Beta Diversity:
Inter-Community Difference

📚EPID 674📚

Brendan J. Kelly, MD, MS

Updated: 11 June 2020

Beta diversity

Pairwise distances

Principal coordinates

PERMANOVA

R's vegan package

Beta diversity

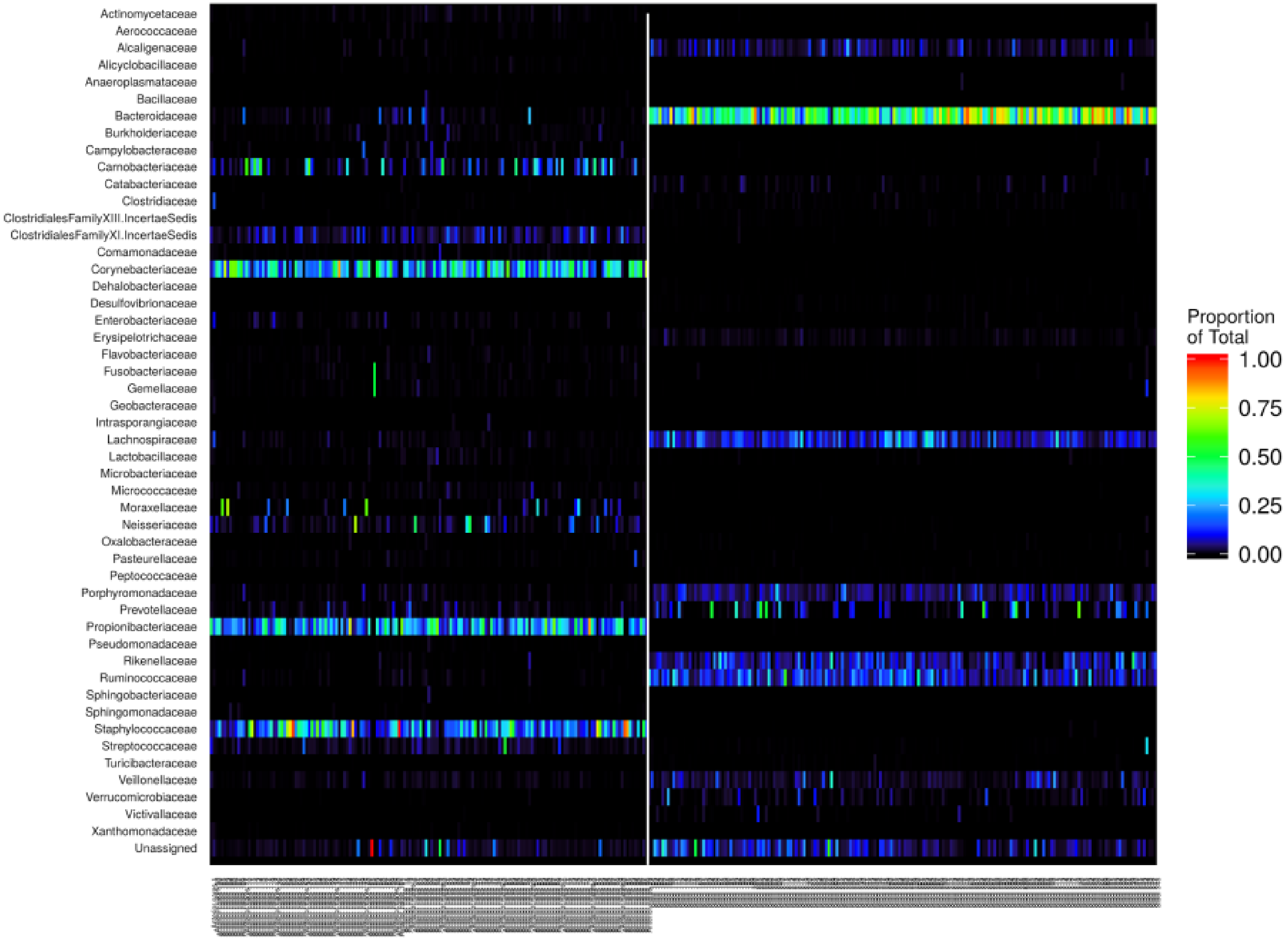
High Dimensional Microbiome Data

```
##              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

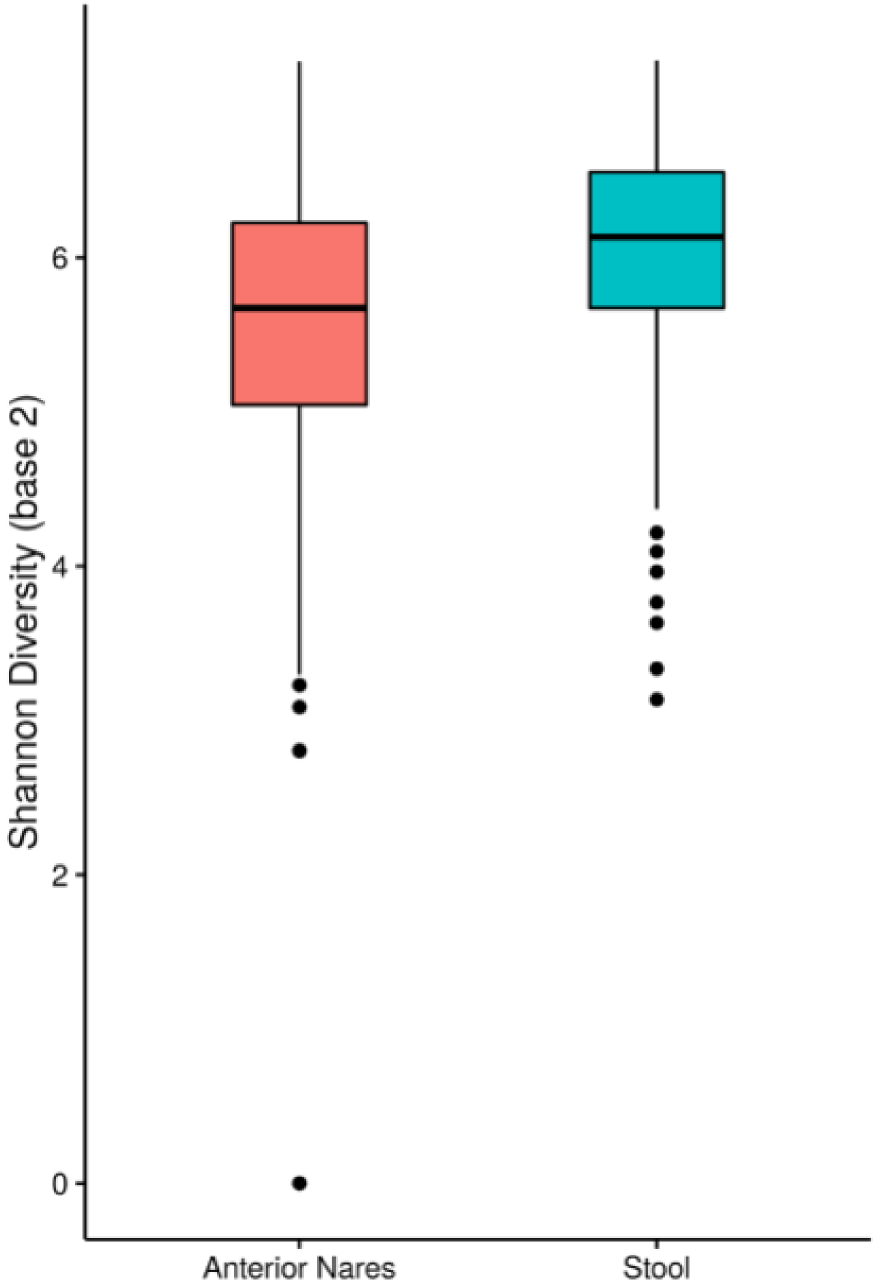
- How to deal with high-dimensional microbiome data?
- **Descriptive (e.g., heatmaps and stacked barplots)**
- Test a priori hypotheses regarding specific OTUs/taxa
- Reduce dimensions:
    - single summary statistic (alpha diversity)
    - pairwise distances (beta diversity) with PCoA or PERMANOVA
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

Anterior Nares vs Stool



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 (base 2)
6
4
2
0
Anterior Nares
Stool

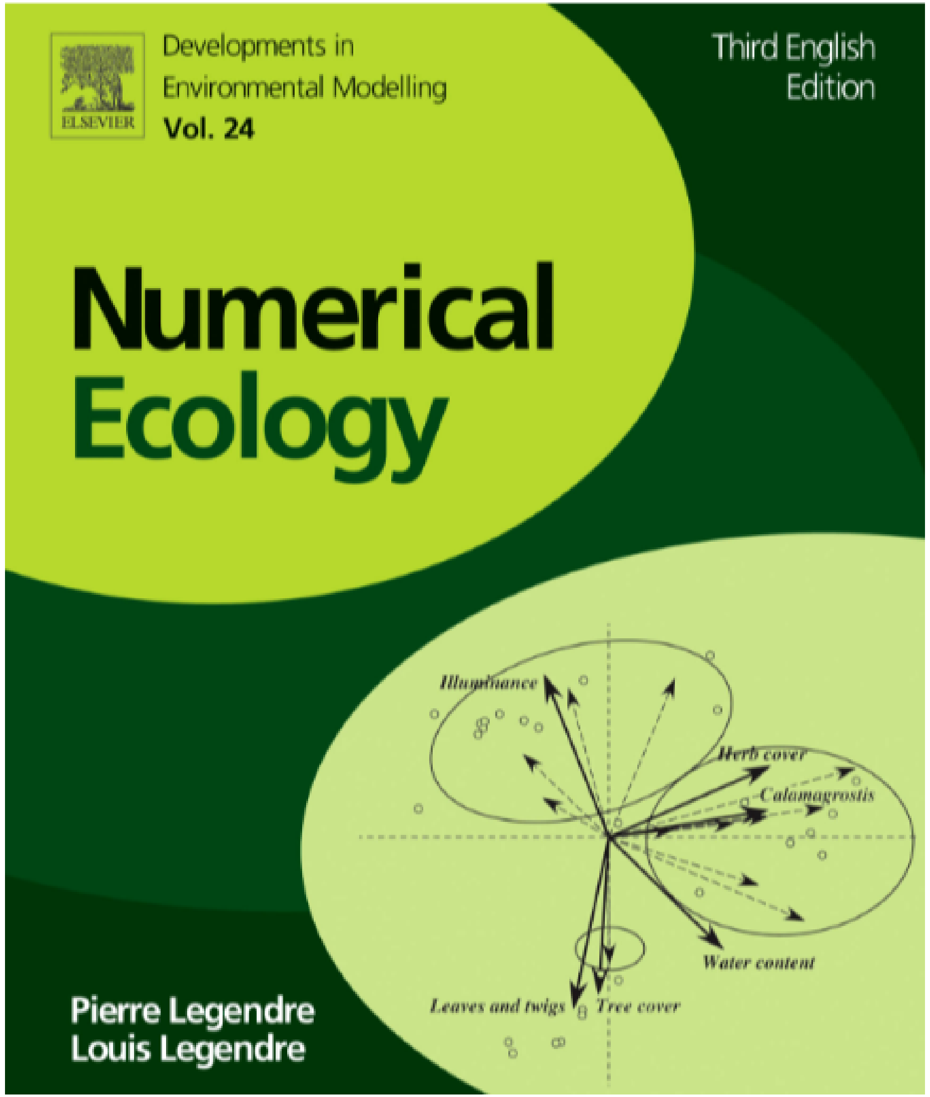
High Dimensional Microbiome Data

- 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

Pairwise distances



Developments in
Environmental Modelling
Vol. 24

Third English
Edition

Numerical Ecology

Pierre Legendre
Louis Legendre

What's in a distance?

