Distances, Trees & Types

Methods for Analyzing High-Dimensional Microbiome Data

Brendan J. Kelly, MD, MS
Infectious Diseases | Epidemiology | Microbiology
University of Pennsylvania
17 June 2024

photo: Adrian Lange @unsplash

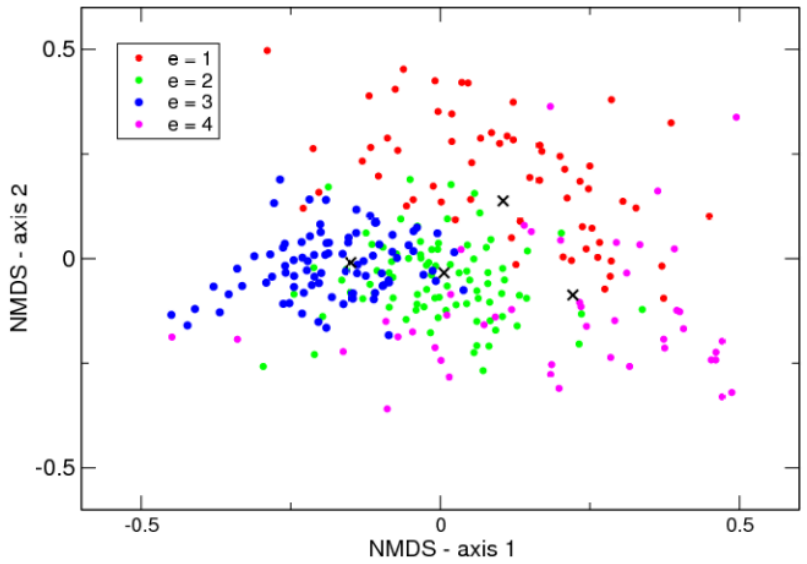
Disclosures

- No conflicts of interest.
- Research supported by:
  - NIAID K23 AI121485
  - CDC BAA 200-2016-91964
  - CDC BAA 200-2016-91937
  - CDC BAA 200-2018-02919
  - CDC BAA 200-2021-10986
  - CDC Prevention Epicenters U54CK000610
  - NIAID U19AI174998

Outline: The Microbiome for Epidemiologists

- The problem: too much data
  - the "p > n problem"
- Reducing dimensions:
  - richness, evenness, and diversity
  - ecological distances (UniFrac)
  - PCA & PCoA
  - PERMANOVA (adonis)

- Cluster analysis:
  - supervised vs unsupervised learning
  - Dirichlet multinomial mixture modeling



Too Much Data

photo: Diana Polekhina @unsplash

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 acids)
- Sequence binning and assignment:
  - operational taxonomic units (OTUs) based on 97% sequence similarity → taxonomic assignment of OTUs
  - assemble contiguous metagenomic sequences → taxonomy
  - unassembled/un-binned reads → taxonomic assignment (e.g., ASVs or SGBs)

Callahan BJ *bioArxiv* 2017

An Example OTU Table

	OTU ID	Specimen_1	Specimen_2	Specimen_3	Specimen_4	Specimen_5	Specimen_6	Specimen_7	Specimen_8	Specimen_9	Specimen_10
1	OTU_1	736	400	989	0	674	380	0	511	24	0
2	OTU_2	826	0	0	697	893	860	460	0	276	0
3	OTU_3	270	564	0	0	252	965	280	348	0	0
4	OTU_4	0	320	571	51	279	937	0	70	668	434
5	OTU_5	989	0	482	658	701	0	674	652	584	704
6..999											
1000	OTU_1000	188	724	461	152	469	459	0	61	496	0

An Example OTU Table: Samples & Variables?

	OTU ID	Specimen_1	Specimen_2	Specimen_3	Specimen_4	Specimen_5	Specimen_6	Specimen_7	Specimen_8	Specimen_9	Specimen_10
1	OTU_1	736	400	989	0	674	380	0	511	24	0
2	OTU_2	826	0	0	697	893	860	460	0	276	0
3	OTU_3	270	564	0	0	252	965	280	348	0	0
4	OTU_4	0	320	571	51	279	937	0	70	668	434
5	OTU_5	989	0	482	658	701	0	674	652	584	704
6..999											
1000	OTU_1000	188	724	461	152	469	459	0	61	496	0

- which are the samples and which are the variables?
- also note: lots of zeros

An Example OTU Table: Samples vs Variables

	Specimen ID	OTU_1	OTU_2	OTU_3	OTU_4	OTU_5	OTU_6	OTU_7	OTU_8	OTU_9	OTU_10	OTU_11	OTU_12	OTU_13	OTU_14	OTU_15	OTU_1
1	Specimen_1	736	826	270	0	989	474	63	636	233	891	0	872	594	0	967	
2	Specimen_2	400	0	564	320	0	9	193	728	947	388	517	168	747	38	0	
3	Specimen_3	989	0	0	571	482	0	506	536	531	312	67	801	0	816	227	65
4	Specimen_4	0	697	0	51	658	142	0	0	198	61	208	0	676	0	22	61
5	Specimen_5	674	893	252	279	701	662	0	0	568	949	139	390	909	0	710	94
6..9																	
10	Specimen_10	0	0	0	434	704	30	0	580	0	600	0	0	604	251	435	63

- OTUs are really *columns* (i.e., variables) & specimens *rows* (i.e., observations)
- 10 rows & 1000 columns → p >> n → overfitting risk

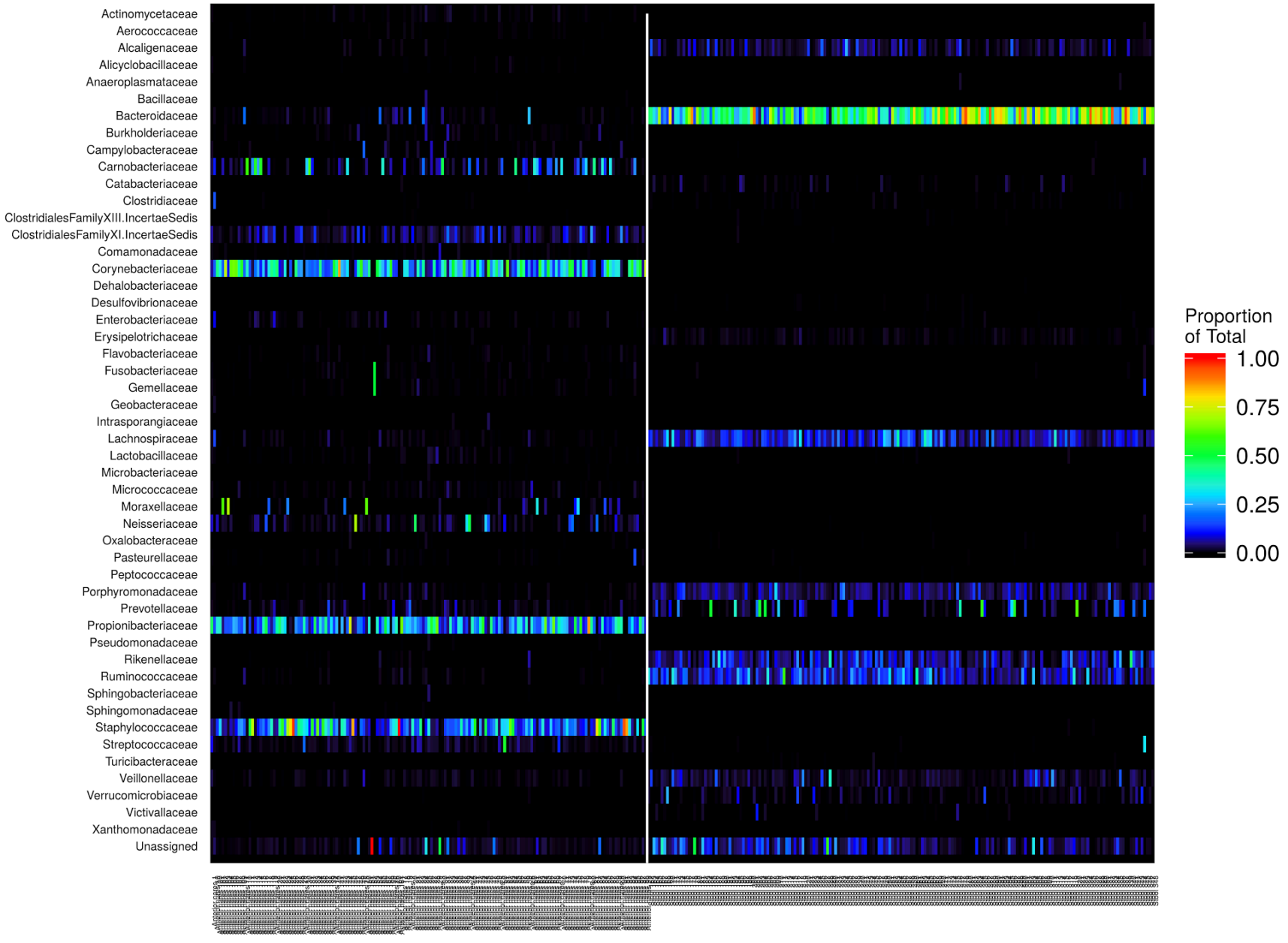
Reducing Dimensions

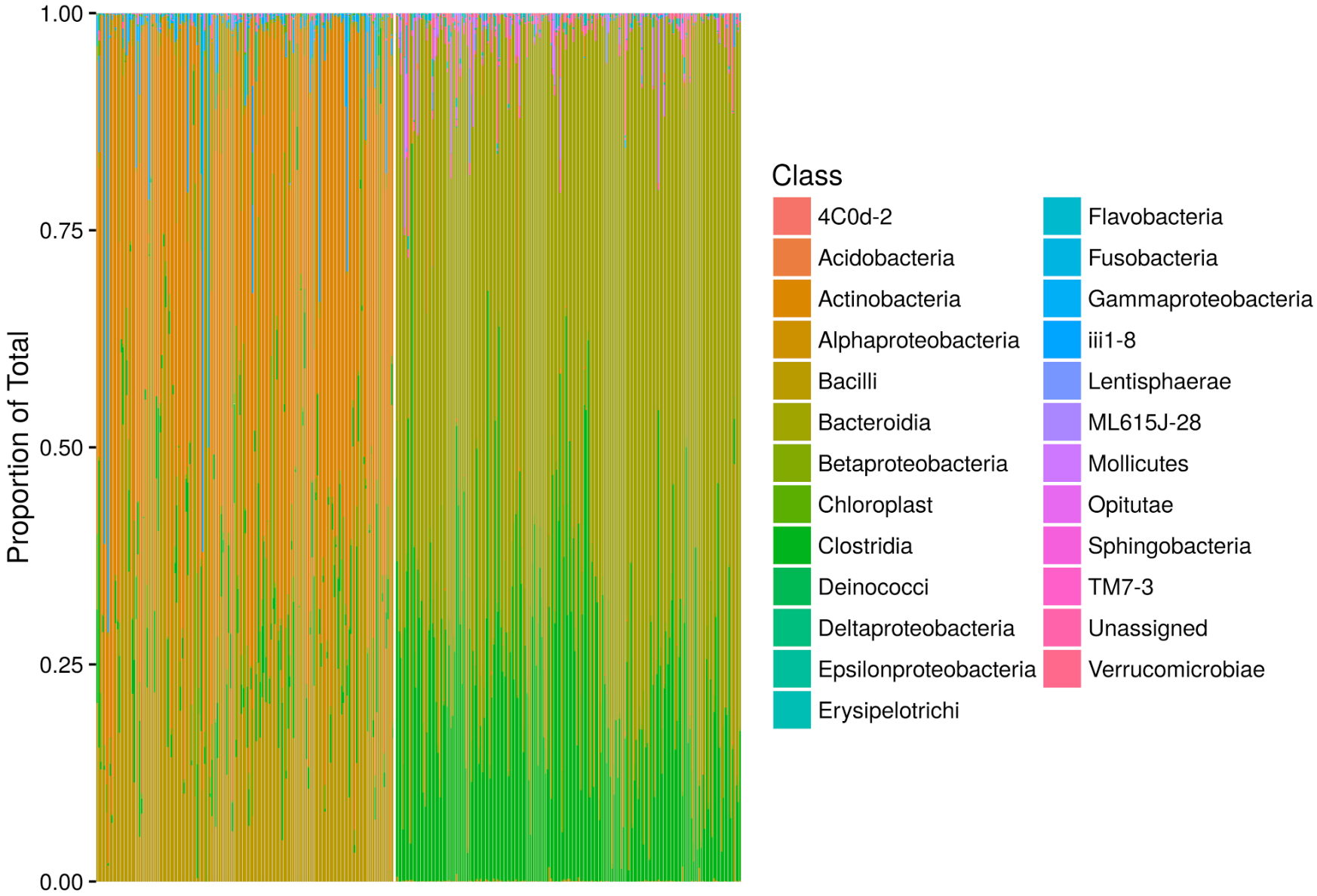
photo: Michael @unsplash

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




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)?
  - functional group (e.g., butyrate production)?
- Hypotheses focusing on specific taxa often fail to account for **possibility of selection bias from culture**.

