Colorado School of Public Health

Analysis of Longitudinal Microbiota Data

Brandie Wagner
Assistant Professor
Department of Biostatistics and Informatics &
Department of Pediatrics

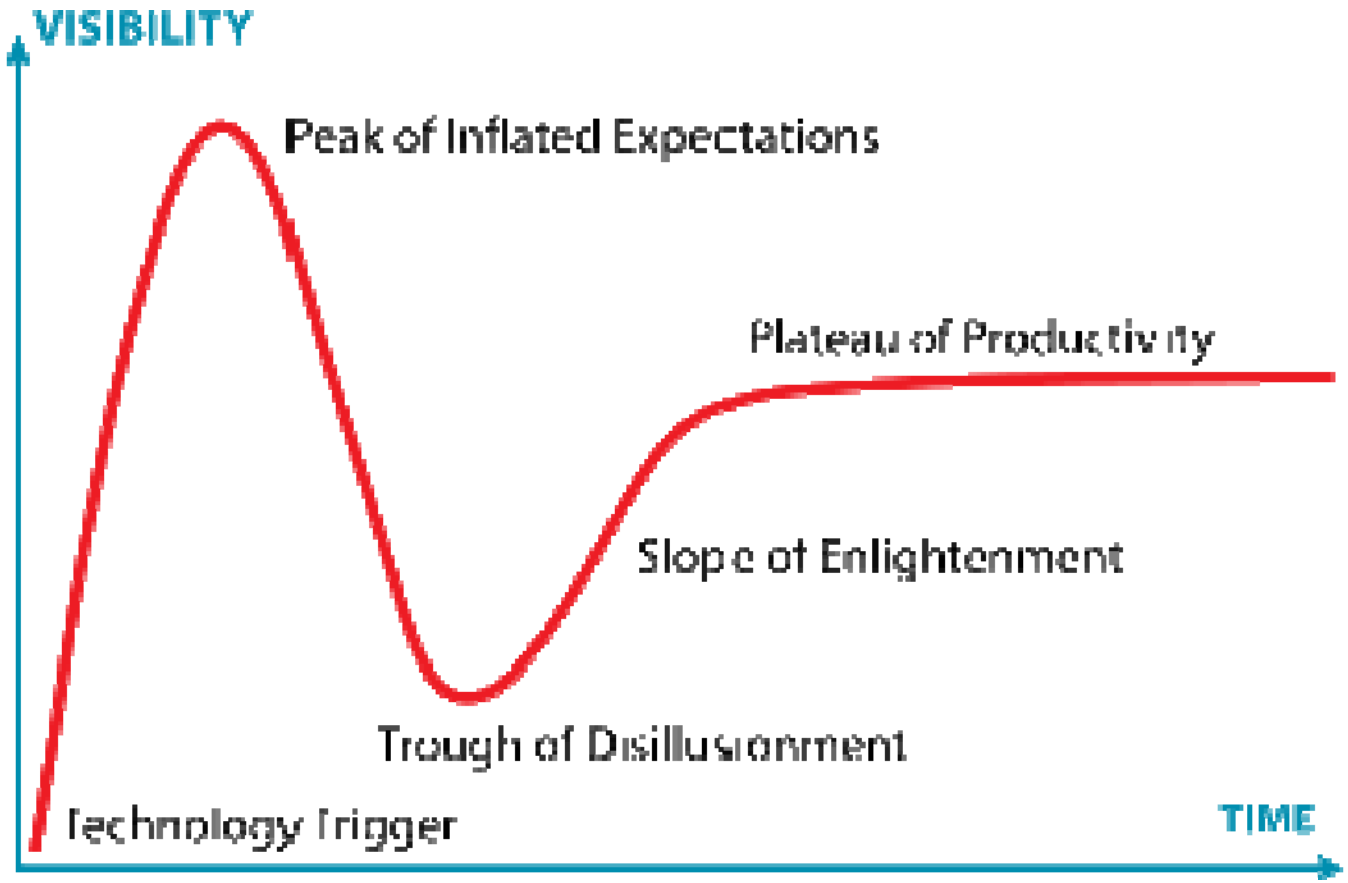
MiRC
Microbiome Research Consortium

Children's Hospital Colorado

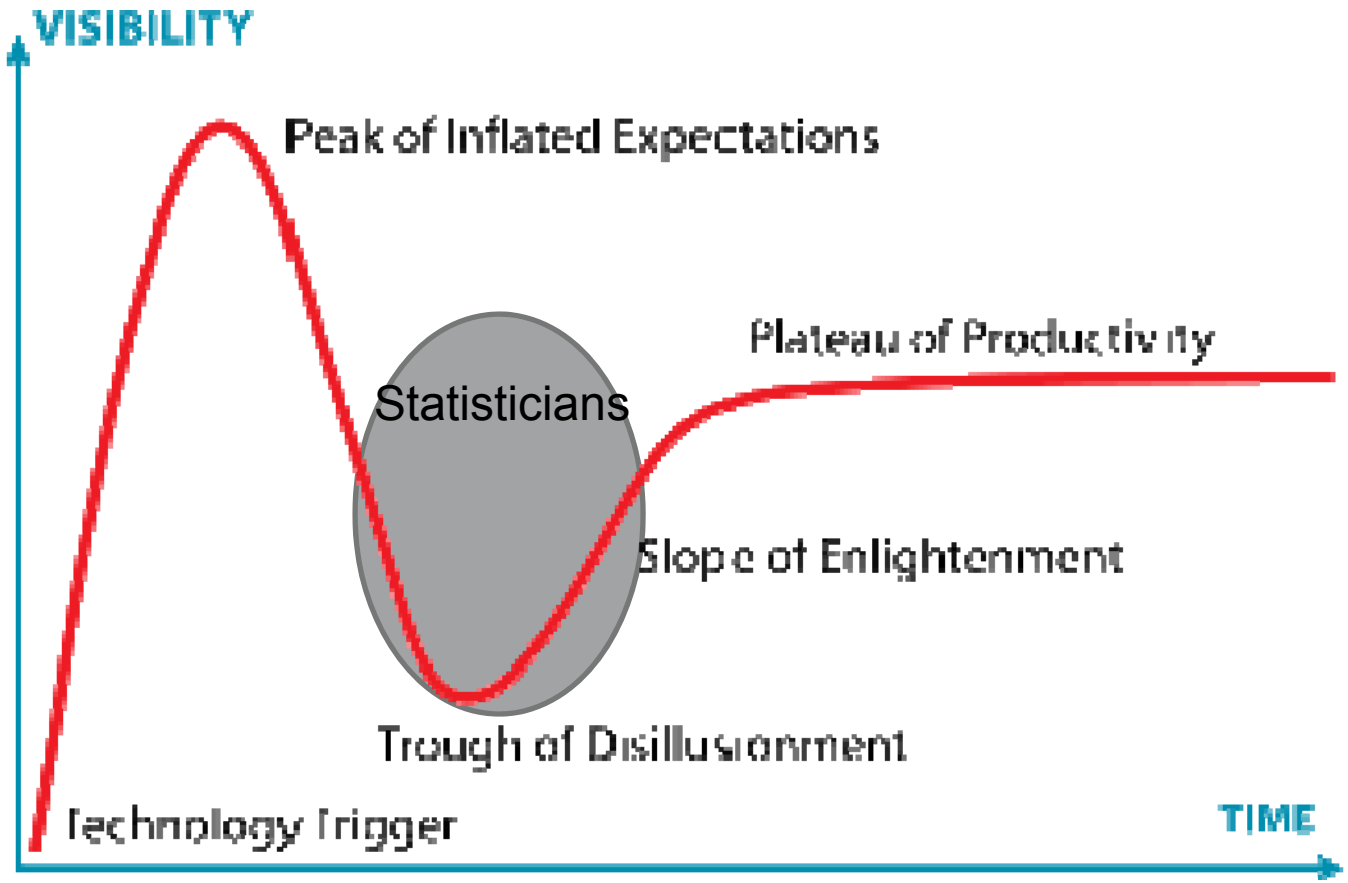
Nuances of microbiome data

- High dimensional & Small sample sizes (p > n)
- Correlated (heirarchical phylogenetic tree)
- Counts
  - *Non-normal (non- continuous; non-negative)*
  - *Bounded by total number of sequences*
- Zero-inflated
  - *Artifact of compiling a dataset from different samples*
  - *Proportion of zeros increases with variability in conditions (niche theory)*
- Variable sequencing effort
  - *Variations in total # sequences obtained for each sample*
- Compositional data (relative abundance)
- Complex designs (time course & repeated measures)

