Differential abundance and correlation analysis of microbiome data: Challenges and some solutions

Huang Lin PhD
(huang.lin@nih.gov)
Biostatistics & Bioinformatics Branch
Division of Intramural Population Health Research

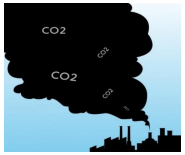
NIH *Eunice Kennedy Shriver* National Institute of Child Health and Human Development

Human Health

External factors

Diet

Physical activity



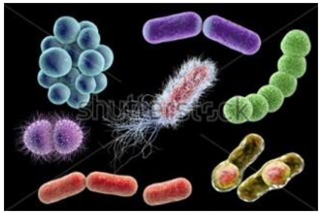
Chemical exposure

Genes

Genes
X
External factors
X
Internal environment

Time

Internal environment: Microbiome



Cells:

- ~10 trillion <u>human cells</u>
- ~100 trillion <u>microbial cells</u>

Genes:

- ~20,000 <u>human genes</u>
- ~2 to 20 million <u>microbial genes</u>

Elephant in the gut!!

Microbiome data …

From Ecosystem to Sample

Ecosystem (e.g., gut):

↓

A random sample

- 16S rRNA gene sequencing and library preparation
- Read counts for each Operational Taxonomic Unit (OTU)/Amplicon Sequence Variant (ASV)
- OTUs/ASVs can be further summarized at different phylogenetic levels (species, family, genus, etc.)

The Data: Abundance Table

Taxon	Sample 1	Sample 2	…	
			…	
			…	
…	…	…	…	…
			…	
Library size			…	

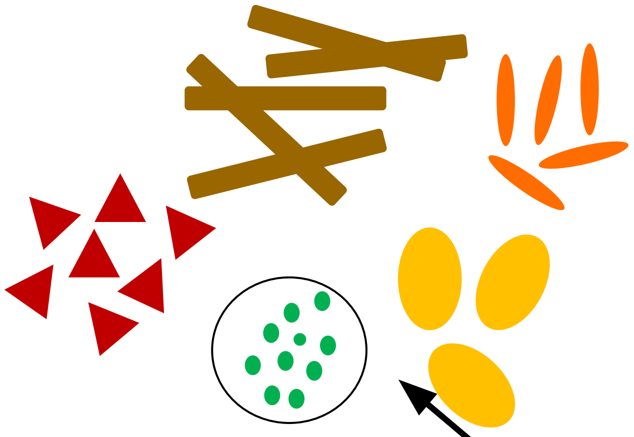
Challenges …

Compositionality

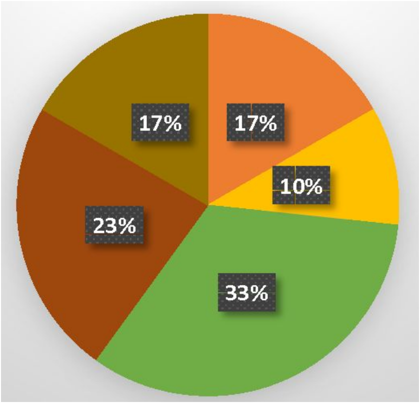
- High-throughput sequencing (HTS) can deliver reads only up to the capacity of the instrument
  - Observed data are relative quantities
  - Hence compositional, i.e., data in a simplex
  - The sum of observed abundance = a fixed constant

A single taxon can change all relative abundances

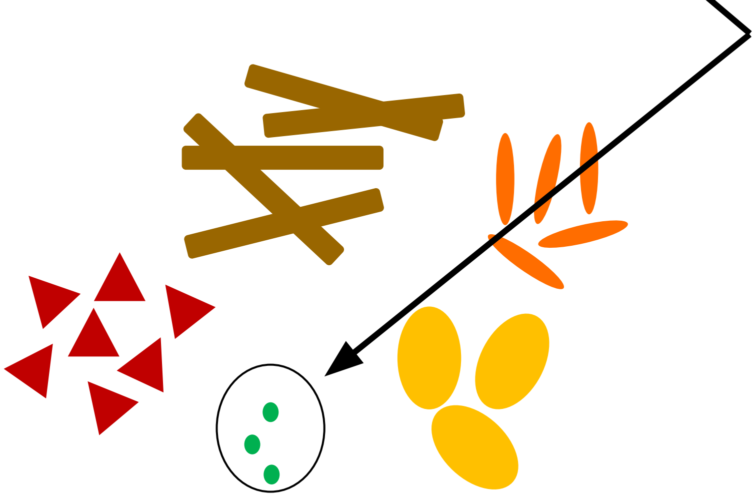
Abundance of 5 taxa: Ecosystem I



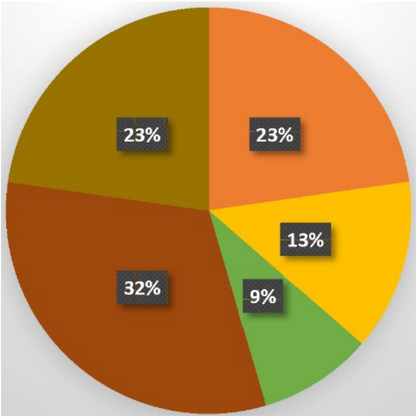
Relative abundance of 5 taxa: Ecosystem I