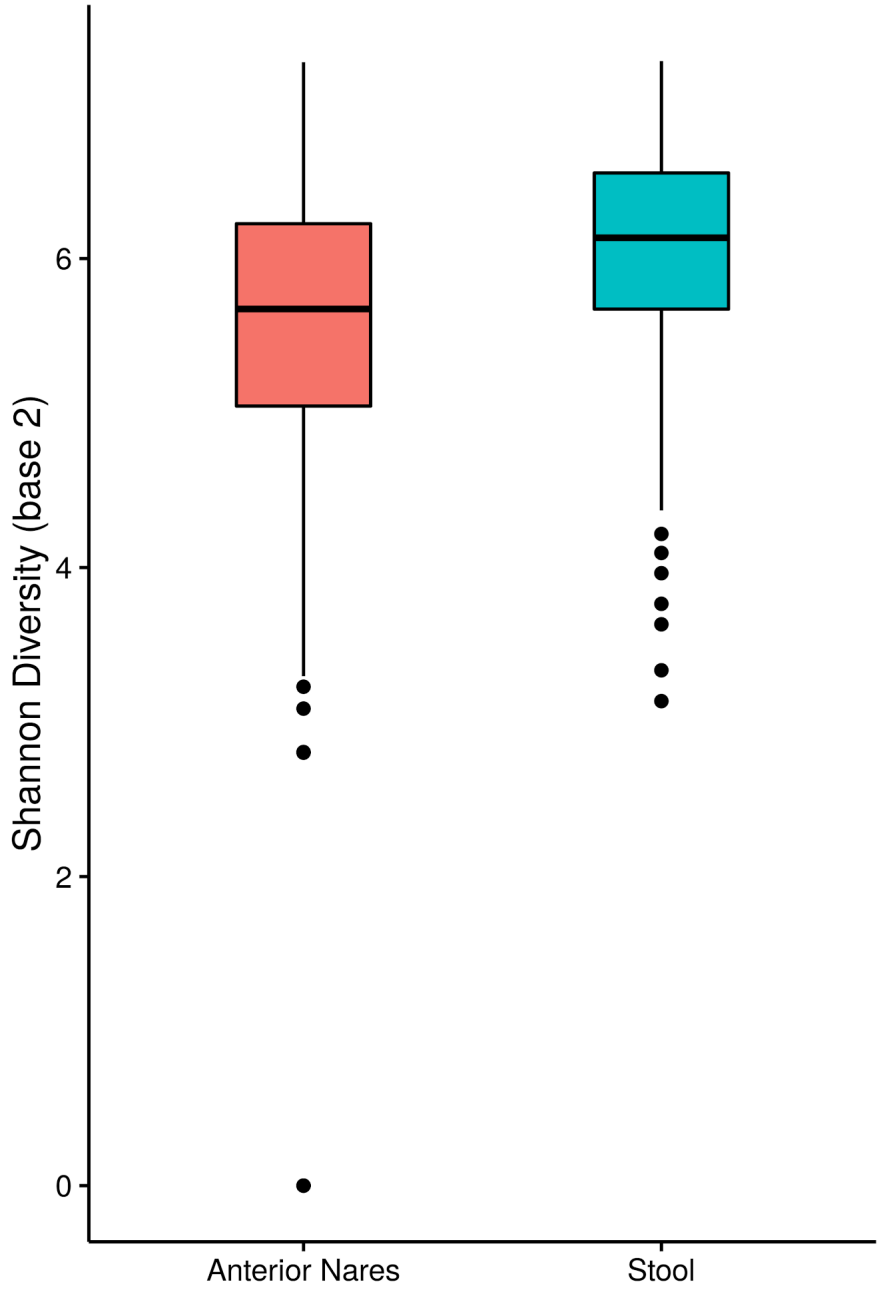
Single-Taxon Hypotheses

OTU	phylum	class	order	family	genus
OTU_97.1	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
OTU_97.10	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Veillonella
OTU_97.100	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	NA
OTU_97.1000	Proteobacteria	Betaproteobacteria	NA	NA	NA
OTU_97.10000	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium
OTU_97.10001	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
OTU_97.10002	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	NA
OTU_97.10003	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium
OTU_97.10004	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium
OTU_97.10005	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides

Dimension Reduction: Alpha Diversity

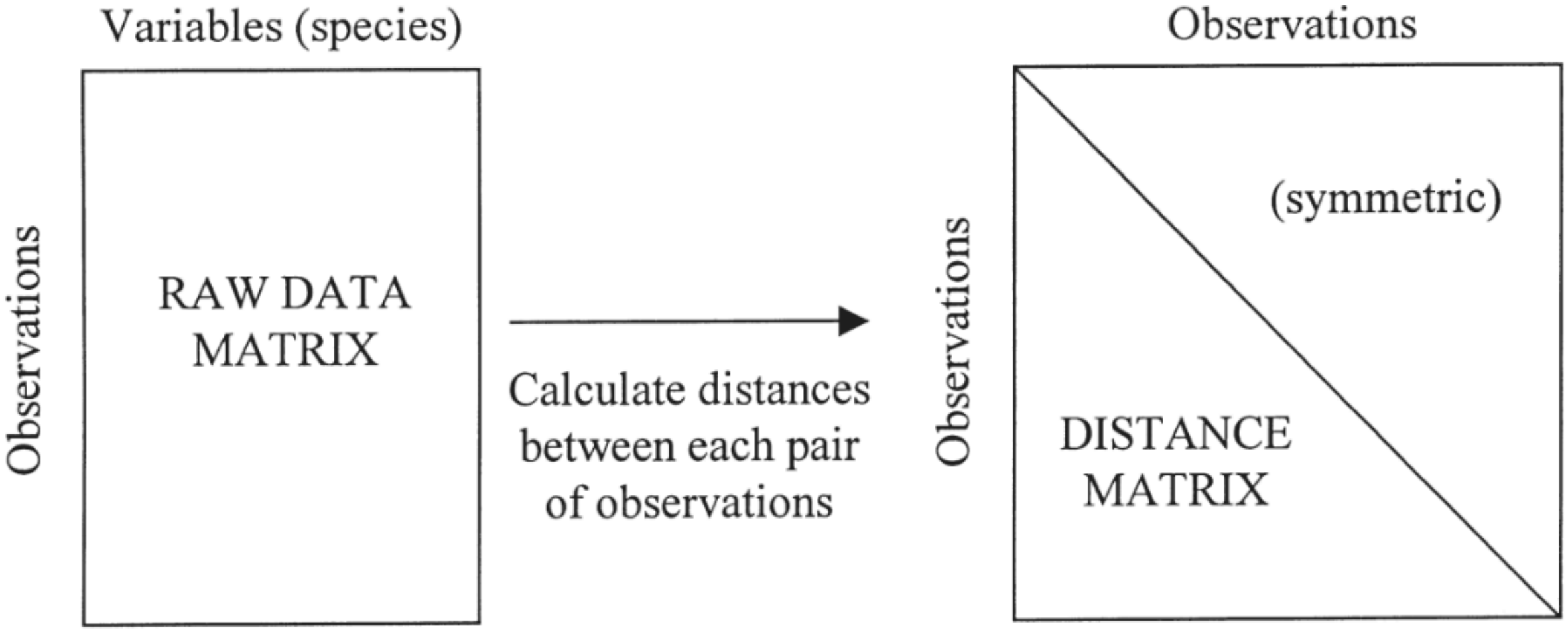
- Summarize each sample's community in a single measure:
  - richness: number of community members
  - evenness: the distribution of member counts
- Many alpha diversity metrics (weight richness/evenness):
  - species number, Chao1 (singletons & doubletons)
  - Shannon diversity: $H' = -\sum_i p_i \cdot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

    (note: may measure similarity or dissimilarity)


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

Dimension Reduction: Beta Diversity

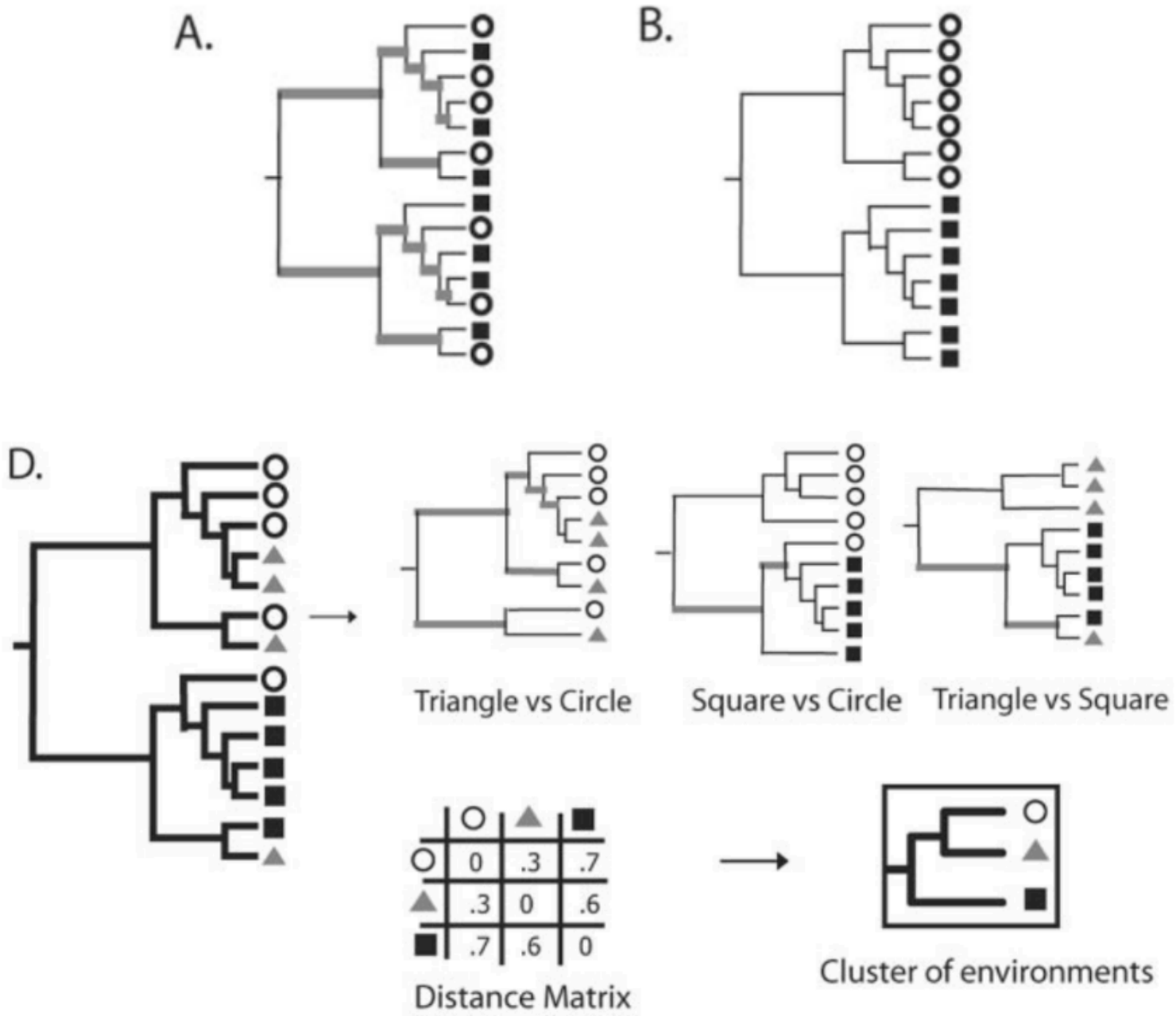
- Summarize each sample's relationship to other samples:
  - pairwise distances
  - OTU table → square matrix
- Many beta diversity metrics:
  - just counts versus counts + phylogeny
  - weighted versus unweighted
  - (Euclidean versus non-Euclidean)



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

Dimension Reduction: Beta Diversity

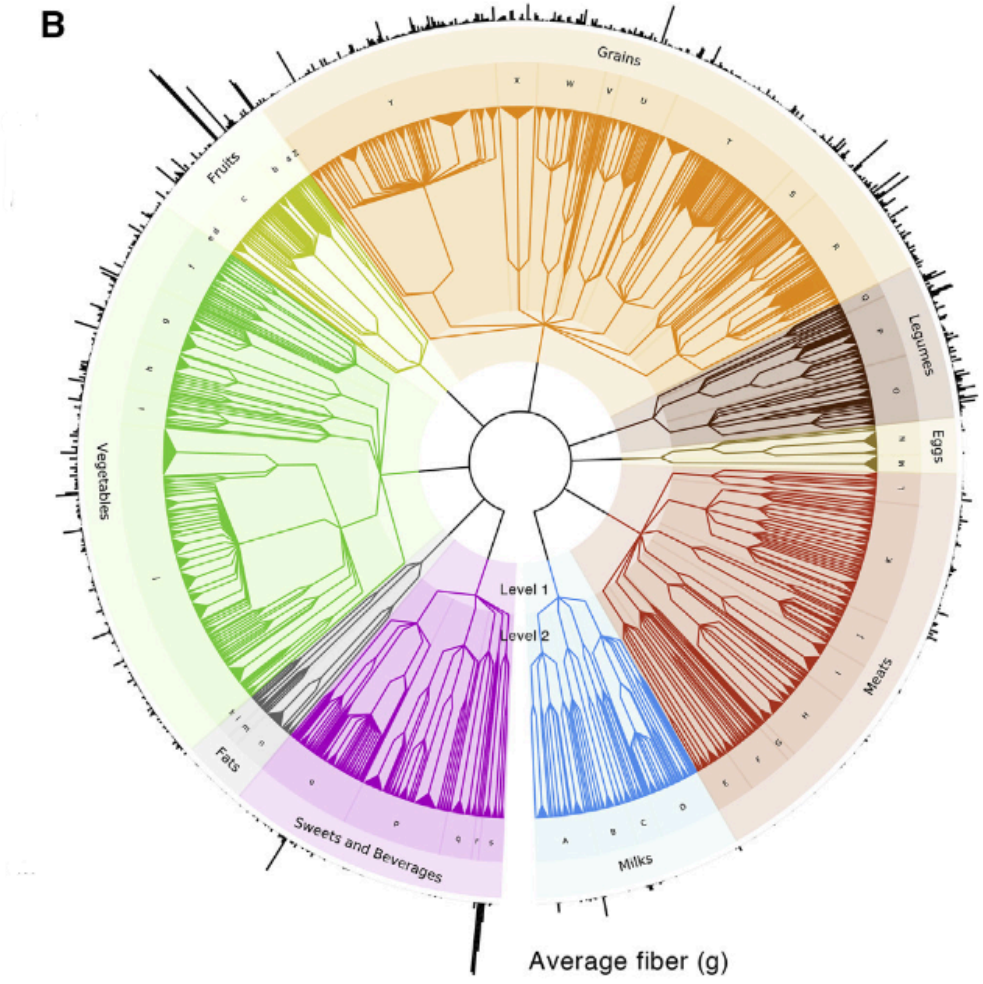
- Just counts versus counts + phylogeny:
  - Jaccard: $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$
  - Jaccard distance: $d_J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$
  - UniFrac: fraction of unique branch length in 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 distances? Unweighted distances?

