Genome Richness: Healthy Dairy Worker study

Amy Willis

2022-09-08

Our goal to estimate species diversity based on shotgun metagenomics data, and compare diversity between community controls and dairy workers.

Preliminaries

Let's first read in the dataset:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(magrittr)
```

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

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

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

```
library(breakaway)
dataframe <- readRDS("HDW_SCG_presence.RDS") %>% as_tibble
dataframe
```

```
## # A tibble: 20,025 x 13
##    sample   gene_name    perce~1 t_dom~2 t_phy~3 t_class t_order t_fam~4 t_genus
##    <chr>    <chr>          <dbl> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1 C02_S104 Ribosomal_L1   100   Bacter~ Bacter~ Bacter~ Bacter~ Bacter~ Prevot~
##  2 C02_S104 Ribosomal_L1    99.6 Bacter~ Firmic~ Clostr~ Oscill~ Oscill~ ER4
##  3 C02_S104 Ribosomal_L1    99.1 Bacter~ Actino~ Coriob~ Coriob~ Coriob~ Collin~
##  4 C02_S104 Ribosomal_L1    94.7 Bacter~ Firmic~ Clostr~ Lachno~ Lachno~ UBA3282
##  5 C02_S104 Ribosomal_L1    99.6 Bacter~ Firmic~ Clostr~ Lachno~ Lachno~ Eubact~
##  6 C02_S104 Ribosomal_L1   100   Bacter~ Bacter~ Bacter~ Bacter~ Bacter~ Prevot~
##  7 C02_S104 Ribosomal_L1    92   Bacter~ Firmic~ Clostr~ 4C28d-~ CAG-727 UBA102~
##  8 C02_S104 Ribosomal_L1    99   Bacter~ Firmic~ Bacilli RF39    CAG-10~ CAG-460
##  9 C02_S104 Ribosomal_L1   100   Bacter~ Firmic~ Negati~ Acidam~ Acidam~ Phasco~
## 10 C02_S104 Ribosomal_L1    99.6 Bacter~ Firmic~ Clostr~ Lachno~ Lachno~ CAG-127
## # ... with 20,015 more rows, 4 more variables: t_species <chr>, present <dbl>,
## #   group <chr>, dairy <dbl>, and abbreviated variable names
## #   1: percent_identity, 2: t_domain, 3: t_phylum, 4: t_family
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

The column present is confusing so let's remove it

```
dataframe %<>%
  filter(present == 1) %>%
  select(-present)
```

Methodology

```
frequency_counts <- dataframe %>%
  mutate(taxon = paste(t_domain, t_phylum, t_class, t_order, t_family, t_genus,
                       t_species)) %>%
  select(-c("t_domain", "t_phylum", "t_class", "t_order", "t_family", "t_genus",
            "t_species")) %>%
  select(-percent_identity) %>%
  distinct %>% # multiple rows correspond to different strains, possibly. Look only at species level
  group_by(taxon, sample) %>%
  summarise(n = n()) %>%
  ungroup
```

```
## `summarise()` has grouped output by 'taxon'. You can override using the
## `.groups` argument.
```

```
frequency_counts
```

```
## # A tibble: 3,023 x 3
##    taxon                                                            sample     n
##    <chr>                                                            <chr>  <int>
##  1 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ C02_S~     3
##  2 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ C02_S~    11
##  3 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ C02_S~    11
##  4 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ C02_S~    11
##  5 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ C02_S~    11
##  6 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ D03_S~    10
##  7 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ D03_S~    11
##  8 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ D03_S~    10
##  9 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ C02_S~     9
## 10 Archaea Euryarchaeota Methanobacteria Methanobacteriales Methan~ C02_S~     1
## # ... with 3,013 more rows
## # i Use `print(n = ...)` to see more rows
```

So we see that *Methanobrevibacter smithii* (the first row) was observed from 3 distinct ribosomal genes in sample C02_S104. Let's double check that:

```r
dataframe %>%
  filter(sample == "C02_S104", t_species == "Methanobrevibacter smithii") %>%
  select(gene_name, percent_identity, t_species)
```

```
## # A tibble: 3 x 3
##   gene_name     percent_identity t_species
##   <chr>                    <dbl> <chr>
## 1 Ribosomal_L1              93.9 Methanobrevibacter smithii
## 2 Ribosomal_L13             97.9 Methanobrevibacter smithii
## 3 Ribosomal_S9             100   Methanobrevibacter smithii
```

