., Mohammed, M.H., Ghosh, T.S., Kanungo, S., Nair, G.B. and Mande, S.S., 2011. Metagenome of the gut of a malnourished child. *Gut Pathogens*, 3(1), p.7.