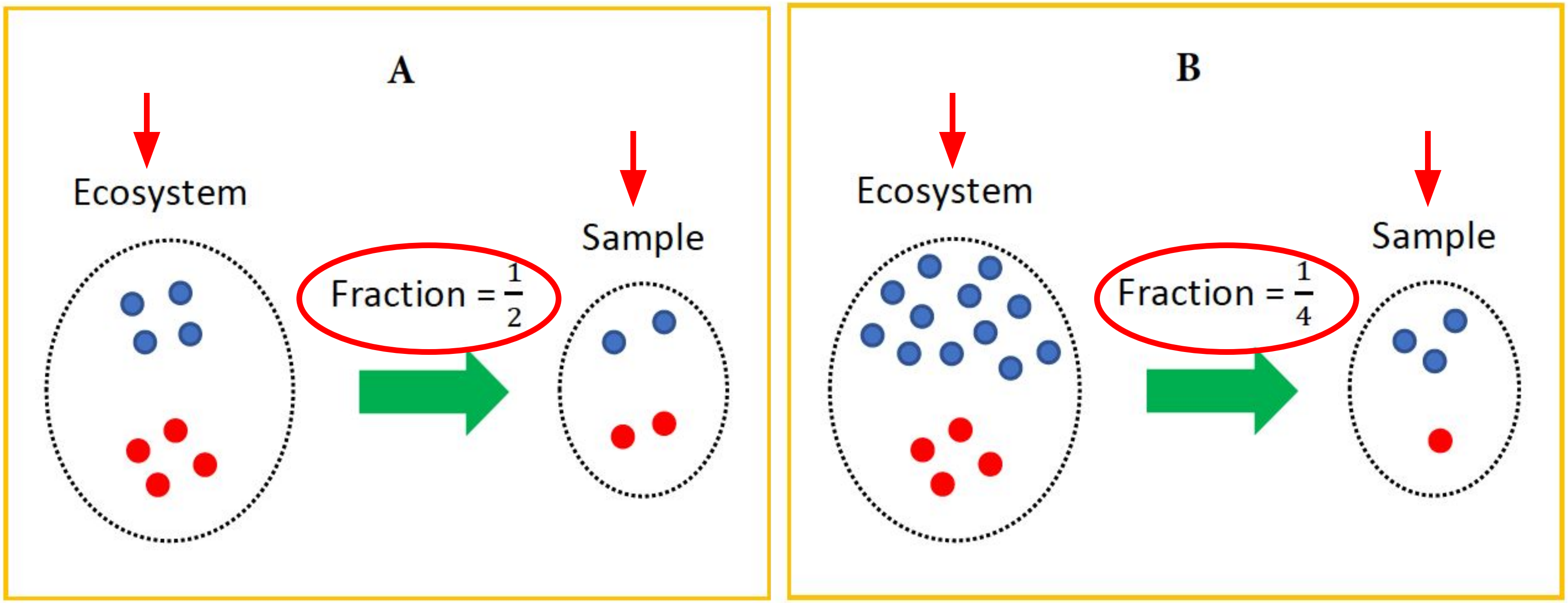
Abundance of 5 taxa: Ecosystem II



Relative abundance of 5 taxa: Ecosystem II



Differential Sampling Fractions



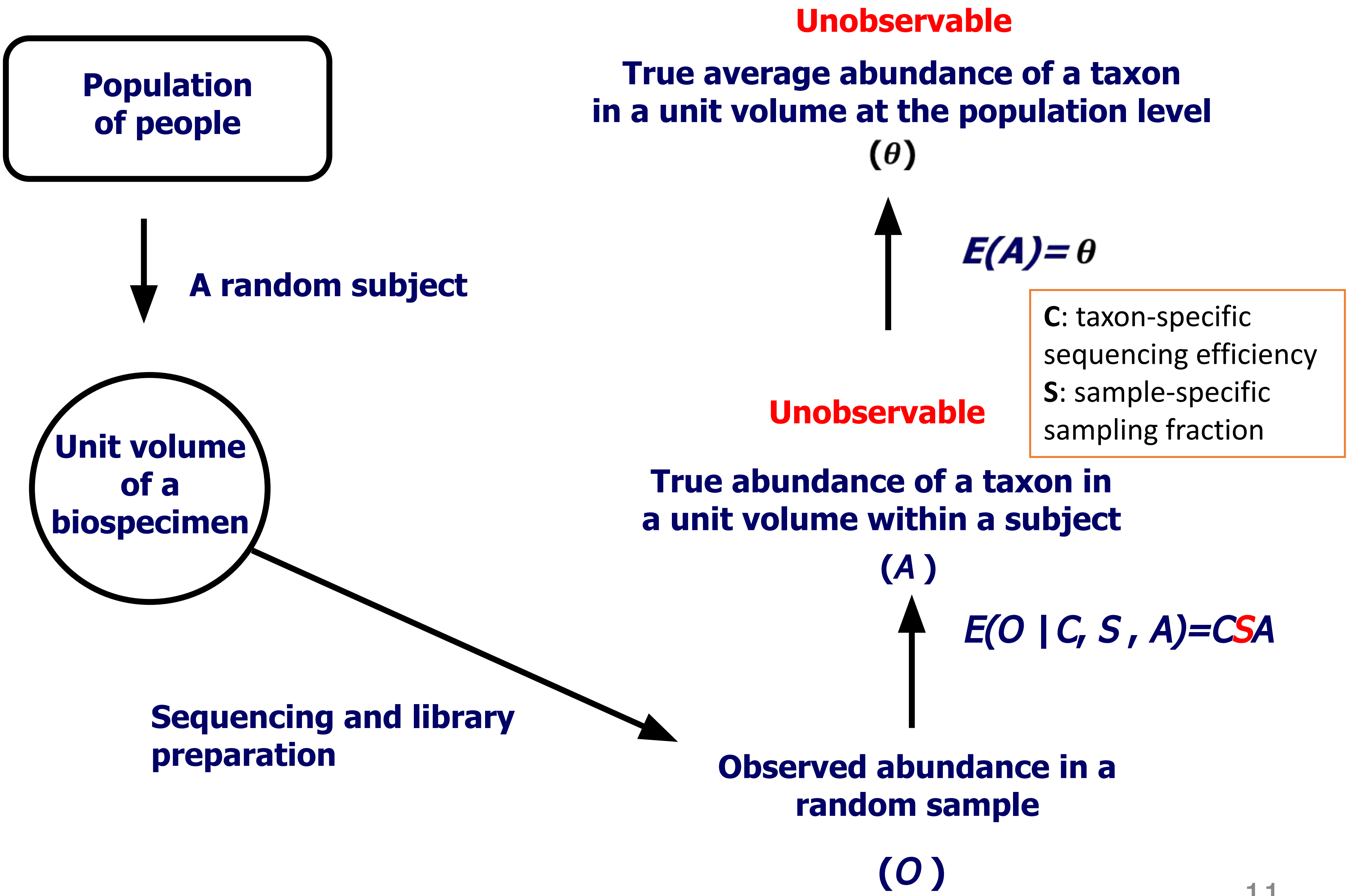
	A		B	
	Blue	Red	Blue	Red
Ecosystem	4	4	12	4
Sample	2	2	3	1

False Positive

Differential Abundance (DA) Analysis …

Lin & Peddada (2020), *Nature Communications*
Lin & Peddada (2020), *NPJ biofilms and microbiomes*

The Set-Up



Population of people

A random subject

Unit volume of a biospecimen

Sequencing and library preparation

Observed abundance in a random sample

(O)

$E(O \mid C, S, A) = CSA$

Unobservable

True abundance of a taxon in a unit volume within a subject

(A)

C: taxon-specific sequencing efficiency

S: sample-specific sampling fraction

$E(A) = \theta$

Unobservable

True average abundance of a taxon in a unit volume at the population level

(θ)

Sampling Fraction *S*

Some popular scaling methods to deal with sampling fraction ***s***

1. DESeq2: MED
2. edgeR: UQ, TMM
3. metagenomeSeq: CSS
4. Wrench

An implicit assumption: A large proportion of features are not differentially expressed.

This may be reasonable for gene expression studies but may not be valid for microbiome.

Lin & Peddada (2020), *NPJ biofilms and microbiomes*

ANCOM-BC Model

Statistical formulation for two groups:

Multiplicative model:

- Observed Abundance $=$ sequencing efficiency × sampling fraction × $\boldsymbol{\theta}$ × random error
- $O = C \times S \times \boldsymbol{\theta} \times \eta$

Additive model (log transformed):

- $o = c + s + \mu + \epsilon$

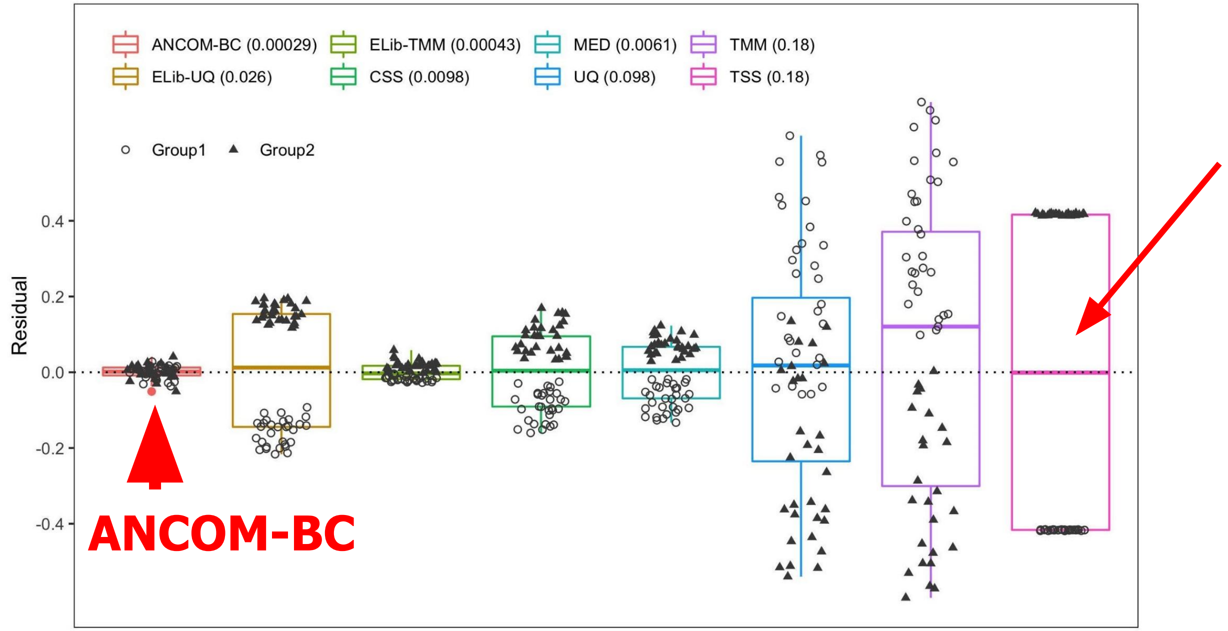
Assumption: Some taxa are non-differentially abundant

Sample specific biasing constant

Lin & Peddada (2020), *Nature Communications*

Simulation Studies

Performance of various normalization methods



ANCOM-BC

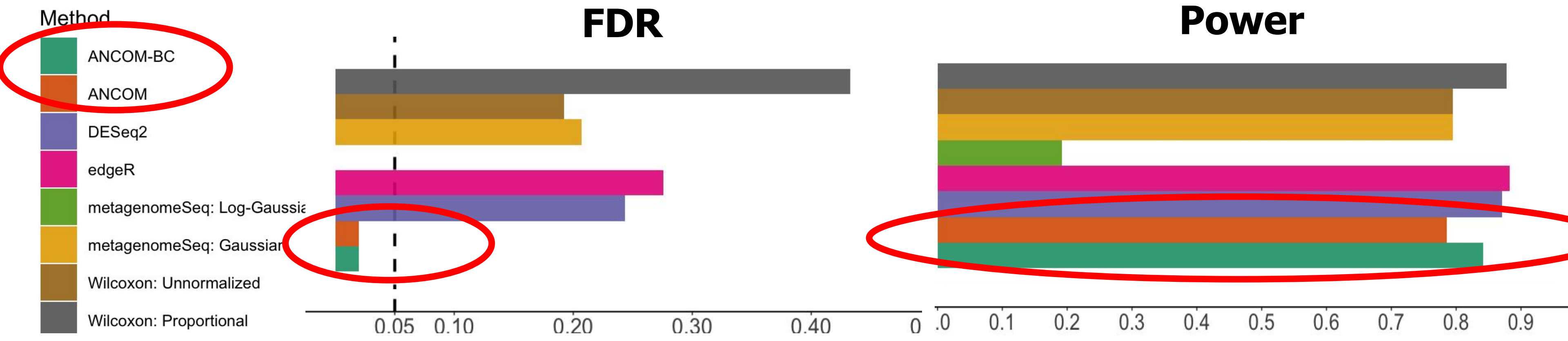


Illustration of ANCOM-BC … role of microbiome and microbial byproducts, namely the short chain fatty acids in HIV

