---

output: pdf_document: default html_document: default word_document: default graphics: yes —

```
library(here)
```

```
## here() starts at /Users/henrikeckermann/workspace/research_master/minor_research_project/article/anal
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(papaja)
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(microbiome)
```

```
## Loading required package: phyloseq
```

```
##
## microbiome R package (microbiome.github.com)
##
##
##
##  Copyright (C) 2011-2018 Leo Lahti et al. <microbiome.github.io>
```

```
##
## Attaching package: 'microbiome'
```

```
## The following object is masked from 'package:base':
##
##     transform
```

```
source("https://raw.githubusercontent.com/HenrikEckermann/in_use/master/reporting.R")
```

```
##
## Attaching package: 'glue'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```r
load(here("data/cc_analyses_workspace.RData"))
```

3. Results

3.1. Microbiota composition

Thirteen genus like groups from the *Actinobacteria*, *Firmicutes*, *Proteobacteria* and phyla showed an average abundance of ≥ 0.05 in at least 20% of the samples (Figure 1). Overall the microbiota was dominated by *Bifidobacterium* with an average abundance of more than 50%. Followed by facultative anaerobic *Bacilli* (*Streptococcus spp*, *Enterococcus*, *Lactobacilllus* and *Granulicatella*). The general variation of relative abundance of all taxa was quite high, with *Bifidobacterium* for instance ranging from 0.2% to 89%.

Figure 2 shows microbiota composition (Aitchison distance) for the first four principal components within the CC (A) and within the noCC (B) group. The starting point of the arrow indicate the microbiota composition at T0, whereas the endpoint corresponds the composition at T1. The plots do not reveal a systematic shift due to CC attendance over time. Instead, microbiota composition development between time points appears to be highly individual.

```r
biplot_time_2 <- biplot(pseq.clr, split = "cc", connect_series = "time", otu_color = "#ef8a62")
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
## Warning in class(x) <- c(subclass, tibble_class): Setting class(x) to
## multiple strings ("tbl_df", "tbl", ...); result will no longer be an S4
## object
```

```r
# ggarrange(
#     biplot_cc[[1]] + xlim(-12.5, 5) + ggtitle('A'),
#     biplot_cc[[2]] + xlim(-12.5, 5) + ggtitle('B'),
#     biplot_cc[[3]] + xlim(-12.5, 5) + ggtitle('C'),
#     biplot_cc[[4]] + xlim(-12.5, 5) + ggtitle('D'),
#     nrow = 2, ncol = 2,
#     common.legend = T)

scaling_factor <- 10
library(patchwork)

(biplot_time_2[[1]] +
  scale_y_continuous(limits = c(-5.25, 8), sec.axis = ~./scaling_factor) +
  scale_x_continuous(limits = c(-12.5, 5), sec.axis = ~./scaling_factor) + ggtitle('A') |
  biplot_time_2[[2]] +
  scale_y_continuous(limits = c(-5.25, 8), sec.axis = ~./scaling_factor) +
  scale_x_continuous(limits = c(-12.5, 5), sec.axis = ~./scaling_factor) + ggtitle('B')) /
  (biplot_time_2[[3]] +
  scale_y_continuous(limits = c(-5.5, 7), sec.axis = ~./scaling_factor) +
  scale_x_continuous(limits = c(-6.25, 5), sec.axis = ~./scaling_factor) +
  ggtitle('C') |
  biplot_time_2[[4]] +
  scale_y_continuous(limits = c(-5.5, 7), sec.axis = ~./scaling_factor) +
  scale_x_continuous(limits = c(-12.5, 5), sec.axis = ~./scaling_factor) + ggtitle('D'))
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```

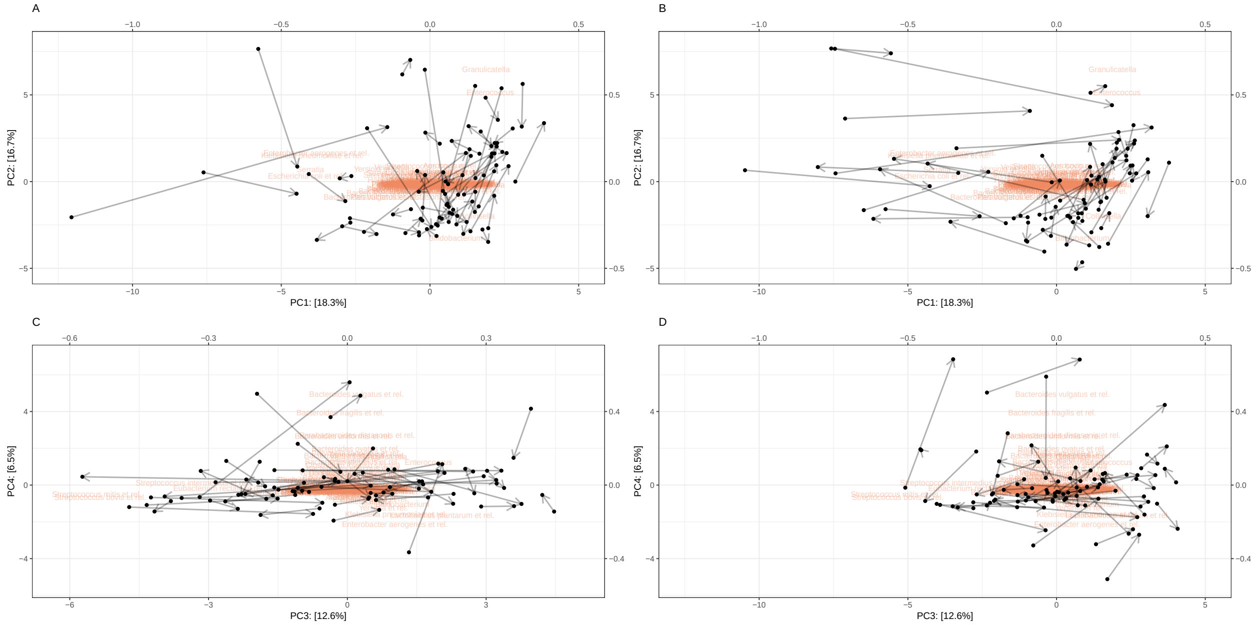

Figure 1: Development of microbiota composition over time within CC (A and C) and no CC (B and D).

