library(here)

## here() starts at /Users/henrikeckermann/workspace/research\_master/minor\_research\_project/article/analyses/bibo

library(ggplot2)  
library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ tibble 1.4.2 ✔ purrr 0.2.5  
## ✔ tidyr 0.8.2 ✔ dplyr 0.7.8  
## ✔ readr 1.3.1 ✔ stringr 1.3.1  
## ✔ tibble 1.4.2 ✔ forcats 0.3.0

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ 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

load(here("data/cc\_analyses\_workspace.RData"))

# 3. Results

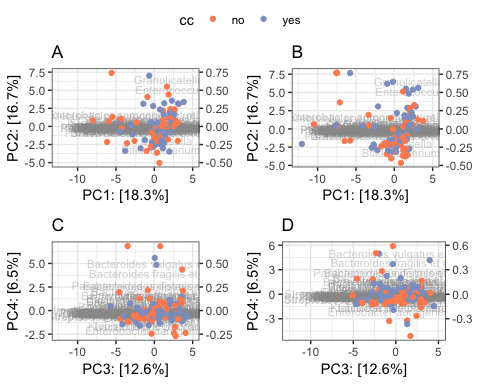
## 3.1. Microbiota composition

Thirteen genus like groups from the *Actinobacteria*, *Firmicutes*, *Proteobacteria* and phyla showed an average abundance of 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.

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)

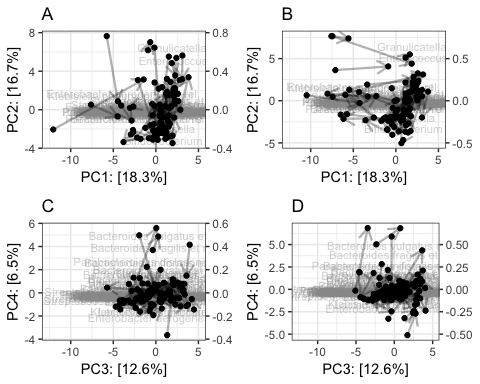
## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.  
## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.  
## Scale for 'x' is already present. Adding another scale for 'x', which  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



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

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

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

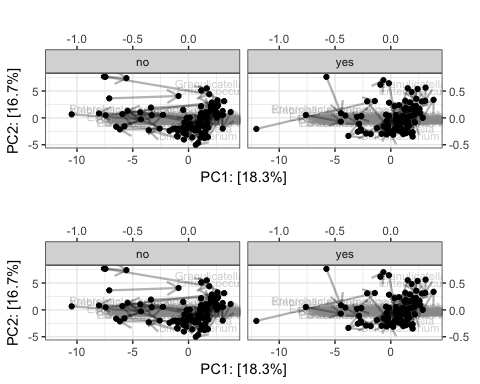


Development of microbiota composition over time within CC (A and C) and no CC (B and D).

biplot\_time\_2 <- biplot(pseq.clr, facet = "cc", connect\_series = "time")

## Warning in class(x) <- c(subclass, "tbl\_df", "tbl", "data.frame"): 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)



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

otus.clr %>% column\_to\_rownames("sample\_id") %>%  
 t() %>% as.data.frame() %>% summarise\_all(sum)