- “The most usual approach to assess the resemblance among objects or descriptors is to first condense all (or the relevant part of) the information available in the ecological data matrix (Section 2.1) into **a square matrix of association** among the objects or descriptors (Section 2.2). In most instances, the association matrix is symmetric.”

- Compare variable-variable: “R-mode” (like Pearson’s r coefficient)

- Compare object-object: “Q-mode”

- Six modes of analysis if incorporate time series (Cattell 1966)

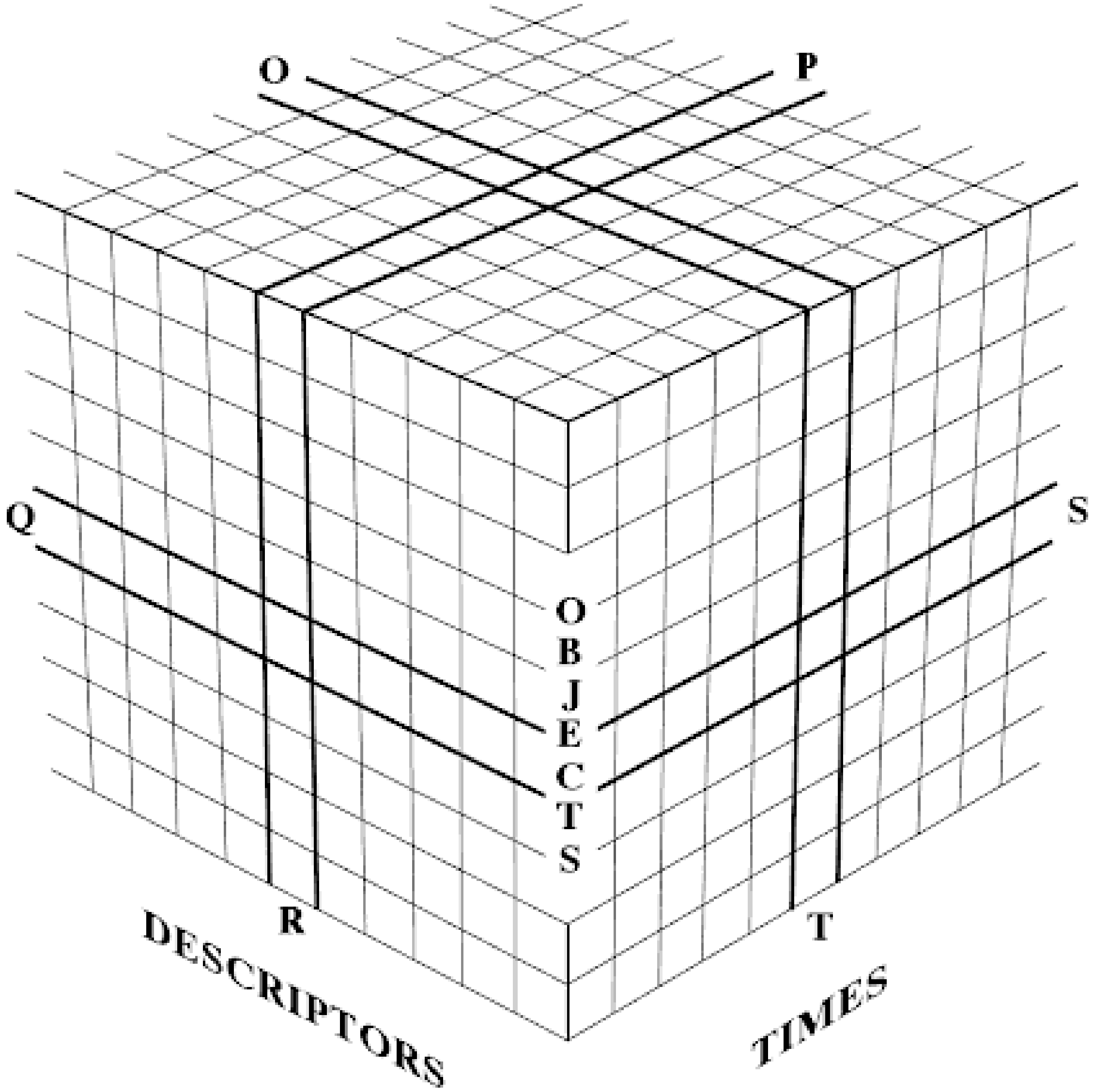
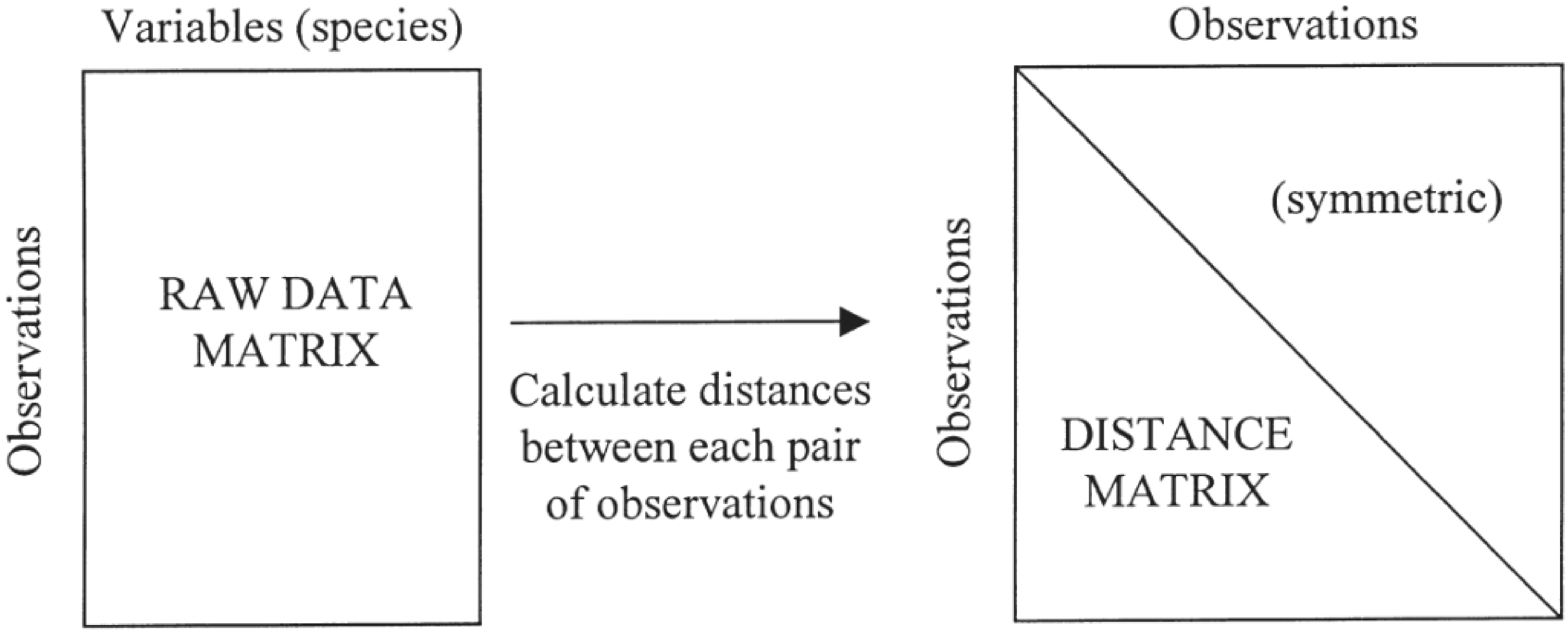


Figure 7.1 The three-dimensional data box (objects × descriptors × times). Adapted from Cattell (1966).

What's in a distance?

- “... association will be used as a general term to describe any measure or coefficient used to quantify the **resemblance or difference** between objects or descriptors, as proposed by Orlóci (1975)."
- Q-mode studies:
  - similarity coefficients (identical = 1)
  - distance (or dissimilarity) coefficients (identical = 0)

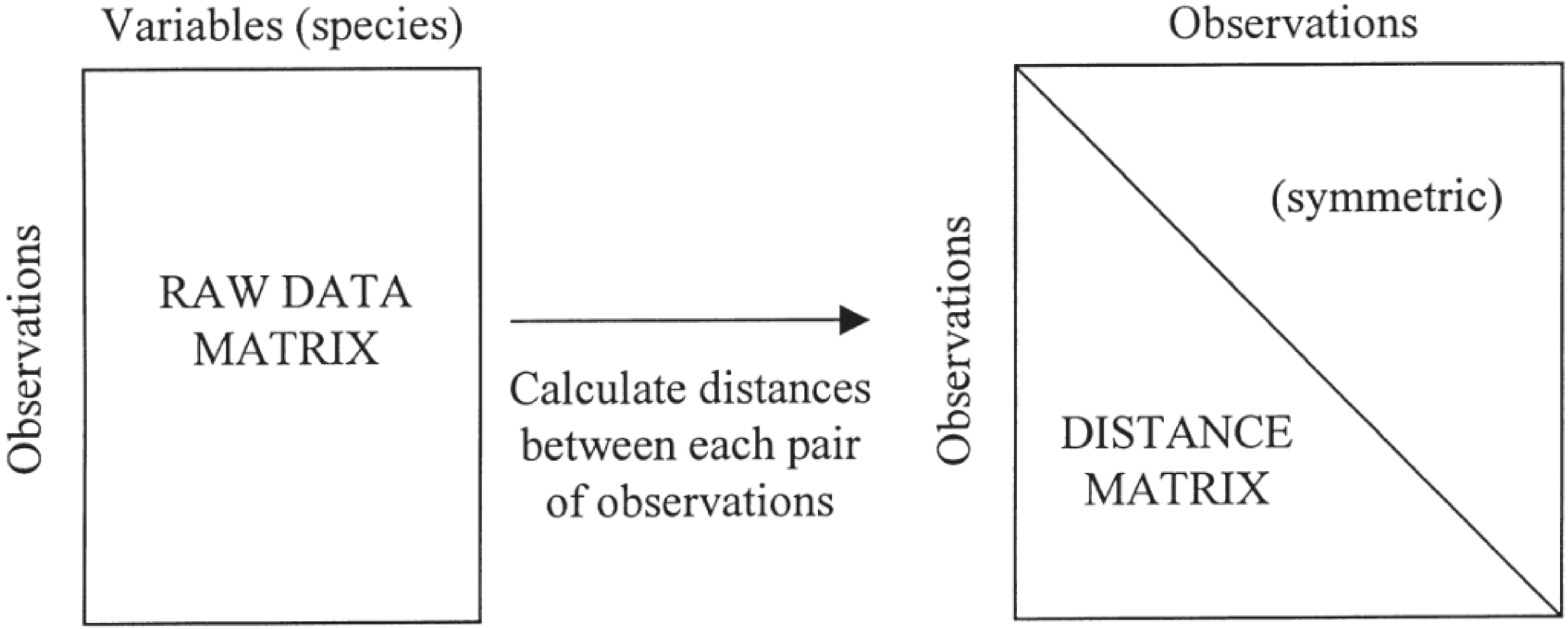

Variables (species)
Observations
RAW DATA MATRIX
Calculate distances between each pair of observations
Observations
Observations
(symmetric)
DISTANCE MATRIX

OTU Table: OTUs x Specimens