Chen, Lin et al. (2021), *Microbiome*, to appear

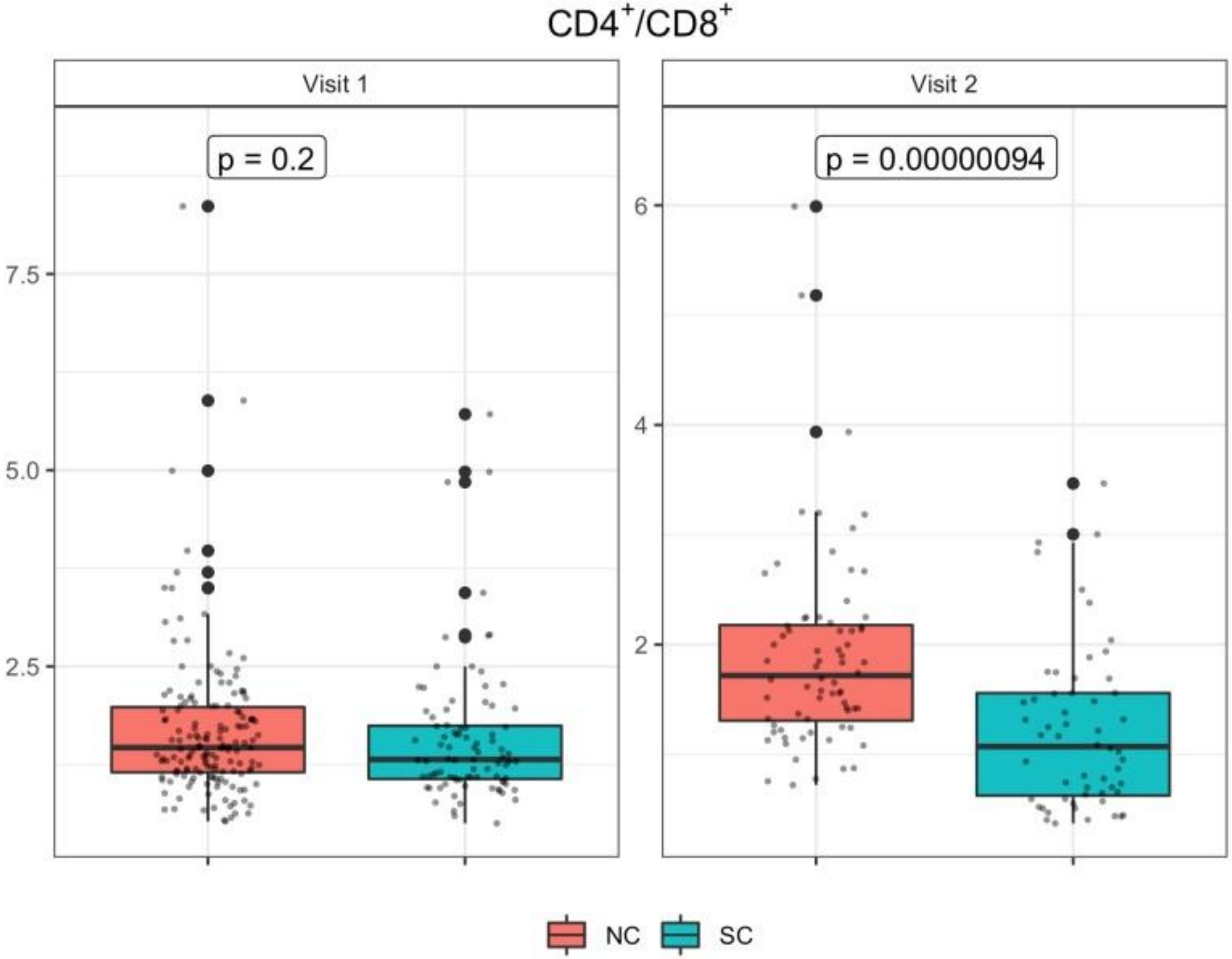
Data on Untreated Male HIV Patients from 1980's

265 Men – ages ranged from 19 to 80

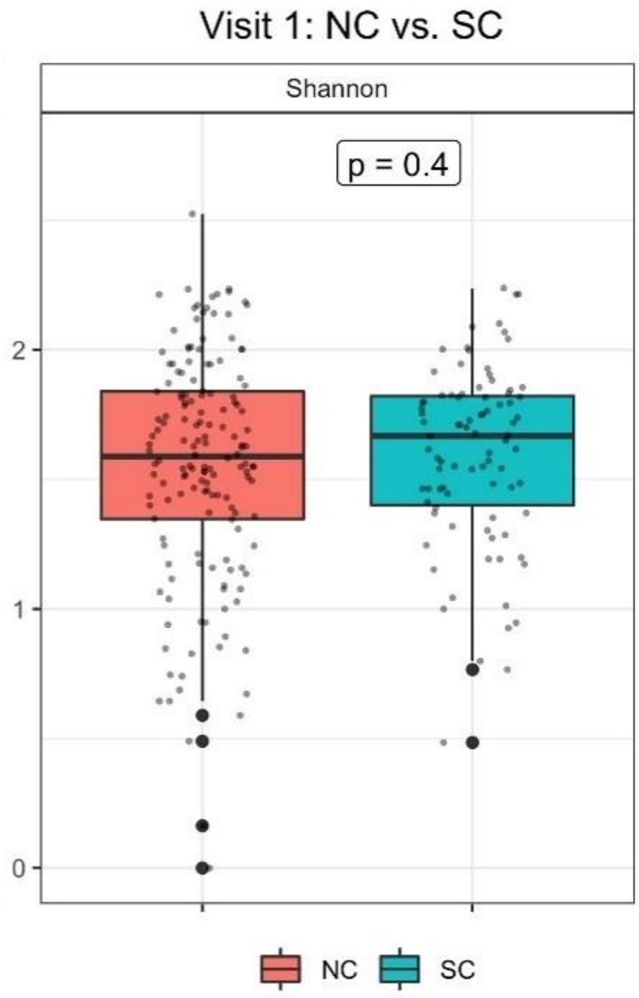
Locations: Baltimore, Chicago, Pittsburgh, LA

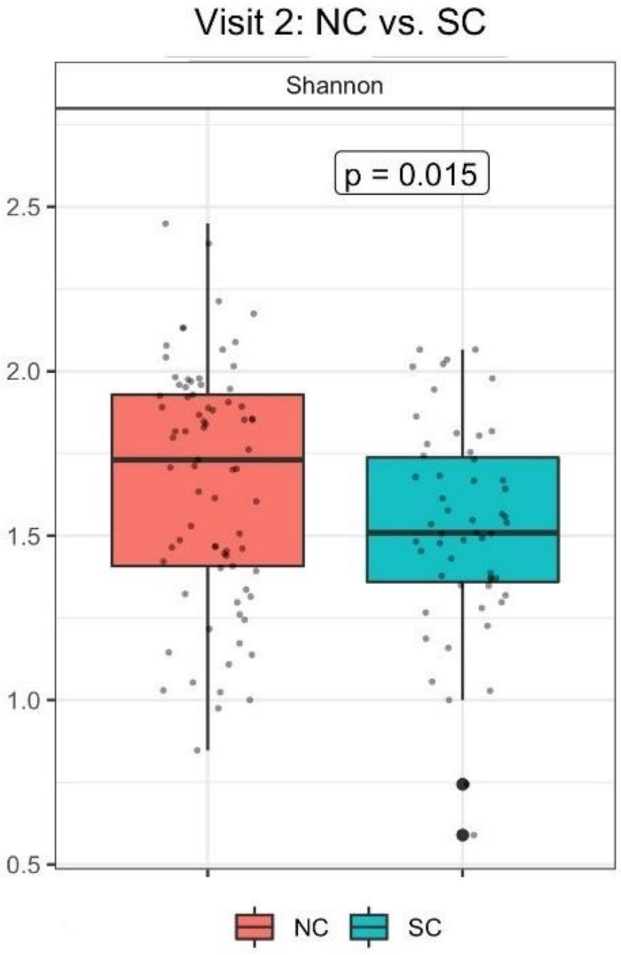
- **Visit 1:** No one seroconverted
- **Visit 2:**
  - 109 men seroconverted (SC)
  - 156 did not seroconvert – Negative Controls (NC)

$CD4^{+}/CD8^{+}$ Ratio of SC and NC at Visit 1/2



Differences in Alpha Diversity of Microbiome





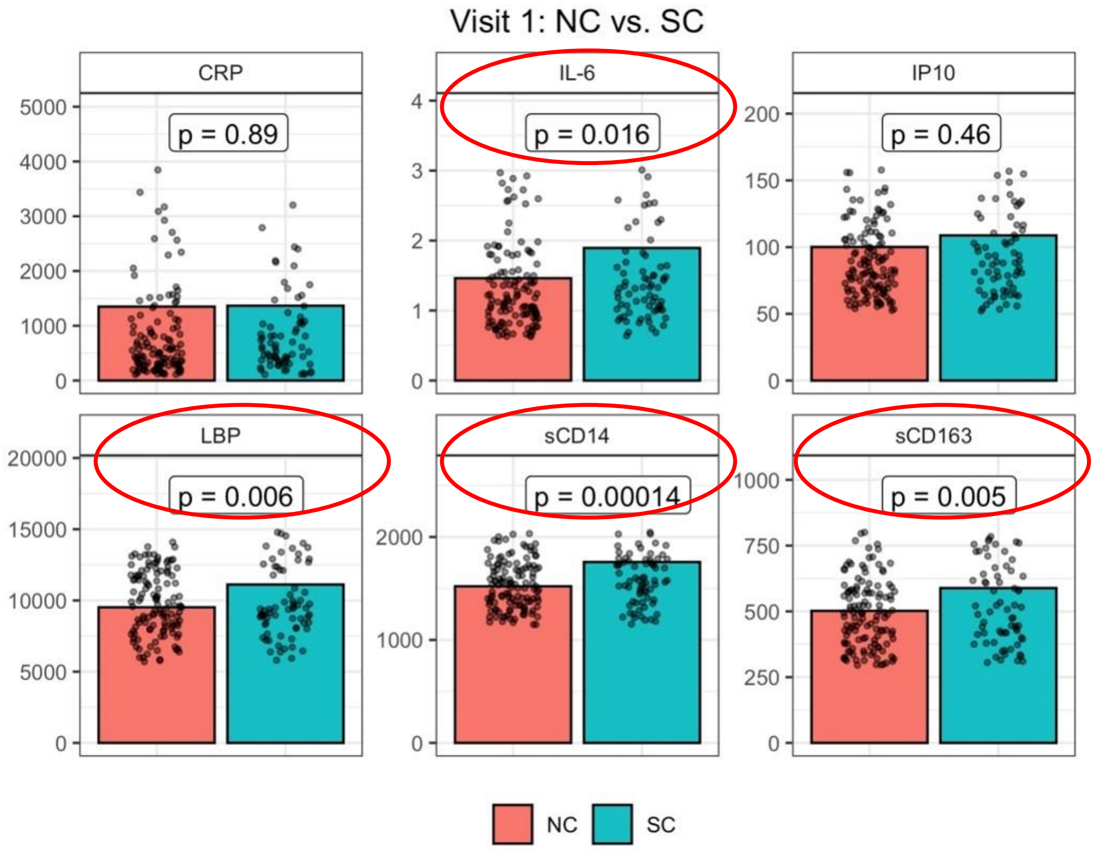
Summary

- At the first visit no one is seroconverted, and there is no difference in
  - $CD4^+/CD8^+$ ratio
  - Microbial alpha diversity

- At the second visit SC group seroconverted but not the NC groups. There is a significant reduction in SC compared to NC in
  - $CD4^+/CD8^+$ ratio
  - Microbial alpha diversity

Questions …

Although there are no differences between the SC and NC group during the first visit because none of them seroconverted, but were men who later seroconverted develop immune deficiency before seroconversion?