Ok, let's run `breakaway`

```r
bas <- frequency_counts %>%
  split(.$sample) %>%
  map(function(df) pull(.data=df, var=n)) %>%
  map(~table(.)) %>%
  map(~as_tibble(.)) %>%
  map(function(df) rename(.data=df, count = 1, f = 2)) %>%
  map(function(df) mutate(.data=df, count = as.numeric(count))) %>%
  map(~breakaway(.))
bt <- bas %>%
  alpha_estimates %>%
  summary %>%
  inner_join(dataframe %>%
               select(sample, dairy) %>%
               distinct, c("sample_names" = "sample")) %>%
  betta(formula = estimate ~ dairy, ses = error, data = .)
```

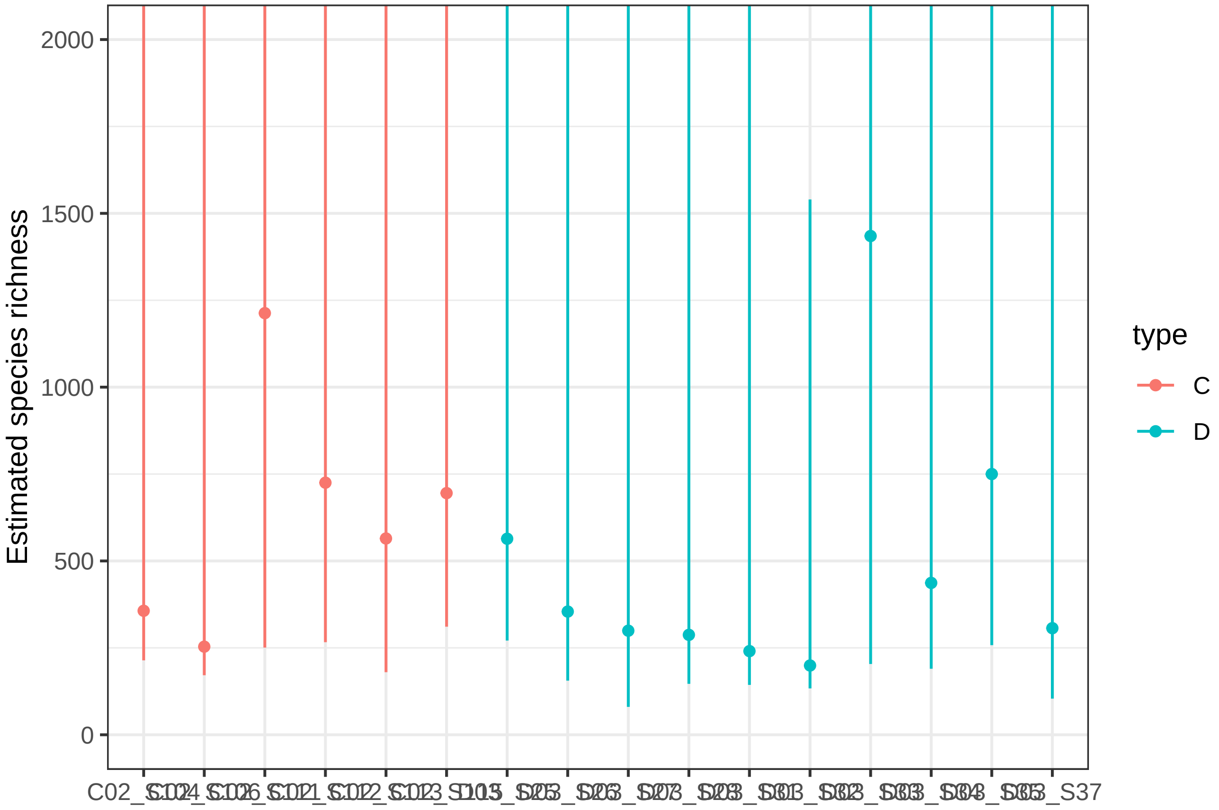
Results

```r
bt$table
```

```
##             Estimates Standard Errors p-values
## (Intercept) 356.49442        43.48546    0.000
## dairy       -54.87796        54.31164    0.312
```

On average, we estimate that there are 55 fewer species for in dairy workers' gut metagenomes compared to gut metagenomes of our community control population ($p = 0.31$).

```r
bas %>%
  alpha_estimates %>%
  summary %>%
  mutate(type = str_sub(sample_names, 1, 1)) %>%
  ggplot(aes(x = sample_names, y = estimate, col = type)) +
  geom_point() +
  geom_segment(aes(y = lower, yend = upper, xend = sample_names)) +
  theme_bw() +
  ylab("Estimated species richness") +
  xlab("") +
  coord_cartesian(ylim=c(0, 2000))
```



It looks like nothing crazy is driving this result.

