Distances, Trees, and Types: Methods for Analyzing High-Dimensional Microbiome Data

Brendan J. Kelly, MD, MSCE
Division of Infectious Diseases
University of Pennsylvania
2 June 2022

Outline

- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

Outline

- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

Generating Microbiome Data

- High-density (next-generation, high-throughput) sequencing:
  - "tag" gene with conserved and variable regions (16S, 18S, ITS)
  - "shotgun" metagenomics (pool of randomly amplified nucleic acid)
- Sequence binning and assignment:
  - operational taxonomic units (OTUs) based on 97% sequence similarity

    → taxonomic assignment of OTUs
  - assemble contiguous metagenomic sequences → taxonomy
  - unassembled reads → taxonomic assignment

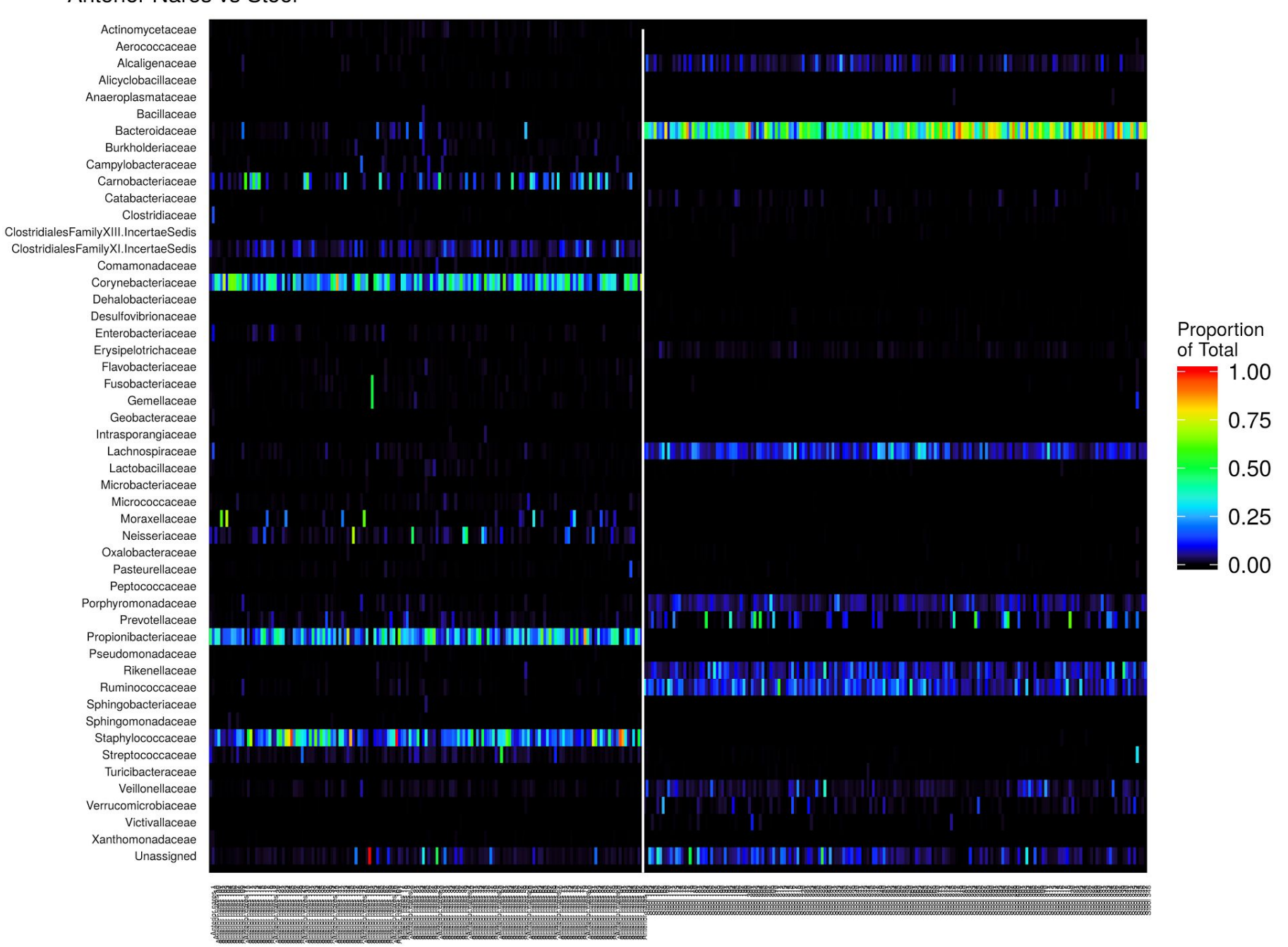
Callahan BJ bioArxiv 2017.

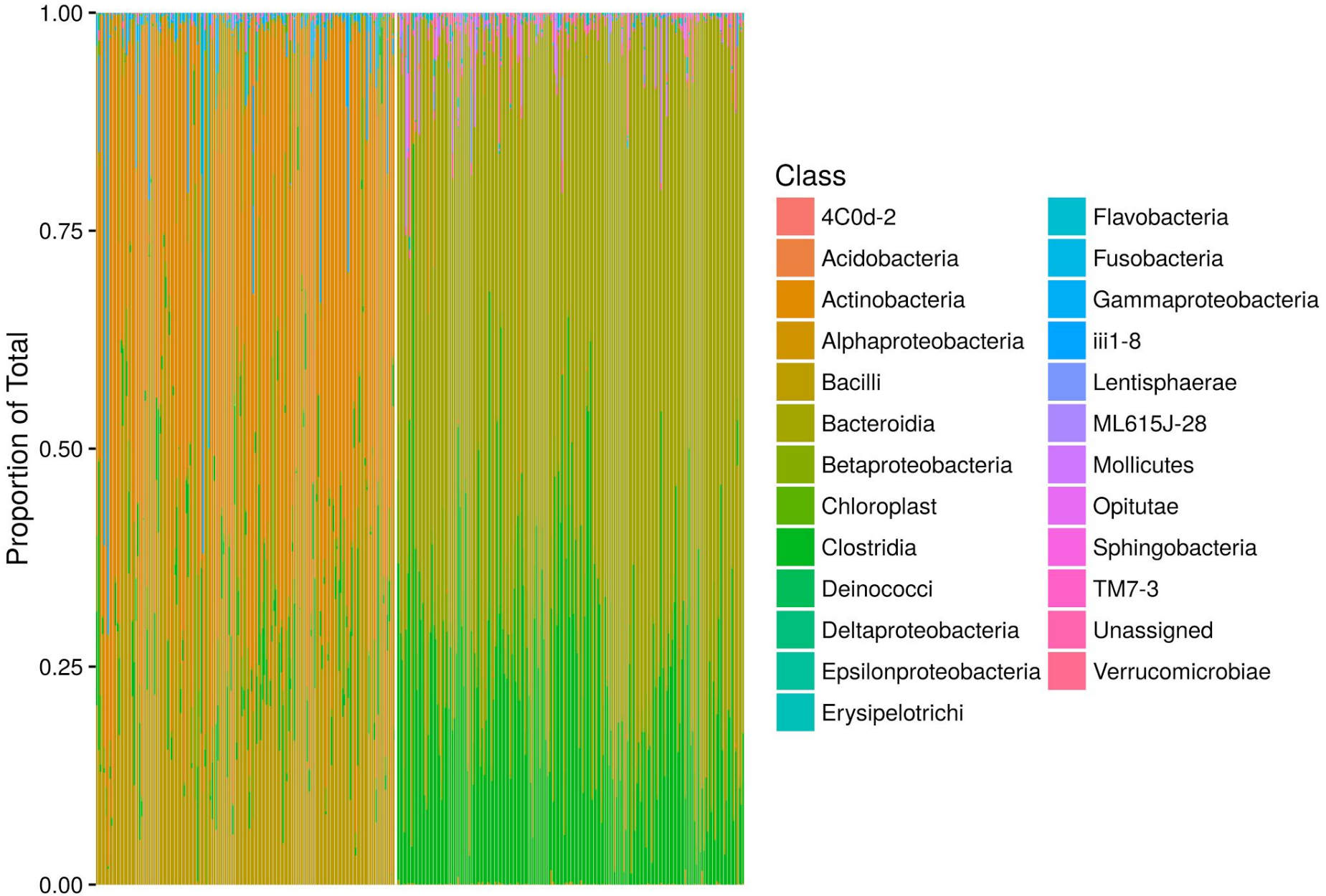
	Stool_1	Stool_2	Stool_3	Stool_4	Stool_5	Stool_6	Stool_7	Stool_8	Stool_9	Stool_10	Stool_11	Stool_12	Stool_13	Stool_14	Stool_15	Stool_16
OTU_97.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10002	0	0	0	0	0	0	0	0	0	4	0	0	0	0	9	0
OTU_97.10005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10008	0	1	0	0	0	0	1	0	1	5	3	1	0	0	1	0
OTU_97.10010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10029	4	12	0	0	0	0	1	0	0	0	0	1	0	6	3	0
OTU_97.10034	0	3	0	0	0	0	0	0	0	1	0	2	0	0	3	0
OTU_97.10035	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0
OTU_97.10037	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
OTU_97.10039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10048	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10057	0	0	2	0	0	0	0	0	0	0	0	1	0	5	0	1
OTU_97.10058	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.10059	0	0	0	0	0	0	4	0	0	0	0	0	0	12	0	0
8048 ROWS - 187 COLUMNS																
OTU_97.9518	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
OTU_97.9523	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0
OTU_97.9535	0	0	0	0	0	0	1	0	0	0	107	0	2	3	12	0
OTU_97.9540	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.9546	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
OTU_97.956	0	0	2	12	0	7	0	0	0	0	0	7	1	0	0	0
OTU_97.957	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.9572	0	1	0	1	0	0	4	0	0	3	0	0	0	0	0	0
OTU_97.9583	0	0	0	0	2	0	0	1	0	0	0	11	0	6	0	0
OTU_97.9588	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.9591	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0
OTU_97.9593	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.9594	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_97.96	0	11	0	0	0	0	0	0	0	0	0	0	0	1	0	0
OTU_97.9600	0	0	0	2	0	0	1	0	0	0	2	0	0	0	0	0
OTU_97.9605	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	OTU_97.10	OTU_97.10002	OTU_97.10005	OTU_97.10008	OTU_97.10010	OTU_97.10016	OTU_97.10018	OTU_97.10029	OTU_97.10034	OTU_97.10035	OTU_97.10037	OTU_97.10039	OTU_97.10048	OTU_97.10057	OTU_97.10058	OTU_97.10059	**187 ROWS - 8048 COLS**	OTU_97.9518	OTU_97.9523	OTU_97.9535	OTU_97.9540	OTU_97.9546	OTU_97.956	OTU_97.957	OTU_97.9572	OTU_97.9583	OTU_97.9588
Stool_1	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
Stool_2	0	0	0	1	0	0	0	12	3	0	0	0	1	0	0	0		0	0	0	0	0	0	0	1	0	0
Stool_3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0		0	0	0	0	1	2	0	0	0	0
Stool_4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	12	0	1	0	0
Stool_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	2	0
Stool_6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	1	7	0	0	0	0
Stool_7	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	4		0	0	1	0	0	0	0	4	0	0
Stool_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	1	0
Stool_9	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
Stool_10	0	4	0	5	0	0	0	0	1	0	0	0	0	0	0	0		0	0	0	0	0	0	0	3	0	0
Stool_11	0	0	0	3	0	0	0	0	0	2	0	0	0	0	0	0		1	0	107	0	0	0	0	0	0	0
Stool_12	0	0	0	1	0	0	0	1	2	0	2	0	0	1	0	0		0	8	0	0	0	7	0	0	11	0
Stool_13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	2	0	0	1	0	0	0	0
Stool_14	0	0	0	0	0	0	0	6	0	0	0	0	0	5	0	12		0	0	3	0	0	0	0	0	6	0
Stool_15	0	9	0	1	0	0	0	3	3	2	0	0	0	0	0	0		0	0	12	0	0	0	0	0	0	0
Stool_16	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		0	0	0	0	0	0	0	0	0	0

High-Dimensional Microbiome Data

- Descriptive:
  - heatmaps
  - 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


1.00
0.75
0.50
0.25
0.00
Proportion of Total
Class
4C0d-2
Acidobacteria
Actinobacteria
Alphaproteobacteria
Bacilli
Bacteroidia
Betaproteobacteria
Chloroplast
Clostridia
Deinococci
Deltaproteobacteria
Epsilonproteobacteria
Erysipelotrichi
Flavobacteria
Fusobacteria
Gammaproteobacteria
iii1-8
Lentisphaerae
ML615J-28
Mollicutes
Opitutae
Sphingobacteria
TM7-3
Unassigned
Verrucomicrobiae

Descriptive: Heatmaps & Barplots

- Visualization of OTU table:
  - typically present counts as a proportion of sample total
  - choice of sample order can highlight group differences
- Limitations:
  - cannot depict full list of OTUs
  - space dictates taxonomic level presented

Single-Taxon Hypotheses

- You suspect *Bacteroides* has a relationship with outcome of interest...
  - *Bacteroides* (genus)?
  - *Bacteroidaceae* (family)?
  - *Bacteroidales* (order)?
  - *Bacteroidetes* (class)?
- Hypotheses focusing on specific taxa often fail to account for possibility of selection bias from culture.

