Regression for Microbiome Data: Moving From Diversity to Inference

Brendan J. Kelly, MD, MS

Updated: 18 June 2020

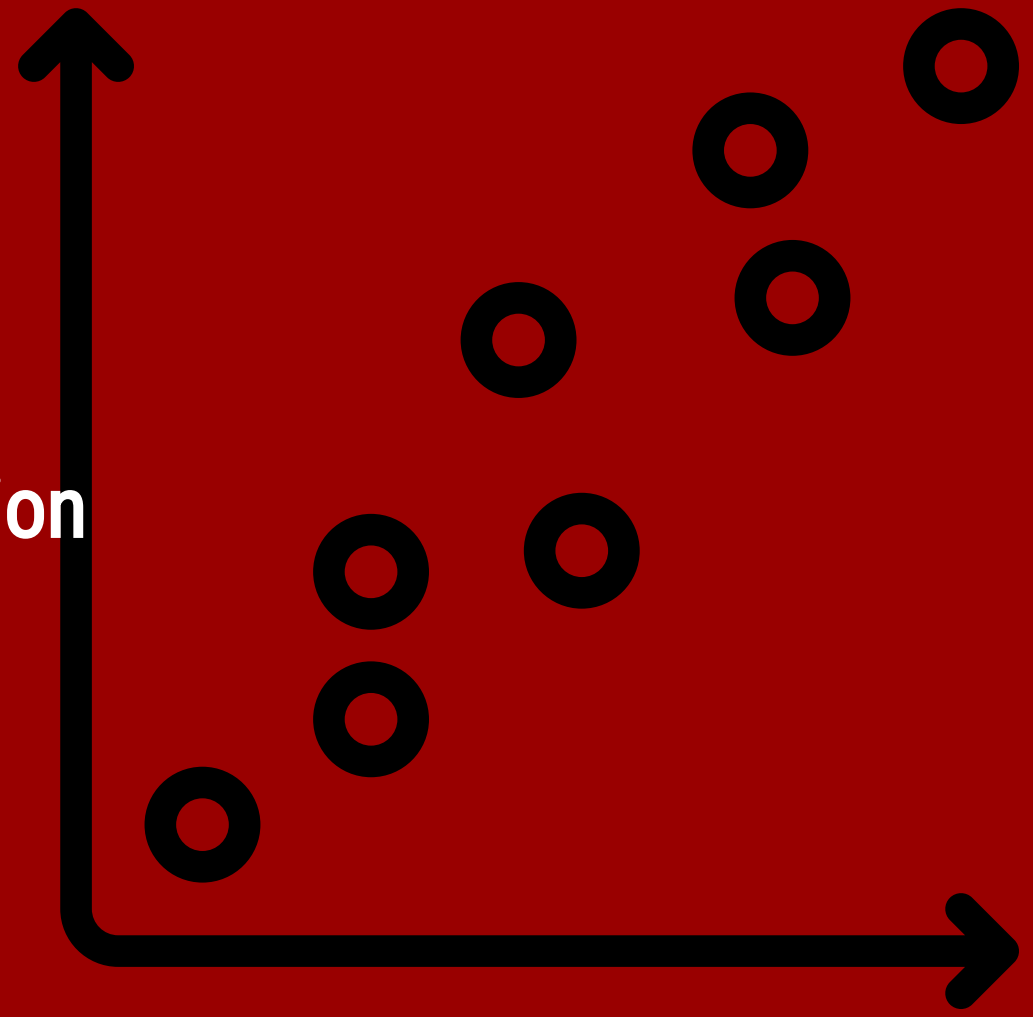
Review α-diversity

Review β-diversity

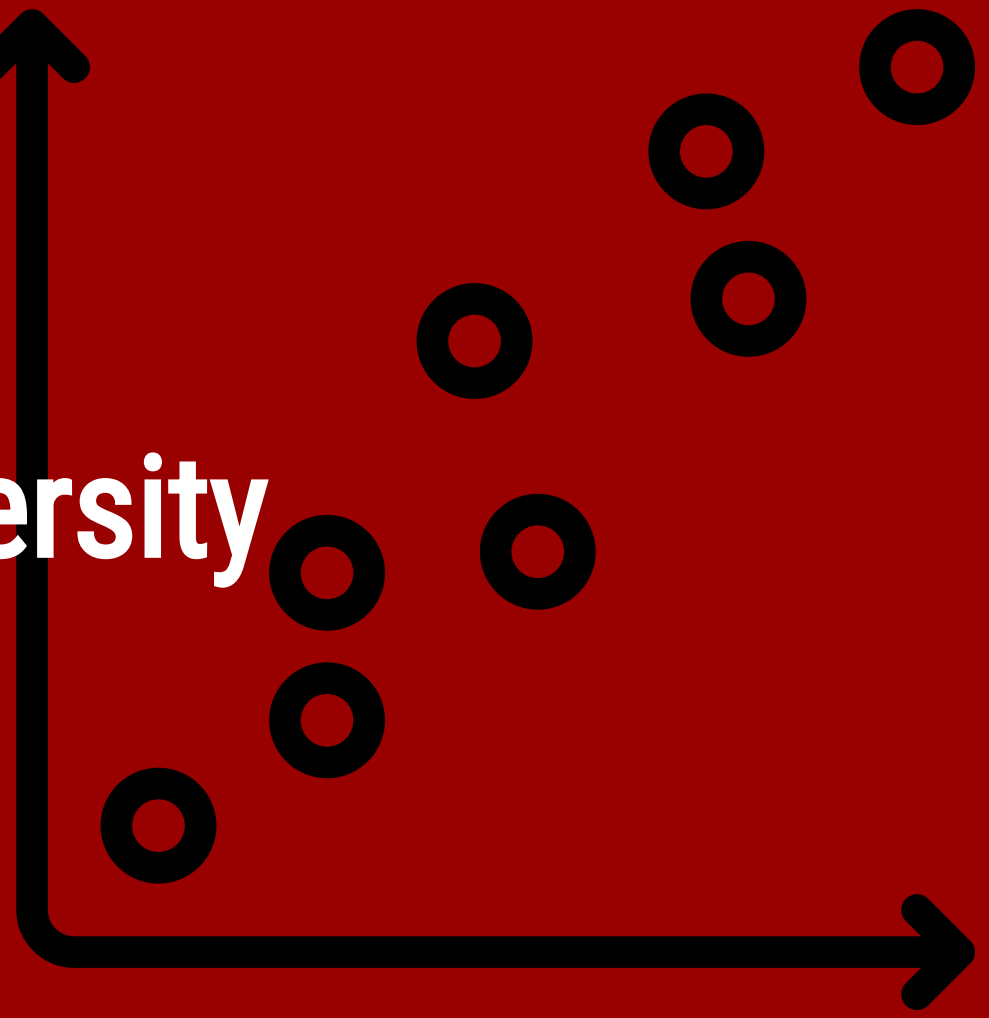
α/β-diversity → linear regression

β-diversity → PERMANOVA

Compositional data?