Questions …

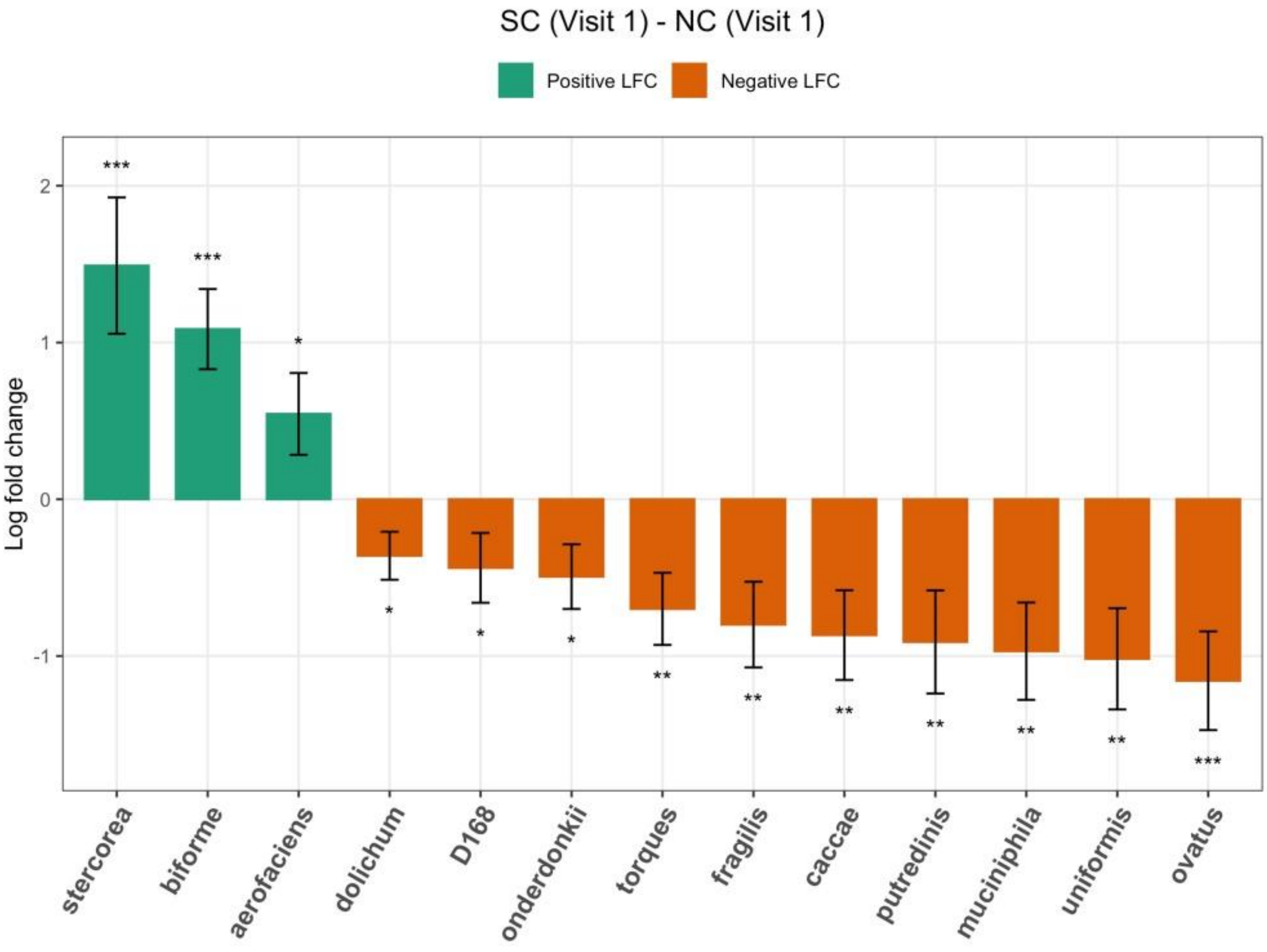
It is well-known that microbiota are involved in inflammation and immune response.

Men who seroconverted, were they pre-disposed to seroconversion because they have a different gut microbiome?

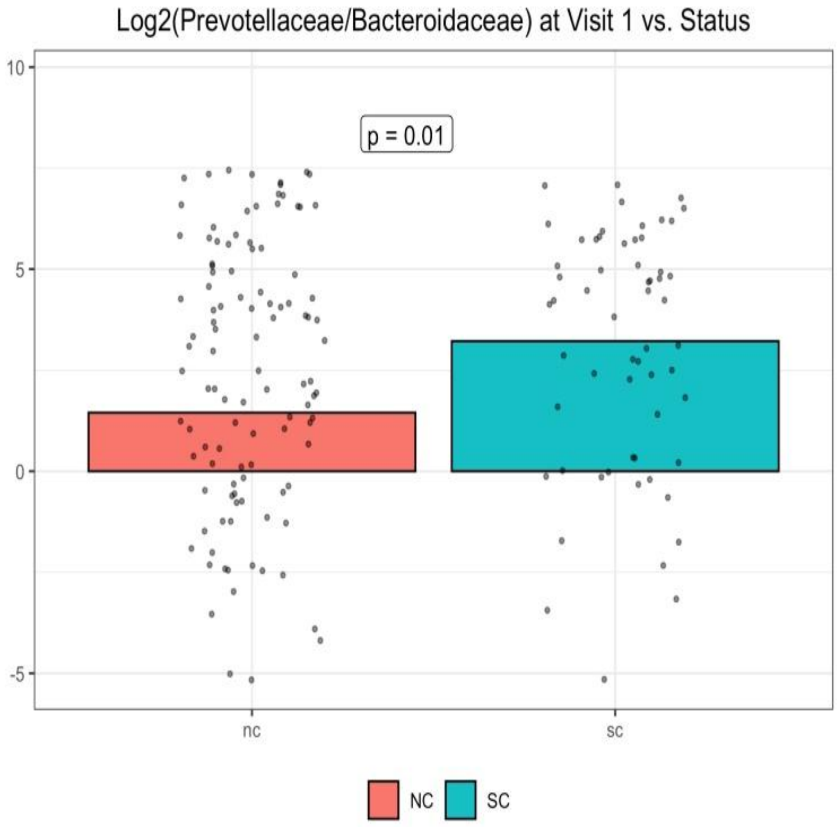
Differentially Abundant Species at Visit 1

	SC minus NC at visit 1						
Family	Genus	Species	LFC	SE	CI	P-value	Adjusted p-value (BH)
Prevotellaceae	Prevotella	stercorea	1.49	0.43	[0.64, 2.34]	0.00	0.01
Erysipelotrichaceae	[Eubacterium]	biforme	1.09	0.26	[0.59, 1.59]	0.00	0.00
Coriobacteriaceae	Collinsella	aerofaciens	0.55	0.26	[0.03, 1.06]	0.04	0.13
Erysipelotrichaceae	[Eubacterium]	dolichum	-0.36	0.15	[-0.66, -0.06]	0.02	0.07
Desulfovibrionaceae	Desulfovibrio	D168	-0.44	0.22	[-0.87, 0.00]	0.05	0.16
Rikenellaceae	Alistipes	onderdonkii	-0.49	0.21	[-0.90, -0.09]	0.02	0.07
Lachnospiraceae	[Ruminococcus]	torques	-0.70	0.23	[-1.15, -0.25]	0.00	0.02
Bacteroidaceae	Bacteroides	fragilis	-0.80	0.27	[-1.33, -0.26]	0.00	0.02
Bacteroidaceae	Bacteroides	caccae	-0.87	0.29	[-1.43, -0.31]	0.00	0.02
Bacteroidaceae	Bacteroides	uniformis	-1.02	0.32	[-1.65, -0.39]	0.00	0.02
Bacteroidaceae	Bacteroides	ovatus	-1.16	0.32	[-1.78, -0.54]	0.00	0.01
Rikenellaceae	Alistipes	putredinis	-0.91	0.33	[-1.55, -0.27]	0.01	0.03
Verrucomicrobiaceae	Akkermansia	muciniphila	-0.97	0.31	[-1.58, -0.36]	0.00	0.02

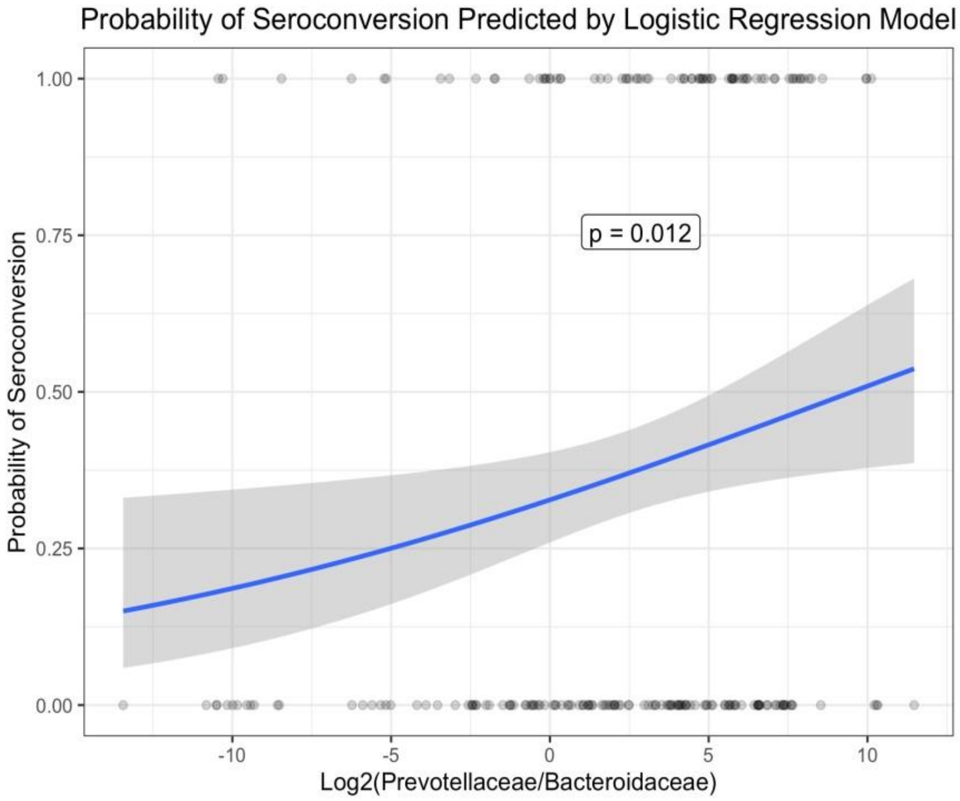
Differentially Abundant Species at Visit 1



