D3C Statistical Analysis - Microbiome Submission

Lexi Ardissone

6/2/2017

Contents

Input & Data Preparation

Packages & Functions

```
source('~/Desktop/D3C/D3C_2.0/writing/Microbiome/Supplement/D3C_functions_5.0.R')
```

Below is a list of the R packages used in this analysis and are called in the source above. Appropriate references for these packages can be found in the *References* section at the end of this doucment. Also, the source provides specific functions that were written for this project; an attempt to note the usage of these functions will be made throughout this document.

- *knitr*
- *phyloseq*
- *ggplot2*
- *gridExtra*
- *plyr*
- *reshape2*
- *lubridate*
- *vegan*
- *geepack*
- *gdata*

Input Data

In this project, the V4 region of the 16S rRNA gene was amplified and sequenced using 2x100 paired-end sequencing on the Illumina HiSeq 2000 platform. The raw sequence data for this project has been deposited in NCBI's SRA under BioProject PRJNA232731.

Sequence preprocessing involved quality filtering and trimming (detailed in Davis-Richardson *et al.*), and the removal of samples with fewer than 20,000 reads (n = 6). This resulted in an average of 337,100 (±181,094) reads per sample with an average length of 2x98 (±3) nucleotides. OTUs were assigned and quantified by aligning sequences to the GreenGenes 97% representatives set version 13.8 (currently provided by Qiime) using the USEARCH program version 6.022. Taxonomic assignment of sequences required a minimum of 97% identity and 95% query coverage identifying a total of 49,455 OTUs after singletons were removed.

OTU filtering was applied based on evaluation of technical replicates; the details of the process can be found in Additional file 3. The primary input file for the analyses presented in this document are provided as Additional document 2.

```
load('~/Desktop/D3C/D3C_2.0/writing/Microbiome/SubmissionMaterials_final/AdditionalFiles/AdditionalFile2
dipp.sam3.genfilt
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 11583 taxa and 1760 samples ]
## sample_data() Sample Data:       [ 1760 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 11583 taxa by 8 taxonomic ranks ]
```

Data Preparation

Handling technical replicates

In the `dipp.sam3.genfilt` phyloseq object, there are `1760` samples, which include technical replicates. However, there are only `1494` unique samples in this data set.

```
#merge samples that were replicated
dipp.TR = merge_samples(dipp.sam3.genfilt, "sample_id")
  ##this function sums read counts, but ultimately proportions will be used,
  ##and differences in counts/sequencing effort would be of little concern.
  ##However, this would affect alpha diversity, so these samples are excluded from those measures
  #Note: illumina_id variable is lost and factor variables are converted to integers
dipp.TR
```

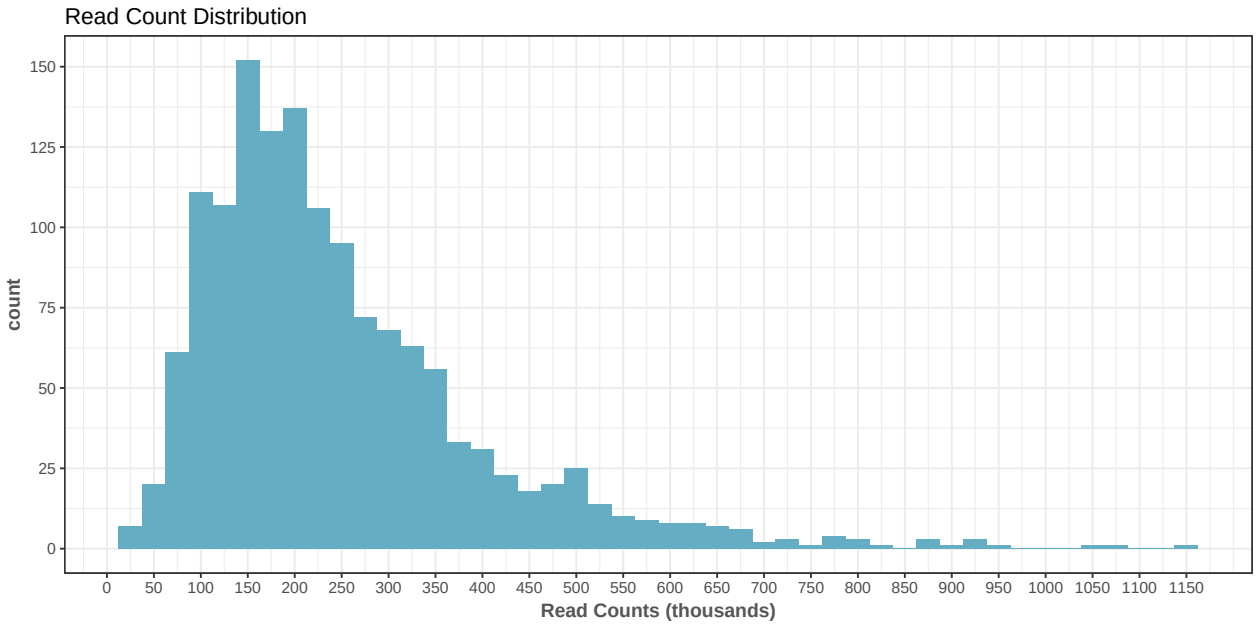
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 11583 taxa and 1494 samples ]
## sample_data() Sample Data:       [ 1494 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 11583 taxa by 8 taxonomic ranks ]
```

Distribution of sample counts

Although samples with fewer than 20,000 rads were initially excluded, this was applied prior to the removal of OTUs with high technical variability. Therefore, this section describes the total number of reads per sample used for further statistical analyses. Note, counts for replicated samples are inflated so are excluded here.

Summary statisitics for all read counts of all samples (N=1,494):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13420  146100  209500  248500  312700 1152000
```



Agglomerate taxa

Now that technical replicates have been dealt with, counts were transformed to proportions and using the `tax_glom()` function in the *phyloseq* package, OTUs were collapsed to phyla, family, genus, and species taxonomies. Warning: this chunk takes ~20-30 minutes to run.

```
dipp.TR.prop = transform_sample_counts(dipp.TR, function(x) x/sum(x))
dipp.p = tax_glom(dipp.TR.prop, 'Phylum')
dipp.f = tax_glom(dipp.TR.prop, 'Family')
dipp.g = tax_glom(dipp.TR.prop, 'Genus')
dipp.s = tax_glom(dipp.TR.prop, 'Species')
```

The table below enumerates the number of features at each taxonomic levels that were subject to analyses:

Level	Number of Taxa
Phylum	12
Family	91
Genus	180
Species	520
OTU	11583

Extract sample/subject data

Subject- and sample-specific meta data was extracted in order to evaluate the cohort. Additional variables were also generated in order to facilitate plotting. These included site-specific inner-quartile ranges (IQRs) and medians for the appearance of the first autoantibody, and a final_status variable. For aesthetic reasons, this code is included in the .Rmd file but not displayed in this document.

Figure 1 - Sampling plot

```
#Define color palette
fresh_swap= c('#233B43', '#65ADC2', '#E84646')

#plot
ggplot(dipp.samp, aes(factor(mask_id), age_at_sampling, shape=final_status, color=final_status)) +
  #PLOT BASICS
  geom_point(size=3) +
  coord_flip() +
  scale_color_manual(values = fresh_swap, name='Final Status') +
  scale_shape_discrete(name='Final Status') +
  theme_bw() +
  theme(legend.position = 'bottom') +
  #guide_legend(title = 'Final Status', title.position = 'bottom') +
  # guide=guide_legend(title='Final Status', title.position='bottom')
  facet_wrap(~site, scales='free') +
  #SC ANNOTATION
  geom_rect(aes(ymin=SeroSite1stIQR, ymax=SeroSite3rdIQR, xmin=-Inf, xmax=Inf),
            fill='grey', alpha=0.2, color='blue', lty='dashed') +
  geom_hline(aes(yintercept=SeroSiteMed), color='blue') +
  #GRID LAYOUT
  geom_hline(yintercept=seq(90,750, by=30.5), linetype='dashed', color='black') +
  geom_vline(xintercept=seq(0,134, by=1), linetype='dashed', color='gray') +
  #AXES LABELS & TITLE
  scale_y_continuous(breaks=seq(90,750, by=30.5), labels=seq(3,24,1)) +
  ylab("Age at sampling (months)") +
  xlab("Subject") +
  ggtitle('DIPP 3 Cities Sampling Frequency for samples included in analyses') +
  geom_point(size=3)
```

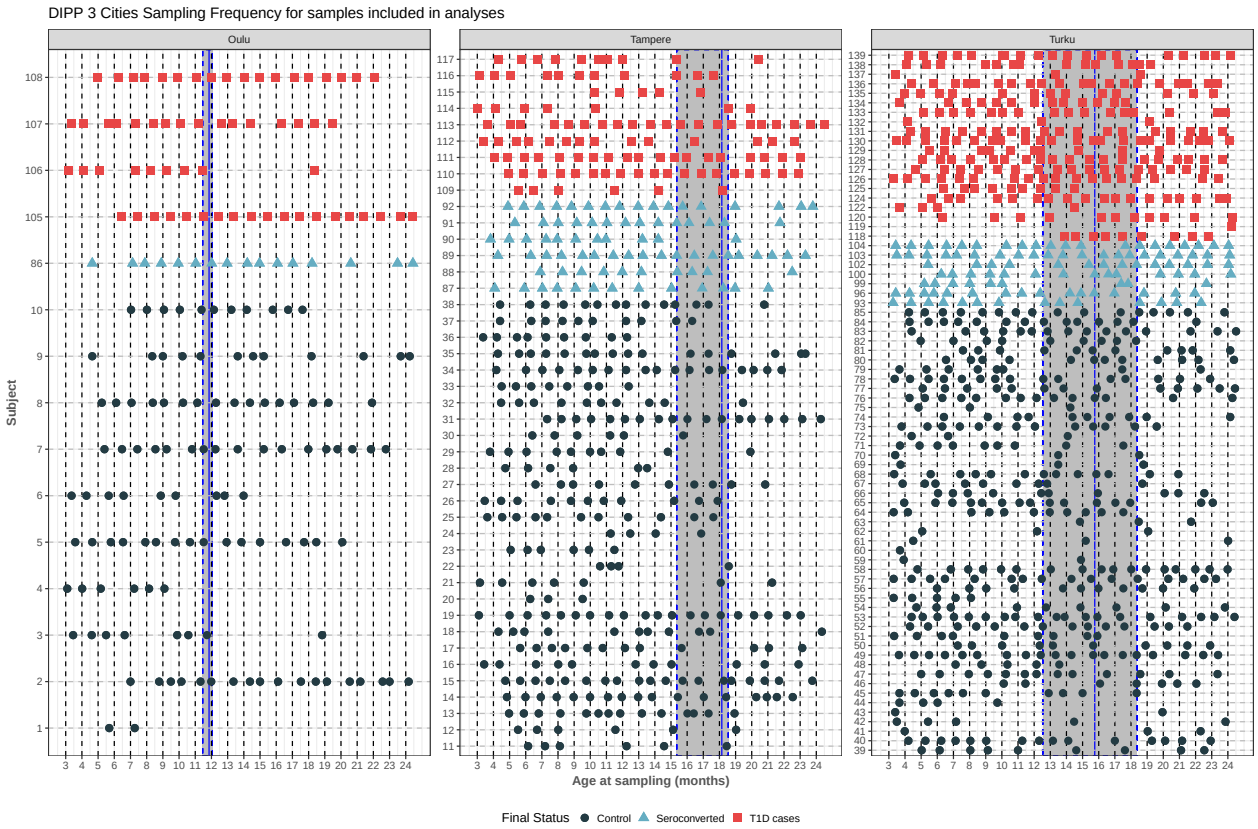


Figure 1. ***Sampling frequency*** A total of 132 subject were sampled from Oulu, Tampere, and Turku, Finland. Samples were collected between 3 and 24 months of age (vertical, dashed lines indicate monthly intervals). Each point represents a unique sample and each row represents a unique subject, which are ordered by those that did not develop autoantibodies (Control, dark blue circles), those that developed 2 or more persistent autoantibodies but did not progress to T1D (Seroconverted, light blue triangles), and those that developed 2 or more autoantibodies and progressed to T1D (T1D cases, red squares). The median age at which the first autoantibody appeared in each site is represented by the solid blue lines, and the inner-quartile range is represented by the light gray, shaded region with blue, dashed perimeter.

Comparison of Sample Data between Cities

Potential Confounders

There are several environmental factors that can influence microbial composition. Therefore, in this section variables of interest are evaluated and any differences in the distribution of these variables and islet autoimmunity outcome are considered. Additionally, differences between cities are also considered as this could confound comparisons between cases and controls between sites.

Code for generating table and plots in this section are supressed for aesthetic reasons but are available in Rmarkdown. Code for all statistical tests is displayed.

HLA risk group

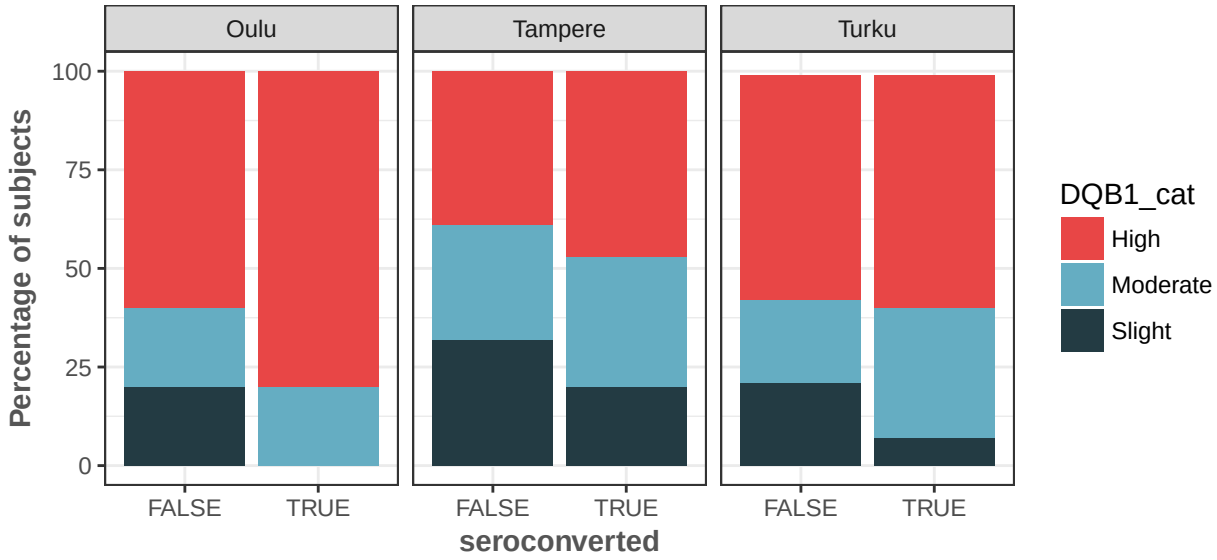
Type 1 diabetes is known to have a strong genetic component with the most risk conferred by HLA genes. Enrollment and selection of subjects to participate in the DIPP cohort was based on HLA predispostion for T1D based on HLA genotype. Initially, the HLA-DQB1 genotype was provided. However to simplfy analysis,

this variable was recategorized based on level of risk as suggested by Kukko **et al.** (2004) and Bakhtadze **et al.** (2006) according to the following:

Risk level category	HLA-DQB1 genotypes
High	02/0302; 0201/0302
Moderate	0302; 0302/0303; 0302/0501; 0302/0603; 0302/0604
Slight	02; 02/0303; 02/0604; 0301/0302; 0301/0303

Overall, there is no difference in HLA-DQB1 genotype risk category between seroconverted subjects and those that do not seroconvert within each site.

	Oulu Controls	Oulu Cases	Tampere Controls	Tampere Cases	Turku Controls	Turku Cases
High (%)	60	80	39	47	57	59
High (N)	6	4	11	7	27	16
Moderate (%)	20	20	29	33	21	33
Moderate (N)	2	1	8	5	10	9
Slight (%)	20	0	32	20	21	7
Slight (N)	2	0	9	3	10	2




```
#different between all cases and controls? --> no quite
fisher.test(table(dipp.sub$seroconverted, dipp.sub$DQB1_cat))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$seroconverted, dipp.sub$DQB1_cat)
## p-value = 0.1278
## alternative hypothesis: two.sided
```

```
#different between sites? --> NO
fisher.test(table(dipp.sub$site, dipp.sub$DQB1_cat))
```

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  table(dipp.sub$site, dipp.sub$DQB1_cat)
## p-value = 0.3582
## alternative hypothesis: two.sided
```

```
#different between cases and controls within site? --> NO
##Oulu
fisher.test(table(dipp.sub[dipp.sub$site == 'Oulu',]$seroconverted, dipp.sub[dipp.sub$site == 'Oulu',]$
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.7602
## alternative hypothesis: two.sided
```

```
##Tampere
fisher.test(table(dipp.sub[dipp.sub$site == 'Tampere',]$seroconverted, dipp.sub[dipp.sub$site == 'Tampe
```

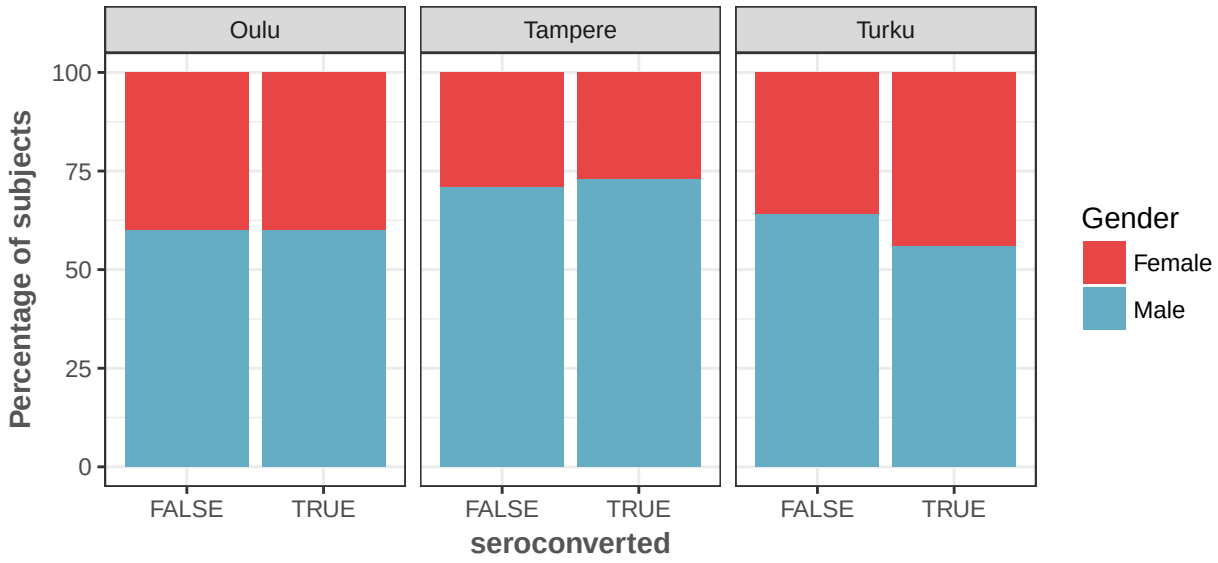
```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.7783
## alternative hypothesis: two.sided
```

