Latent variables explain dependencies in bacterial communities

Susan Holmes
http://www-stat.stanford.edu/~susan/
@SherlockpHolmes

Bio-X and Statistics, Stanford University

IPAM, January, 23rd 2019
NIH-TR01

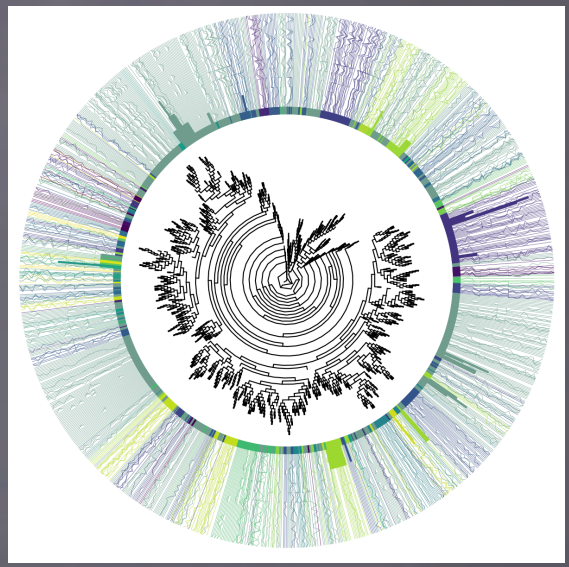
Multidomain, multiway data

Human Microbiome

Joint work with David Relman and his Lab, funded by NIH TR01: Perturbations and Resilience of the Human Microbiome and March of Dimes.

- Effect of Antibiotics.
- Colonic Cleanout.
- Diet perturbations.
- .....and March of Dimes study of pregnancy.

Challenges when working with Longitudinal Multidomain data



Keeping all the data together

- Data and Heterogeneity.
- Graph or Tree integration.
- Longitudinal data are dependent (less information).
- Reproducibility of results across labs, experimental conditions and users.

Paths in thinking about these heterogeneous systems

- You can use distances between very general objects (trees, tables, matrices, graphs).
- You can use Graph or Trees to ”influence” these distances (Structured high-dimensionality).
- Mixtures are everywhere (not one parametric population).
- Latent variables or factors are an enormous resource.
- Don’t stress about choices, they are not forever (because of reproducible workflows).
- Think very carefully when you are throwing out information of any sort.
- Be lazy: re-use and recycle methods, vocabulary and infrastructure.

Heterogeneity of Data

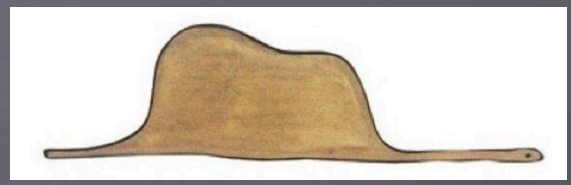
- Status : response/ explanatory.
- Hidden (latent)/measured.
- Types :
  - Continuous
  - Binary, categorical
  - Graphs/ Trees
  - Images
  - Maps/ Spatial Information
  - Rankings
- Amounts of dependency: independent/time series/spatial.
- Different technologies used (Sanger, 454, Illumina, MassSpec, RNA-seq, Imaging, CyTOF).

Human Microbiome: What are the data?

DNA	The Genomic material present (16sRNA-gene especially).
RNA	What genes are being turned on (gene expression), transcriptomics.
Mass Spec	Specific signatures of chemical compounds present.
Clinical	Multivariate information about patients' clinical status, medication, weight.
Environmental	Location, nutrition, time.
Domain Knowledge	Metabolic networks, phylogenetic trees, gene ontologies.

Paths in thinking about these heterogeneous systems

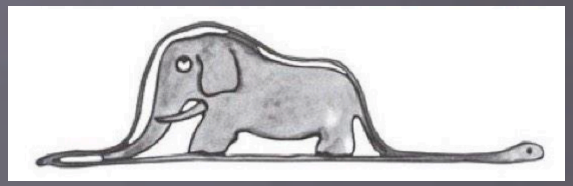
- Think in layers: latent variables or factors enable interpretation.



hidden variables.

Paths in thinking about these heterogeneous systems

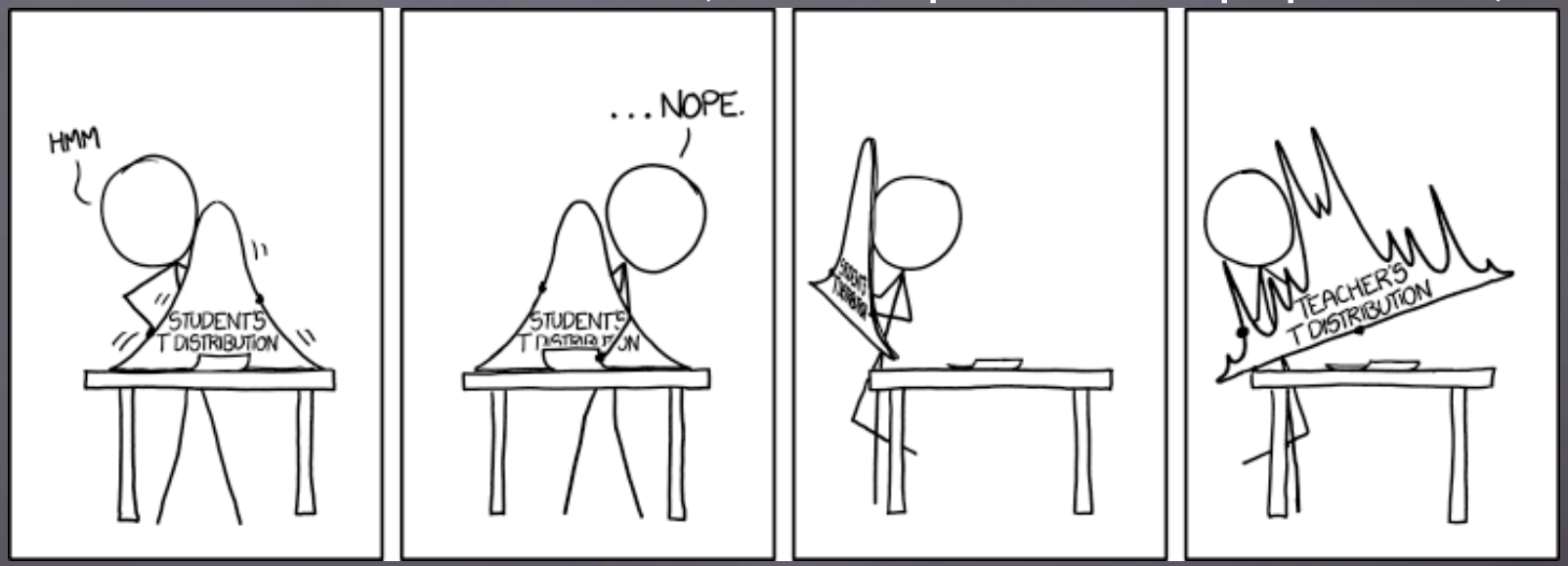
- Think in layers: latent variables or factors enable interpretation.



hidden variables.