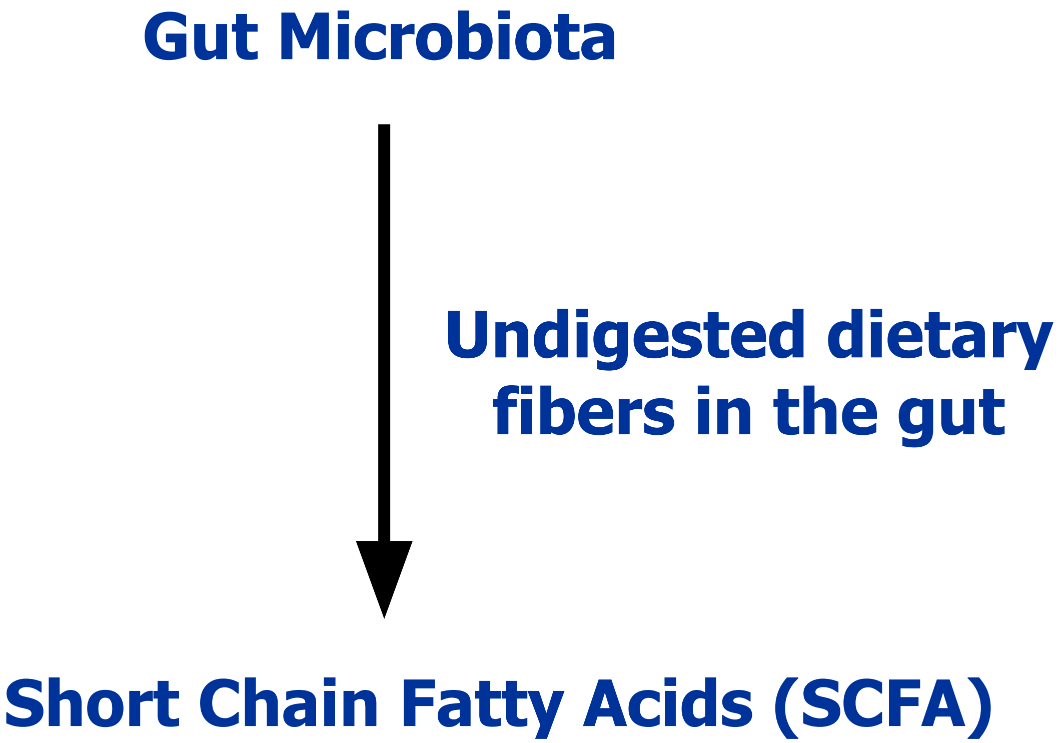
Prevotellaceae/Bacteroidaceae Predictor of Future Seroconversion



Log2(Prevotellaceae/Bacteroidaceae) at Visit 1 vs. Status



Probability of Seroconversion Predicted by Logistic Regression Model



Gut Microbiota

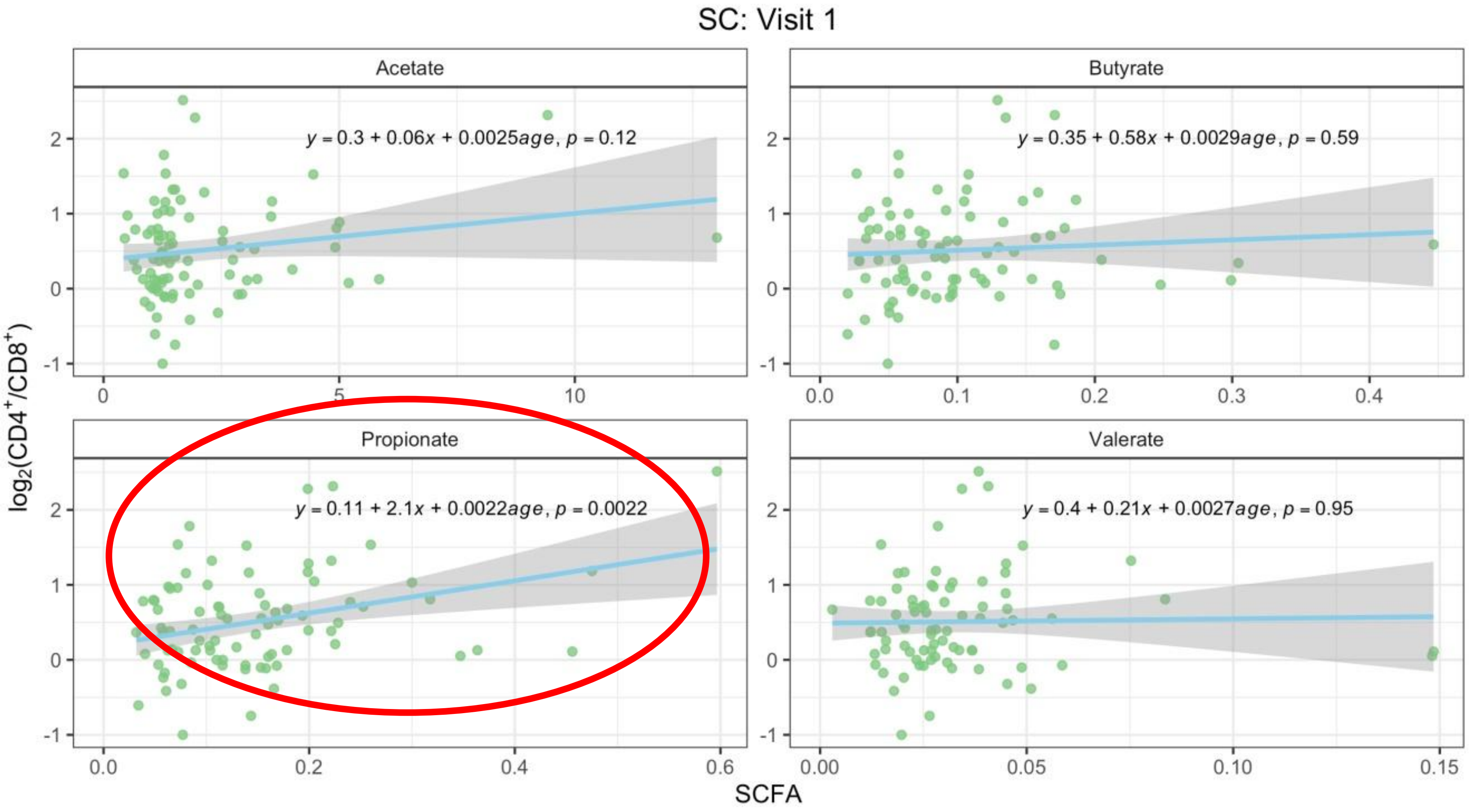
Undigested dietary fibers in the gut

Short Chain Fatty Acids (SCFA)