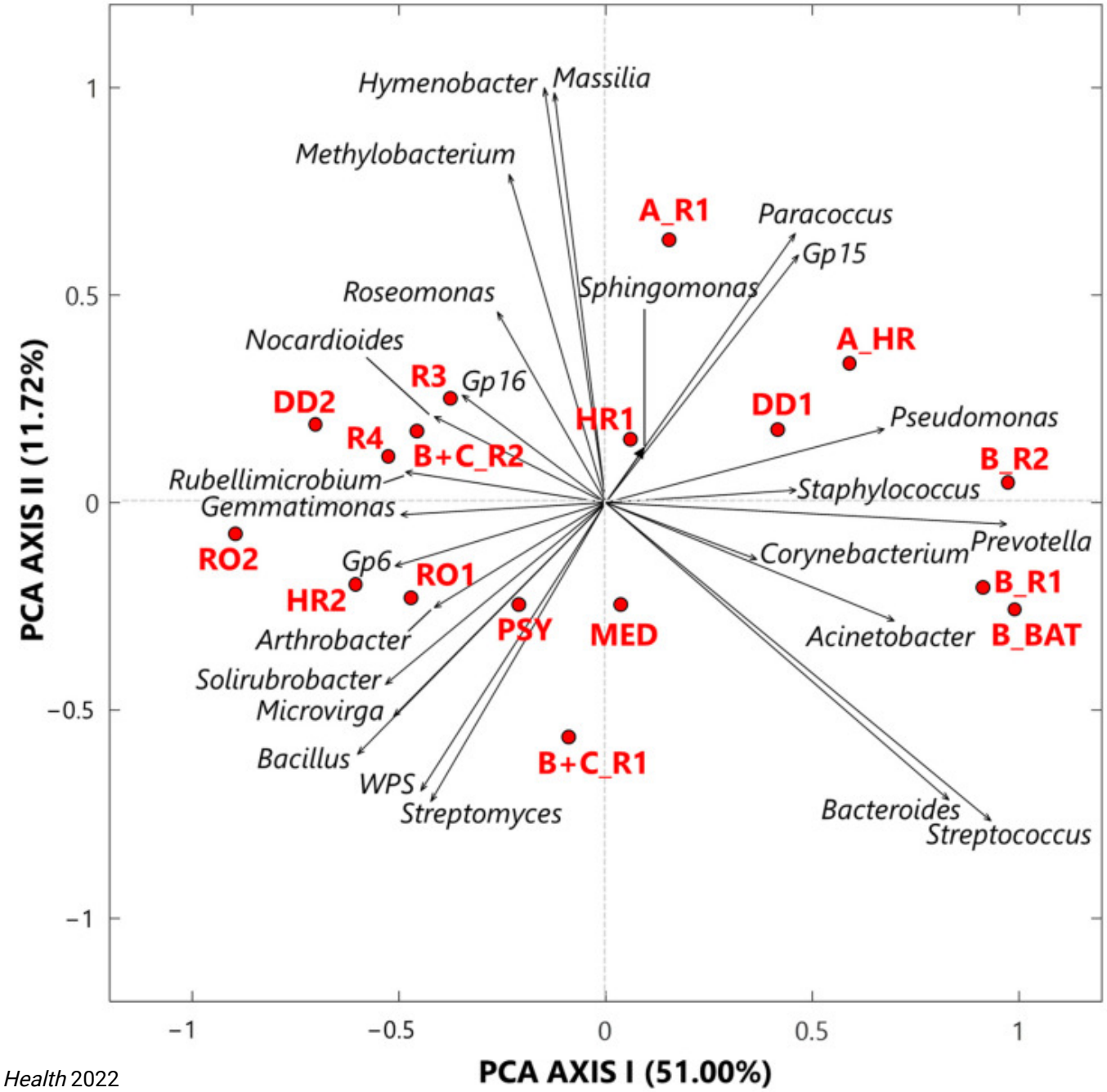
Thinking Like a Tree



Johnson AJ et al. *Cell Host Microbe* 2019

Dimension Reduction: PCA

- PCA: principal component analysis
  - uses original descriptors (e.g., OTU abundance)
  - rigid rotation for successive directions of maximum variance
  - lots of restrictions (Euclidean)
  - but allows projection of original descriptors in PCA space

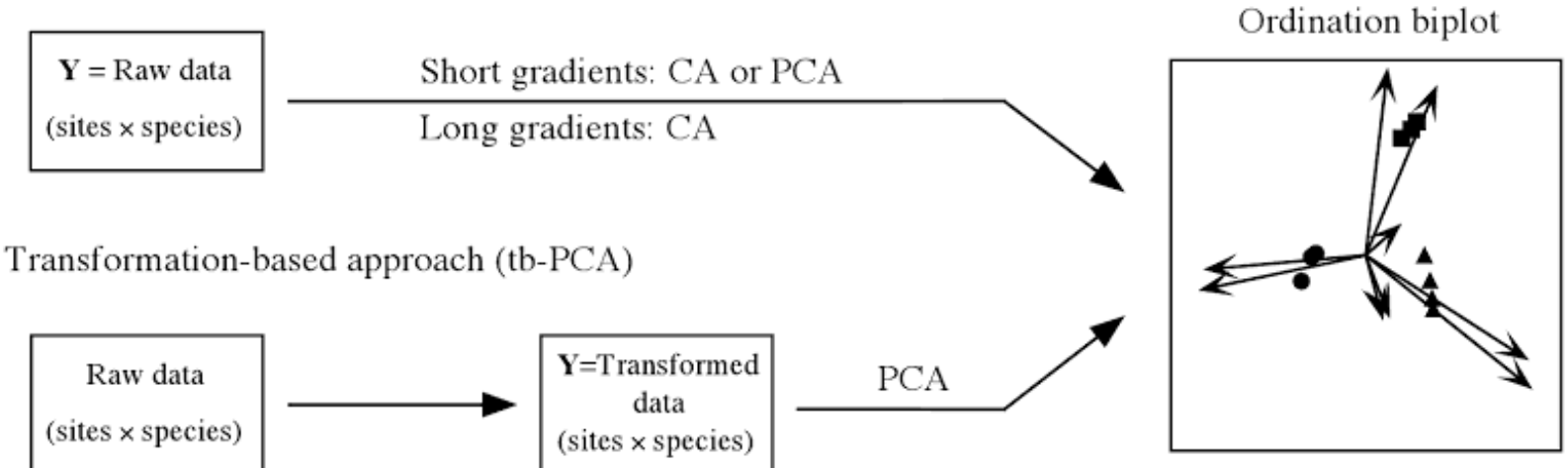


Perrone MR et al *Int J Environ Res Public Health* 2022

Dimension Reduction: PCoA

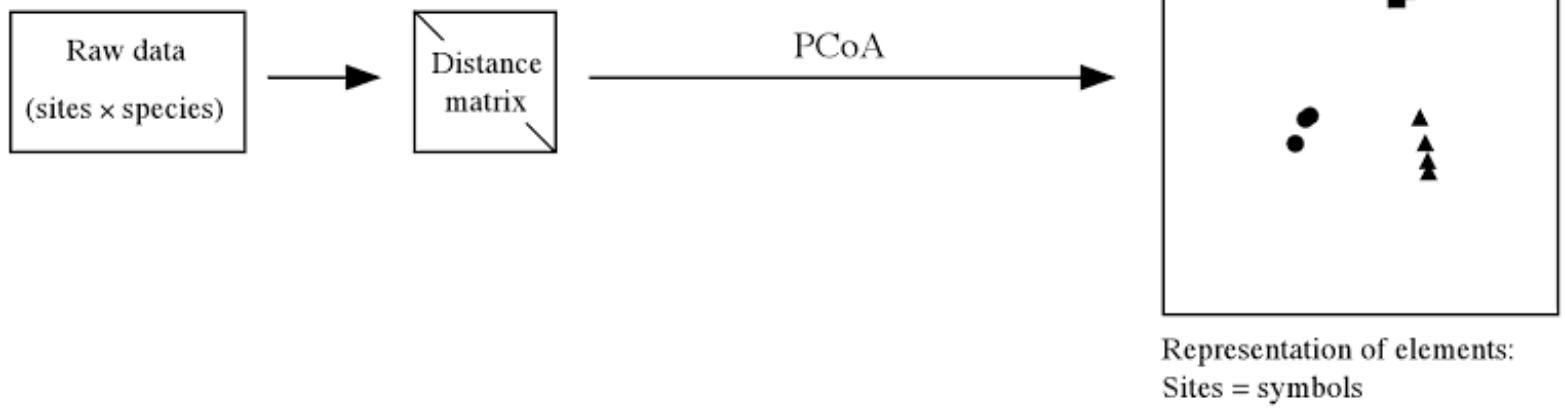
- PCoA: principal coordinate analysis
  - uses pairwise distances
  - any metric distance, even if non-Euclidean
  - like PCA, eigenvalue decomposition (maximum variance) but mediated by distance function (no original descriptors)
  - unlike PCA, does not allow projection of original descriptors in reduced-dimension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

