# Detection of Bilirubin Reductase in the Human Gut

## Quality Control and classification of samples

Metagenomic data table consists of sample information (run ID, bioproject, geographic location, etc.), metadata (host ids, associated diseases, ages), and mapping data (reads mapped per gene, total reads in sample).

The processing consists of the following steps: 1) calculating the CPM values for both genes for each sample. 2) removing any samples with less than or equal to one million reads. 3) classifying each sample as having bilirubin reductase as present or absent based on a 5 CPM threshold.

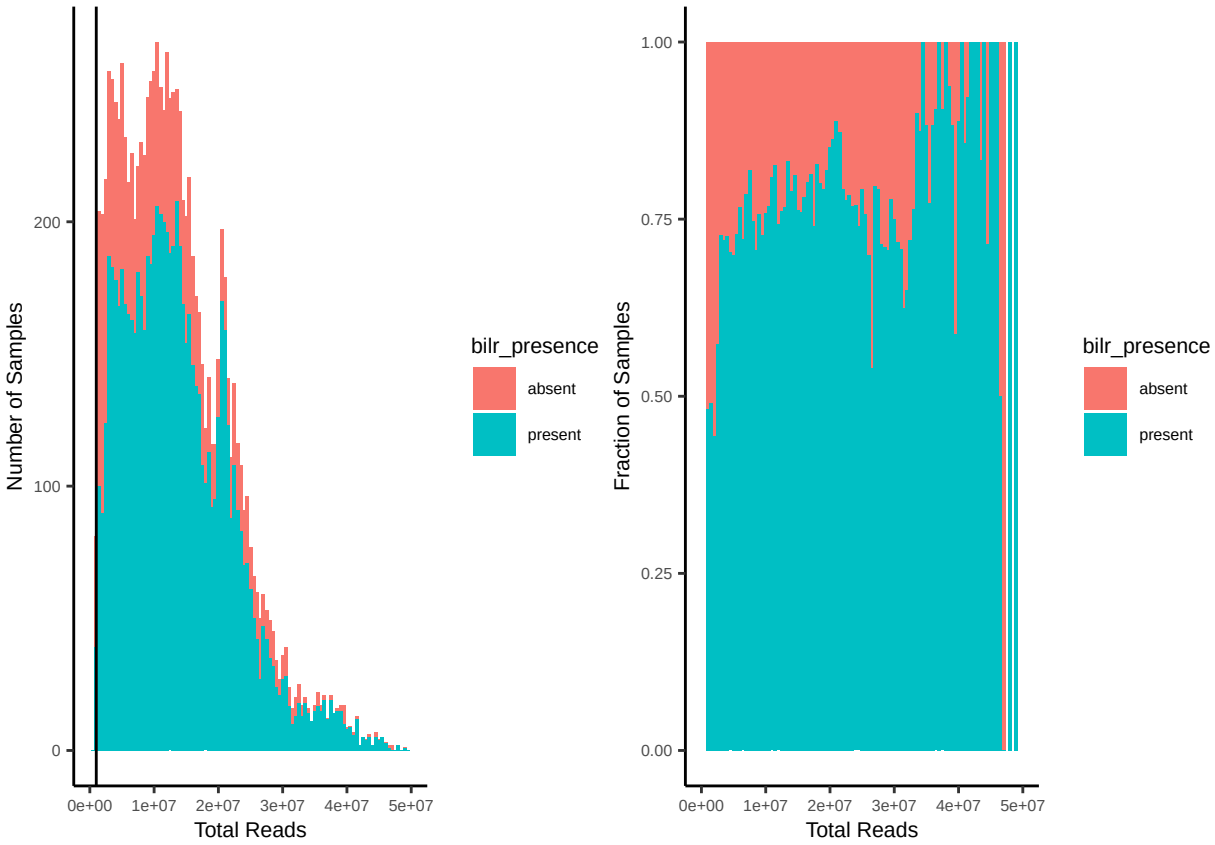
```
# Read in overall data file
project_data <- read.csv('./bilirubin_mgx_data.csv', row.names = 1)

# Calculate CPM values for the bilR gene. Remove samples with less than 1
# million reads
project_data$bilr_tot <- (project_data$bilR/project_data$total_reads)*1000000
project_data <- project_data[project_data$total_reads > 1000000,]


# Classify each sample as having bilR present or absent if CPM is greater than 5
cpm_threshold = 5
project_data$bilr_presence[project_data$bilr_tot <= cpm_threshold] <- 'absent'
project_data$bilr_presence[project_data$bilr_tot > cpm_threshold] <- 'present'
```