```
##Turku
fisher.test(table(dipp.sub[dipp.sub$site == 'Turku',]$seroconverted, dipp.sub[dipp.sub$site == 'Turku',
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.2504
## alternative hypothesis: two.sided
```

Gender

	Oulu Controls	Oulu Cases	Tampere Controls	Tampere Cases	Turku Controls	Turku Cases
Male (%)	60	60	71	73	64	56
Male (N)	6	3	20	11	30	15
Female (%)	40	40	29	27	36	44
Female (N)	4	2	8	4	17	12




```
#different between all cases and controls? --> NO
fisher.test(table(dipp.sub$seroconverted, dipp.sub$Gender))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$seroconverted, dipp.sub$Gender)
## p-value = 0.7053
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5318747 2.6668205
## sample estimates:
## odds ratio
##   1.196905
```

```
#different between sites? --> NO
fisher.test(table(dipp.sub$site, dipp.sub$Gender))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$site, dipp.sub$Gender)
## p-value = 0.4703
## alternative hypothesis: two.sided
```

```
#different between cases and controls within site? --> NO
##Oulu
fisher.test(table(dipp.sub[dipp.sub$site == 'Oulu',]$seroconverted, dipp.sub[dipp.sub$site == 'Oulu',]$
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.05780824 13.78938072
## sample estimates:
```

```
## odds ratio
##          1
```

```
##Tampere
fisher.test(table(dipp.sub[dipp.sub$site == 'Tampere',]$seroconverted, dipp.sub[dipp.sub$site == 'Tampe
```

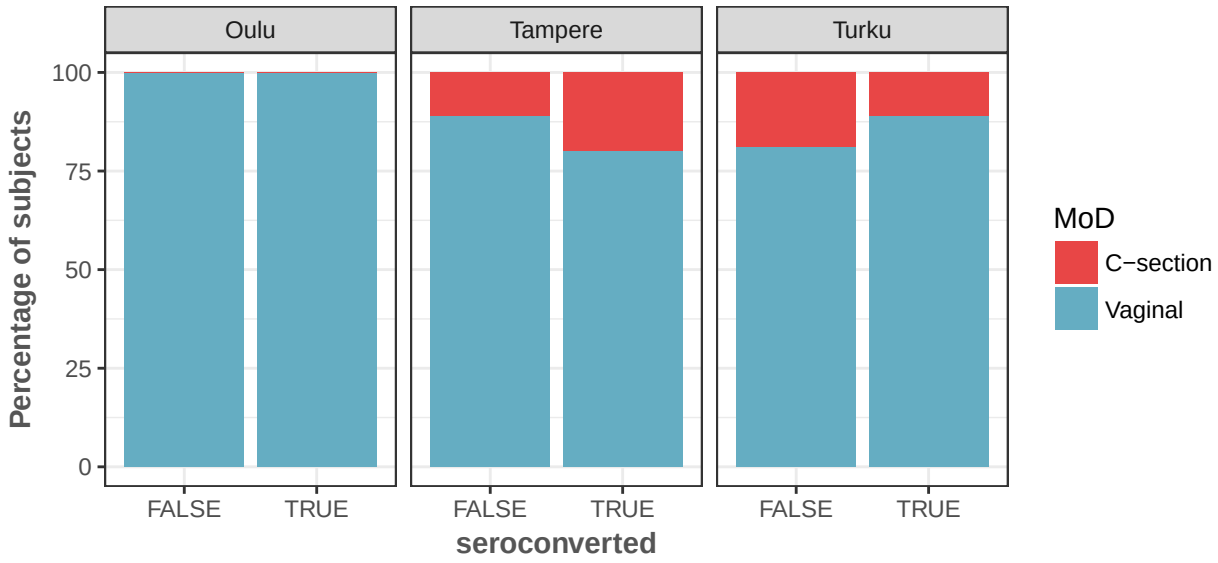
```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.162575 4.427609
## sample estimates:
## odds ratio
##  0.9111252
```

```
##Turku
fisher.test(table(dipp.sub[dipp.sub$site == 'Turku',]$seroconverted, dipp.sub[dipp.sub$site == 'Turku',
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.6214
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4793906 4.1099509
## sample estimates:
## odds ratio
##   1.405073
```

Mode of delivery

	Oulu Controls	Oulu Cases	Tampere Controls	Tampere Cases	Turku Controls	Turku Cases
Vaginal (%)	100	100	89	80	81	89
Vaginal (N)	10	5	25	12	38	24
C-section (%)	0	0	11	20	19	11
C-section (N)	0	0	3	3	9	3




```
#different between all cases and controls? --> NO
fisher.test(table(dipp.sub$seroconverted, dipp.sub$MoD_simp))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$seroconverted, dipp.sub$MoD_simp)
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2545691 2.8010698
## sample estimates:
## odds ratio
##  0.8909943
```

```
#different between sites? --> NO
fisher.test(table(dipp.sub$site, dipp.sub$MoD_simp))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$site, dipp.sub$MoD_simp)
## p-value = 0.3093
## alternative hypothesis: two.sided
```

```
#different between cases and controls within site? --> NO
##Oulu
fisher.test(table(dipp.sub[dipp.sub$site == 'Oulu',]$seroconverted, dipp.sub[dipp.sub$site == 'Oulu',]$M
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0 Inf
## sample estimates:
```

```
## odds ratio
##          0
```

```
##Tampere
fisher.test(table(dipp.sub[dipp.sub$site == 'Tampere',]$seroconverted, dipp.sub[dipp.sub$site == 'Tampe
```

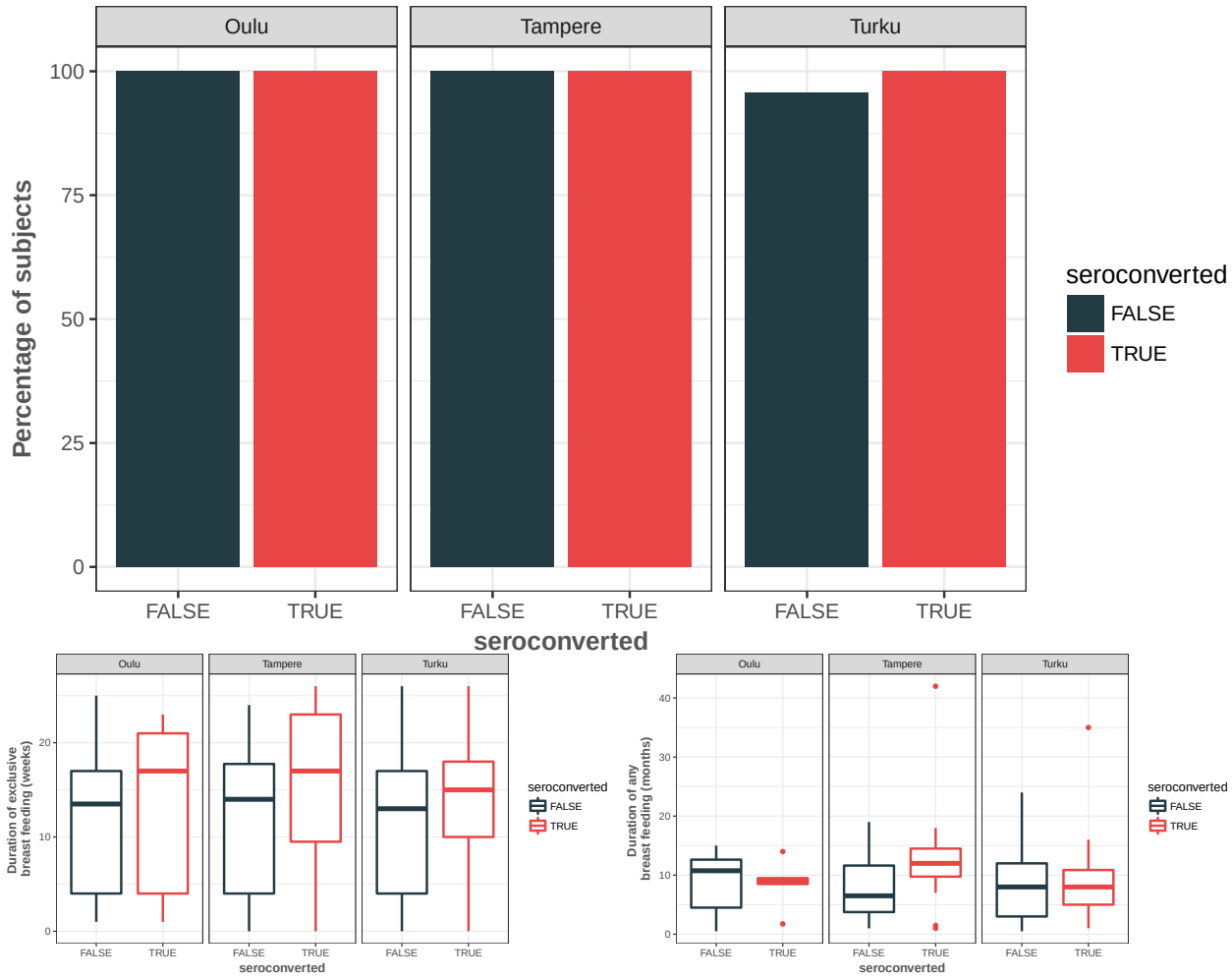
```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.6474
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.2378316 17.6753714
## sample estimates:
## odds ratio
##   2.045612
```

```
##Turku
fisher.test(table(dipp.sub[dipp.sub$site == 'Turku',]$seroconverted, dipp.sub[dipp.sub$site == 'Turku',
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.5171
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.08427106 2.41940189
## sample estimates:
## odds ratio
##  0.5320567
```

Early feeding habits

	Oulu Controls	Oulu Cases	Tampere Controls	Tampere Cases	T
Breast feeding prevalence (%)	100	100	100	100	9
(N)	10	5	28	15	4
Median duration of exclusive breastfeeding (weeks)	14	17	14	17	1
IQR duration of exclusive breastfeeding (weeks)	(4:17)	(4:21)	(4:18)	(10:23)	(
Median duration of any breastfeeding (months)	11	9	6	12	8
IQR duration of any breastfeeding (months)	(4:13)	(8:10)	(4:12)	(10:14)	(




```
#Breast feeding prevalence
#different between all cases and controls? --> NO
fisher.test(table(dipp.sub$seroconverted, dipp.sub$Breast_feeding_any))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$seroconverted, dipp.sub$Breast_feeding_any)
## p-value = 0.5379
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1036302       Inf
## sample estimates:
## odds ratio
##        Inf
```

```
#different between sites? --> NO
fisher.test(table(dipp.sub$site, dipp.sub$Breast_feeding_any))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$site, dipp.sub$Breast_feeding_any)
```

```
## p-value = 0.632
## alternative hypothesis: two.sided
```

```
#different between cases and controls within site? --> NO
##Oulu
fisher.test(table(dipp.sub[dipp.sub$site == 'Oulu',]$seroconverted, dipp.sub[dipp.sub$site == 'Oulu',]$
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0 Inf
## sample estimates:
## odds ratio
##          0
```

```
##Tampere
fisher.test(table(dipp.sub[dipp.sub$site == 'Tampere',]$seroconverted, dipp.sub[dipp.sub$site == 'Tampe
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0 Inf
## sample estimates:
## odds ratio
##          0
```

```
##Turku
fisher.test(table(dipp.sub[dipp.sub$site == 'Turku',]$seroconverted, dipp.sub[dipp.sub$site == 'Turku',
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.5302
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1075345       Inf
## sample estimates:
## odds ratio
##        Inf
```

```
################################################
#Duration of exclusive breast feeding
#Difference between all cases and controls? --> NOT QUITE
kruskal.test(Duration_exclusive_breast_feeding_weeks~seroconverted, data=dipp.sub.BF)
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Duration_exclusive_breast_feeding_weeks by seroconverted
## Kruskal-Wallis chi-squared = 2.7009, df = 1, p-value = 0.1003
```

```
#Difference between sites? --> NO
kruskal.test(Duration_exclusive_breast_feeding_weeks~site, data=dipp.sub.BF)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_exclusive_breast_feeding_weeks by site
## Kruskal-Wallis chi-squared = 0.41211, df = 2, p-value = 0.8138
```

```
#Difference between cases and controls within each site? --> NO
##Oulu
kruskal.test(Duration_exclusive_breast_feeding_weeks~seroconverted, data=dipp.sub.BF[dipp.sub.BF$site ==
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_exclusive_breast_feeding_weeks by seroconverted
## Kruskal-Wallis chi-squared = 0.24526, df = 1, p-value = 0.6204
```

```
##Tampere
kruskal.test(Duration_exclusive_breast_feeding_weeks~seroconverted, data=dipp.sub.BF[dipp.sub.BF$site ==
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_exclusive_breast_feeding_weeks by seroconverted
## Kruskal-Wallis chi-squared = 1.8818, df = 1, p-value = 0.1701
```

```
##Turku
kruskal.test(Duration_exclusive_breast_feeding_weeks~seroconverted, data=dipp.sub.BF[dipp.sub.BF$site ==
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_exclusive_breast_feeding_weeks by seroconverted
## Kruskal-Wallis chi-squared = 1.0118, df = 1, p-value = 0.3145
```

```
##############################################
#Duration of any breast feeding
#Difference between all cases and controls? --> NOT QUITE
kruskal.test(Duration_Breast_Feeding_months~seroconverted, data=dipp.sub.BF)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_Breast_Feeding_months by seroconverted
## Kruskal-Wallis chi-squared = 1.3963, df = 1, p-value = 0.2373
```

```
#Difference between sites? --> NO
kruskal.test(Duration_Breast_Feeding_months~site, data=dipp.sub.BF)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_Breast_Feeding_months by site
## Kruskal-Wallis chi-squared = 0.20663, df = 2, p-value = 0.9018
```

```
#Difference between cases and controls within each site? --> YES IN TAMPERE!!!
##Oulu
kruskal.test(Duration_Breast_Feeding_months~seroconverted, data=dipp.sub.BF[dipp.sub.BF$site == 'Oulu',]
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_Breast_Feeding_months by seroconverted
## Kruskal-Wallis chi-squared = 0.18441, df = 1, p-value = 0.6676
```

```
##Tampere
kruskal.test(Duration_Breast_Feeding_months~seroconverted, data=dipp.sub.BF[dipp.sub.BF$site == 'Tampere
```

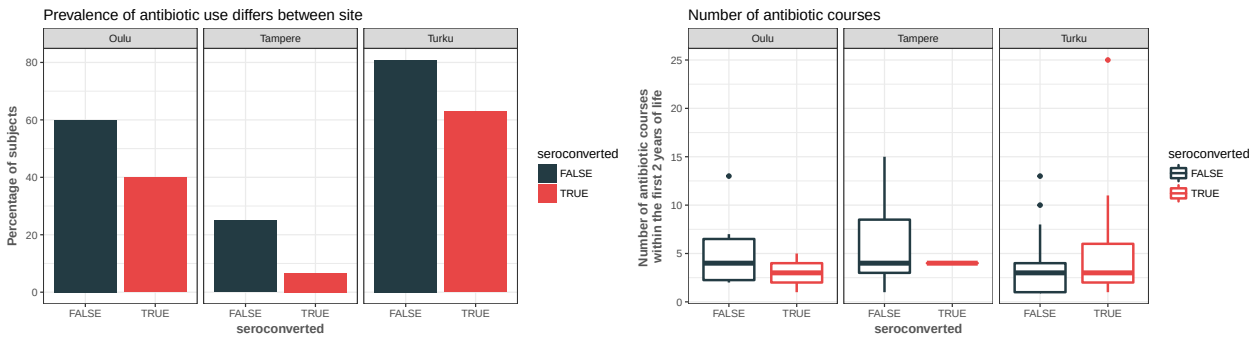
```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_Breast_Feeding_months by seroconverted
## Kruskal-Wallis chi-squared = 5.5392, df = 1, p-value = 0.0186
```

```
##Turku
kruskal.test(Duration_Breast_Feeding_months~seroconverted, data=dipp.sub.BF[dipp.sub.BF$site == 'Turku'
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Duration_Breast_Feeding_months by seroconverted
## Kruskal-Wallis chi-squared = 0.048998, df = 1, p-value = 0.8248
```

Antibiotics in the first 2 years of life

	Oulu Controls	Oulu Cases	Tampere Controls	Tampere Cases	Turku Controls	T
Prevalence of antibiotic use (%)	60	40	25	6.67	80.85	6
(N)	6	2	7	1	38	1
Median number of courses	4	3	4	4	3	3
IQR of number of courses	(2:6)	(2:4)	(3:8)	(4:4)	(1:4)	(2




```
#Abx prevalence
#different between all cases and controls? --> NO
fisher.test(table(dipp.sub$seroconverted, dipp.sub$antibiotics))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$seroconverted, dipp.sub$antibiotics)
```

```
## p-value = 0.06865
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2245418 1.0814417
## sample estimates:
## odds ratio
##   0.496532
```

```
#different between sites? --> YES
fisher.test(table(dipp.sub$site, dipp.sub$antibiotics))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub$site, dipp.sub$antibiotics)
## p-value = 8.526e-09
## alternative hypothesis: two.sided
```

```
#different between cases and controls within site? --> NO
##Oulu
fisher.test(table(dipp.sub[dipp.sub$site == 'Oulu',]$seroconverted, dipp.sub[dipp.sub$site == 'Oulu',]$a
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.6084
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02687938 6.23767632
## sample estimates:
## odds ratio
##  0.4698172
```

```
##Tampere
fisher.test(table(dipp.sub[dipp.sub$site == 'Tampere',]$seroconverted, dipp.sub[dipp.sub$site == 'Tampe
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.2261
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.004448037 2.037263989
## sample estimates:
## odds ratio
##  0.2206875
```

```
##Turku
fisher.test(table(dipp.sub[dipp.sub$site == 'Turku',]$seroconverted, dipp.sub[dipp.sub$site == 'Turku',]
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:
## p-value = 0.1047
```

```
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1210627 1.3430674
## sample estimates:
## odds ratio
##   0.407974
```

```
#######################################
#Number of courses
#Difference between all cases and controls? --> NO
kruskal.test(antibiotic_courses~seroconverted, data=dipp.sub.Abx)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  antibiotic_courses by seroconverted
## Kruskal-Wallis chi-squared = 0.14087, df = 1, p-value = 0.7074
```

```
#Difference between sites? --> NO
kruskal.test(antibiotic_courses~site, data=dipp.sub.Abx)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  antibiotic_courses by site
## Kruskal-Wallis chi-squared = 2.2513, df = 2, p-value = 0.3244
```

```
#Difference between cases and controls within each site? --> NO
##Oulu
kruskal.test(antibiotic_courses~seroconverted, data=dipp.sub.Abx[dipp.sub.Abx$site == 'Oulu',])
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  antibiotic_courses by seroconverted
## Kruskal-Wallis chi-squared = 0.71138, df = 1, p-value = 0.399
```