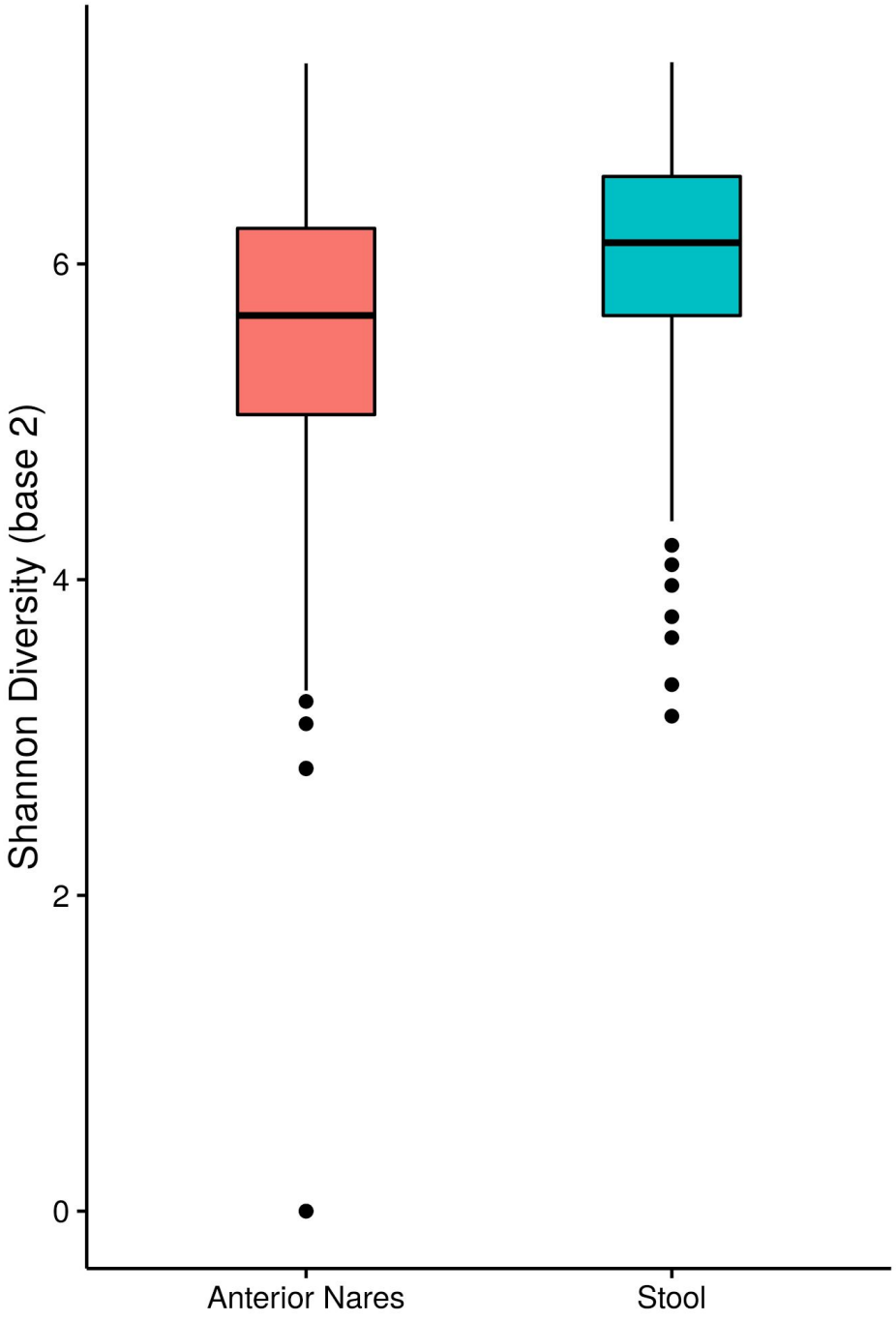
OTU	assignment	phylum	class	order	family	genus
OTU_97.1	Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
OTU_97.10	Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Veillonella	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Veillonella
OTU_97.100	Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	NA
OTU_97.1000	Root;p__Proteobacteria;c__Betaproteobacteria	Proteobacteria	Betaproteobacteria	NA	NA	NA
OTU_97.10000	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium
OTU_97.10001	Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
OTU_97.10002	Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	NA
OTU_97.10003	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium
OTU_97.10004	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium
OTU_97.10005	Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides

Outline

- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

Dimension Reduction: Alpha Diversity

- Summarize each sample's community in a single measure:
  - richness: number of community members
  - evenness: the distribution of member counts
- Many alpha diversity metrics (weight richness/evenness):
  - species number, Chao1 (singletons & doubletons)
  - Shannon diversity: $H' = -\sum_i p_i \log_b p_i$
  - (note: may measure similarity or dissimilarity)

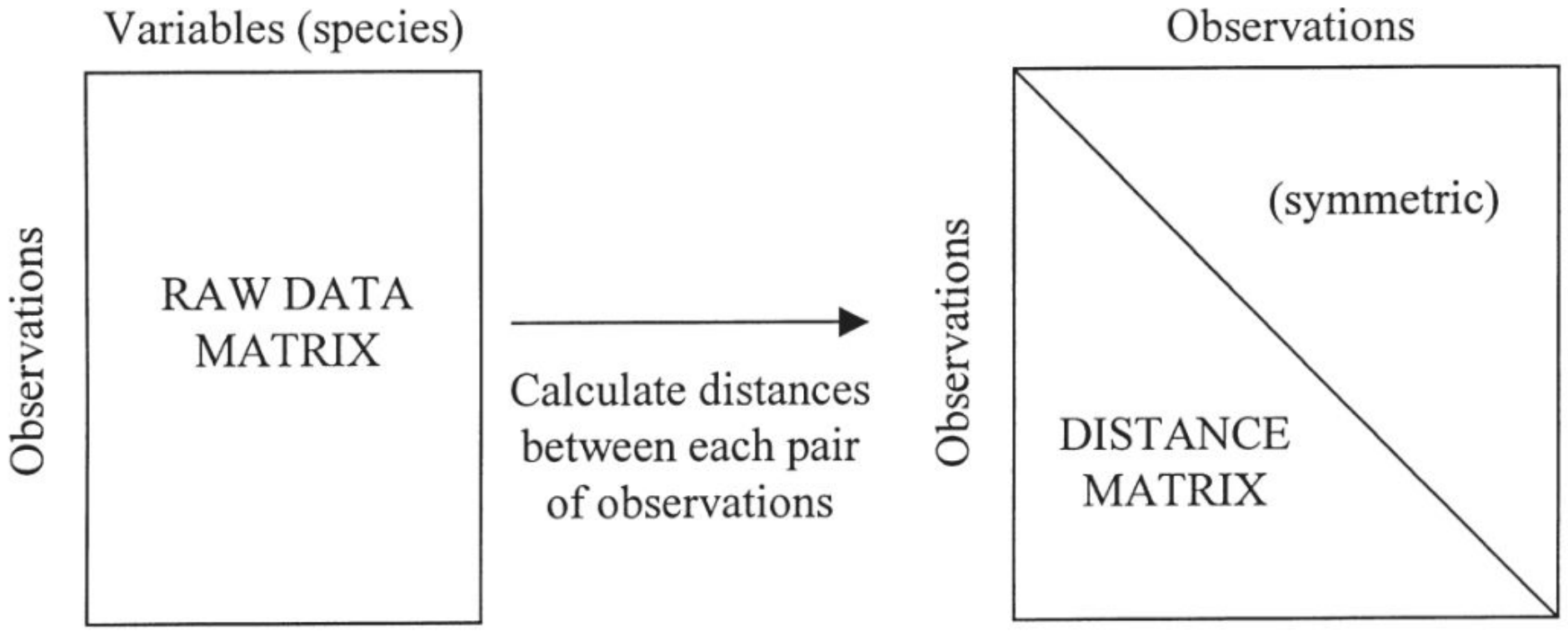

Shannon Diversity (base 2)
0
2
4
6
Anterior Nares
Stool

Outline

- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

Dimension Reduction: Beta Diversity

- 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
  - (Euclidean versus non-Euclidean)

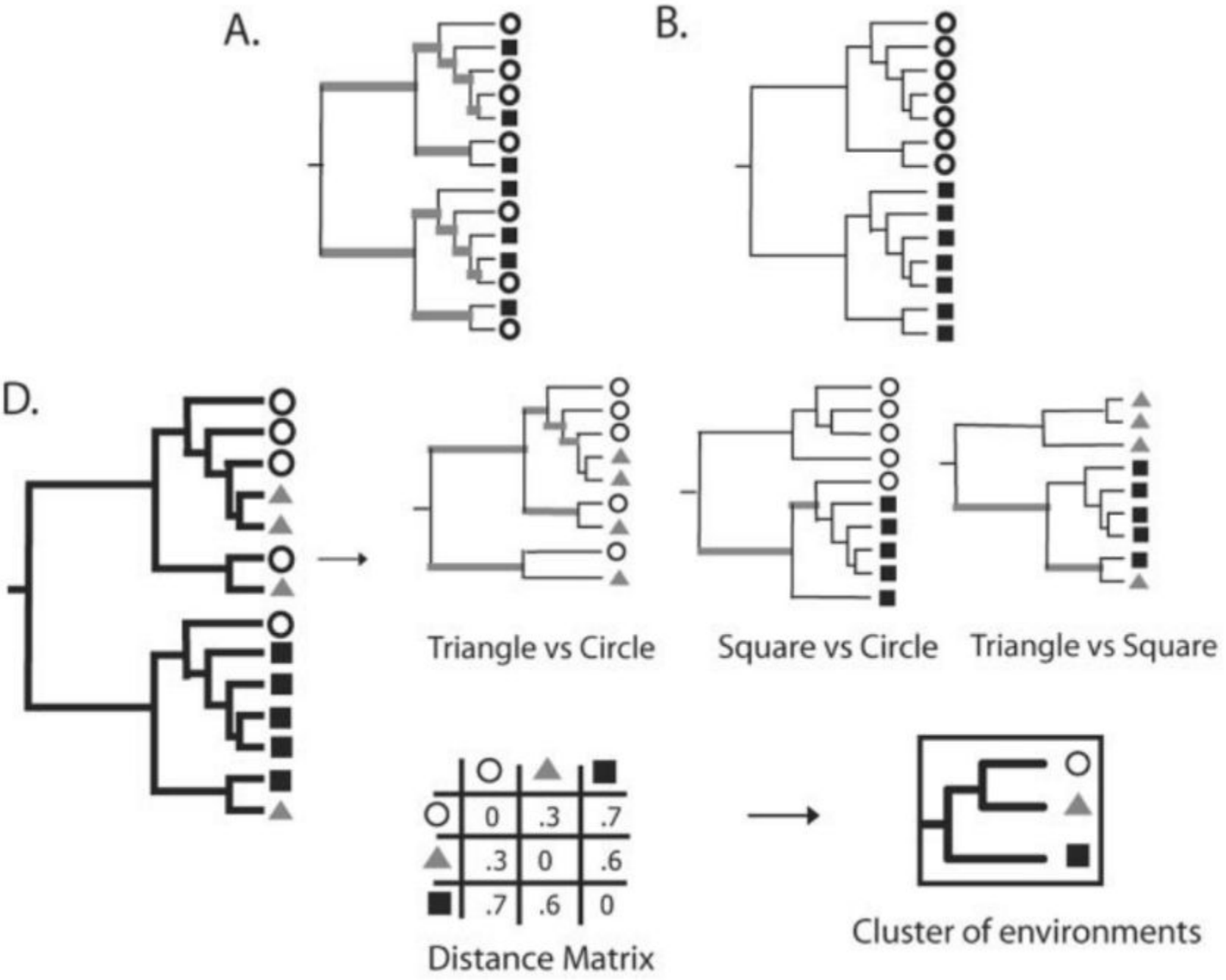


Anderson MJ et al. Austral Ecol 2001;26(1):32-46.

Dimension Reduction: Beta Diversity

- Just counts versus counts + phylogeny:
  - Jaccard: $$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

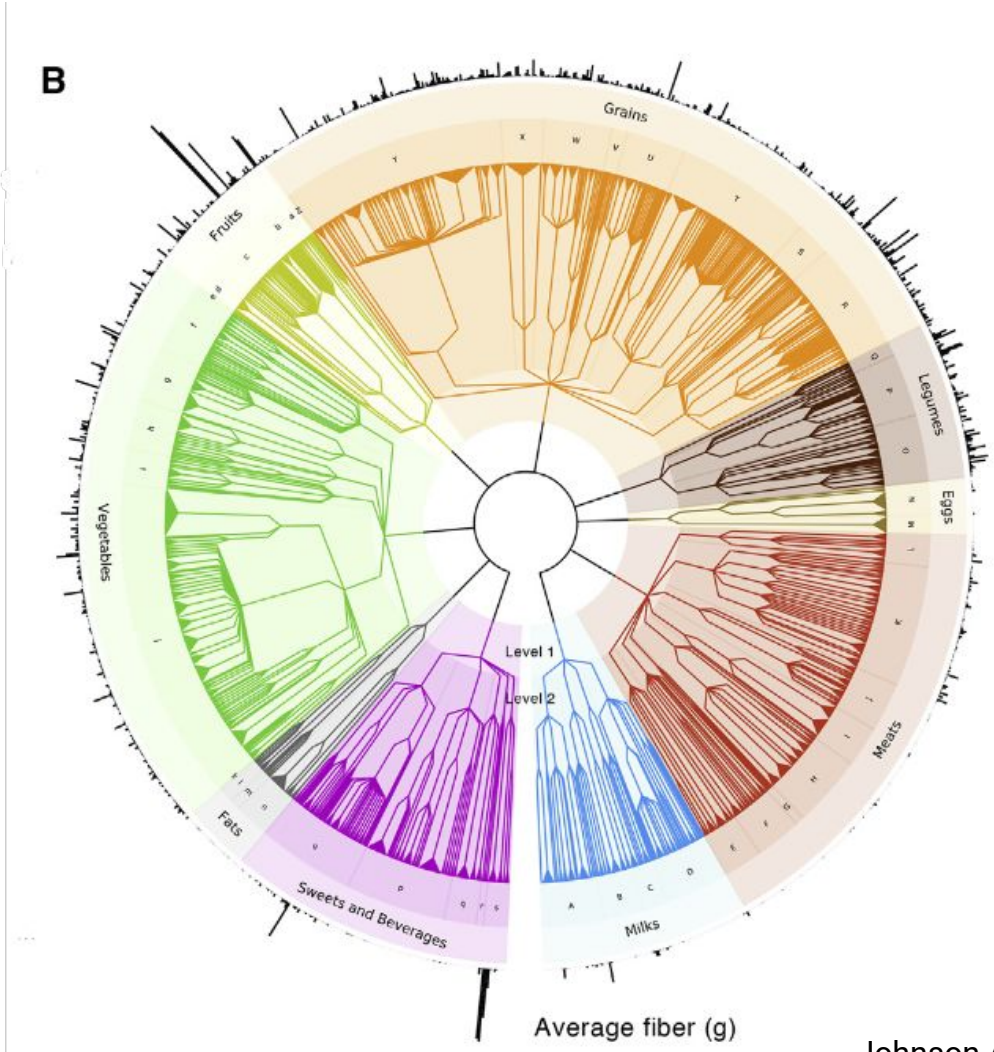
    $$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$
  - UniFrac: fraction of unique branch length in tree
- Weighted versus unweighted:
  - weighted: counts matter
  - unweighted: binary (presence-absence)


A.
B.
D.
Triangle vs Circle
Square vs Circle
Triangle vs Square
0 .3 .7
.3 0 .6
.7 .6 0
Distance Matrix
Cluster of environments

Beta Diversity: How to Choose?