α-diversity

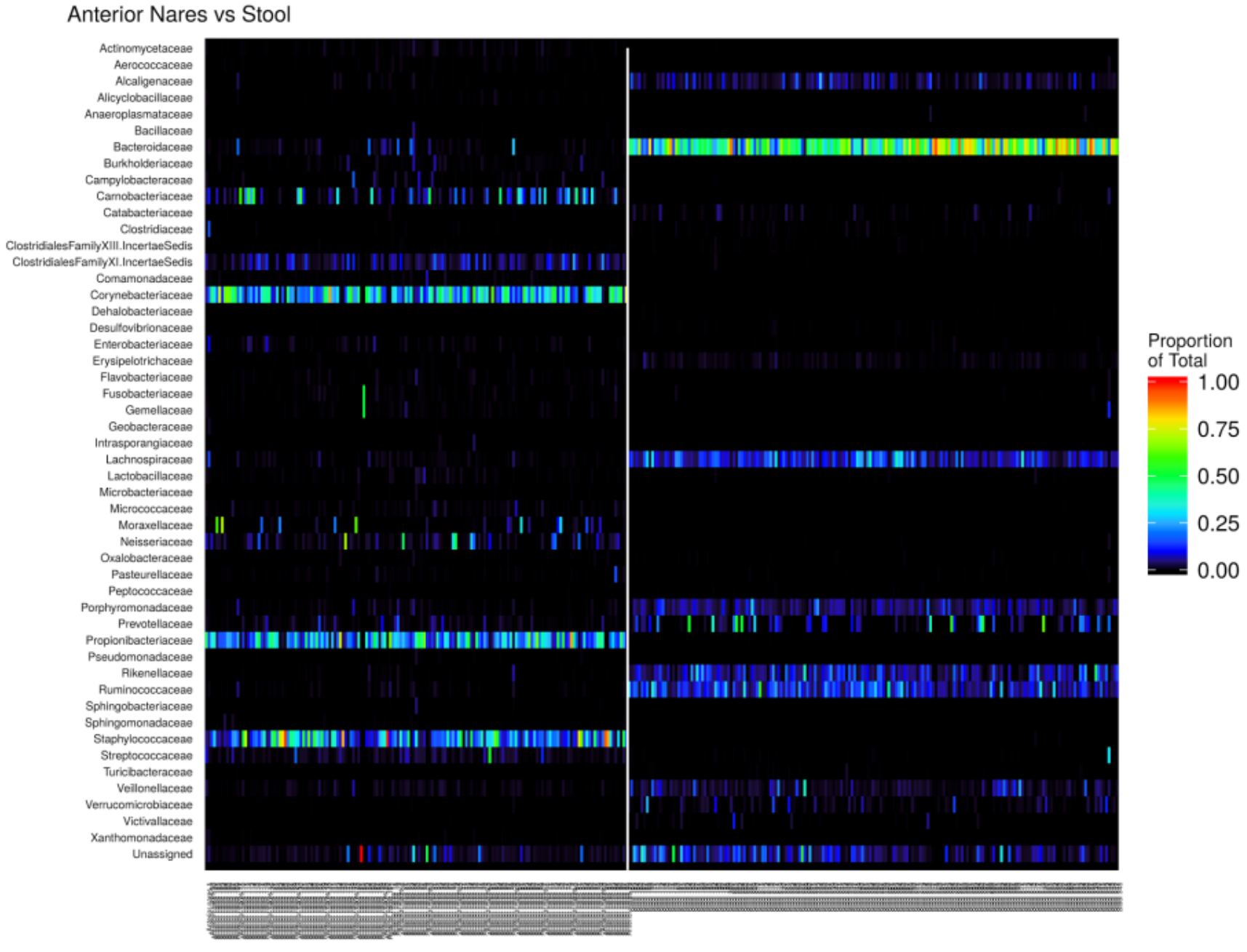


High Dimensional Microbiome Data

```
##              700013549 700014386 700014403 700014409 700014412 700014415
## OTU_97.1             0         0         0         0         0         0
## OTU_97.10            0         0         6         4         1         5
## OTU_97.100           0         0       133         7         1         4
## OTU_97.1000          0         0         0         0         0         0
## OTU_97.10000         0         0         0         0         0         0
## OTU_97.10001         0         0         0         0         0         1
## OTU_97.10002         0         0         0         0         0         0
## OTU_97.10003         0         0         0         0         0         0
## OTU_97.10004         0         0         0         0         0         0
## OTU_97.10005         0         0         0         0         0         0
## OTU_97.10006         0         0         0         0         0         0
## OTU_97.10007         0         0         0         0         0         0
## OTU_97.10008         0         1         0         0         0         0
## OTU_97.10009         0         0         1         0         0         0
## OTU_97.1001          0         0         0         0         0         0
## OTU_97.10010         0         0         0         0         0         0
```

High Dimensional Microbiome Data

- How to deal with high-dimensional microbiome data?
- **Descriptive (e.g., heatmaps and stacked barplots)**
- Test a priori hypotheses regarding specific OTUs/taxa
- Reduce dimensions:
  - single summary statistic (alpha diversity)
  - pairwise distances (beta diversity) with PCoA or PERMANOVA
  - community types (mixture modeling)


Anterior Nares vs Stool
Actinomycetaceae
Aerococcaceae
Alcaligenaceae
Alicyclobacillaceae
Anaeroplasmataceae
Bacillaceae
Bacteroidaceae
Burkholderiaceae
Campylobacteraceae
Carnobacteriaceae
Catabacteriaceae
Clostridiaceae
ClostridialesFamilyXIII.IncertaeSedis
ClostridialesFamilyXI.IncertaeSedis
Comamonadaceae
Corynebacteriaceae
Dehalobacteriaceae
Desulfovibrionaceae
Enterobacteriaceae
Erysipelotrichaceae
Flavobacteriaceae
Fusobacteriaceae
Gemellaceae
Geobacteraceae
Intrasporangiaceae
Lachnospiraceae
Lactobacillaceae
Microbacteriaceae
Micrococcaceae
Moraxellaceae
Neisseriaceae
Oxalobacteraceae
Pasteurellaceae
Peptococcaceae
Porphyromonadaceae
Prevotellaceae
Propionibacteriaceae
Pseudomonadaceae
Rikenellaceae
Ruminococcaceae
Sphingobacteriaceae
Sphingomonadaceae
Staphylococcaceae
Streptococcaceae
Turicibacteraceae
Veillonellaceae
Verrucomicrobiaceae
Victivallaceae
Xanthomonadaceae
Unassigned
Proportion of Total
1.00
0.75
0.50
0.25
0.00

High Dimensional Microbiome Data

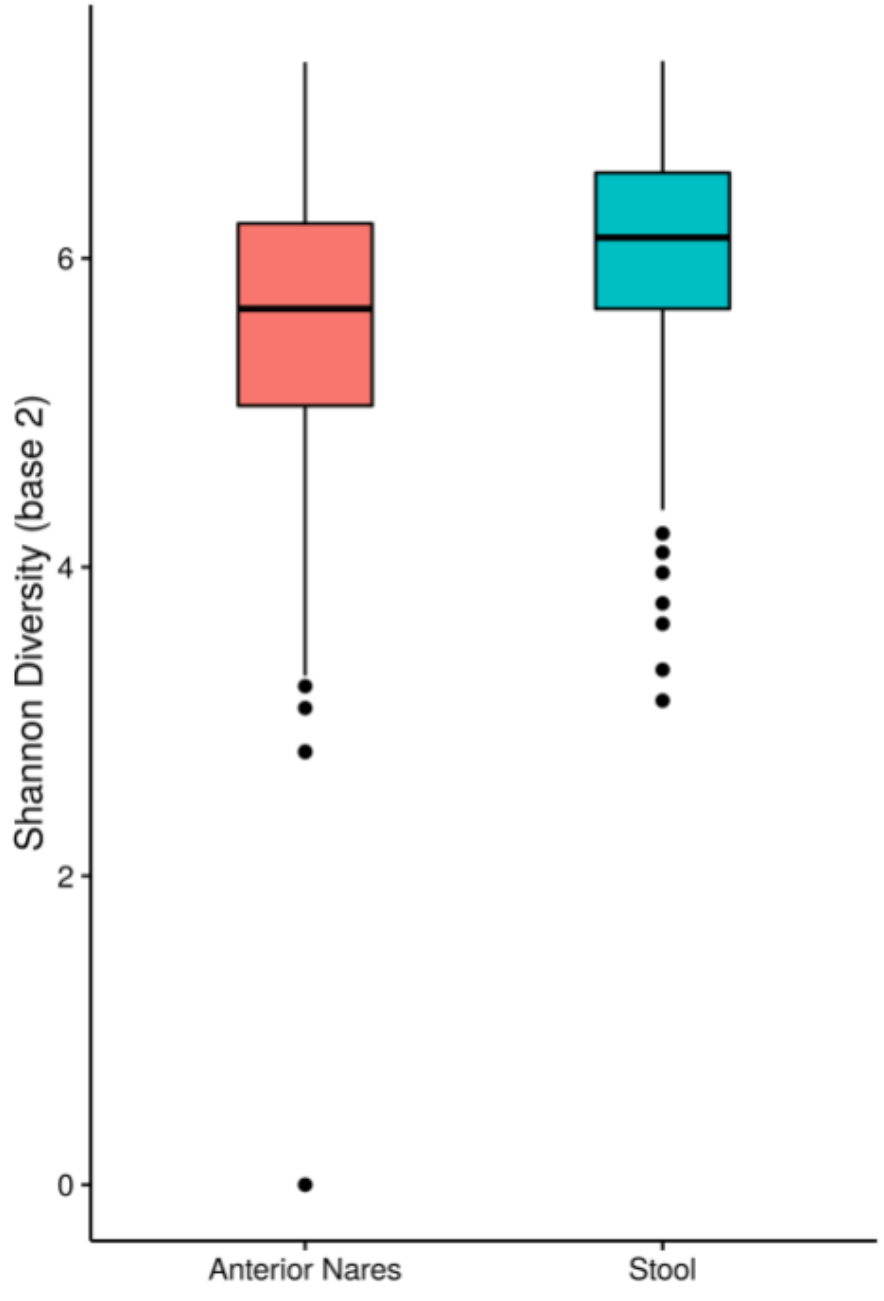
- How to deal with high-dimensional microbiome data?
- Descriptive (e.g., heatmaps and stacked barplots)
- Test a priori hypotheses regarding specific OTUs/taxa
- **Reduce dimensions:**
  - **single summary statistic (alpha diversity)**
  - pairwise distances (beta diversity) with PCoA or PERMANOVA
  - community types (mixture modeling)