```
##Tampere
kruskal.test(antibiotic_courses~seroconverted, data=dipp.sub.Abx[dipp.sub.Abx$site == 'Tampere',])
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  antibiotic_courses by seroconverted
## Kruskal-Wallis chi-squared = 0.05, df = 1, p-value = 0.8231
```

```
##Turku
kruskal.test(antibiotic_courses~seroconverted, data=dipp.sub.Abx[dipp.sub.Abx$site == 'Turku',])
```

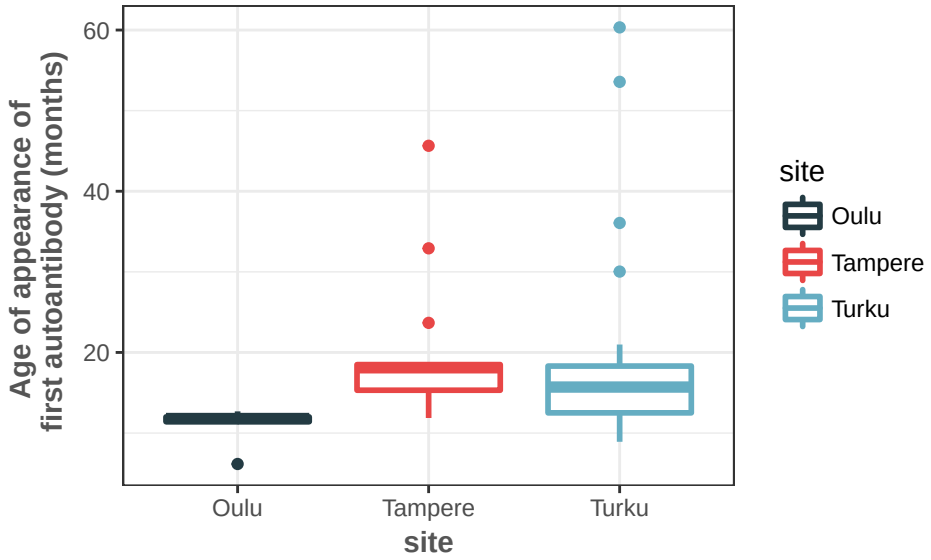
```
##
##  Kruskal-Wallis rank sum test
##
## data:  antibiotic_courses by seroconverted
## Kruskal-Wallis chi-squared = 0.80588, df = 1, p-value = 0.3693
```

Natural History

Islet Autoantibodies

Appearance of first autoantibody

	Oulu Cases	Tampere Cases	Turku Cases
Median age of first autoantibody (months)	12	18	16
IQR of age of first autoantibody (months)	(11:12)	(15:18)	(13:18)

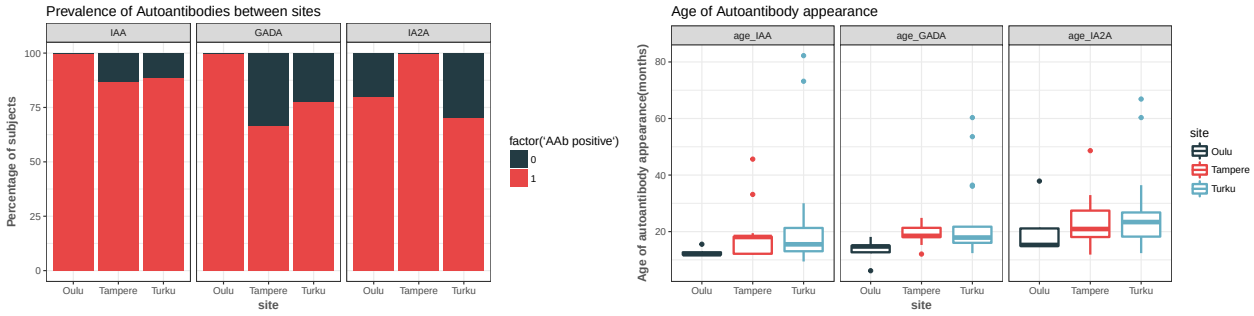



```
#Difference between sites? --> YES!
kruskal.test(age_first_sc~site, data=dipp.sub.cases)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age_first_sc by site
## Kruskal-Wallis chi-squared = 8.6048, df = 2, p-value = 0.01354
```

Prevalence and median age of specific autoantibodies

	Oulu	Tampere	Turku
Prevalence IAA (%)	100	87	89
(N of IAA)	5	13	24
Median IAA age	12	18	16
IQR IAA age	(12:13)	(12:18)	(13:21)
Prevalence GADA (%)	100	67	78
(N of GADA)	5	10	21
Median GADA age	15	18	18
IQR GADA age	(13:15)	(18:21)	(16:22)
Prevalence IA2A (%)	80	100	70
(N of IA2A)	4	15	19
Median IA2A age	15	21	23
IQR IA2A age	(15:21)	(18:27)	(18:27)




```
#Difference between sites?
##IAA prevalence --> NO
fisher.test(table(dipp.sub.cases$site, dipp.sub.cases$aa_IAA))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub.cases$site, dipp.sub.cases$aa_IAA)
## p-value = 1
## alternative hypothesis: two.sided
```

```
##GADA prevalence --> NO
fisher.test(table(dipp.sub.cases$site, dipp.sub.cases$aa_GADA))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub.cases$site, dipp.sub.cases$aa_GADA)
## p-value = 0.3098
## alternative hypothesis: two.sided
```

```
##IA2A prevalence --> YES
fisher.test(table(dipp.sub.cases$site, dipp.sub.cases$aa_IA2A))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub.cases$site, dipp.sub.cases$aa_IA2A)
## p-value = 0.0453
## alternative hypothesis: two.sided
```

```
#Difference between age of T1D AAb appearance
##IAA --> NOT QUITE
kruskal.test(age_IAA~site, dipp.sub.cases[dipp.sub.cases$aa_IAA == 1,])
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age_IAA by site
## Kruskal-Wallis chi-squared = 5.6726, df = 2, p-value = 0.05864
```

```
##IAA --> NOT QUITE
kruskal.test(age_GADA~site, dipp.sub.cases[dipp.sub.cases$aa_GADA == 1,])
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  age_GADA by site
## Kruskal-Wallis chi-squared = 5.7063, df = 2, p-value = 0.05766
```

```
##IA2A --> NOT QUITE
kruskal.test(age_IAA~site, dipp.sub.cases[dipp.sub.cases$aa_IA2A == 1,])
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age_IAA by site
## Kruskal-Wallis chi-squared = 4.3176, df = 2, p-value = 0.1155
```

Type 1 Diabetes

	Oulu	Tampere	Turku
Prevalence T1D (%)	80	60	74
(N of T1D)	4	9	20
Median T1D age	28	67	48
IQR T1D age	(20:39)	(36:97)	(35:75)


```
#Prevalence
#Difference between sites? --> NO
fisher.test(table(dipp.sub.cases$site, dipp.sub.cases$has_T1D))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dipp.sub.cases$site, dipp.sub.cases$has_T1D)
## p-value = 0.642
## alternative hypothesis: two.sided
```

```
#Difference between age of T1D onset --> NOT QUITE
kruskal.test(age_T1D~site, dipp.sub.t1d)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age_T1D by site
## Kruskal-Wallis chi-squared = 4.0418, df = 2, p-value = 0.1325
```

Summary of Confounders

Overall, the main demographic differences were in antibiotic use. Meanwhile, there were differences in autoimmune progression between the 3 cities:

- Antibiotic use within the first 2 years of life weas more prevalent in Turku than in Tampere. However, of the subjects that did receive anitbiotics, there was no differences in the number of coursess received.
- There were no differences in gender, mode of delivery, or prevalence of breast-feeding between cities or between cases and controls within cities.
- In Tampere, cases received breast milk for a longer period of time compared to controls (median of 12 vs. 7 months, respectively).
- There was no difference in HLA-DQB1 between cities or between cases and controls within each city.
- The prevalence on IAA and GADA were similar across cities, but...
- IA-2A was more prevalent in Tampere compared to Oulu and Turku.
- The total number of autoantibodies a case developed were similar between the 3 cities.
- There was no difference in which autoantibody appeared first acorss the 3 cities, however...
- the first autoantibody appeared earlier in Oulu compared to Tampere and Turku.
- There were no differences in the prevalence or age of onset of type 1 diabetes.

General Bacterial Profiles

A basic view of the bacterial profiles within each site, across age was considered. To prevent any subject with larger sampling from overly influencing the profile of the sample population, samples from the same subject within an age bin (width = 2 months, sampled every month) were averaged prior to taking the population average This calculation was performed within each site separately. Profiles were only considered at the phylum level, though this code can be used to perform calculations at any level.

Based on previous studies, it is expected that microbial composition will be transitioning within the first year of life and this will likely be characterized by an increase in Firmicutes and Bacteroidetes, while Proteobacteria and Actinobacteria will be in decline. These phyla are typically the dominant phyla found in human feces.

Figure 2A

```
#melt data
dipp.p.m = psmelt(dipp.p)

#generate age bins - bins all samples within a 2 month window every month
dipp.p.agebins = generate_subsets(dipp.p.m, 'age_at_sampling', seq(91.5,732, by=30.5), 60, 'Phylum', 'Al

#get summary stats within each age bin
##consider cases and controls separately so even weight will be given to each group
dipp.p.median.sc = ddply(dipp.p.agebins, ~site+seroconverted+window+Phylum,
                      function(x) c(med_age = median(x$age_at_sampling),
                                    med_Abundance = median(x$value),
                                    avg_Abundance = mean(x$value),
                                    sd_Abundance = sd(x$value),
                                    var_Abundance = var(x$value)))

#reorder phyla by abundance
##calculate sum of avg_Abundance for each phyla
phyla = ddply(dipp.p.median.sc, ~Phylum, function(x) sum(x$avg_Abundance))
##order by decreasing Phylum (class=data.frame)
```

```
phyla.ord = phyla[order(phyla$V1, decreasing=TRUE),]
dipp.p.median.sc$Phylum = factor(dipp.p.median.sc$Phylum, levels=c(as.character(phyla.ord$Phylum)))

#get median of case and control medians
dipp.p.median = ddply(dipp.p.median.sc, ~site+window+Phylum,
                      function(x) c(med_age = median(x$med_age),
                                    med_Abundance = median(x$med_Abundance),
                                    avg_Abundance = median(x$avg_Abundance),
                                    sd_Abundance = median(x$sd_Abundance),
                                    var_Abundance = median(x$var_Abundance)))

#get phyla that average greater than 1%
phyla.1per = phyla.ord[phyla.ord$V1 >= 1,]$Phylum

dipp.p.median.1per = subset(dipp.p.median, Phylum %in% phyla.1per)
dipp.p.median.1per$Phylum = factor(dipp.p.median.1per$Phylum)
dipp.p.median.1per$Phylum = factor(dipp.p.median.1per$Phylum, levels=rev(levels(dipp.p.median.1per$Phylu

fresh_swap = rev(c('#65ADC2', '#233B43', '#E84646', '#C29365', '#362C21'))


ggplot(dipp.p.median.1per, aes(window, avg_Abundance, fill=Phylum)) +
  geom_area(stat='identity', position='stack') +
  facet_wrap(~site) +
  theme_classic() +
  xlab('Age (months)') +
  ylab('Mean relative abundnace (proportion)') +
  scale_x_continuous(breaks=unique(dipp.p.median.1per$window)[c(seq(1,22,3))],
                     labels=seq(3,24,3),
                     expand=c(0,0.25)) +
  scale_y_continuous(breaks=seq(0,1,0.2), labels=seq(0,1, 0.2), expand=c(0,0)) +
  coord_cartesian(ylim = c(0,1)) +
  theme(panel.grid=element_blank(),
        panel.border=element_blank()
        #legend.position='none'
        ) +
  ggtitle('Phylum-Level Taxonomy') +
  scale_fill_manual(values=fresh_swap, name='Phyla') +
  scale_color_manual(values=fresh_swap)
```

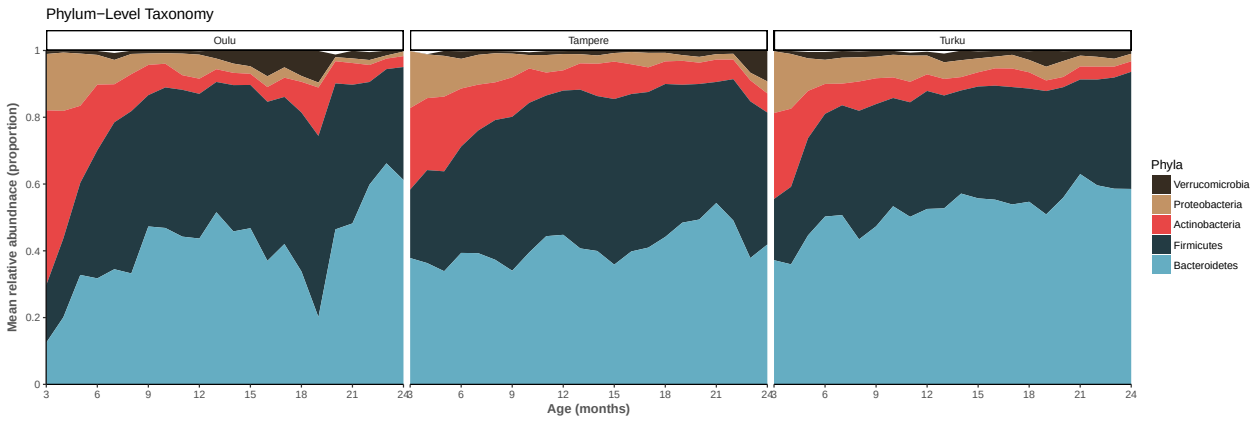


Figure 2A. ***Phyla profile by site*** On average, samples across all sites are dominated by Firmicutes and Bacteroidetes by 1 year of life, as the relative abundances of these phyla increased with age. Also, the relative abundances of Proteobacteria and Actinobacteria decreased with age.

Tri-monthly summary statistics by site

4 major phyla

Analyses were focued on taxa assigned to the 4 most dominant phyla: Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria. On average, these phyla accounted for the following cumulative relative abundance across all samples:

Site	Cumulative relative abundance of 4 dominant phyla (%)
Oulu	97.19
Tampere	98.23
Turku	97.89

Summary statistics for 4 major phyla by site

site	Bacteroidetes	Firmicutes	Actinobacteria	Proteobacteria
Oulu	43.96	42.50	6.57	3.16
Tampere	39.88	42.23	8.40	3.85
Turku	53.04	33.39	5.08	5.03

Summary statistics for 4 major phyla at 3-5 and 12 months across all sites

Phylum	site	3-5 months	12 months
Bacteroidetes	Oulu	0.2178648	0.4424937
Bacteroidetes	Tampere	0.3601895	0.4438849
Bacteroidetes	Turku	0.3926088	0.5254225
Firmicutes	Oulu	0.2279610	0.4332768
Firmicutes	Tampere	0.2608443	0.4319792
Firmicutes	Turku	0.2355455	0.3429745
Actinobacteria	Oulu	0.3788775	0.0454374
Actinobacteria	Tampere	0.2278716	0.0686698
Actinobacteria	Turku	0.2109596	0.0613660
Proteobacteria	Oulu	0.1666179	0.0647666
Proteobacteria	Tampere	0.1411320	0.0487834
Proteobacteria	Turku	0.1486890	0.0678331

Phyla associations with age

It is well documented that taxonomic profiles transition during the first year of life. In this section, the relationship of age and the 4 major phyla are formally tested using generalized linear models.

The relative abundances of Bacteroidetes and Firmicutes increased with age (latter not quite significant, p=0.15), while the relative abundances of Proteobacteria and Actinobacteria decreased. This is also indicated in Figure 2A and the summary statistics for the 3-5 and 12 month time points presented above.

```
#prep long-formatted table
dipp.p.m = psmelt(dipp.p)
#select abundant phyla
dipp.p.m4 = subset(dipp.p.m, Phylum %in% maj.phy)
```

```
#add less variable age
dipp.p.m4$mo_at_sampling = round(dipp.p.m4$age_at_sampling/30.5, 0)
#run GEE: abundance ~ Age^2 + site + gender, strata = subject
gee.age.result = ddply(dipp.p.m4, ~Phylum, function(x) test_age_gee(x, 'Abundance'))
gee.age.result2 = gee.age.result[gee.age.result$effect != '(Intercept)',]
row.names(gee.age.result2) = NULL
kable(gee.age.result2)
```

Phylum	Estimate	Std.err	Wald	p_value	p.adjust	effect
Actinobacteria	-0.0002204	2.10e-05	109.748442	0.000000	0.000000	I(mo_at_sampling^2)
Bacteroidetes	0.0003136	4.84e-05	41.984160	0.000000	0.000000	I(mo_at_sampling^2)
Firmicutes	0.0000566	3.94e-05	2.064767	0.150738	0.150738	I(mo_at_sampling^2)
Proteobacteria	-0.0001695	1.41e-05	143.932424	0.000000	0.000000	I(mo_at_sampling^2)

Diversity Analysis

In initial exploratory analyses, sample diversity was considered. In this section, 3 aspects of diversity were assessed:

- sequencing depth using rarefaction curves
- alpha diversity (species richness and evenness)
- beta diversity

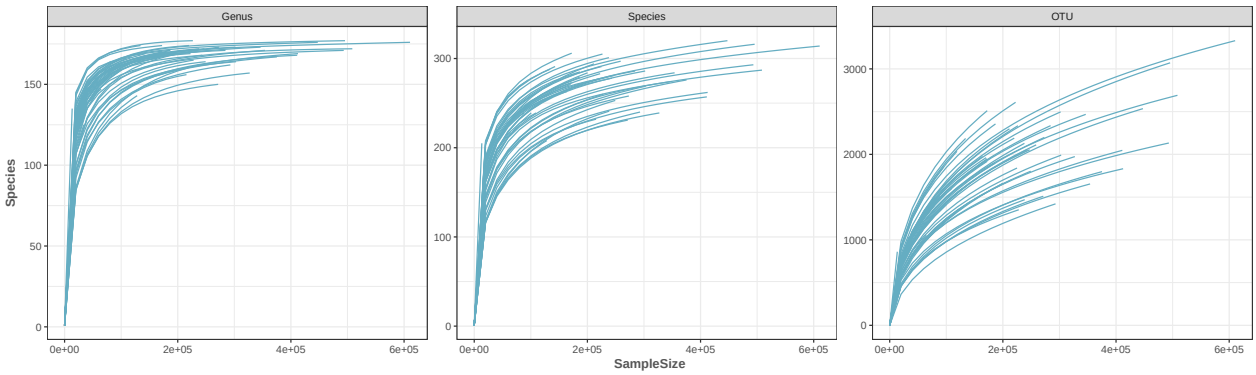
Rarefaction Curve

Rarefaction curves were examined in order to determine whether sufficient coverage for each sample was obtained in order to obtain adequate representation of taxa. That is, taxa that dominate in a sample are more likely to be sequenced than rare taxa. Therefore, the lower sequence coverage of a sample you have, the more likely you are to miss low abundant or rare taxa. Ideally, with increased coverage, the number of unique OTUs or species identified will reach an asymptote, and one could say they have sufficient coverage to detect the majority of unique taxa expected to be present in a sample. With stool microbiome studies, it has been suggested that the minimum amount of coverage should be ~10,000 reads per sample. Of course, this number is debateable, and it is best to assess the validity of this within a given experiment. Below, 50 samples are randomly selected and a rarefaction curve generated.