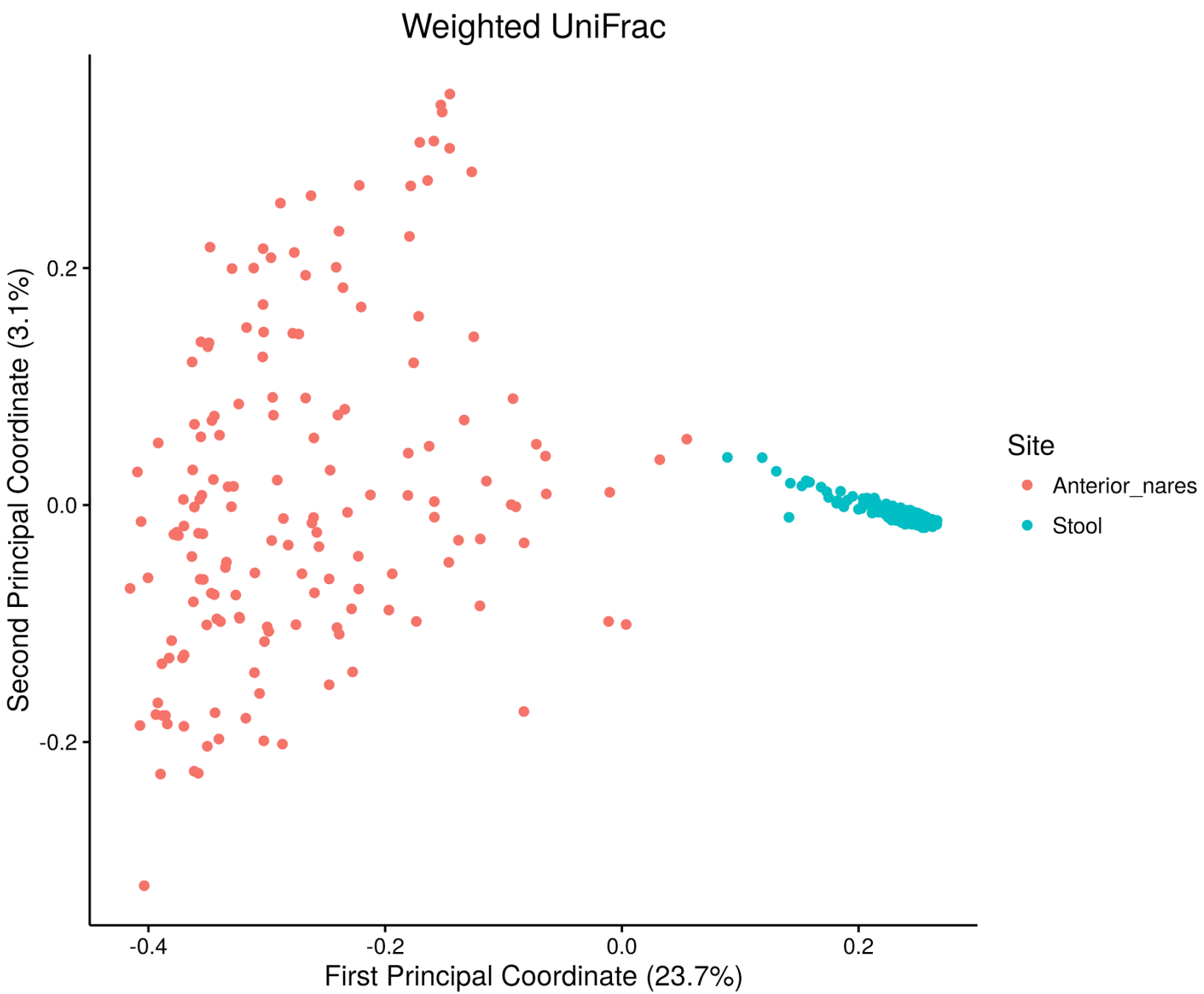
(a) Classical approach

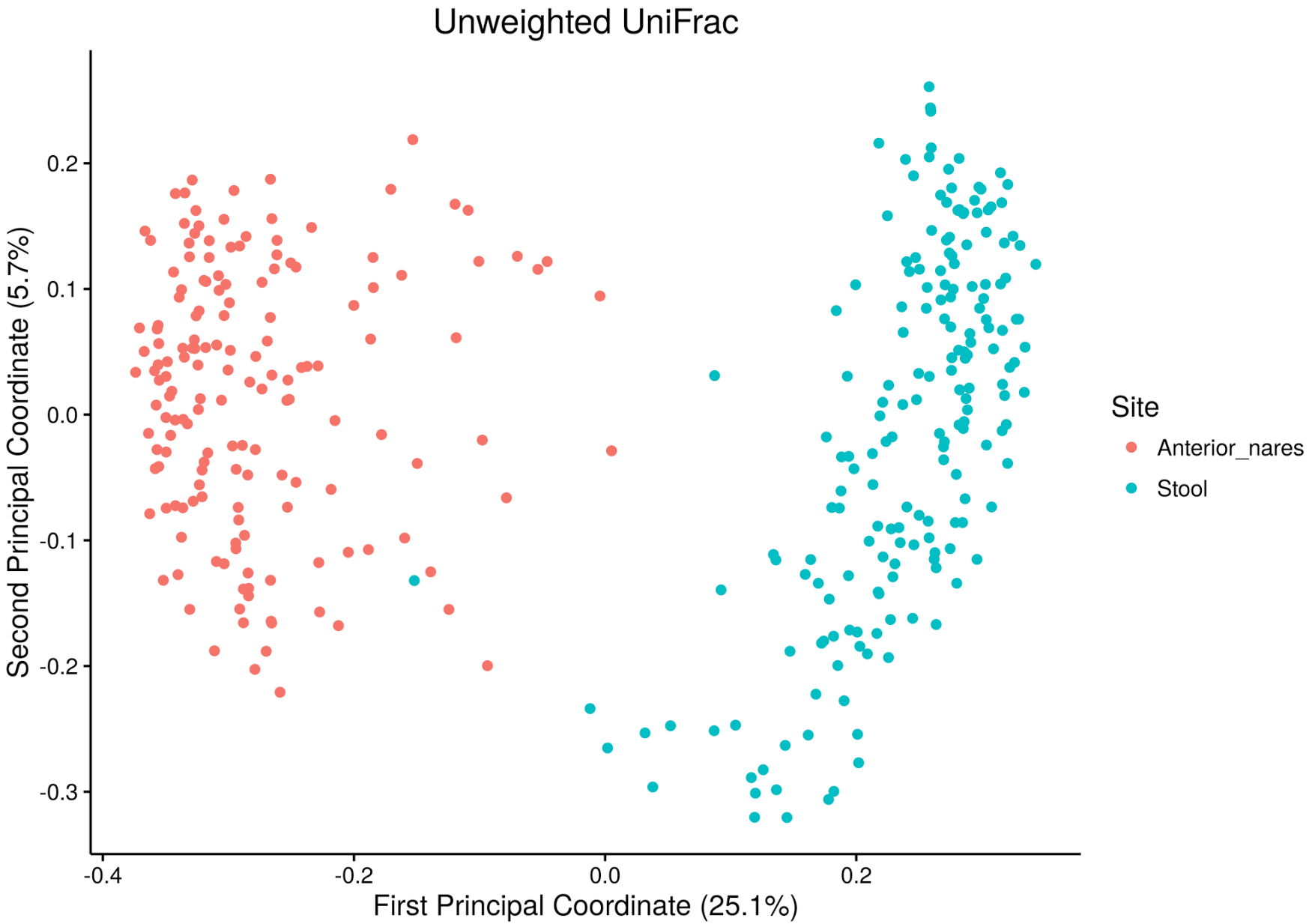


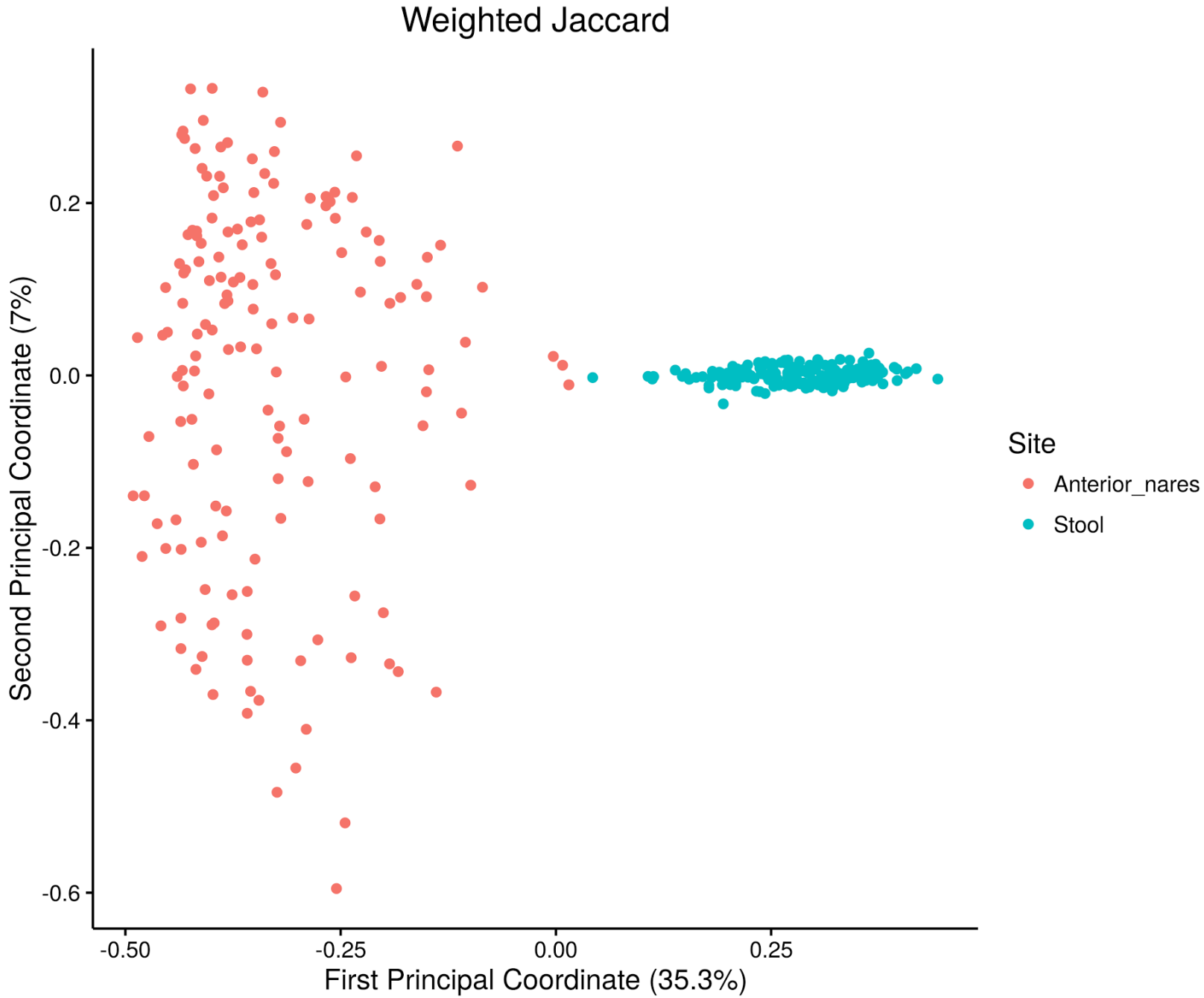
(b) Transformation-based approach (tb-PCA)

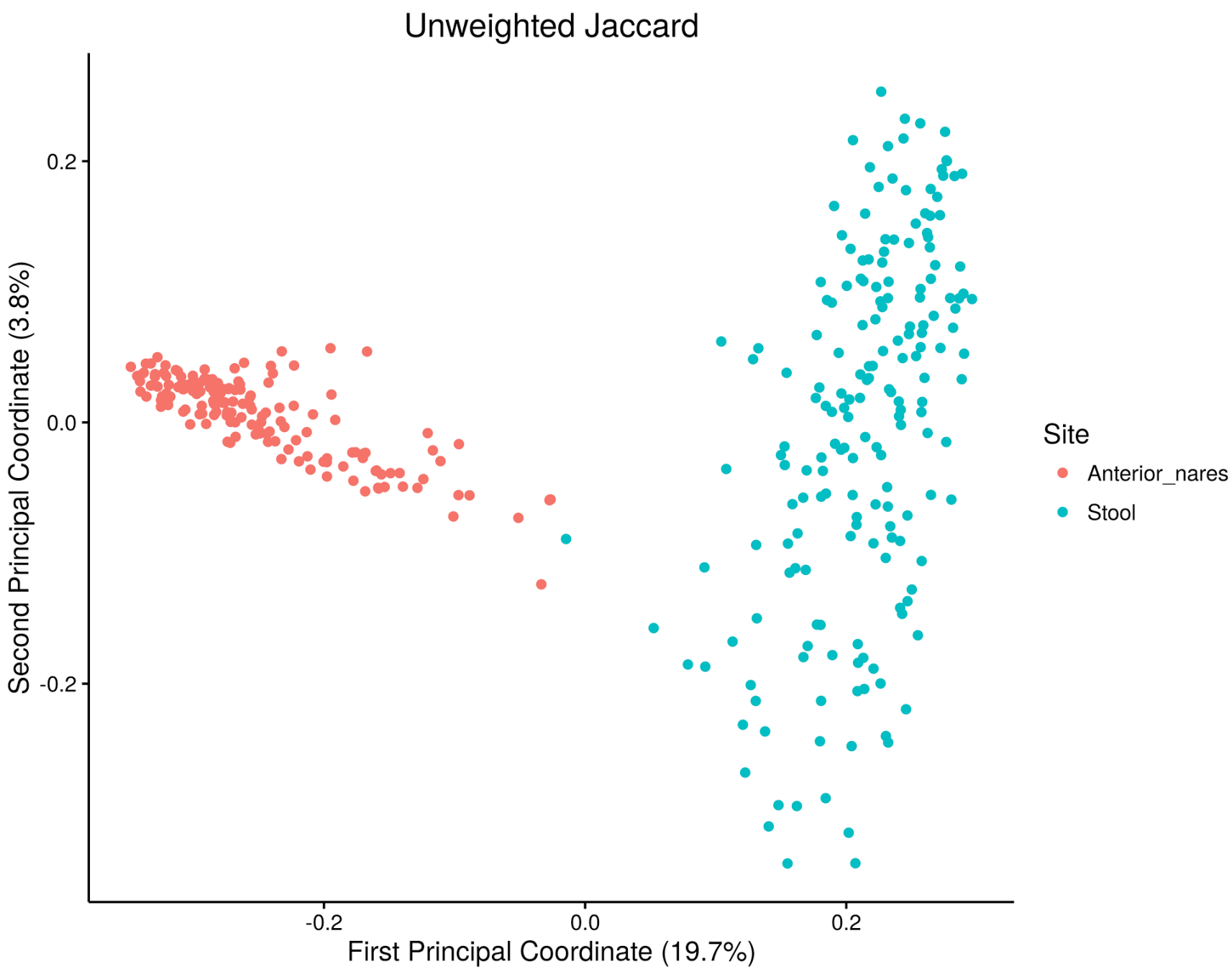
(c) Distance-based approach (PCoA)

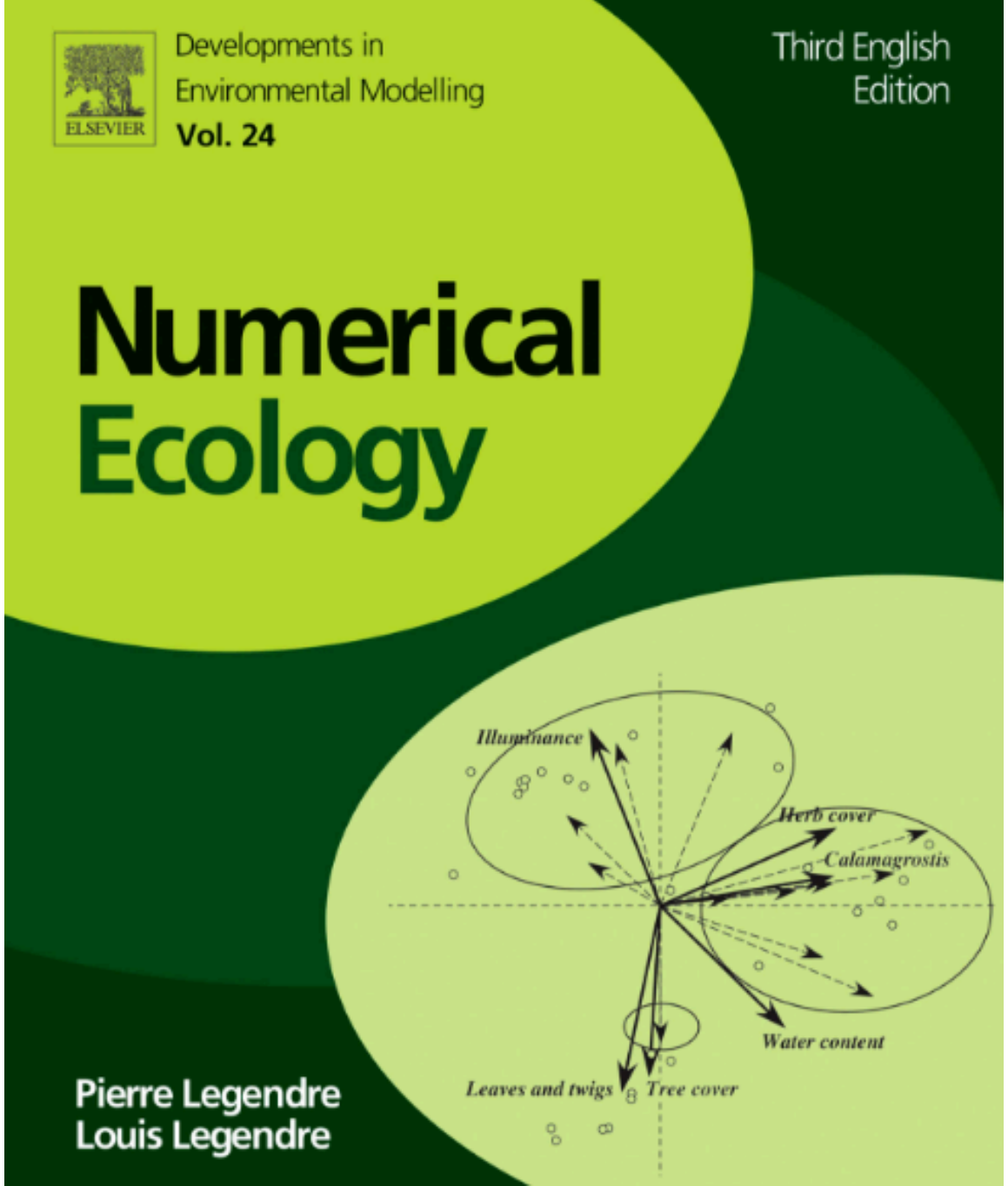



Weighted UniFrac
Second Principal Coordinate (3.1%)
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
First Principal Coordinate (23.7%)
Site
Anterior_nares
Stool


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

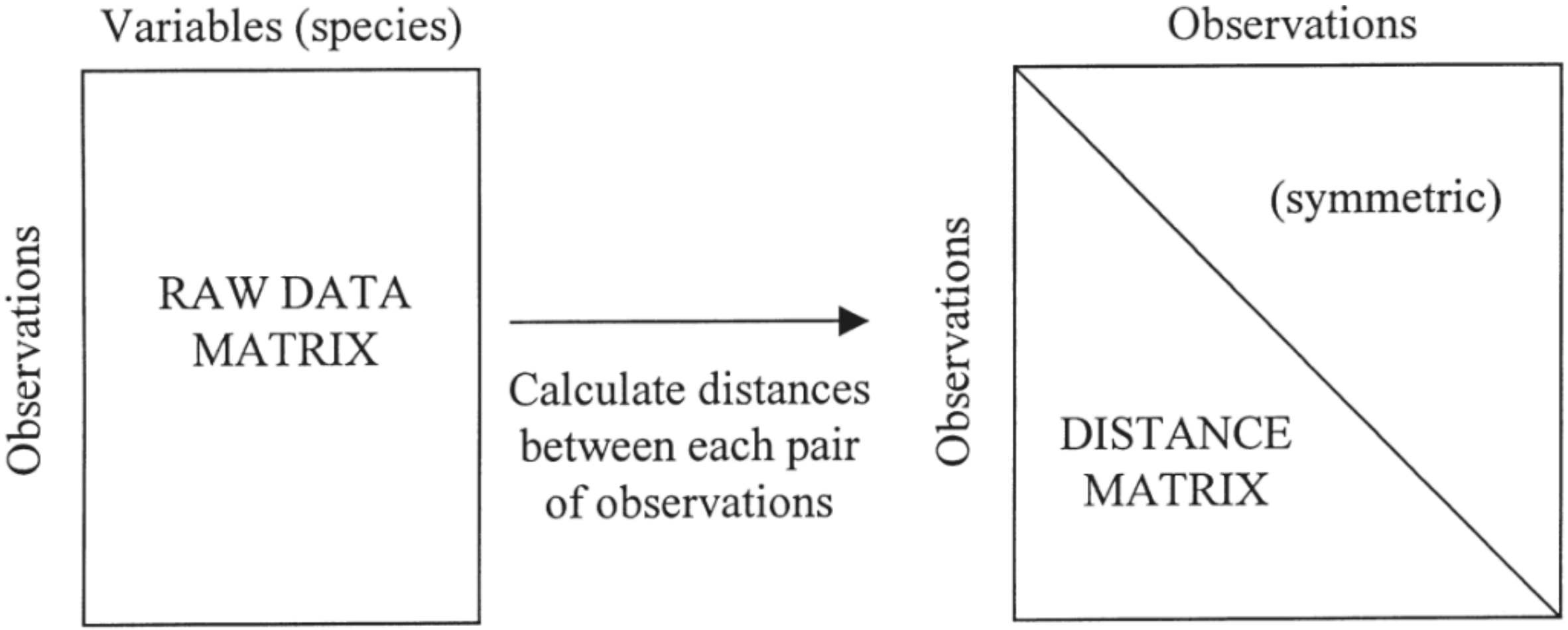

Weighted Jaccard
Second Principal Coordinate (7%)
0.2
0.0
-0.2
-0.4
-0.6
-0.50
-0.25
0.00
0.25
First Principal Coordinate (35.3%)
Site
Anterior_nares
Stool