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

Thinking Like a Tree



Johnson AJ et al. Cell Host Microbe 2019.

Outline

- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

Original Descriptors: 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

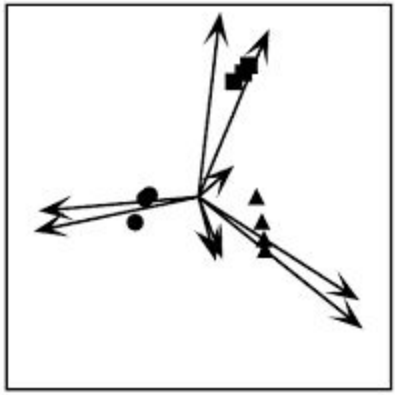
(a) Classical approach

Y = Raw data (sites × species)

Short gradients: CA or PCA

Long gradients: CA

Ordination biplot



Representation of elements:
Species = arrows
Sites = symbols

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

Raw data (sites × species)

Y=Transformed data (sites × species)

PCA

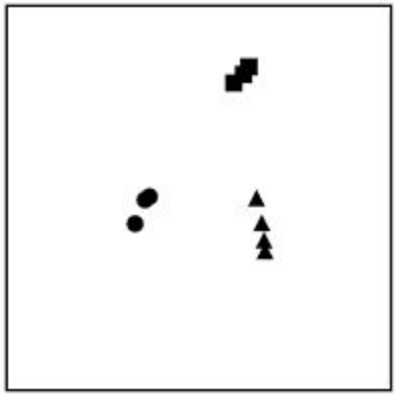
(c) Distance-based approach (PCoA)

Raw data (sites × species)