Gartner Hype Cycle



Gartner Hype Cycle



Importance of Longitudinal Studies

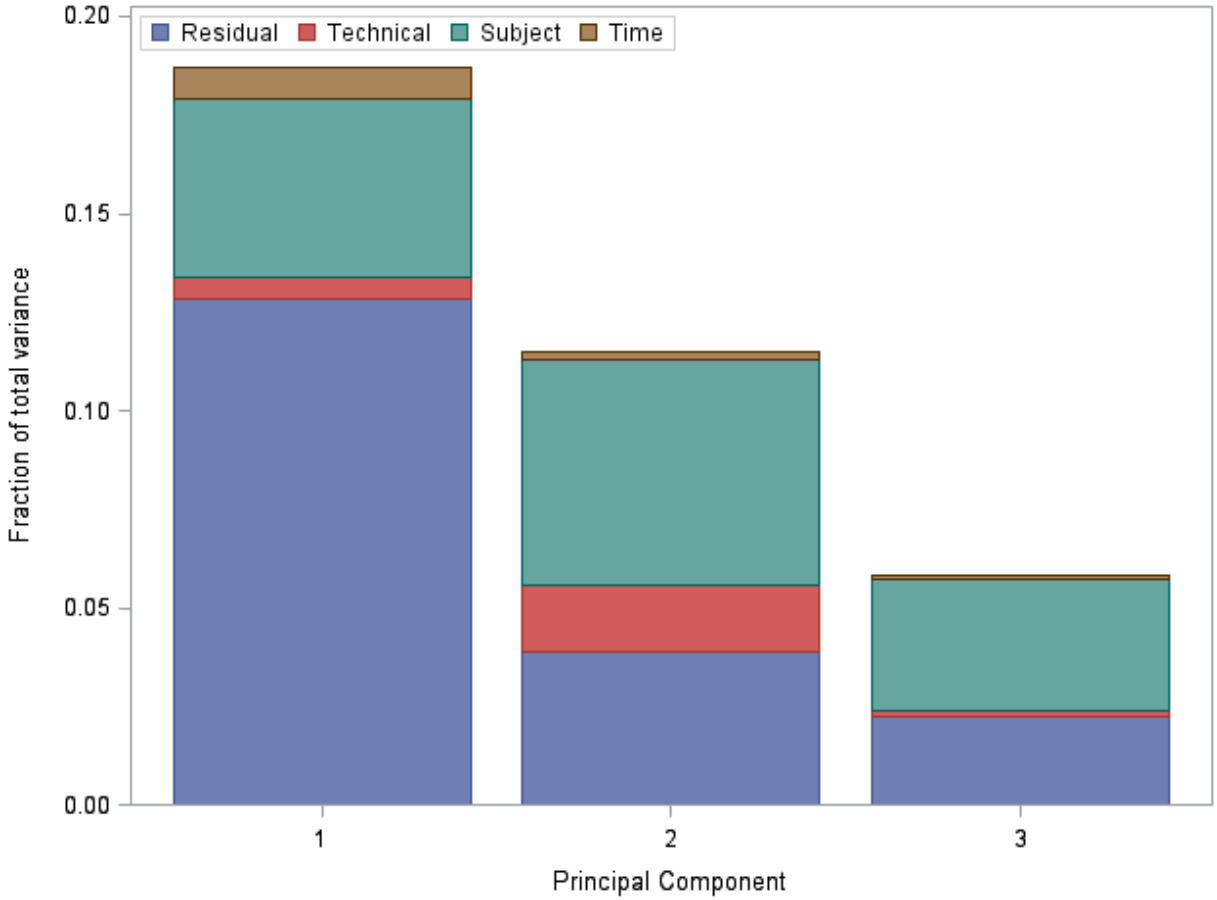
- Large subject-to-subject variability, many factors contribute to variability (diet, region, exposure, genetics)
- Valuable information is likely to come from within subject variations
- Compare to increased power of cross-over designs compared to parallel designs
- Flores *et al* 2014 supports importance of temporal variability in human microbiome studies

Flores *et al.* Temporal variability is a Personalized Feature of the Human Microbiome. Genome Biol. 2014:15(12);531

Importance of Longitudinal Studies

- samples were collected from 16 subjects over 3 time points.
- At each time point, the sample was separated into two aliquots.
- These duplicate aliquots were processed simultaneously to assess the amount of technical variability for the entire sample processing procedure.
- Principal variance components analysis was applied

Harrison et al. Principal variance components analysis of crop composition data. J. Agric Food Chem. 2013:16;6412-6422



- The first 3 principal components (PC) explained 36% of the variability
- The between subject variability is the largest source of explained variability, bigger than technical variability.
- The relative contribution of the within subject variability over time is smaller than the between subject variability.

Simple design analysis strategy

- Summarize communities using ecological parameters Alpha Diversity
  - *Richness*
  - *Evenness*
- Ordination
  - *PCA*
  - *PCoA*
  - *MDS*
- Univariate comparisons
  - *Non-parametric*
  - *Negative binomial (DEseq - genome biology 2010)*
  - *Beta-Binomial (BBseq - bioinformatics 2011)*

Simple design analysis strategy

- apply statistical tests individually to each taxon
  - useful for identifying important organisms and performing dimension reduction
  - it ignores the interactions and multivariate structure of the bacterial communities.
  - These microbes are not present in isolation and understanding the entire community is important.
- Multivariate methods
- Longitudinal data – adds a level of complexity
  - Correlation within subjects

Microbial Diversity

Quick overview - Diversity

- Alpha – lots of different measures, most of which are related using Renyi equation, Hill's numbers make them interpretable
- Whittaker (1960) decomposes diversity into gamma, alpha and beta components
- Beta – compares compositions of multiple libraries
- Tuomisto and Jost similarly relate beta diversity measures ("True" beta diversity)

Generalization to longitudinal data
Beta diversity

- Partition diversity into independent components
  - evenness and richness
  - collections of libraries (beta diversity).
- Similarly, differing results are obtained across beta diversity indices
  - due to differences in weighting of the components.
  - a unifying theory called new beta.
  - The most popular similarity indices (Jaccard, Sorenson, Horn and Morisita-Horn) have been shown to be monotonic transformations of new beta.
- Although this area is still under development there is general consensus on the ideal properties of a beta diversity measure

  1) alpha and beta components are mathematically independent, that is, the value of alpha should not affect the value of beta and vice versa;

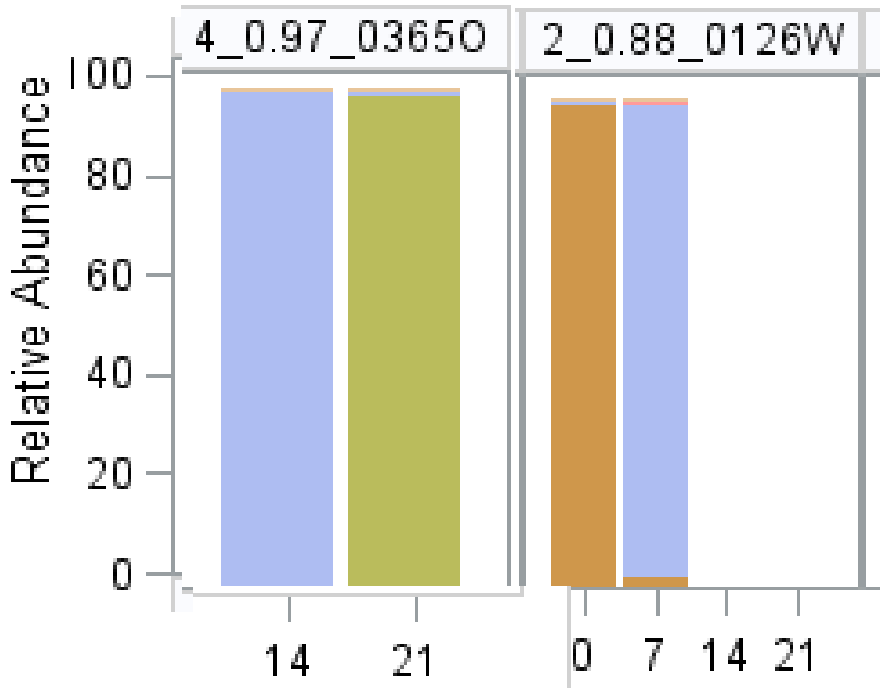
  2) gamma is completely determined by alpha and beta; and

  3) alpha is never greater than gamma.