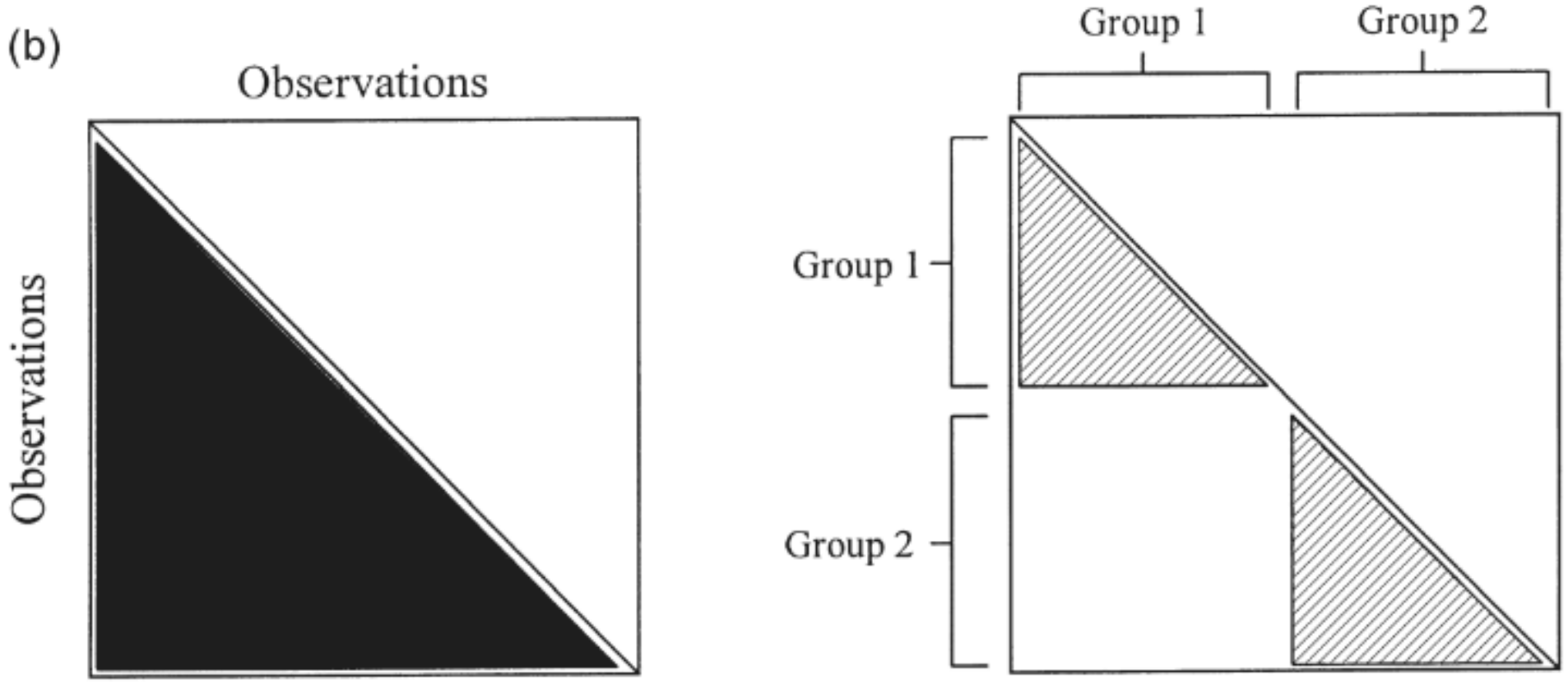

Unweighted Jaccard
Second Principal Coordinate (3.8%)
0.2
0.0
-0.2
-0.2
0.0
0.2
First Principal Coordinate (19.7%)
Site
Anterior_nares
Stool


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

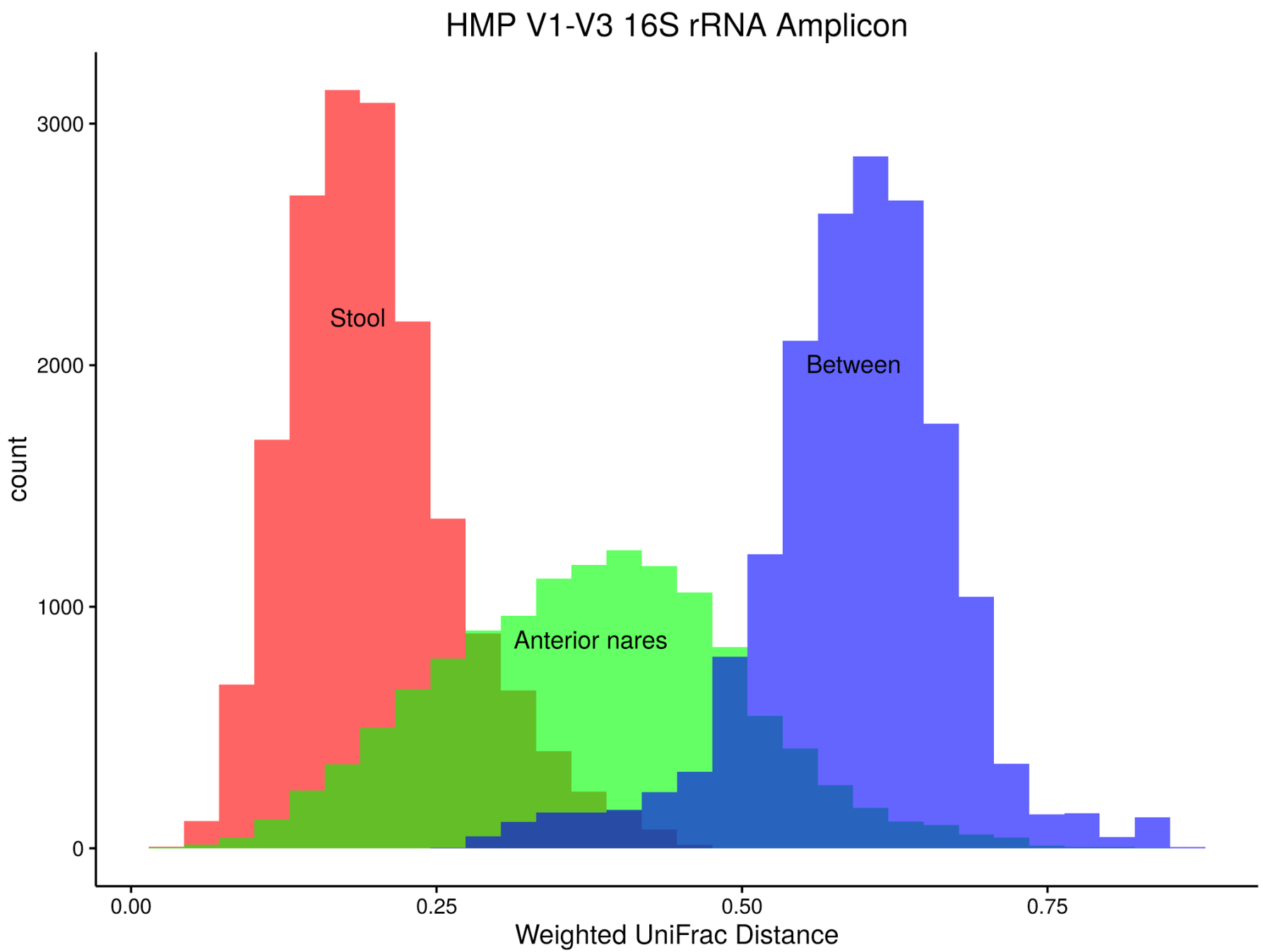
Dimension Reduction: PERMANOVA (adonis)

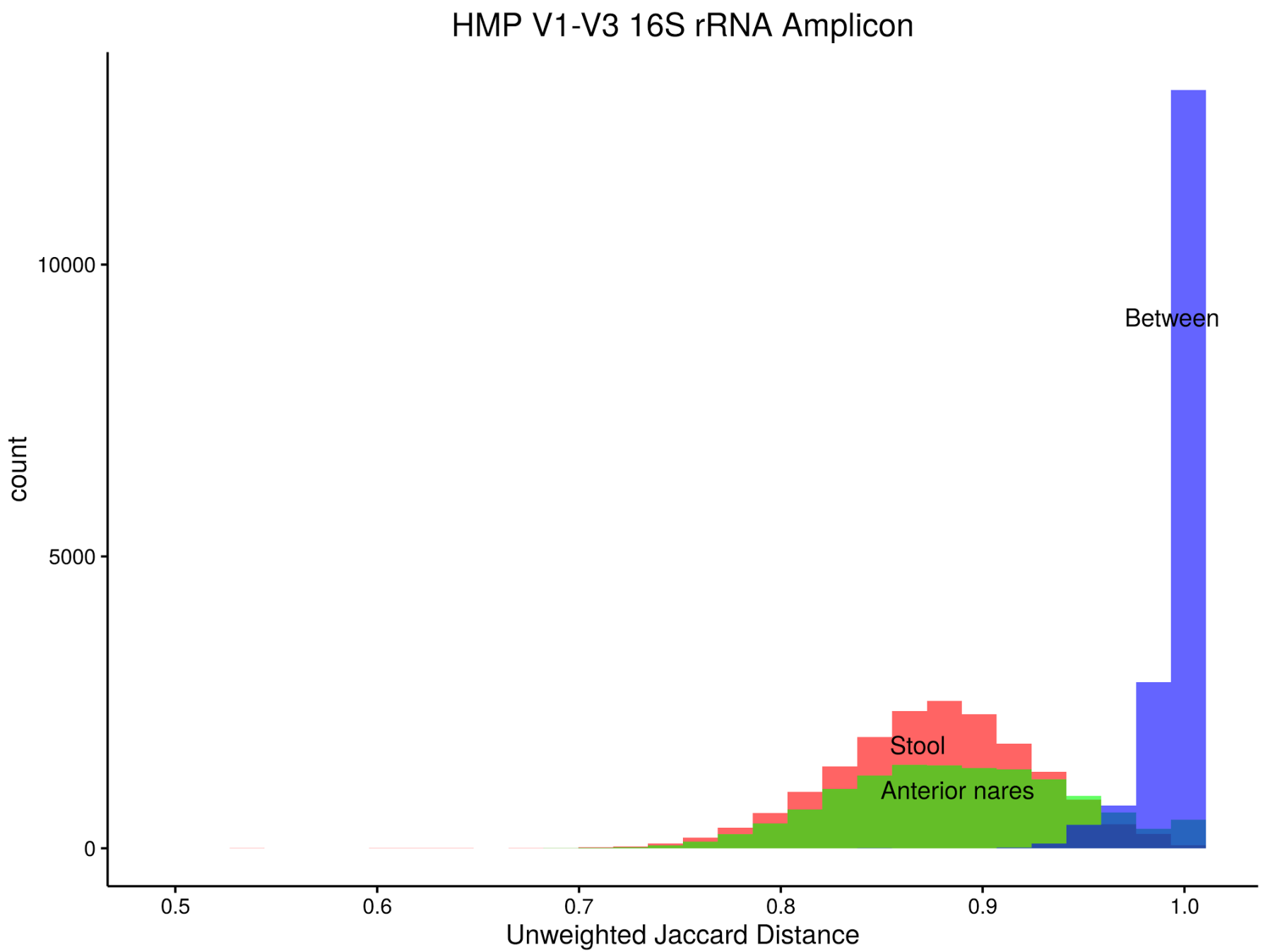
- Pairwise distance matrix can be partitioned by group assignment and ANOVA-like analysis can be applied to detect difference between groups.
- PERMANOVA: permutational ANOVA (aka, adonis)
  - pseudo F-ratio: conceptually similar but not F-distributed
  - testing by label permutation
  - quantification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

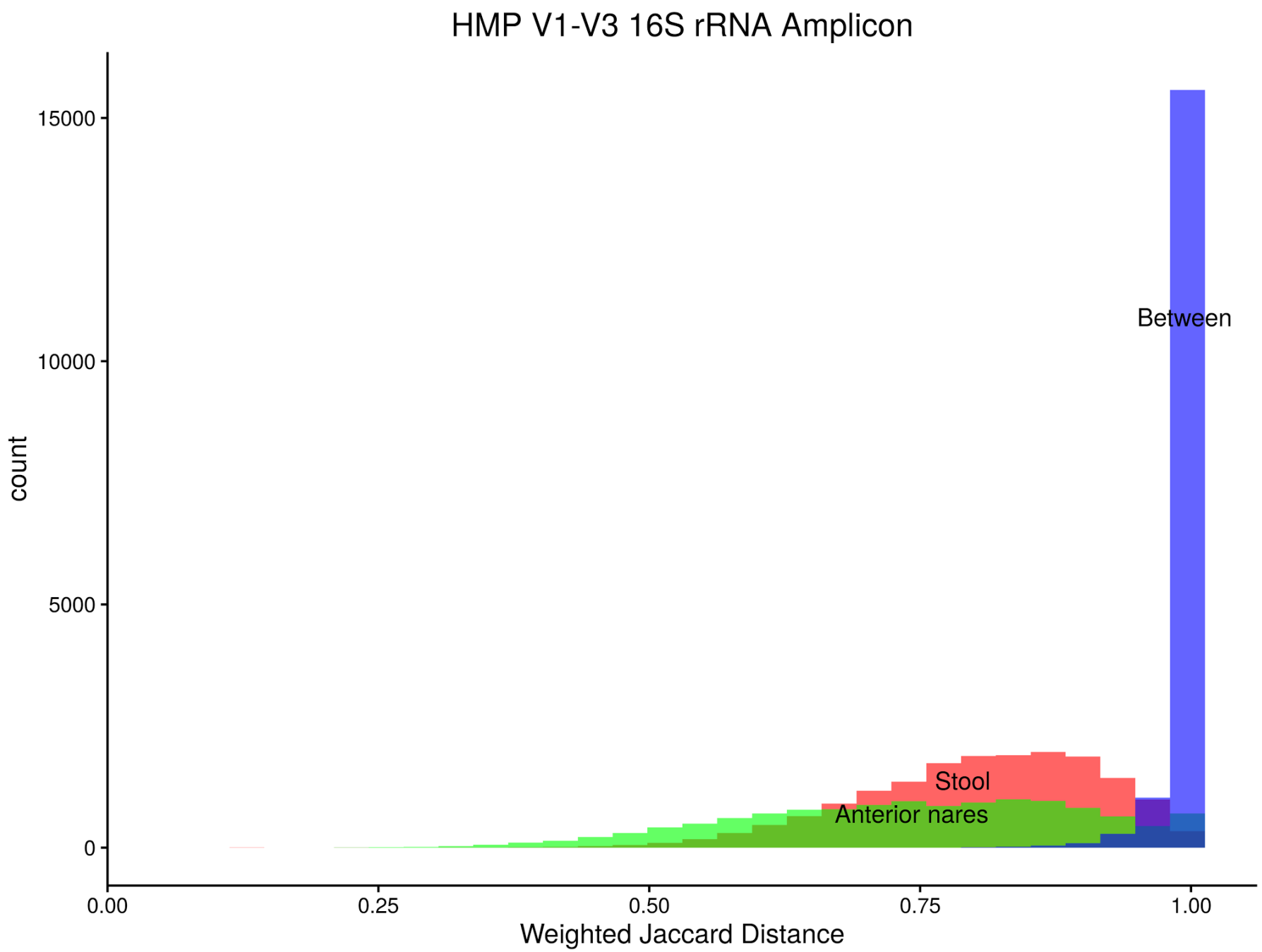

Variables (species)
Observations
RAW DATA
MATRIX
Calculate distances
between each pair
of observations
Observations
Observations
(symmetric)
DISTANCE
MATRIX