## Visual inspection of data

A visual inspection of the data was performed to inspect if there was an obvious relationship between total reads in the sample and bilR presence in the sample. This analysis showed that the presnce of bilR was fairly stable over most of the data range, especially in samples with between 2 and 20 million reads where most of the samples fall.

```
# Plot lower abundance bins where most data points fall.
bilr_hist <- ggplot(project_data, aes(x=total_reads, fill=bilr_presence)) +
  geom_histogram(binwidth = 500000) + pub_theme() + xlim(0,5.0e07) +
  xlab('Total Reads') + ylab('Number of Samples') + geom_vline(xintercept = 1000000)
bilr_hist_stack <- ggplot(project_data, aes(x=total_reads, fill=bilr_presence)) +
  geom_histogram(binwidth = 500000, position='fill') + pub_theme() + xlim(0,5.0e07) +
  xlab('Total Reads') + ylab('Fraction of Samples')
bilr_hist | bilr_hist_stack
```

```
## Warning: Removed 96 rows containing non-finite values (stat_bin).

## Warning: Removed 4 rows containing missing values (geom_bar).

## Warning: Removed 96 rows containing non-finite values (stat_bin).

## Warning: Removed 12 rows containing missing values (geom_bar).
```

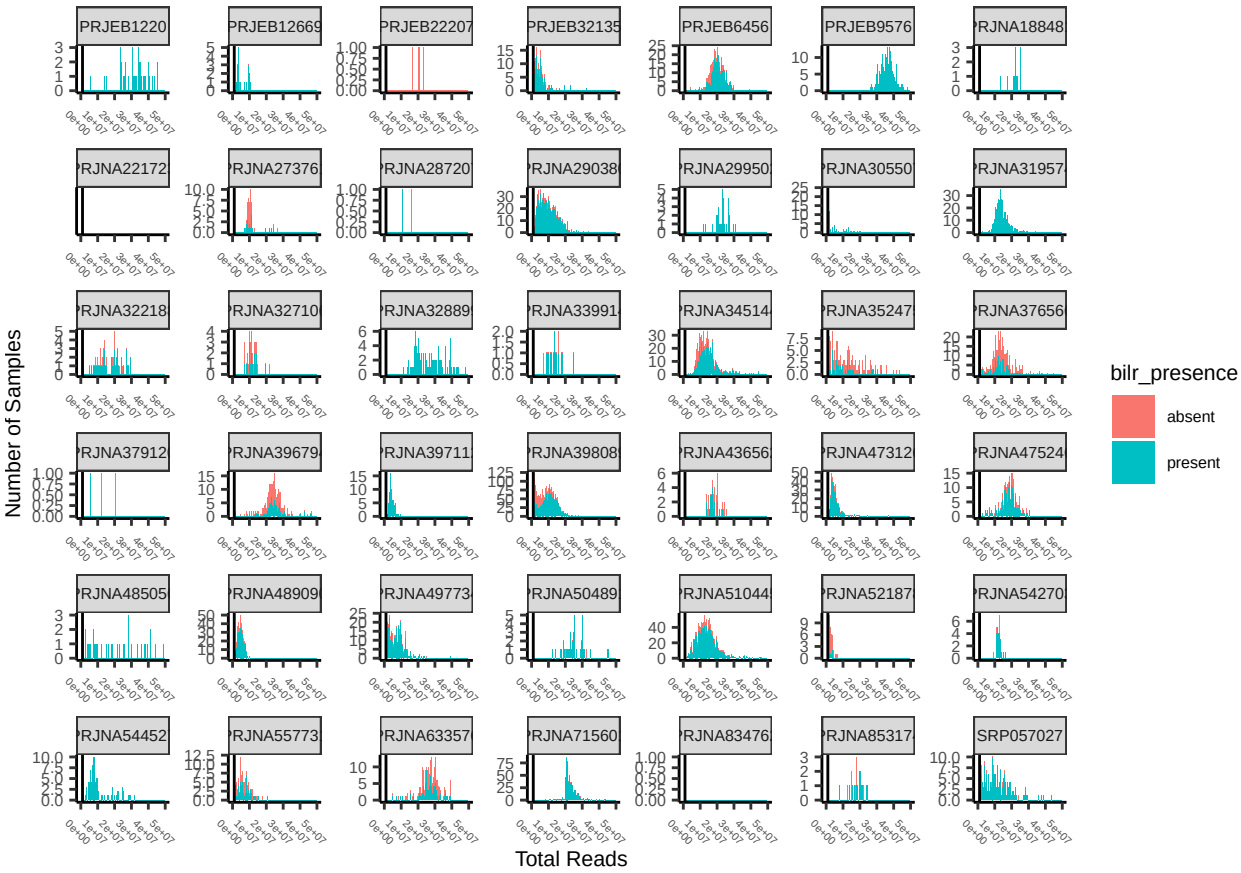



In addition to checking if there were any trends bewteen presence of bilR and total reads, we wanted to see if any studies in particulare were extreme outliers in terms of bilirubin presence or absence. Based on our preliminary analysis multiple studies related to probiotic testing were extreme outliers having bilR absent in almost all samples. These studies did not end up being included in the final dataset. No other clear outliers were found in the datasets.

```
# Plot lower abundance bins where most data points fall.
bilr_hist_panels <- ggplot(project_data, aes(x=total_reads, fill=bilr_presence)) +
  geom_histogram(binwidth = 500000) + pub_theme() + xlim(0,5.0e07) +
  xlab('Total Reads') + ylab('Number of Samples') + geom_vline(xintercept = 1000000) +
  facet_wrap(~bioproject, scales = 'free')  + theme(axis.text.x = element_text(size = 4, angle = -45))

bilr_hist_panels
```

```
## Warning: Removed 96 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 140 rows containing missing values (geom_bar).
```




## Infant data analysis

The infant metagenome data was first curated to remove samples with no age metadata and to subset the data to only include samples taken within the first year of life.