```
##              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
```

OTU Table: Specimens x OTUs

```
##           OTU_97.1 OTU_97.10 OTU_97.100 OTU_97.1000 OTU_97.10000 OTU_97.10001
## 700013549        0         0          0           0            0            0
## 700014386        0         0          0           0            0            0
## 700014403        0         6        133           0            0            0
## 700014409        0         4          7           0            0            0
## 700014412        0         1          1           0            0            0
## 700014415        0         5          4           0            0            1
## 700014418        0         2          0           0            0            0
## 700014421        0         3         25           0            0            0
## 700014424        0         1          5           0            0            0
## 700014427        0         1          0           0            0            0
## 700014430        0         6          0           0            0            0
## 700014445        0         0          0           0            0            0
## 700014501        0         2          1           0            0            0
## 700014515        0         0          0           0            0            0
## 700014516        0         0          0           0            0            0
## 700014517        0         0          0           0            0            0
```



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)

The "Double Zero" Problem

- “The proportion of zeros in community composition data generally increases with the variability in environmental conditions among the sampling sites. If sampling has been conducted along one or several environmental axes, the species present are likely to differ at least partly from site to site. **Including double zeros in the comparison between sites would result in high values of similarity for the many pairs of sites holding only a few species**, these pairs presenting many double zeros; this would not provide a correct ecological assessment of the situation.”