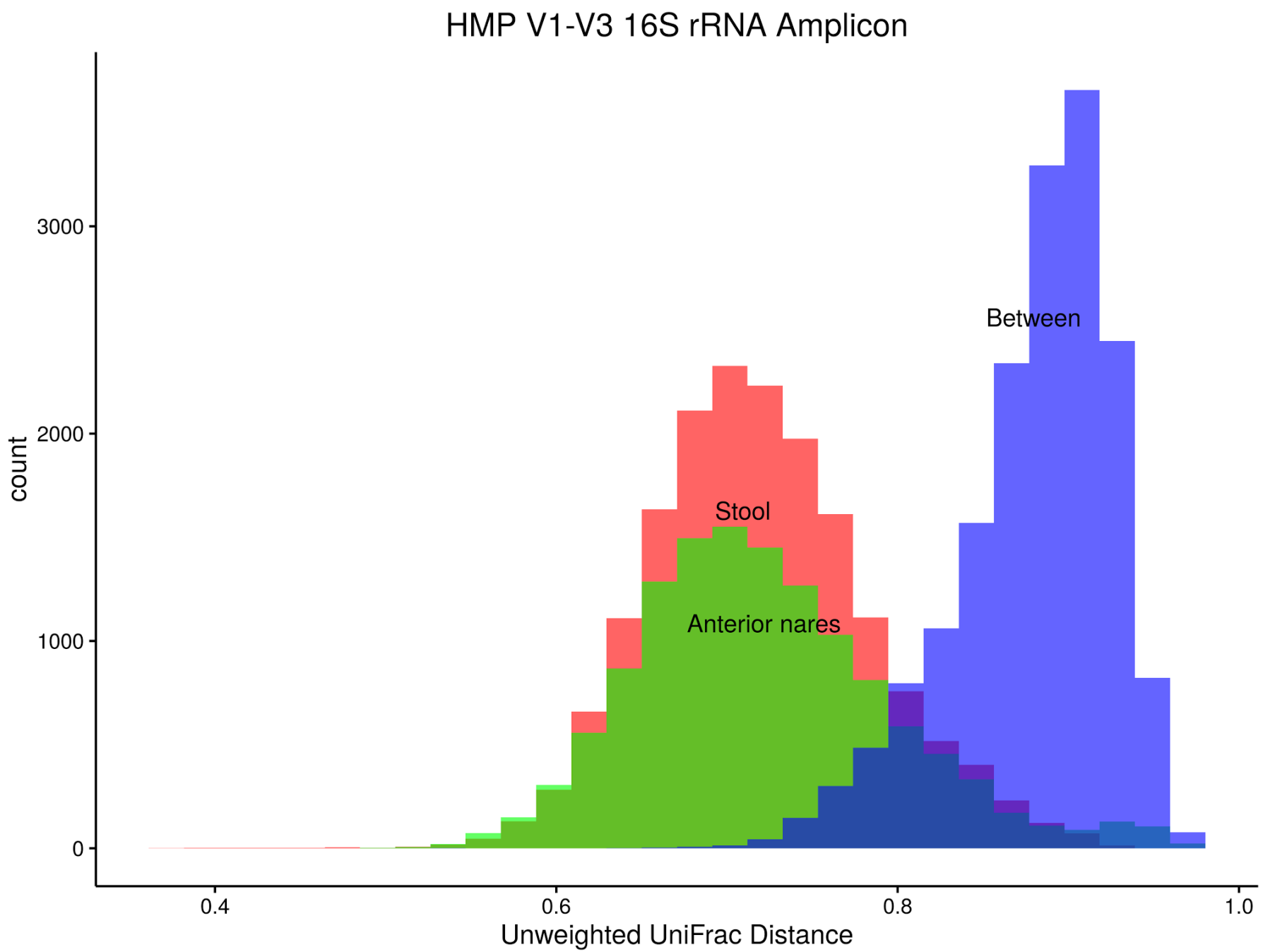


$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$


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
10000
5000
0
Between
Stool
Anterior nares
0.5
0.6
0.7
0.8
0.9
1.0
Unweighted Jaccard 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
3000
2000
1000
0
Between
Stool
Anterior nares
0.4
0.6
0.8
1.0
Unweighted UniFrac Distance

PERMANOVA: Effect Sizes

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

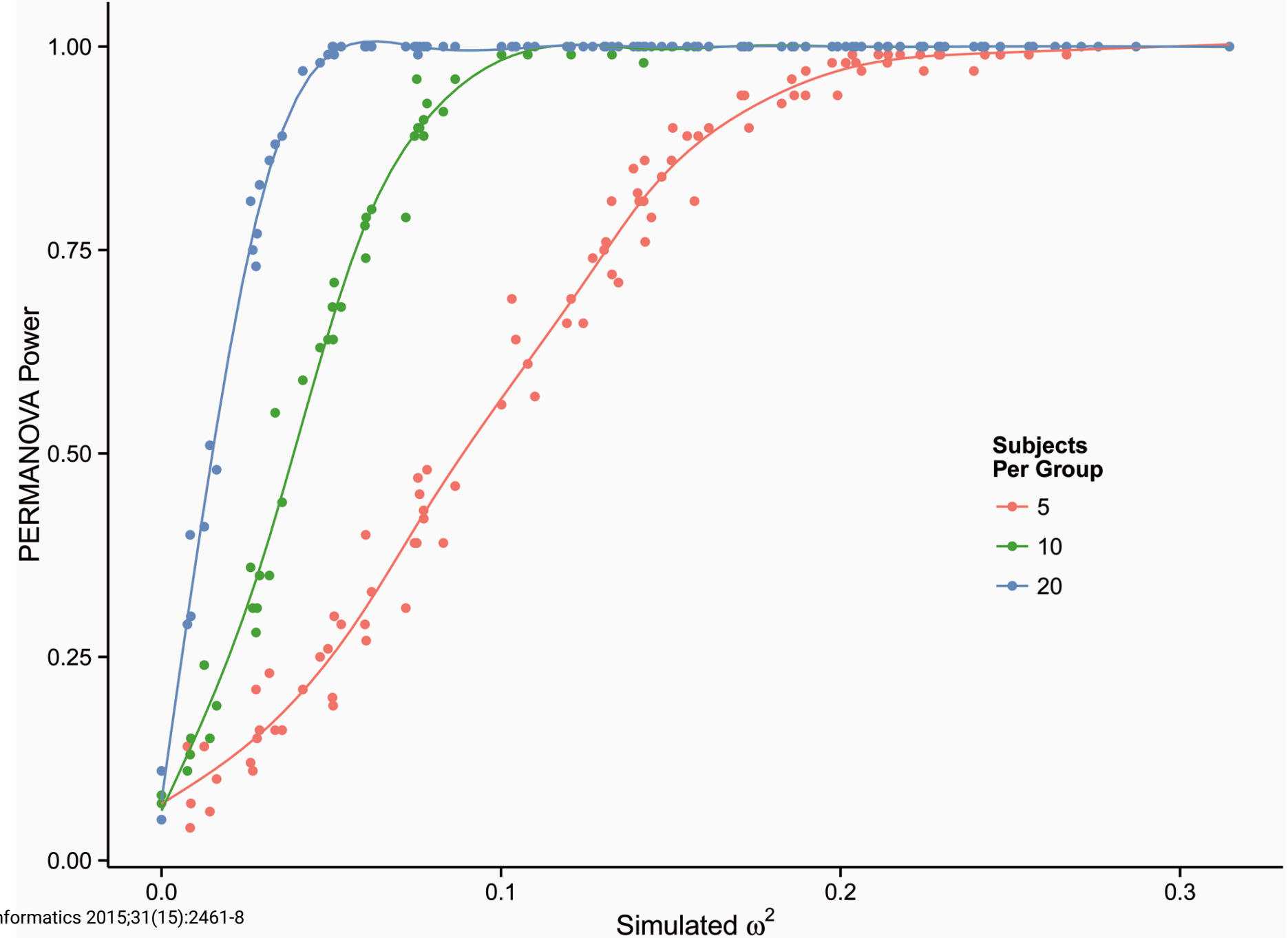
$$\omega^2 = \frac{SS_A - (a-1) \cdot \frac{SS_W}{N-a}}{SS_T + \frac{SS_W}{N-a}}$$

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)

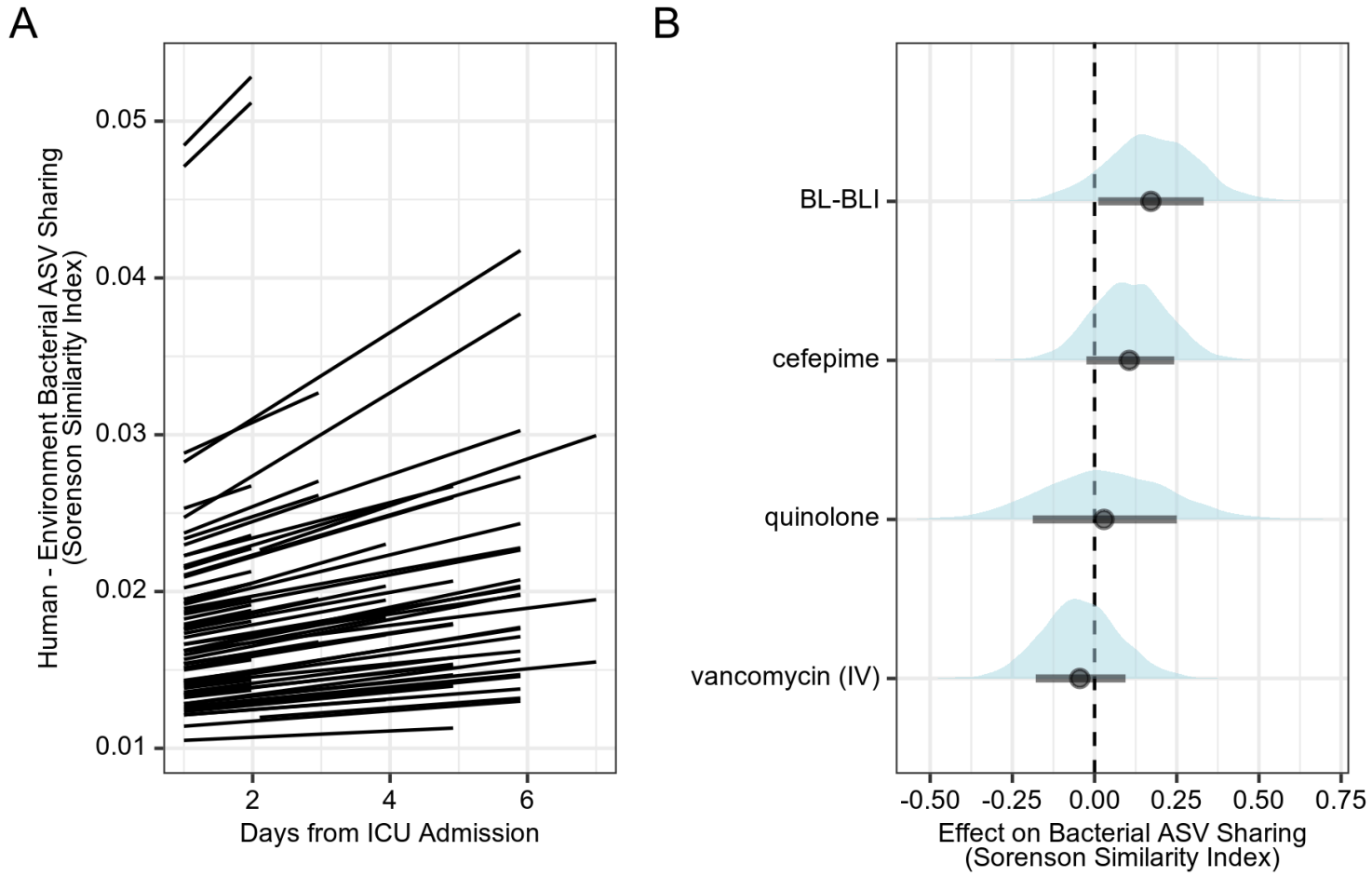
The range of observed effect sizes differs according to the metric of pairwise distance chosen for analysis. HMP data are shown to demonstrate a large effect (the degree of difference between two different human microbiome sampling sites) and a negligible effect (the degree of difference between skin sampling in the left versus right retroauricular crease)