Samples were binned into age groups of 30 days and the presence and absence of bilR within these groups was summarized and plotted.

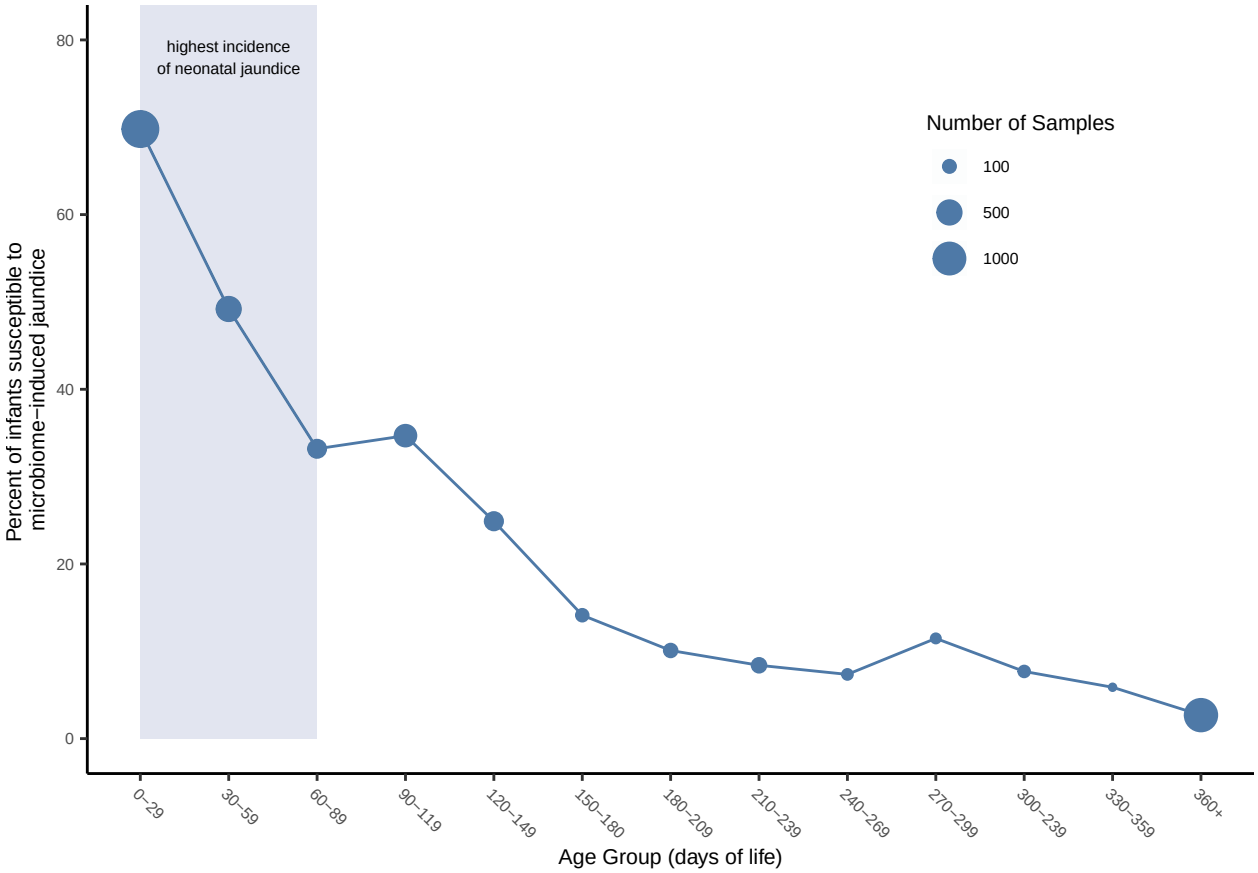
There was a clear trend observed over the first year of life where bilR was absent in a high fraction of the samples over the first few months of life but then gradually became present in nearly all of the samples by 1 year of life. This initial period of time where bilirubin reductase is often missing corresponds with a period where infants are generally succeptible to developing jaundice.

```
# Subset project data to only include samples from infant data sets in first year
infant_data <- project_data[project_data$host_disease == 'infant',]
infant_data$host_age <- as.numeric(infant_data$host_age)
```

```
## Warning: NAs introduced by coercion
```

```
infant_data <- infant_data[!is.na(infant_data$host_age), ]
infant_data <- infant_data[infant_data$host_age <= 365,]

# Bin first year infant samples into bins of 30 days
infant_data$bin <- cut(infant_data$host_age, seq(-1, max(infant_data$host_age)+30, 30))
```

```
infant_data$bin <- factor(infant_data$bin, levels = c('(-1,29]', '(29,59]', '(59,89]',
                                                      '(89,119]', '(119,149]', '(149,179]
                                                      '(179,209]', '(209,239]', '(239,269]
                                                      '(269,299]', '(299,329]', '(329,359]
                                                      '(359,389]'),
                          labels = c('0-29', '30-59', '60-89',
                                     '90-119', '120-149', '150-180',
                                     '180-209', '210-239', '240-269',
                                     '270-299', '300-239', '330-359',
                                     '360+'))


# Calculate fraction of samples with bilR present across infant age bins.
means_bilr <- as.data.frame(infant_data %>%
                            group_by(bin) %>%
                            summarise(across(bilr_tot, mean), sample_size = n()))
colnames(means_bilr) <- c('age', 'avg', 'Number of Samples')
t <- as.data.frame(table(infant_data[c('bin', 'bilr_presence')]))
t <- t[t$bilr_presence == 'present',]
colnames(t) <- c('age', 'presence', 'present_count')
means_bilr <- merge(means_bilr, t, by='age')
means_bilr$frac <- means_bilr$present_count / means_bilr$`Number of Samples`

# Plot bilR abscence (1-presence) over the first year of life.
bilR_infants <- ggplot(means_bilr, aes(x=factor(age), y=(1-frac)*100, size=`Number of Samples`, group='
  geom_line(size=0.5, color='#4E79A7') +  pub_theme() + xlab('Age Group (days of life)') +
  ylab('Percent of infants susceptible to \nmicrobiome-induced jaundice') +
  theme(axis.text.x=element_text(angle = -45, hjust = 0), legend.position = c(0.80,0.75), legend.backgr
        plot.margin= margin(0,0,0,0,'cm')) +
  ylim(0,80) + scale_color_tableau(palette = "Tableau 10", type = "regular",direction = 1) +
  geom_rect(aes(xmin='0-29', xmax='60-89', ymin=0, ymax=Inf), fill='#4E79A7', alpha=0.01) +
  annotate(geom='text', x='30-59', y=78, label='highest incidence\nof neonatal jaundice', size=2) + scal

infant_frac <- means_bilr
infant_frac <- infant_frac[c('age', 'frac')]

bilR_infants
```