Current analytic approach

- Diversity over time
  - Some have simply modeled alpha diversity over time
    - *This approach addresses whether diversity is changing over time*
    - *This is not the same as whether the community composition is changing temporally*
    - *Alpha diversity can remain unchanged despite large changes in microbial composition*

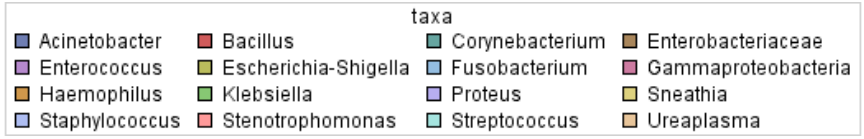
Alpha remains unchanged



Alpha = 0.04 0.06 0.29 0.42
Hill Beta = 1.97 1.88

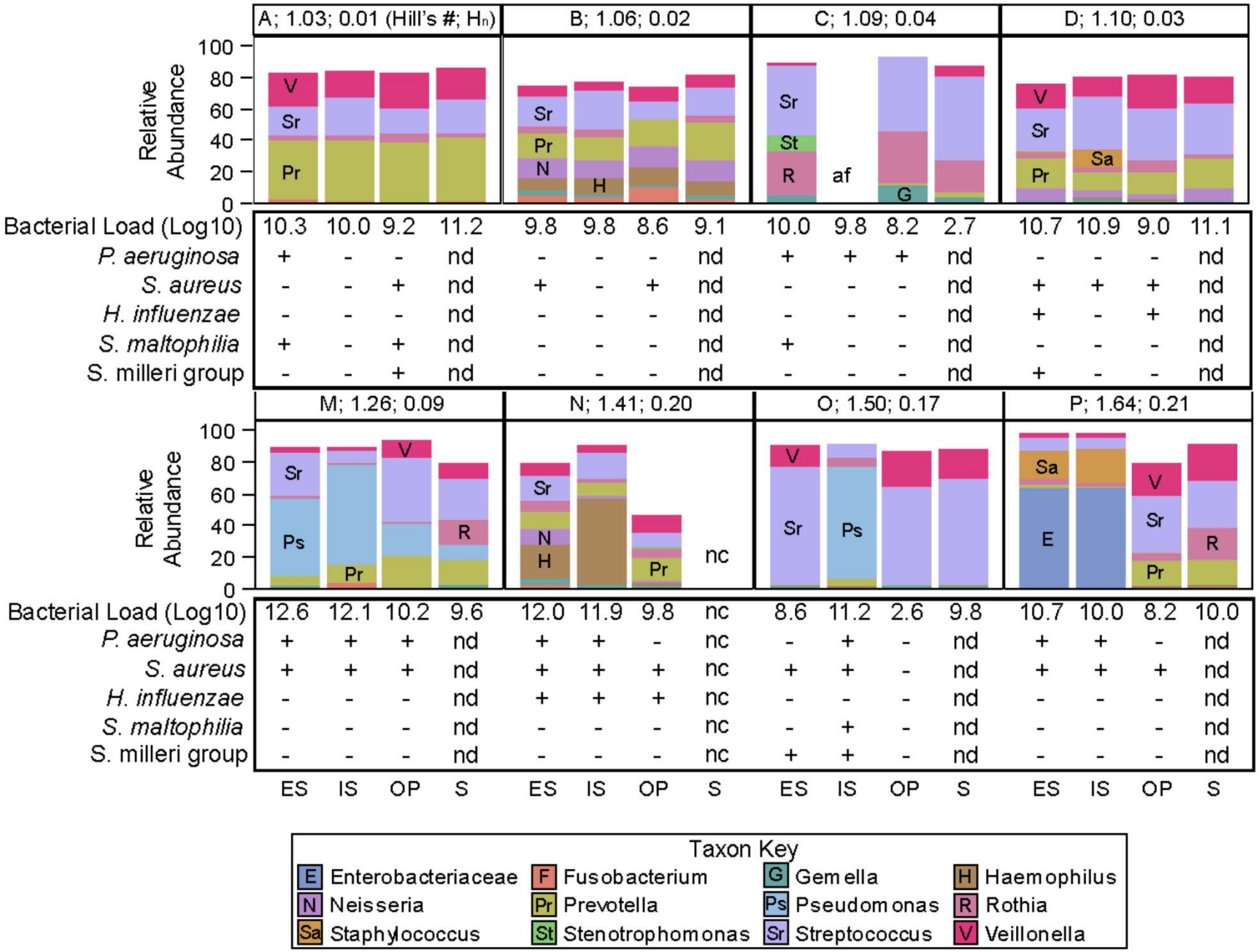
Multiple trachial aspirate samples collected from ventilated premature infants

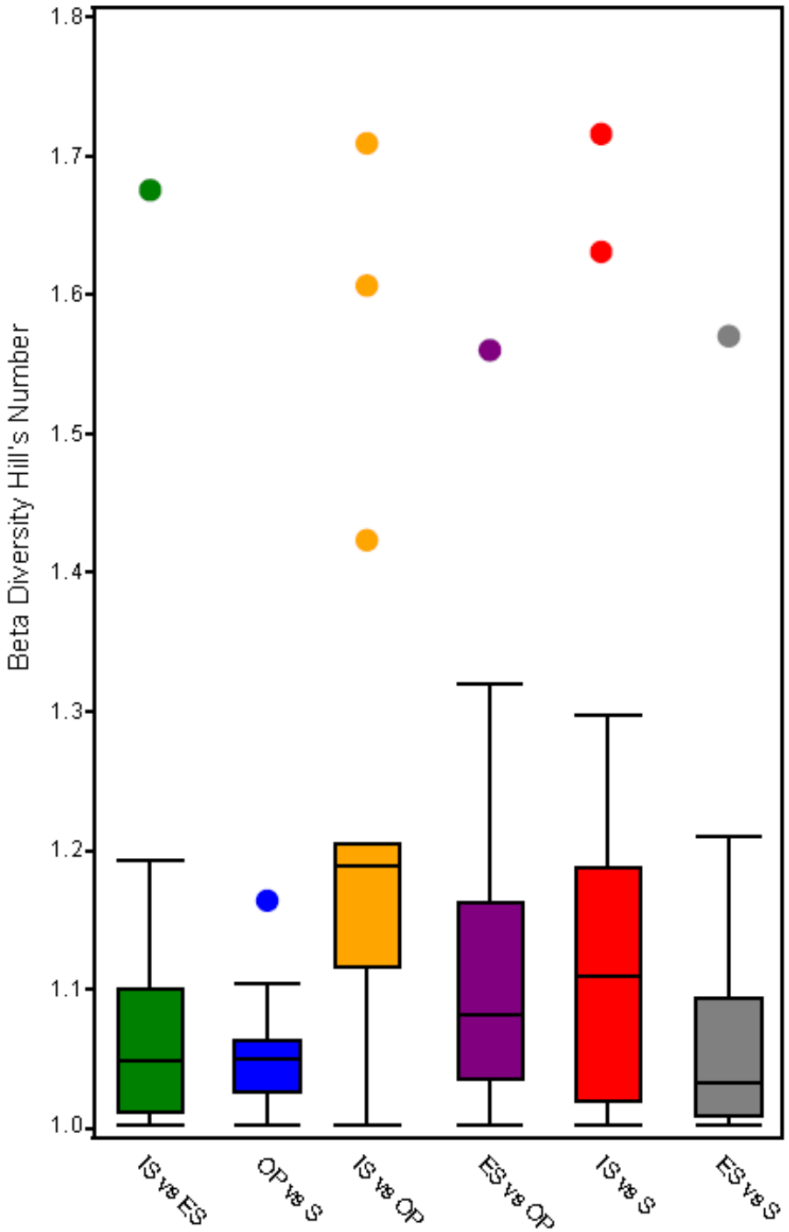
For ease of clinical interpretation, Shannon Beta is expressed as a Hill's number which indicates the effective number of communities represented by the group of samples.