```
## Scale for 'x' is already present. Adding another scale for 'x', which
## will replace the existing scale.
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which
## will replace the existing scale.
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which
## will replace the existing scale.
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which
## will replace the existing scale.
```

```
biplot_time_2 <- biplot(pseq.clr, facet = "cc", connect_series = "time")
```

```
## Warning in class(x) <- c(subclass, tibble_class): Setting class(x) to
## multiple strings ("tbl_df", "tbl", ...); result will no longer be an S4
## object
```

```
ggarrange(biplot_time_2[[1]], biplot_time_2[[1]], nrow = 2)
```

3.2 Permutational multivariate ANOVA

We performed PERMANOVA for 10 imputed datasets (see methods), which all yield similar results. We compared the overall community composition using PERMANOVA based on Aitchison distance metric. An

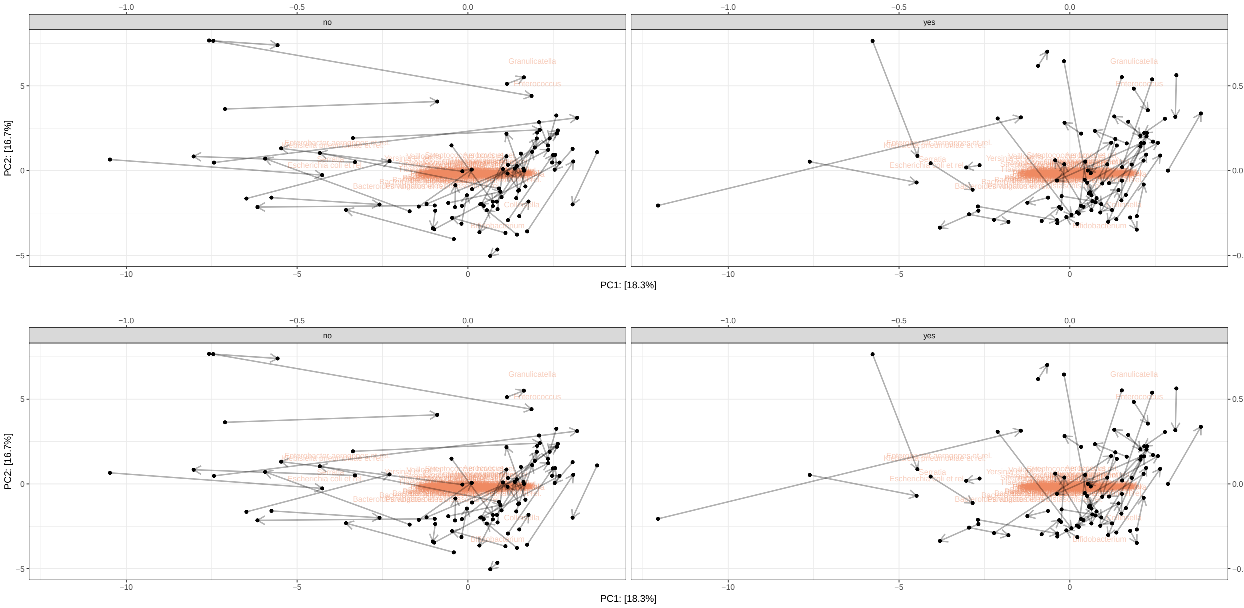

Figure 2: Development of microbiota composition over time within CC (A and C) and no CC (B and D).

assumption for PERMANOVA is multivariate homogeneity of group dispersions (variances) [@Anderson2017]. We used the function *betadisper* [@vegan2017], which utilizes the *PERMDISP2* procedure as implemented by Marti Anderson and found that this assumption was met for the factors *childcare* $F(1{,}194) = 0.19$, $p =$ .188, *time* $F(1{,}194) = 1.73$, $p = .190$ and the subgroups that result out of the interaction of *time* and *cc* $F(3{,}192) = 1.19$, $p = .313$. According to PERMANOVA, breastfeeding and age significantly predicted overall community composition (see table x). Figure 5 shows the genera that mostly changed as a function of each predictor. There is no significant effect of CC over time on overall community composition.