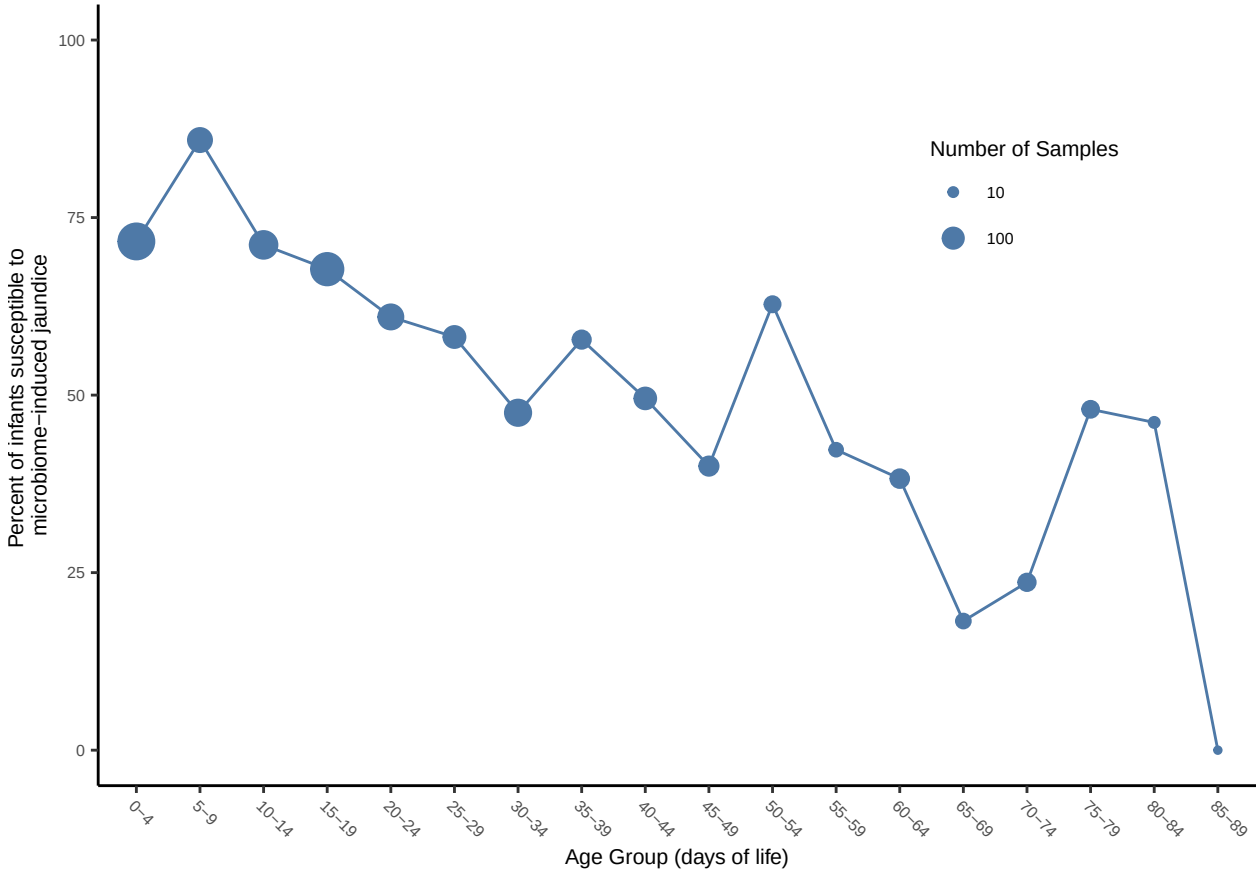

When examinig this trend at a finer time scale over the first three month of life it can be seen that the fraction of samples with bilR missing stays relatively high during this time period but may start to trend lower by the end of the third month.

```
# Subset project data to only include samples from infant data sets in first year
infant_data_3m <- project_data[project_data$host_disease == 'infant',]
infant_data_3m$host_age <- as.numeric(infant_data_3m$host_age)
```

```
## Warning: NAs introduced by coercion
```

```
infant_data_3m <- infant_data_3m[!is.na(infant_data_3m$host_age), ]
infant_data_3m <- infant_data_3m[infant_data_3m$host_age <= 89,]

# Bin first year infant samples into bins of 30 days
infant_data_3m$bin <- cut(infant_data_3m$host_age, seq(-1, max(infant_data_3m$host_age)+5, 5))

infant_data_3m$bin <- factor(infant_data_3m$bin, levels = c('(-1,4]','(4,9]','(9,14]','(14,19]','(19,24
                                labels = c('0-4','5-9','10-14','15-19','20-24','25-29','30-34','35-39'


# Plot bilR presence over time for infants
means_bilr <- as.data.frame(infant_data_3m %>%
                              group_by(bin) %>%
                              summarise(across(bilr_tot, mean), sample_size = n()))
colnames(means_bilr) <- c('age', 'avg', 'Number of Samples')
t <- as.data.frame(table(infant_data_3m[c('bin', 'bilr_presence')]))
```

```
t <- t[t$bilr_presence == 'present',]
colnames(t) <- c('age', 'presence', 'present_count')
means_bilr <- merge(means_bilr, t, by='age')
means_bilr$frac <- means_bilr$present_count / means_bilr$`Number of Samples`
means_bilr <- means_bilr[order(means_bilr$age),]

bilR_infants_3m <- ggplot(means_bilr, aes(x=age, y=(1-frac)*100, size=`Number of Samples`, group='Gene'
  geom_line(size=0.5, color='#4E79A7') +  pub_theme() + xlab('Age Group (days of life)') +
  ylab('Percent of infants susceptible to \nmicrobiome-induced jaundice') +
  theme(axis.text.x=element_text(angle = -45, hjust = 0), legend.position = c(0.80,0.75), legend.backgr
        plot.margin= margin(0,0,0,0,'cm')) + scale_x_discrete(limits=c('0-4','5-9','10-14','15-19','20-
  ylim(0,100) + scale_color_tableau(palette = "Tableau 10", type = "regular",direction = 1) + scale_siz
ggsave('S4Figure.png', bg='transparent', width=7, height=4)
bilR_infants_3m
```