Beta diversity, simple example

- 16 subjects, single time point, samples collected from 4 sites
- Beta diversity measure used to assess similarity across sites for each subject
- Beta diversity relative to the number of samples, so a normalization was used to compare measures across individuals with different number of samples

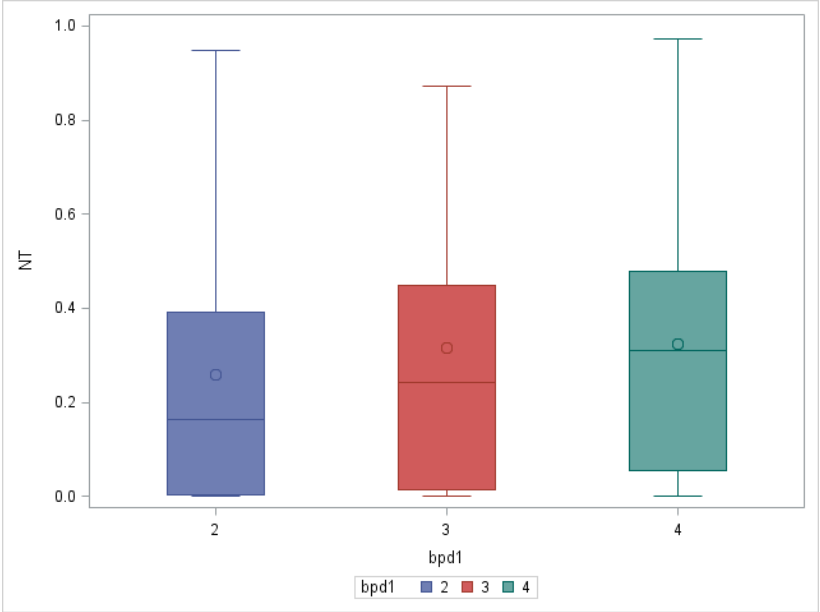

A; 1.03; 0.01 (Hill's #; Hn)
B; 1.06; 0.02
C; 1.09; 0.04
D; 1.10; 0.03
Relative Abundance
100
80
60
40
20
0
V
Sr
Pr
N
H
St
R
af
G
Sa
Bacterial Load (Log10) 10.3 10.0 9.2 11.2 9.8 9.8 8.6 9.1 10.0 9.8 8.2 2.7 10.7 10.9 9.0 11.1
P. aeruginosa + - - nd - - - nd + + + nd - - - nd
S. aureus - - + nd + - + nd - - - nd + + + nd
H. influenzae - - - nd - - - nd - - - nd + - + nd
S. maltophilia + - + nd - - - nd + - - nd - - - nd
S. milleri group - - + nd - - - nd - - - nd + - - nd
M; 1.26; 0.09
N; 1.41; 0.20
O; 1.50; 0.17
P; 1.64; 0.21
Ps
Pr
R
nc
E
Bacterial Load (Log10) 12.6 12.1 10.2 9.6 12.0 11.9 9.8 nc 8.6 11.2 2.6 9.8 10.7 10.0 8.2 10.0
P. aeruginosa + + + nd + + - nc - + - nd + + - nd
S. aureus + + + nd + + + nc + + - nd + + + nd
H. influenzae - - - nd + + + nc - - - nd - - - nd
S. maltophilia - - - nd - - - nc - + - nd - - - nd
S. milleri group - - - nd - - - nc + + - nd - - - nd
ES IS OP S ES IS OP S ES IS OP S ES IS OP S
Taxon Key
E Enterobacteriaceae
F Fusobacterium
G Gemella
H Haemophilus
N Neisseria
Pr Prevotella
Ps Pseudomonas
R Rothia
Sa Staphylococcus
St Stenotrophomonas
Sr Streptococcus
V Veillonella



Can also look at distributions of pairwise beta diversities across sites

BPD dataset – association between beta diversity and disease severity

- 302 trach aspirate samples were sequenced from 162 subjects up to 4 time points per subject ranging from 0 to 28 days of life.



This plot displays the normalized values from subjects with at least 2 or more samples. There is a trend towards greater community turnover across the groups (Blue n = 25, Red N = 39 and GreenN = 30).

Nelson et al

- Dice similarity was calculated to evaluate the percentage similarity of composition over time between consecutive sampling points for individual patients and associated with IV antibiotics