## sa\_10000 sa\_10038 sa\_10118 sa\_10211 sa\_10230  
## 1 2.131628e-14 -5.329071e-15 3.730349e-14 -1.199041e-14 7.105427e-15  
## sa\_10247 sa\_1025 sa\_10268 sa\_1028 sa\_1030  
## 1 -3.375078e-14 -3.019807e-14 -2.04281e-14 -4.973799e-14 4.440892e-16  
## sa\_1045 sa\_1054 sa\_10657 sa\_10742 sa\_10747  
## 1 -7.549517e-15 1.465494e-14 -1.554312e-14 -7.105427e-15 5.77316e-15  
## sa\_108 sa\_10812 sa\_10840 sa\_10944 sa\_11040  
## 1 1.776357e-14 3.907985e-14 5.373479e-14 7.549517e-15 -3.019807e-14  
## sa\_11136 sa\_11180 sa\_11183 sa\_11190 sa\_11210  
## 1 1.865175e-14 2.753353e-14 3.019807e-14 -1.110223e-14 -4.174439e-14  
## sa\_1127 sa\_11392 sa\_11488 sa\_11498 sa\_115  
## 1 -7.105427e-15 -5.151435e-14 2.220446e-14 -3.996803e-14 -5.595524e-14  
## sa\_11517 sa\_11615 sa\_11628 sa\_1164 sa\_11659  
## 1 1.865175e-14 7.105427e-15 -8.881784e-15 3.108624e-14 -5.684342e-14  
## sa\_1294 sa\_13 sa\_1389 sa\_1559 sa\_1577  
## 1 -4.884981e-14 -5.329071e-15 -5.151435e-14 -4.440892e-16 5.506706e-14  
## sa\_1659 sa\_1678 sa\_1743 sa\_1857 sa\_1866  
## 1 -2.975398e-14 2.131628e-14 1.199041e-14 -2.664535e-14 1.24345e-14  
## sa\_1939 sa\_1972 sa\_1999 sa\_2052 sa\_2116  
## 1 2.797762e-14 -1.909584e-14 1.021405e-14 7.993606e-15 4.707346e-14  
## sa\_2124 sa\_2298 sa\_2337 sa\_2384 sa\_2424  
## 1 8.437695e-15 4.796163e-14 7.993606e-15 -3.108624e-15 -2.442491e-14  
## sa\_2431 sa\_2495 sa\_2511 sa\_2541 sa\_2691  
## 1 5.595524e-14 5.062617e-14 1.776357e-14 -3.197442e-14 1.509903e-14  
## sa\_2717 sa\_2826 sa\_2864 sa\_2908 sa\_2952  
## 1 1.776357e-15 9.769963e-15 3.419487e-14 -4.440892e-14 -3.774758e-14  
## sa\_3146 sa\_3212 sa\_3232 sa\_3238 sa\_3407  
## 1 2.309264e-14 1.776357e-14 -6.217249e-15 7.105427e-15 -2.842171e-14  
## sa\_3434 sa\_3535 sa\_3548 sa\_3573 sa\_358  
## 1 8.881784e-15 -1.731948e-14 8.881784e-16 4.796163e-14 -5.373479e-14  
## sa\_3634 sa\_372 sa\_3758 sa\_3791 sa\_3833  
## 1 -2.176037e-14 4.618528e-14 -1.421085e-14 -5.018208e-14 -5.506706e-14  
## sa\_3873 sa\_3887 sa\_3892 sa\_3935 sa\_3947  
## 1 4.440892e-15 4.796163e-14 -2.131628e-14 -2.664535e-15 -9.325873e-15  
## sa\_4003 sa\_426 sa\_433 sa\_438 sa\_4416  
## 1 -2.575717e-14 -2.087219e-14 2.398082e-14 2.131628e-14 -5.373479e-14  
## sa\_4478 sa\_4511 sa\_4574 sa\_4630 sa\_4779  
## 1 -3.907985e-14 -2.664535e-14 -1.154632e-14 -4.263256e-14 1.776357e-14  
## sa\_4941 sa\_4966 sa\_4990 sa\_5002 sa\_5045  
## 1 -3.730349e-14 -3.019807e-14 5.684342e-14 -5.595524e-14 -2.88658e-14  
## sa\_5073 sa\_5131 sa\_5288 sa\_5374 sa\_5450  
## 1 5.551115e-14 -2.220446e-14 4.396483e-14 -1.776357e-15 -4.973799e-14  
## sa\_5453 sa\_5469 sa\_549 sa\_5512 sa\_553  
## 1 -2.176037e-14 5.506706e-14 -4.218847e-14 -4.973799e-14 2.04281e-14  
## sa\_5551 sa\_5637 sa\_5725 sa\_5740 sa\_5780  
## 1 9.325873e-15 -1.154632e-14 1.021405e-14 3.508305e-14 3.463896e-14  
## sa\_5818 sa\_5861 sa\_5864 sa\_5875 sa\_5906  
## 1 5.417888e-14 1.332268e-14 -3.68594e-14 4.796163e-14 -4.485301e-14  
## sa\_5928 sa\_5986 sa\_6067 sa\_6195 sa\_6223  
## 1 -4.396483e-14 -1.598721e-14 1.421085e-14 1.065814e-14 5.062617e-14  
## sa\_6328 sa\_6332 sa\_6369 sa\_6380 sa\_6430  
## 1 2.975398e-14 -3.197442e-14 5.506706e-14 4.707346e-14 -5.595524e-14  
## sa\_6648 sa\_6661 sa\_6743 sa\_6773 sa\_6852  
## 1 -1.776357e-15 2.930989e-14 -3.730349e-14 -4.218847e-14 3.996803e-14  
## sa\_6976 sa\_6989 sa\_7040 sa\_7116 sa\_7134  
## 1 -3.552714e-14 -1.509903e-14 7.105427e-15 -1.64313e-14 3.28626e-14  
## sa\_7161 sa\_7166 sa\_7168 sa\_7198 sa\_726  
## 1 1.332268e-14 3.064216e-14 -4.440892e-16 3.552714e-15 -5.151435e-14  
## sa\_7322 sa\_7340 sa\_7341 sa\_740 sa\_7417  
## 1 1.909584e-14 2.753353e-14 5.728751e-14 2.531308e-14 1.509903e-14  
## sa\_7422 sa\_7451 sa\_7459 sa\_7477 sa\_7538  
## 1 -4.796163e-14 -4.707346e-14 4.884981e-14 -1.865175e-14 -2.309264e-14  
## sa\_7599 sa\_7679 sa\_7759 sa\_7768 sa\_7778  
## 1 3.108624e-14 4.840572e-14 -4.352074e-14 0 1.065814e-14  
## sa\_7972 sa\_8026 sa\_8061 sa\_832 sa\_8342  
## 1 3.419487e-14 -3.108624e-15 -1.154632e-14 2.398082e-14 1.24345e-14  
## sa\_8657 sa\_8678 sa\_8703 sa\_8715 sa\_881  
## 1 3.28626e-14 -2.575717e-14 -9.325873e-15 1.332268e-14 -2.930989e-14  
## sa\_8818 sa\_8833 sa\_8856 sa\_886 sa\_8869  
## 1 -5.151435e-14 -3.463896e-14 1.820766e-14 3.952394e-14 2.664535e-14  
## sa\_8998 sa\_9001 sa\_9022 sa\_9030 sa\_9080  
## 1 1.24345e-14 -4.796163e-14 -2.753353e-14 -1.731948e-14 0  
## sa\_9091 sa\_9232 sa\_9343 sa\_9352 sa\_9408  
## 1 -1.776357e-14 3.730349e-14 6.217249e-15 -4.440892e-14 2.220446e-14  
## sa\_9412 sa\_9462 sa\_953 sa\_9543 sa\_9691  
## 1 1.776357e-15 2.975398e-14 -2.309264e-14 4.396483e-14 2.531308e-14  
## sa\_987 sa\_9913 sa\_9924 sa\_9951 sa\_9953  
## 1 4.52971e-14 5.329071e-15 3.996803e-14 -4.396483e-14 -5.684342e-14  
## sa\_9994  
## 1 -5.329071e-15