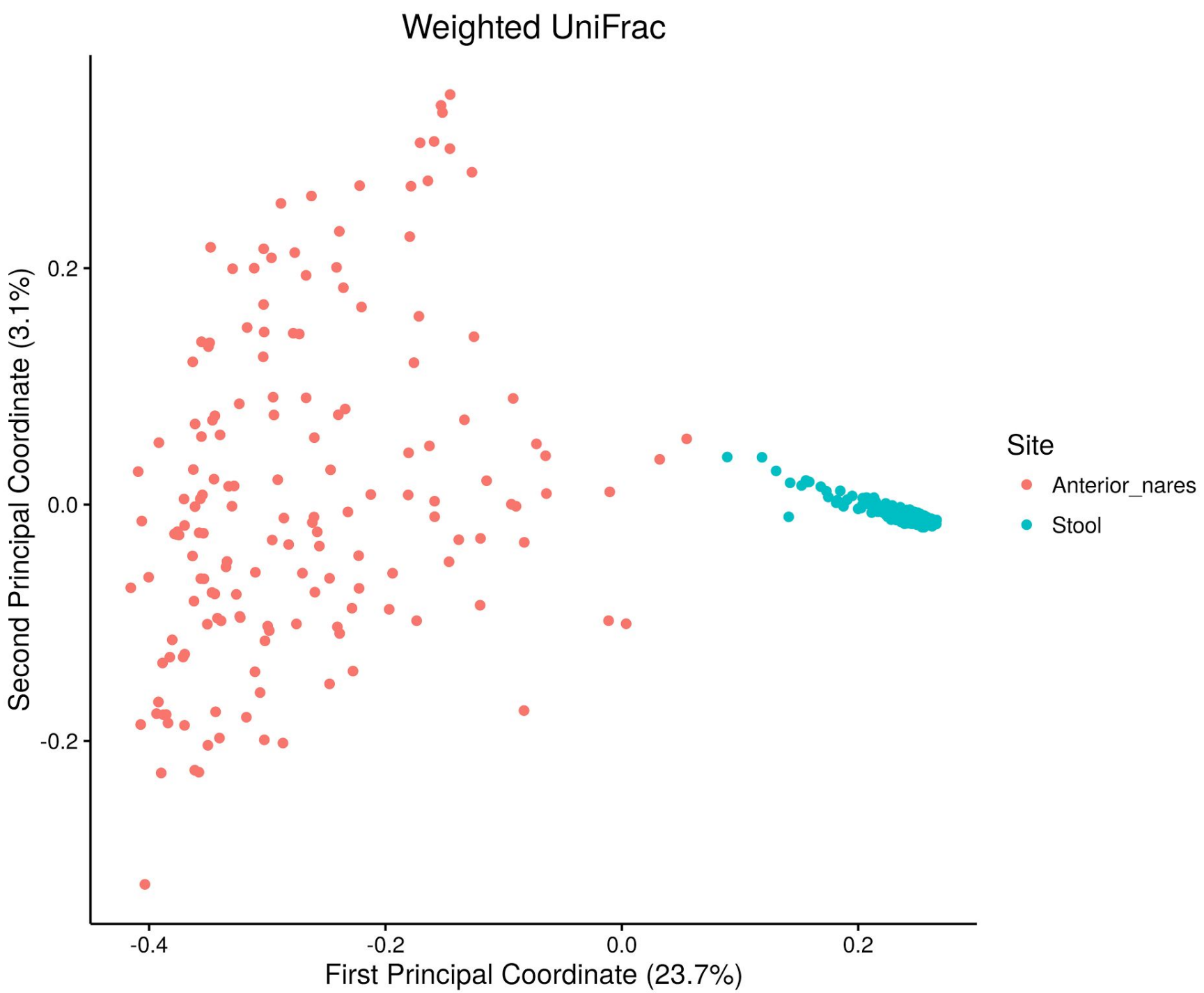
Distance matrix

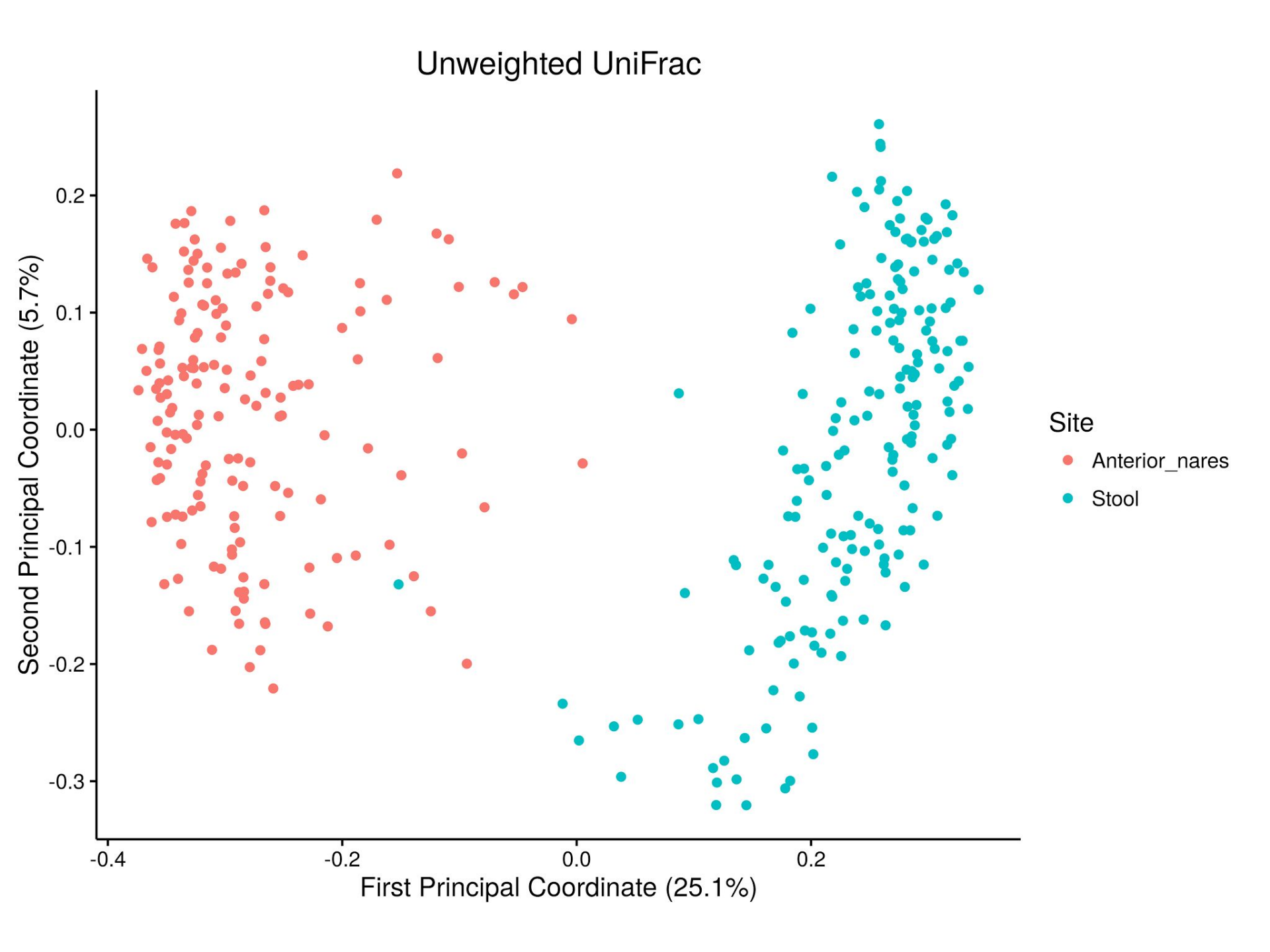
PCoA

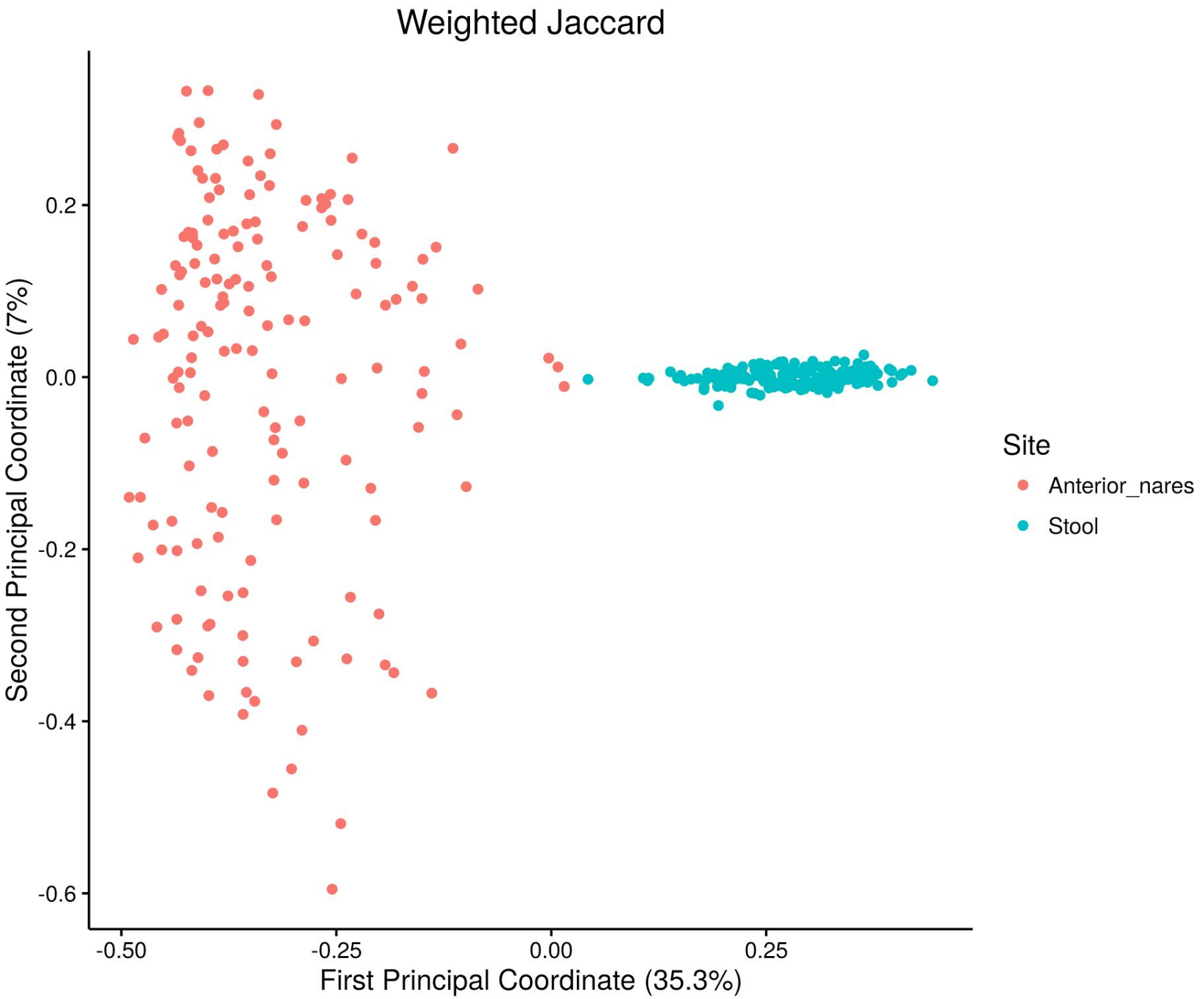
Ordination of sites

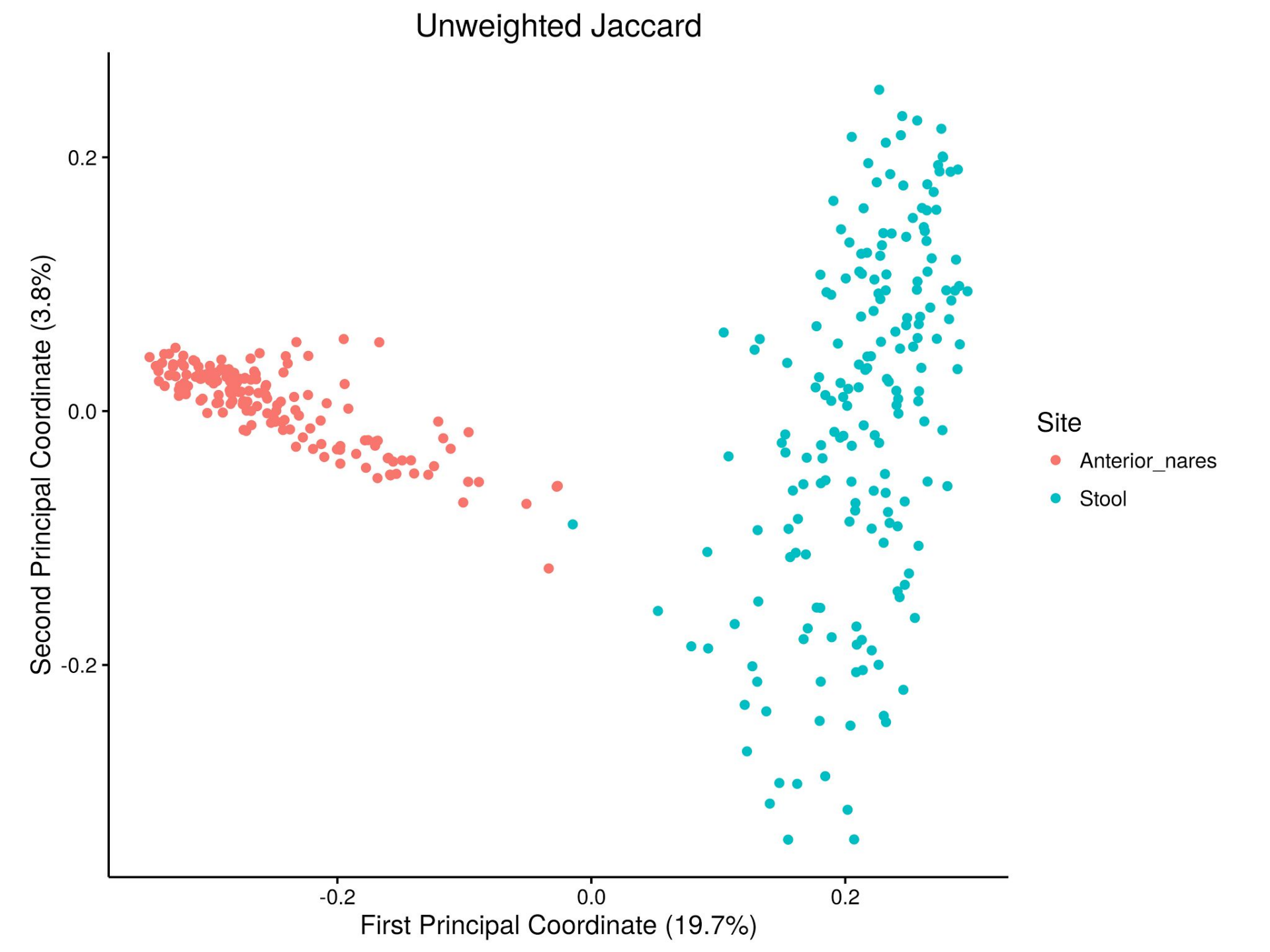


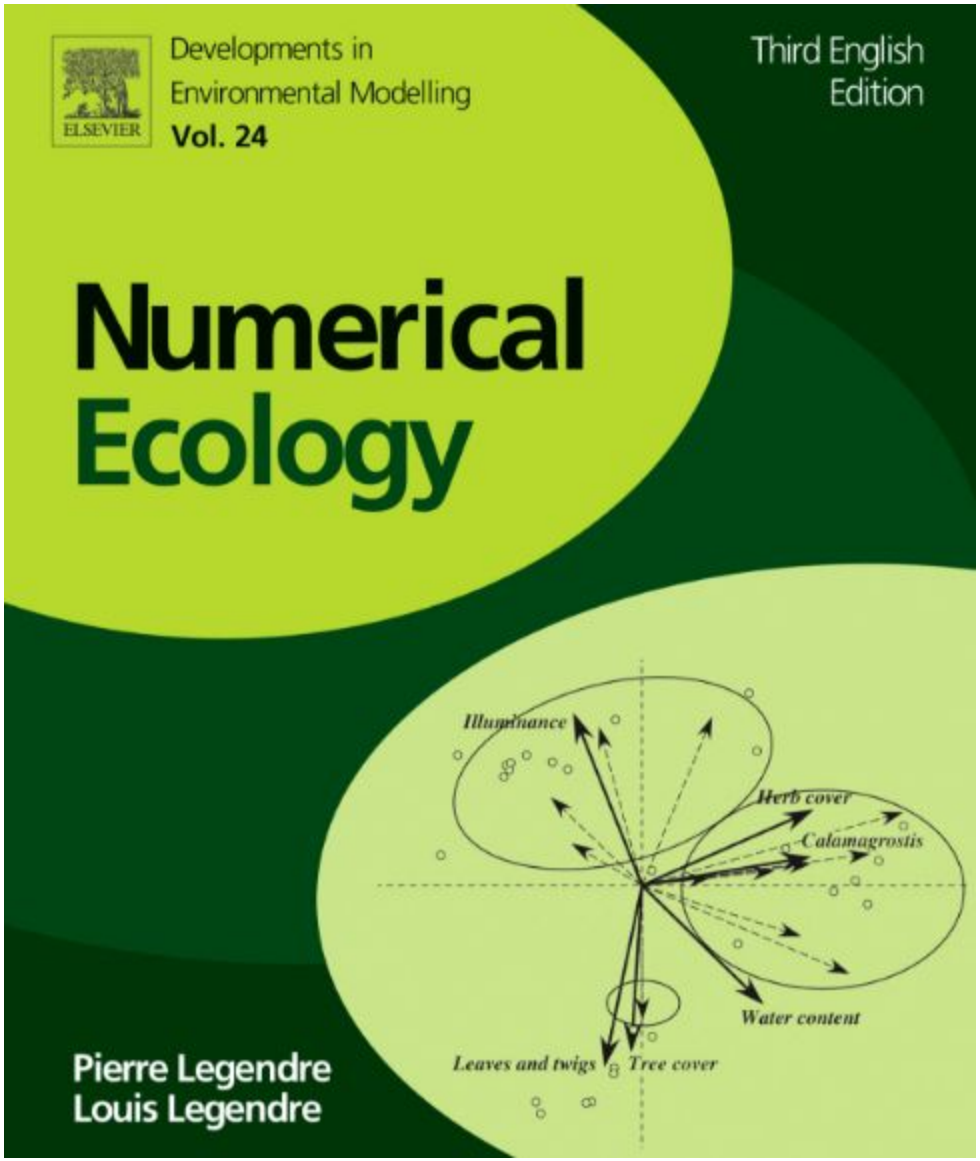
Representation of elements:
Sites = symbols


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%)
First Principal Coordinate (25.1%)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.2
0.0
0.2
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
-0.2
0.0
0.2
First Principal Coordinate (19.7%)
Site
Anterior_nares
Stool

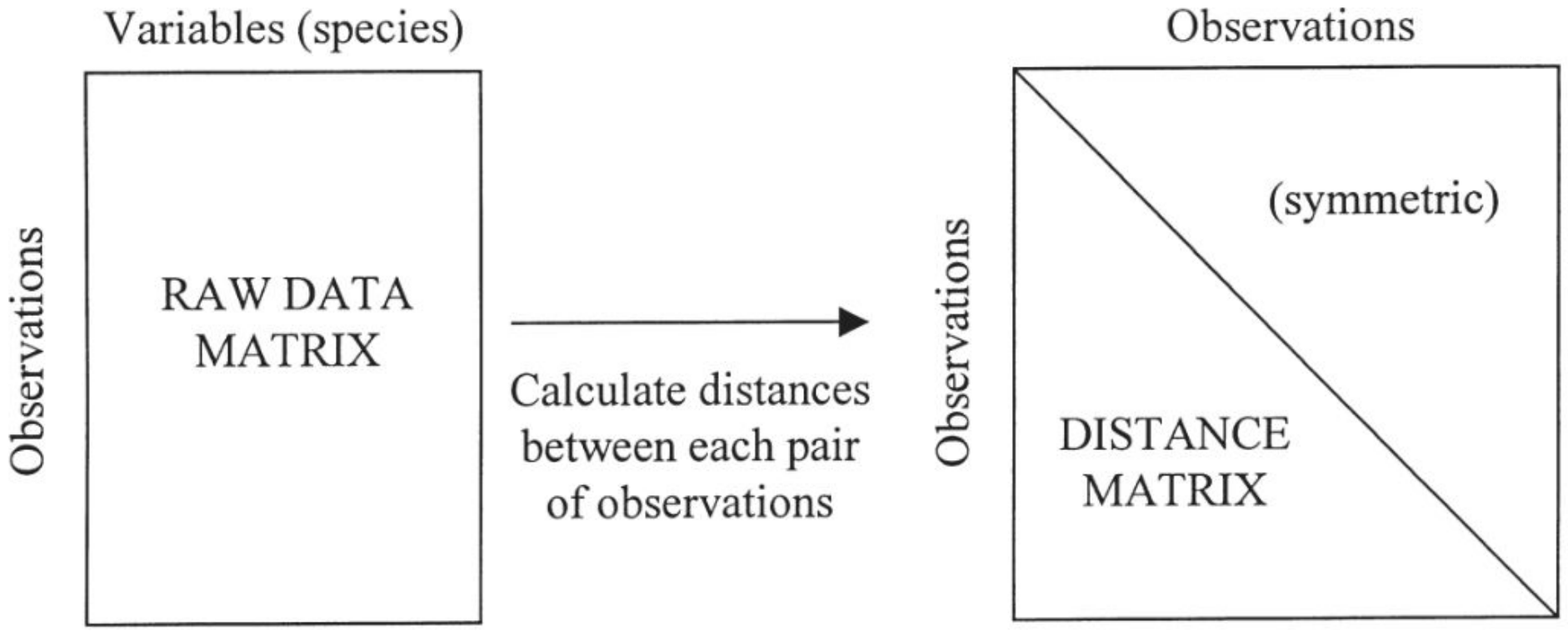

Developments in
Environmental Modelling
Vol. 24
Third English
Edition
Numerical
Ecology
Illuminance
Herb cover
Calamagrostis
Water content
Leaves and twigs
Tree cover
Pierre Legendre
Louis Legendre