```
#randomly select 50 sample ids (excluding technical replicates)
#not worried about documenting rng.seed here
x = sample(sample_data(dipp.nTR)$sample_id, 50)

#extract these samples
dipp.rarcurv = prune_samples(sample_data(dipp.nTR)$sample_id %in% x, dipp.nTR)

tax.levels = c("OTU", "Genus", "Species")

dipp.rarcurv.df = data.frame()

for (tax in tax.levels) {
  if (tax == 'OTU') {
    phy = dipp.rarcurv
```

```
  } else {
    phy = tax_glom(dipp.rarcurv, tax)
  }

  #function defined in D3C_functions_5.0.R
  ##extracts rarefaction values from rarecurve()
  rar = get_rarefaction(data.frame(otu_table(phy)), 20000)
  #add group variable indicating tax.level
  rar$group = factor(c(tax))
  #combine rarefied output from each tax.levels
  dipp.rarcurv.df = rbind(dipp.rarcurv.df, rar)
}
```

```
dipp.rarcurv.df$group = factor(dipp.rarcurv.df$group, levels=c('Genus', 'Species', 'OTU'))
ggplot(dipp.rarcurv.df, aes(SampleSize, Species, group=ID)) +
  geom_line() +
  facet_wrap(~group, scales='free') +
  theme_bw()
```



CONCLUSIONS: Sequencing depth appears sufficient for genus discovery. However, at the species and OTU levels, a saturation point was not clearly achieved.

Alpha Diversity

Alpha diversity metrics were calculated using raw read counts for the OTUs. This does introduce the bias of higher alpha diversity for samples that had a greater sequencing depth. Therefore, alpha diversity was also calculated on rarefied counts. By basing the alpha diversity measure on a subselection of reads, samples with high read counts will have an under-estimated alpha diversity whereas the alpha metric would not be expected to deviate much in samples with lower total read counts.

Regardless of the the format of input count values, alpha diversity increases with age and is generally not different between cases and controls. Although, there is some evidence of increased alpha diversity in controls in Tampere. However, this only occurred at select time points and was not observed in the other 2 cities.

Alpha metrics were calculated for each site separately using `estimate_richness` functions from the *phyloseq*`* package, and plots were generated uing`ggplot2“‘. Again samples that served as technical replicates were removed. The following alpha metrics were considered:

- **Chao1:** a measure of species richness and estimates the number of species present in a community. With the understanding that rare species are less likely to be detected, Chao1 is typically greater than the observed richness within a sample.
- **Shannon:** a measure of species eveness, combining richness and abundance into a single value. The actual value, is somewhat arbitrary but is relative to other samples. Samples that are dominated by one

or a few taxa will have a lower Shannon Diversity Index (eveness) compared to those where abundance is distributed more equally among many taxa. However, eveness is not indicative of composition. For example, you may have 2 samples with a similar eveness dominated by 2 species, based on beta diversity metric alone, you would not be able to distinguish these samples, but one sample could be dominated by completely different species than the other.

- **Simpson's Index of Diversity (1-D):** a measure that accounts for the number of species present as well as the abundances of species. The value ranges from 0 to 1 and represents the probability that two reads randomly selected from a given samples will belong to different species. Thus, the greater the value, the greater the diversity.

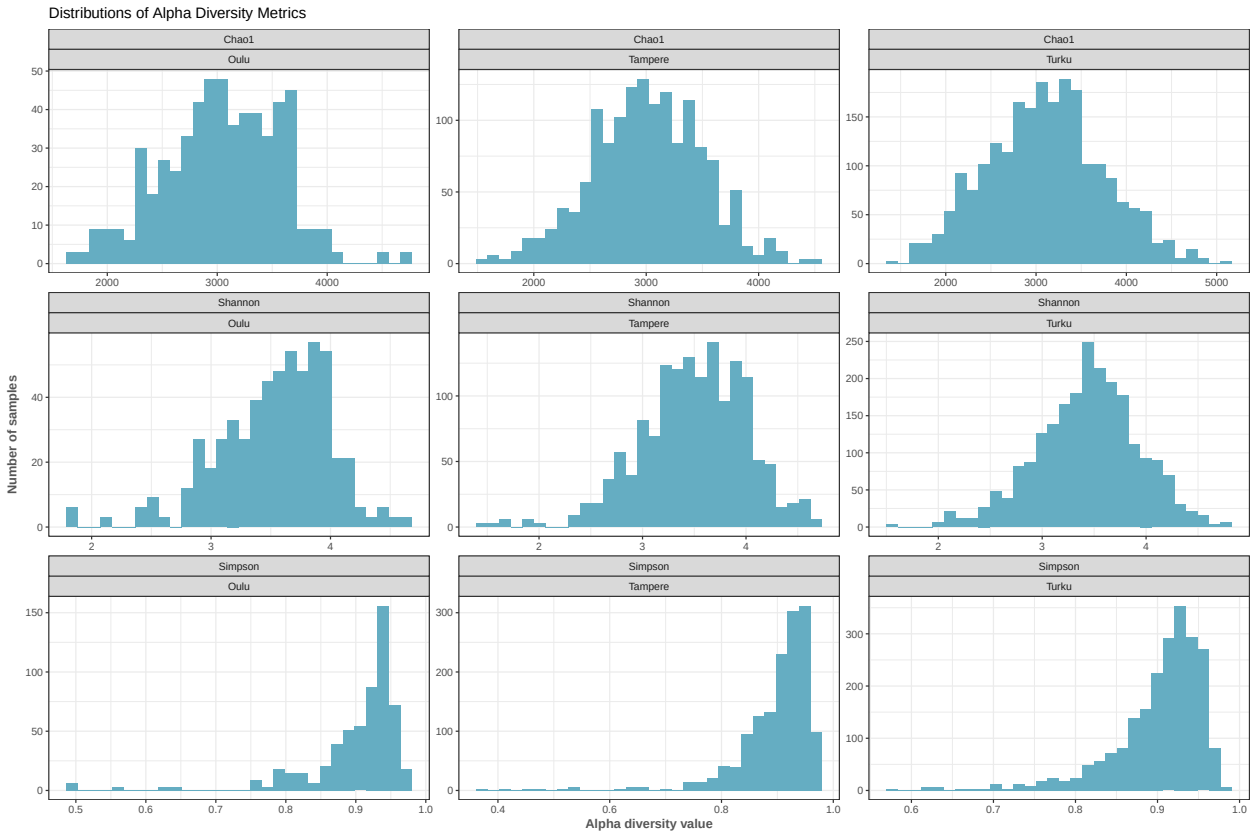
```
#define arguments
alphas = c('Chao1', 'Shannon', 'Simpson')
sites = c('Oulu', 'Tampere', 'Turku')

#calculate alpha metrics
dipp.alpha = data.frame()
for (site in sites) {
  #subset by site
  phy = subset_samples(dipp.nTR, site==site)
  #calculate alpha metrics
  phy.alpha = estimate_richness(phy, measures = alphas)
  #add sample_data
  phy.div = data.frame(phy.alpha, sample_data(phy))
  #combine results
  dipp.alpha = rbind(dipp.alpha, phy.div)
}

#get long-format of data.frame for plotting
dipp.alpha.m = melt(dipp.alpha, measure.vars = c('Chao1', 'Shannon', 'Simpson'))
```

Differences in alpha diversity between seroconversion status and confounders

Normality appears acceptable for Chao1 and Shannon measures, so no transformation will be applied to these values prior to performing statistical test.

Two different statistical methods were applied to identify differences in alpha diversity: a time independent approach simply using the Mann-Whitney test for differences of means in non-parametric data, and a regression approach fitting a generalized linear model. In the former, samples were binned monthly and the median of samples from the same subject within the same age bin were obtained in order to account for repeated measures.

```
#DISTRIBUTION
ggplot(dipp.alpha.m, aes(value)) +
  geom_histogram() +
  facet_wrap(~variable+site, scales='free') +
  theme_bw() +
  xlab('Alpha diversity value') +
  ylab('Number of samples') +
  ggtitle('Distributions of Alpha Diversity Metrics')
```



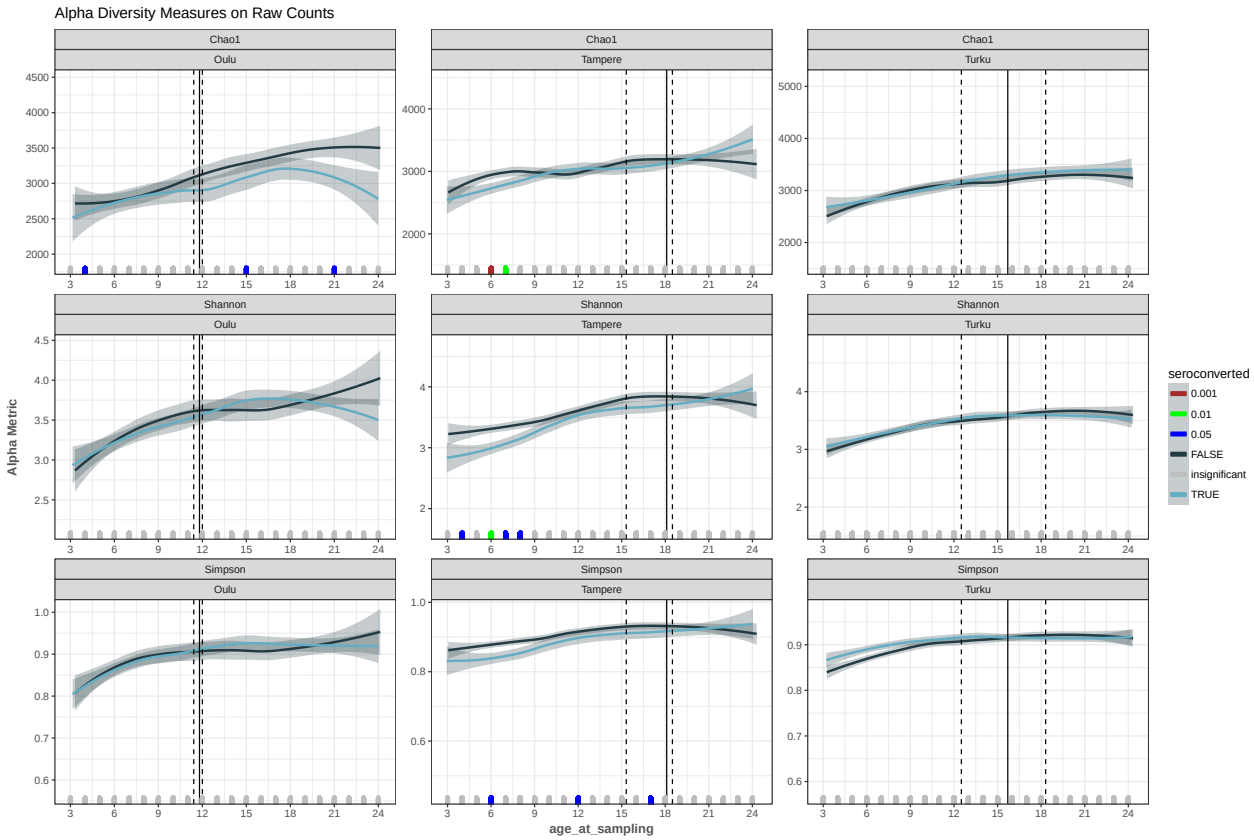

```
#TIME-SPECIFIC ANLAYSIS (MANN-WHITNEY)
##generate_subsets() defined in D3C_functions_5.0.R
  ##returns a data.frame where samples have been assigned to monthy age bins
  ##and repeated measures within an age bin accounted for
dipp.alpha.agebins = generate_subsets(dipp.alpha.m, 'age_at_sampling', seq(91.5,732, by=30.5), 60, 'vari
##perform Mann-Whitney test within each age bin; function defined in D3C_functions_5.0.R
dipp.alpha.mw = test_mw(dipp.alpha.agebins, 'value', 'window', 'variable', 'seroconverted', 'site')

#add MW results to data.frame
dipp.alpha.m2 = merge(dipp.alpha.agebins, dipp.alpha.mw, by=intersect(names(dipp.alpha.mw), names(dipp.a

#GLM
dipp.alpha.m$mask_id = factor(dipp.alpha.m$mask_id)
alpha.glm = list()
for (site in sites) {
  for (var in alphas) {
    #subset by site
    alpha.df = dipp.alpha.m[dipp.alpha.m$site == site,]
    #subset by alpha metric
    alpha.df.var = alpha.df[alpha.df$variable == var,]
    #scale variables
    alpha.df.var$scaled.value = scale(alpha.df.var$value)
    alpha.df.var$scaled.age = scale(alpha.df.var$age_at_sampling)
    x = glm(value~I(age_at_sampling^2)*seroconverted, data = alpha.df.var)
  alpha.glm[[paste(site, var, sep='_')]] = x
    }
}
```

```
#extract GLM summary results
glm.table = list()
#glm.table = data.frame()
for (i in names(alpha.glm)) {
  x = summary(alpha.glm[[i]])$coefficients[,4]
  #xx = x$coefficients
  #cbind(glm.table, data.frame(x))
  glm.table[[i]] = x
  #glm.table[,i] = data.frame(x)
}
```

GLM table of p-values

	Oulu_Chao1	Oulu_Shannon	Oulu_Simpson	Tampere_Chao1	Tampe
(Intercept)	0.0000000	0.0000000	0.0000000	0.0000000	
I(age_at_sampling^2)	0.0000000	0.0000000	0.0000000	0.0000000	
seroconvertedTRUE	0.9349006	0.0917950	0.2237766	0.0000056	
I(age_at_sampling^2):seroconvertedTRUE	0.0007374	0.0048169	0.1940119	0.0015244	

```
#plots
ggplot(dipp.alpha.m2, aes(age_at_sampling, value, color=seroconverted)) +
  geom_smooth(alpha=0.25, method='loess') +
  facet_wrap(~variable+site, scales='free') +
  scale_x_continuous(breaks=seq(0,747,91.5), labels=seq(0,24,3)) +
  ylab('Alpha Metric') +
  theme_bw() +
  geom_vline(aes(xintercept=SeroSiteMed)) +
  geom_vline(aes(xintercept=SeroSite1stIQR), lty='dashed') +
  geom_vline(aes(xintercept=SeroSite3rdIQR), lty='dashed') +
  ggtitle('Alpha Diversity Measures on Raw Counts') +
  scale_color_manual(values = c("brown", "green", "blue", "#233B43", "grey", "#65ADC2")) +
  #geom_rug(aes(color=significance), sides ='b', size=1)
  geom_rug(aes(x=window, color=significance), sides ='b', size=2)
```



CONCLUSIONS:

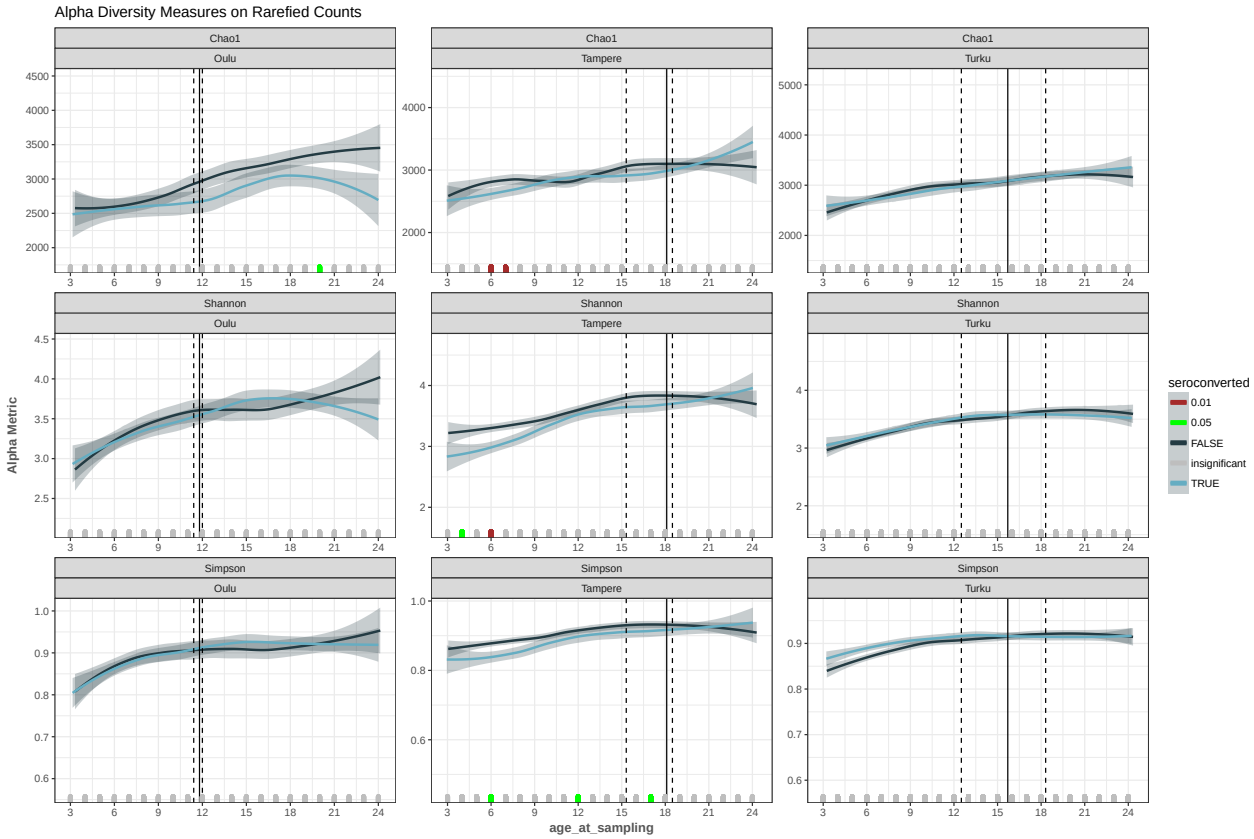
- Age had the most effect on alpha diversity indicated by both statistical methods applied and regardless of site or alpha diversity measure investigated (GLM, p-value < 0.05 for all metrics).
- Time-specific anaylsis (Mann-Whitney test with FDR correction) indicated that there is little to no difference between alpha diversity and case-control status regardless of the metric considered, except for possibly in Tampere.
- There may be a small difference in alpha diversity between cases and controls in Tampere. However, this may be confounded by a difference in breast feeding as described in the *Comparison of Potential Confounders between Cities* section above.

Effect of rarefying

There is debate about the 'best' way to normalize read counts. As previously mentioned, alpha diversity metrics are subject to total read count biases. Therefore, to consider an alternative, read counts were rarefied to 20,000 reads per sample and then alpha diversity metrics were calculated. Again, technical replicate samples were omitted in order to remain consistent with the non-rarefied alpha diversity analysis presented above, although their inclusion should not affect the results.