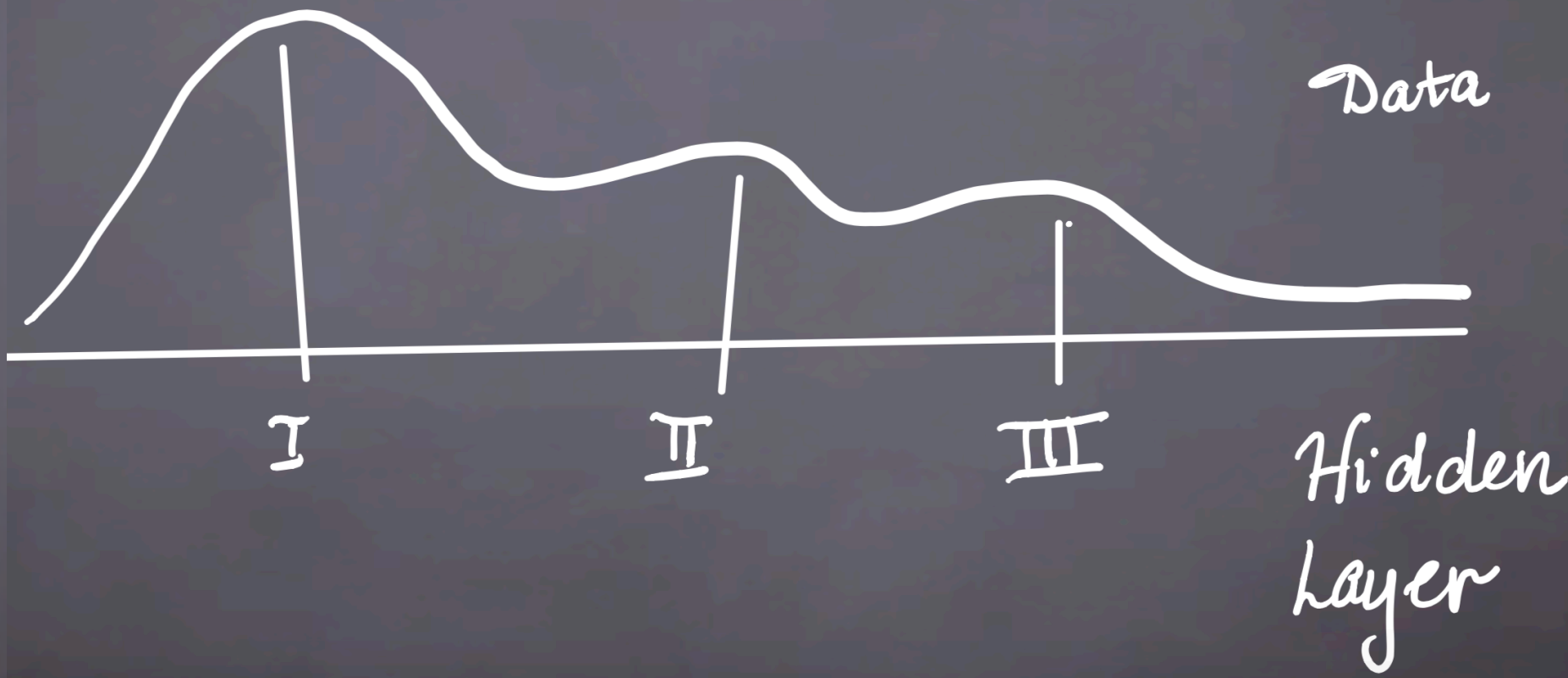
Paths in thinking about these heterogeneous systems

- Think in terms of mixtures (not one parametric population).



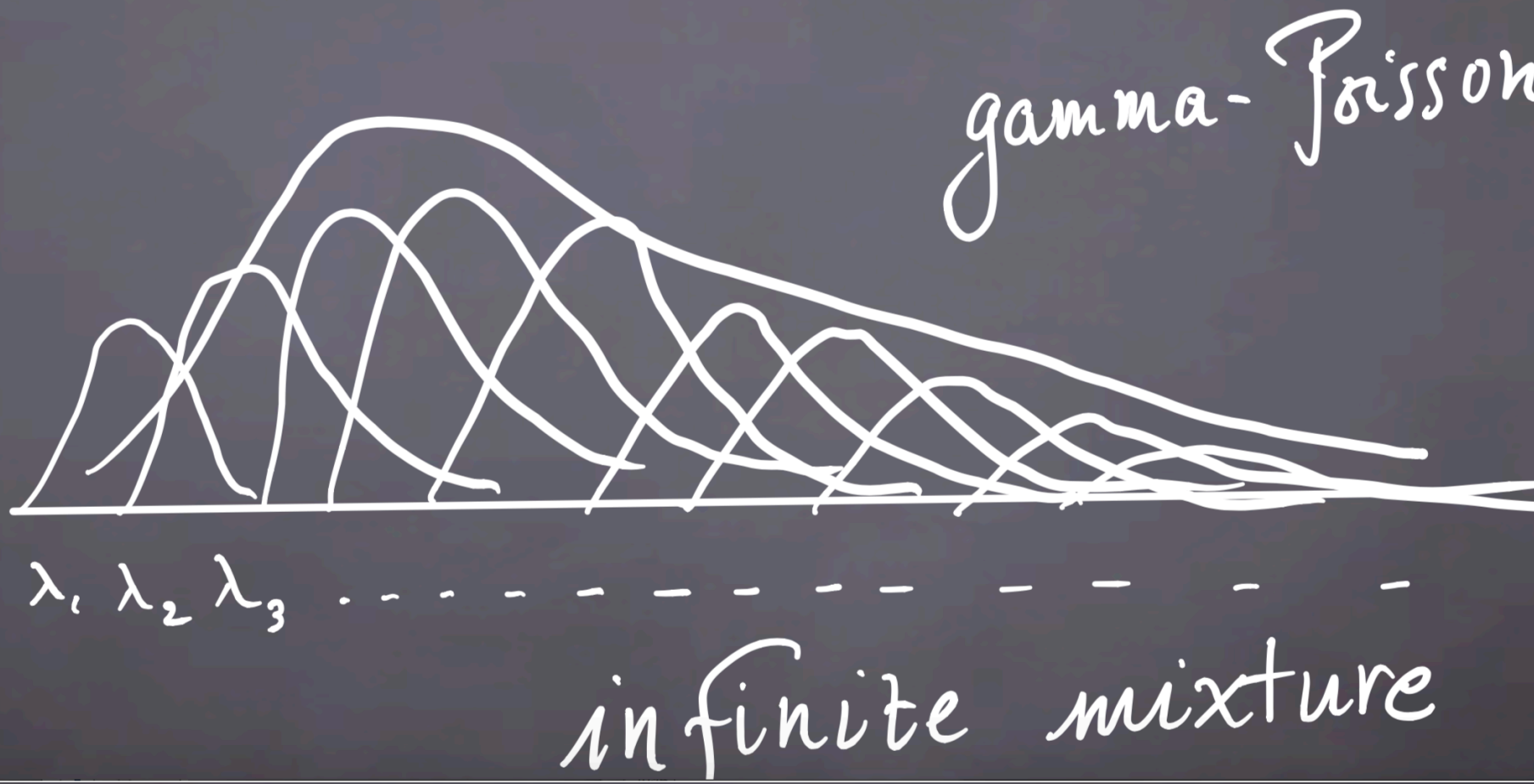
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.



Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation



Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences. Different taxa are identified as Amplicon Strain Variant (ASV) generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p_{tt}} = (p_1, p_2, \ldots p_J) \qquad \text{For} \quad J \ \text{ASV's}$$

$$\mathbf{p_{ctl}} = (p_1, p_2, \ldots p_J) \qquad \Delta = \text{diff}(\mathbf{p_{tt}} - \mathbf{p_{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors.