The "Double Zero" Problem

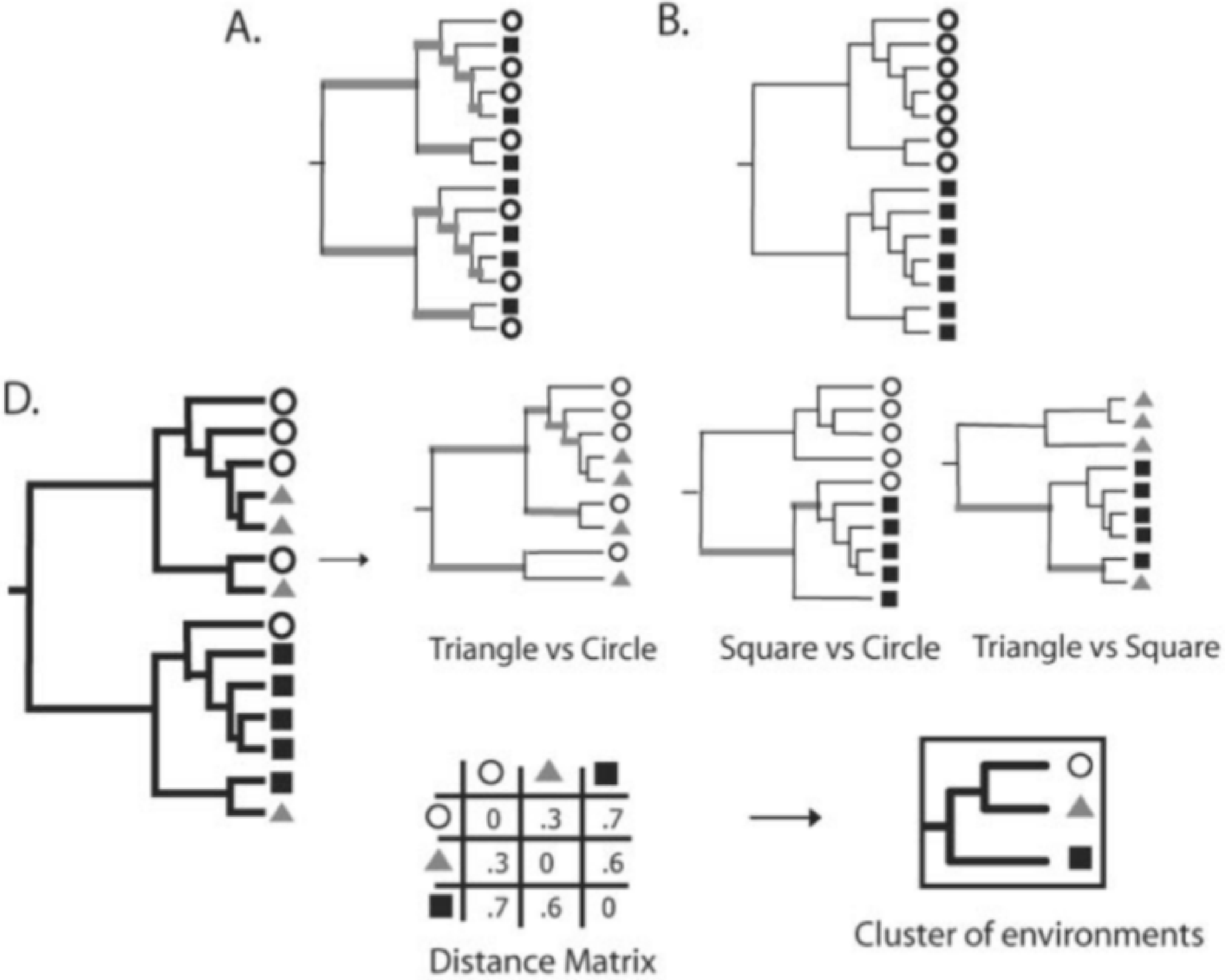
- “Because double zeros are not informative, their interpretation generates **the double zero problem: is the value of an association coefficient affected by inclusion of double zeros in its calculation?** When choosing an association coefficient, ecologists must pay attention to the interpretation of double zeros: except in very limited cases (e.g. controlled experiments involving very few species and with small uncontrolled ecological variation), it is preferable to draw no ecological conclusion from the simultaneous absence of a species at two sites.... In numerical terms, this means to **skip double zeros when computing similarity or distance coefficients** using species presence-absence or abundance data.”

UniFrac

- UniFrac measures the distance between communities based on the lineages they contain.
- Satisfies the technical requirements of a distance metric:
    - always positive
    - transitive
    - satisfies the triangle inequality
- Can thus be used with standard multivariate statistics (e.g., UPGMA, clustering, and PCoA).

UniFrac

- UniFrac “exploits the different degrees of similarity between sequences”:
  - “the unique fraction metric, or UniFrac, measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both”
  - “captures the total amount of evolution that is unique to each state, presumably reflecting adaptation to one environment that would be deleterious in the other” (designed to be based on rRNA)



	○	▲	■
○	0	.3	.7
▲	.3	0	.6
■	.7	.6	0

Lozupone C and Knight R. AEM 2005:8228-35.