Shannon Diversity

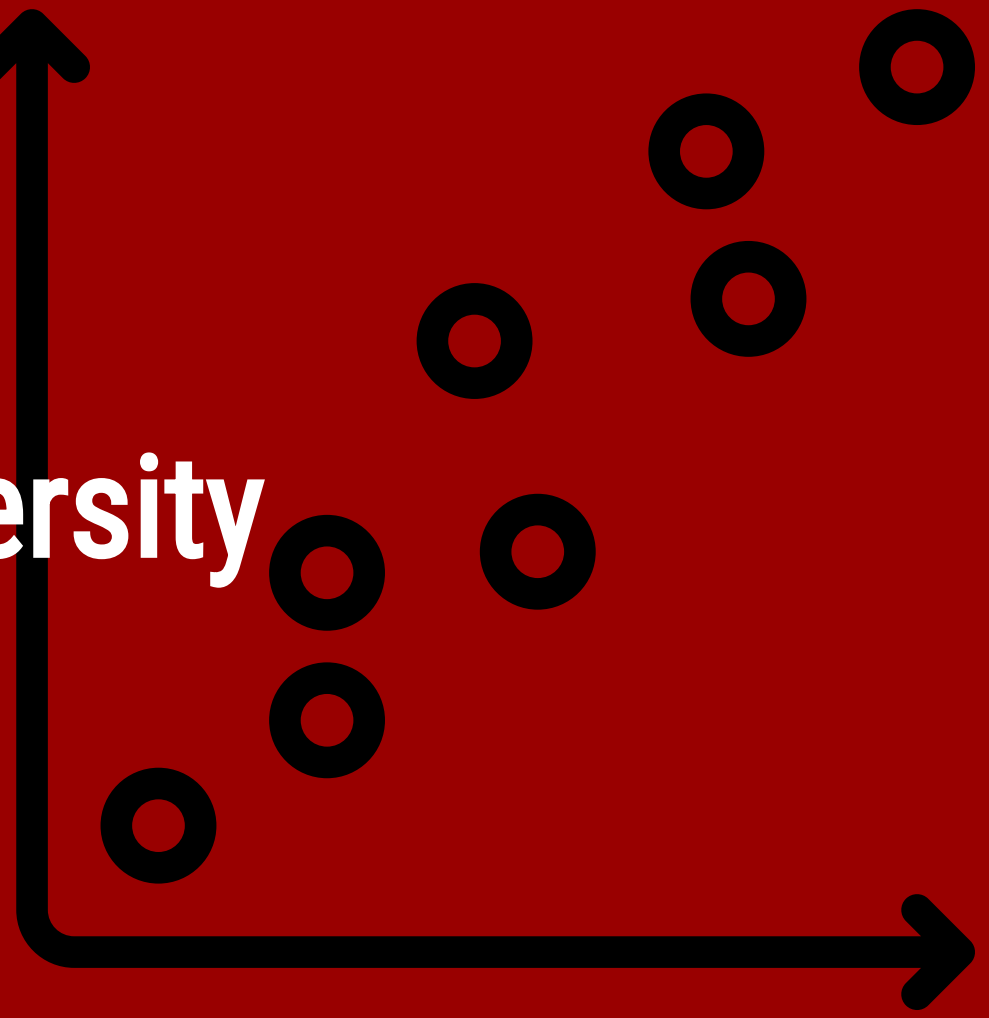
- Richness & evenness
- Shannon diversity:

  $$H' = - \sum{ p_{i} * \log_{b}{(p_{i})} }$$

- "The uncertainty contained in a probability distribution is the average log-probability of an event." (McElreath *Statistical Rethinking, 2nd* 2020)


Shannon Diversity (base 2)
0
2
4
6
Anterior Nares
Stool

β-diversity

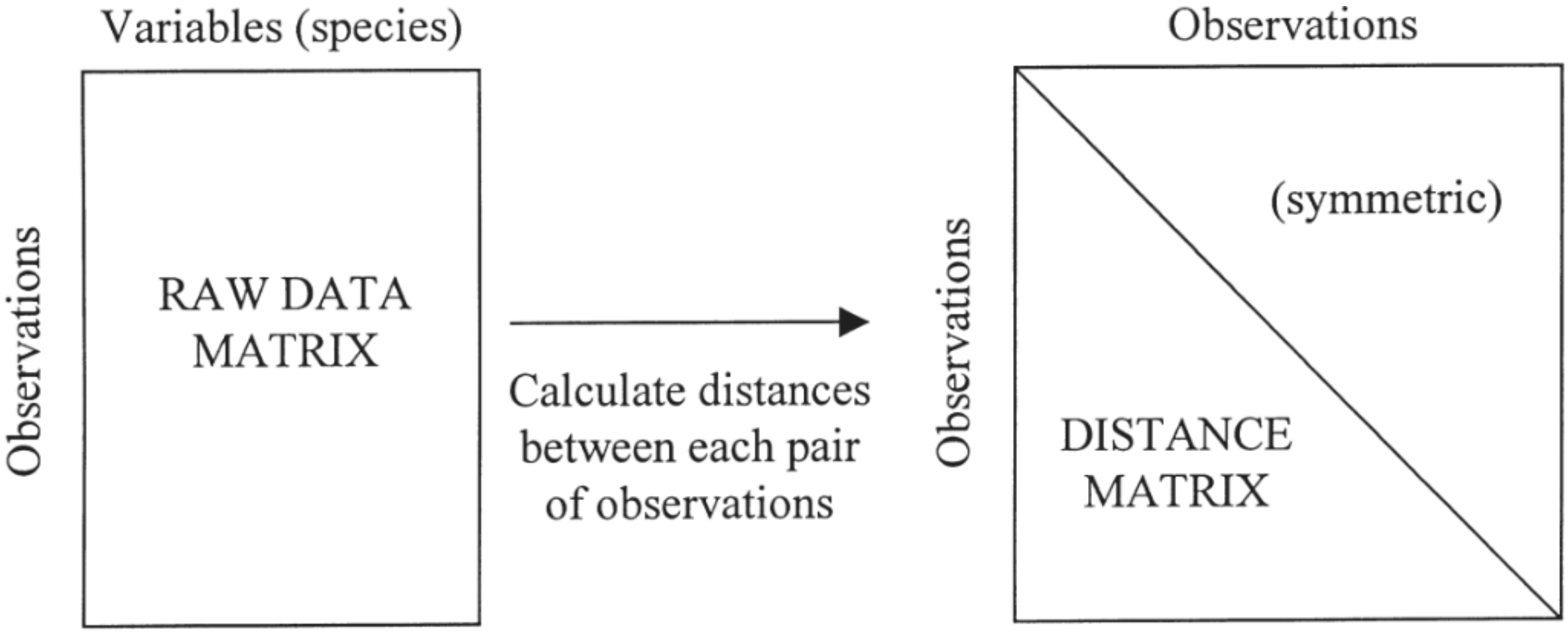


High Dimensional Microbiome Data

- How to deal with high-dimensional microbiome data?
- Descriptive (e.g., heatmaps and stacked barplots)
- Test a priori hypotheses regarding specific OTUs/taxa
- **Reduce dimensions:**
  - single summary statistic (alpha diversity)
  - **pairwise distances (beta diversity) with PCoA or PERMANOVA**
  - community types (mixture modeling)

Beta Diversity as Dimension Reduction

- Summarize each sample's relationship to other samples:
  - pairwise distances
  - OTU table → square matrix
- Many beta diversity metrics:
  - just counts versus counts + phylogeny
  - weighted versus unweighted

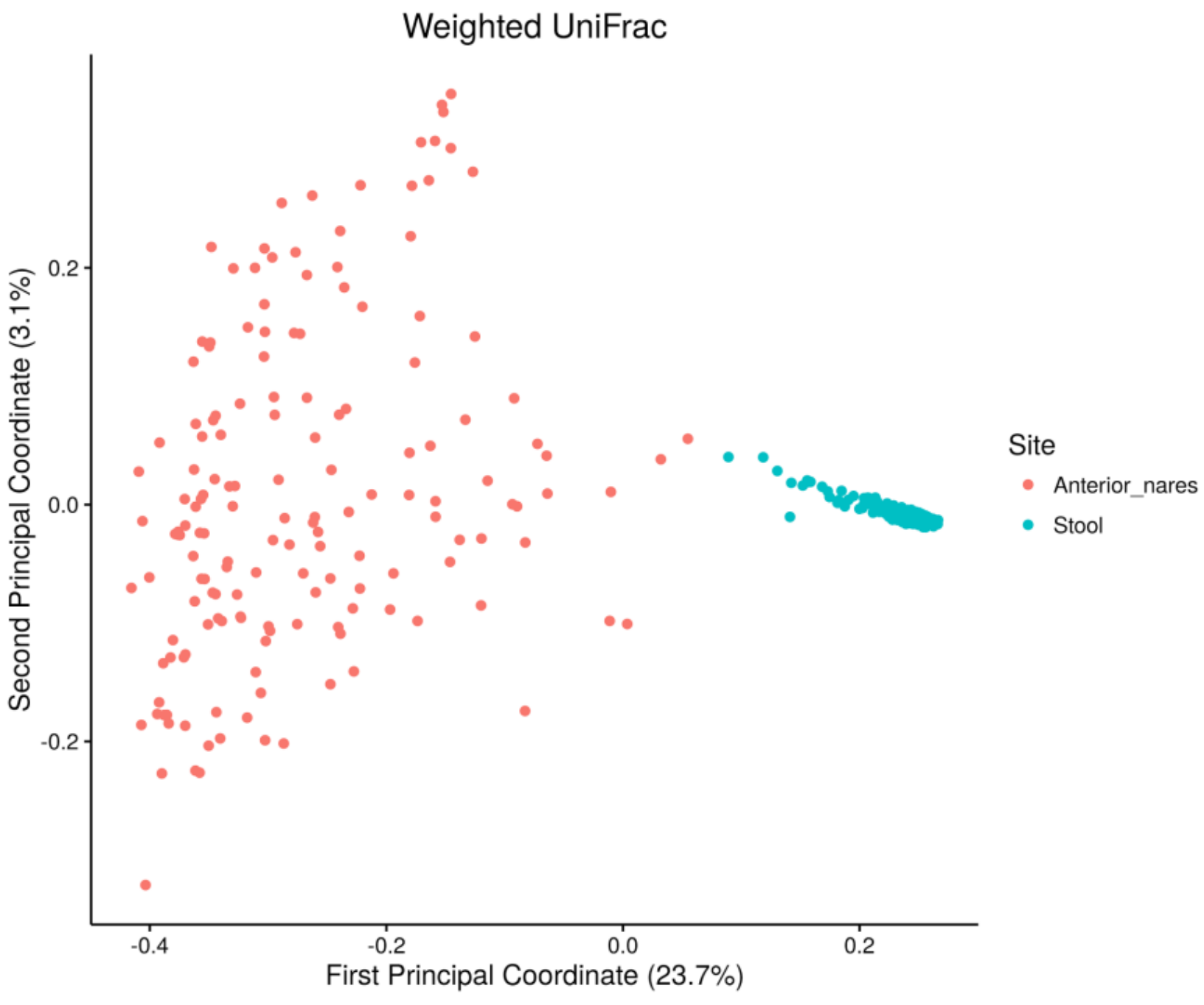

Variables (species)
Observations
RAW DATA MATRIX
Calculate distances between each pair of observations
Observations
Observations
(symmetric)
DISTANCE MATRIX