```
#rarefy counts
dipp.nTR.rar20K = rarefy_even_depth(dipp.nTR, rngseed = TRUE, replace = FALSE, sample.size = 20000)

#calculate alpha metrics
dipp.alpha.rar = data.frame()
for (site in sites) {
  #subset by site
  phy = subset_samples(dipp.nTR.rar20K, site==site)
  #calculate alpha metrics
```

```
  phy.alpha = estimate_richness(phy, measures = alphas)
  #add sample_data
  phy.div = data.frame(phy.alpha, sample_data(phy))
  #combine results
  dipp.alpha.rar = rbind(dipp.alpha, phy.div)
}

dipp.alpha.rar.m = melt(dipp.alpha.rar, measure.vars = c('Chao1', 'Shannon', 'Simpson'))
```



CONCLUSIONS:

The same conclusions regarding alpha diversity were drawn when reads were rarefied and thus has little effect on on results. The code used in this portion is hidden in the final document for aesthetic reasons but ressembles that performed on non-rarefied data and can be found in the provided Rmarkdown file.

Beta Diversity

Beta diversity compares the composition of different microbial communities. To do this, first a pair-wise distance or dissimilarity matrix for all samples is calculated. Ordination and clustering methods are then used to visually compare community composition for exploratory purposes. Differences in microbial composition can be formally tested using permutational/non-parametric MANOVA implemented with the `adonis` function. The following distance/dissimilarity methods were applied:

- **Bray-Curtis:** quantifies the compositional dissimilarity between 2 samples (considers presence/absence & abundance)
- **Weighted UniFrac:** a quantitative (accounts for abundance) assessment of microbial composition which accounts for phylogenetic relatedness of community members

- **Unweighted UniFrac:** a qualitative (presence or absence) assessment of microbial composition which accounts for phylogenetic relatedness of community members

The latter two methods require a phylogenetic tree...

```
#load phylogenetic tree needed for unifrac methods
gg_tree = read_tree_greengenes('~/Desktop/D3C/D3C_2.0/1_inputs/input_files/gg_13_8.97_otus.tree')
dipp.TR.tree = merge_phyloseq(dipp.TR, gg_tree)

#even sampling depth for each sample
dipp.TR.tree.depth = transform_sample_counts(dipp.TR.tree, function(x) 1e+06 * x/sum(x))

#add tree
dipp.g.tree = merge_phyloseq(dipp.g, gg_tree)
#even sampling depth for each sample
dipp.g.100K = transform_sample_counts(dipp.g.tree, function(x) 1e+06*x)

#split phyloseq object by site
turku.depth = subset_samples(dipp.g.100K, site == 'Turku') #808 samples
tampere.depth = subset_samples(dipp.g.100K, site == 'Tampere') #493 samples
oulu.depth = subset_samples(dipp.g.100K, site == 'Oulu') #193 samples
```

The following code was used to generate the dissimilarity/distance matrices. However, generating these matrices takes quite a bit of time, so the resulting matrices are provided in `AdditionalFile4/distance_matrices` for convenience.

```
#apply 3 methods --> used for visualization
dipp.WUF = ordinate(dipp.g.100K, "PCoA", "unifrac", weighted=TRUE)
dipp.UUF = ordinate(dipp.g.100K, "PCoA", "unifrac", weighted=FALSE)
dipp.NB = ordinate(dipp.g.100K, "NMDS", "bray")

#get distance matrices (weighted unifrac only) --> for PERMANOVA and dispersion sections
WUF.dist.all = phyloseq::distance(dipp.g.100K, method='wunifrac', type='samples')
WUF.dist.turku = phyloseq::distance(turku.depth, method='wunifrac', type='samples')
WUF.dist.tampere = phyloseq::distance(tampere.depth, method='wunifrac', type='samples')
WUF.dist.oulu = phyloseq::distance(oulu.depth, method='wunifrac', type='samples')
```

PERMANOVA Results

In order to test differences in global microbial composition, PERMANOVA (implemented with the `adonis` function) was used to test the effects of relevant variables on the microbiota using weighted Unifrac distances.

```
#extract meta data
meta = data.frame(sample_data(dipp.g.100K))

#change variables to a factors
meta$mask_id = factor(meta$mask_id)
meta$age_month = factor(meta$age_month)
meta$Gender = factor(meta$Gender, labels=c('Female', 'Male'))
meta$MoD_simp = ifelse(meta$MoD_simp > 1, c('Vaginal'), c('C-section'))
meta$antibiotics = as.logical(meta$antibiotics)
```

weighted unifrac ~ site, strata = age

Initially, we tested if microbial variation differed by site, while accounting for age (specified as the `strata` argument). The latter would correct for differing age distribution of samples between sites. Site had

a significant effect (p-value=0.001) on microbial variation, accounting for ~2.1% of microbial variation. Therefore, the site variable was used as strata for subsequent PERMANOVA tests.

```
adon.site = adonis(WUF.dist.all~site, data = meta, strata=meta$age_month)
adon.site
```

```
##
## Call:
## adonis(formula = WUF.dist.all ~ site, data = meta, strata = meta$age_month)
##
## Blocks:  strata
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##             Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## site         2     4.092 2.04594  15.691 0.02061  0.001 ***
## Residuals 1491   194.416 0.13039         0.97939
## Total     1493   198.508                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

weighted unifrac ~ age*mask_id, strata = site

Previous literature of human microbiome studies indicate that subject and age contribute a large amount to microbiota variation. This was also the observed in the this cohort.

```
adon.IDage = adonis(WUF.dist.all~mask_id*age_month, data = meta, strata=meta$site)
adon.IDage
```

```
##
## Call:
## adonis(formula = WUF.dist.all ~ mask_id * age_month, data = meta,      strata = meta$site)
##
## Blocks:  strata
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                     Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## mask_id            131    70.798 0.54044 13.5441 0.35665  0.001 ***
## age_month           21    13.272 0.63200 15.8385 0.06686  0.001 ***
## mask_id:age_month 1266   111.445 0.08803  2.2061 0.56141  0.001 ***
## Residuals           75     2.993 0.03990         0.01508
## Total             1493   198.508                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

weighted unifrac ~ site + seroconverted + Gender + breast-feeding + antibiotics, strata = age

Additional variables of interest were tested in order to determine if they contributed to microbiota variation. Variables tested included: seroconversion status, mode of delivery, gender, whether or not a patient was receiving breast milk at the time of sampling, and whether or not a subject received antibiotics within the first 2 years of life. All variables tested had a small but significant effect on microbiota variation.

```
adon.vars = adonis(WUF.dist.all~site + seroconverted + MoD_simp + Gender + receiving_breast_milk + antil
adon.vars
```

```
##
## Call:
## adonis(formula = WUF.dist.all ~ site + seroconverted + MoD_simp +      Gender + receiving_breast_mill
##
## Blocks:  strata
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## site                      2     4.092  2.0459  16.472 0.02061  0.001 ***
## seroconverted             1     1.716  1.7158  13.814 0.00864  0.001 ***
## MoD_simp                  1     1.949  1.9487  15.690 0.00982  0.001 ***
## Gender                    1     1.187  1.1872   9.558 0.00598  0.001 ***
## receiving_breast_milk     1     4.713  4.7128  37.944 0.02374  0.003 **
## antibiotics               1     0.283  0.2826   2.275 0.00142  0.100 .
## Residuals              1486   184.569  0.1242         0.92978
## Total                  1493   198.508                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PERMANOVA results summary

Subject and age had the largest effect on microbial variation. Site also had a significant, although small effect on microbial variation, which justifies stratifying analyses by site. Other variables (mode of delivery, breast-feeding, antibiotic use, gender, and seroconversion status) were also tested for their effects on microbial variation. All, except antibiotic use, had a small although significant effect on microbial variation. This warranted a taxa-independent approach for identifying taxa associated with seroconversion.

	R2	p-value
mask_id	0.3567	0.001
age_month	0.0669	0.001
mask_id:age_month	0.5614	0.001
site	0.0206	0.001
seroconverted	0.0086	0.001
MoD_simp	0.0098	0.001
Gender	0.0060	0.001
receiving_breast_milk	0.0237	0.003
antibiotics	0.0014	0.100

Visualization

Figure 2C

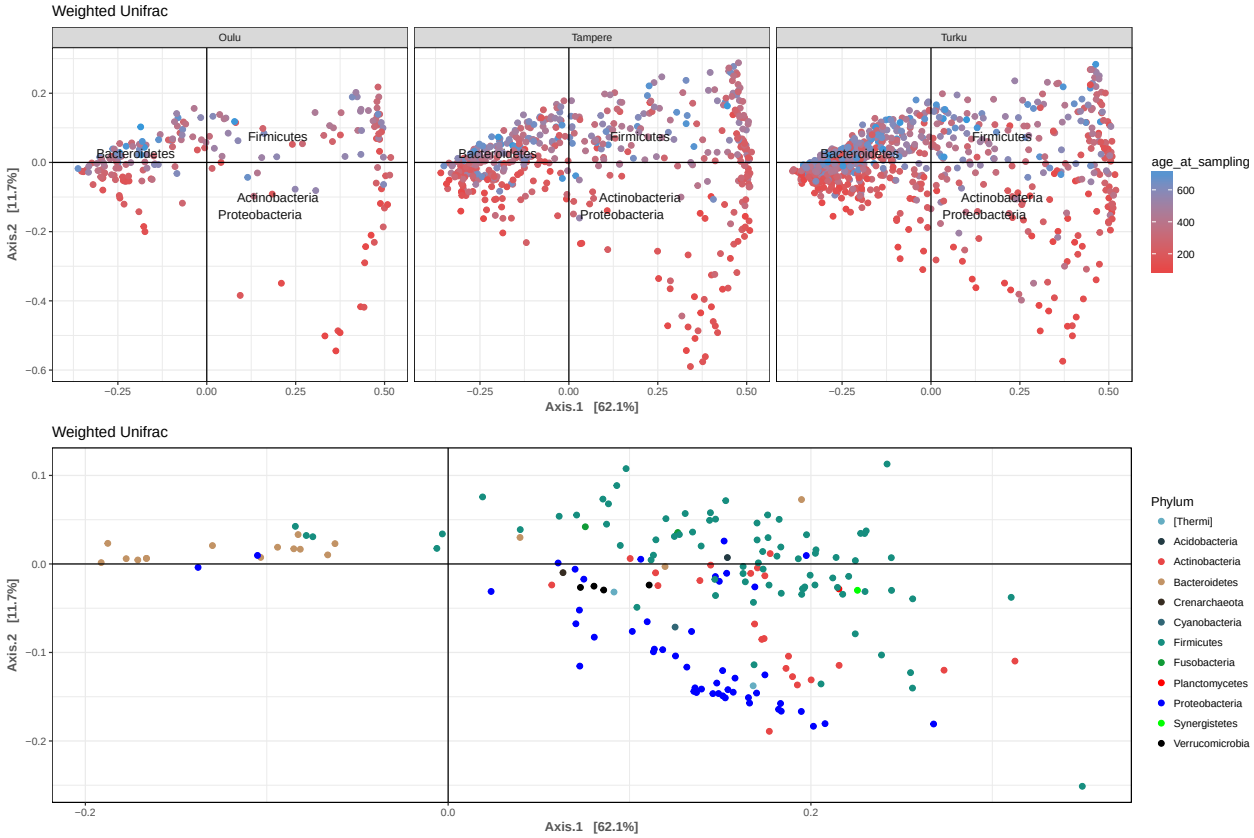
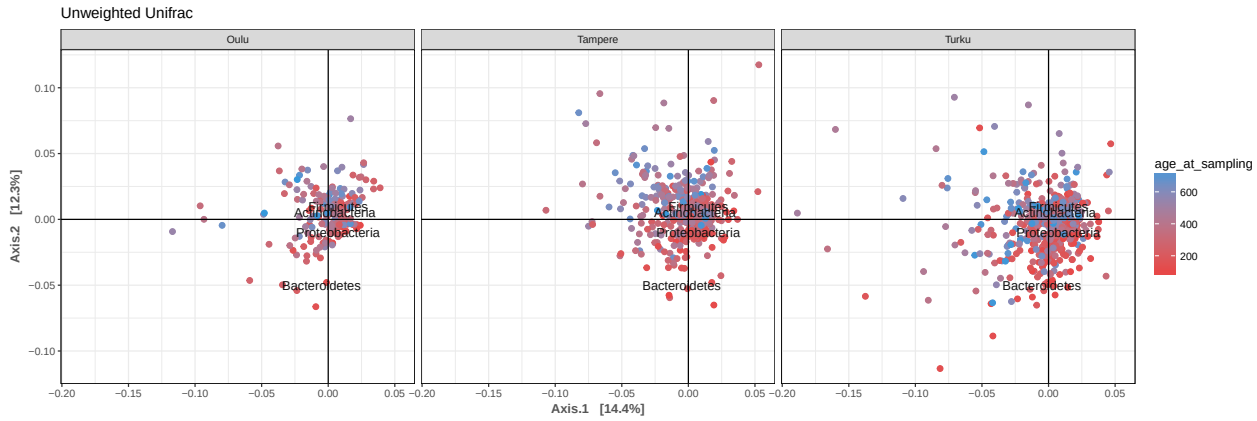
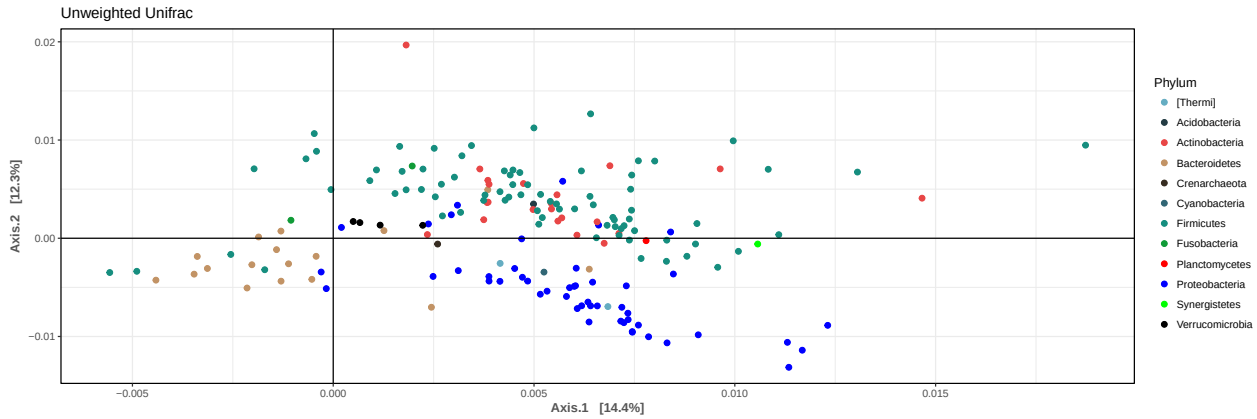
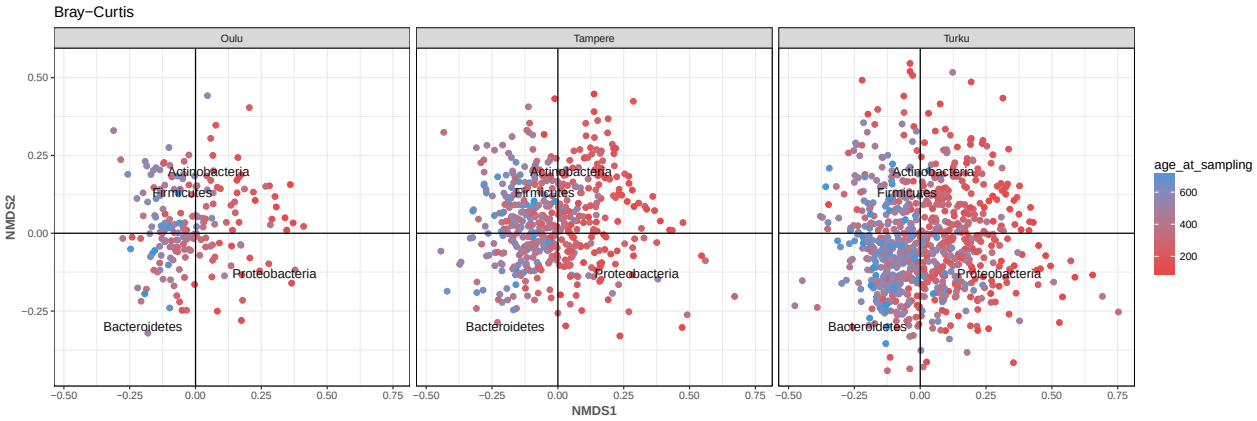


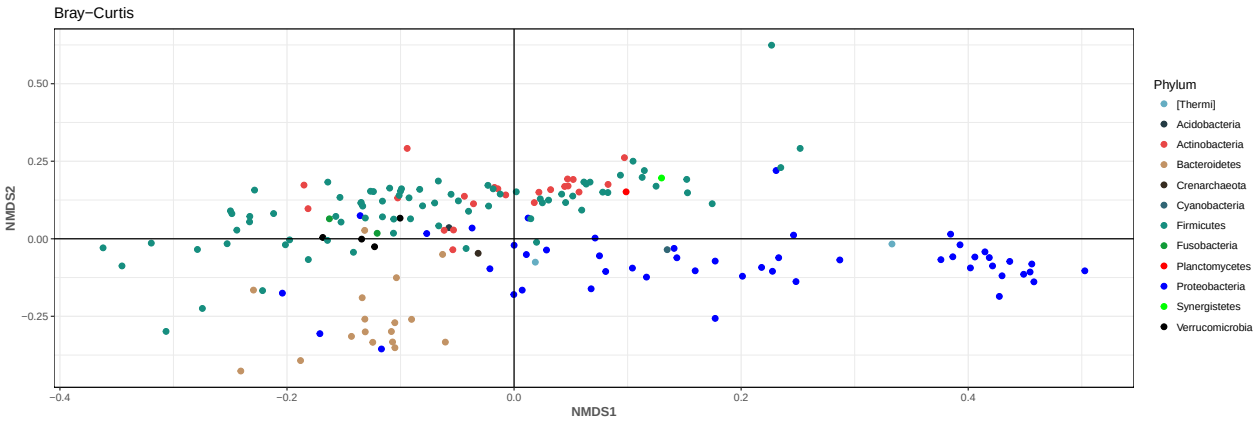
Figure 2C Principle coordinate analysis of weighted-Unifrac distances indicated that subject and age were the dominant sources of variance in global microbial composition.

Note: plots appear in a different orientation from those presented in the manuscript but relationships/distances are the same.