Example in microbiome: unknown parameters?

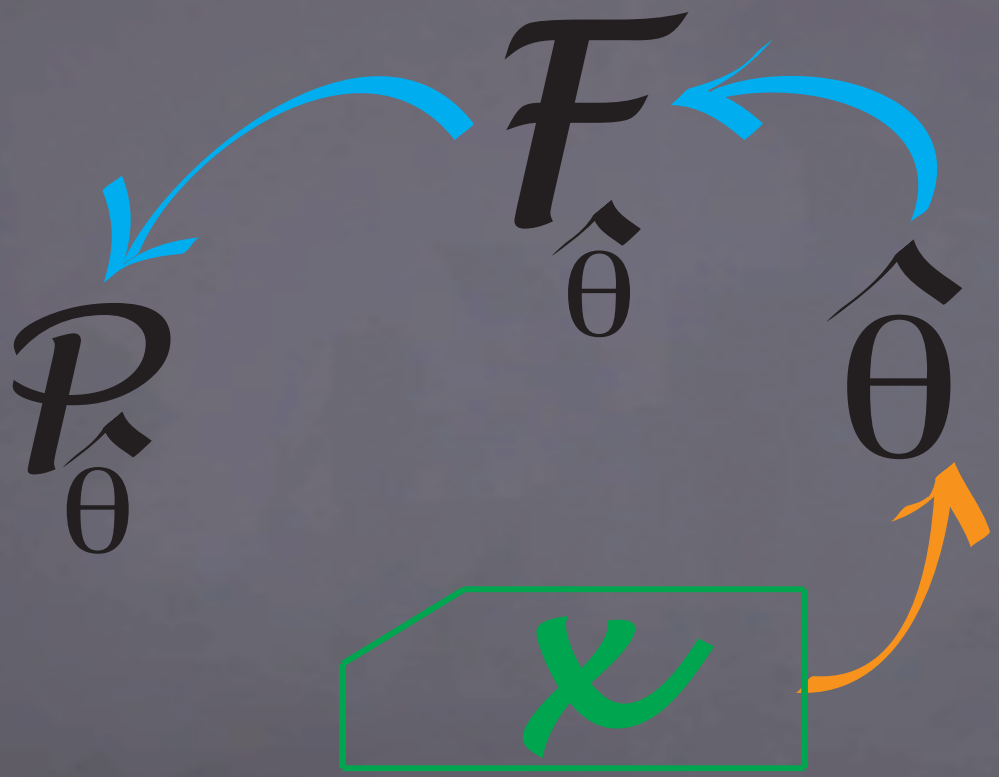
The relative abundances of bacteria and their differences. Different taxa are identified as Amplicon Strain Variant (ASV) generated with **DADA2** (Callahan et al., 2017)

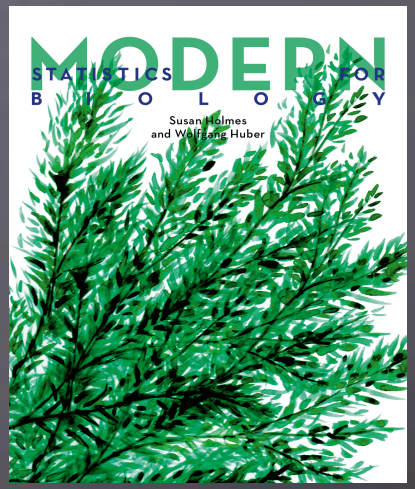
$$\mathbf{p_{tt}} = (p_1, p_2, \ldots p_J) \qquad \text{For} \quad J \ \text{ASV's}$$

$$\mathbf{p_{ctl}} = (p_1, p_2, \ldots p_J) \qquad \Delta = \text{diff}(\mathbf{p_{tt}} - \mathbf{p_{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors. We need to quantify the uncertainty we have on the parameters.

Statistics: separate the model from the data





See a complete book:

`http://bios221.stanford.edu/book/`

Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences.

Different taxa are identified as Amplicon Strain Variant (ASV) generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p_{tt}} = (p_1, p_2, \ldots p_J) \qquad \text{For} \quad J \ \text{ASV's}$$

$$\mathbf{p_{ctl}} = (p_1, p_2, \ldots p_J) \qquad \Delta = \text{diff}(\mathbf{p_{tt}} - \mathbf{p_{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors.
Models for noise: hierarchical Gamma-Poisson: we know how to transform the data to stabilize the variance (Delta-method).
McMurdie and Holmes (2014) "Waste Not, Want Not: Why rarefying microbiome data is inadmissible", PLOS Comp.Bio.

Read data are counts, the data are not compositional.

We do not summarize them to ratios or “relative abundance”.

- After perturbations amounts of bacteria go up & down.
- Remove contaminants using read numbers (`decontam`).
- Estimating depth bias requires read numbers.
- We need the read depths for variability/standard error estimation and uncertainty quantification.
- Transform the data to equalize the variance.

Some real data (Caporoso et al, 2011)

```
> GlobalPatterns

phyloseq-class experiment-level object
otu_table()  OTU Table:         [ 19216 taxa and 26 samples ]
sample_data()Sample Data:       [ 26 samples by 7 sample variable
tax_table()Taxonomy Table:    [ 19216 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree:[ 19216 tips and 19215 internal nod

> sample_sums(GlobalPatterns)
     CL3      CC1      SV1  M31Fcsw  M11Fcsw  M31Plmr  M11Plmr
  864077  1135457   697509  1543451  2076476   718943   433894
 .....
     NP3      NP5  TRRsed1  TRRsed2  TRRsed3     TS28     TS29
 1478965  1652754    58688   493126   279704   937466  1211071

> summary(sample_sums(GlobalPatterns))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  58690  567100 1107000 1085000 1527000 2357000
```

Part I

Improving data quality using frequencies

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan[1], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo A Johnson[2] & Susan P Holmes[1]

We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjjneb/dada2). DADA2 infers sample sequences exactly and resolves differences of as little as

We previously introduced the Divisive Amplicon Denoising Algorithm (DADA), a model-based approach for correcting amplicon errors without constructing OTUs[5]. DADA identified fine-scale variation in 454-sequenced amplicon data while outputting few false positives[2,5].