## Healthy adults and IBD patients

In contrast to infants we expected bilR to be nearly universally present in healthy adults. In IBD patients there is some evidence of lower concentrations of gut urobilin, suggesting that bilirubin metabolism may be altered in these environments.

For these samples we did not include metagenomic samples that were from non-IBD inflammatory gut diseases collected in the HMP2 study.

A clear difference in the presence of bilR was seen when comparing IBD patients to healthy gut metagenomes. This difference was seen to be significant based on a test of equal proportions.

```r
# Subset data to only include samples from healthy adults, CD, or UC patients.
ibd_data <- project_data[project_data$host_disease %in% c('CD', 'UC', 'healthy'),]


bilr_disease <- as.data.frame.matrix(table(ibd_data[c('host_disease', 'bilr_presence')]))
bilr_disease$host_disease <- row.names(bilr_disease)
bilr_disease$host_disease <- factor(bilr_disease$host_disease, levels=c('CD', 'UC', 'healthy'), labels=c
bilr_disease$total <- bilr_disease$absent + bilr_disease$present
bilr_disease$absent_frac <- bilr_disease$absent / bilr_disease$total

project_sub <-  ibd_data[ibd_data$host_disease %in% c('CD', 'UC', 'healthy'),]
project_sub$host_disease <- factor(project_sub$host_disease, levels=c('CD', 'UC', 'healthy'))
prop.test(table(project_sub$host_disease, project_sub$bilr_presence), correct=TRUE)
```

```
##
##  3-sample test for equality of proportions without continuity
##  correction
##
## data:  table(project_sub$host_disease, project_sub$bilr_presence)
## X-squared = 471.1, df = 2, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##      prop 1      prop 2      prop 3
## 0.318627451 0.331768388 0.001689189
```

```r
project_sub_sub <- ibd_data[ibd_data$host_disease %in% c('CD', 'UC'),]
project_sub_sub$host_disease <- factor(project_sub_sub$host_disease, levels=c('CD', 'UC'))
prop.test(table(project_sub_sub$host_disease, project_sub_sub$bilr_presence), correct=TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(project_sub_sub$host_disease, project_sub_sub$bilr_presence)
## X-squared = 0.27411, df = 1, p-value = 0.6006
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05921114  0.03292927
## sample estimates:
##    prop 1    prop 2
## 0.3186275 0.3317684
```

```r
project_sub_sub <- ibd_data[ibd_data$host_disease %in% c('CD', 'healthy'),]
project_sub_sub$host_disease <- factor(project_sub_sub$host_disease, levels=c('CD', 'healthy'))
prop.test(table(project_sub_sub$host_disease, project_sub_sub$bilr_presence), correct=TRUE)
```