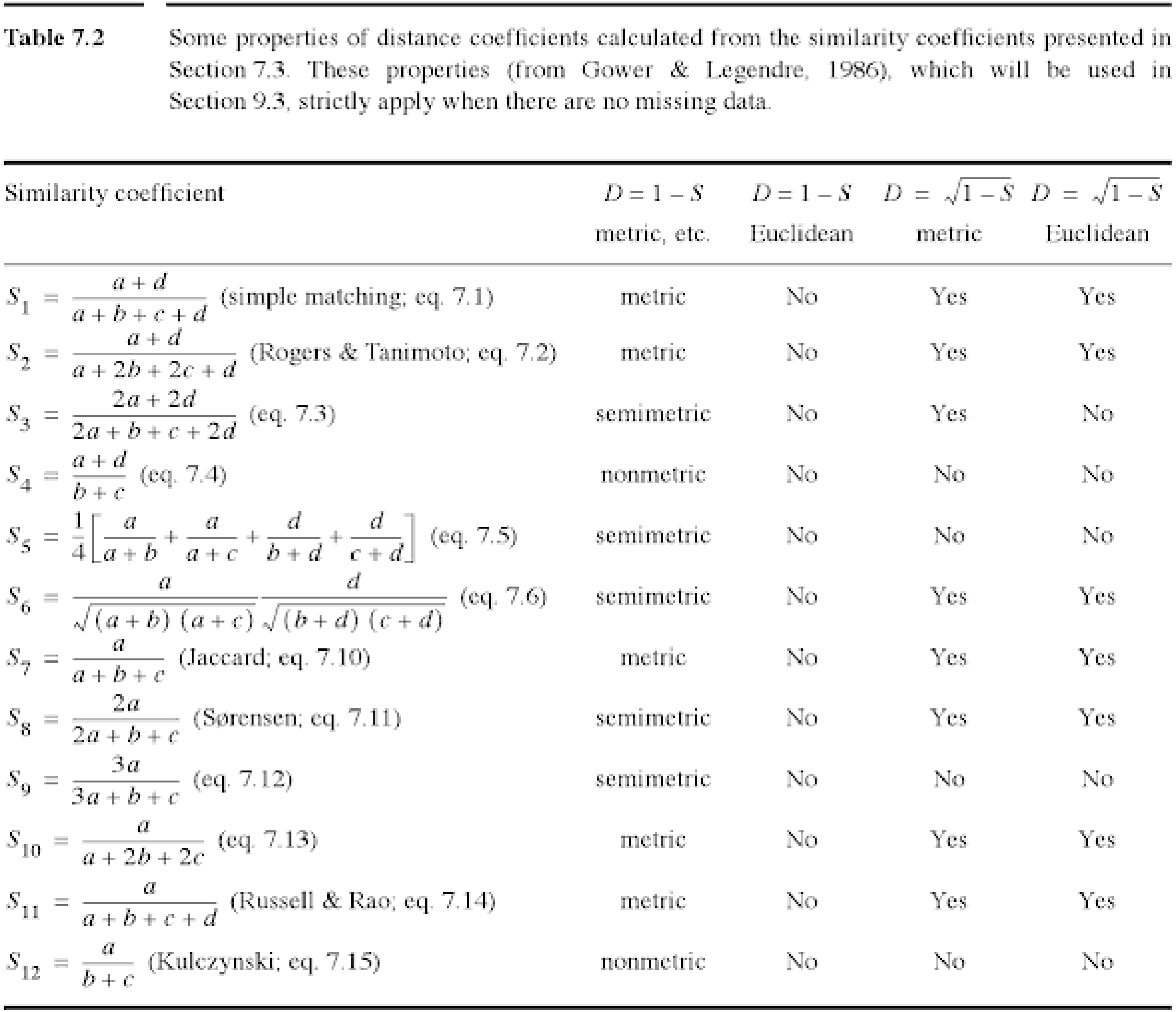


Table 7.2 Some properties of distance coefficients calculated from the similarity coefficients presented in Section 7.3. These properties (from Gower & Legendre, 1986), which will be used in Section 9.3, strictly apply when there are no missing data.

Similarity coefficient	$D = 1 - S$ metric, etc.	$D = 1 - S$ Euclidean	$D = \sqrt{1-S}$ metric	$D = \sqrt{1-S}$ Euclidean
$S_1 = \frac{a+d}{a+b+c+d}$ (simple matching; eq. 7.1)	metric	No	Yes	Yes
$S_2 = \frac{a+d}{a+2b+2c+d}$ (Rogers & Tanimoto; eq. 7.2)	metric	No	Yes	Yes
$S_3 = \frac{2a+2d}{2a+b+c+2d}$ (eq. 7.3)	semimetric	No	Yes	No
$S_4 = \frac{a+d}{b+c}$ (eq. 7.4)	nonmetric	No	No	No
$S_5 = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$ (eq. 7.5)	semimetric	No	No	No
$S_6 = \frac{a}{\sqrt{(a+b)\,(a+c)}}\frac{d}{\sqrt{(b+d)\,(c+d)}}$ (eq. 7.6)	semimetric	No	Yes	Yes
$S_7 = \frac{a}{a+b+c}$ (Jaccard; eq. 7.10)	metric	No	Yes	Yes
$S_8 = \frac{2a}{2a+b+c}$ (Sørensen; eq. 7.11)	semimetric	No	Yes	Yes
$S_9 = \frac{3a}{3a+b+c}$ (eq. 7.12)	semimetric	No	No	No
$S_{10} = \frac{a}{a+2b+2c}$ (eq. 7.13)	metric	No	Yes	Yes
$S_{11} = \frac{a}{a+b+c+d}$ (Russell & Rao; eq. 7.14)	metric	No	Yes	Yes
$S_{12} = \frac{a}{b+c}$ (Kulczynski; eq. 7.15)	nonmetric	No	No	No

Beta Diversity: Which Distance Metric?

- Why use Jaccard? UniFrac?
- Why use weighted? Unweighted?

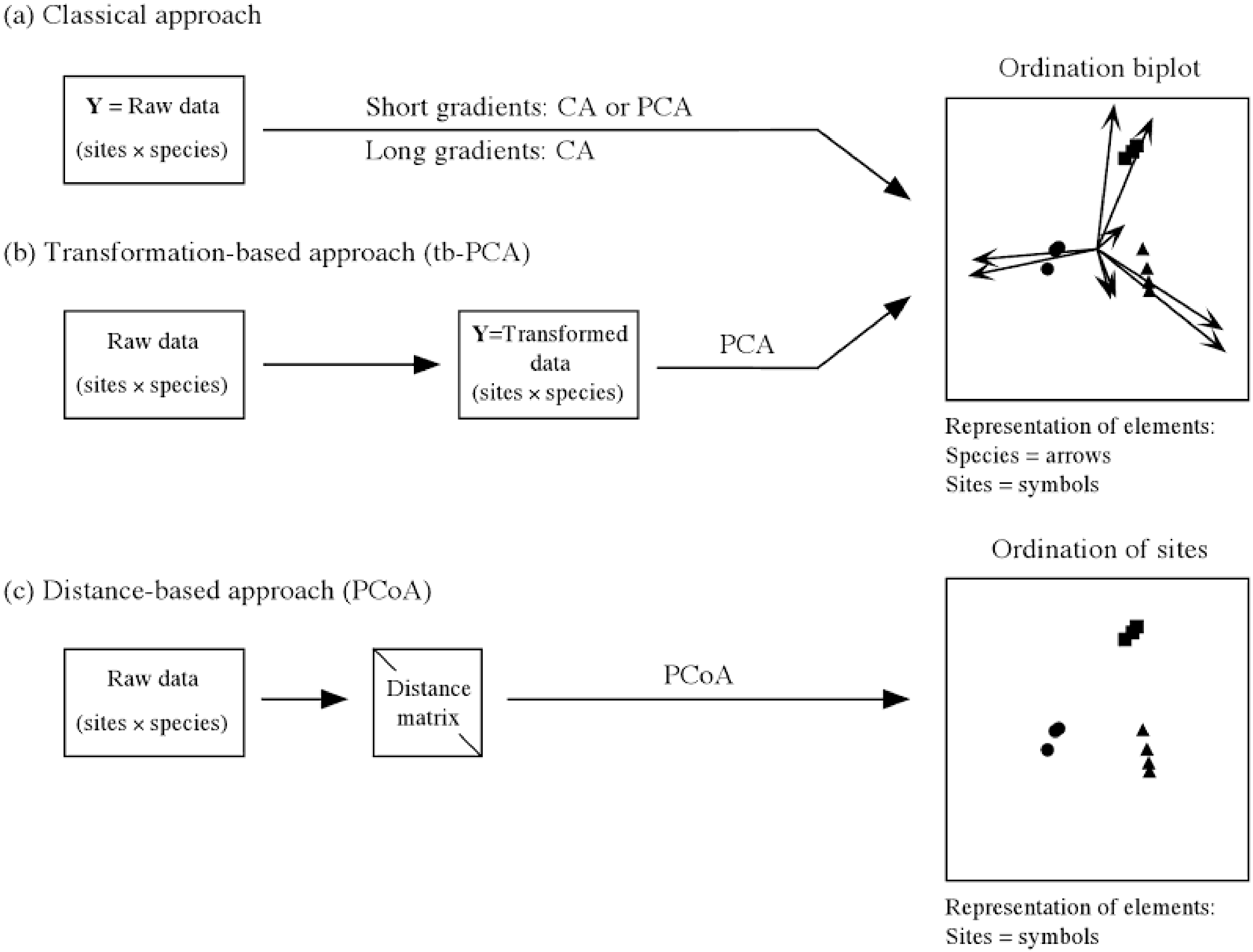
Principal Coordinates

Original Discriptors → PCA

- PCA: principal component analysis
    - rigid rotation for successive directions of maximum variance
    - lots of restrictions (Euclidean)
    - but allows projection of original descriptors in PCA space

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



(a) Classical approach

Y = Raw data

(sites × species)

Short gradients: CA or PCA

Long gradients: CA

Ordination biplot

(b) Transformation-based approach (tb-PCA)