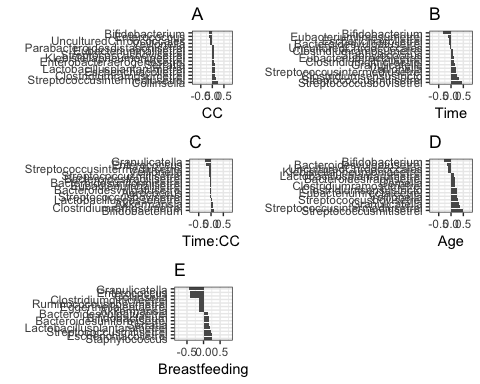
## 3.2 Permutational multivariate ANOVA

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 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.

apa\_table(pm\_table)

## <caption>(\#tab:unnamed-chunk-4)</caption>  
##   
## <caption>\*\*</caption>  
##   
##   
##   
## Model Parameter Sum of Squares Mean Sum of Squares F Df p R Square   
## ---------------- --------------- -------------------- ------ ------- ------ ---------  
## time 58.44 58.444 1.469 1.00 0.12 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 0.026 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 0.505 0.00   
## Residuals 7,558.55 39.782 - 190.00 - 0.95   
## Total 7,982.47 - - 195.00 - 1.00

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)



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

## 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 (). The LMEs indicate a small signal of CC on the abundance of some genera.

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 0.12  
## 2 cc 52.636 52.636 1.323 1 0.181  
## 3 age\_d\_s 79.553 79.553 2 1 0.026  
## 4 bf\_count\_s 197.060 197.06 4.954 1 0.001  
## 5 time:cc 36.229 36.229 0.911 1 0.505  
## 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

apa\_table(pm\_table)

## <caption>(\#tab:unnamed-chunk-6)</caption>  
##   
## <caption>\*\*</caption>  
##   
##   
##   
## Model Parameter Sum of Squares Mean Sum of Squares F Df p R Square   
## ---------------- --------------- -------------------- ------ ------- ------ ---------  
## time 58.44 58.444 1.469 1.00 0.12 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 0.026 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 0.505 0.00   
## Residuals 7,558.55 39.782 - 190.00 - 0.95   
## Total 7,982.47 - - 195.00 - 1.00

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 and skewness parameter . As for the LMEs, the parameter is modeled as a linear function of the predictors “cc”, “time”, “breastfeeding” and “age”. Note that no longer represents the mean as . In line with our hypothesis, we found that there were no differences in between CC and noCC at T0 as well as between T0 and T1 within the noCC group. But 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 for those comparisons including 95% highest posterior density interval (HPDI). Given our assumptions, if the probability that 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 of the generalized normal distribution between groups for each diversity index. Assuming alpha-diversity is gaussian distributed with constant parameters (variance) and (skewness parameter), the p-value represents the probability that 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 (red dots) including the narrowest interval containing 95% of the probability mass (highest posterior density interval). We see that 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 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.