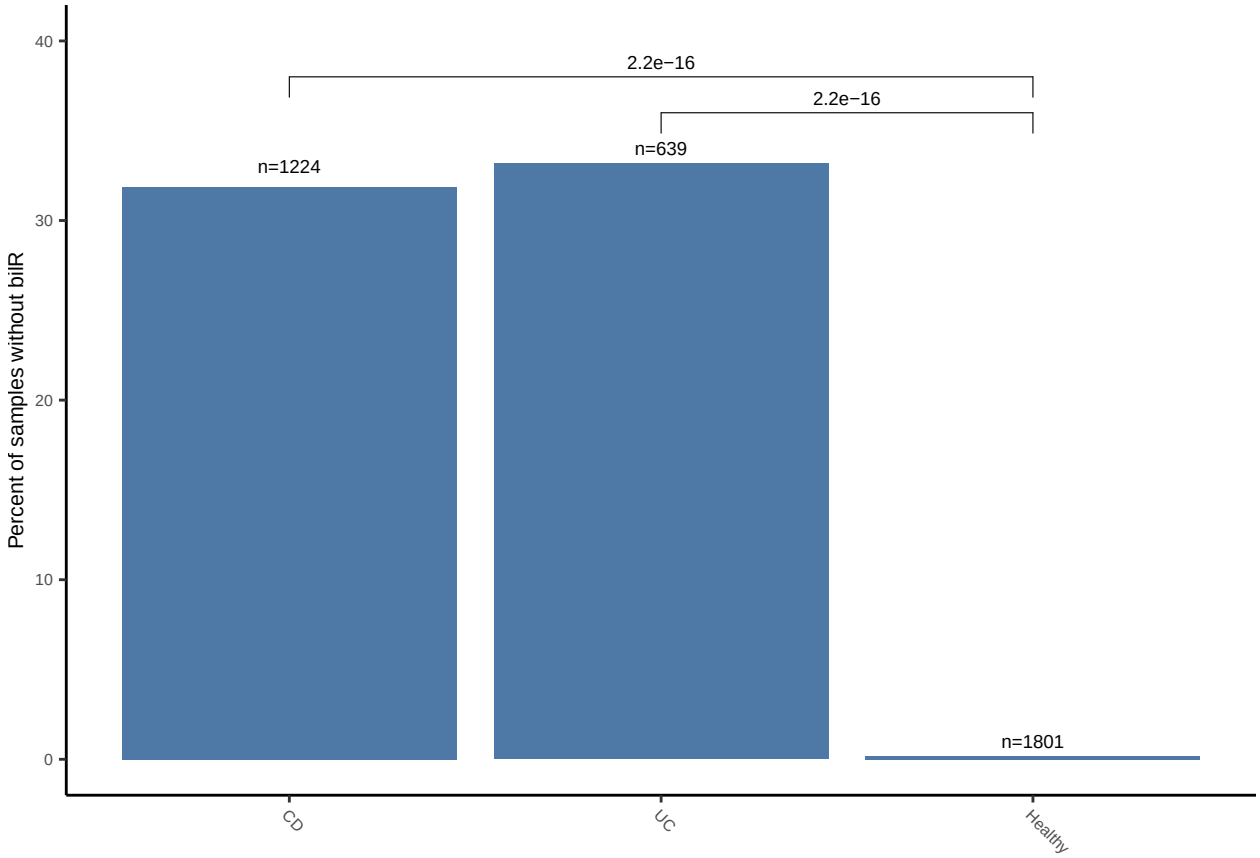
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(project_sub_sub$host_disease, project_sub_sub$bilr_presence)
## X-squared = 441.25, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
```

```
##  0.2898998 0.3439767
## sample estimates:
##      prop 1      prop 2
## 0.318627451 0.001689189
```

```
project_sub_sub <- ibd_data[ibd_data$host_disease %in% c('UC', 'healthy'),]
project_sub_sub$host_disease <- factor(project_sub_sub$host_disease, levels=c('UC', 'healthy'))
prop.test(table(project_sub_sub$host_disease, project_sub_sub$bilr_presence), correct=TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(project_sub_sub$host_disease, project_sub_sub$bilr_presence)
## X-squared = 433.24, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2922924 0.3678660
## sample estimates:
##      prop 1      prop 2
## 0.331768388 0.001689189
```