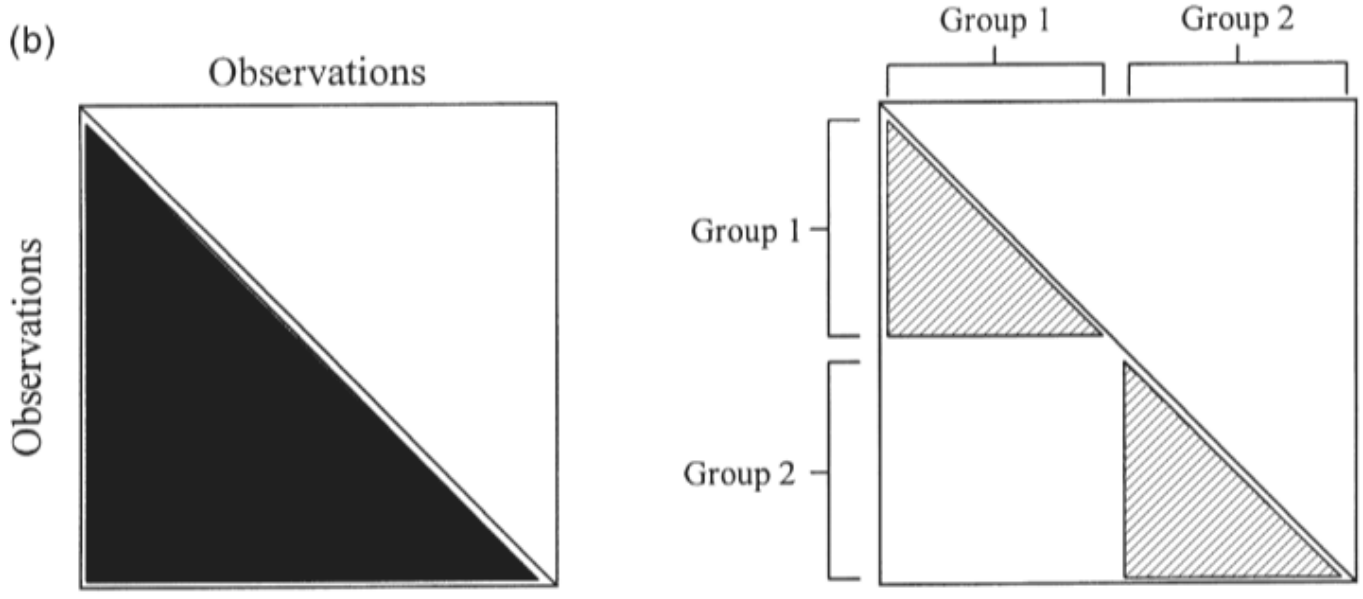
Distance Metrics for Beta Diversity

- Just counts versus counts + phylogeny:
  - Jaccard: $J(A,B) = \frac{A \cap B}{A \cup B}$ & $d_{J}(A,B) = 1 - J(A,B)$
  - UniFrac: fraction of unique branch length in tree
- Weighted versus unweighted:
  - weighted: counts matter
  - unweighted: binary (presence-absence)

Pairwise Distances → PCoA

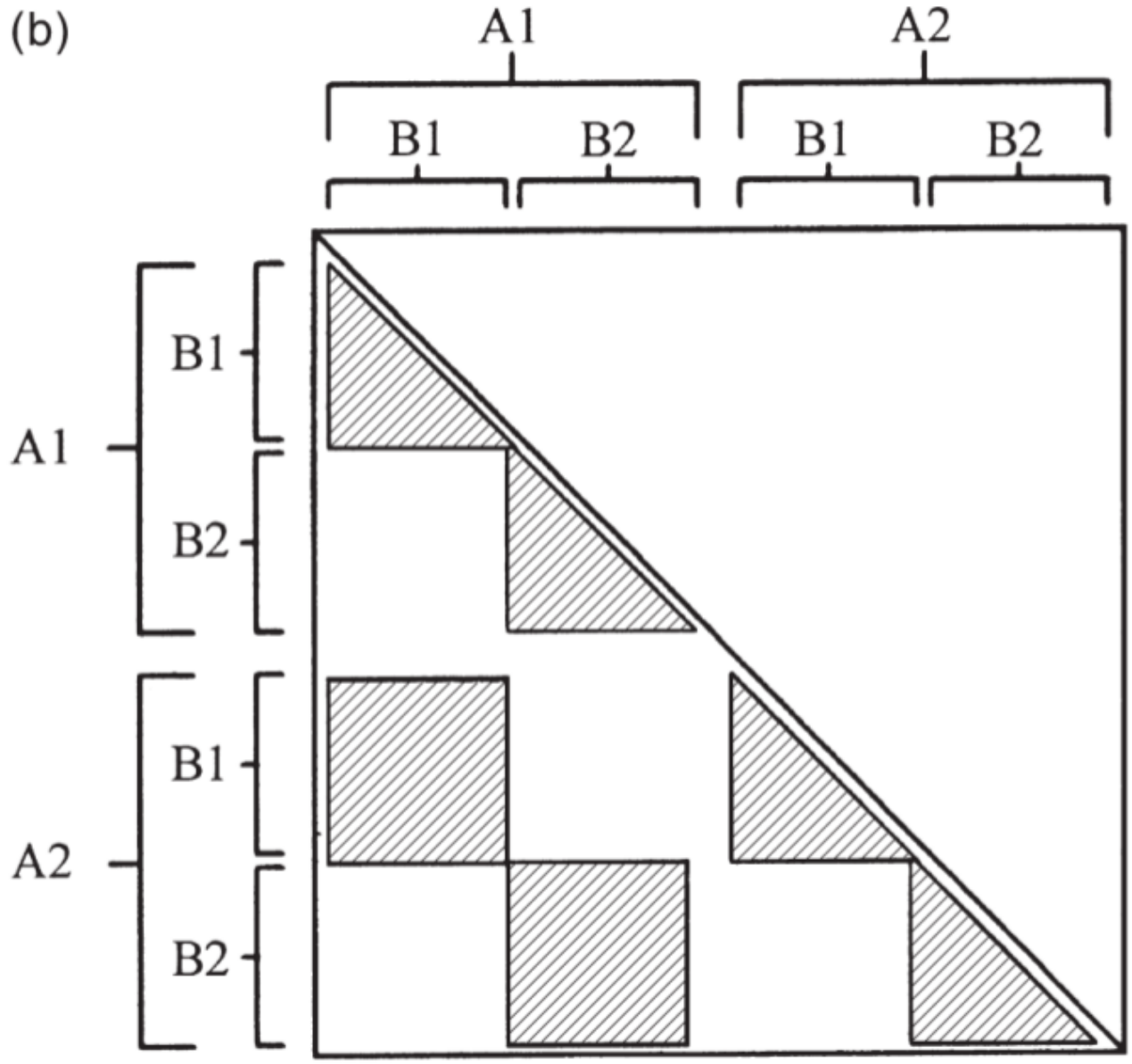
- PCoA: principal coordinate analysis
  - any metric distance, even if non-Euclidean
  - like PCA, eigenvalue decomposition (maximum variance) but mediated by distance function (no original descriptors)
  - unlike PCA, does not allow projection of original descriptors in reduced-dimension space


Weighted UniFrac
Second Principal Coordinate (3.1%)
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
First Principal Coordinate (23.7%)
Site
Anterior_nares
Stool



$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

Anderson MJ et al. Austral Ecol 2001;26(1):32-46.



Anderson MJ et al. Austral Ecol 2001;26(1):32-46.