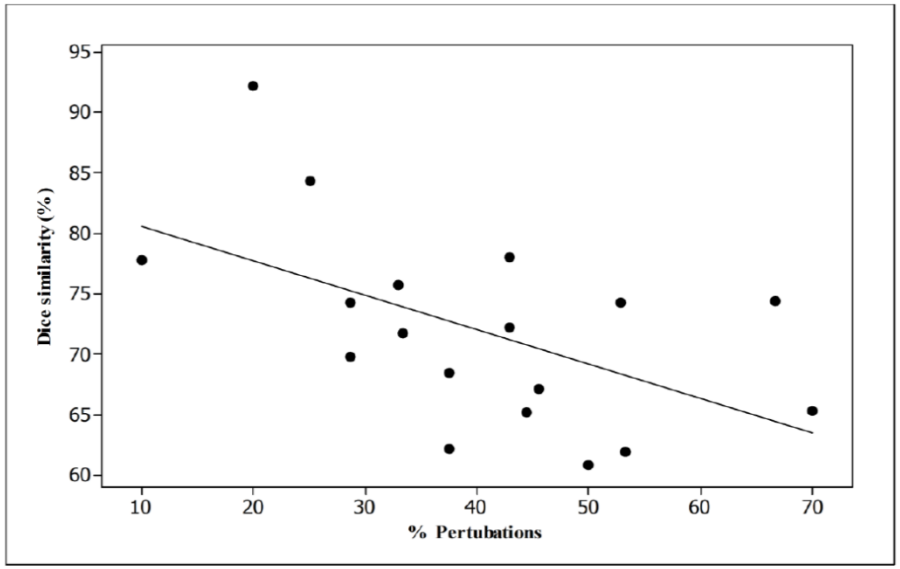
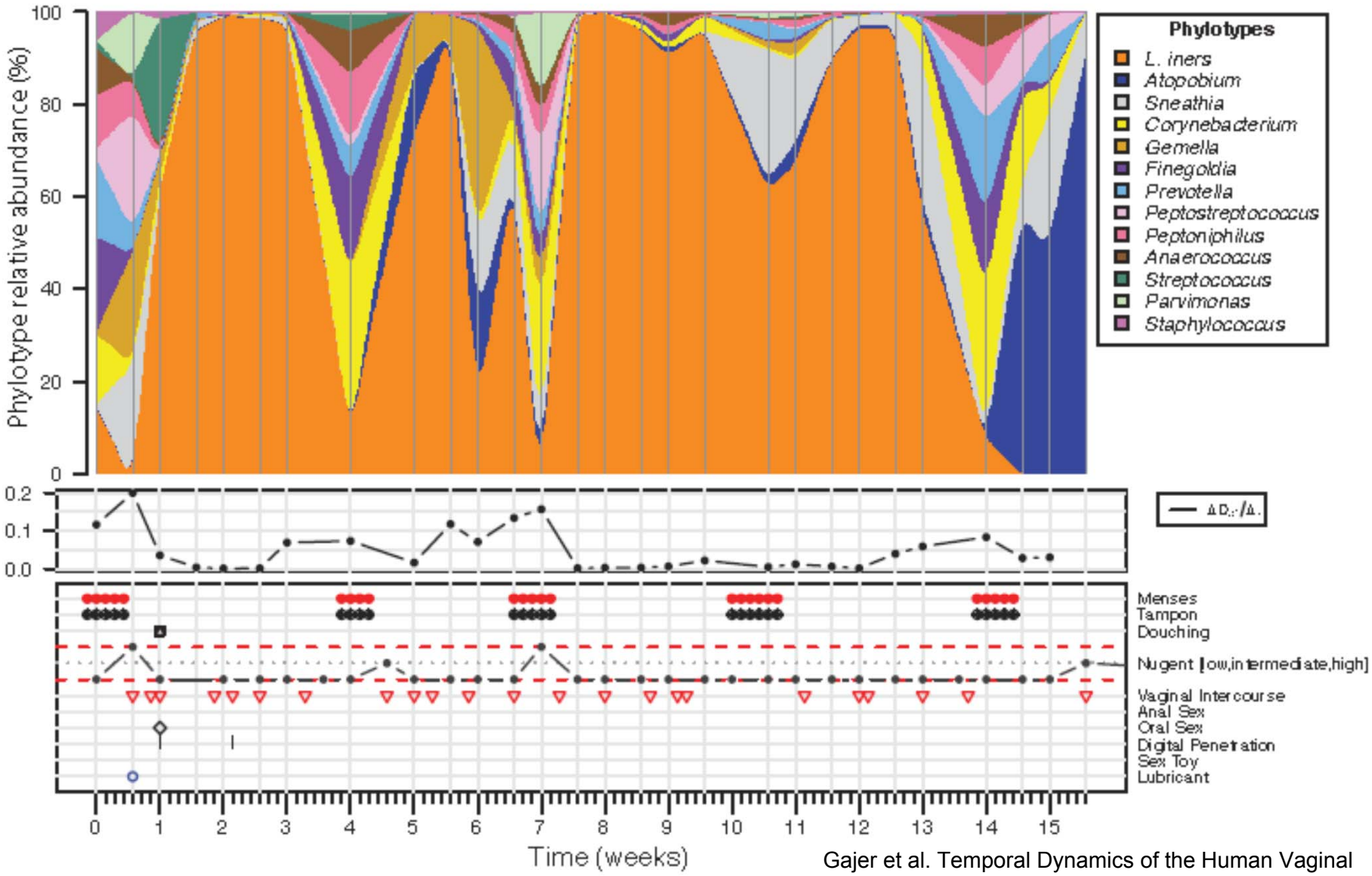


Fig 3 Relationship between IV antibiotics and bacterial community similarity



Gajer et al. Temporal Dynamics of the Human Vaginal Microbiome. Sci Transl Med. 2012:4(132);132ra52

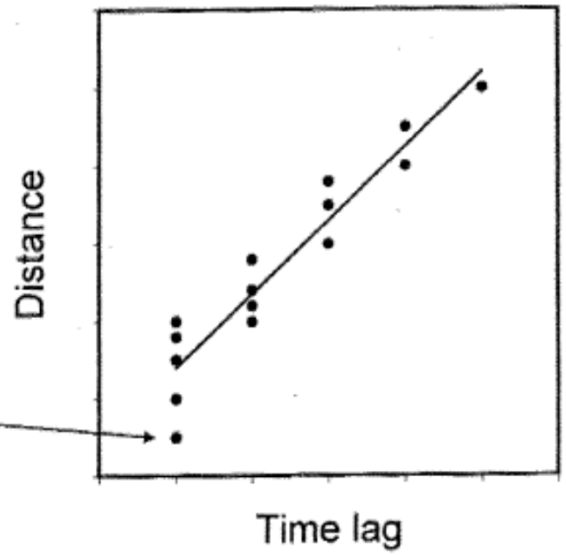
Extensions to longitudinal

- Previous analyses calculate a single beta diversity measure for each subject describing the overall similarity across communities
- Also of interest to use beta diversity to evaluate turnover in community composition over time.
- Collins (2000)

Beta diversity (pairwise) is regressed on a time lag variable using a time series model.

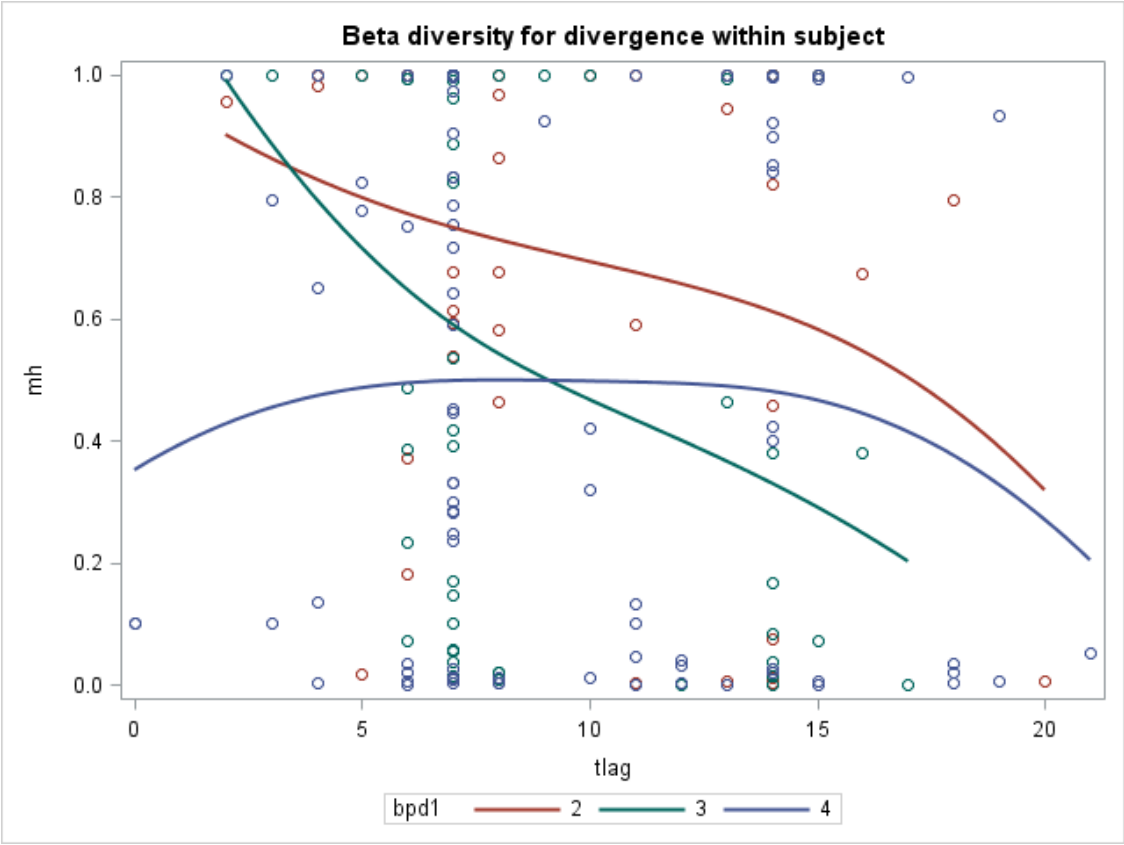
Distance matrix

0.0					
1.0	0.0				
2.0	1.5	0.0			
3.0	2.2	0.5	0.0		
4.0	3.5	2.4	2.0	0.0	
5.0	4.5	3.8	2.8	1.8	0.0



Collins et al. A Method to determine rates and patterns of variability in ecological communities. Oikos. 2000:91(2);285-293.