```r
apa_table(pm_table)
```

```
##
##
## \begin{table}[tbp]
## \begin{center}
## \begin{threeparttable}
## \caption{\label{tab:unnamed-chunk-3}}
## \begin{tabular}{lllllll}
## \toprule
## Model Parameter & \multicolumn{1}{c}{Sum of Squares} & \multicolumn{1}{c}{Mean Sum of Squares} & \mul
## \midrule
## time & 58.44 & 58.444 & 1.469 & 1.00 & 0.137 & 0.01\\
## cc & 52.64 & 52.636 & 1.323 & 1.00 & 0.181 & 0.01\\
## age\_d\_s & 79.55 & 79.553 & 2 & 1.00 & 0.037 & 0.01\\
## bf\_count\_s & 197.06 & 197.06 & 4.954 & 1.00 & 0.001 & 0.02\\
## time:cc & 36.23 & 36.229 & 0.911 & 1.00 & 0.486 & 0.00\\
## Residuals & 7,558.55 & 39.782 & - & 190.00 & - & 0.95\\
## Total & 7,982.47 & - & - & 195.00 & - & 1.00\\
## \bottomrule
## \end{tabular}
## \end{threeparttable}
## \end{center}
```

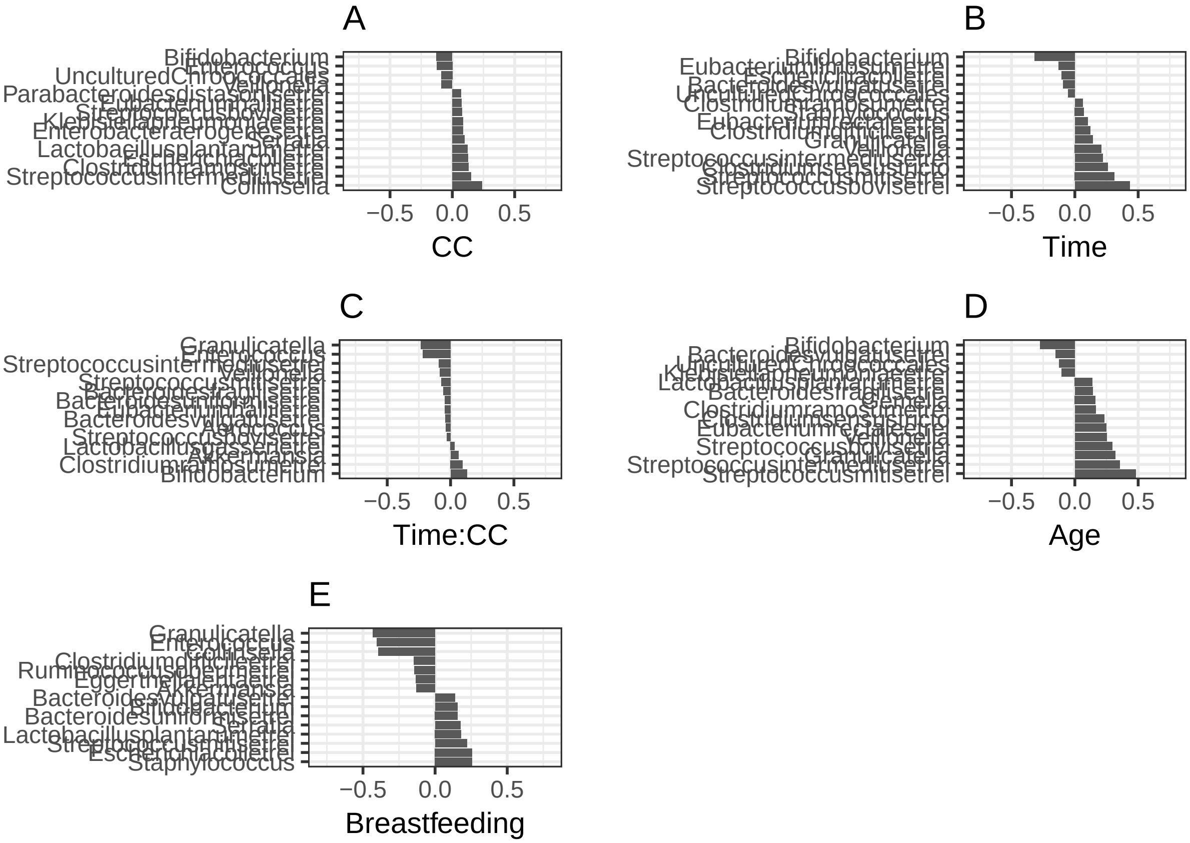

Figure 3: Top Taxa that differ most as a function of each predictor. CC = childcare.