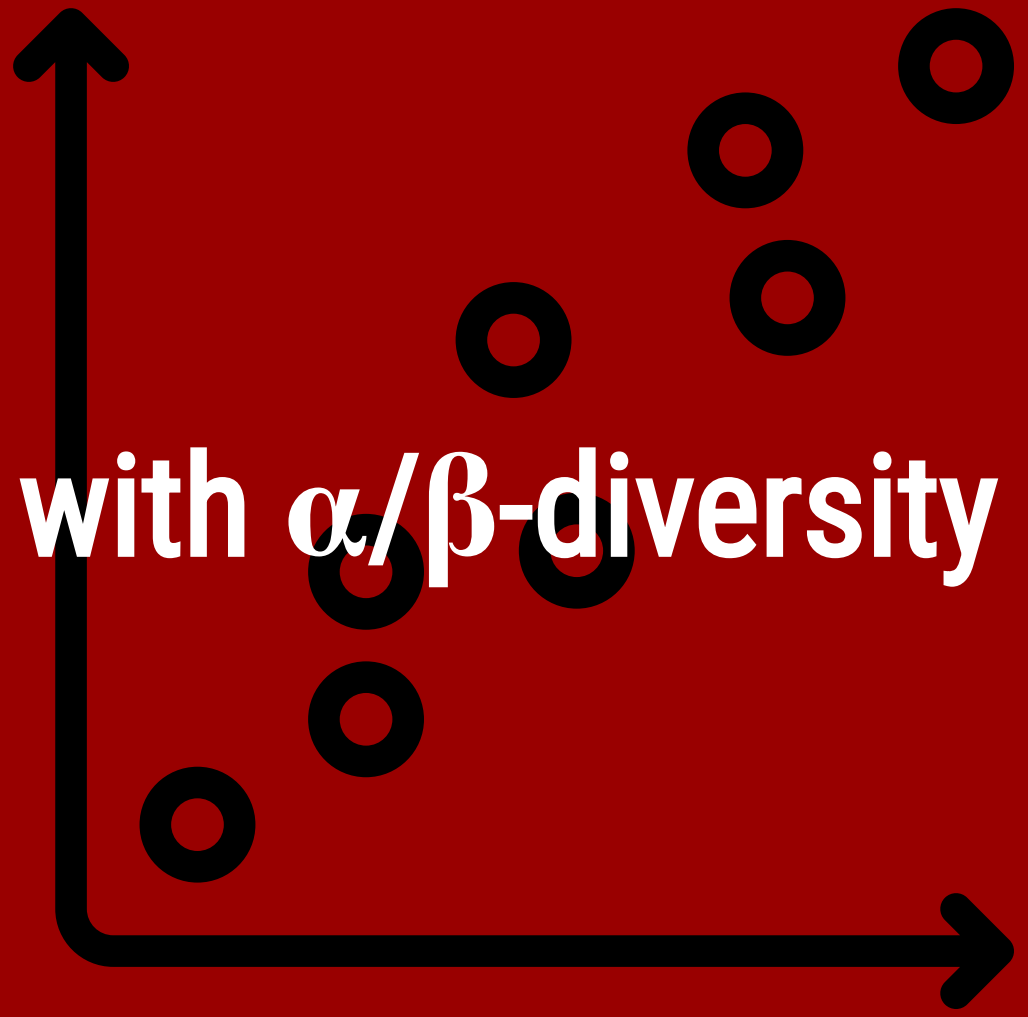
$$R^2 = 1 - \frac{SS_W}{SS_W + SS_A} = \frac{SS_A}{SS_T} \qquad \omega^2 = \frac{SS_A - (a-1)\frac{SS_W}{N-a}}{SS_T + \frac{SS_W}{N-a}}$$

Table 1. Effect sizes observed from various exposures/interventions in studies of various microbiome sampling sites are shown as measured by omega-squared (ω^2) statistics, together with the *P*-values from PERMANOVA test

Site	Comparison groups		ω^2/*P*-value				
	Control	Exposure	Weighted UniFrac	Unweighted UniFrac	Weighted Jaccard	Unweighted Jaccard	Reference
Nares	Non-smoker (33)	Smoker (29)	0.042/0.001	0.009/0.001	0.023/0.001	0.007/0.001	Charlson *et al.* (2010)
Oral	Non-smoker (33)	Smoker (29)	0.032/0.001	0.008/0.001	0.024/0.001	0.007/0.001	Charlson *et al.* (2010)
Gut	Before feeding (10)	After feeding (10)	0.056/0.138	0.013/0.986	0/0.989	0.014/0.985	Wu *et al.* (2011)
Oral	No azithromycin (42)	Azithromycin (6)	0.063/0.01	0.039/0.001	0.099/0.004	0.032/0.001	Charlson *et al.* (2012)
Lung	No azithromycin (34)	Azithromycin (6)	0.065/0.005	0.038/0.001	0.019/0.089	0.033/0.001	Charlson *et al.* (2012)
Skin	Left retroauricular (186)	Right retroauricular (187)	0.000/0.828	0.0001/0.327	0.000/0.986	0.000/1.000	HMP Consortium (2012b)
Human	Anterior nares (161)	Stool (187)	0.567/0.001	0.201/0.001	0.230/0.001	0.117/0.001	HMP Consortium (2012b)

Kelly BJ et al. Bioinformatics 2015;31(15):2461-8.

Linear regression with α/β-diversity



Linear Regression with `lm()`

```
# install.packages("tidyverse")
library(tidyverse)

# install.packages("vegan")
library(vegan)

# install.packages("ape")
library(ape)

set.seed(16)

otu_tab <- read_rds(
"./data/HMP_OTU_table_matrix_stool_nares.rds"
)

otu_tab %>%
  str(vec.len = 3)
```

```
##  num [1:43140, 1:10] 0 0 0 0 0 0 0 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:43140] "OTU_97.1" "OTU_97.10" "OTU_97.100" ...
##   ..$ : chr [1:10] "700014718" "700014767" "700014923" ...
```

Linear Regression with `lm()`

```
otu_tab %>%
  as_tibble(rownames = "otu_id") %>%
  gather(key = "specimen_id",
         value = "read_count",
         -otu_id) %>%
  distinct() -> otu_long

otu_long
```

```
## # A tibble: 431,400 x 3
##    otu_id       specimen_id read_count
##    <chr>        <chr>             <dbl>
##  1 OTU_97.1     700014718             0
##  2 OTU_97.10    700014718             0
##  3 OTU_97.100   700014718             0
##  4 OTU_97.1000  700014718             0
##  5 OTU_97.10000 700014718             0
##  6 OTU_97.10001 700014718             0
##  7 OTU_97.10002 700014718             0
##  8 OTU_97.10003 700014718             0
##  9 OTU_97.10004 700014718             0
## 10 OTU_97.10005 700014718             0
## # … with 431,390 more rows
```

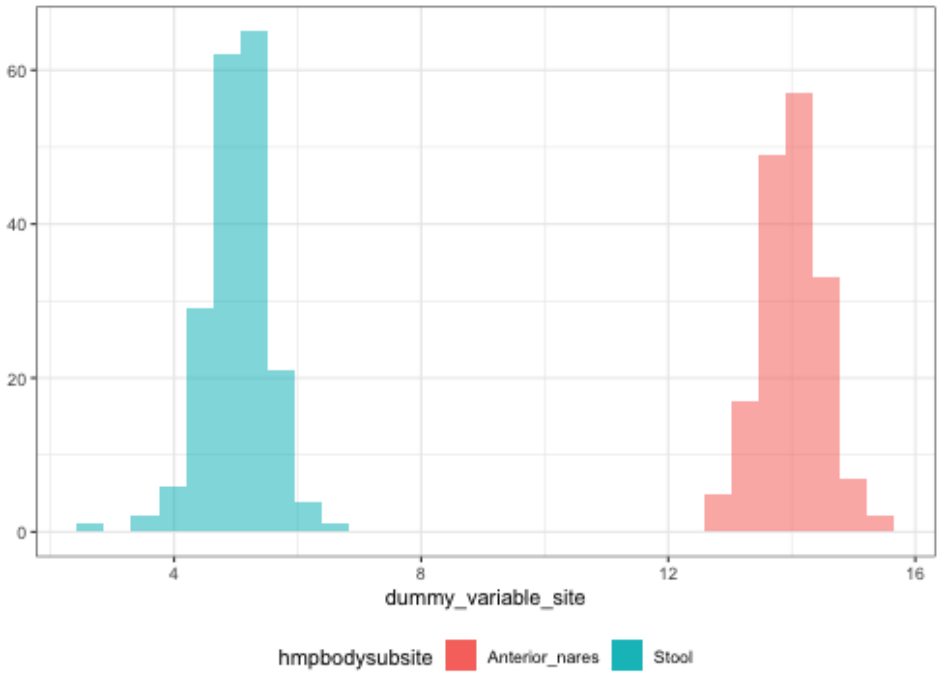
Linear Regression with `lm()`

```
read_tsv(file = "./data/v13_map_uniquebyPSN.txt.bz2") %>%
  rename_all(.funs = ~ gsub("#","",tolower(.x))) %>%
  rename(specimen_id = sampleid) %>%
  distinct() -> specimen_data

specimen_data %>%
  group_by(hmpbodysubsite) %>%
  mutate(dummy_variable_site =
           rnorm(n = length(hmpbodysubsite),
                 mean = nchar(unique(hmpbodysubsite)),
                 sd = 0.5)) %>%
  ungroup() %>%
  filter(hmpbodysubsite %in%
           c("Anterior_nares","Stool")) %>%
  select(specimen_id,
         hmpbodysubsite,
         dummy_variable_site) %>%
  mutate(specimen_id = as.character(specimen_id)) %>%
  distinct() -> specimen_data

specimen_data
```

```
## # A tibble: 361 x 3
##    specimen_id hmpbodysubsite dummy_variable_site
##    <chr>       <chr>                       <dbl>
##  1 700013549   Stool                        5.87
##  2 700014386   Stool                        4.68
##  3 700014445   Anterior_nares              14.2
##  4 700014488   Stool                        4.95
##  5 700014497   Stool                        5.36
##  6 700014527   Anterior_nares              13.9
##  7 700014555   Stool                        4.39
##  8 700014718   Stool                        4.10
##  9 700014767   Anterior_nares              14.5
## 10 700014797   Anterior_nares              13.3
## # … with 351 more rows
```

Linear Regression with `lm()`

```
specimen_data %>%
  qplot(data = .,
        x = dummy_variable_site,
        fill = hmpbodysubsite,
        alpha = 0.8,
        geom = "histogram",
        position = "identity") +
  scale_alpha(guide = FALSE) +
  theme_bw() +
  theme(legend.position = "bottom")
```