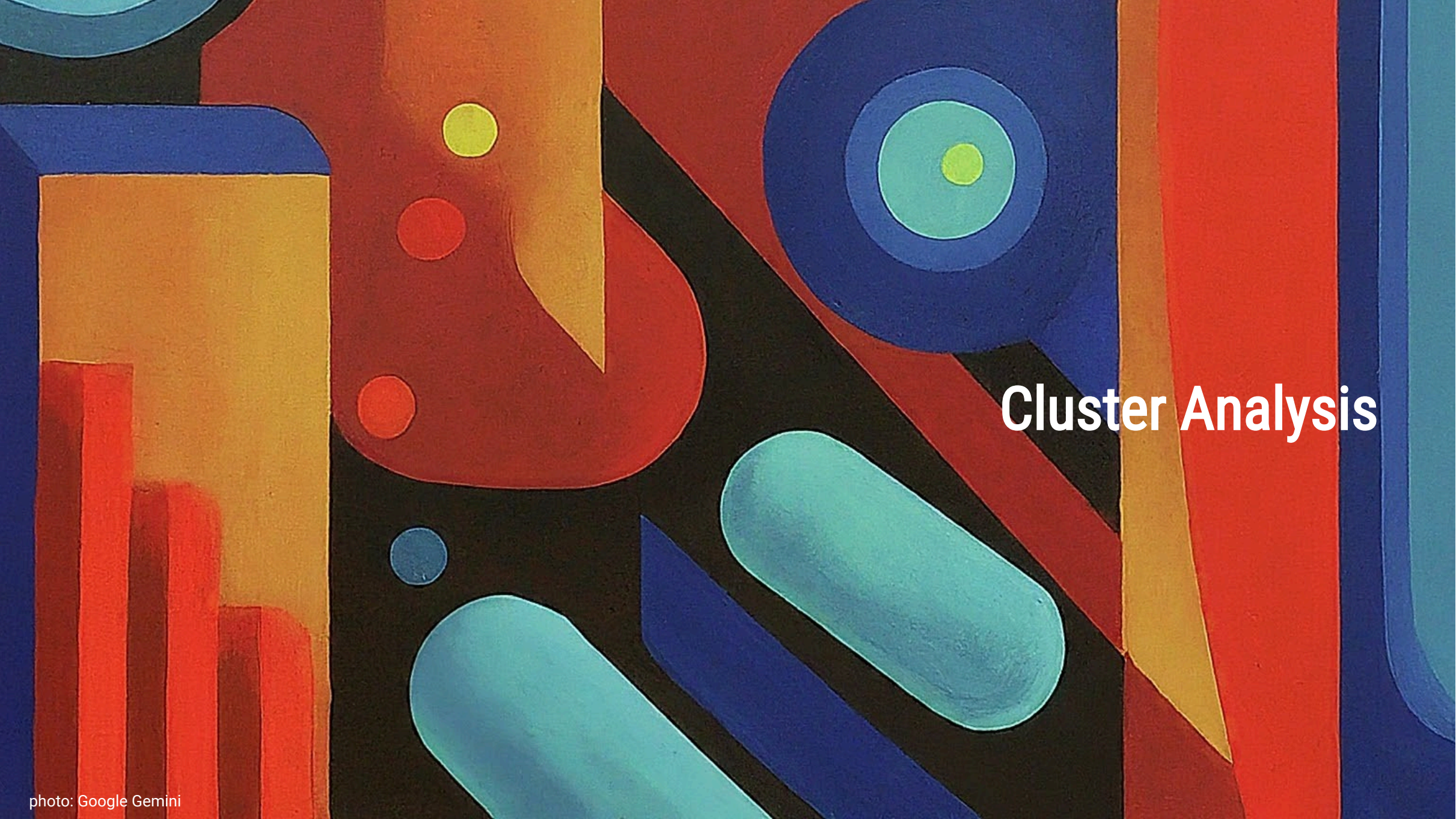
Kelly BJ et al. Bioinformatics 2015;31(15):2461-8



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

Other Approaches to Modeling Distances





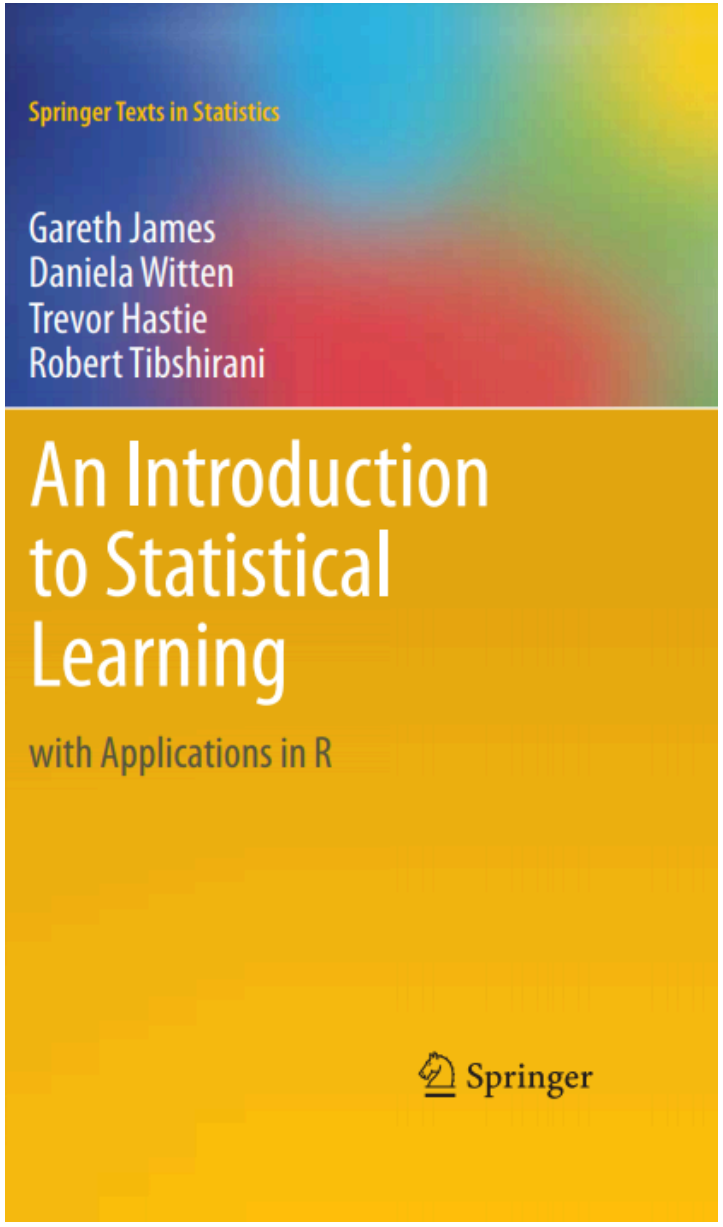
Cluster Analysis

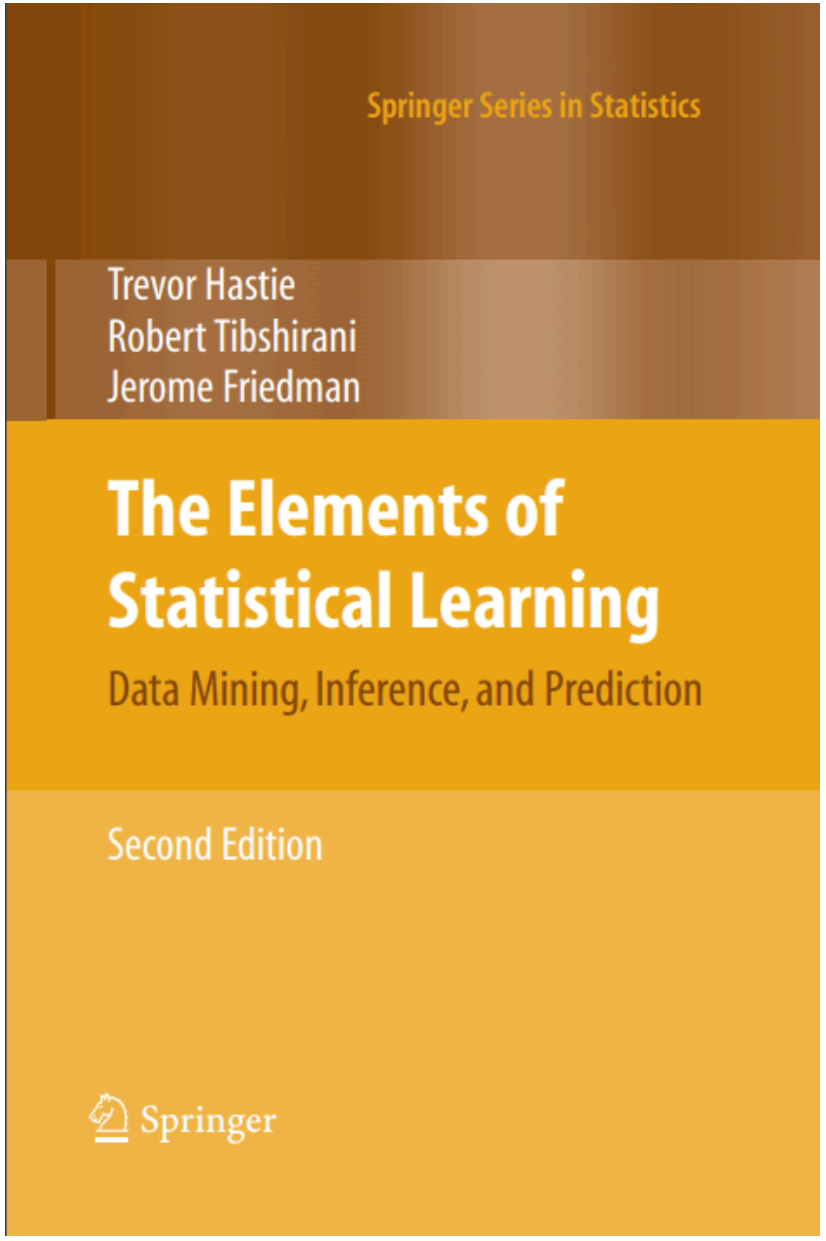
photo: Google Gemini

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

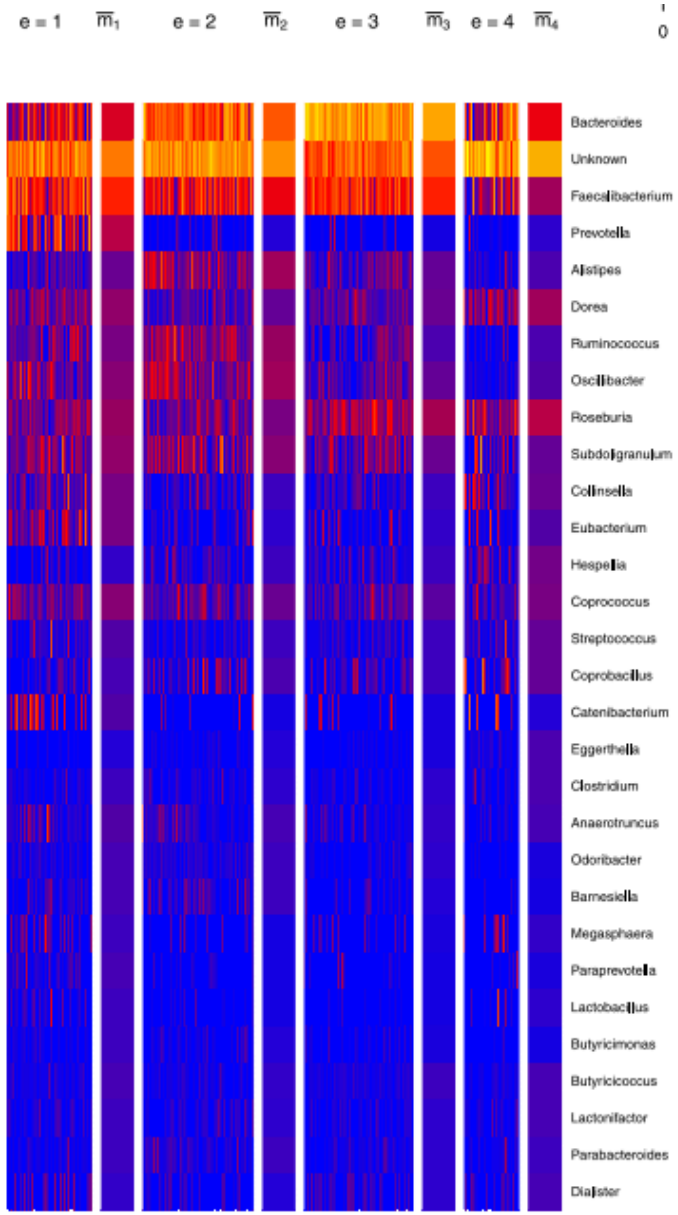

Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer

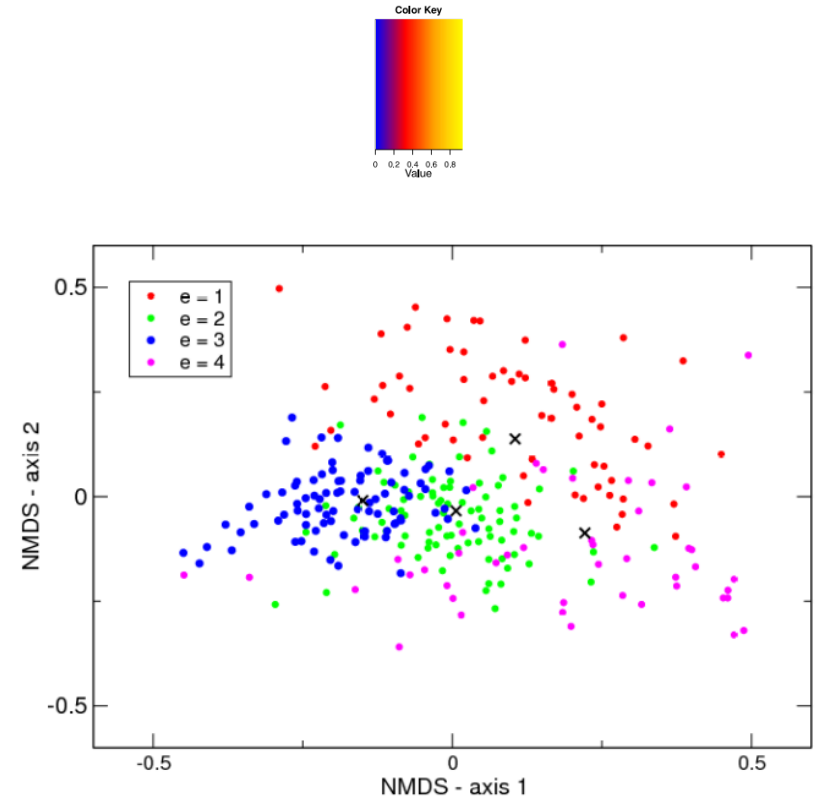

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Dimension Reduction: Mixture Models