```
## \end{table}
```

```
ggarrange(
    pmps[[1]] + ggtitle('A'),
    pmps[[2]] + ggtitle('B'),
    pmps[[3]] + ggtitle('C'),
    pmps[[4]] + ggtitle('D'),
    pmps[[5]] + ggtitle('E'),
    nrow = 3, ncol = 2, common.legend = T)
```

3.3 Hierarchical linear models

3.3.1 Differential abundance with LME

We evaluated the model assumptions of homogeneity of variance and normality of residuals individually for each model by visual inspection of residual- and quantile-quantile plots. Importantly, most of our models showed moderate violations of the homogeneity of variance and normality of residual assumptions. Several models violated both model assumptions severely. In contrast, the Bayesian generalized linear models that will be described (3.2.2) are more flexible and seem more appropriate to model the skewed and long tailed distributions of clr-transformed bacterial abundance.

We set contrasts in R such that the intercept reflects the CC group at timepoint “post” and the coefficients of the the model (cc and time) reflect the desired group comparisons. In the frequentist paradigm, our hypothesis would predict that both coefficients are significantly different from zero for more bacterial genus abundances than expected by chance. We adjusted for multiple testing based on @benjamini1995 allowing for 20% of false discoveries. Table x shows all coefficients (incl. covariates) that remained significant after adjusting for multiple testing ($q \leq .2$). The LMEs indicate a small signal of CC on the abundance of some genera.

```r
library(papaja)
library(knitr)
library(kableExtra)
pm_table
```

```
##   Model Parameter Sum of Squares Mean Sum of Squares     F  Df     p
## 1            time         58.444              58.444 1.469   1 0.137
## 2              cc         52.636              52.636 1.323   1 0.181
## 3         age_d_s         79.553              79.553     2   1 0.037
## 4      bf_count_s        197.060              197.06 4.954   1 0.001
## 5         time:cc         36.229              36.229 0.911   1 0.486
## 6       Residuals       7558.549              39.782     - 190     -
## 7           Total       7982.471                   -     - 195     -
##   R Square
## 1    0.007
## 2    0.007
## 3    0.010
## 4    0.025
## 5    0.005
## 6    0.947
## 7    1.000
```

```r
apa_table(pm_table)
```

```
## 
## 
## \begin{table}[tbp]
## \begin{center}
## \begin{threeparttable}
## \caption{\label{tab:unnamed-chunk-5}}
## \begin{tabular}{lllllll}
## \toprule
## Model Parameter & \multicolumn{1}{c}{Sum of Squares} & \multicolumn{1}{c}{Mean Sum of Squares} & \mul
## \midrule
## time & 58.44 & 58.444 & 1.469 & 1.00 & 0.137 & 0.01\\
## cc & 52.64 & 52.636 & 1.323 & 1.00 & 0.181 & 0.01\\
## age\_d\_s & 79.55 & 79.553 & 2 & 1.00 & 0.037 & 0.01\\
## bf\_count\_s & 197.06 & 197.06 & 4.954 & 1.00 & 0.001 & 0.02\\
## time:cc & 36.23 & 36.229 & 0.911 & 1.00 & 0.486 & 0.00\\
## Residuals & 7,558.55 & 39.782 & - & 190.00 & - & 0.95\\
## Total & 7,982.47 & - & - & 195.00 & - & 1.00\\
## \bottomrule
## \end{tabular}
## \end{threeparttable}
## \end{center}
## \end{table}
```

```r
comp_all_df <- comp_all_df %>% arrange(comparison)