BPD dataset



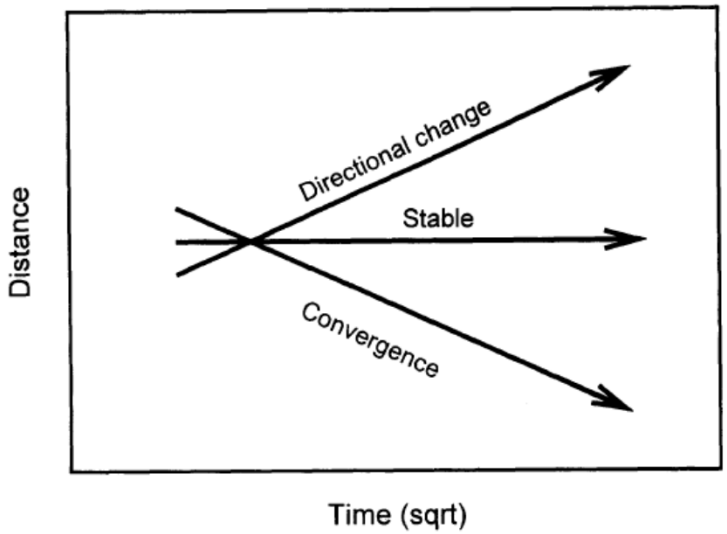


Fig. 2. Some theoretical possibilities with time-series data. If the distance between samples does not change as time-lags increase, then the community is considered to be stable. If sample distance increases over time, the community is unstable and undergoing directional change. If sample distance decreases over time, then the community is unstable and undergoing convergence.

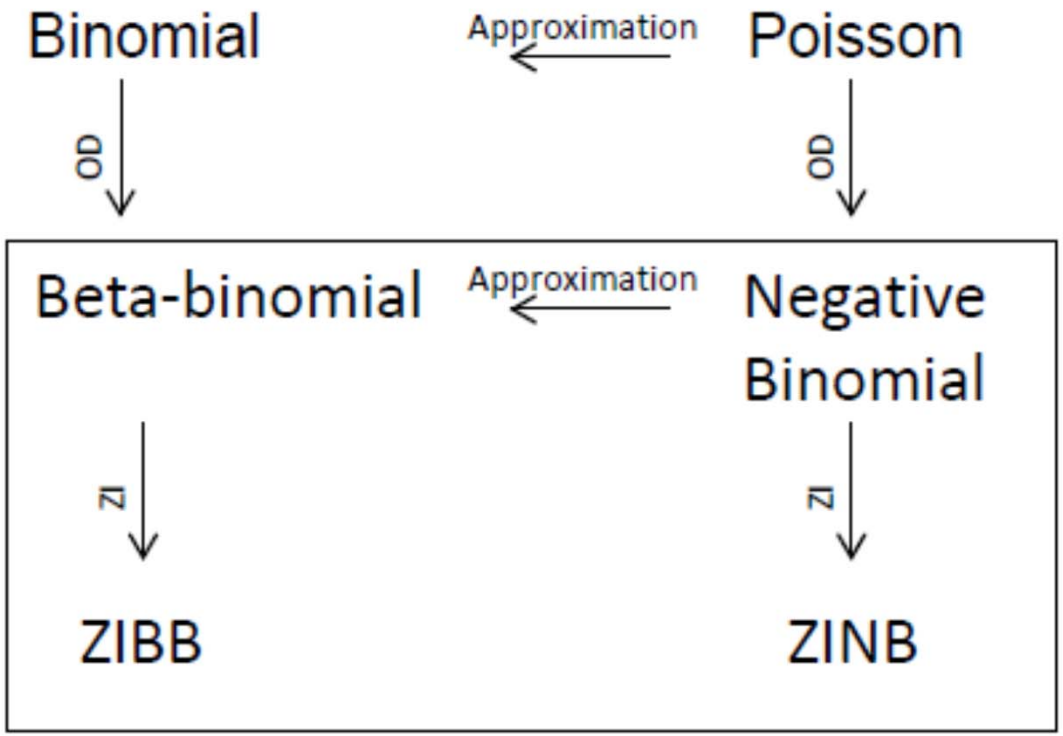
Univariate

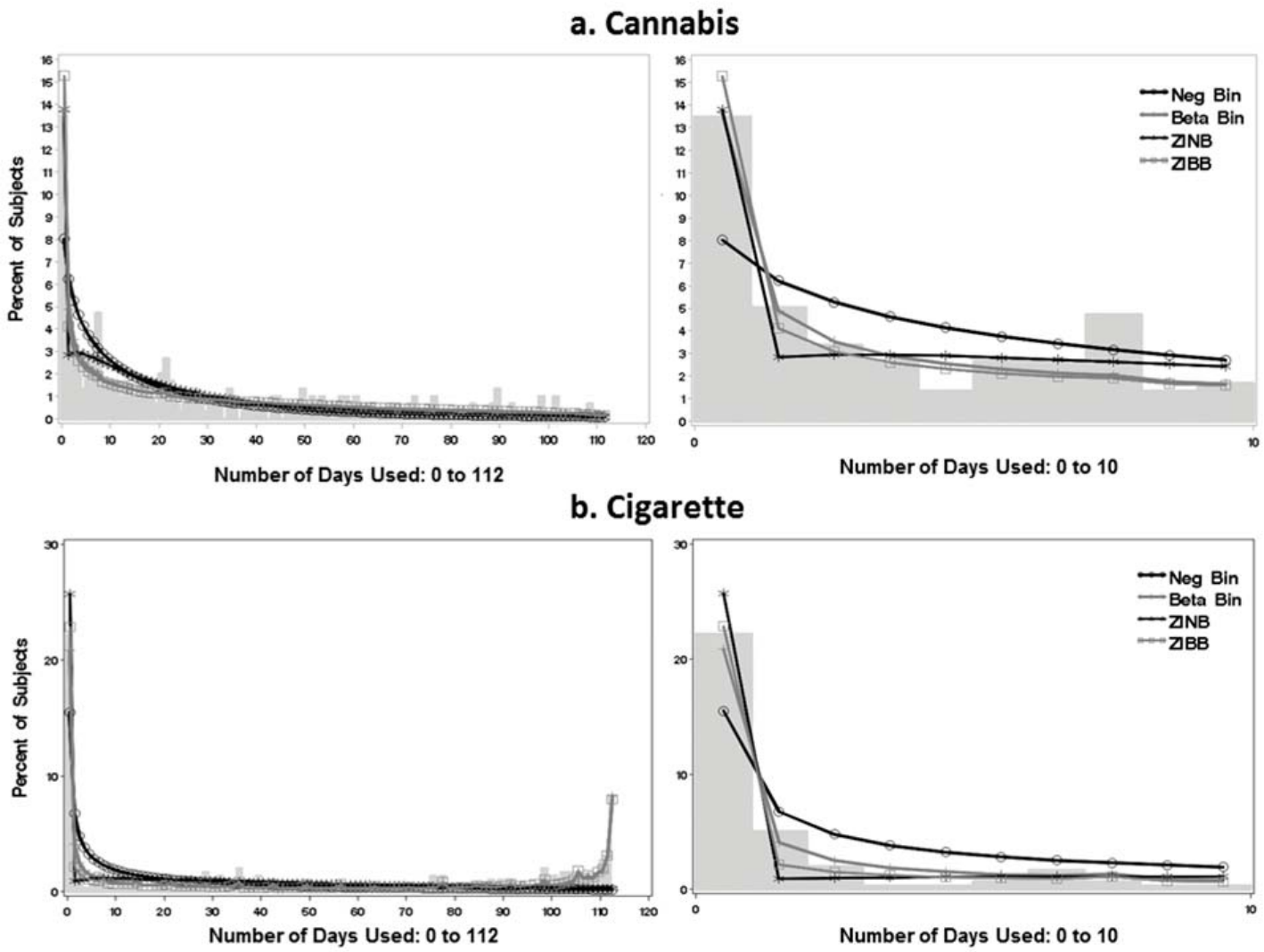
Zero-inflated regression: overview

- provides an explicit way of modeling the excess zeros, as well as allowing the variance of the outcome to differ from the mean.
- Mixture distribution of binary and ordinary count distribution such as Poisson or negative binomial (ZIP or ZINB).
- Extend to random-effects zero-inflated regression models.
  - accounts for the subject to subject variation to directly model correlation among the repeated measures within a subject.