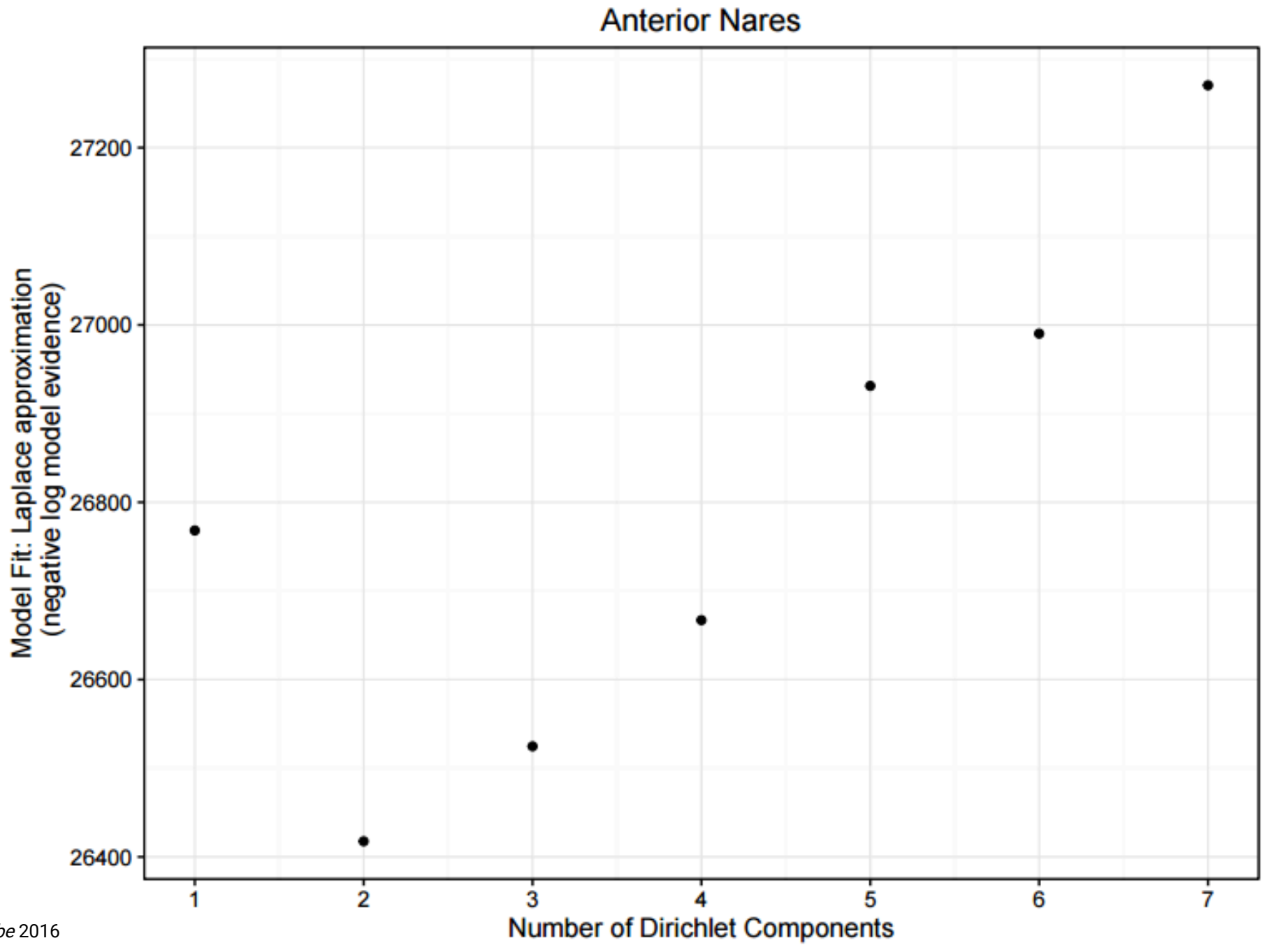
- Dirichlet-multinomial distribution:
  - compound probability distribution: probability vector drawn from Dirichlet distribution (generalized beta) → observation drawn from multinomial distribution (generalized binomial)
- D-M mixture modelling:
  - each sample ~ multinomial from one Dirichlet vector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

Holmes I et al *PLoS One* 2012

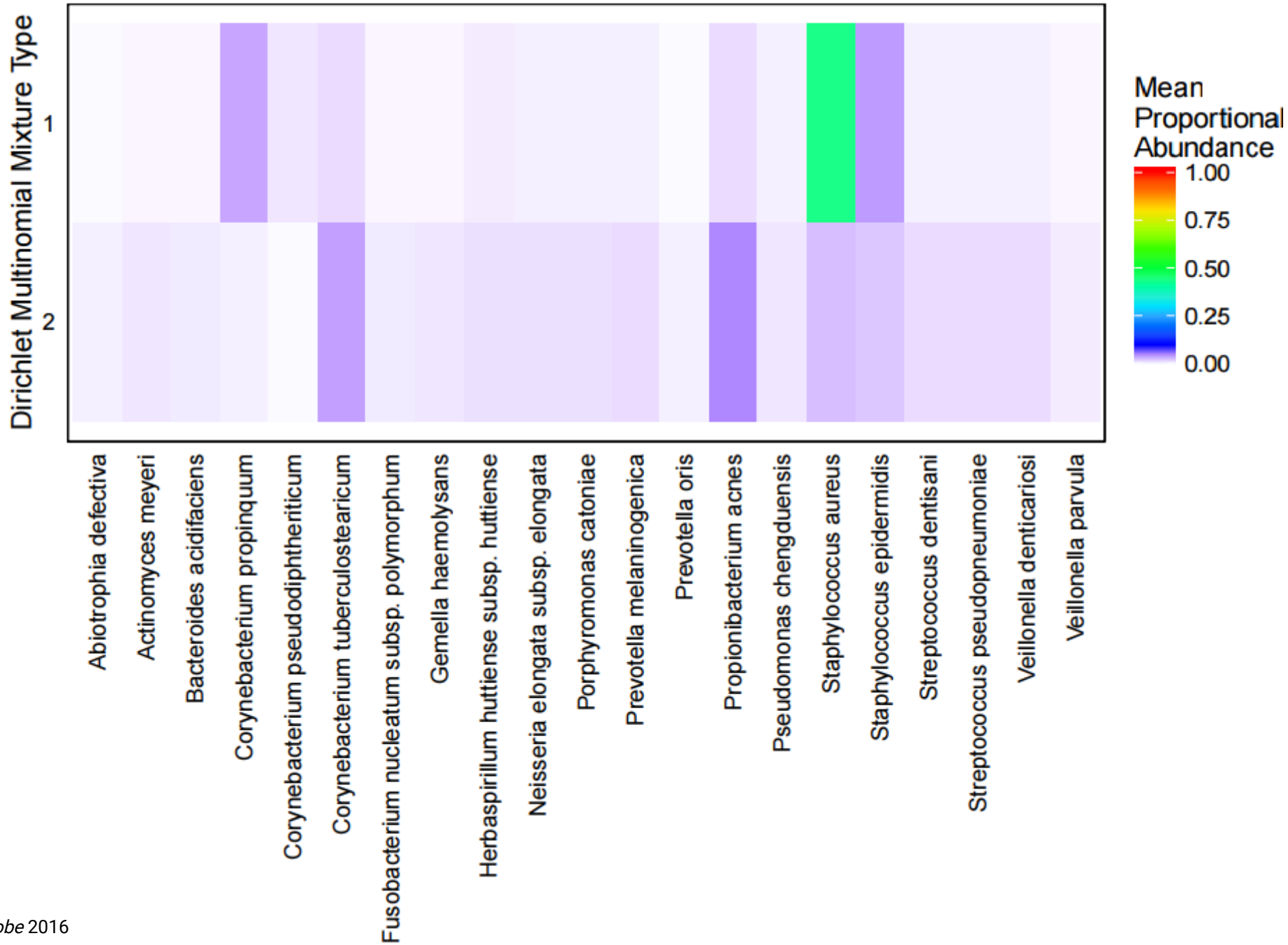




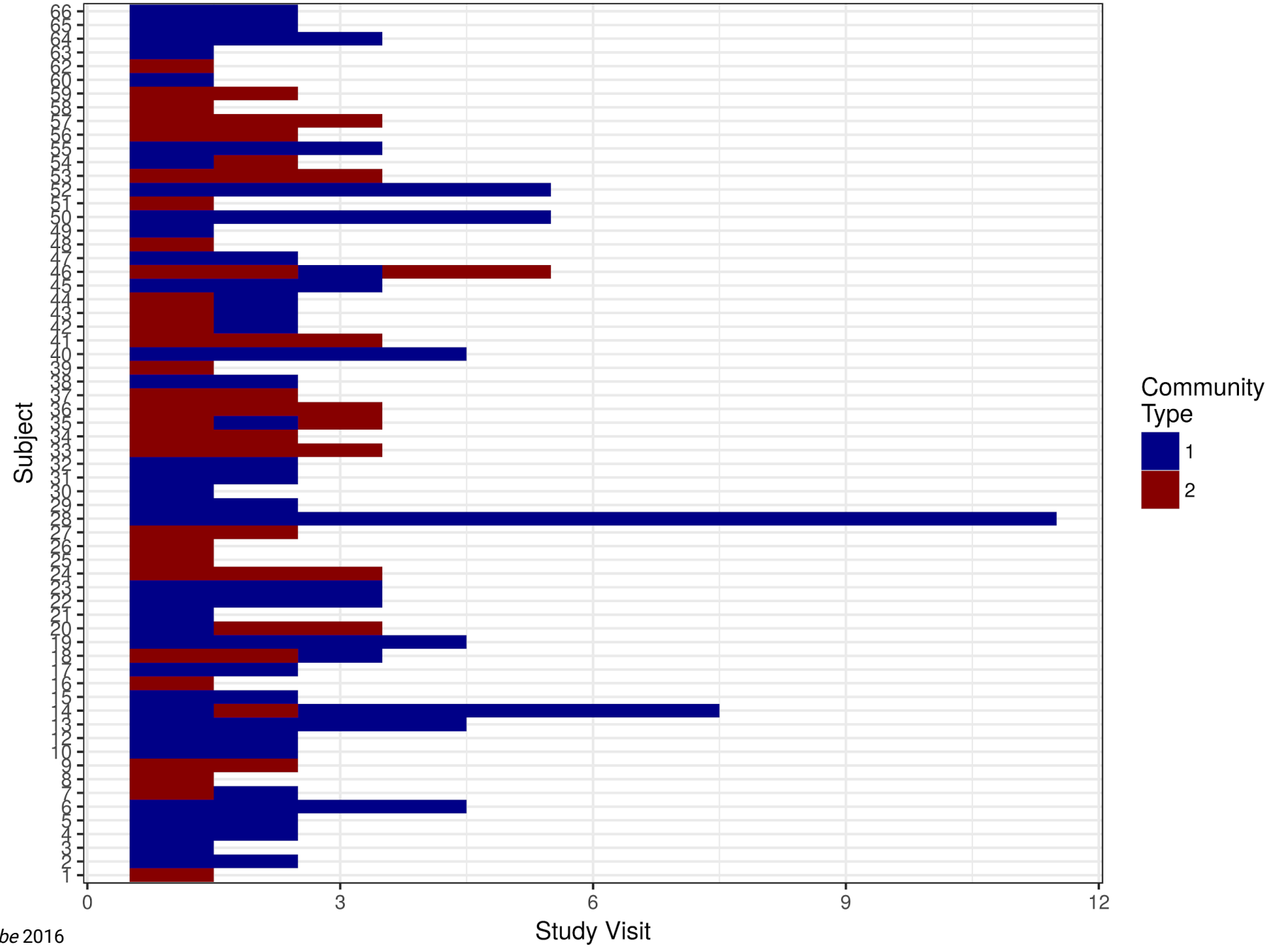
Holmes I et al *PLoS One* 2012



Kelly et al *ASM Microbe* 2016



Kelly et al *ASM Microbe* 2016



Kelly et al *ASM Microbe* 2016

Reading Selection

LETTER

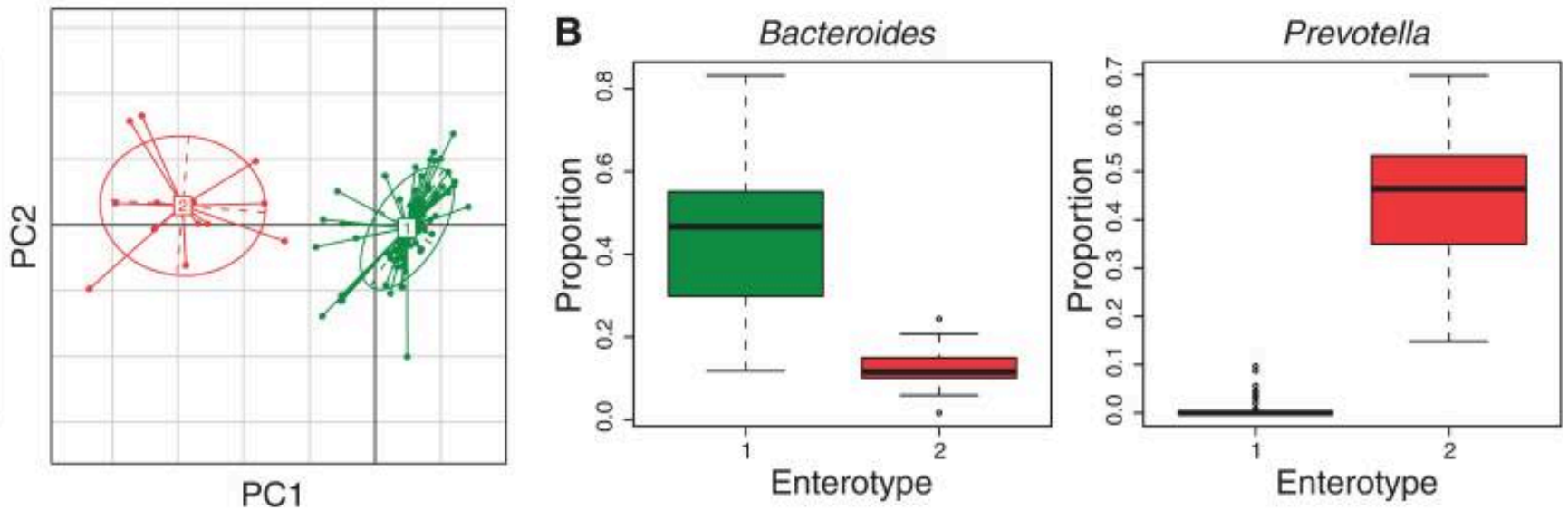
doi:10.1038/nature13178

Dynamics and associations of microbial community types across the human body

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

Ding & Schloss *Nature* 2014

Reference Standard for DMM Approach



Wu et al *Science* 2011

photo: Geranimo @unsplash

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

photo: Adrian Lange @unsplash