Raw data

(sites × species)

Y=Transformed data

(sites × species)

PCA

Representation of elements:
Species = arrows
Sites = symbols

Ordination of sites

(c) Distance-based approach (PCoA)

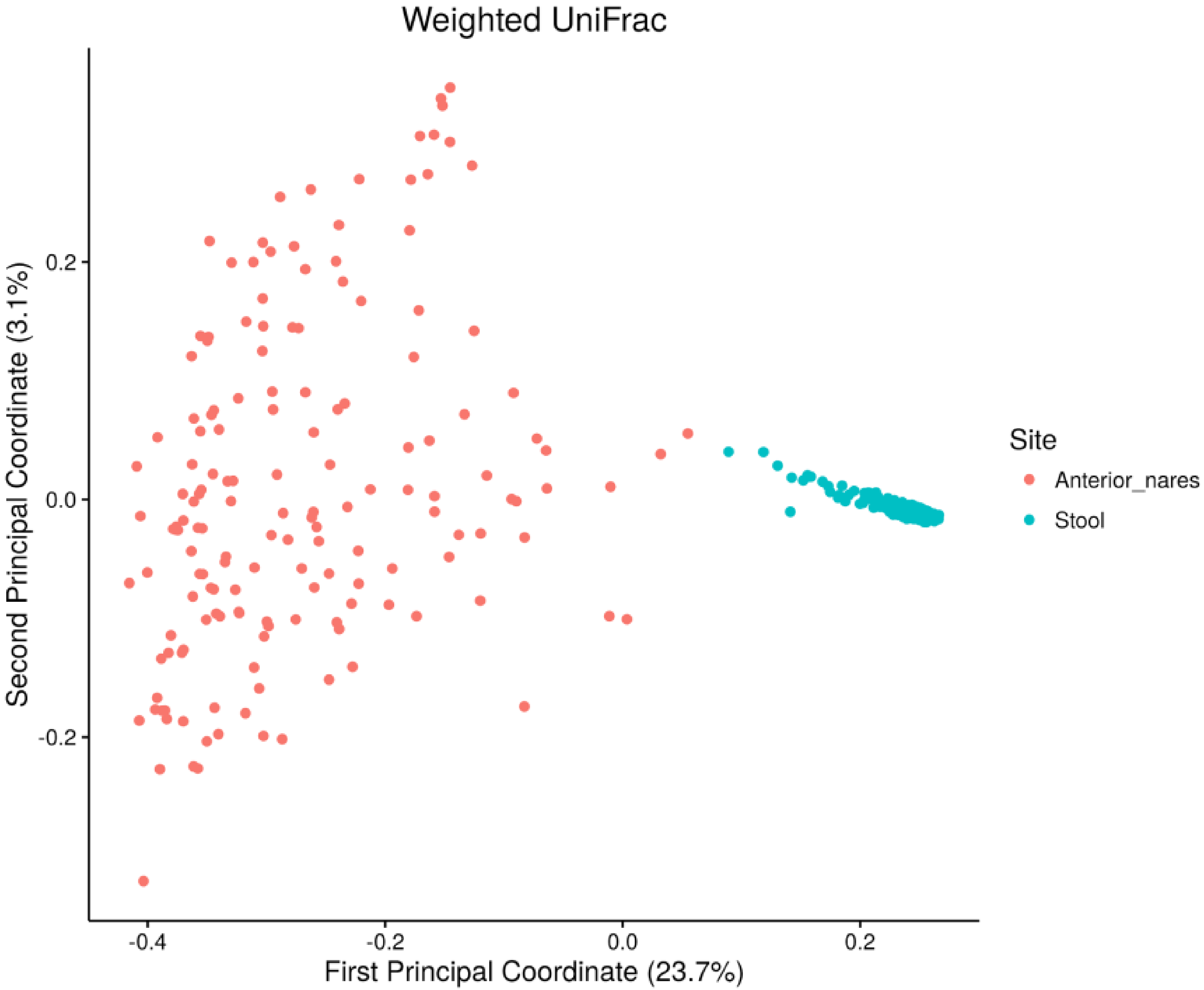
Raw data

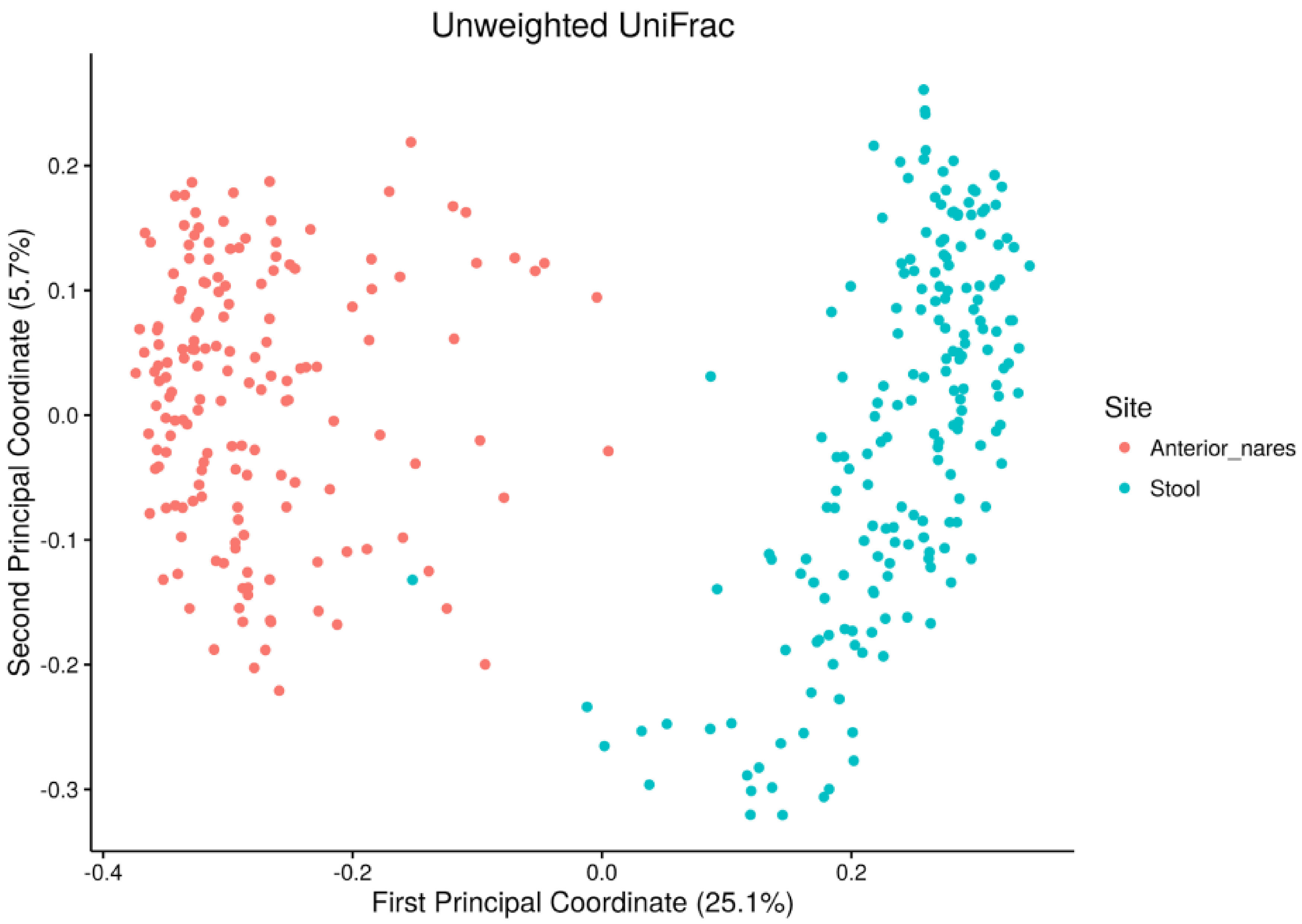
(sites × species)

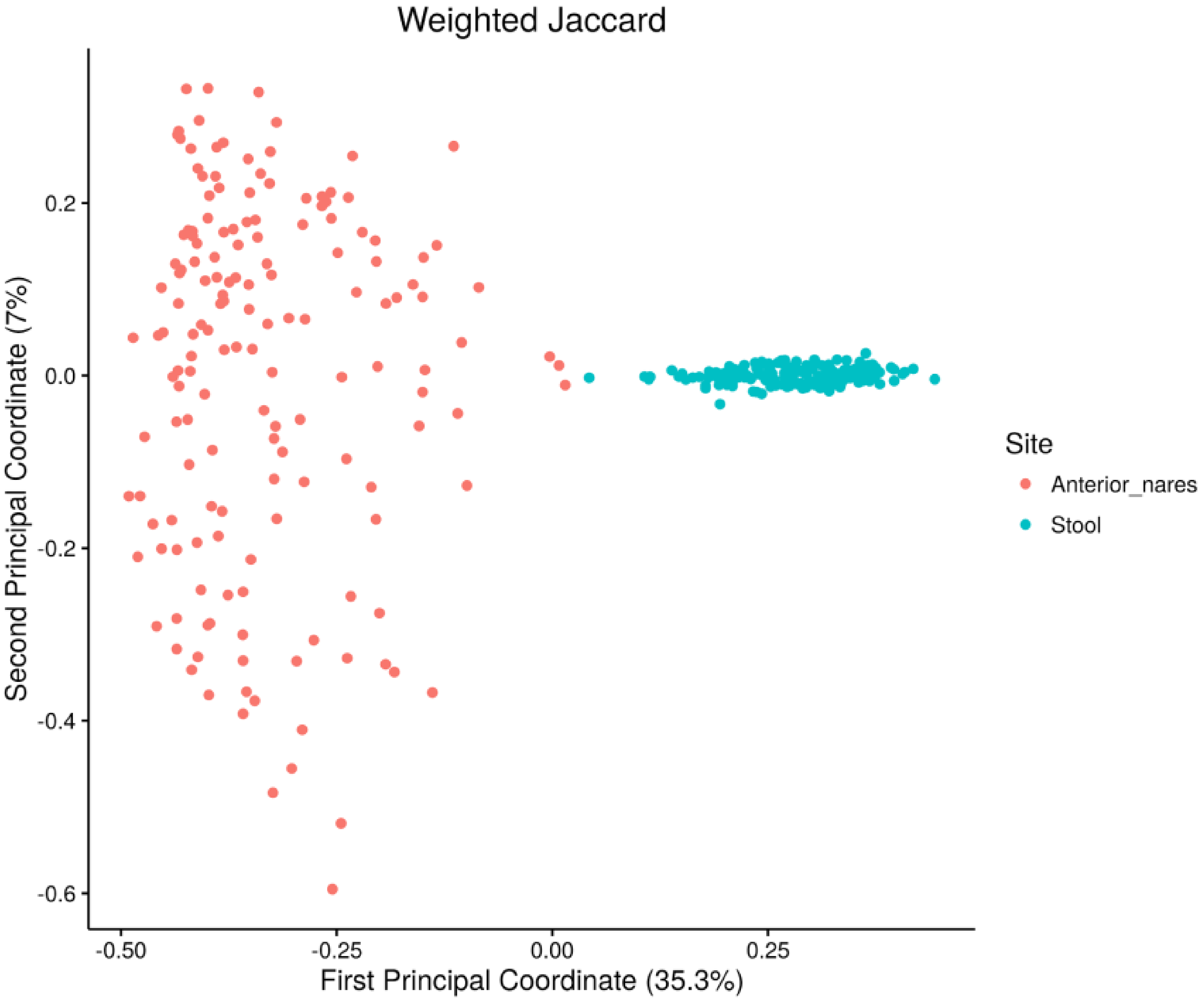
Distance matrix

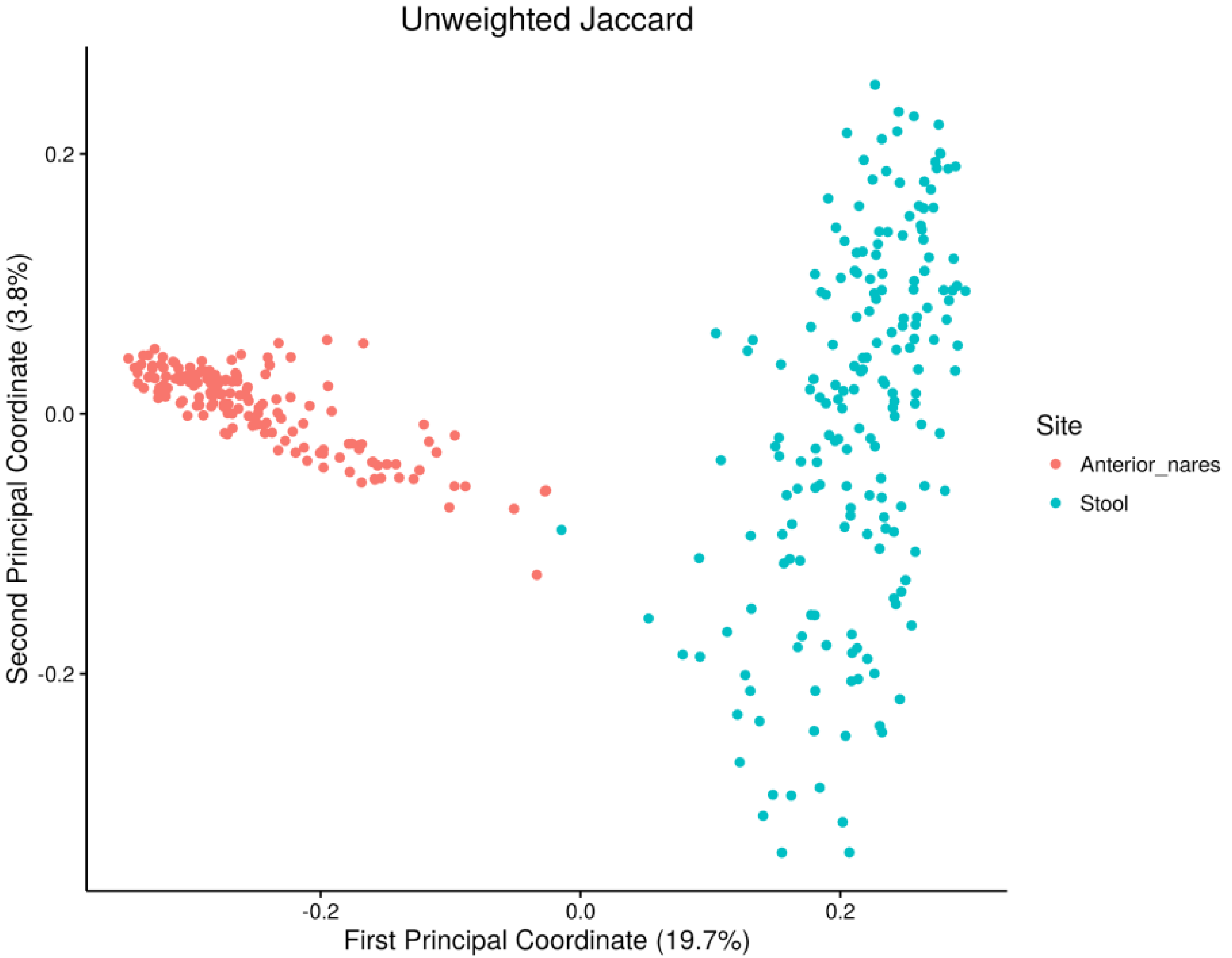
PCoA

Representation of elements:
Sites = symbols


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


Unweighted UniFrac
Second Principal Coordinate (5.7%)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.2
0.0
0.2
First Principal Coordinate (25.1%)
Site
Anterior_nares
Stool

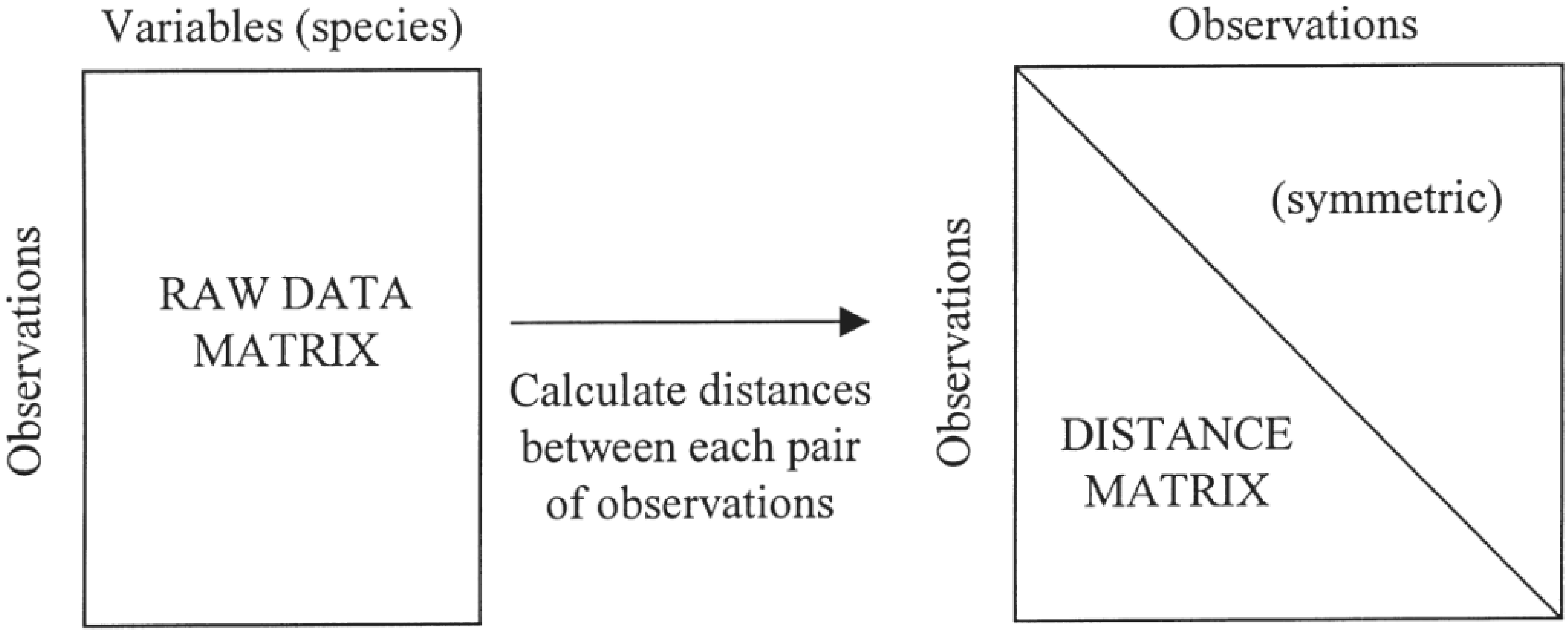

Weighted Jaccard
Second Principal Coordinate (7%)
0.2
0.0
-0.2
-0.4
-0.6
-0.50
-0.25
0.00
0.25
First Principal Coordinate (35.3%)
Site
● Anterior_nares
● Stool


Unweighted Jaccard
Second Principal Coordinate (3.8%)
0.2
0.0
-0.2
First Principal Coordinate (19.7%)
-0.2
0.0
0.2
Site
Anterior_nares
Stool

PERMANOVA

Pairwise Distances → PERMANOVA

- Pairwise distance matrix can be partitioned by group assignment and ANOVA-like analysis can be applied to detect difference between groups
- PERMANOVA: permutational ANOVA (aka, adonis)
  - pseudo F-ratio: conceptually similar but not F-distributed
  - testing by label permutation
  - quantification of effect size by R-squared or omega-squared
  - (the latter a less biased estimator of true effect)



Variables (species)

Observations

RAW DATA MATRIX

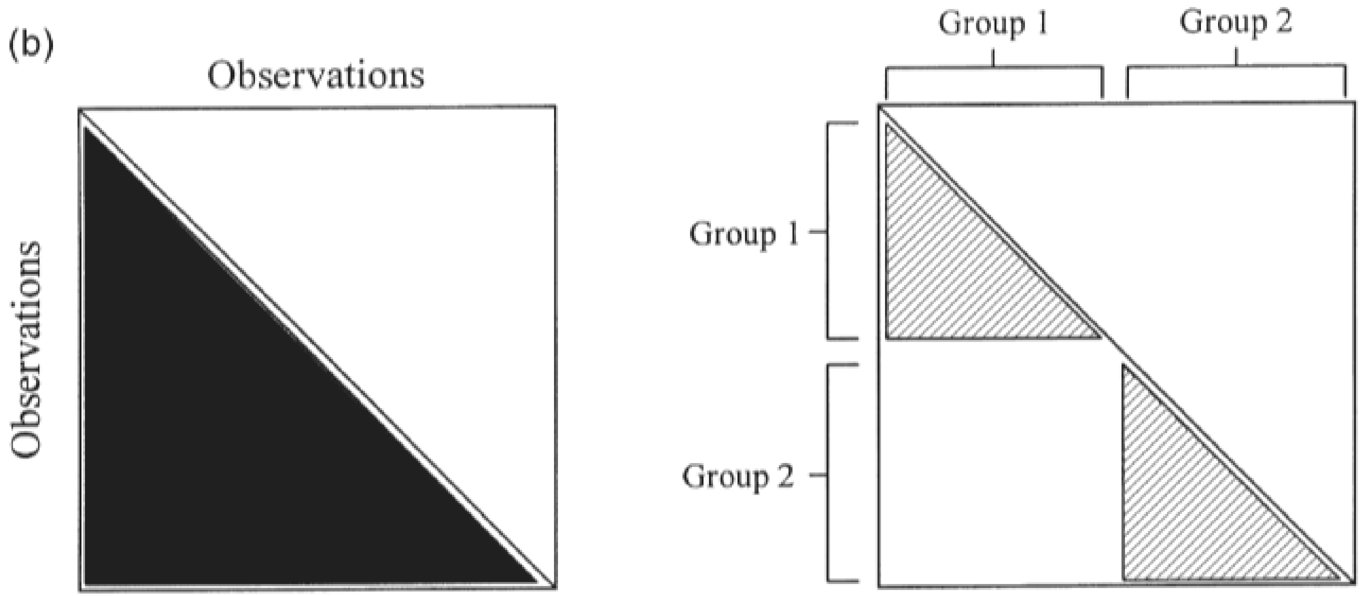
Calculate distances between each pair of observations

Observations

Observations

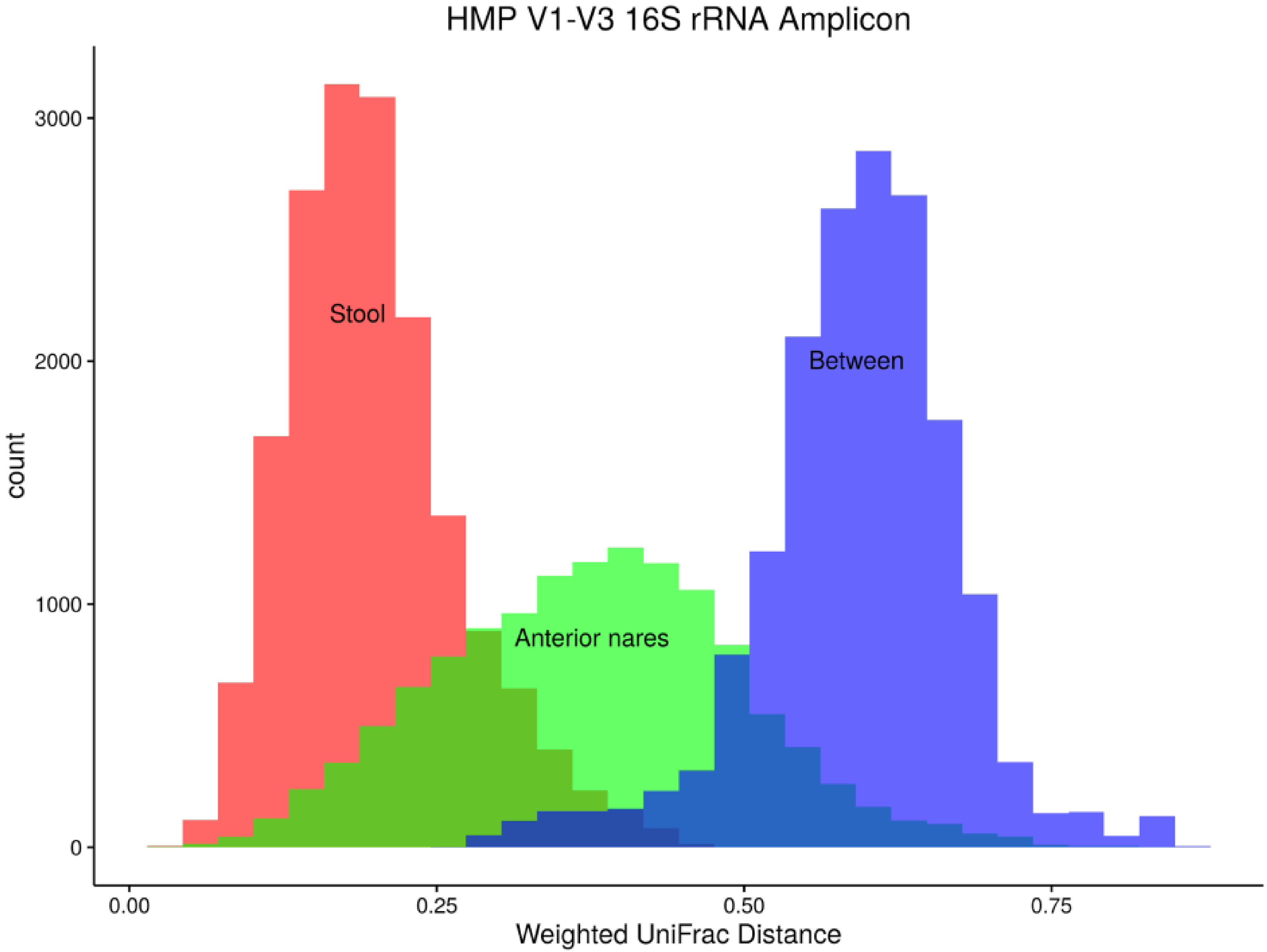
(symmetric)

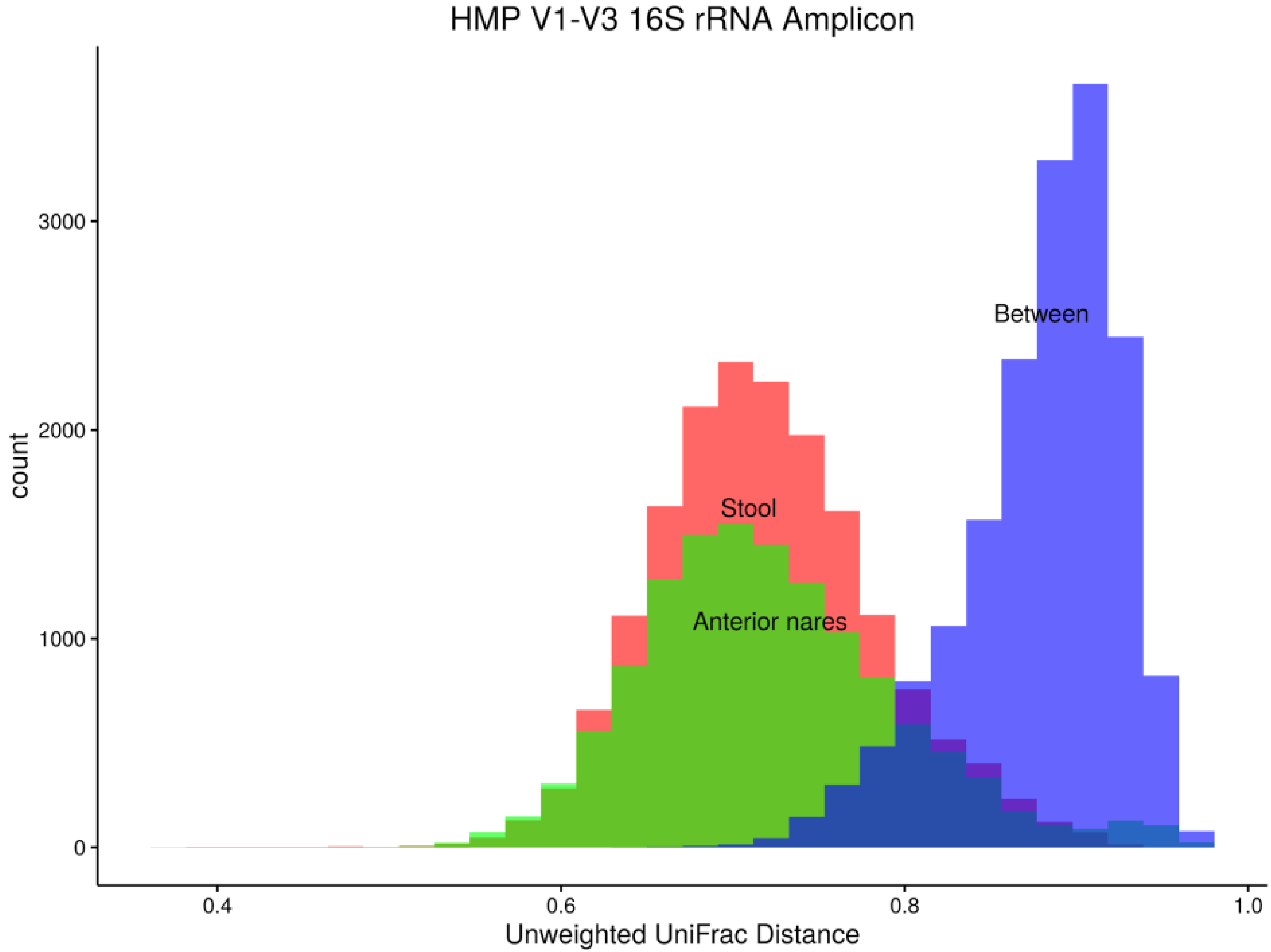
DISTANCE MATRIX

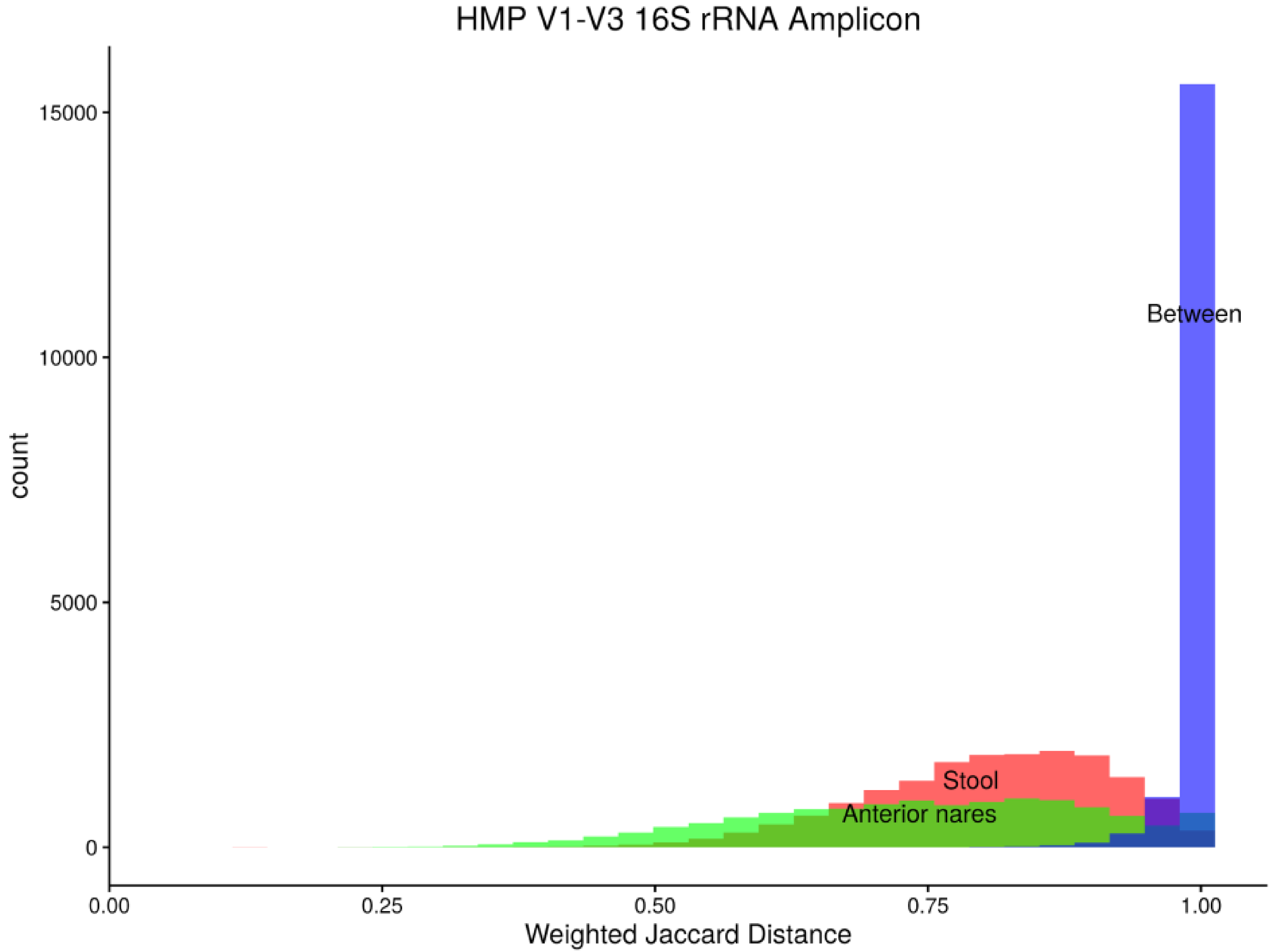


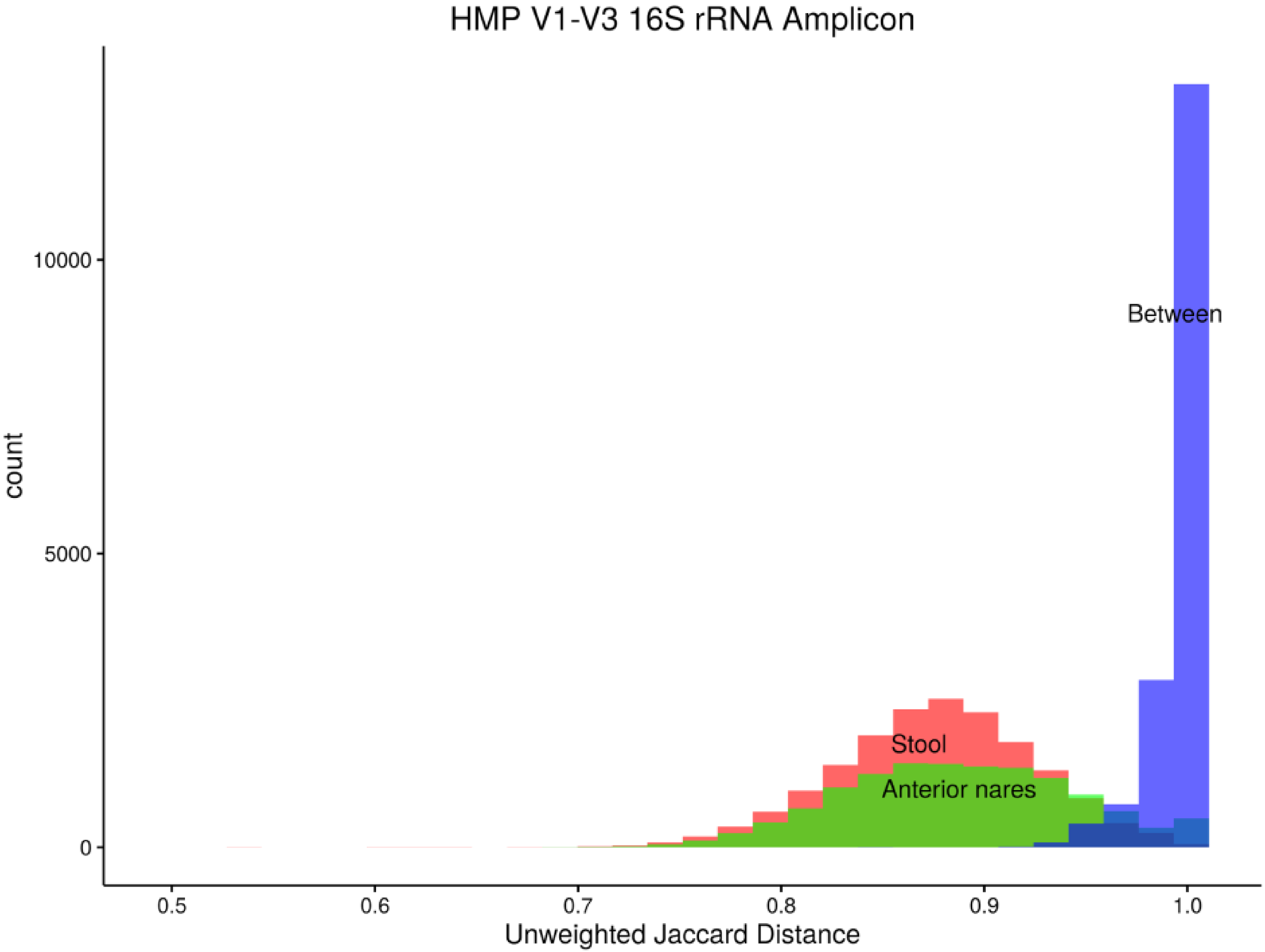