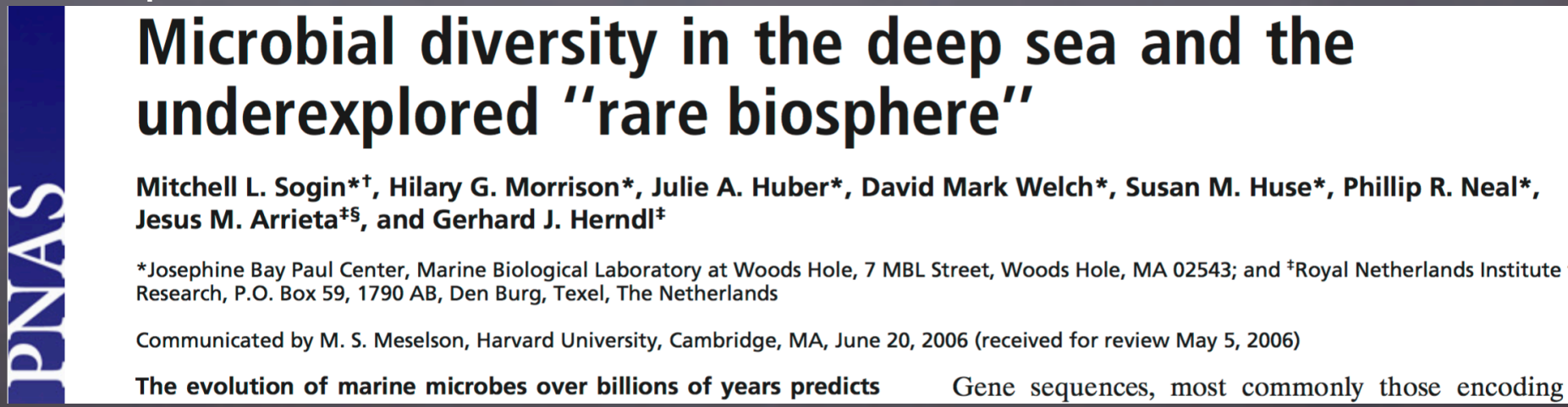
Here we present DADA2, an open-source R package (https://github.com/benjjneb/dada2, **Supplementary Software**) that extends and improves the DADA algorithm. DADA2 implements a new quality-aware model of Illumina amplicon errors. Sample composition is inferred by dividing amplicon reads into partitions consistent with the error model (Online Methods). DADA2 is reference free and applicable to any genetic locus. The DADA2 R package implements the full amplicon workflow: filtering, dereplication, sample inference, chimera identification, and merging of paired-end reads.

We compared DADA2 to four algorithms (Online Methods):

Diversities in the microbiome depend on the number of taxa

- α-diversity: Number of 'species'-taxa in a biological sample (from one location).
- β-diversity: Differentiation in diversity among different samples from different locations.

Extremely sensitive to noise.
Fake species:



PNAS

Microbial diversity in the deep sea and the underexplored "rare biosphere"

Mitchell L. Sogin*†, Hilary G. Morrison*, Julie A. Huber*, David Mark Welch*, Susan M. Huse*, Phillip R. Neal*, Jesus M. Arrieta‡§, and Gerhard J. Herndl‡

*Josephine Bay Paul Center, Marine Biological Laboratory at Woods Hole, 7 MBL Street, Woods Hole, MA 02543; and ‡Royal Netherlands Institute Research, P.O. Box 59, 1790 AB, Den Burg, Texel, The Netherlands

Communicated by M. S. Meselson, Harvard University, Cambridge, MA, June 20, 2006 (received for review May 5, 2006)

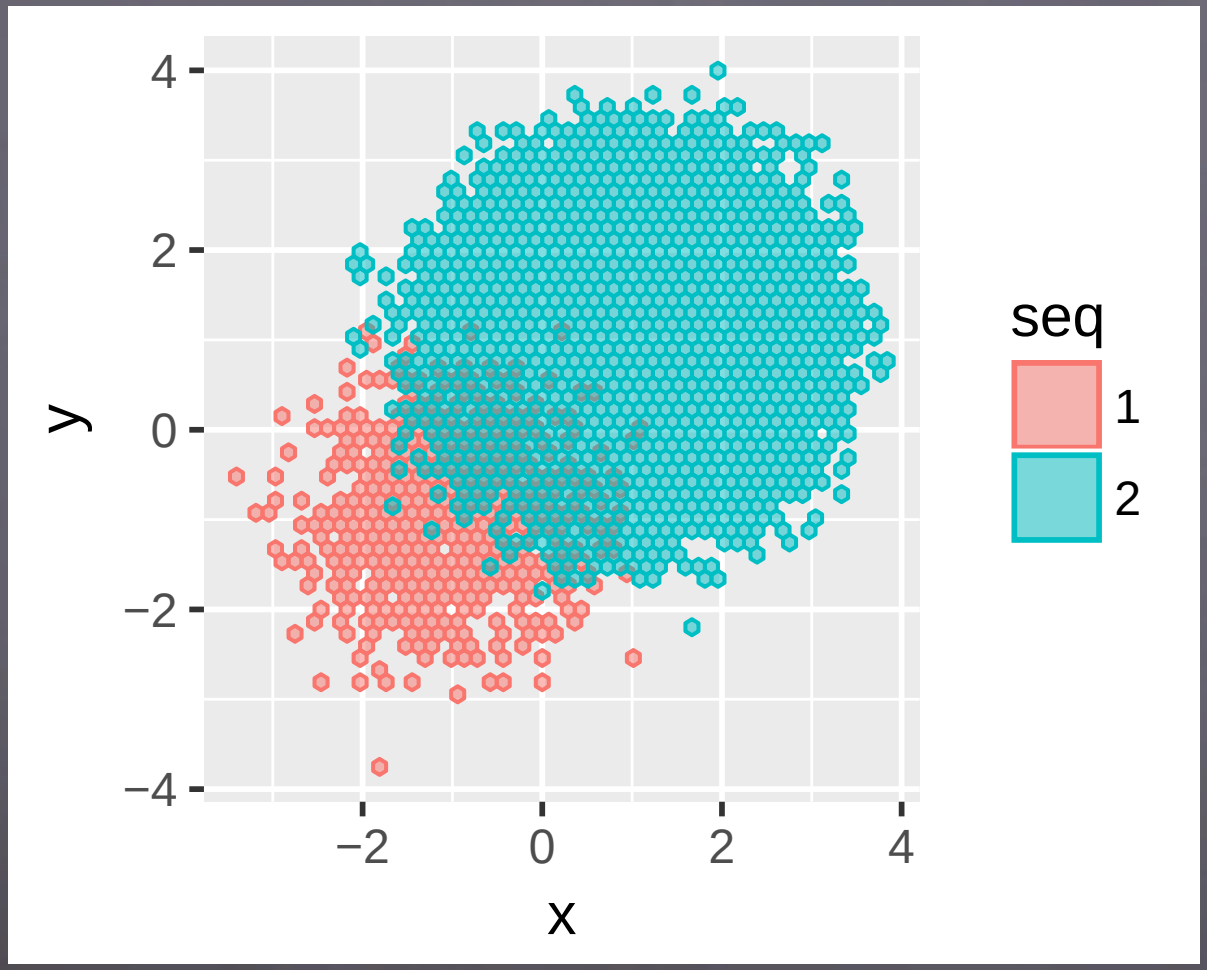
The evolution of marine microbes over billions of years predicts Gene sequences, most commonly those encoding

How many words does Professor D. know?