Unweighted Unifrac
Oulu
Tampere
Turku
Axis.2 [12.3%]
Axis.1 [14.4%]
Actinobacteria
Firmicutes
Proteobacteria
Bacteroidetes
age_at_sampling
600
400
200


Unweighted Unifrac
Axis.2 [12.3%]
Axis.1 [14.4%]
Phylum
[Thermi]
Acidobacteria
Actinobacteria
Bacteroidetes
Crenarchaeota
Cyanobacteria
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Synergistetes
Verrucomicrobia

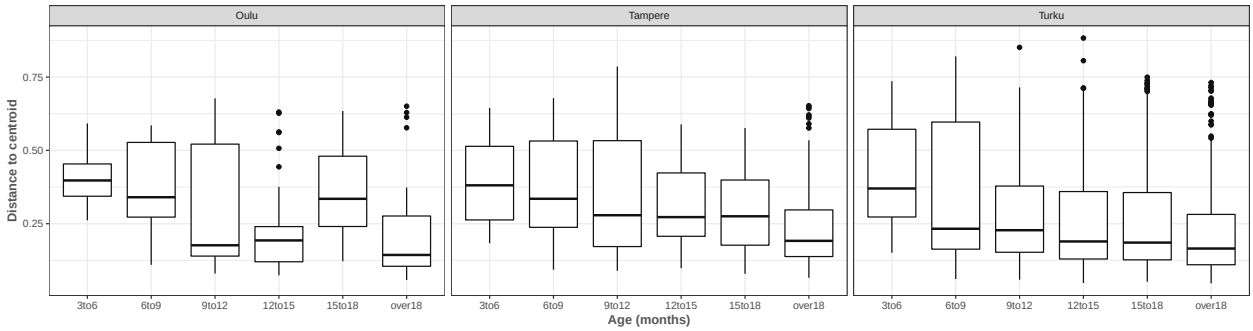

Bray−Curtis
Oulu
Tampere
Turku
NMDS2
NMDS1
Actinobacteria
Firmicutes
Proteobacteria
Bacteroidetes
age_at_sampling
600
400
200


Bray−Curtis
NMDS2
NMDS1
Phylum
[Thermi]
Acidobacteria
Actinobacteria
Bacteroidetes
Crenarchaeota
Cyanobacteria
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Synergistetes
Verrucomicrobia

Dispersion

To determine if samples within more discrete ages were more distant to samples in other age groups, pair-wise distances were (only considered weighted Unifrac distance here) calculated and the dispersion between age groups was considered. In early age groups, the within age group sample distances are larger than at later age groups. This indicates that the microbiota in early age groups are more sporadic, whereas at later age groups, microbial composition becomes more similar.

```
##make 3 month age categories
meta$age_month  = as.numeric(meta$age_month)+2
meta$age_3mo[meta$age_month >= 3 & meta$age_month <= 5] <- '3to6'
meta$age_3mo[meta$age_month >= 6 & meta$age_month <= 8] <- '6to9'
meta$age_3mo[meta$age_month >= 9 & meta$age_month <= 11] <- '9to12'
meta$age_3mo[meta$age_month >= 12 & meta$age_month <= 14] <- '12to15'
meta$age_3mo[meta$age_month >= 15 & meta$age_month <= 18] <- '15to18'
meta$age_3mo[meta$age_month > 18] <- 'over18'
meta$age_3mo = factor(meta$age_3mo, levels = c('3to6', '6to9', '9to12', '12to15', '15to18', 'over18'))
##subset meta data by site
turku.meta = meta[meta$site == 'Turku',]
tampere.meta = meta[meta$site == 'Tampere',]
oulu.meta = meta[meta$site == 'Oulu',]

#betadisper - dispersion or homogeneity of groups
disp.turku = with(turku.meta, betadisper(WUF.dist.turku, turku.meta$age_3mo))
disp.tampere = with(tampere.meta, betadisper(WUF.dist.tampere, tampere.meta$age_3mo))
disp.oulu = with(oulu.meta, betadisper(WUF.dist.oulu, oulu.meta$age_3mo))

#merge betadisper results to plot
disp.oulu.df = data.frame(site = rep('Oulu', length(disp.oulu$group)),
                          group = disp.oulu$group,
                          distances = disp.oulu$distances)
disp.tampere.df = data.frame(site = rep('Tampere', length(disp.tampere$group)),
                          group = disp.tampere$group,
                          distances = disp.tampere$distances)
disp.turku.df = data.frame(site = rep('Turku', length(disp.turku$group)),
                          group = disp.turku$group,
                          distances = disp.turku$distances)
disp.all.df = rbind(disp.oulu.df, disp.tampere.df, disp.turku.df)

#dispersion plot
fresh_swap = c('#65ADC2', '#233B43', '#E84646', '#C29365', '#362C21', '#316675', '#168E7F', '#109B37')
ggplot(disp.all.df, aes(group, distances)) +
  geom_boxplot(fill='white') +
  facet_wrap(~site) +
  ylab('Distance to centroid') +
  xlab('Age (months)') +
  scale_color_manual(values=fresh_swap, name='Age at sampling (months)') +
  theme_bw()
```



Microbiota dispersion between subjects declines with age. Pair-wise weighted Unifrac distances were calculated for all samples and stratified by site and collected in 3 month groups. Across all sites, dispersion of microbial composition was greatest in the earliest age group, 3 to 6 months, and declined and stabilized by 9 to 12 months of age.

Identifying Differentially Abundant Taxa

Approach

In the following section, generalized estimating equations (GEEs) are used to identify taxa associated with seroconversion. GEEs are applied at different taxonomic levels (all taxa at Phylum and Family levels, but only genera and species of significant Bacteroidetes and Proteobacteria families). An effect is considered significant if the FDR-adjusted p-value is <0.05.

Initially, only the simple formula `Abundance~Age^2*serconverted, id=mask_id` was considered. An adjusted formula was then tested for any taxa in which serocnversion status had a significant effect.

Results were displayed graphically by sliding, smoothed plots of significant taxa from adjusted models. In an attempt to identify specific ages where the relative abundance of taxa had the greatest differences between cases and controls, as well as a secondary approach to validate the GEE results, Mann-Whitney test (with FDR correction, $p<0.05$) was performed. Mann-Whitney results were represented along the x-axis of the sliding smoothed plots, implemented with the `geom_rug` layer.

Generalized estimating equations is a method which aims to estimate the average response over the population ('population-aaveraged' effects). Because high inter-individual variation in the microbiota was expected and demonstrated, GEEs were selected to test the population average relative abundances of taxa between cases and controls rather than attempting to model individual-specific effects. This method is commonly used in epidemiological studies, especially multi-site cohort studies because they can handle many types of unmeasured dependence between outcome.

Phylum

Proteobacteria & Bacteroidetes are associated with seroconversion

Initially, the simple model `Abundance ~ Age^2*seroconverted, id = mask_id` was considered in order to identify any phyla that are affected by seroconversion status. Based on PERMANOVA results, age and subject greatly influence microbial variation. Thus, subject effects were accounted for by specifying the argument `id = mask_id`, and age was included in the model formula as an interaction term.

Only significicant (FDR adjusted p-value < 0.05) are displayed.

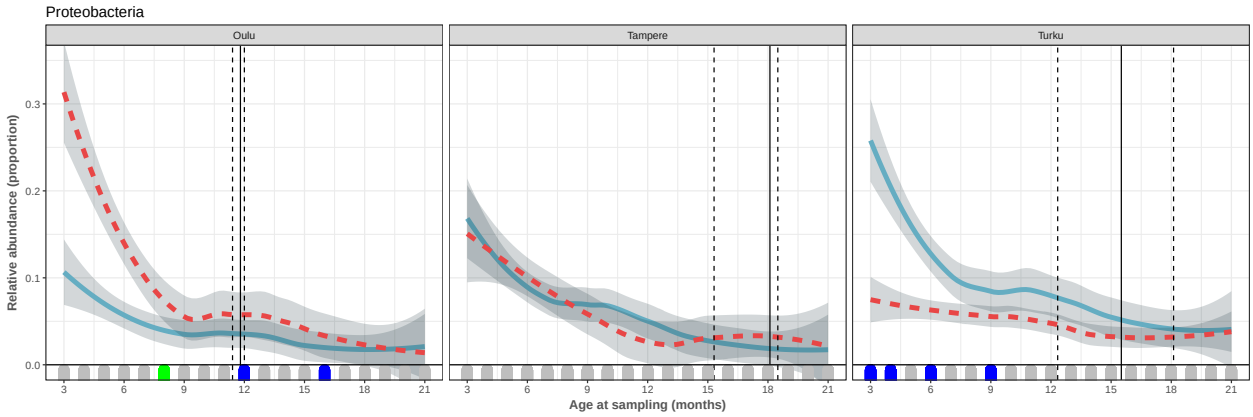
```
#long-format
dipp.p.m = psmelt(dipp.p)
dipp.p.m$mask_id  = factor(dipp.p.m$mask_id)
```

```
#run GEE
gee.p.sc = ddply(dipp.p.m, ~Phylum, function(x) test_sero_gee(x, 'Abundance'))

#extract significant results
sig.phy.sc.list = get_gee_sig_taxa(gee.p.sc, 0.05, c('seroconvertedTRUE', 'I(age_month^2):seroconverted

kable(subset(sig.phy.sc.list, Phylum %in% c('Bacteroidetes', 'Proteobacteria')))
```

	Phylum	Estimate	Std.err	Wald	p_value	p.adjust	effect
15	Bacteroidetes	0.0785997	0.0269471	8.507798	0.0035363	0.0047150	seroconvertedTRUE
39	Proteobacteria	-0.0275369	0.0092612	8.840844	0.0029456	0.0039275	seroconvertedTRUE
40	Proteobacteria	0.0000675	0.0000272	6.142433	0.0131976	0.0131976	I(age_month^2):seroconvertedTRUE

Figure 3A-B

```
dipp.p.m2 = dipp.p.m[dipp.p.m$age_month <=21,]
Prot = plot_gee_sig_taxa('Proteobacteria', dipp.p.m2, 'Abundance', 'Phylum', 'seroconverted', 'seroconv
Prot + coord_cartesian(ylim=c(0,0.35)) + geom_hline(yintercept=0, color='black') + ggtitle('Proteobacte
```




```
#samples from C-section subjects were removed; see confounders section
Bact = plot_gee_sig_taxa('Bacteroidetes', dipp.p.m2[dipp.p.m2$MoD_simp == 'Vaginal',], 'Abundance', 'Phy
Bact + coord_cartesian(ylim=c(0,0.8)) + geom_hline(yintercept=0, color='black') + ggtitle('Bacteroidetes
```

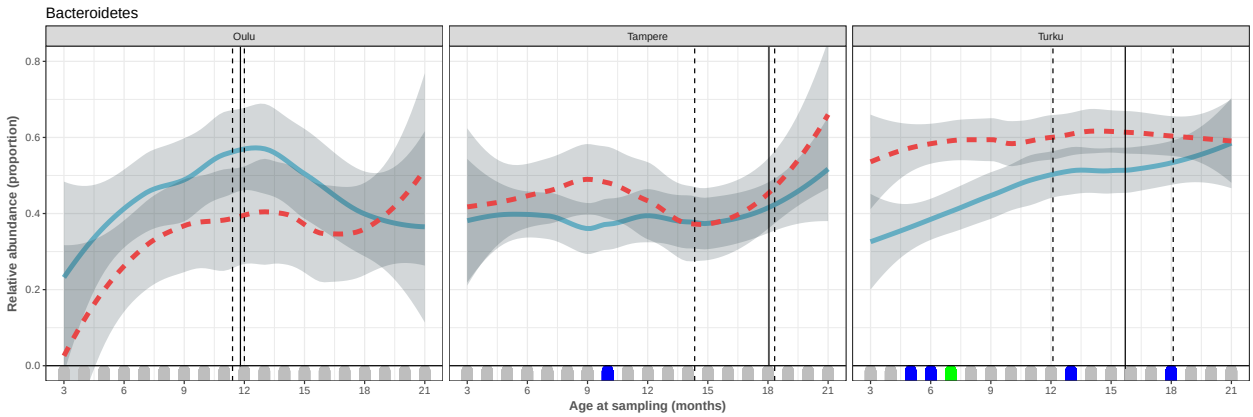


Figure 3A-B *Proteobacteria and Bacteroidetes* While the relative abundance of Proteobacteria declines with age, it was associated with cases in Oulu, but controls in Turku, while no association was observed in Tampere. Meanwhile, the relative abundance of Bacteroidetes increased with age and was significantly associated with cases in Turku and to a lesser extent in Tampere, while the inverse relationship was significant in Oulu.

Confounders & Bacteroidetes

Based on previous results, an adjusted model was tested considering additional variables that had an effect on overall microbial variation (identified in the PERMANOVA section). While there was a distinction of initial trends by site, additional confounders did have a significant effect on Bacteroidetes but not Proteobacteria.

Figure 3C

```
dipp.p.m.noC = dipp.p.m2[dipp.p.m2$MoD_simp =='Vaginal',]
#variable of interest must be TRUE/FALSE == MALE/FEMALE
dipp.p.m.noC$Gender = as.logical(as.numeric(dipp.p.m.noC$Gender)-1)
#gender
Bact.gender = plot_gee_sig_taxa_var('Bacteroidetes', dipp.p.m.noC, 'Abundance', 'Phylum', 'Gender', 'ag

Bact.gender + coord_cartesian(ylim=c(0, 0.65)) + geom_hline(yintercept=0, color='black') + theme(legend
```

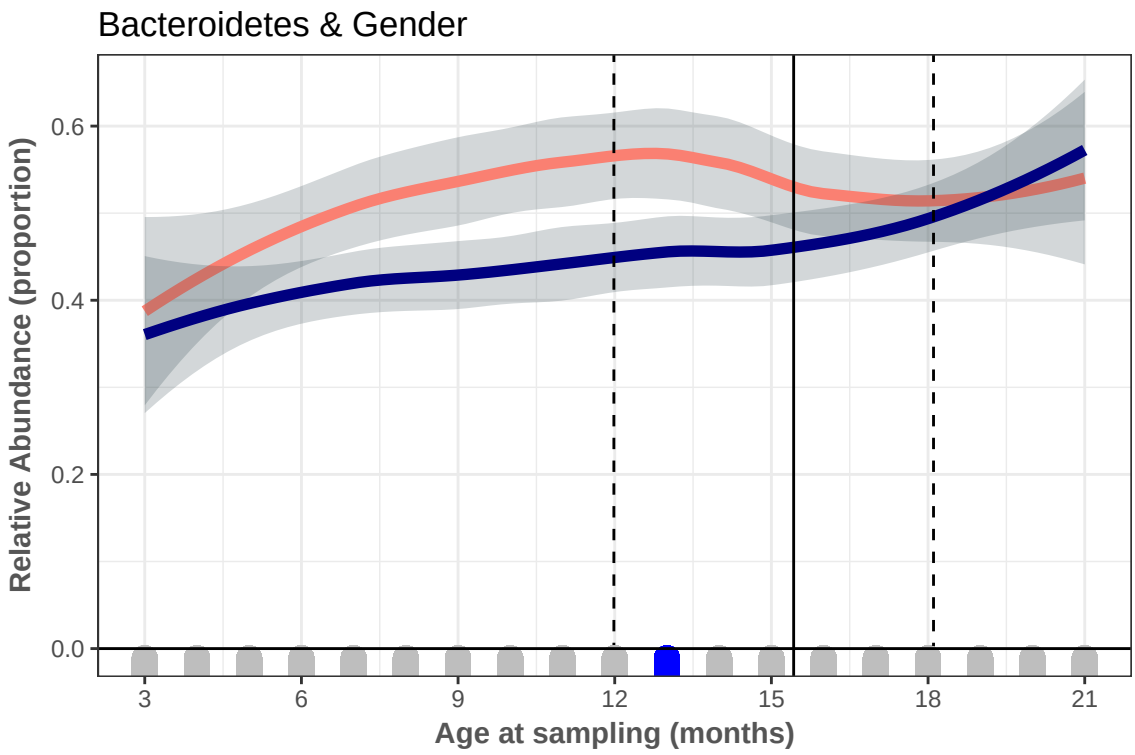
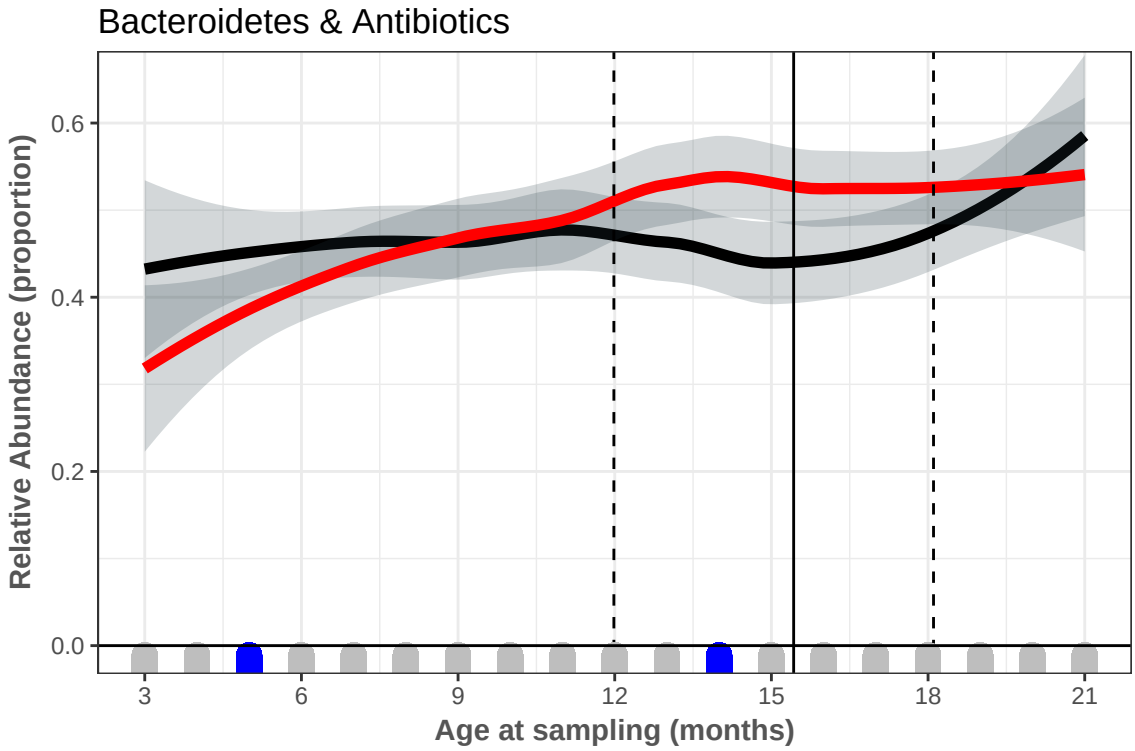


Figure 3C *Bacteroidetes and Gender* Females (light pink color) tend to have higher relative abundance of Bacteroidetes compared to males (navy blue color). However, this difference was most evident near 13 months of age and did not persist across all ages sampled.

```
#antibiotics
Bact.abx = plot_gee_sig_taxa_var('Bacteroidetes', dipp.p.m.noC, 'Abundance', 'Phylum', 'antibiotics', '

Bact.abx + coord_cartesian(ylim=c(0,0.65)) + geom_hline(yintercept=0, color='black') + theme(legend.pos
```



*Subjects that received antibiotics within the first 2 years of life are represented in **red**, those that did not receive antibiotics are in **black**.