```
healthy_ibd <- ggplot(bilr_disease, aes(x=host_disease, y=absent_frac*100), ) +
  geom_bar(stat = 'identity', position='stack', fill='#4E79A7') + ylim(0,40) +
  theme_bw() +
  ylab('Percent of samples without bilR') + pub_theme() + xlab(element_blank()) +
  theme(axis.text.x=element_text(angle = -45, hjust = 0), legend.position='none;') + scale_fill_tableau
  geom_bracket(xmin=c('CD', 'UC'), xmax=c('Healthy', 'Healthy'), label=c('2.2e-16', '2.2e-16'), y.posit
  annotate(geom='text', x=c('CD', 'UC', 'Healthy'), y=c(33, 34, 1),label=c('n=1224', 'n=639', 'n=1801')

disease_frac <- bilr_disease[c('absent_frac')]


healthy_ibd
```

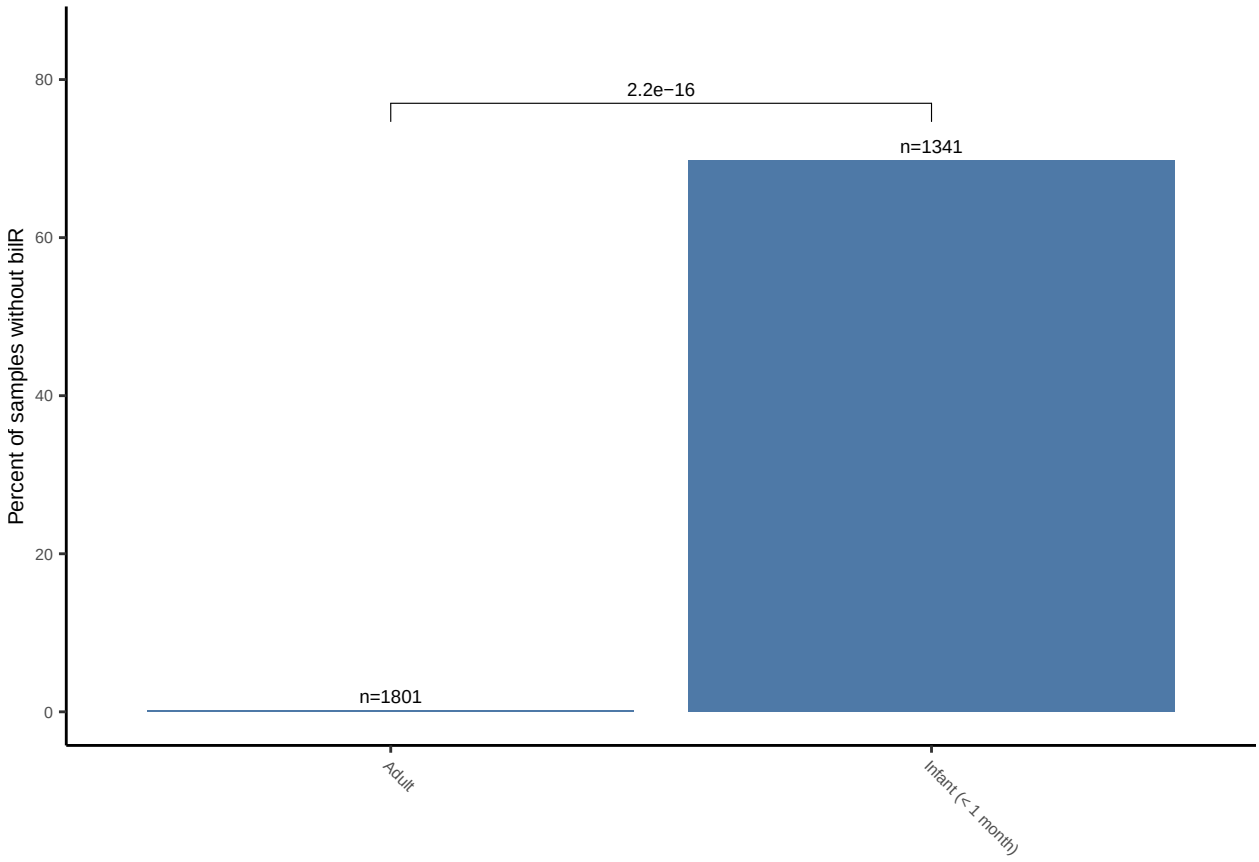



When comparing the ealry life infants to healthy adult the difference was also evident, with infants having a significantly higher fraction of samples missing bilR than the healthy adults.