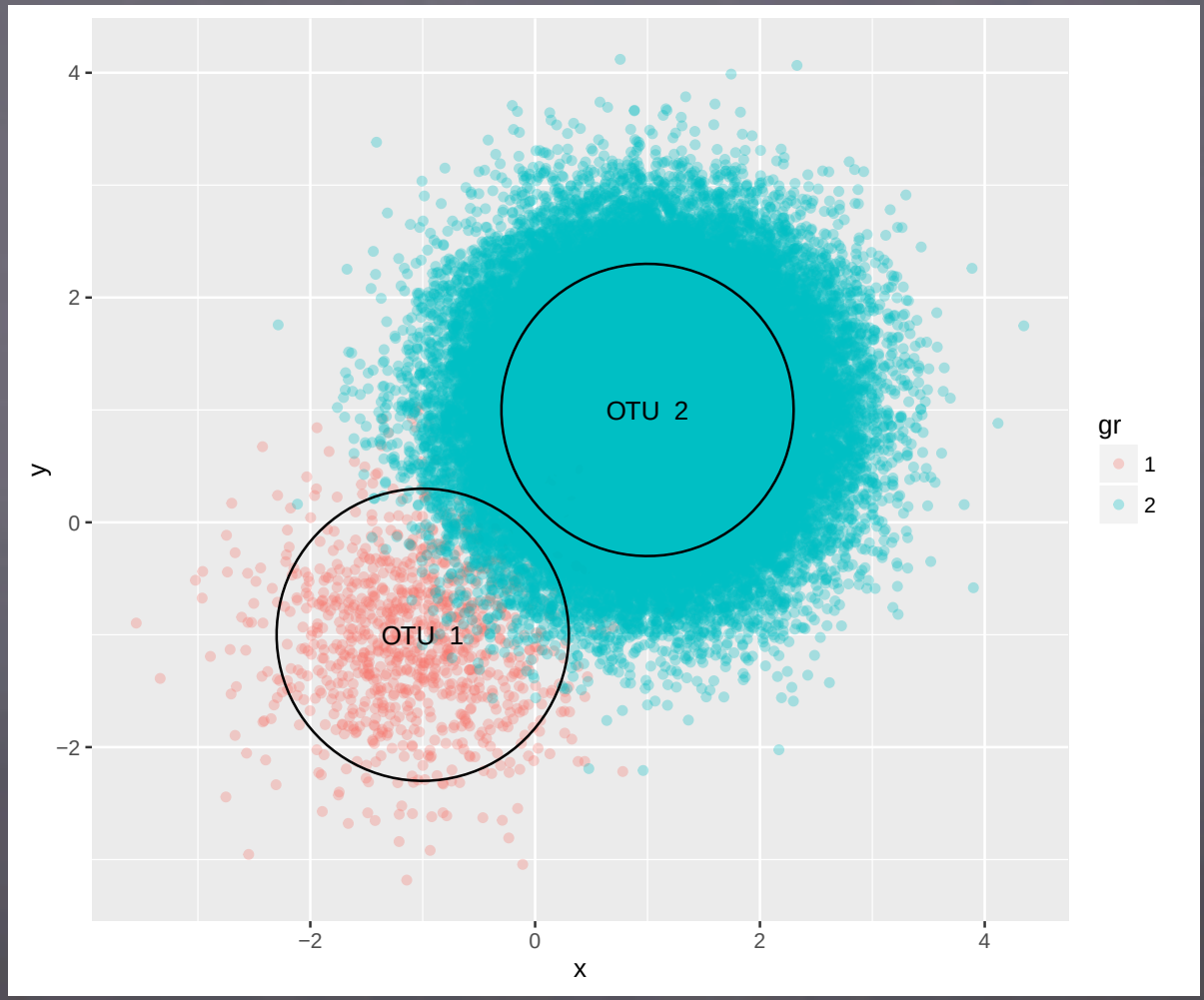
- Maybe 15,000, 20,000?
- Start sampling..... banana, bannana, bannanna, orange, orenge, muscle, musel, muscel, foreign, forene, forane,.....
- How many real words does Prof D. know?
- Use more information than the spelling....

The success of `dada2` is in it's use of the **frequencies**, often forgotten or hidden from the user if you only inventory the different sequences.

From reads to Operational Taxonomic Units

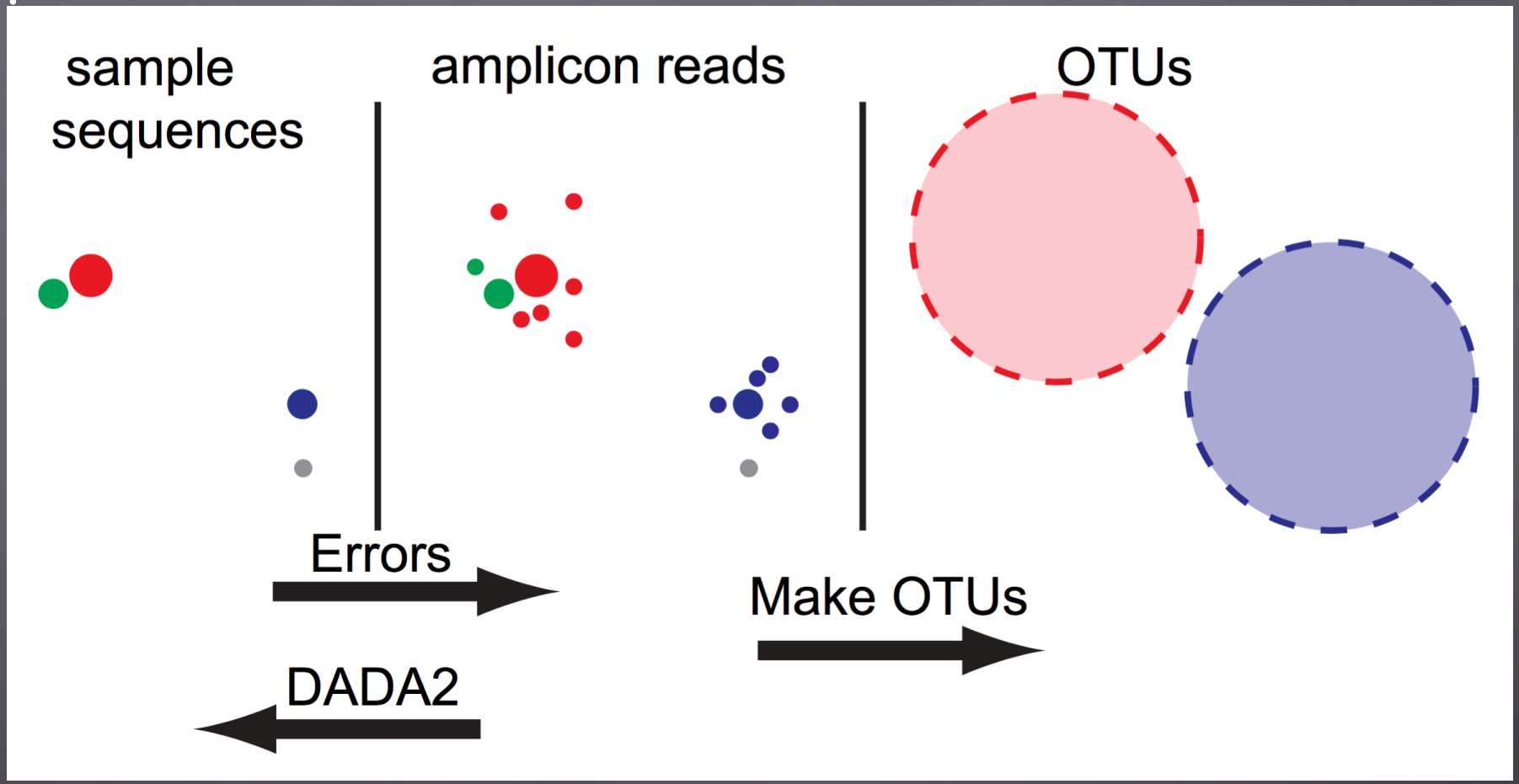


From reads to Operational Taxonomic Units



Curent practice (`qiime`, `mothur`, `rdp`, ...): 97% similarity.