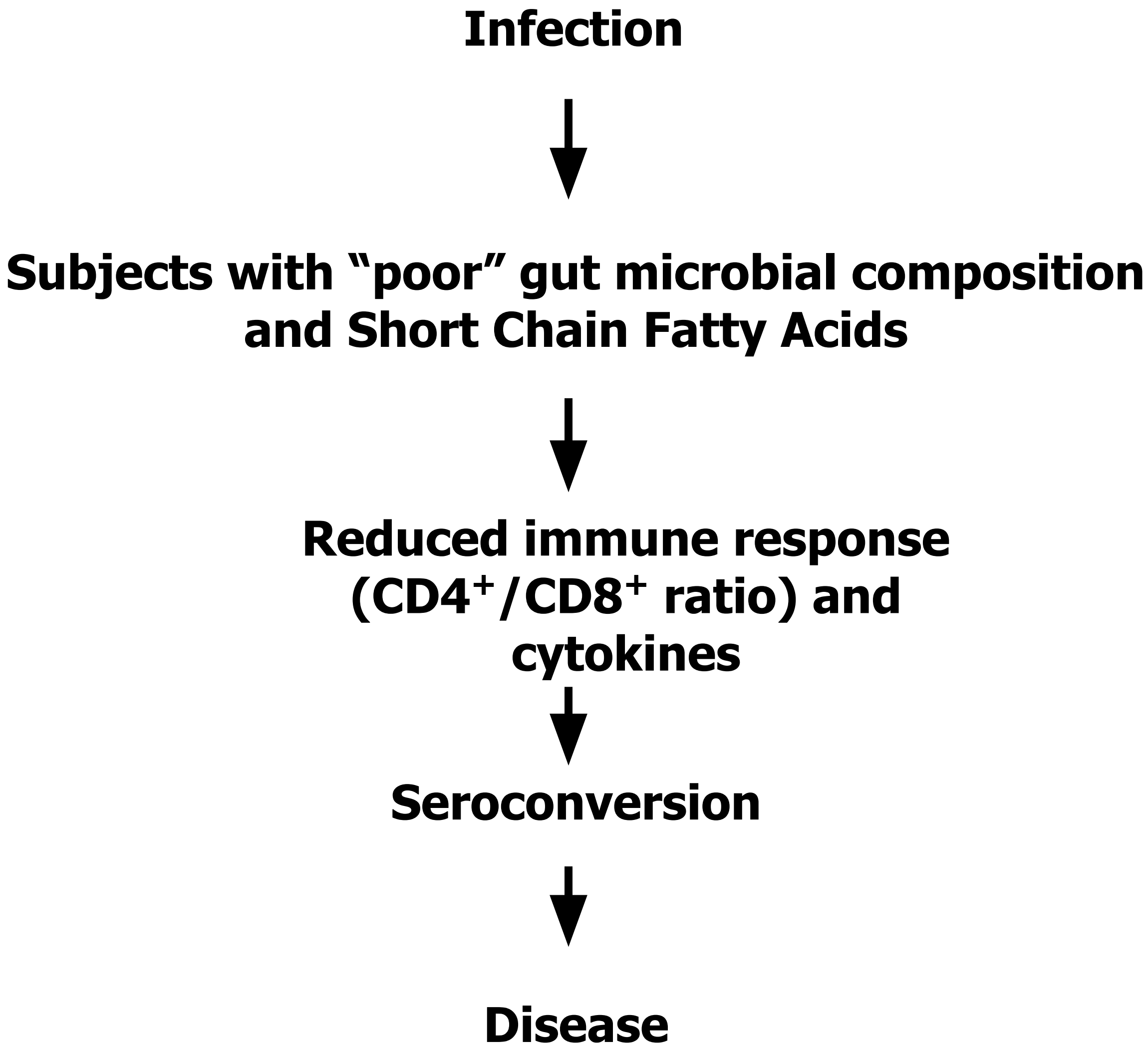
Short Chain Fatty Acids

SCFA	Bacteria involved in the production of SCFA	Function
Acetic Acid (Acetate)	Bifidobacteria, Lactobacillus, Akkermansia muciniphila, Prevotella spp., Ruminococcus spp.	Regulates gut pH controls appetite nourishes butyrate-producing bacteria; protects against pathogens
Propionic Acid (Propionate)	Bacteroidetes, Firmicutes, Lachnospiraceae	regulates appetite, anti-inflammatory;
Butyeric Acid (Butyrate)	Faecalibacterium prausnitzii, Eubacterium rectale and Roseburia spp.	energy source for colon, prevent leaky gut, anti-inflammatory, anti-oxidant properties

Propionic Acid Positively Correlates with $CD4^+/CD8^+$ Ratio at Visit 1



Potential hypothesis?



Infection

Subjects with “poor” gut microbial composition and Short Chain Fatty Acids

Reduced immune response (CD4$^+$/CD8$^+$ ratio) and cytokines

Seroconversion

Disease

Correlation Analysis for Microbiome …

Lin & Peddada (2021), under preparation

Motivation

- In many microbiome studies, the primary goal is to discover interactions among microbiota within or between ecosystems. Correlation analysis is an important starting point

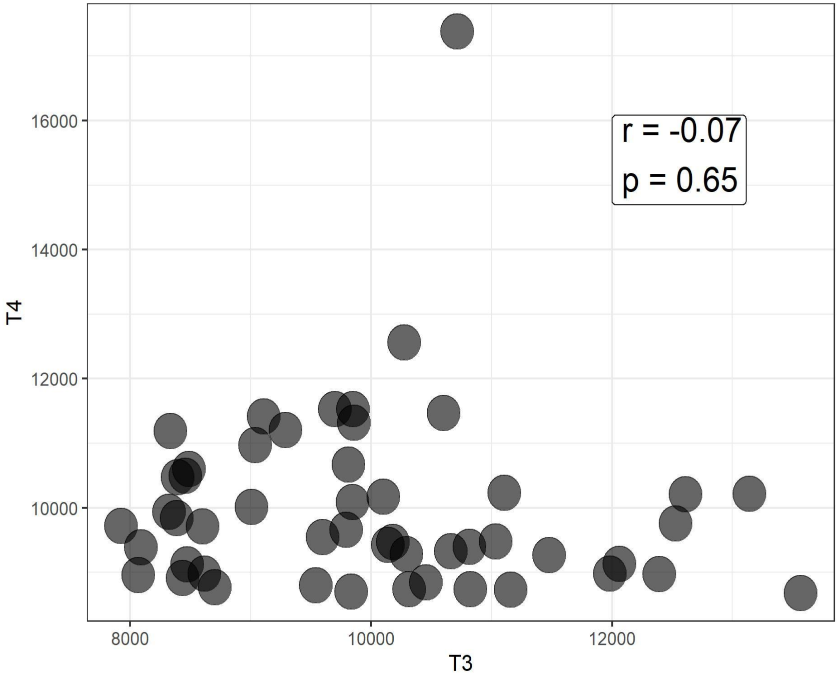
- However, estimation of correlations among microbiota is a challenging problem due to the unique features of microbiome data

- Most taxa are uncorrelated, which means the correlation matrix should be a sparse matrix

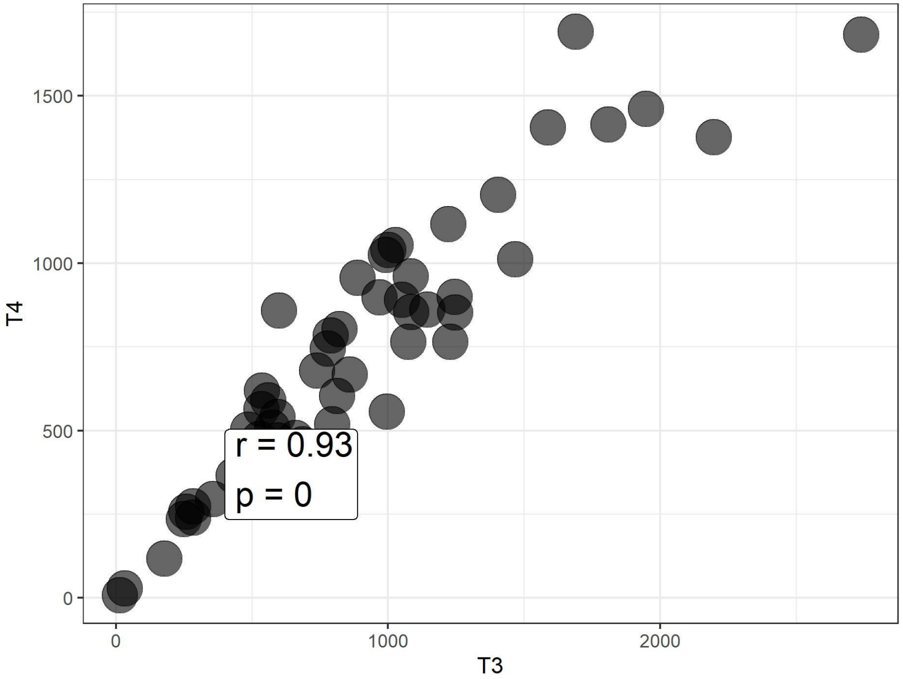
- Not all interactions among taxa are linear