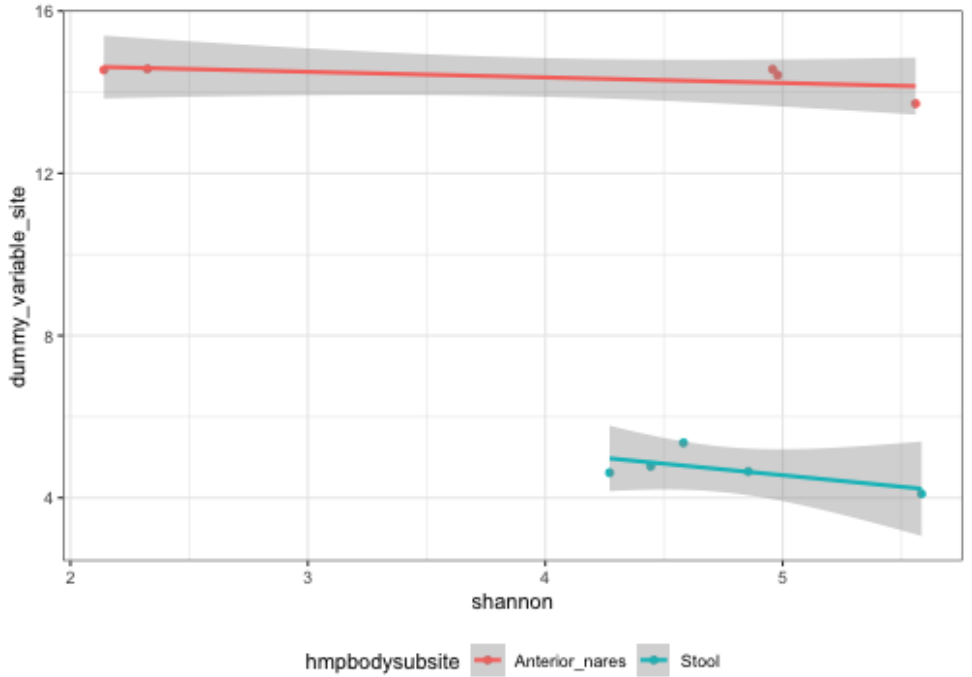
Linear Regression with `lm()`

```
otu_long %>%
  group_by(specimen_id) %>%
  summarise(shannon = diversity(x = read_count,
                                index = "shannon")) %>%
  ungroup() %>%
  left_join(specimen_data, by = "specimen_id") %>%
  mutate(dummy_variable_shannon =
           rnorm(n = length(shannon),
                 mean = 0,
                 sd = 0.2) +
           shannon) %>%
  distinct() -> shannon_summary

shannon_summary
```

```
## # A tibble: 10 x 5
##    specimen_id shannon hmpbodysubsite dummy_variable_site dummy_var
##    <chr>         <dbl> <chr>                         <dbl>
##  1 700014718      5.58 Stool                          4.10
##  2 700014767      2.14 Anterior_nares                14.5
##  3 700014923      2.32 Anterior_nares                14.6
##  4 700016920      4.96 Anterior_nares                14.6
##  5 700023706      4.98 Anterior_nares                14.4
##  6 700038343      5.56 Anterior_nares                13.7
##  7 700095956      4.86 Stool                          4.65
##  8 700105834      4.58 Stool                          5.36
##  9 700107189      4.27 Stool                          4.62
## 10 700109383      4.44 Stool                          4.78
```

Linear Regression with `lm()`

```
shannon_summary %>%
    qplot(data = .,
          x = shannon,
          y = dummy_variable_site,
          color = hmpbodysubsite,
          geom = c("point","smooth"),
          method = "lm") +
  theme_bw() +
  theme(legend.position = "bottom")
```



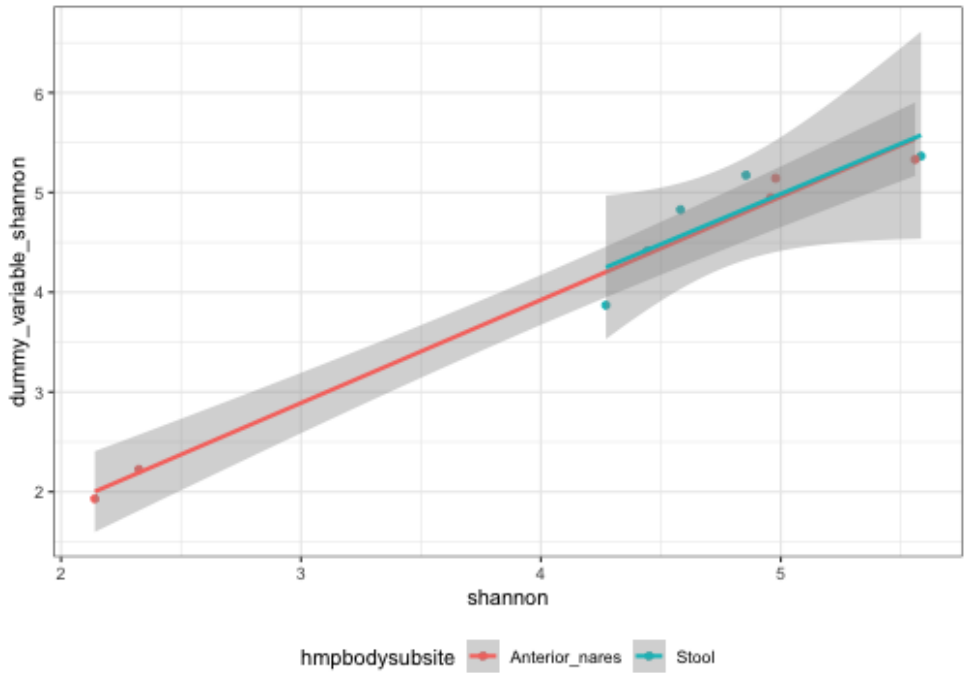
Linear Regression with `lm()`

```
shannon_summary %>%
  lm(formula = dummy_variable_site ~ shannon,
     data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = dummy_variable_site ~ shannon, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0672 -4.0586 -0.9973  4.8409  6.0355
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.328      6.301   2.591    0.032 *
## shannon       -1.555      1.395  -1.115    0.297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.041 on 8 degrees of freedom
## Multiple R-squared:  0.1345,    Adjusted R-squared:  0.02627
## F-statistic: 1.243 on 1 and 8 DF,  p-value: 0.2973
```

Linear Regression with `lm()`

```
shannon_summary %>%
    qplot(data = .,
          x = shannon,
          y = dummy_variable_shannon,
          color = hmpbodysubsite,
          geom = c("point","smooth"),
          method = "lm") +
  theme_bw() +
  theme(legend.position = "bottom")
```



Linear Regression with `lm()`

```
shannon_summary %>%
  lm(formula = dummy_variable_shannon ~ shannon,
     data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = dummy_variable_shannon ~ shannon, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34993 -0.18375  0.01602  0.14710  0.34814
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.20539    0.30165  -0.681    0.515
## shannon      1.03619    0.06678  15.515 2.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2413 on 8 degrees of freedom
## Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9638
## F-statistic: 240.7 on 1 and 8 DF,  p-value: 2.965e-07
```

Linear Regression with `lm()`

```
shannon_summary %>%
  lm(formula = dummy_variable_shannon ~ shannon +
       hmpbodysubsite,
     data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = dummy_variable_shannon ~ shannon + hmpbodysubsite,
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3688 -0.1690  0.0134  0.1674  0.3323
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.20055    0.32224  -0.622    0.553
## shannon              1.03088    0.07540  13.672 2.64e-06 ***
## hmpbodysubsiteStool  0.03673    0.17233   0.213    0.837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2572 on 7 degrees of freedom
## Multiple R-squared:  0.968,    Adjusted R-squared:  0.9589
## F-statistic:   106 on 2 and 7 DF,  p-value: 5.834e-06
```