Outline

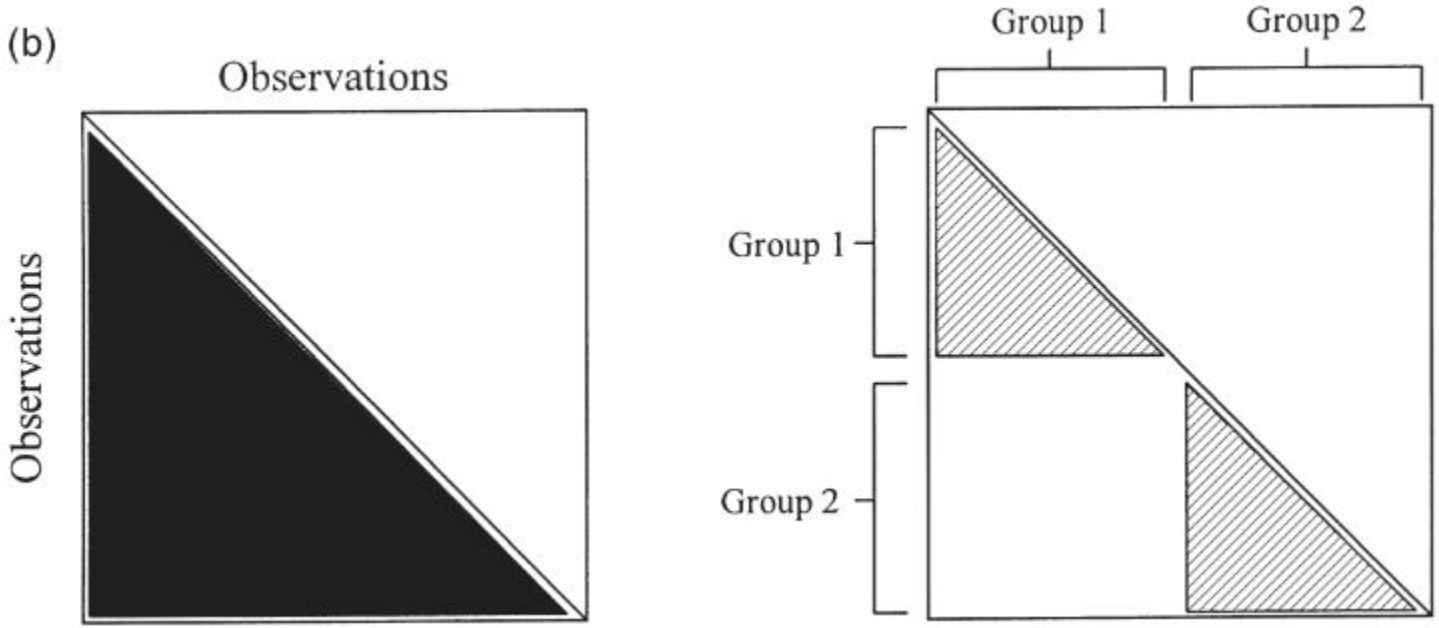
- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

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)

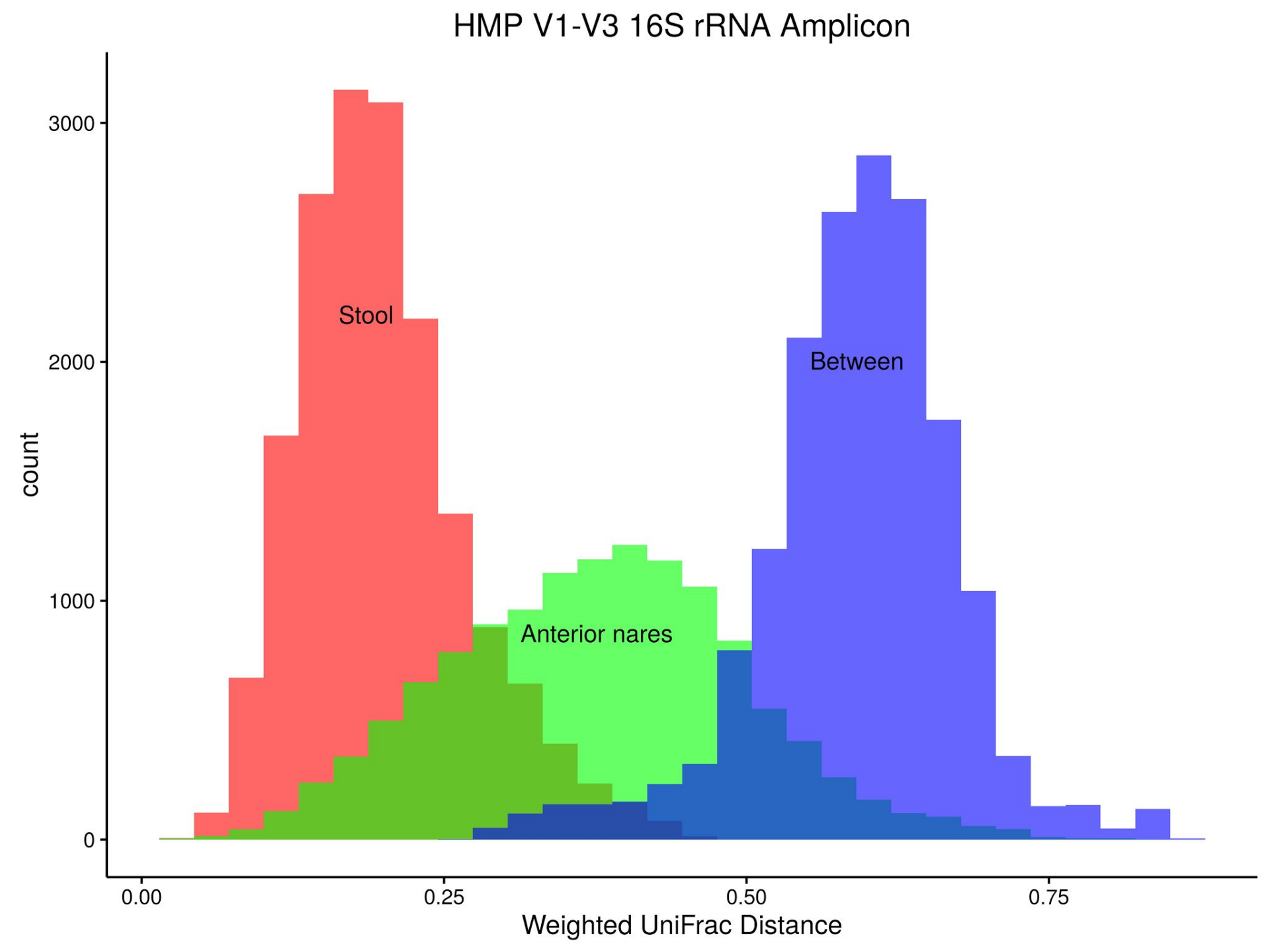


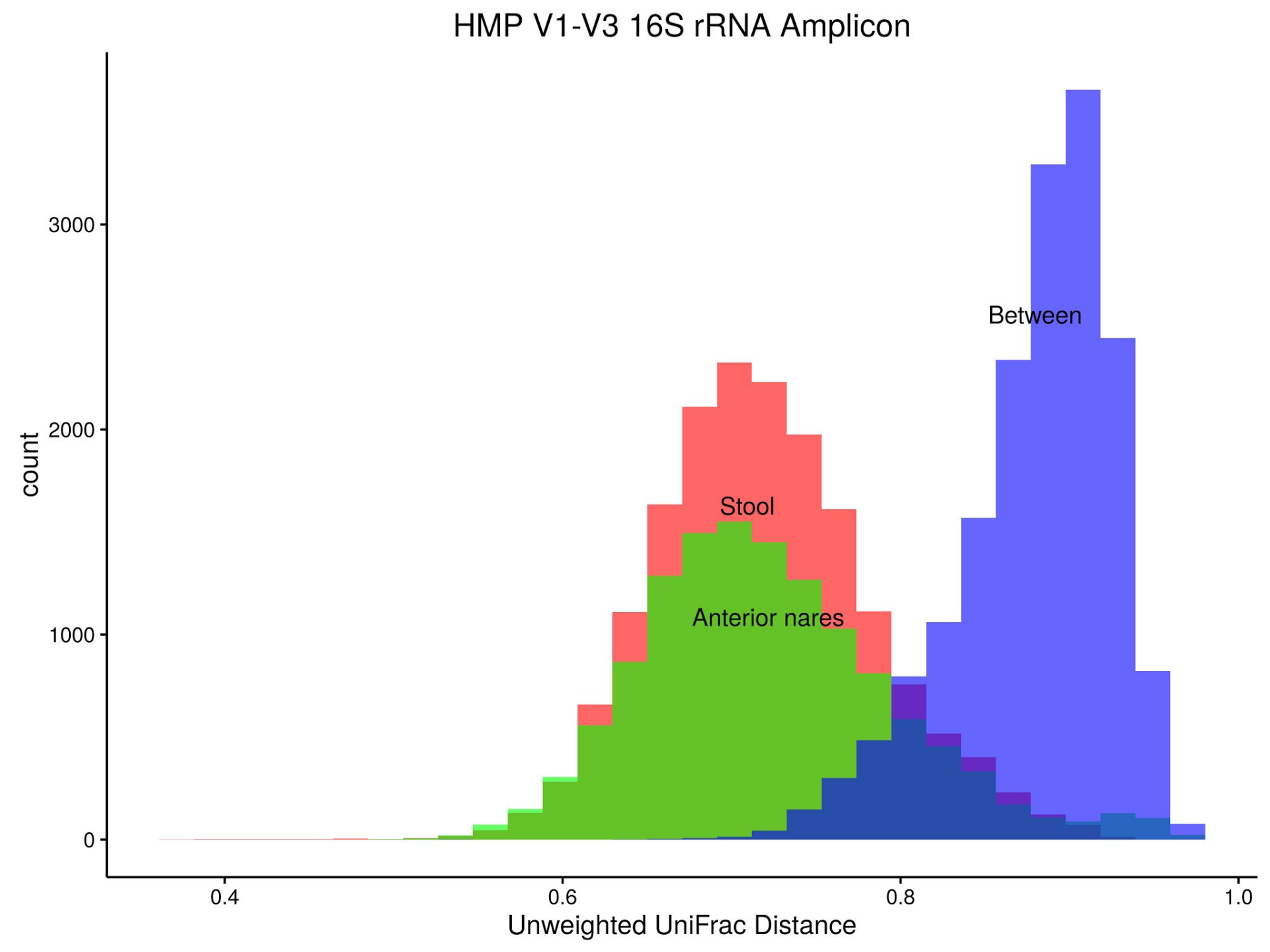
Anderson MJ et al. Austral Ecol 2001;26(1):32-46.