An Illustrative Example

Underlying True Relationships

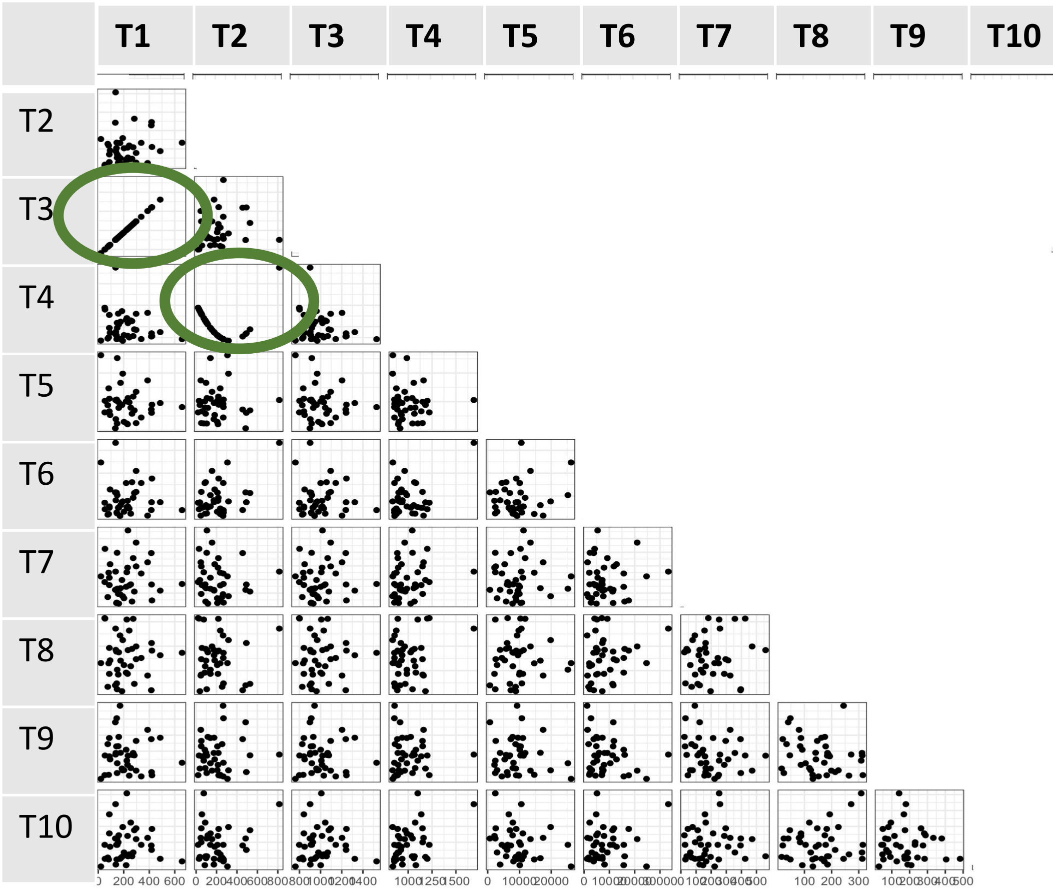


Observed Relationships

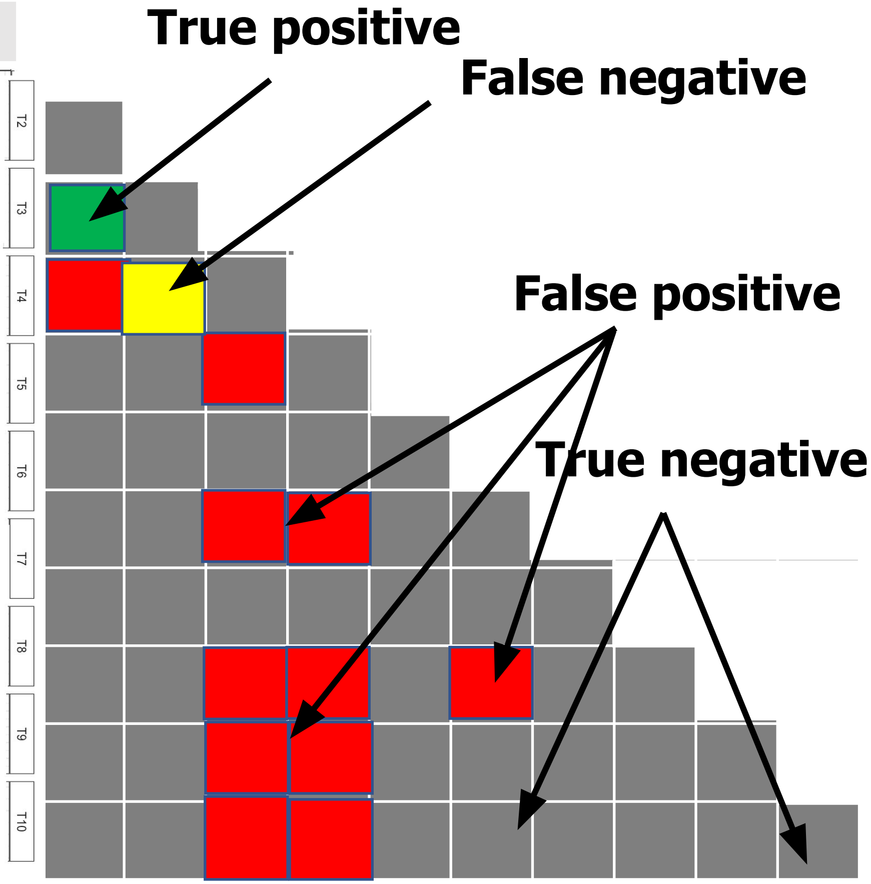


Standard Spearman Correlation Coefficient

Truth in the ecosystem



Spearman correlations



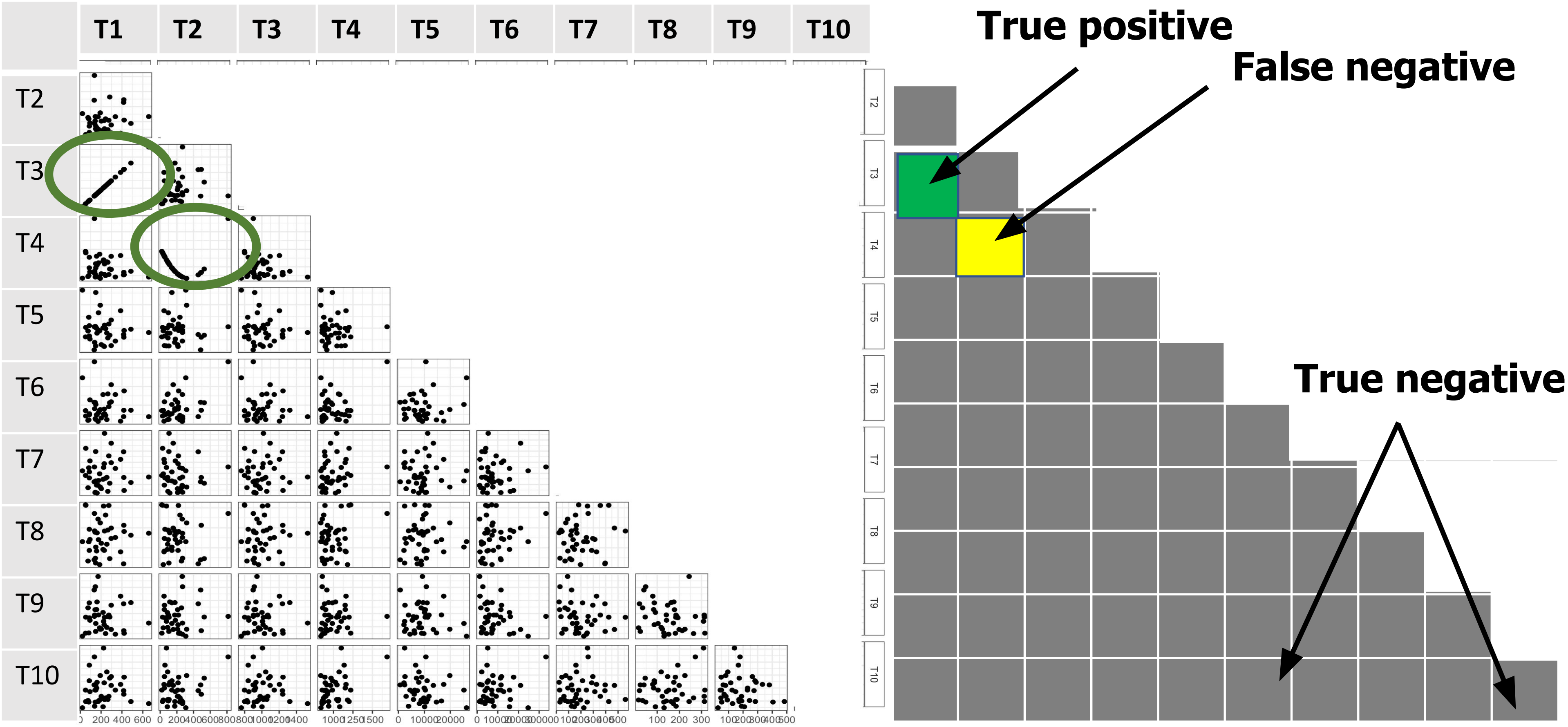
True positive

False negative

False positive

True negative

SparCC



Truth in the ecosystem

SparCC

True positive

False negative

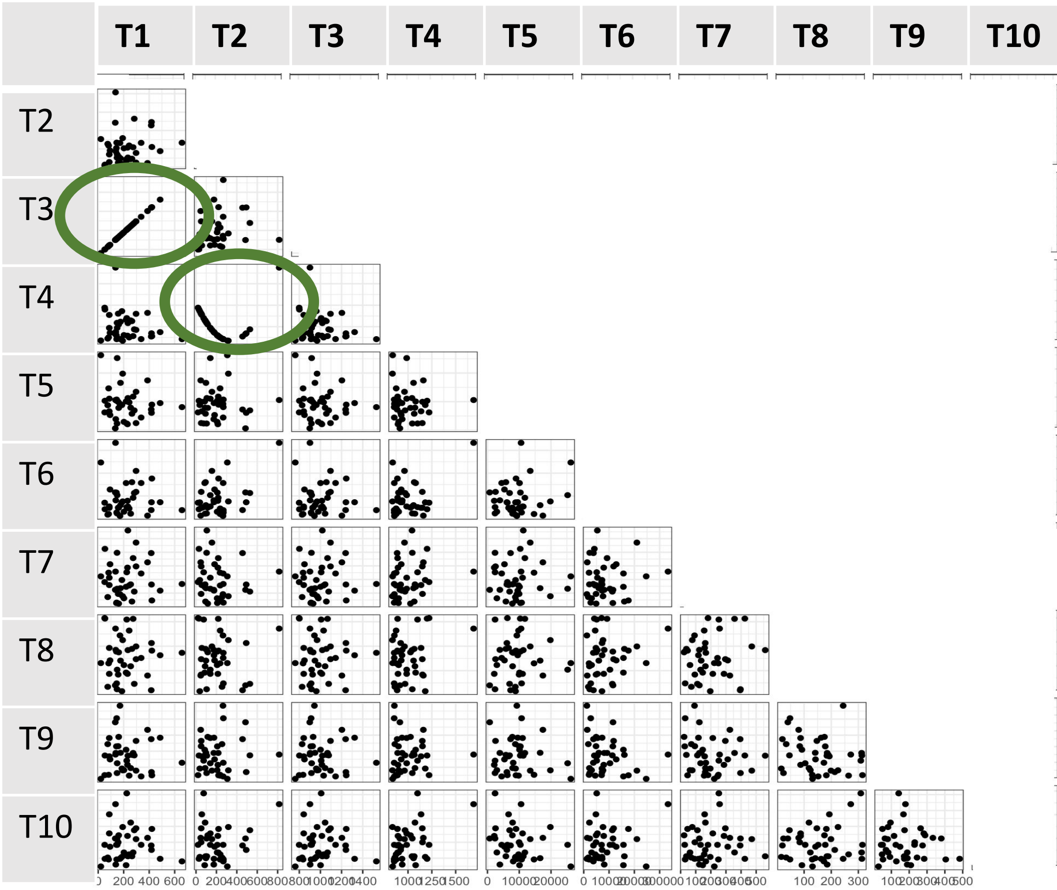
True negative

Distance Correlation

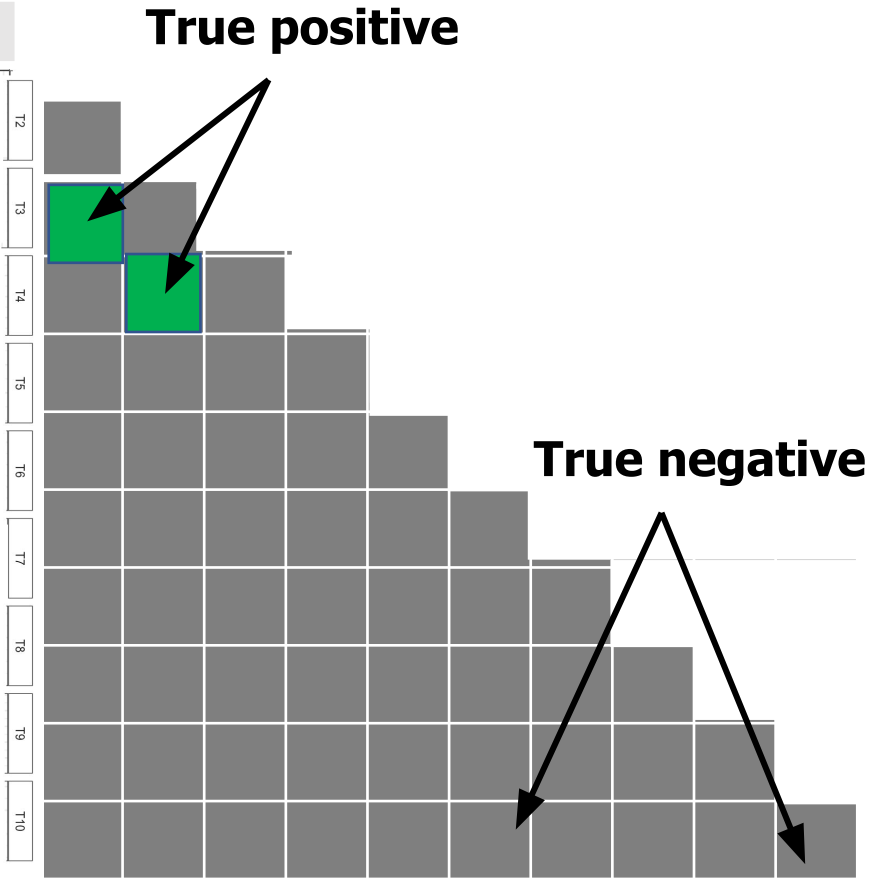
- The Pearson/Spearman can only test for linearity in relationship
- Distance Correlation: Szekely, 2007 Annals of Statistics allows us to test for non-linear relationships
- Generalizing the concept of distance correlation for microbiome data by accounting for:
  - Bias due to compositionality
  - Differential sampling fractions
  - Excess zeros

SECOM

Truth in the ecosystem



SECOM


True positive

True negative

Illustration of SECOM using global gut microbiota data …

Cross-Cultural Infant Gut Microbiome Data

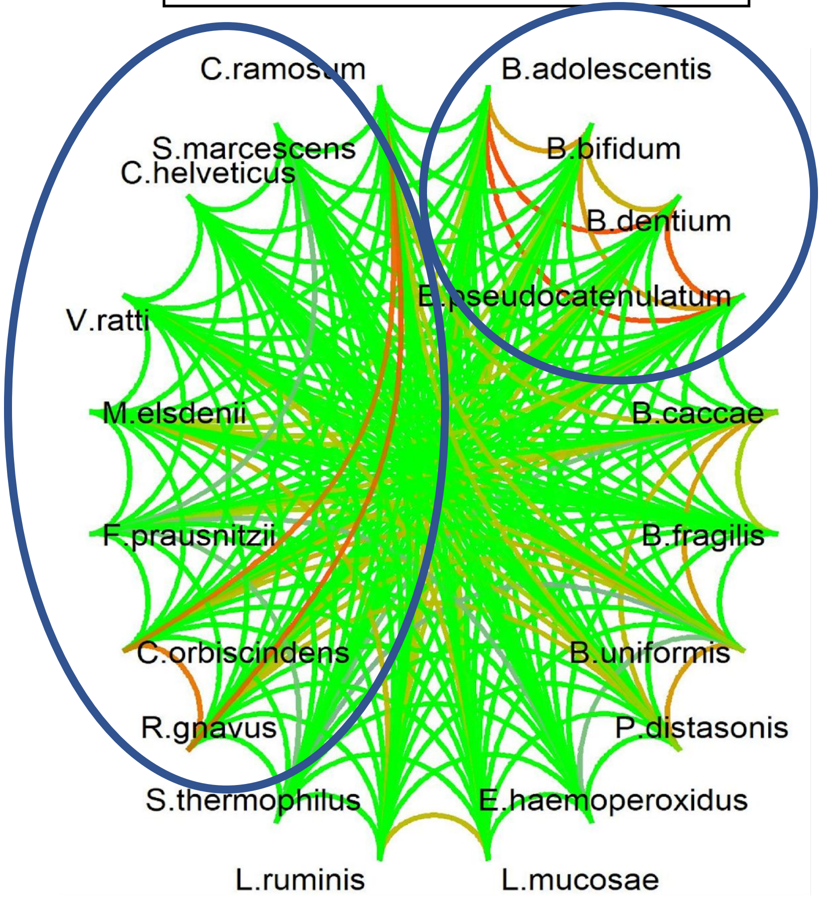
- 11,905 OTUs obtained from subjects in three locations:
  - **The US** (n = 317), **Malawi** (n = 114), and **Venezuela**(n = 99)

- Infants (age <= 2 years old) were selected for analysis
  - **The US** (n = 50), **Malawi** (n = 47), and **Venezuela**(n = 27)

- The OTU data were aggregated into **species level** for analysis, and the **top 20** most abundant species were selected for visualization

Summary

Age <= 2



- For infants, more pairs of species appear to be positively correlated

- *B.adolescentis, B.bifidum, B.dentium*, and *B.pseudocatenulatum,* which belong to genera *Bifidobacterium* that is commonly available in breast milk, show a strong positive correlation among each other

- *R.gnavus*, *C.orbiscindens*, and *C.ramosum*, which are all anaerobic bacteria, are also grouped together

Conclusion

- Analyzing microbiome data is a challenging problem due to their unique features
  - Compositionality, differential sampling fractions, excess zeros

- ANCOM-BC and SECOM correct bias due to these features and lead to an unbiased differential abundance (DA) analysis and correlation analysis, respectively

- ANCOM-BC and SECOM are designed for drawing inferences on absolute abundance and not relative abundance