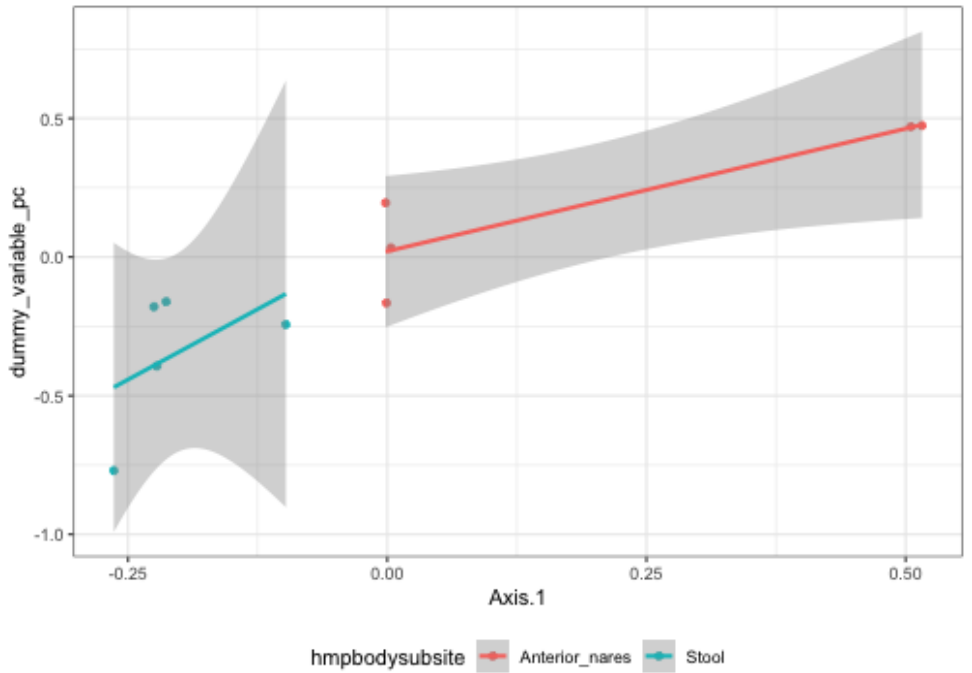
Linear Regression with `lm()`

```
otu_tab %>%
  t() %>%  # TRANSPOSE
  vegdist(x = ., method = "jaccard") %>%
  pcoa(D = .) -> pc

pc$vectors %>%
  as_tibble(rownames = "specimen_id") %>%
  select(specimen_id, Axis.1, Axis.2) %>%
  left_join(shannon_summary, by = "specimen_id") %>%
  mutate(dummy_variable_pc =
           rnorm(n = length(shannon),
                 mean = 0,
                 sd = 0.2) +
           Axis.1) %>%
  distinct() -> pc_summary

pc_summary
```

```
## # A tibble: 10 x 8
##    specimen_id   Axis.1  Axis.2 shannon hmpbodysubsite dummy_variab
##    <chr>          <dbl>   <dbl>   <dbl> <chr>                    <
##  1 700014718   -2.64e-1  0.222     5.58 Stool
##  2 700014767    5.16e-1  0.174     2.14 Anterior_nares           1
##  3 700014923    5.05e-1  0.183     2.32 Anterior_nares           1
##  4 700016920    3.50e-3 -0.434     4.96 Anterior_nares           1
##  5 700023706   -8.28e-4 -0.290     4.98 Anterior_nares           1
##  6 700038343   -1.54e-3 -0.435     5.56 Anterior_nares           1
##  7 700095956   -2.22e-1  0.189     4.86 Stool
##  8 700105834   -2.25e-1  0.174     4.58 Stool
##  9 700107189   -9.78e-2  0.0363    4.27 Stool
## 10 700109383   -2.13e-1  0.180     4.44 Stool
## # … with 2 more variables: dummy_variable_shannon <dbl>,
## #   dummy_variable_pc <dbl>
```

Linear Regression with `lm()`

```
pc_summary %>%
  qplot(data = .,
        x = Axis.1,
        y = dummy_variable_pc,
        color = hmpbodysubsite,
        geom = c("point","smooth"),
        method = "lm") +
  theme_bw() +
  theme(legend.position = "bottom")
```



Linear Regression with `lm()`

```
pc_summary %>%
  lm(formula = dummy_variable_pc ~ Axis.1,
     data = .) %>%
  summary()
```

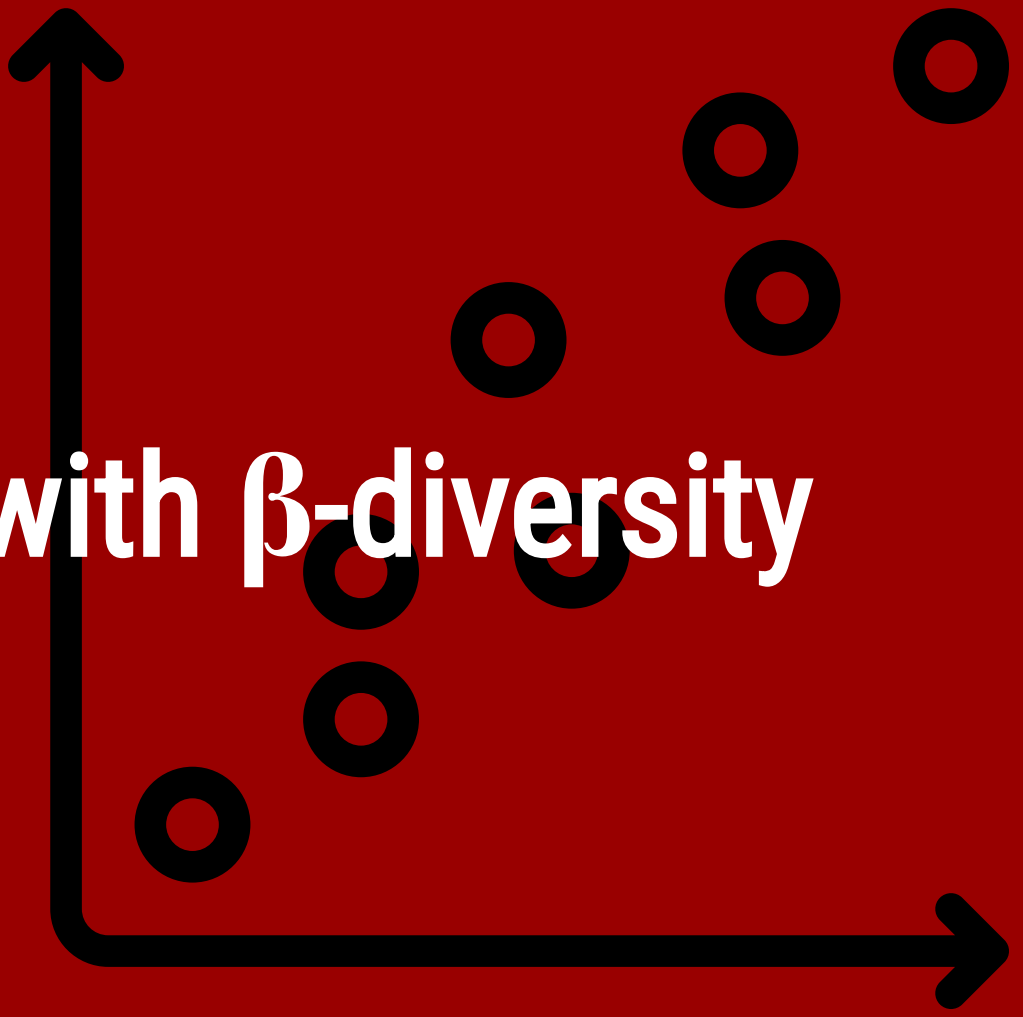
```
##
## Call:
## lm(formula = dummy_variable_pc ~ Axis.1, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38811 -0.05846 -0.04998  0.14369  0.27116
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.07387    0.06180  -1.195 0.266241
## Axis.1       1.16777    0.22633   5.160 0.000864 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1954 on 8 degrees of freedom
## Multiple R-squared:  0.7689,    Adjusted R-squared:   0.74
## F-statistic: 26.62 on 1 and 8 DF,  p-value: 0.0008641
```

Linear Regression with `lm()`

```
pc_summary %>%
  lm(formula = dummy_variable_pc ~ Axis.2,
     data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = dummy_variable_pc ~ Axis.2, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63113 -0.17200 -0.04452  0.10175  0.59896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.07387    0.12574  -0.587    0.573
## Axis.2      -0.29223    0.48408  -0.604    0.563
##
## Residual standard error: 0.3976 on 8 degrees of freedom
## Multiple R-squared:  0.04357,    Adjusted R-squared:  -0.07598
## F-statistic: 0.3644 on 1 and 8 DF,  p-value: 0.5628
```

PERMANOVA with β-diversity



PERMANOVA with `adonis()`

```
otu_tab %>%
  t() %>%  # TRANSPOSE
  vegdist(x = ., method = "jaccard") -> otu_dist

otu_dist %>%
  str(vec.len = 2)
```

```
##  'dist' num [1:45] 1 0.996 ...
##  - attr(*, "Size")= int 10
##  - attr(*, "Labels")= chr [1:10] "700014718" "700014767" ...
##  - attr(*, "Diag")= logi FALSE
##  - attr(*, "Upper")= logi FALSE
##  - attr(*, "method")= chr "jaccard"
##  - attr(*, "call")= language vegdist(x = ., method = "jaccard")
```

PERMANOVA with adonis()

```
labels(otu_dist) %>% #match order from dist
  enframe(value = "specimen_id") %>%
  select(specimen_id) %>%
  left_join(pc_summary, by = "specimen_id") %>%
  mutate(dummy_category = Axis.1 > mean(Axis.1)) %>%
  distinct() -> sorted_summary

sorted_summary
```

```
## # A tibble: 10 x 9
##    specimen_id   Axis.1  Axis.2 shannon hmpbodysubsite dummy_variab
##    <chr>          <dbl>   <dbl>   <dbl> <chr>                    <
##  1 700014718   -2.64e-1  0.222     5.58 Stool
##  2 700014767    5.16e-1  0.174     2.14 Anterior_nares           1
##  3 700014923    5.05e-1  0.183     2.32 Anterior_nares           1
##  4 700016920    3.50e-3 -0.434     4.96 Anterior_nares           1
##  5 700023706   -8.28e-4 -0.290     4.98 Anterior_nares           1
##  6 700038343   -1.54e-3 -0.435     5.56 Anterior_nares           1
##  7 700095956   -2.22e-1  0.189     4.86 Stool
##  8 700105834   -2.25e-1  0.174     4.58 Stool
##  9 700107189   -9.78e-2  0.0363    4.27 Stool
## 10 700109383   -2.13e-1  0.180     4.44 Stool
## # … with 3 more variables: dummy_variable_shannon <dbl>,
## #   dummy_variable_pc <dbl>, dummy_category <lgl>
```

PERMANOVA with adonis()

```
# distance matrix is response variable
adonis(otu_dist ~ hmpbodysubsite,
       data = sorted_summary)
```

```
##
## Call:
## adonis(formula = otu_dist ~ hmpbodysubsite, data = sorted_summary)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## hmpbodysubsite  1    0.7036 0.70364  1.5719 0.16422  0.014 *
## Residuals       8    3.5811 0.44764         0.83578
## Total           9    4.2847                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PERMANOVA with adonis()