Probabilistic Model

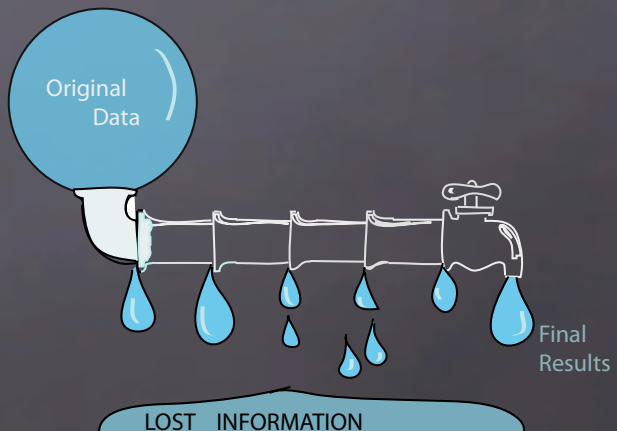


Error Model

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                 |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

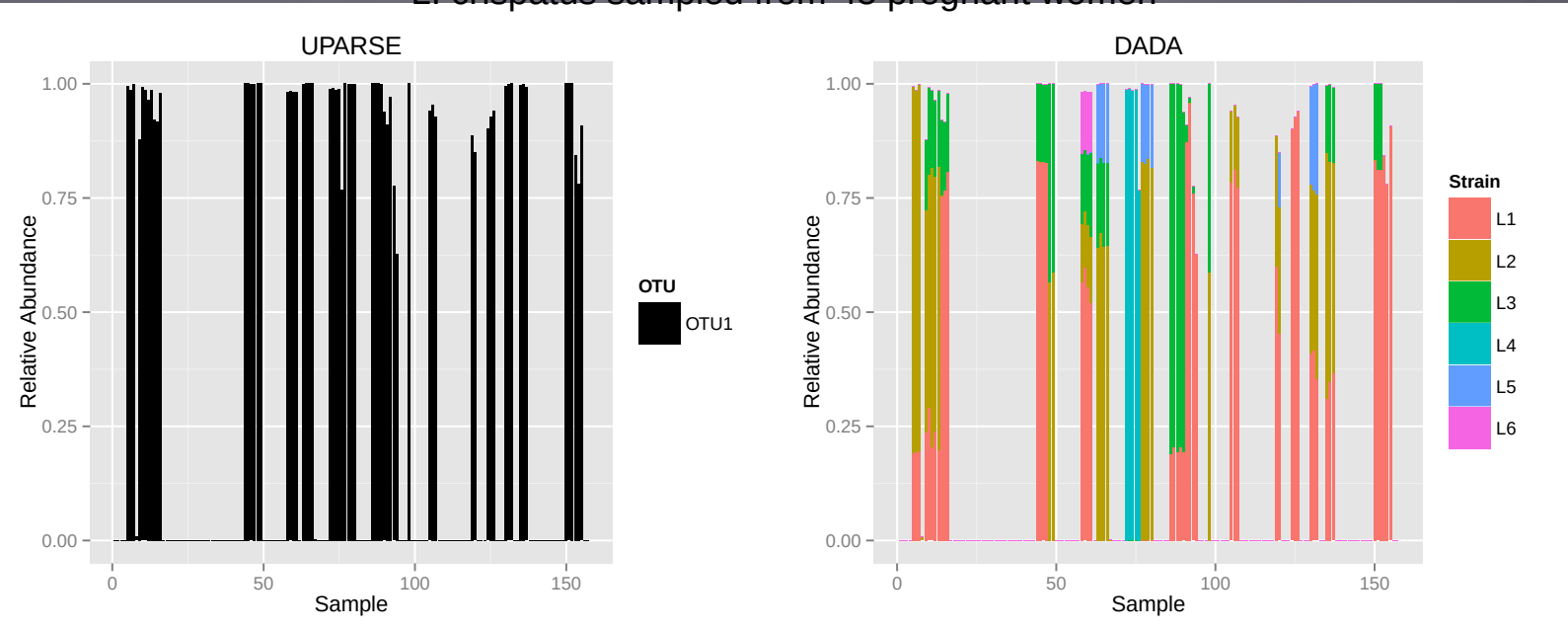
$$P(r|s) = \prod_{i=1}^{L} P(r(i)|s(i), q_r(i), Z)$$

P probabilities of substitutions ($A- > C$)
q Quality score (Q=30) Batch effect (run)
Use the denoised sequence instead of the OTU[?].



Higher resolution strain clustering: DADA2

L. crispatus sampled from 45 pregnant women



R package:

`http://benjjneb.github.io/dada2/R/tutorial.html`

Part II

Communities and transitions

Pregnancy data: perturbation, stability and preterm birth

A case-control study of 49 pregnant women:

- 15 delivered preterm.
- From 40 of these women: 3,766 specimens collected weekly during gestation, and monthly after delivery.
- Sites:vagina, distal gut, saliva, and tooth/gum.
- 9 women: validation set collected after the first study was complete.

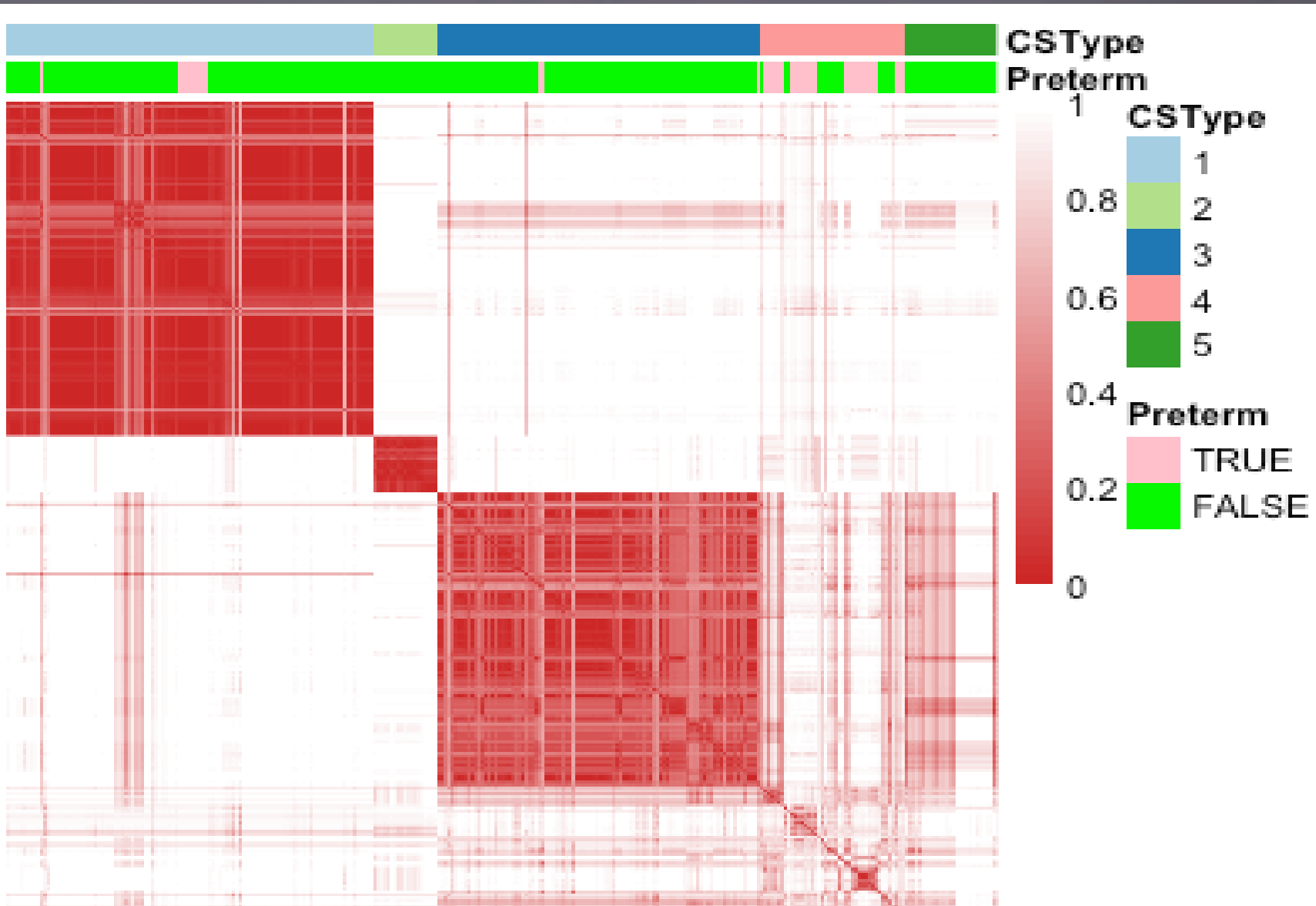
Methods used: variance stabilization through negative binomial, testing perturbations through linear mixed-effects modeling. Preterm prediction through medoid-based clustering and simple Markov chain.
Provided: Simple community temporal trends, community structure, and vaginal community state transitions.
DiGiulio DB, Callahan BJ, McMurdie PJ, ... & Holmes, SP and