- SECOM can not only identify linear correlations but also detect non-linear correlations among microbiota, thus, fill an important gap in the literature

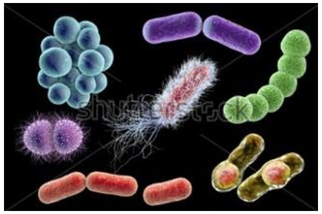
Acknowledgments

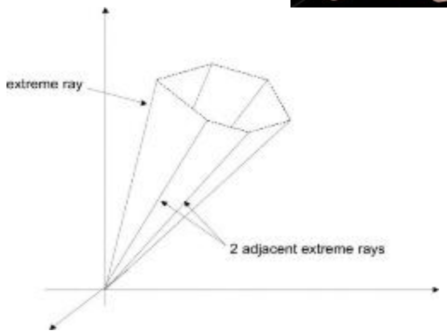
Shyamal Das Peddada, Senior Investigator and Branch Chief
Biostatistics & Bioinformatics Branch
Division of Intramural Population Health Research
NICHD, NIH

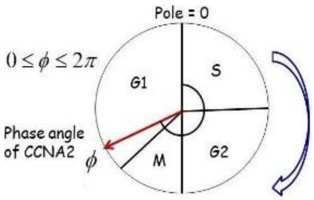
Charles Rinaldo, Professor and former Chair
Department of Infectious Disease and Microbiology
University of Pittsburgh

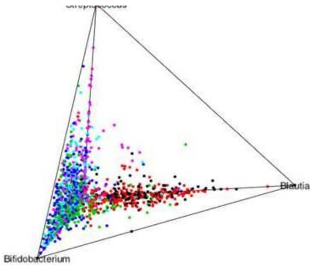
Yue Chen, Assistant Professor
Department of Infectious Disease and Microbiology
University of Pittsburgh

Thank you !




extreme ray
2 adjacent extreme rays


Pole = 0
$0 \le \phi \le 2\pi$
G1
S
Phase angle of CCNA2
ϕ
M
G2


Streptococcus
Blautia
Bifidobacterium

ANCOM-BC Model

Offset-based log linear model:

$$o_{ij} = s_i + \beta_j^T x_i + \varepsilon_{ij}$$

i: sample
j: taxon
x_i: covariates
$o_{ij} = \log(O_{ij})$

Principle of estimation: β and s are not estimable individually. We attempt to pool information across taxa to estimate them iteratively (**Algorithm 1: Iterative least square regression**).

Upon convergence:

$$\begin{cases} E(s^*) = s - X\delta \\ E\left(\beta_j^*\right) = \boxed{\beta_j + \delta} \end{cases}$$

Biased estimators

ANCOM-BC Model

$E(\beta_j^*) = \beta_j + \delta$

j: taxon

k: covariate

Algorithm 2: E-M algorithm for bias correction:

- Gaussian mixture model

$$f(\beta_{jk}^*) = \pi_0 \phi\left(\frac{\beta_{jk}^* - \delta_k}{\nu_{i0}}\right) + \pi_1 \phi\left(\frac{\beta_{jk}^* - (\delta_k + l_1)}{\nu_{i1}}\right) + \pi_2 \phi\left(\frac{\beta_{jk}^* - (\delta_k + l_2)}{\nu_{i2}}\right)$$

- $\hat{\beta}_{jk} \leftarrow \beta_{jk}^* - \hat{\delta}_k^{EM}$

Unbiased estimators

Hypothesis Testing

Hypothesis:

$$H_0\colon A\beta_j = A\beta_j^{(0)},$$
$$H_1\colon A\beta_j \neq A\beta_j^{(0)}.$$

The test statistic:

$$W_j = \left(A\hat{\beta}_j - A\beta_j\right)^T \left(A\hat{\Sigma}_j A^T\right)^{-1}\left(A\hat{\beta}_j - A\beta_j\right) \rightarrow_d \chi_q^2, \text{ as } n \rightarrow \infty.$$

where $q = rank(A)$.

n: sample size
d: # taxa

Estimator for sampling fractions:

$$\hat{s} = \frac{1}{d}\sum_{j=1}^{d}\left(o_j - X\hat{\beta}_j\right) \rightarrow_p s, \text{ as } n, d \rightarrow \infty.$$