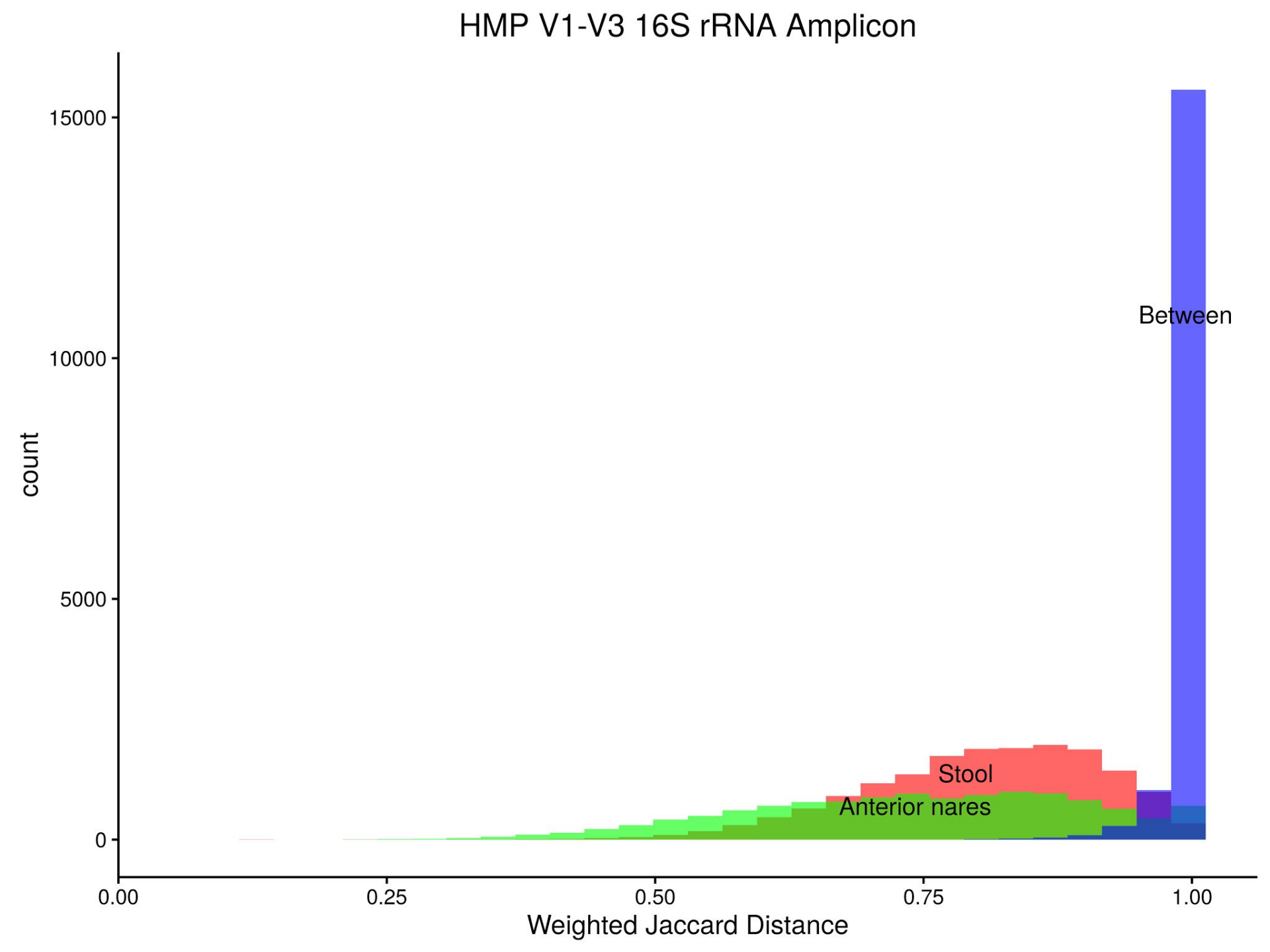


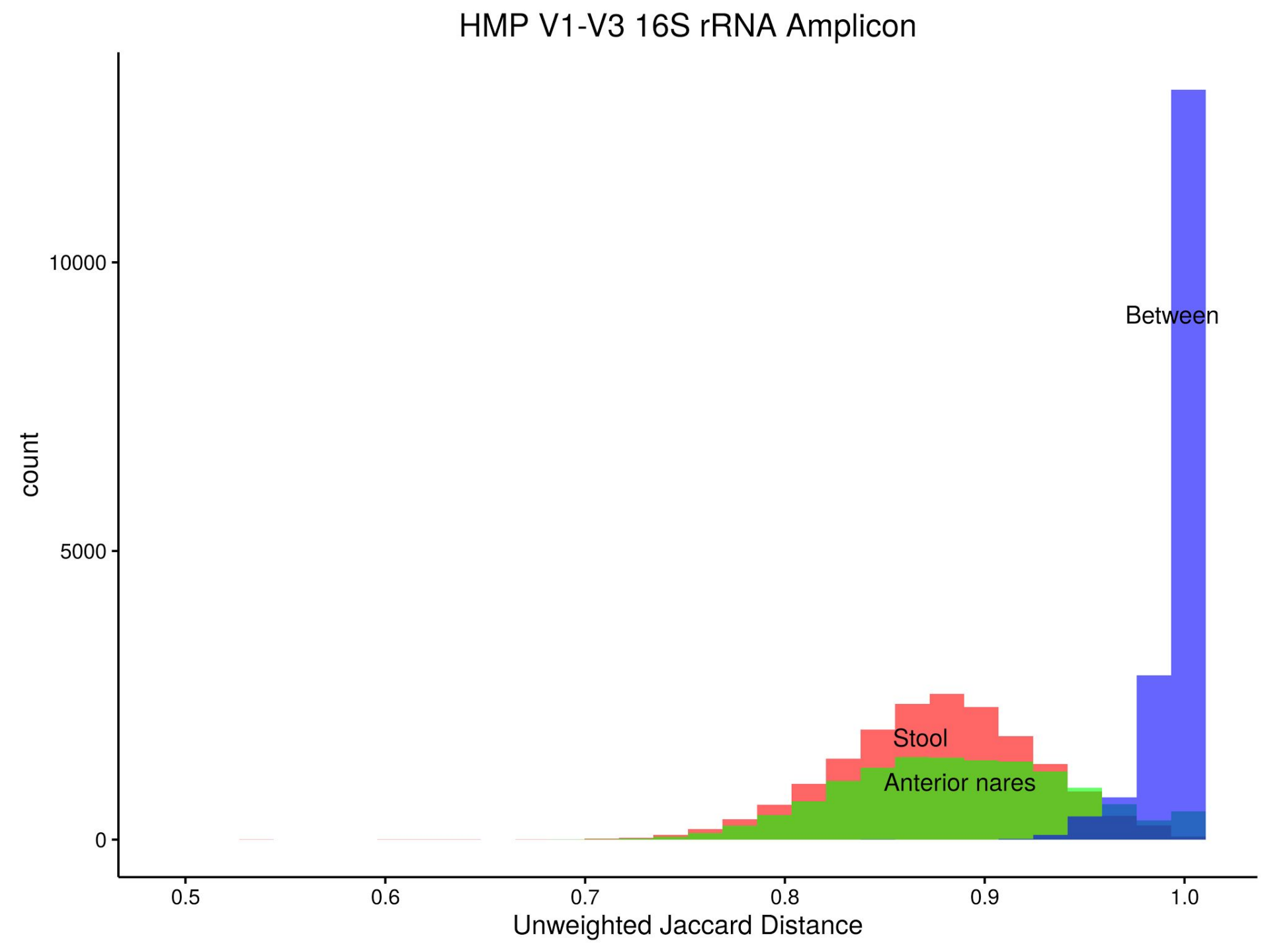
$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

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


HMP V1-V3 16S rRNA Amplicon
3000
2000
1000
0
count
Stool
Between
Anterior nares
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

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

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)

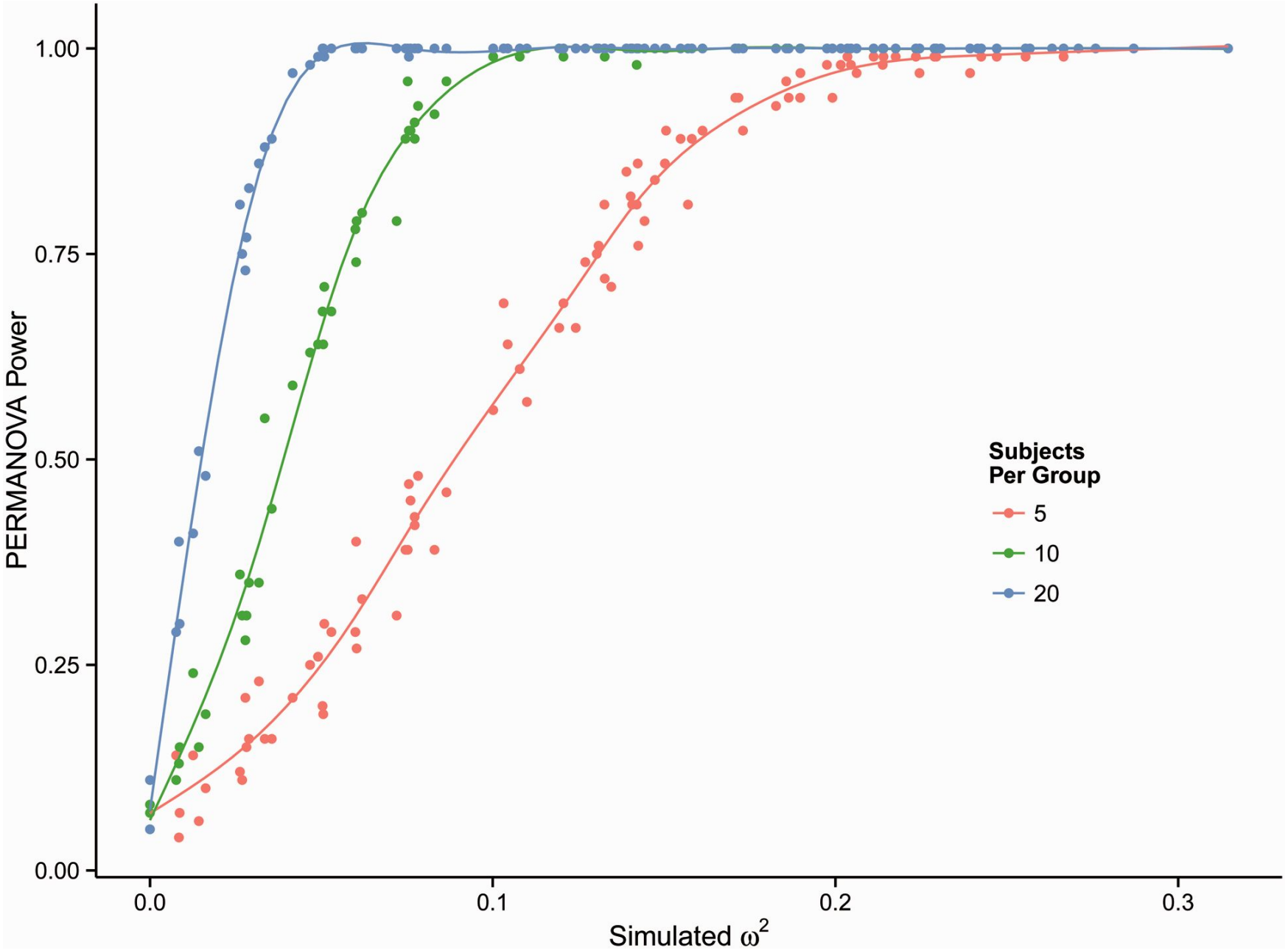
Kelly BJ et al. Bioinformatics 2015;31(15):2461-8.

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

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

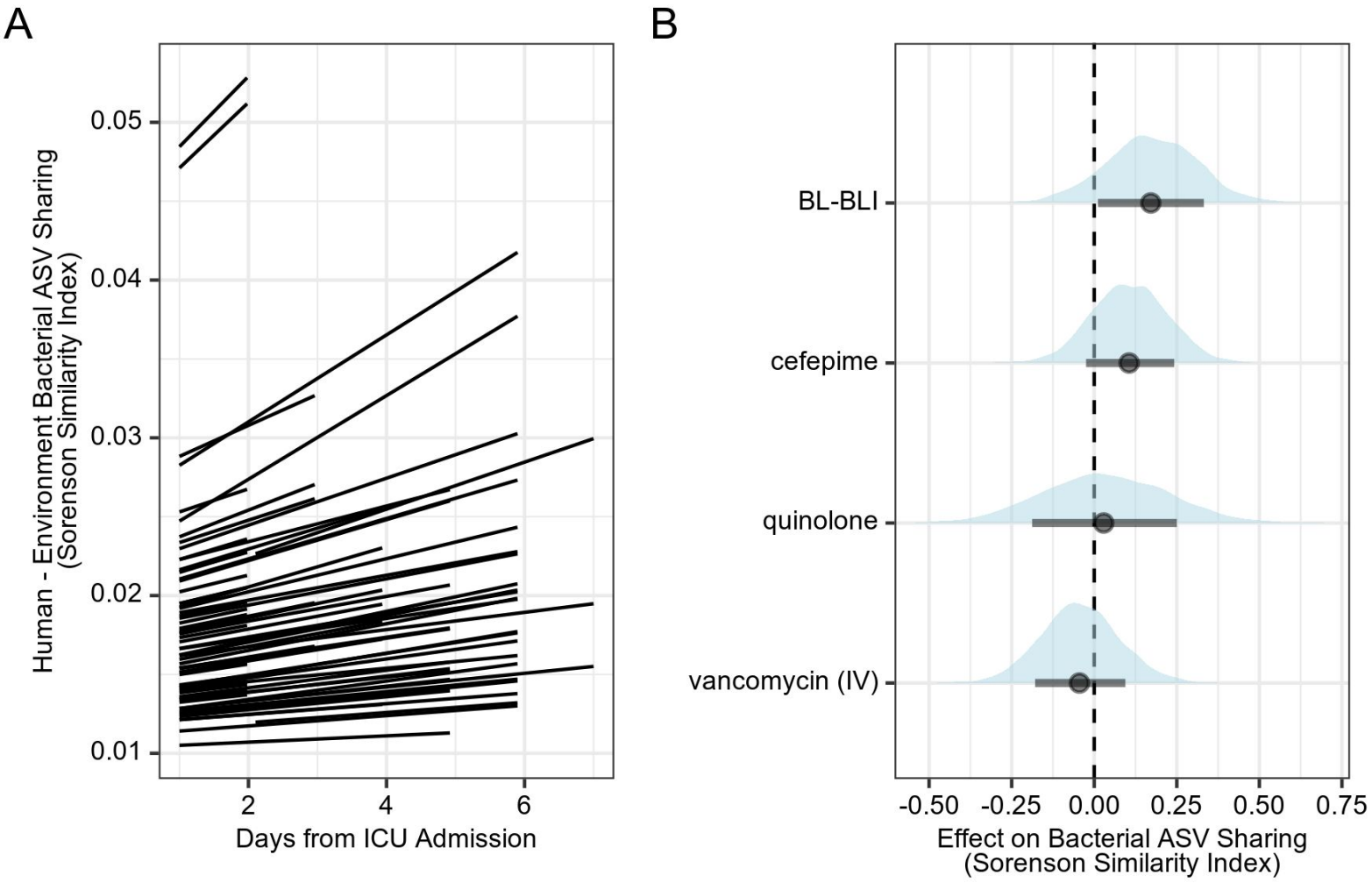
Site	Comparison groups		ω^2/*P*-value				
	Control	Exposure	Weighted UniFrac	Unweighted UniFrac	Weighted Jaccard	Unweighted Jaccard	Reference
Nares	Non-smoker (33)	Smoker (29)	0.042/0.001	0.009/0.001	0.023/0.001	0.007/0.001	Charlson *et al.* (2010)
Oral	Non-smoker (33)	Smoker (29)	0.032/0.001	0.008/0.001	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.