$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

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


HMP V1-V3 16S rRNA Amplicon
count
3000
2000
1000
0
Stool
Anterior nares
Between
0.00
0.25
0.50
0.75
Weighted UniFrac Distance


HMP V1-V3 16S rRNA Amplicon
count
3000
2000
1000
0
Between
Stool
Anterior nares
0.4
0.6
0.8
1.0
Unweighted UniFrac Distance


HMP V1-V3 16S rRNA Amplicon
15000
10000
5000
0
count
Between
Stool
Anterior nares
0.00
0.25
0.50
0.75
1.00
Weighted Jaccard Distance


HMP V1-V3 16S rRNA Amplicon
count
10000
5000
0
Between
Stool
Anterior nares
0.5
0.6
0.7
0.8
0.9
1.0
Unweighted Jaccard Distance

$$R^2 = 1 - \frac{SS_W}{SS_W + SS_A} = \frac{SS_A}{SS_T}$$

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

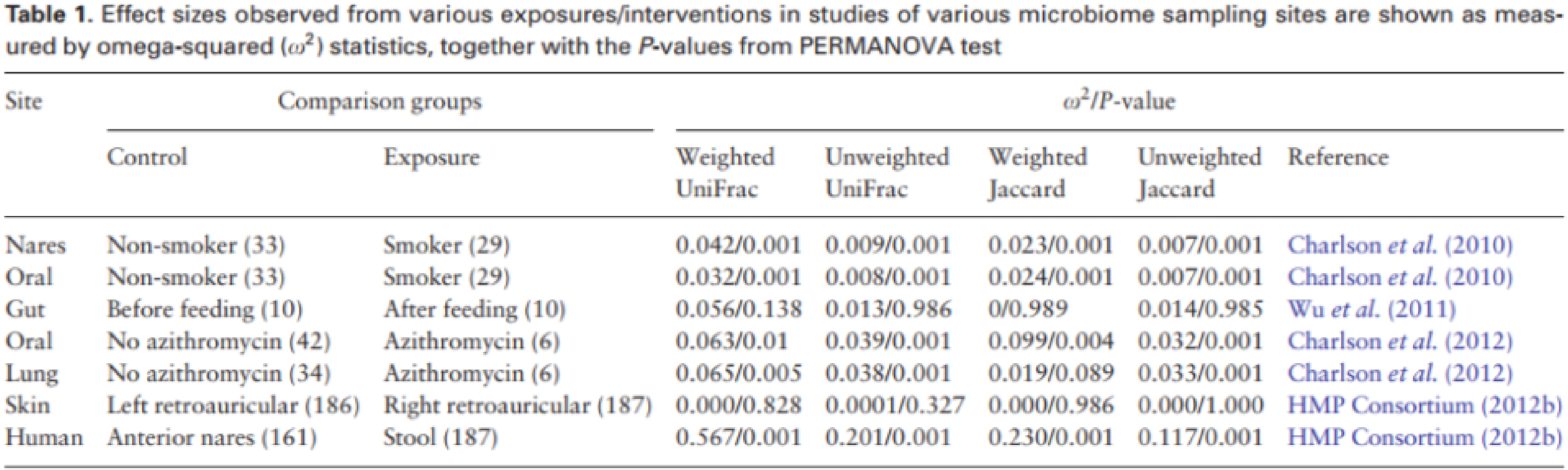


Table 1. Effect sizes observed from various exposures/interventions in studies of various microbiome sampling sites are shown as measured by omega-squared (ω^2) statistics, together with the *P*-values from PERMANOVA test

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.

$$R^2 = 1 - \frac{\mathrm{SS_W}}{\mathrm{SS_W} + \mathrm{SS_A}} = \frac{\mathrm{SS_A}}{\mathrm{SS_T}}$$

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

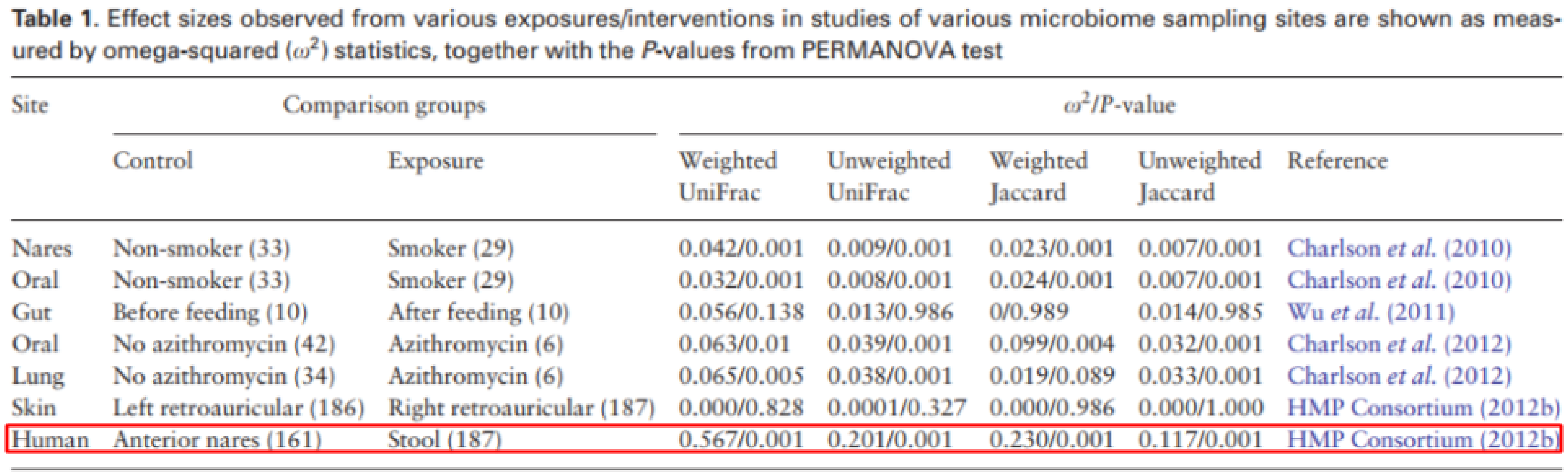
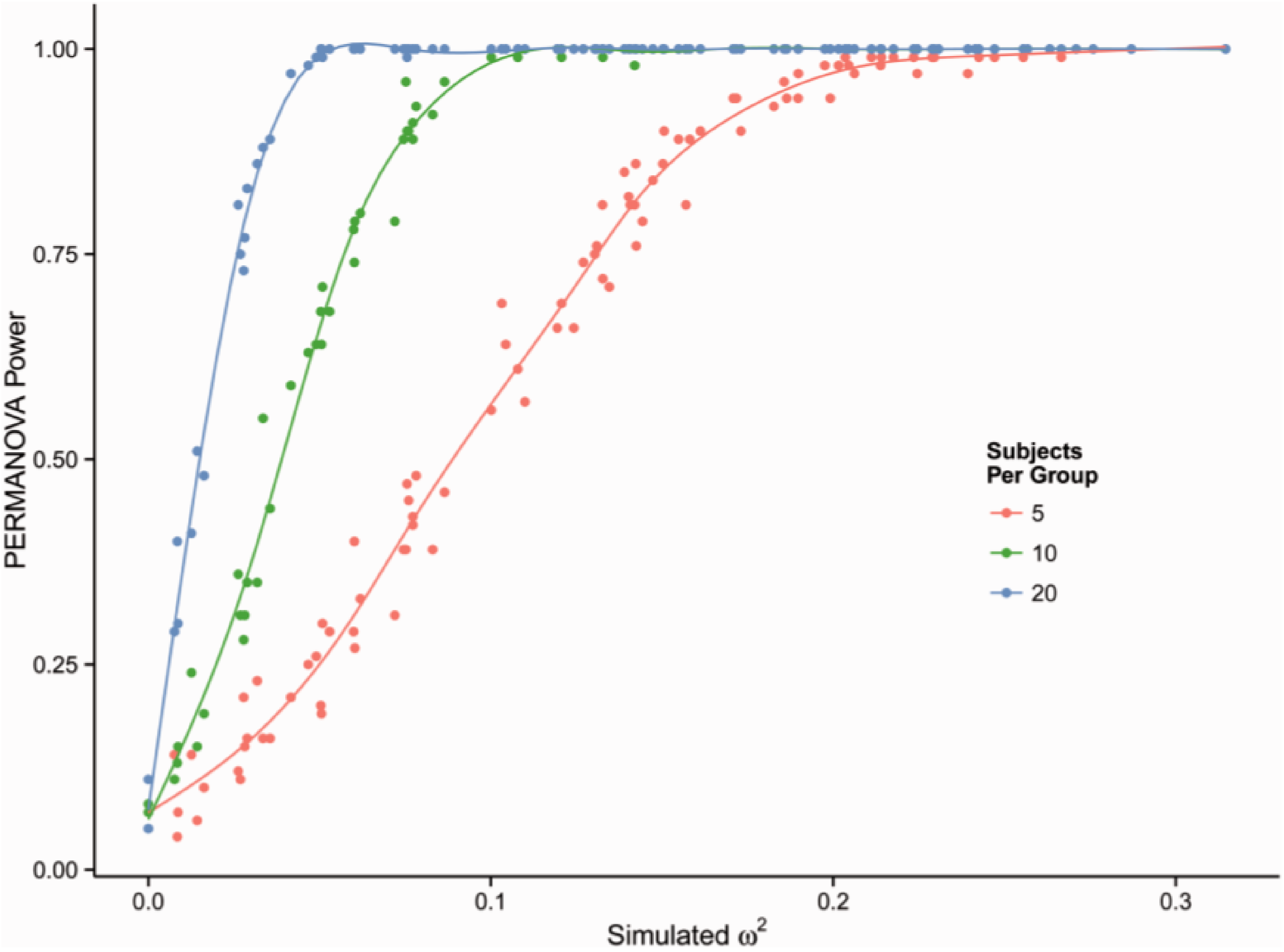


Table 1. Effect sizes observed from various exposures/interventions in studies of various microbiome sampling sites are shown as measured by omega-squared (ω^2) statistics, together with the *P*-values from PERMANOVA test

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.



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

R's vegan package

vegan::vegdist()

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

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

otu_long <- read_csv(
  "./data/HMP_OTU_table_longformat_stool_nares.c
)

otu_long
```