kable(comp_all_df, "latex", caption = "Comparison of mu between groups",
  booktabs = T) %>%
    kable_styling() %>%
    group_rows("Age", 1, 20) %>%
    group_rows("Breastfeeding", 21, 25)
```

3.3.2 Differential abundance with Bayesian GLM

We modeled clr-transformed bacterial abundance using the generalized gaussian distribution with constant variance σ and skewness parameter α. As for the LMEs, the parameter μ is modeled as a linear function of the predictors “cc”, “time”, “breastfeeding” and “age”. Note that μ no longer represents the mean as $\alpha \neq 0$. In line with our hypothesis, we found that there were no differences in μ between CC and noCC at T0 as well as between T0 and T1 within the noCC group. But μ was different between “cc” and “nocc” at time “post” and when comparing “pre” to “post” within CC for more genera than could be expected by chance. Figures x-x show the mean differences of μ for those comparisons including 95% highest posterior density interval (HPDI). Given our assumptions, if the probability that μ of one group is different from the other is greater than 95%, 99% or 99.9%, this is indicated by “*“,”**" and “***“, respectively.

3.3.2 Alpha diversity

Alpha diversity indices (Shannon, Inverse Simpson and Gini-Simpson) were calculated using the *microbiome* package before the clr-transformation. LMEs and Bayesian hierarchical linear regression reached similar results. Here, we only present results of the Bayesian models as pooling of the parameter estimates after multiple imputation is straightforward. We found that after controlling for breastfeeding and age, the alpha-diversity was slightly lower for infants after CC attendance compared to all other groups. Table x shows the estimated difference in the parameter μ of the generalized normal distribution between groups for each diversity index. Assuming alpha-diversity is gaussian distributed with constant parameters σ (variance) and α (skewness parameter), the p-value represents the probability that μ is lower for the group in the left column compared to the group in the right column. Figure 3 shows alpha diversity for each subject (black rectangles) for each subgroup. It furthermore shows the posterior distribution of μ (red dots) including the narrowest interval containing 95% of the probability mass (highest posterior density interval). We see that μ was highest in the CC group before entrance and lowest of all groups after CC attendance. However, we see large individual variation within each group and the difference in μ is small.