Co-occurrence networks

Dual networks:

- Edges are created between taxa if in more than a certain proportion of samples share that taxa.
  This can be seen as a geometric graph with the distance being the Jaccard distance.
- Edges are created between samples if they share more than a certain proportion of taxa in common.

Communities of bacteria organize into 5 different types

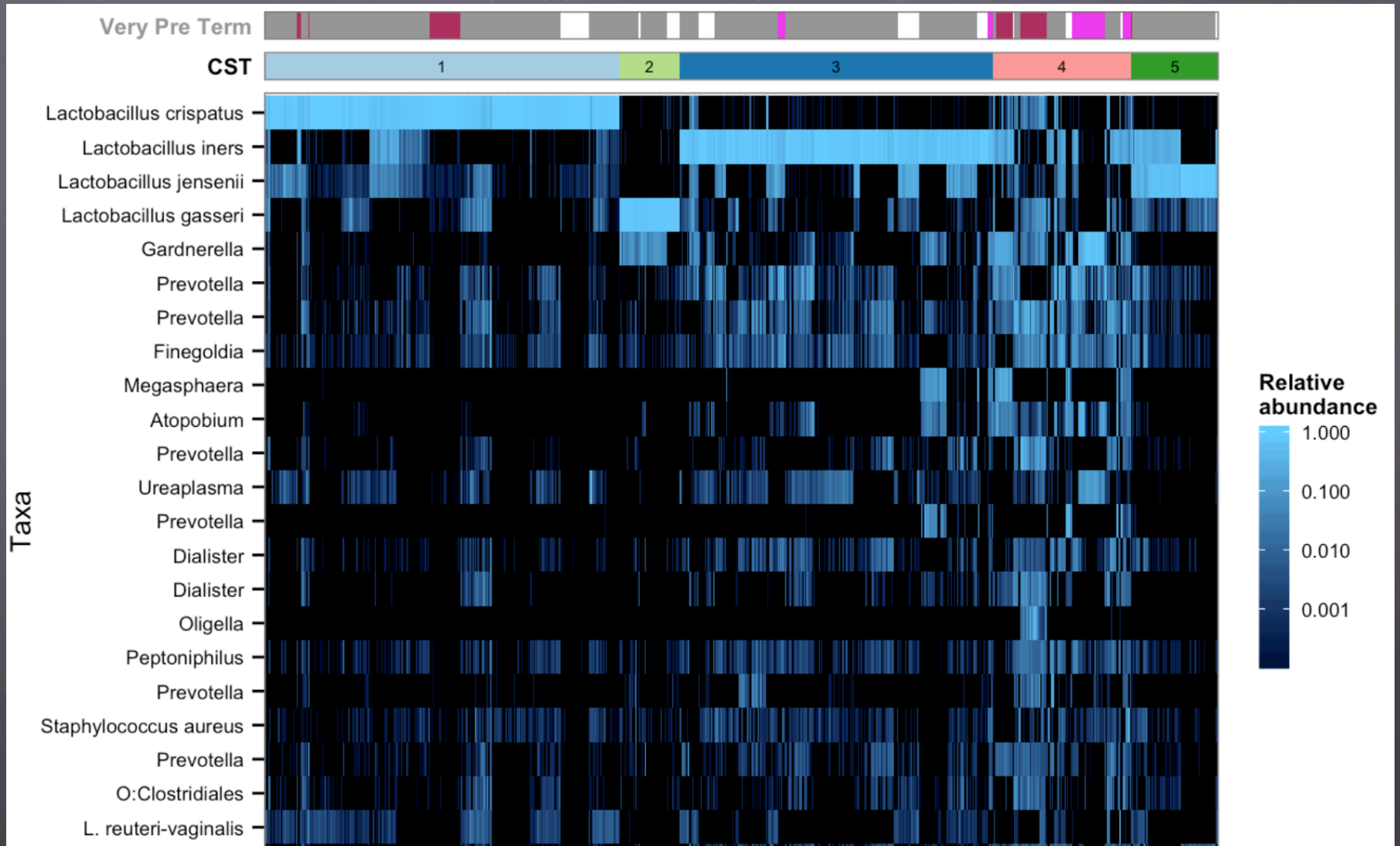


Questions asked?