```
# multivariable possible...
adonis(otu_dist ~ hmpbodysubsite + dummy_category,
       data = sorted_summary)
```

```
##
## Call:
## adonis(formula = otu_dist ~ hmpbodysubsite + dummy_category,      data = sorted_summary)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## hmpbodysubsite  1    0.7036 0.70364  1.6100 0.16422  0.002 **
## dummy_category  1    0.5218 0.52180  1.1939 0.12178  0.125
## Residuals       7    3.0593 0.43704         0.71400
## Total           9    4.2847                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PERMANOVA with adonis()

```
# ... but order matters!!!
adonis(otu_dist ~ dummy_category + hmpbodysubsite,
       data = sorted_summary)
```

```
##
## Call:
## adonis(formula = otu_dist ~ dummy_category + hmpbodysubsite,      data = sorted_summary)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## dummy_category  1    0.6329 0.63293  1.4482 0.14772  0.018 *
## hmpbodysubsite  1    0.5925 0.59251  1.3557 0.13828  0.036 *
## Residuals       7    3.0593 0.43704         0.71400
## Total           9    4.2847                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PERMANOVA with adonis()

```
# ... do you mean strata?
adonis(otu_dist ~ dummy_category,
       strata = sorted_summary$hmpbodysubsite,
       data = sorted_summary)
```

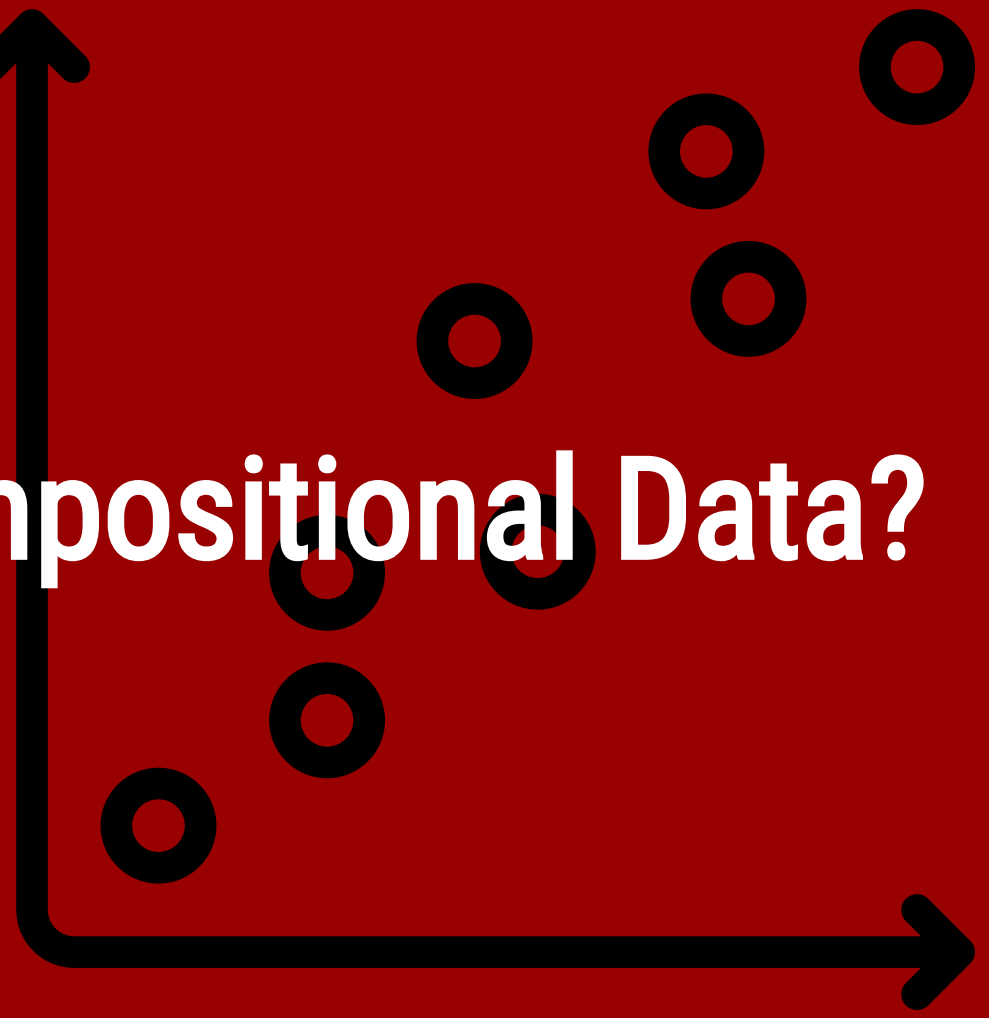
```
##
## Call:
## adonis(formula = otu_dist ~ dummy_category, data = sorted_summary,      strata = sorted_summary$hmpbodysubsite)
##
## Blocks:  strata
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## dummy_category  1    0.6329 0.63293  1.3865 0.14772  0.291
## Residuals       8    3.6518 0.45648         0.85228
## Total           9    4.2847                 1.00000
```

PERMANOVA with adonis()

```
# ... or do you mean nestedness?
adonis(otu_dist ~ dummy_category / hmpbodysubsite,
       data = sorted_summary)
```

```
##
## Call:
## adonis(formula = otu_dist ~ dummy_category/hmpbodysubsite, data = sorted_summary)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                               Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## dummy_category                 1    0.6329 0.63293  1.4482 0.14772  0.009 **
## dummy_category:hmpbodysubsite  1    0.5925 0.59251  1.3557 0.13828  0.023 *
## Residuals                      7    3.0593 0.43704         0.71400
## Total                          9    4.2847                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression & Compositional Data?

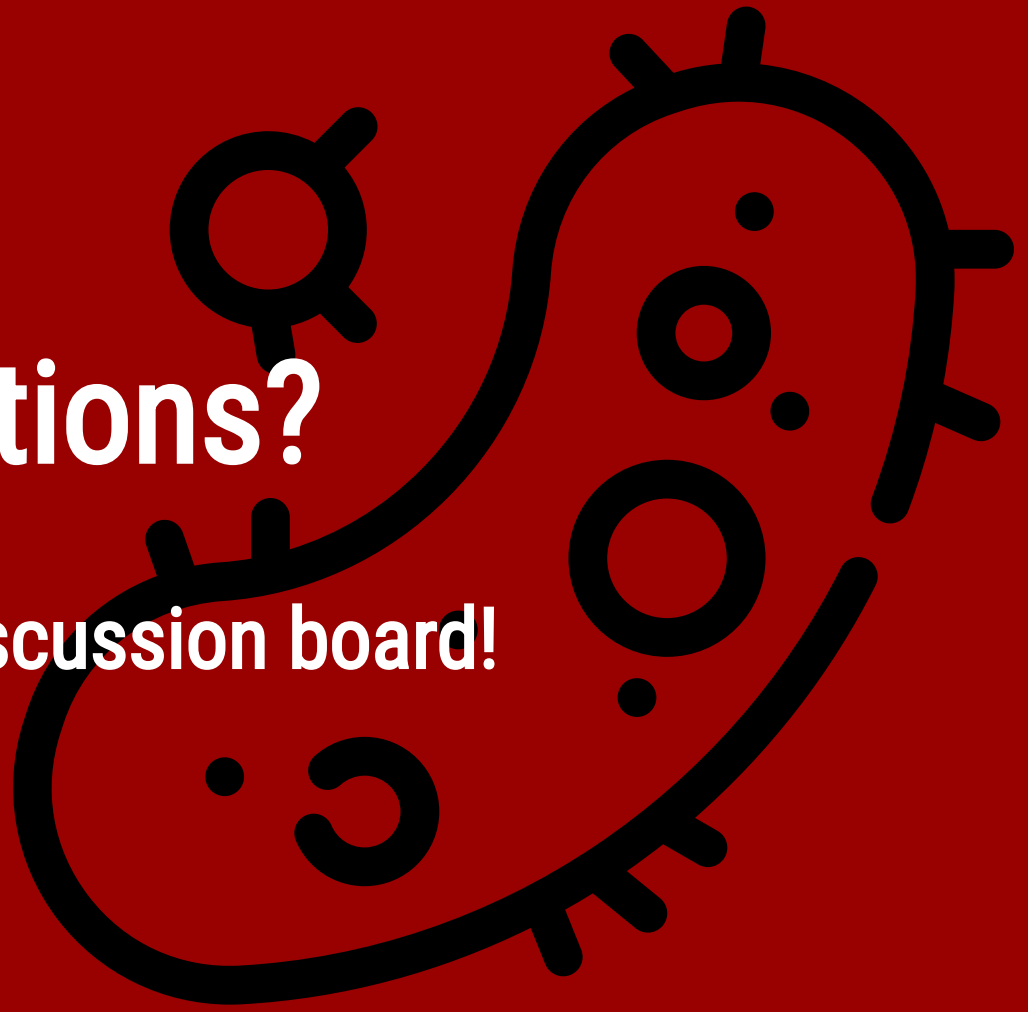


Regression & Compositional Data?

- Compositional data approaches correct OTU dependency:
  - e.g., `compositions::clr()` or `philr::philr()`
  - p >> n challenges persist
- Must pair compositional transform with regularization:
  - `glmnet::glmnet` for LASSO/ridge/elastic net
  - Bayesian methods

Questions?

Post to the discussion board!



Thank you!

Slides available: github.com/bjklab

brendank@pennmedicine.upenn.edu