Initial results indicated that there was an effect of mode of delivery on Bacteroidetes. However, due to the small number of subjects born by C-section, they were excluded here.

For Bacteroidetes, as previously reported, seroconversion was associated with higher relative abundance in Turku and Tampere, but the inverse was observed in Oulu when accounting for site, gender, receiving breast milk, and antibiotics within the first 2 years of life. Additionally, gender has a significant effect (observed across all sites) and tended to be higher in females. While antibiotic use also had a significant effect.

While Proteobacteria initially seemed to be associated with controls, the inverse effect was observed in Oulu. Meanwhile, there were no additional effects of any of the variables included in the adjusted model.

```
#make Turku the reference group
dipp.p.m$site = factor(dipp.p.m$site, levels=c('Turku', 'Tampere', 'Oulu'))
#extract only relevant phyla
dipp.p.m.BP = subset(dipp.p.m, Phylum %in% c('Bacteroidetes', 'Proteobacteria'))
#exclude C-section subjects
dipp.p.m.BP.noC = dipp.p.m.BP[dipp.p.m.BP$MoD_simp == 'Vaginal',]

#run adjusted GEE
gee.p.adj = ddply(dipp.p.m.BP.noC, ~Phylum, function(x) test_adj_gee(x, 'Abundance'))


gee.p.adj[gee.p.adj$p.adjust <=0.05,]
```

```
##           Phylum      Estimate      Std.err       Wald      p_value
## 1   Bacteroidetes  0.4952114715 3.669010e-02 182.172827 0.000000e+00
## 3   Bacteroidetes  0.1626559850 3.704812e-02  19.275589 1.131437e-05
## 6   Bacteroidetes -0.0939158402 2.820363e-02  11.088369 8.687087e-04
## 8   Bacteroidetes -0.0513651095 1.927386e-02   7.102298 7.698517e-03
```

```
## 13  Bacteroidetes -0.3923400627 7.774706e-02  25.465802 4.502942e-07
## 14  Bacteroidetes  0.0002693036 1.059451e-04   6.461338 1.102466e-02
## 16  Bacteroidetes  0.0008937316 2.802412e-04  10.170689 1.426910e-03
## 17 Proteobacteria  0.1028089452 1.502228e-02  46.837113 7.713830e-12
## 18 Proteobacteria -0.0002165938 4.088899e-05  28.059453 1.176451e-07
## 19 Proteobacteria -0.0629432284 1.391036e-02  20.474878 6.041907e-06
## 21 Proteobacteria -0.0672042894 1.475690e-02  20.739771 5.261186e-06
## 25 Proteobacteria  0.0001603334 3.942693e-05  16.537202 4.770487e-05
## 27 Proteobacteria  0.0001424052 4.290426e-05  11.016679 9.029569e-04
## 29 Proteobacteria  0.1340773833 2.372255e-02  31.943914 1.586887e-08
## 32 Proteobacteria -0.0003496849 7.030217e-05  24.740948 6.557605e-07
##        p.adjust                                    effect
## 1  0.000000e+00                                (Intercept)
## 3  6.034331e-05                          seroconvertedTRUE
## 6  3.474835e-03                                 GenderMale
## 8  2.052938e-02                              antibioticsTRUE
## 13 3.602354e-06                 seroconvertedTRUE:siteOulu
## 14 2.519922e-02                  I(age_month^2):GenderMale
## 16 4.566111e-03 I(age_month^2):seroconvertedTRUE:siteOulu
## 17 1.234213e-10                                (Intercept)
## 18 6.274403e-07                             I(age_month^2)
## 19 1.611175e-05                          seroconvertedTRUE
## 21 1.611175e-05                                    siteOulu
## 25 1.090397e-04           I(age_month^2):seroconvertedTRUE
## 27 1.805914e-03                    I(age_month^2):siteOulu
## 29 1.269510e-07                 seroconvertedTRUE:siteOulu
## 32 2.623042e-06 I(age_month^2):seroconvertedTRUE:siteOulu
```

Family

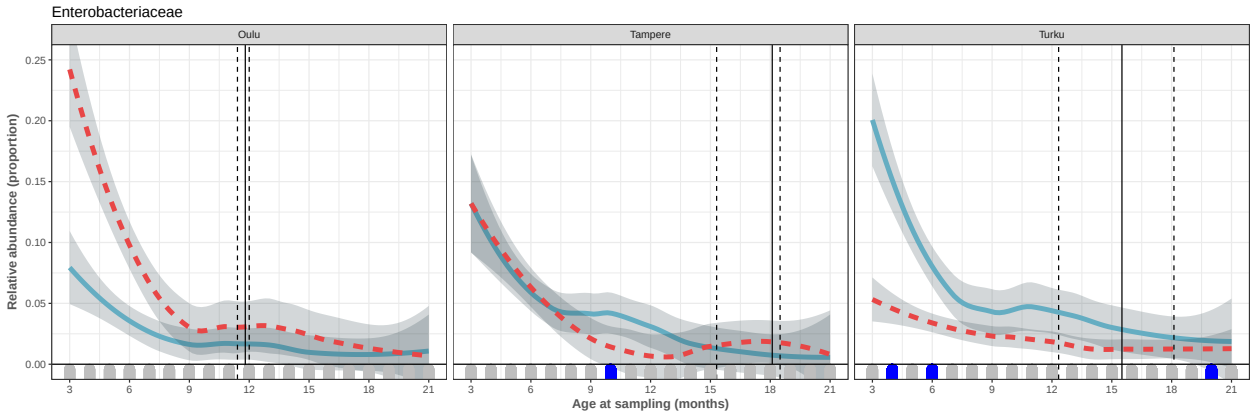
After testing all 91 families, 25 had a significant association with seroconversion status using the simple formula `Abundance ~ Age^2*seroconverted, id = mask_id`. Most of these families (17 of 25) were Proteobacteria, followed by Bacteroidetes (4 of 25) and Firmicutes (2 of 25). Focus was given to the most abundant members of Bacteroidetes and Proteobacteria:

```
#long-format
dipp.f.m = psmelt(dipp.f)
dipp.f.m$mask_id  = factor(dipp.f.m$mask_id)
```

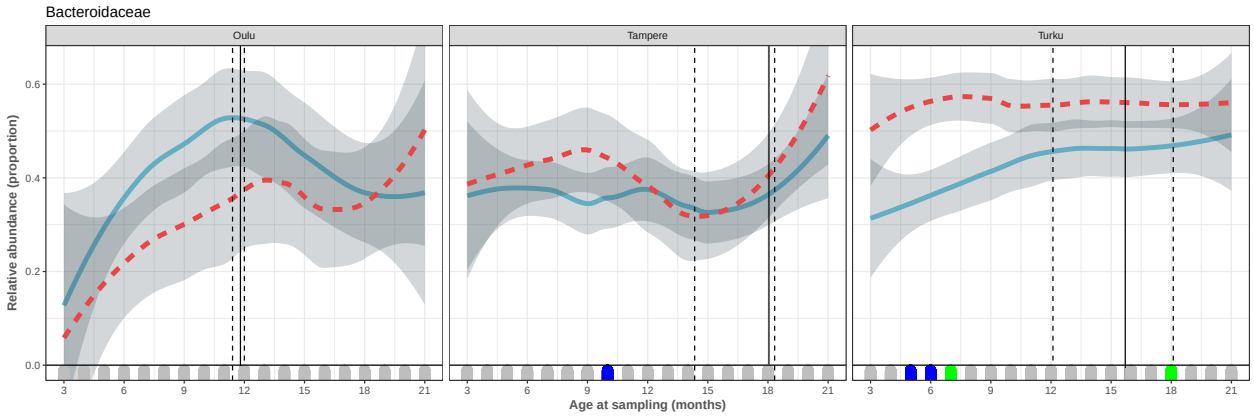
```
#run GEE
gee.f.sc = ddply(dipp.f.m, ~Phylum+Family, function(x) test_sero_gee(x, 'Abundance'))

#extract significant results
sig.fam.sc.list = get_gee_sig_taxa(gee.f.sc, 0.05, c('seroconvertedTRUE', 'I(age_month^2):seroconverted

kable(subset(sig.fam.sc.list, Family %in% c('Bacteroidaceae', '[Paraprevotellaceae]', 'Enterobacteriacea
```

	Phylum	Family	Estimate	Std.err	Wald	p_value	p.adjust	effect
67	Bacteroidetes	[Paraprevotellaceae]	0.0018498	0.0006488	8.129481	0.0043551	0.0058068	seroconvertedT
71	Bacteroidetes	Bacteroidaceae	0.0736456	0.0261911	7.906539	0.0049256	0.0065675	seroconvertedT
283	Proteobacteria	Enterobacteriaceae	-0.0231042	0.0073367	9.917098	0.0016375	0.0021833	seroconvertedT
284	Proteobacteria	Enterobacteriaceae	0.0000534	0.0000218	6.025046	0.0141043	0.0141043	I(age_month^2

```
dipp.f.m2 = dipp.f.m[dipp.f.m$age_month <=21,]

Ent = plot_gee_sig_taxa('Enterobacteriaceae', dipp.f.m2, 'Abundance', 'Family', 'seroconverted', 'seroc
Ent + coord_cartesian(ylim=c(0,0.25)) + geom_hline(yintercept=0, color='black') + ggtitle('Enterobacter
```




```
#samples from C-section subjects were removed; see confounders section
Bact.f = plot_gee_sig_taxa('Bacteroidaceae', dipp.f.m2[dipp.f.m2$MoD_simp == 'Vaginal',], 'Abundance',
Bact.f + coord_cartesian(ylim=c(0,0.65)) + geom_hline(yintercept=0, color='black') + ggtitle('Bacteroida
```



Taxonomic differences were most apparent within Bacteroidetes (specifically the families Bacteroidaceae and [Paraprevotellaceae]) and Proteobacteria (specifically the family Enterobacteriaceae). Therefore, these families were selected and investigated at the genus and species levels.

Bacteroidetes - A Closer Look

Bacteroidetes were extracted, taxa compressed to genus and species levels, then only taxa assigned to *Bacteroidaceae* and *[Paraprevotellaceae]* were considered in further analyses.

Genus

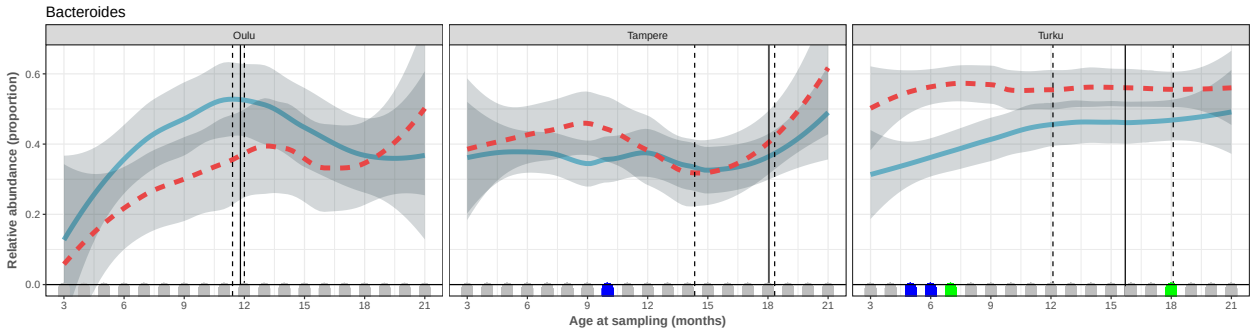
At the genus level, Bacteroides and [Prevotella] dominated their respective families and the relationship between cases and controls remained significant.

```
#run GEE
gee.gB.sc = ddply(dipp.g.BP.m, ~Phylum+Family+Genus, function(x) test_sero_gee(x, 'Abundance'))
```

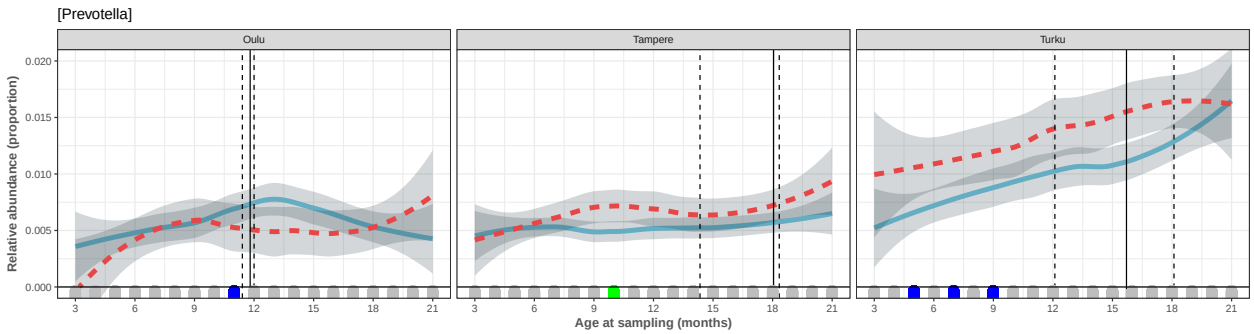
```
#extract significant results
sig.genBP.sc.list = get_gee_sig_taxa(gee.gB.sc, 0.05, c('seroconvertedTRUE', 'I(age_month^2):seroconver

kable(subset(sig.genBP.sc.list, Genus %in% c('Bacteroides', '[Prevotella]')))
```

	Phylum	Family	Genus	Estimate	Std.err	Wald	p_value	p.adjust	effec
3	Bacteroidetes	[Paraprevotellaceae]	[Prevotella]	0.0018739	0.0006338	8.742830	0.0031082	0.0041443	seroc
19	Bacteroidetes	Bacteroidaceae	Bacteroides	0.0735252	0.0263636	7.777892	0.0052889	0.0070519	seroc

```
dipp.g.BP.m2 = dipp.g.BP.m[dipp.g.BP.m$age_month <=21,]
dipp.g.BP.m.noC = dipp.g.BP.m2[dipp.g.BP.m2$MoD_simp == 'Vaginal',]

Bact.g = plot_gee_sig_taxa('Bacteroides', dipp.g.BP.m.noC, 'Abundance', 'Genus', 'seroconverted', 'seroc
Bact.g + coord_cartesian(ylim=c(0,0.65)) + geom_hline(yintercept=0, color='black') + ggtitle('Bacteroid
```




```
#samples from C-section subjects were removed; see confounders section
Prev.g = plot_gee_sig_taxa('[Prevotella]', dipp.g.BP.m.noC, 'Abundance', 'Genus', 'seroconverted', 'sero
Prev.g + coord_cartesian(ylim=c(0,0.02)) + geom_hline(yintercept=0, color='black') + ggtitle('[Prevotel
```



Adjusted models were considered for Bacteroides and [Prevotella]. Gender and antibiotics also had an effect on the relative abundance of Bacteroidetes at the genus level for Bacteroides. Antibiotics, but not gender, had an effect on the relative abundance of [Prevotella].

```
#make Turku the reference group
dipp.g.BP.m.noC$site = factor(dipp.g.BP.m.noC$site, levels=c('Turku', 'Tampere', 'Oulu'))
#extract only relevant genera
dipp.g.BP.m.noC = subset(dipp.g.BP.m.noC, Genus %in% c('Bacteroides', '[Prevotella]'))

#run adjusted GEE
gee.g.BP.adj = ddply(dipp.g.BP.m.noC, ~Genus, function(x) test_adj_gee(x, 'Abundance'))
```

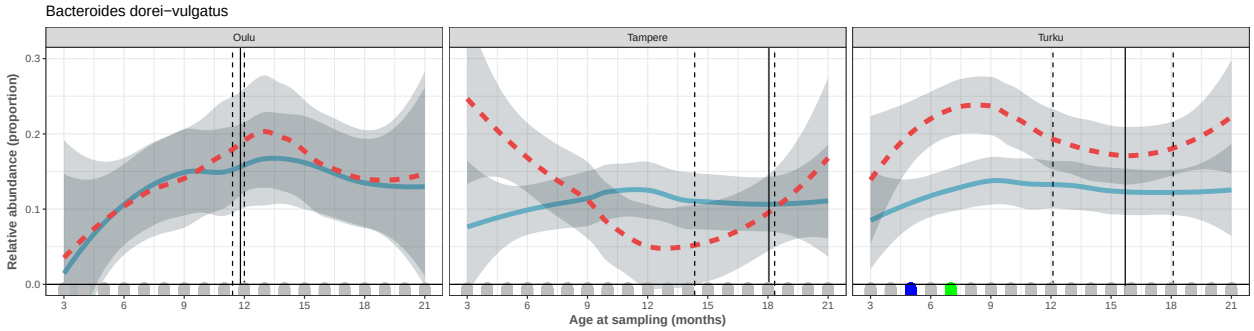
```
#extract significant results implicating seroconversion
gee.g.BP.adj[gee.g.BP.adj$p.adjust <= 0.05,]
```

```
##           Genus      Estimate      Std.err       Wald      p_value
## 1  [Prevotella]  8.504048e-03 1.047921e-03  65.855793 4.440892e-16
## 2  [Prevotella]  1.766157e-05 5.162240e-06  11.705289 6.232272e-04
## 3  [Prevotella]  3.883629e-03 1.114259e-03  12.147947 4.914193e-04
## 4  [Prevotella] -2.191464e-03 9.522824e-04   5.295869 2.137606e-02
## 8  [Prevotella] -1.175497e-03 5.192630e-04   5.124700 2.358765e-02
## 10 [Prevotella] -1.880662e-05 5.649421e-06  11.081878 8.717545e-04
## 11 [Prevotella] -2.047321e-05 5.817857e-06  12.383571 4.331280e-04
## 12 [Prevotella] -3.242658e-03 1.384274e-03   5.487289 1.915522e-02
## 13 [Prevotella] -5.930190e-03 1.642586e-03  13.034100 3.058702e-04
## 17  Bacteroides  4.608557e-01 3.799176e-02 147.146835 0.000000e+00
## 19  Bacteroides  1.762350e-01 3.914659e-02  20.267321 6.734103e-06
## 22  Bacteroides -7.949322e-02 2.975835e-02   7.135796 7.556011e-03
## 24  Bacteroides -6.502807e-02 1.956985e-02  11.041461 8.909662e-04
## 29  Bacteroides -3.732459e-01 8.378233e-02  19.846558 8.391367e-06
##        p.adjust                         effect
## 1  7.105427e-15                    (Intercept)
## 2  1.994327e-03                 I(age_month^2)
## 3  1.965677e-03              seroconvertedTRUE
## 4  4.193359e-02                   siteTampere
## 8  4.193359e-02                antibioticsTRUE
## 10 2.324679e-03    I(age_month^2):siteTampere
## 11 1.965677e-03       I(age_month^2):siteOulu
## 12 4.193359e-02 seroconvertedTRUE:siteTampere
## 13 1.965677e-03    seroconvertedTRUE:siteOulu
## 17 0.000000e+00                    (Intercept)
## 19 4.475396e-05              seroconvertedTRUE
## 22 2.417924e-02                    GenderMale
## 24 3.563865e-03                antibioticsTRUE
## 29 4.475396e-05    seroconvertedTRUE:siteOulu
```

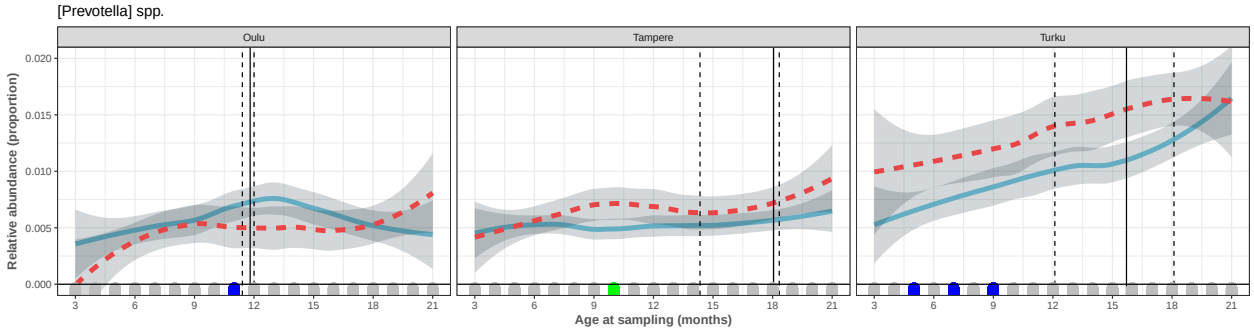
Species

At the species level, the strong association of Bacteroides and cases is mainly attributed to *Bacteroides dorei-vulgatus* in Turku, although this trend was not observed in the other 2 sites. The relative abundance of *[Prevotella] spp.* (referred to in the manuscript as *Paraprevotella spp.*) was associated with cases in Tampere and Turku. Meanwhile the inverse association of Bacteroides in Oulu was primarily attributed to *Bacteroides fragilis*, whose relative abundance was higher in controls in both Oulu and Tampere, but not observed in Turku.