```
# ggarrange(p_div[[1]] + ggtitle("A"),
#           p_div[[2]] + ggtitle("B"),
#           p_div[[3]] + ggtitle("C"),
#           ncol = 3)
```

3.4 Random Forest

RF is a tree based ensemble learning method that is well suited for classification based on microbiome data [@knightsSupervisedClassificationHuman2011]. We randomly selected 80% of the collected samples that constituted the training data set. We first tuned the RF models based on out-of-bag error. Node splitting was based on the gini criterion. Then we evaluated whether we can correctly classify CC based on 130 clr-transformed genus abundances using the hold out set. According to our hypotheses, we would expect to be able to classify whether an infant in the test data set belongs to the CC group at T1. In contrast, at T0 we would expect prediction accuracy to be lower since there were no differences between CC groups based on the (demographic) variables obtained. However, neither the T0 model, nor the model for T1 achieved a high prediction accuracy suggesting that there was no systematic effect of childcare entrance on microbiota composition. Table 2 shows the confusion matrix for each model.

Table 1: Comparison of mu between groups

comparison	mean	lower	upper	prob	genus
Age					
age	-0.0322475	-0.1067309	0.0458077	0.801075	Actinomycetaceae
age	0.0527438	-0.0201235	0.1215064	0.073400	Aerococcus
age	-0.0179310	-0.0444726	0.0090222	0.906600	Aeromonas
age	-0.0294980	-0.1211939	0.0603634	0.741300	Akkermansia
age	-0.0255466	-0.0675923	0.0159011	0.887500	Alcaligenesfaecalisetrel
age	-0.0521341	-0.1093420	0.0061089	0.962200	Allistipesetrel
age	-0.0283239	-0.0596302	0.0030984	0.963025	Anaerobiospirillum
age	0.0026660	-0.0462504	0.0510208	0.453275	Anaerofustis
age	0.0076369	-0.0579420	0.0741669	0.409650	Anaerostipescaccaeetrel
age	-0.0104363	-0.0334147	0.0134173	0.809325	Anaerotruncuscolihominisetrel
age	0.0009257	-0.0396629	0.0397777	0.482325	Anaerovoraxodorimutansetrel
age	-0.0096374	-0.0347590	0.0145815	0.778775	Aneurinibacillus
age	-0.0569183	-0.0966163	-0.0159325	0.996850	Aquabacterium
age	-0.0133486	-0.0363521	0.0101089	0.873900	Asteroleplasmaetrel
age	-0.0450606	-0.0771109	-0.0137424	0.996750	Bacillus
age	-0.0166799	-0.1149051	0.0878193	0.637250	Bacteroidesfragilisetrel
age	-0.0473330	-0.1021340	0.0087002	0.952600	Bacteroidesintestinalisetrel
age	-0.0445769	-0.1149947	0.0277532	0.888400	Bacteroidesovatusetrel
age	-0.0411351	-0.0880522	0.0056354	0.957475	Bacteroidesplebeiusetrel