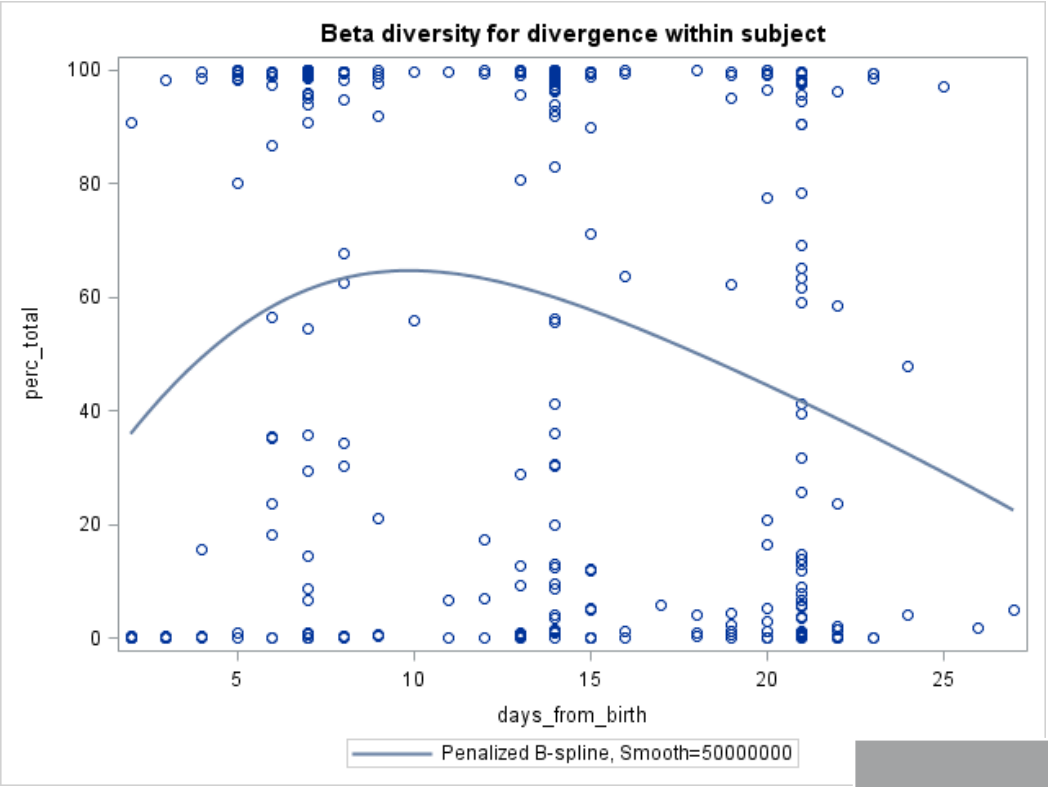
General approach to modeling

- Add t




a. Cannabis
Percent of Subjects
Number of Days Used: 0 to 112
Number of Days Used: 0 to 10
Neg Bin
Beta Bin
ZINB
ZIBB
b. Cigarette
Percent of Subjects
Number of Days Used: 0 to 112
Number of Days Used: 0 to 10
Neg Bin
Beta Bin
ZINB
ZIBB

BPD dataset - Staphylococcus



Piece-wise linear regression with knot at 10 days

OD parameter – $p < 0.01$ for both

	NB	BB
Slope <10 days	0.08 (0.04)	0.12 (0.05)
Δ slope > 10 days	-0.13 (0.05)	-0.19 (0.07)

The ZINB model for human microbiota sequence data

- λ_{ij} is the mean of sequence count for ith subject and jth time
- π_{ij} is the zero-inflated factor for ith subject and jth time
- Linear predictor for negative binomial component

  $\log(\lambda_{ij}) = \beta 0 + \beta 1 \times time_{ij} + (ltotal)_{ij} + ui$

  ltotal is the natural logarithm of the total sequence counts (offset)
- Linear predictor for binary component

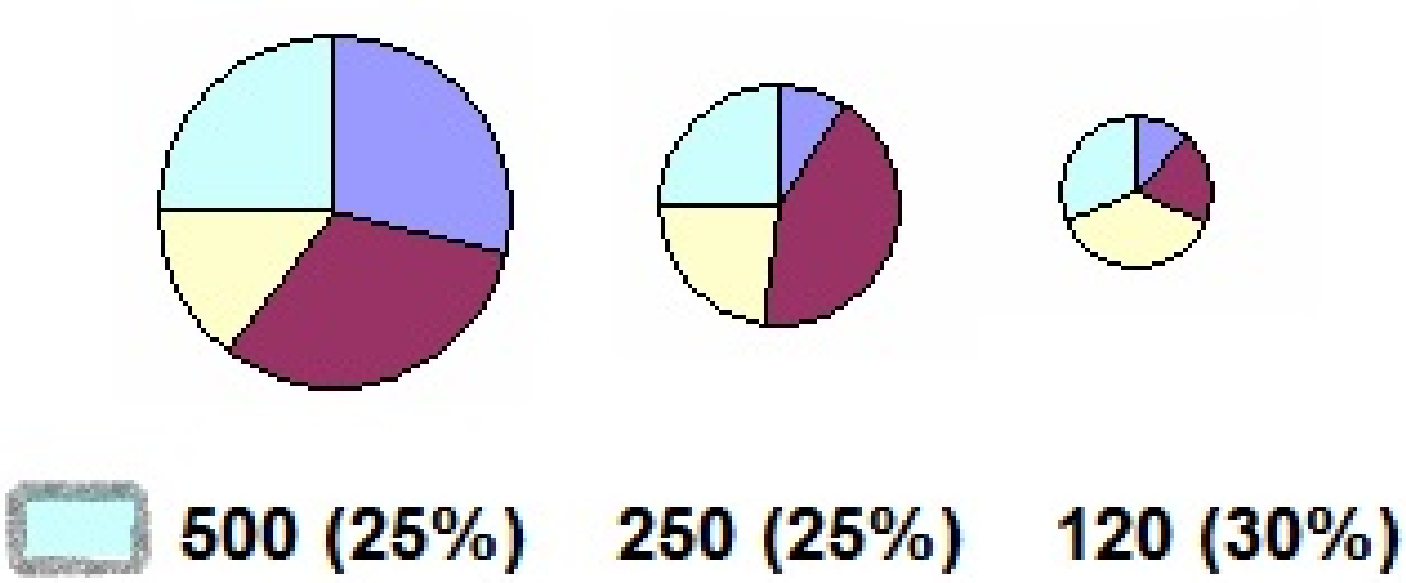
  $\log(\pi_{ij}/(1-\pi_{ij})) = \gamma 0 + \gamma 1 \times time_{ij} + vi$
- Here, ui and vi are the random intercepts and they are assumed to be independent and follow the bivariate normal distributions as