```
colnames(infant_frac) <- c('group', 'absent_frac')
infant_frac$absent_frac <- 1-infant_frac$absent_frac
disease_frac$group <- rownames(disease_frac)
healthy_infant <- rbind(infant_frac, disease_frac)
healthy_infant <- healthy_infant[healthy_infant$group %in% c('healthy', '0-29'),]
healthy_infant$group <- factor(healthy_infant$group, levels=c('healthy', '0-29'),
                               labels=c('Adult', 'Infant (< 1 month)'))

healthy_infant_dat <- rbind(infant_data[infant_data$host_age <= 30,][c('host_disease', 'bilr_presence')]
healthy_infant_dat$host_disease <- factor(healthy_infant_dat$host_disease, levels=c('infant', 'healthy'
prop.test(table(healthy_infant_dat$host_disease, healthy_infant_dat$bilr_presence), correct=TRUE)
```

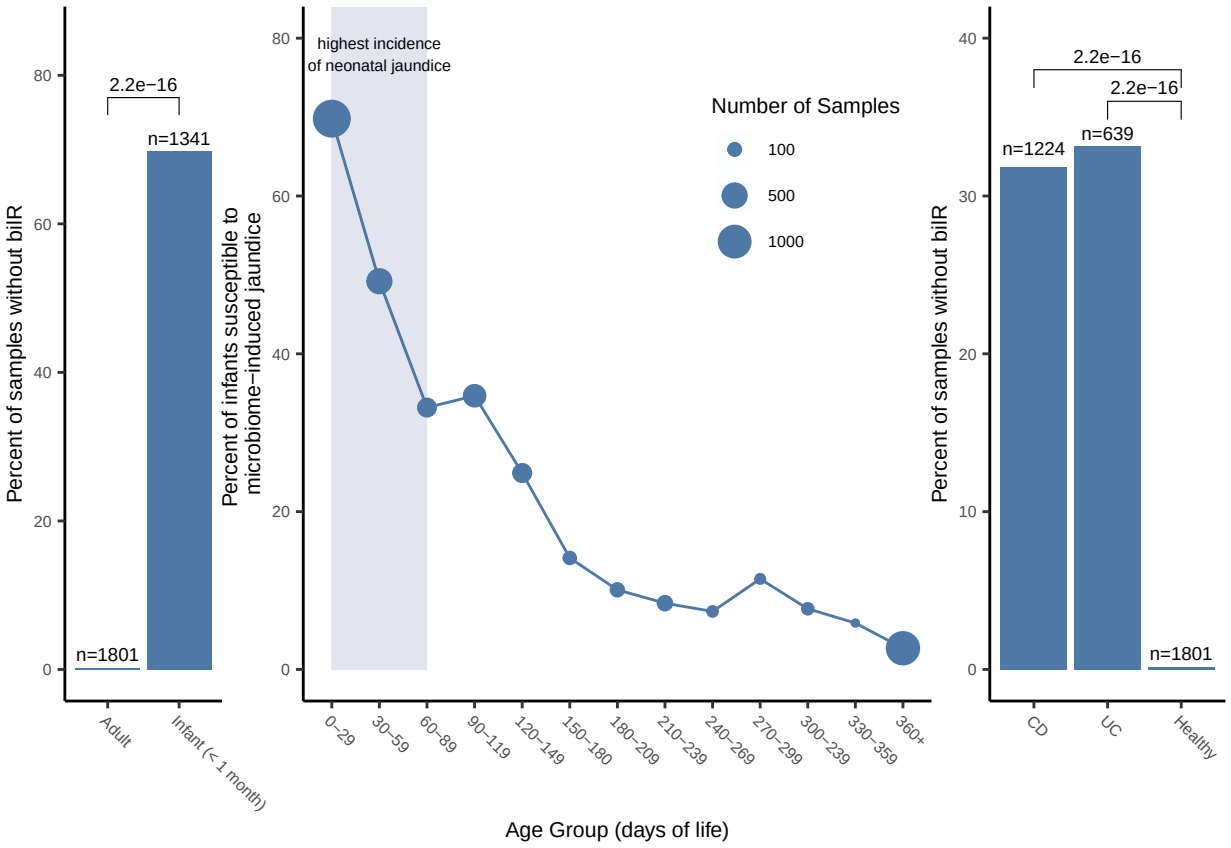
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(healthy_infant_dat$host_disease, healthy_infant_dat$bilr_presence)
## X-squared = 1262.7, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.6462998 0.6957763
## sample estimates:
##      prop 1      prop 2
## 0.672727273 0.001689189
```

```
health_infant <- ggplot(healthy_infant, aes(x=group, y=absent_frac*100), ) +
  geom_bar(stat = 'identity', position='stack', fill='#4E79A7') + ylim(0,85) +
  theme_bw() +
  ylab('Percent of samples without bilR') + pub_theme() + xlab(element_blank()) +
  theme(axis.text.x=element_text(angle = -45, hjust = 0), legend.position='none', plot.margin= margin(0
  geom_bracket(xmin=c('Infant (< 1 month)'), xmax=c('Adult'), label=c('2.2e-16'), y.position = c(77), si
  annotate(geom='text', x=c('Infant (< 1 month)', 'Adult'), y=c(71.5, 2),label=c('n=1341','n=1801') , si
health_infant
```




When put together these data paint a picture of bilirubin reductase being present in nearly all healthy adults, but often being absent in ealry life infants, as the gut micorbiome is first developing, and during inflammatory gut disorders, which are known to cause significant microbial dysbiosis in the gut.

```
health_infant + bilR_infants + healthy_ibd + plot_layout(widths=c(1,4,1.5))
```




```
ggsave('Figure6.svg', bg='transparent', width=7, height=4)
```