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

Other Approaches to Modeling Distance



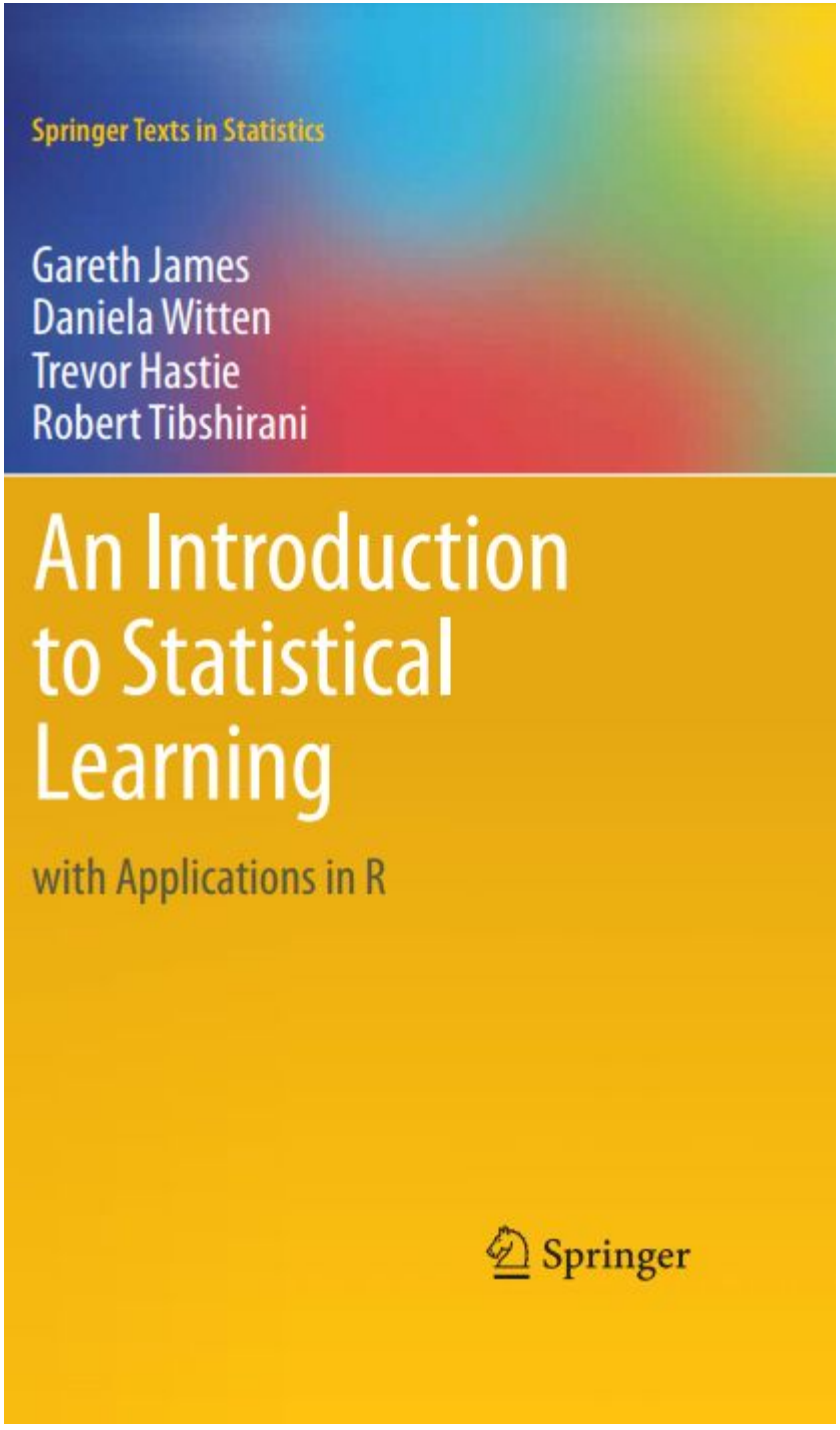
Outline

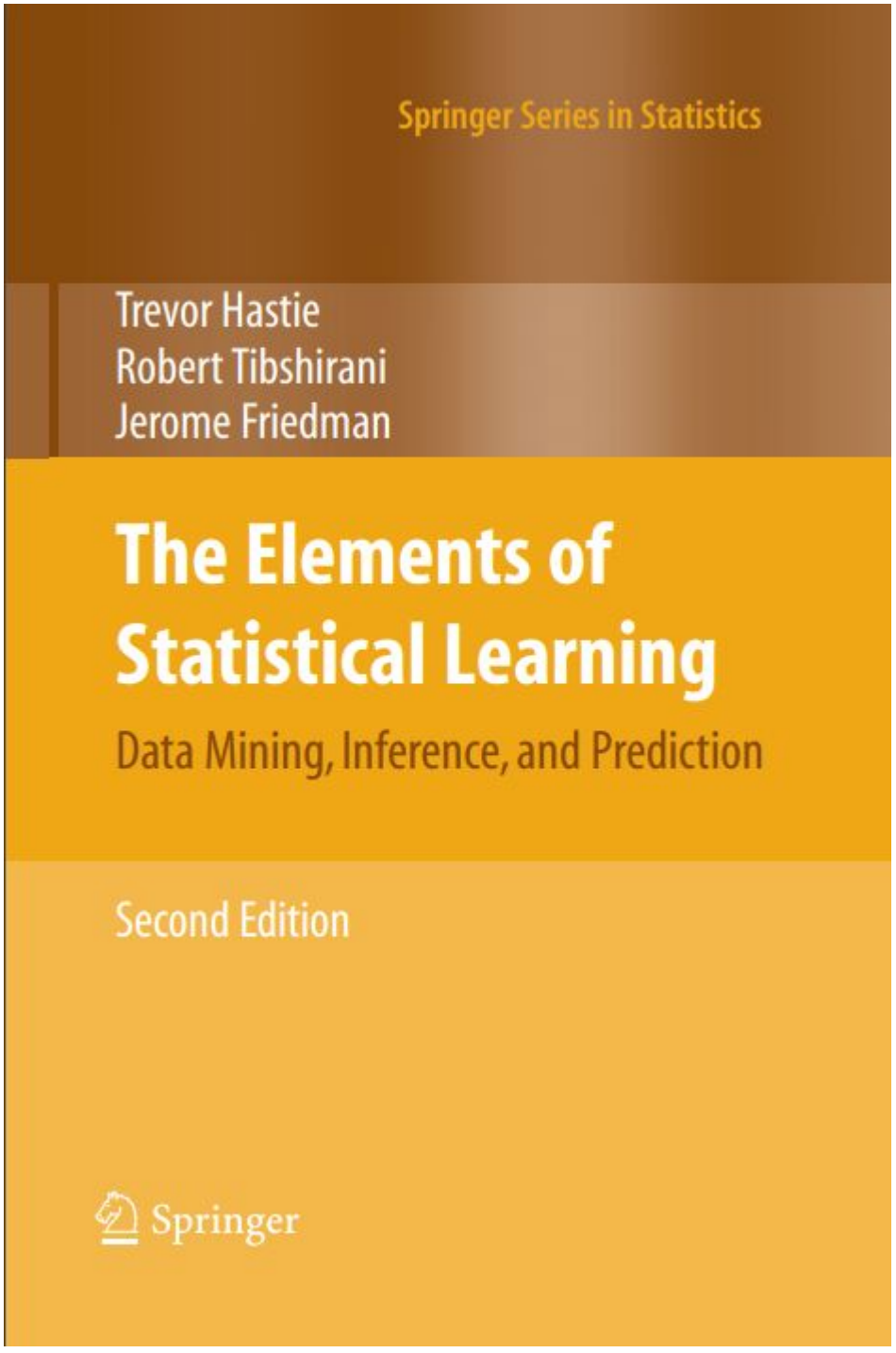
- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

Statistical / Machine Learning

- Supervised learning:
  - exposure and outcome
  - regression, linear discriminant analysis, KNN clustering
  - test & training data; cross-validation
- Unsupervised learning:
  - understand relationships between observations or variables
  - can we reduce the dimensions of microbiome data?

Holmes I et al. PLoS One 2012.


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

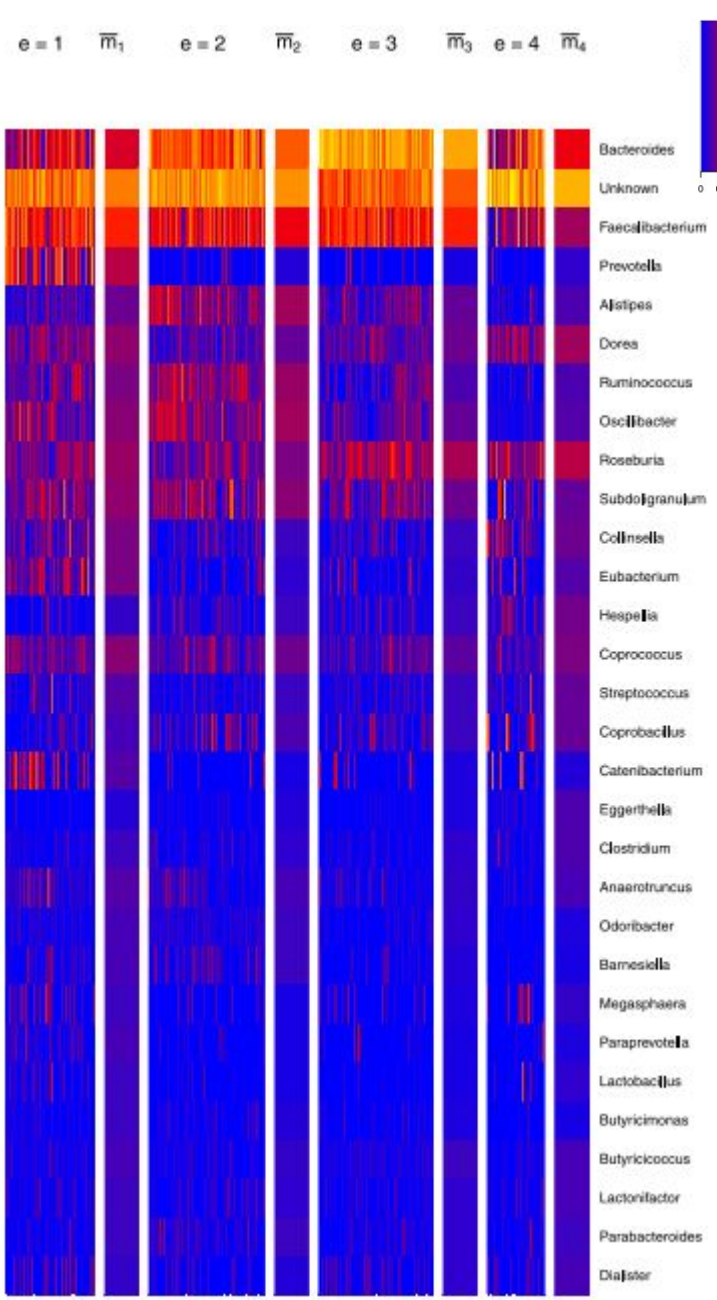
Outline

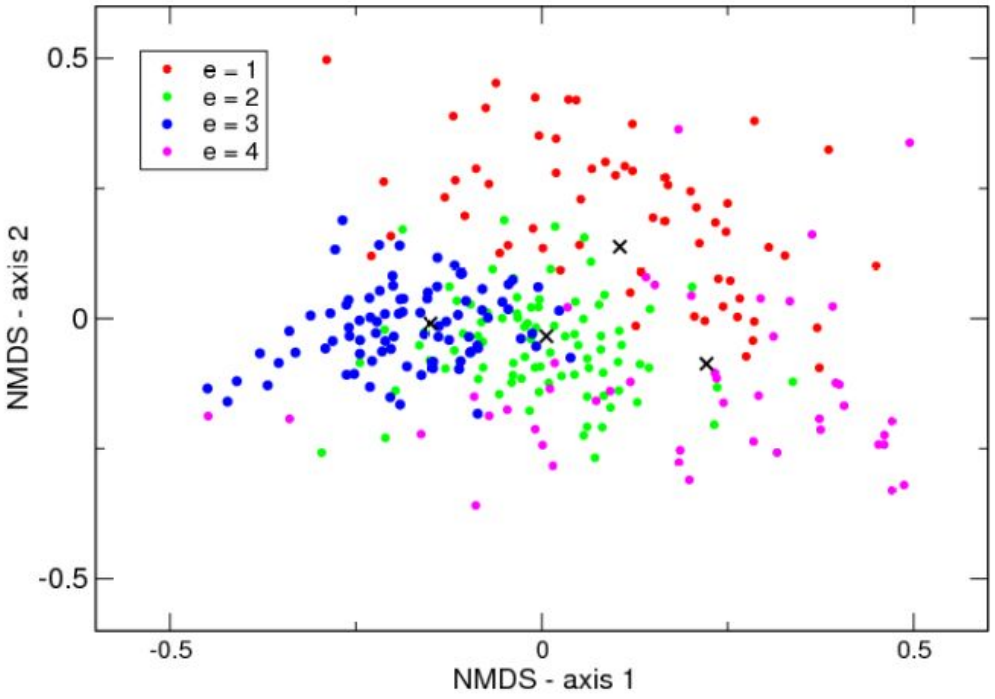
- The problem: too much data.
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)
- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling

Dimension Reduction: Mixture Models

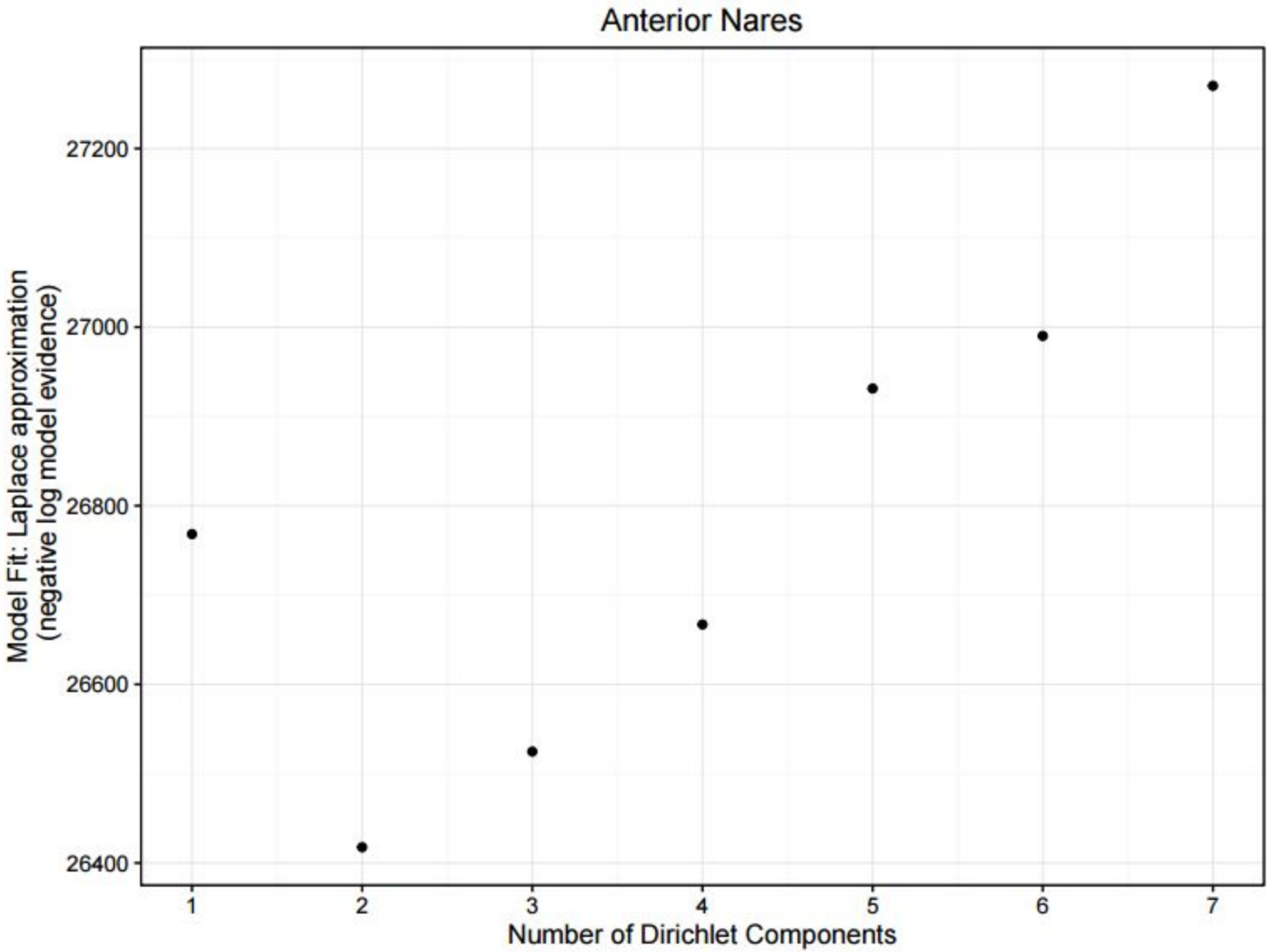
- Dirichlet-multinomial distribution:
  - compound probability distribution: probability vector drawn from Dirichlet distribution (generalized beta) → observation drawn from multinomial distribution (generalized binomial)
- D-M mixture modelling:
  - each sample ~ multinomial from one Dirichlet vector
  - # Dirichlet vectors: minimize -log(model evidence, Laplace approx)
  - Dirichlet probability vectors = “community types”

Holmes I et al. PLoS One 2012.

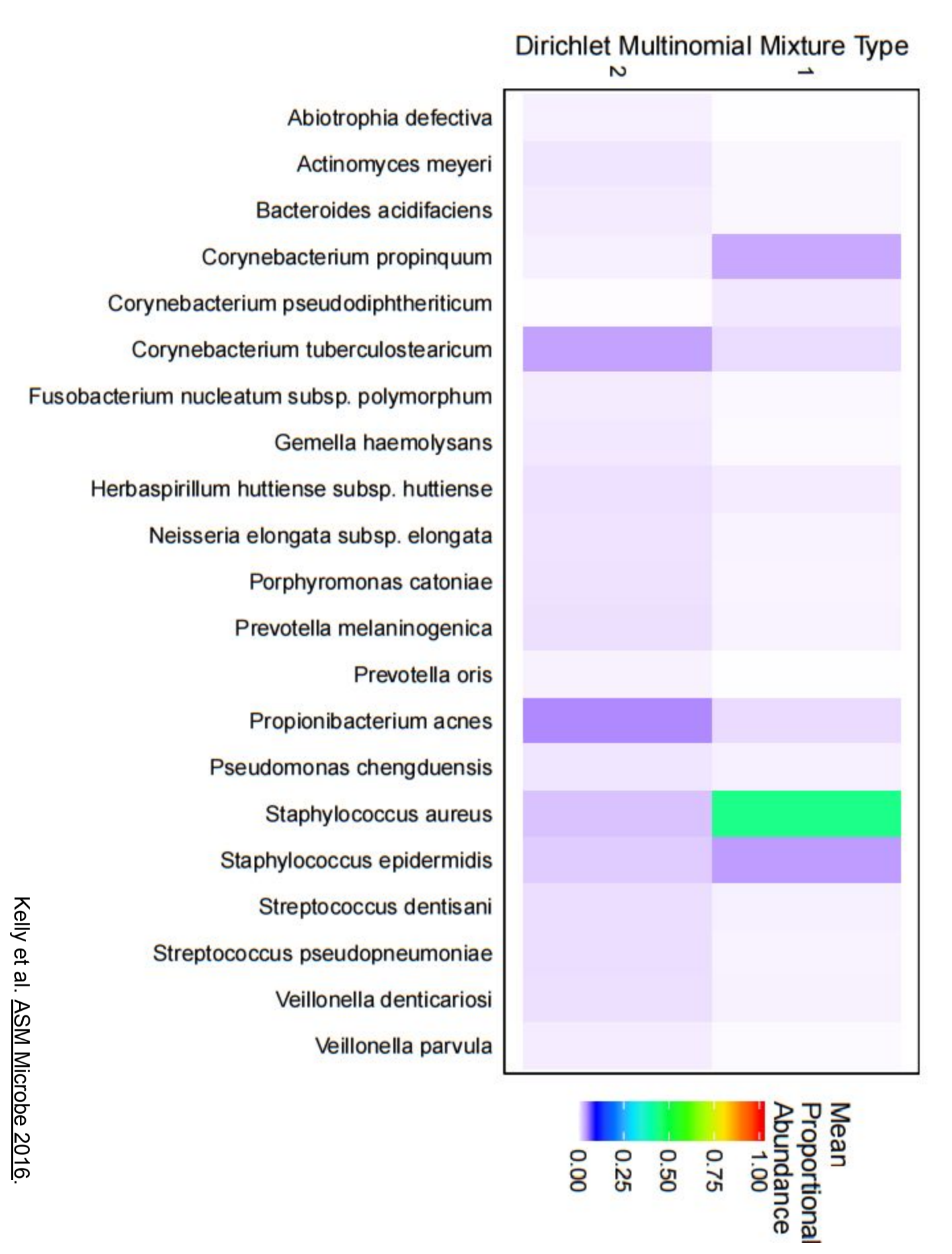




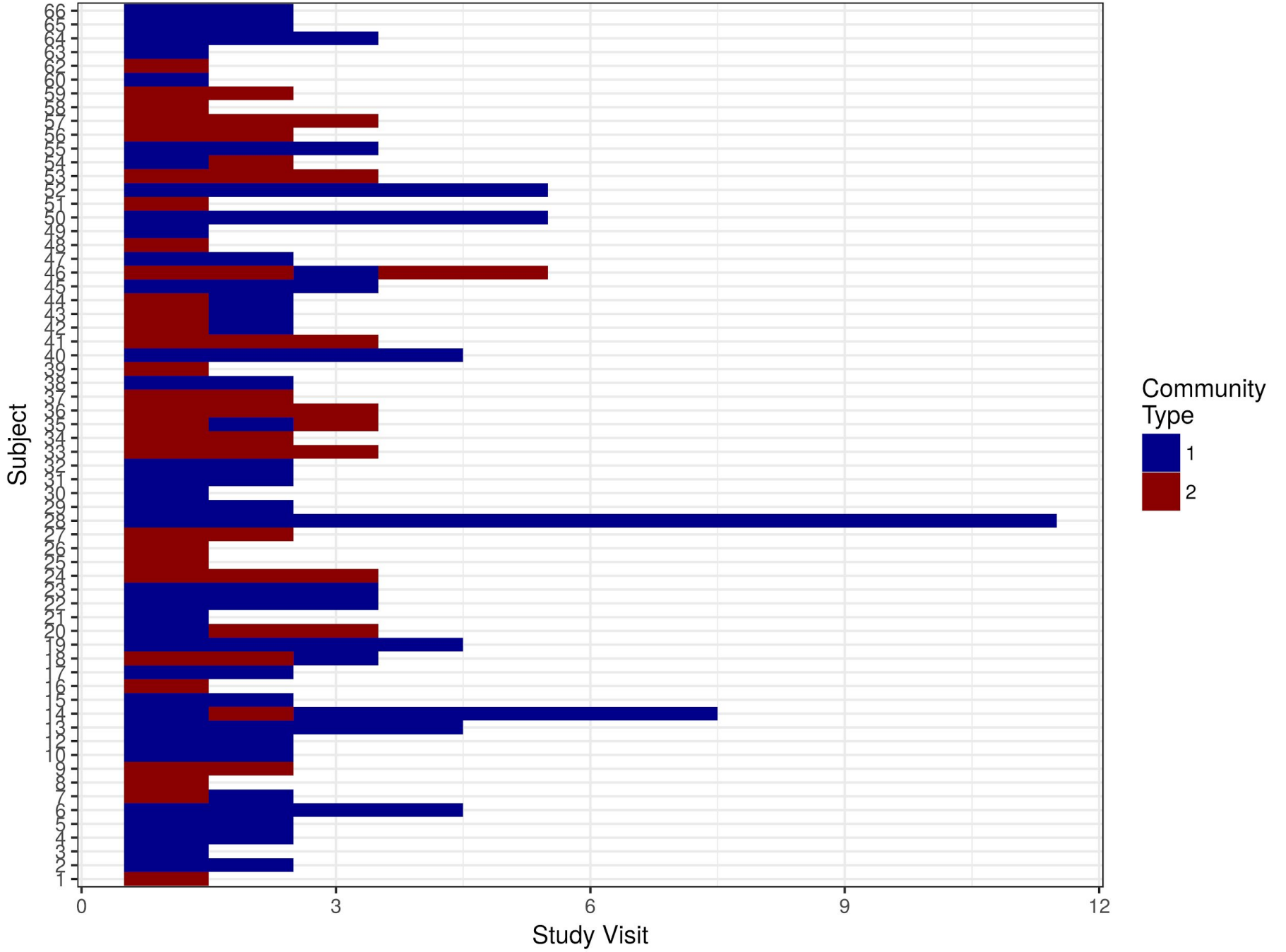
Holmes I et al. PLoS One 2012.



Kelly et al. ASM Microbe 2016.



Kelly et al. ASM Microbe 2016.


Subject
Study Visit
0
3
6
9
12
Community Type
1
2

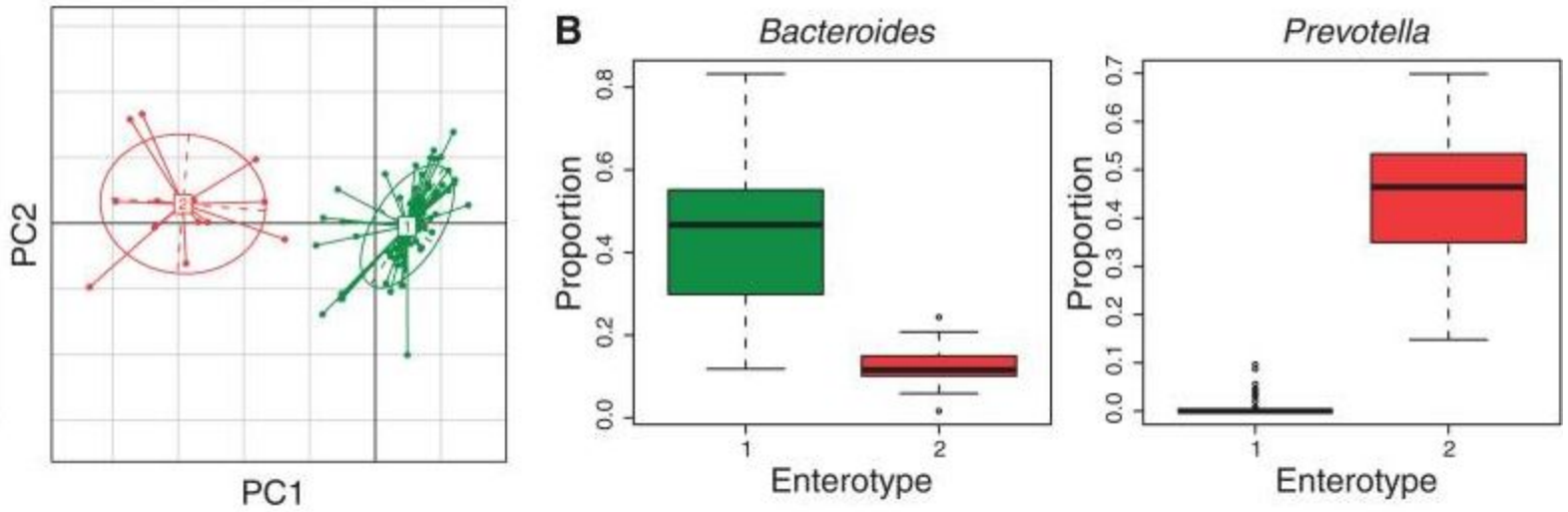
LETTER

doi:10.1038/nature13178

Dynamics and associations of microbial community types across the human body

Tao Ding[1] & Patrick D. Schloss[1]

Ding &Schloss Nature 2014.



Wu et al Science 2011.

Methods for Microbiome Data

- Visualization: heatmaps and barplots.
- Single-taxon hypothesis.
- Alpha diversity: richness and evenness.
- Pairwise distances: count/phylogeny, weighted/unweighted.
- Ordination: PCA & PCoA.
- PERMANOVA: categorical exposure & microbiome outcome

  (allows quantification of effect size)
- DMM models: unsupervised clustering > “community types”

  (identify relationships among variables/OTUs)

Conclusions

- Distance-based analysis and adonis/PERMANOVA testing:
  - microbiome outcome measures
  - *omega2* to define effect size of exposure/intervention
  - power estimation
- Dirichlet-multinomial mixtures:
  - categorical analysis may correspond with biologic community types
  - identify key species
  - discovery / validation design