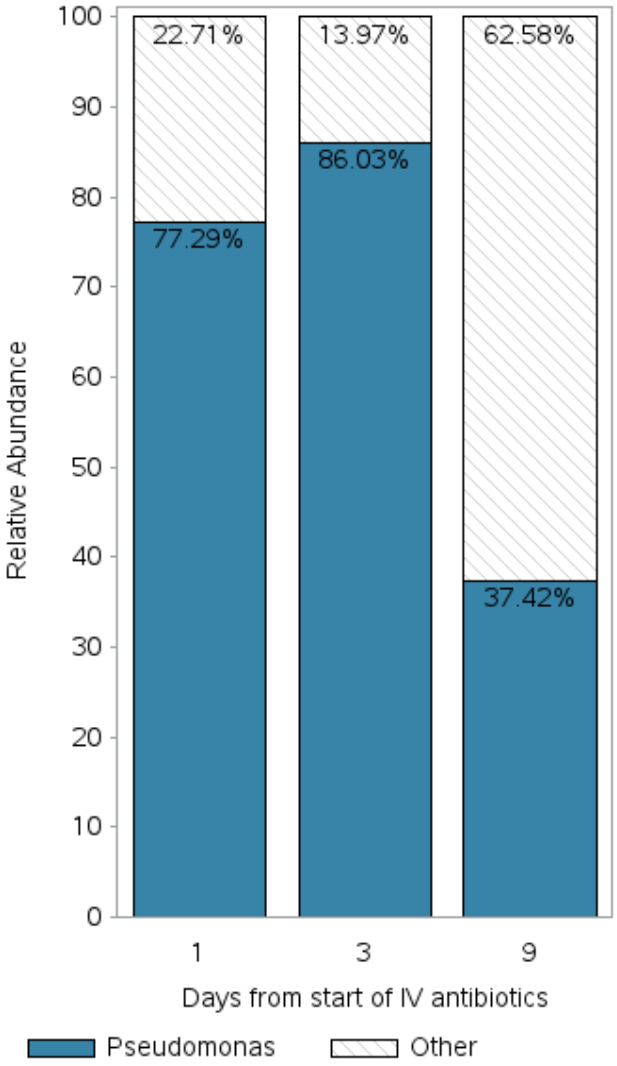
$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix}\right).$$

Independence of random effects

- For simplicity, we assume the two random effects are independent
- Although this is not a necessary assumption, it is commonly used in the previous literature regarding ZIP/ZINB with random effects.
- We believe that the process that generates the structural zeros (dependent on sequencing depth) is independent of the process that generates the counts

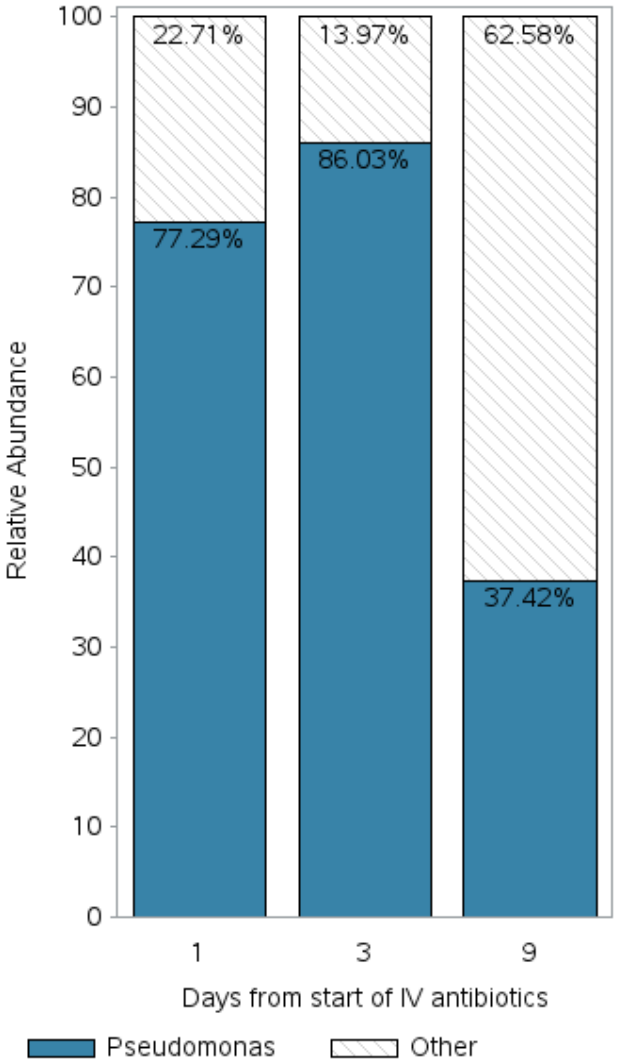
Other Considerations - Relative Abundance vs. Absolute Abundance





The use of relative abundance indicates an increase in *Pseudomonas* between days 1 and 3 (left)

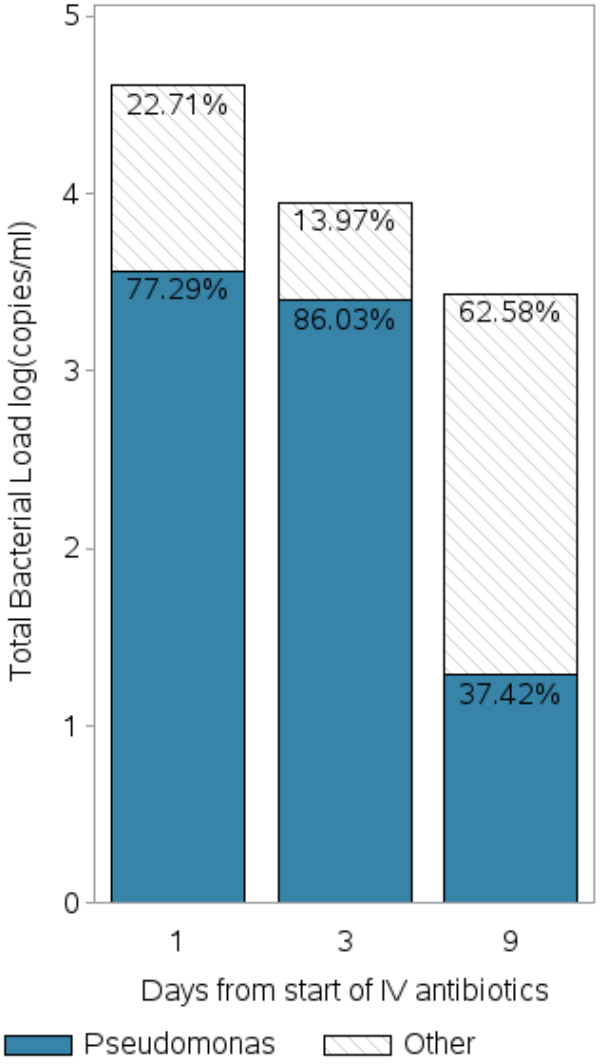
Simple example: 3 samples collected over time for a single subject. Community is simplified to Pseudomonas vs everything else.



The use of relative abundance indicates an increase in *Pseudomonas* between days 1 and 3 (left)

whereas if total bacterial load is considered, the absolute abundance decreases slightly (right).

In addition to community composition, there is evidence to suggest also looking at overall community size



Thanks

MiRC Collaborators:

J. Kirk Harris

Charles Robertson

Dan Frank

Graduate Students:

Rui Fang

Miranda Kroehl

Clinical collaborators:

Frank Accurso

Edith Zemanick

Pete Mourani

Marci Sontag

Sophie Fillon

Kristen Demourelle

References

- Nelson *et al.* Temporal Profiling of the Bacterial and Fungal Communities in DeltaF508 Adult Cystic Fibrosis Sputum. PeerJ. *In Press*
- Flores *et al.* Temporal variability is a Personalized Feature of the Human Microbiome. Genome Biol. 2014:15(12);531
- Zemanick et al. Assessment of Airway Microbiota and Inflammation in CF children using multiple sampling techniques. The Annals of the American Thoracic Society. In Press.
- Zemanick et al. Inflammation and Airway Microbiota during Cystic Fibrosis Pulmonary Exacerbations. PLoS ONE 8(4): e62917.
- Gajer et al. Temporal Dynamics of the Human Vaginal Microbiome. Sci Transl Med. 2012:4(132);132ra52
- Harrison et al. Principal variance components analysis of crop composition data. J. Agric Food Chem. 2013:16;6412-6422
- Collins et al. A Method to determine rates and patterns of variability in ecological communities. Oikos. 2000:91(2);285-293.