Lets check out sample species richness

```
frequency_counts %>%
  split(.$sample) %>%
  map(function(df) pull(.data=df, var=n)) %>%
  map(~table(.)) %>%
  map(~as_tibble(.)) %>%
  map(function(df) rename(.data=df, count = 1, f = 2)) %>%
  map(function(df) mutate(.data=df, count = as.numeric(count))) %>%
  map(~sample_richness(.)) %>%
  alpha_estimates %>%
  summary %>%
  inner_join(dataframe %>%
               select(sample, dairy) %>%
               distinct, c("sample_names" = "sample")) %>%
  lm(estimate ~ dairy, data = .) %>%
  summary
```

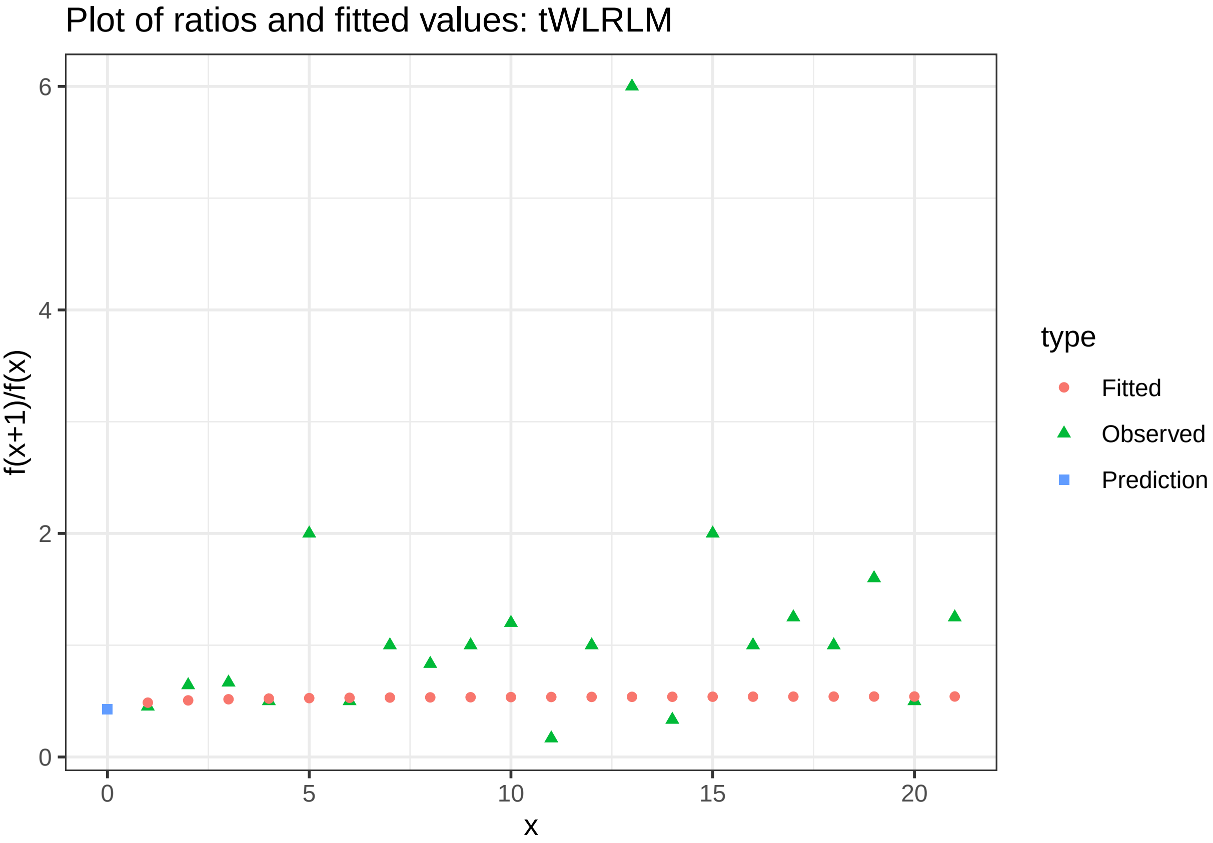
```
##
## Call:
## lm(formula = estimate ~ dairy, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -86.40 -39.34 -14.78  34.02 103.60
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    228.17      23.74    9.61 1.53e-07 ***
## dairy          -62.77      30.03   -2.09   0.0553 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 58.16 on 14 degrees of freedom
## Multiple R-squared:  0.2378, Adjusted R-squared:  0.1834 
## F-statistic: 4.368 on 1 and 14 DF,  p-value: 0.05534
```

So we don't get a very different result, but if we ignored uncertainty in estimating richness we'd probably conclude a significant difference between the groups.

Let's just do some diagnostics

```
frequency_counts %>%
  split(.$sample) %>%
  map(function(df) pull(.data=df, var=n)) %>%
  map(~table(.)) %>%
  map(~as_tibble(.)) %>%
  map(function(df) rename(.data=df, count = 1, f = 2)) %>%
  map(function(df) mutate(.data=df, count = as.numeric(count))) %>%
  nth(1) %>%
  breakaway %>%
  plot
```



Yep looks reasonable!