age	-0.0341478	-0.0684289	-0.0013961	0.977275	Bacteroidessplachnicusetrel
Breastfeeding					
age	-0.0413594	-0.0808608	0.0003953	0.977125	Bacteroidesstercorisetrel
age	-0.0346522	-0.1224964	0.0540577	0.785250	Bacteroidesuniformisetrel
age	-0.0421769	-0.1675406	0.0799988	0.751800	Bacteroidesvulgatusetrel
age	-0.0917980	-0.2657059	0.0888700	0.846325	Bifidobacterium
age	-0.0252226	-0.0525612	0.0010603	0.966300	Bilophilaetrel
age	-0.0066925	-0.0339682	0.0195620	0.688175	Brachyspira
age	0.0291910	-0.0322600	0.0915677	0.175700	Bryantellaformatexigensetrel
age	-0.0271341	-0.0753811	0.0233852	0.862825	Bulleidiamooreietrel
age	-0.0790701	-0.1618650	0.0010880	0.970425	Burkholderia
age	-0.0087939	-0.0317142	0.0138210	0.780025	Campylobacter
age	-0.0214210	-0.0424097	-0.0009025	0.977250	Clostridiumcellulosietrel
age	-0.0122520	-0.0350221	0.0096257	0.858850	Clostridiumcolinumetrel
age	-0.0046310	-0.0863117	0.0789585	0.541850	Clostridiumdifficileetrel
age	-0.0191654	-0.0418997	0.0032304	0.954525	Clostridiumfelsineumetrel
age	-0.0073150	-0.0494394	0.0354963	0.633125	Clostridiumleptumetrel
age	-0.0035319	-0.0595272	0.0520426	0.549900	Clostridiumorbiscindensetrel
age	0.0060026	-0.0920970	0.0985439	0.444125	Clostridiumramosumetrel
age	0.0449726	-0.0356547	0.1233454	0.132225	Clostridiumsensustricto
age	0.0122540	-0.0262963	0.0510914	0.267150	Clostridiumsphenoidesetrel
age	-0.0369770	-0.0669962	-0.0078347	0.993325	Clostridiumstercorariumetrel
age	0.0385096	-0.0147057	0.0897579	0.075350	Clostridiumsymbiosumetrel
age	-0.0222484	-0.0463215	0.0015739	0.963675	Clostridiumthermocellumetrel
age	-0.0803421	-0.2326696	0.0704158	0.855975	Collinsella
age	-0.0267265	-0.0586701	0.0052510	0.949175	Coprobacilluscatenaformisetrel
age	0.0356865	-0.0322667	0.1043238	0.151300	Coprococcuseutactusetrel
age	-0.0628704	-0.1101462	-0.0139812	0.993375	Corynebacterium
age	-0.0259504	-0.0593647	0.0069217	0.937950	Desulfovibrioetrel
age	-0.0101323	-0.0335226	0.0120286	0.810225	Dialister
age	0.0258282	-0.0410088	0.0949632	0.224875	Doreaformicigeneransetrel
age	-0.0275933	-0.1372484	0.0804536	0.691900	Eggerthellalentaetrel
age	-0.0650985	-0.1978826	0.0644075	0.838625	Enterobacteraerogenesetrel
age	0.2103005	0.0010579	0.4219027	0.024650	Enterococcus
age	-0.0253141	-0.1667674	0.1124008	0.642850	Escherichiacolietrel
age	-0.0224829	-0.0604072	0.0163216	0.879050	Eubacteriumbiformeetrel
age	-0.0271455	-0.0775703	0.0241278	0.854800	Eubacteriumcylindroidesetrel