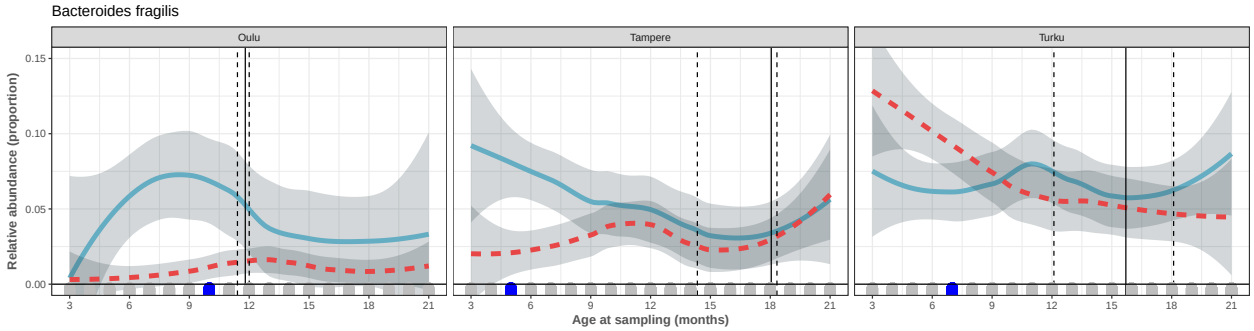
```
dipp.s.BP.m2 = dipp.s.BP.m[dipp.s.BP.m$age_month <=21,]
dipp.s.BP.m.noC = dipp.s.BP.m2[dipp.s.BP.m2$MoD_simp == 'Vaginal',]

Bdv.s = plot_gee_sig_taxa('Bacteroides_dorei-vulgatus', dipp.s.BP.m.noC, 'Abundance', 'species', 'seroc
Bdv.s + coord_cartesian(ylim=c(0,0.30)) + geom_hline(yintercept=0, color='black') + ggtitle('Bacteroides
```




```
Prev.s = plot_gee_sig_taxa('[Prevotella]_', dipp.s.BP.m.noC, 'Abundance', 'species', 'seroconverted', '
Prev.s + coord_cartesian(ylim=c(0,0.02)) + geom_hline(yintercept=0, color='black') + ggtitle('[Prevotel
```




```
B.frag = plot_gee_sig_taxa('Bacteroides_fragilis', dipp.s.BP.m.noC, 'Abundance', 'species', 'seroconver
B.frag + coord_cartesian(ylim=c(0,0.15)) + geom_hline(yintercept=0, color='black') + ggtitle('Bacteroid
```




```
#relevel site
dipp.s.BP.m.noC$site = factor(dipp.s.BP.m.noC$site, levels=c('Turku', 'Tampere', 'Oulu'))
#extract only Bacteroides & [Prevotella]
dipp.s.BP.m.noC.BP = subset(dipp.s.BP.m.noC, Genus %in% c('Bacteroides', '[Prevotella]'))

#run GEE
gee.s.BP.sc = ddply(dipp.s.BP.m.noC.BP, ~Phylum+Family+Genus+species, function(x) test_sero_gee(x, 'Abu
#run adjusted GEE
gee.s.BP.adj = ddply(dipp.s.BP.m.noC.BP, ~Phylum+Family+Genus+species, function(x) test_adj_gee(x, 'Abu

#extract significant results
sig.spec.BP.sc.list = get_gee_sig_taxa(gee.s.BP.sc, 0.05, c('seroconvertedTRUE', 'I(age_month^2):seroco
  ##only Bacteroides dorei-vulgatus and [Prevotella] have a significant association under the simple fo
sig.spec.BP.adj.list = get_gee_sig_taxa(gee.s.BP.adj, 0.05, c('seroconvertedTRUE', 'seroconvertedTRUE:s

kable(subset(sig.spec.BP.adj.list, species %in% c('Bacteroides_dorei-vulgatus', 'Bacteroides_fragilis',
```

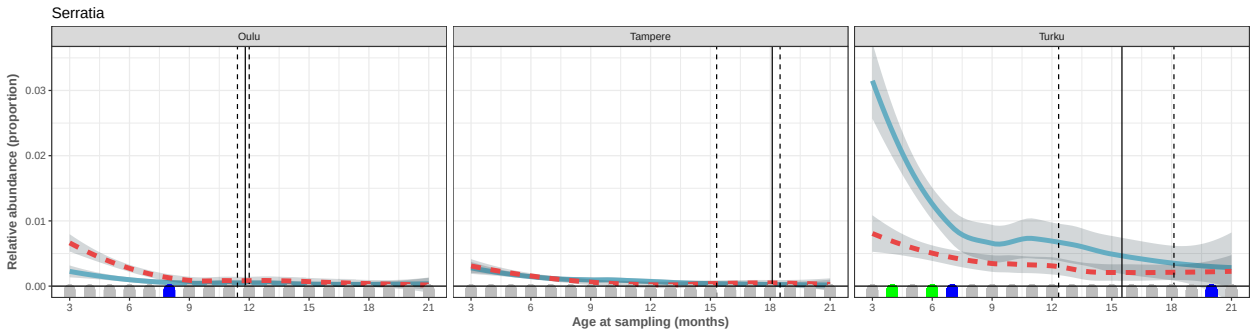
	Phylum	Family	Genus	species	Estimate	Std.err	Wal
3	Bacteroidetes	[Paraprevotellaceae]	[Prevotella]	[Prevotella]_	0.0040108	0.0011148	12.94365
12	Bacteroidetes	[Paraprevotellaceae]	[Prevotella]	[Prevotella]_	-0.0034222	0.0013819	6.13263
13	Bacteroidetes	[Paraprevotellaceae]	[Prevotella]	[Prevotella]_	-0.0063941	0.0015531	16.95049
131	Bacteroidetes	Bacteroidaceae	Bacteroides	Bacteroides_dorei-vulgatus	0.1010234	0.0248011	16.59218
172	Bacteroidetes	Bacteroidaceae	Bacteroides	Bacteroides_fragilis	-0.0823066	0.0200895	16.78535
173	Bacteroidetes	Bacteroidaceae	Bacteroides	Bacteroides_fragilis	-0.0862327	0.0243413	12.55038

Proteobacteria - A Closer Look

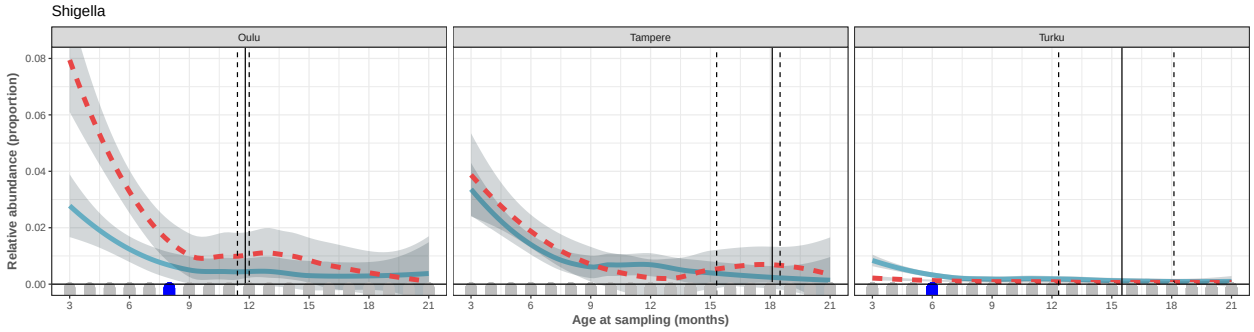
Genus

There were 11 genera assigned to *Enterobacteriaceae*. **Serratia** and **Shigella** was selected for further investigation based on a significant GEE result and median relative abundance >0.1%. However, these trends varied by site, while similar to at the phylum level, additional variables considered in the adjusted models did not have a significant effect. Shigella was associated with controls in Turku, but this trend was not observed in the other 2 sites. Meanwhile, Shigella was associated with cases in Oulu, but not observed in Tampere or Turku.

```
dipp.TR.Ent.g.m2 = dipp.TR.Ent.g.m[dipp.TR.Ent.g.m$age_month <=21,]

Serr.g = plot_gee_sig_taxa('Serratia', dipp.TR.Ent.g.m2, 'Abundance', 'Genus', 'seroconverted', 'seroco
Serr.g + coord_cartesian(ylim=c(0,0.035)) + geom_hline(yintercept=0, color='black') +ggtitle('Serratia'
```




```
Shig.g = plot_gee_sig_taxa('Shigella', dipp.TR.Ent.g.m2, 'Abundance', 'Genus', 'seroconverted', 'seroco
Shig.g + coord_cartesian(ylim=c(0,0.08)) + geom_hline(yintercept=0, color='black') + ggtitle('Shigella'
```




```
#relevel site
dipp.TR.Ent.g.m$site = factor(dipp.TR.Ent.g.m$site, levels=c('Turku', 'Tampere', 'Oulu'))

#run adjusted GEE
```

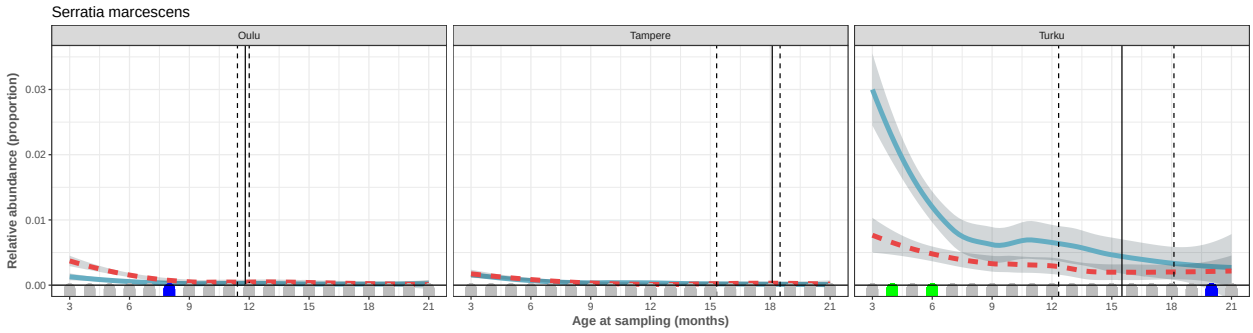
```
gee.g.Ent.adj = ddply(dipp.TR.Ent.g.m, ~Phylum+Family+Genus, function(x) test_adj_gee(x, 'Abundance'))

sig.gen.Ent.adj.list = get_gee_sig_taxa(gee.g.Ent.adj, 0.05, c('seroconvertedTRUE', 'seroconvertedTRUE:s
gee.adj.Serr = gee.g.Ent.adj[gee.g.Ent.adj$Genus == 'Serratia',]
gee.adj.Shig = gee.g.Ent.adj[gee.g.Ent.adj$Genus == 'Shigella',]

kable(subset(sig.gen.Ent.adj.list, Genus %in% c('Serratia', 'Shigella')))
```

	Phylum	Family	Genus	Estimate	Std.err	Wald	p_value	p.adjust	effect
147	Proteobacteria	Enterobacteriaceae	Serratia	-0.0079043	0.0017224	21.06104	0.0000044	0.0000089	seroco
156	Proteobacteria	Enterobacteriaceae	Serratia	0.0077145	0.0018137	18.09269	0.0000210	0.0000337	seroco
157	Proteobacteria	Enterobacteriaceae	Serratia	0.0096040	0.0017850	28.95017	0.0000001	0.0000003	seroco
163	Proteobacteria	Enterobacteriaceae	Shigella	-0.0021399	0.0005956	12.90959	0.0003269	0.0017435	seroco
173	Proteobacteria	Enterobacteriaceae	Shigella	0.0191117	0.0059712	10.24412	0.0013712	0.0054848	seroco

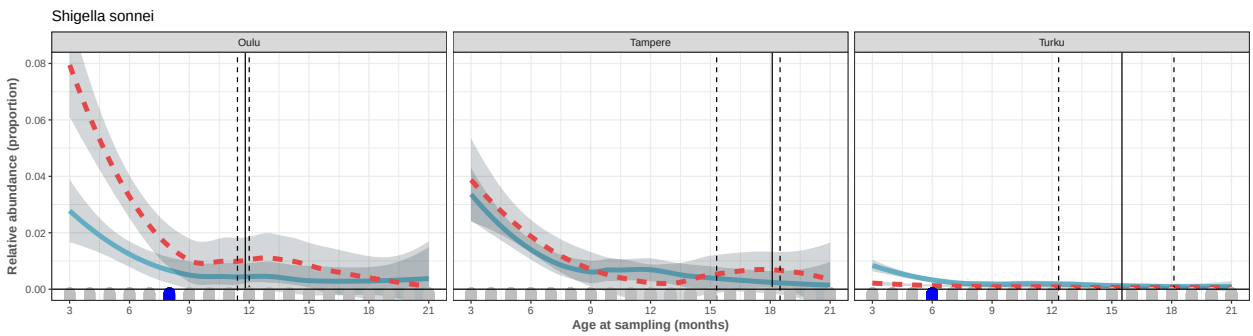
Species

The majority of reads assigned to *Serratia* and *Shigella* were attributed to *Serratia marcescens* and *Shigella sonnei*, respectively. These species were selected for their significant association with seroconversion based on GEE results and for having a median relative abundance >0.1%. It should be noted that there are inverse effects across sites. However, due to the low relative abundance of *Serratia* in Oulu and Tampere and *Shigella* in Turku, we cannot reliably quantify these species in those respective sites.

```
dipp.TR.Ent.s.m2 = dipp.TR.Ent.s.m[dipp.TR.Ent.s.m$age_month <=21,]

SerrM.g = plot_gee_sig_taxa('Serratia_marcescens', dipp.TR.Ent.s.m2, 'Abundance', 'Species', 'seroconver
SerrM.g + coord_cartesian(ylim=c(0,0.035)) + geom_hline(yintercept=0, color='black') +ggtitle('Serratia
```




```
ShigS.g = plot_gee_sig_taxa('Shigella_sonnei', dipp.TR.Ent.s.m2, 'Abundance', 'Species', 'seroconverted
ShigS.g + coord_cartesian(ylim=c(0,0.08)) + geom_hline(yintercept=0, color='black') + ggtitle('Shigella
```



```
#relevel site
dipp.TR.Ent.s.m$site = factor(dipp.TR.Ent.s.m$site, levels=c('Turku', 'Tampere', 'Oulu'))
#extract only Shigella and Serratia
dipp.TR.Ent.s.m.SS = subset(dipp.TR.Ent.s.m, Genus %in% c('Serratia', 'Shigella'))

#run GEE
gee.s.Ent.sc = ddply(dipp.TR.Ent.s.m.SS, ~Phylum+Family+Genus+Species, function(x) test_sero_gee(x, 'Abu
#run adjusted GEE
gee.s.Ent.adj = ddply(dipp.TR.Ent.s.m.SS, ~Phylum+Family+Genus+Species, function(x) test_adj_gee(x, 'Abu

#extract significant results
sig.spec.Ent.sc.list = get_gee_sig_taxa(gee.s.Ent.sc, 0.05, c('seroconvertedTRUE', 'I(age_month^2):sero
sig.spec.Ent.adj.list = get_gee_sig_taxa(gee.s.Ent.adj, 0.05, c('seroconvertedTRUE', 'seroconvertedTRUE

gee.adj.Serr.s = gee.g.Ent.adj[gee.g.Ent.adj$Genus == 'Serratia',]
gee.adj.Shig.s = gee.g.Ent.adj[gee.g.Ent.adj$Genus == 'Shigella',]


kable(subset(sig.spec.Ent.adj.list, Species %in% c('Serratia_marcescens', 'Shigella_sonnei')))
```

	Phylum	Family	Genus	Species	Estimate	Std.err	Wald	p_val
19	Proteobacteria	Enterobacteriaceae	Serratia	Serratia_marcescens	-0.0075211	0.0016351	21.15728	0.00000
28	Proteobacteria	Enterobacteriaceae	Serratia	Serratia_marcescens	0.0075170	0.0017017	19.51397	0.00001
29	Proteobacteria	Enterobacteriaceae	Serratia	Serratia_marcescens	0.0086192	0.0016616	26.90759	0.00000
99	Proteobacteria	Enterobacteriaceae	Shigella	Shigella_sonnei	-0.0021399	0.0005956	12.90962	0.00032
109	Proteobacteria	Enterobacteriaceae	Shigella	Shigella_sonnei	0.0191117	0.0059712	10.24412	0.00137

References

- knitr Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.15.1.
- phyloseq phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. Paul J. McMurdie and Susan Holmes (2013) PLoS ONE 8(4):e61217.
- vegan Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2017). vegan: Community Ecology Package. R package version 2.4-2.
- geepack HÃ¸jsgaard, S., Halekoh, U. & Yan J. (2006) The R Package geepack for Generalized Estimating Equations Journal of Statistical Software, 15, 2, pp1–11
- Unifrac Lozupone, C. A.; Hamady, M; Kelley, S. T.; Knight, R. (2007). "Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities". Applied and Environmental Microbiology 73 (5): 1576–85. doi:10.1128/AEM.01996-06.
- **Bray-Curtis** Legendre, Pierre, and Louis Legendre. "Numerical Ecology". 3rd ed. Vol. 24: Elsevier, 2012. Print.
- R R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.