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

vegan::vegdist()

```
otu_matrix <- read_rds(
  "./data/HMP_OTU_table_matrix_stool_nares.rds"
)

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

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

vegan::vegdist()

```r
otu_matrix <- read_rds(
  "./data/HMP_OTU_table_matrix_stool_nares.rds"
)

otu_matrix %>%
  t() %>% # TRANSPOSE
  str(vec.len = 2)
```

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

vegan::vegdist()

```r
otu_matrix <- read_rds(
  "./data/HMP_OTU_table_matrix_stool_nares.rds"
)

otu_matrix %>%
  t() %>% #TRANSPOSE
  vegdist(x = .,
          method = "jaccard",
          binary = TRUE) %>%
  str(vec.len=2)
```

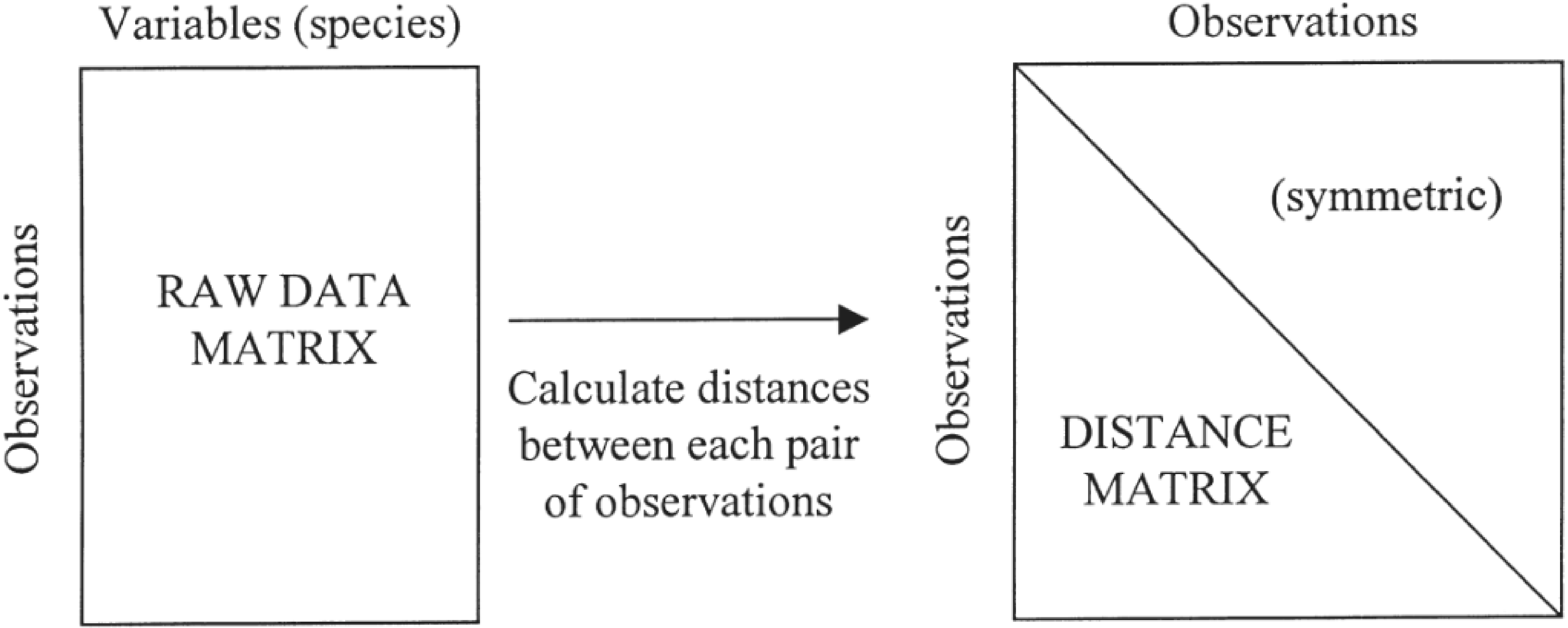
```
##  'dist' num [1:45] 1 0.982 ...
##  - attr(*, "Size")= int 10
##  - attr(*, "Labels")= chr [1:10] "700014718" "700014767
##  - attr(*, "Diag")= logi FALSE
##  - attr(*, "Upper")= logi FALSE
##  - attr(*, "method")= chr "binary jaccard"
##  - attr(*, "call")= language vegdist(x = ., method = "j
```

vegan::vegdist()

```
otu_matrix <- read_rds(
  "./data/HMP_OTU_table_matrix_stool_nares.rds"
)

otu_matrix %>%
  t() %>% #TRANSPOSE
  vegdist(x = .,
          method = "jaccard",
          binary = TRUE) %>%
  as.matrix() %>%
  str(vec.len=2)
```

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



Variables (species)

Observations

RAW DATA MATRIX

Calculate distances between each pair of observations

Observations

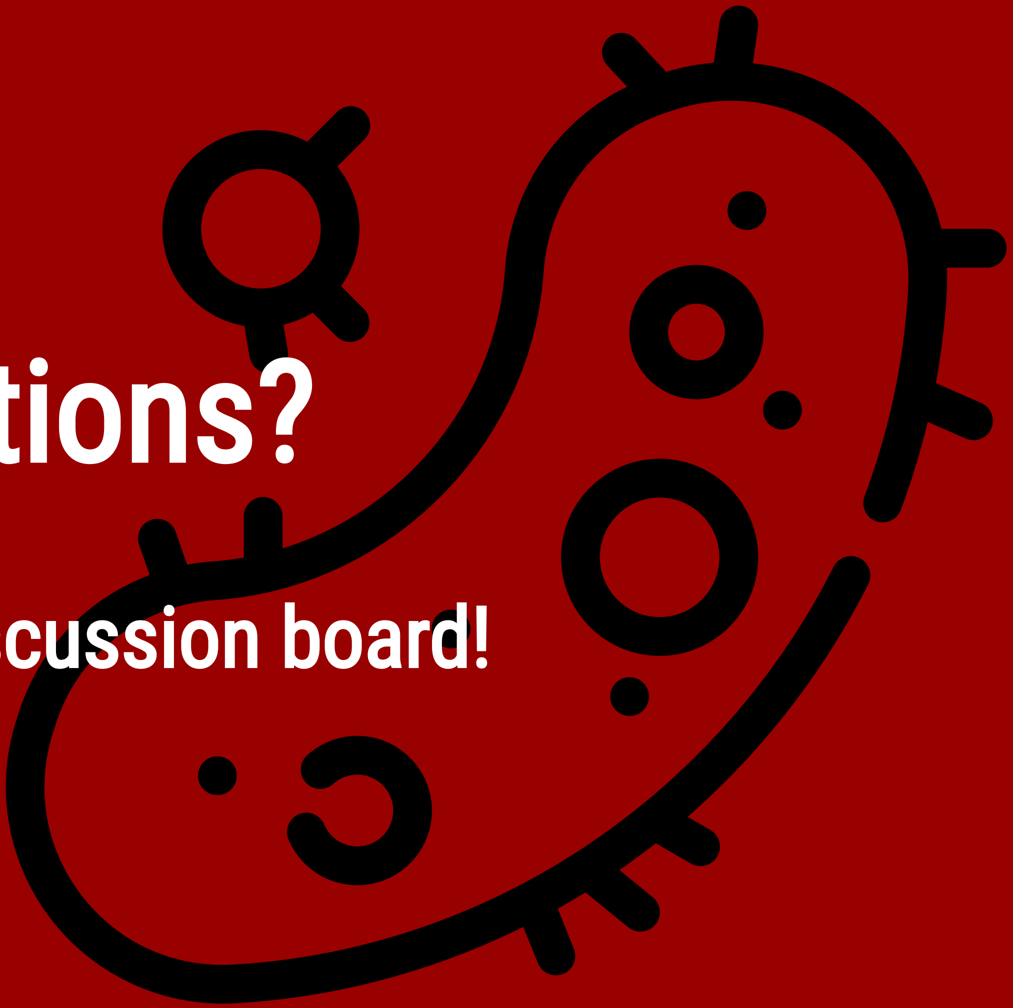
Observations

(symmetric)

DISTANCE MATRIX

Questions?

Post to the discussion board!



Thank you!

Slides available: github.com/bjklab

brendank@pennmedicine.upenn.edu