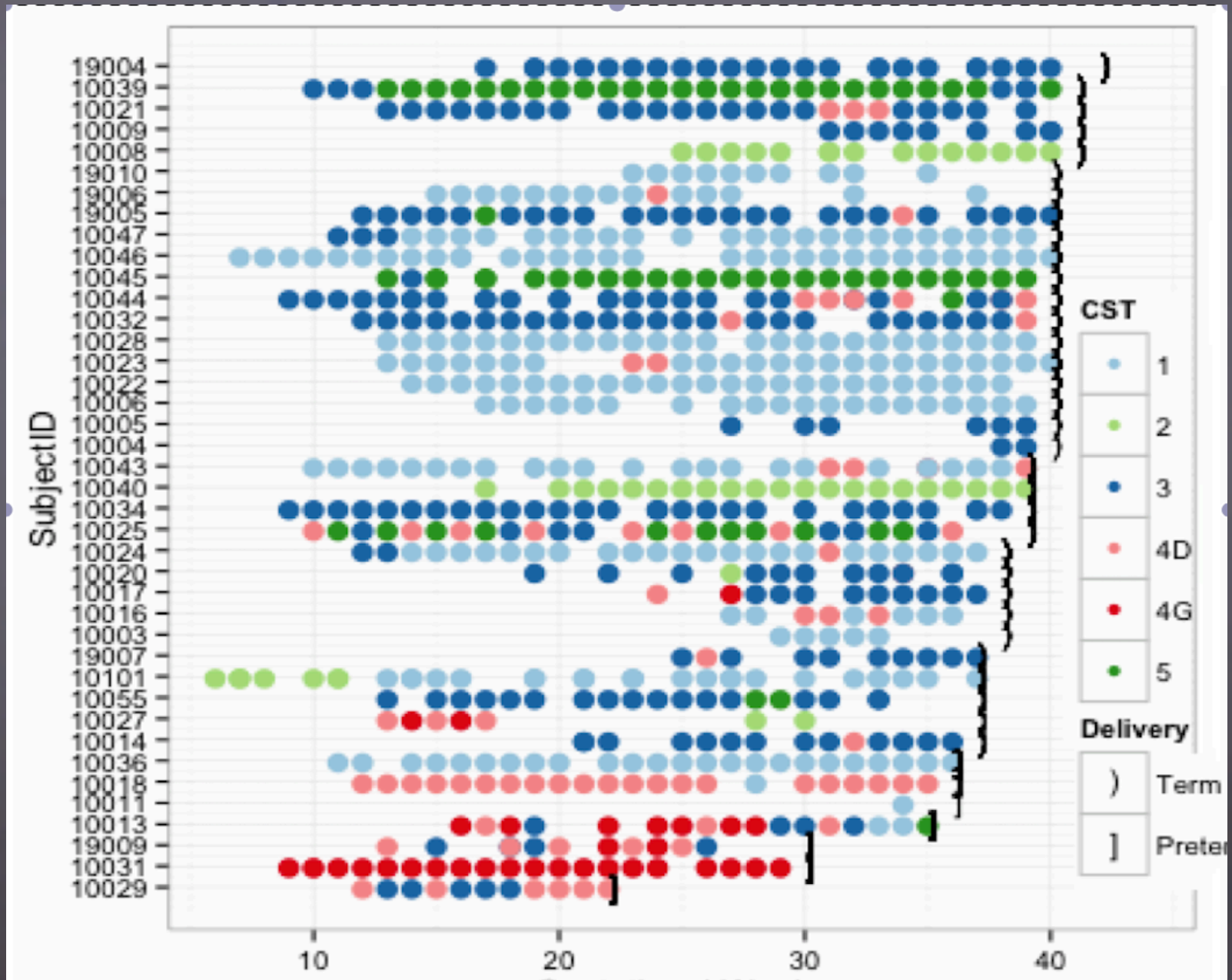
- Are the community state types the same as seen in previous studies?
- How stable are the communities within each individual during pregnancy?
- What alterations of the vaginal microbiome predict preterm birth?
- How early do these alterations occur?

Previously known Microbial Community State Types: Latent categorical variable.

Samples into community types and species patterns associated.



Longitudinal Analyses



Markov Chain Model

Transitions between states, as in simple ecological models.

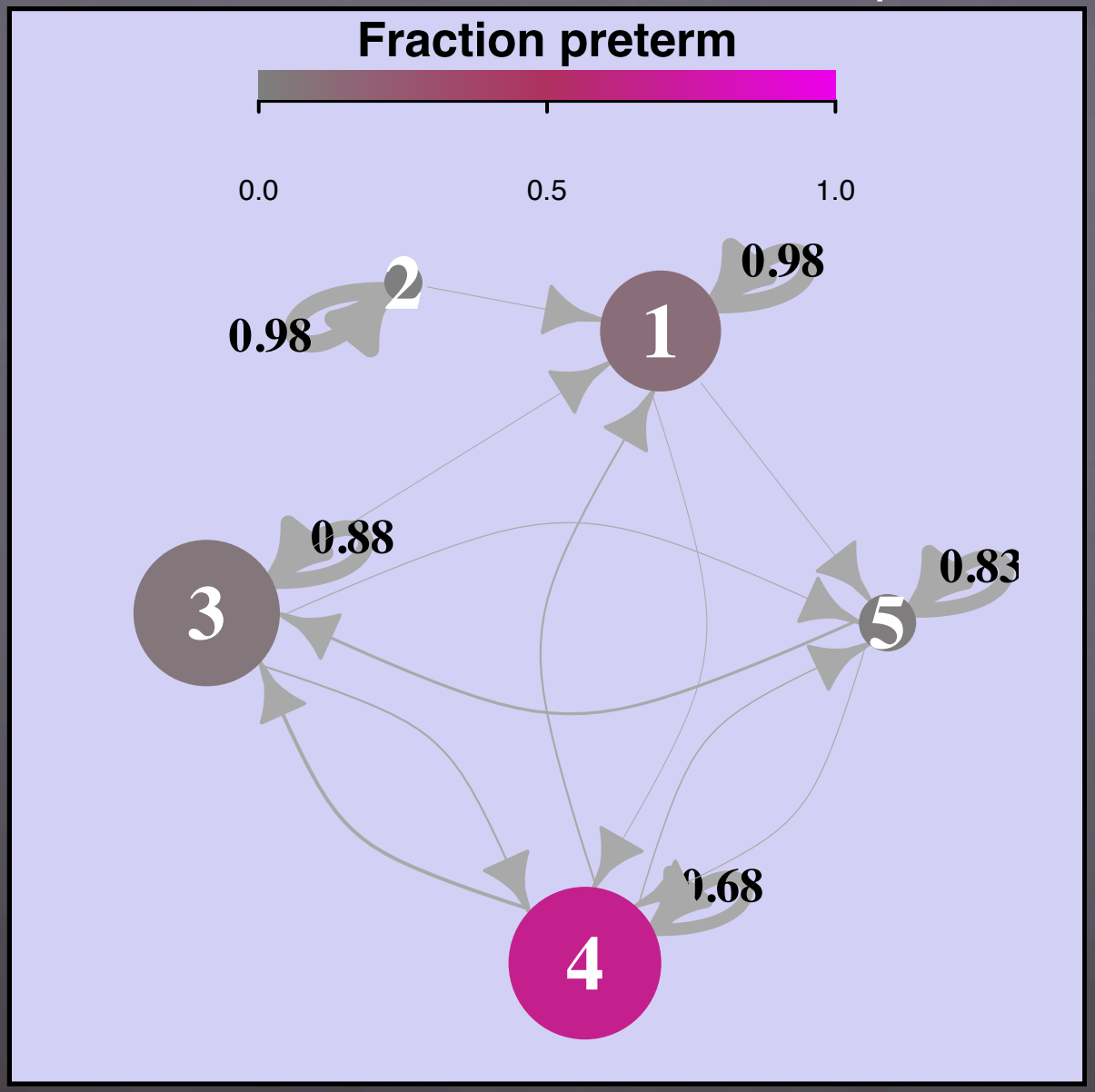
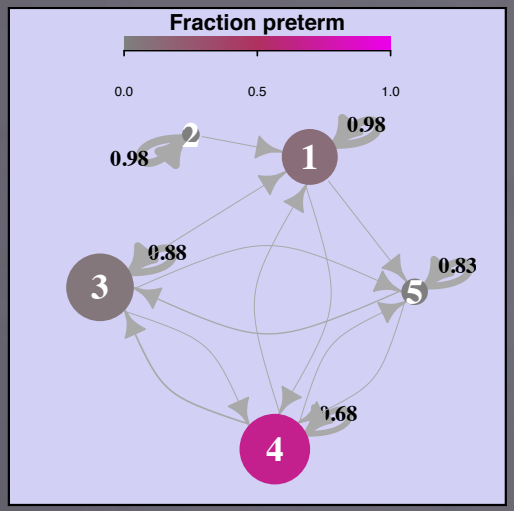
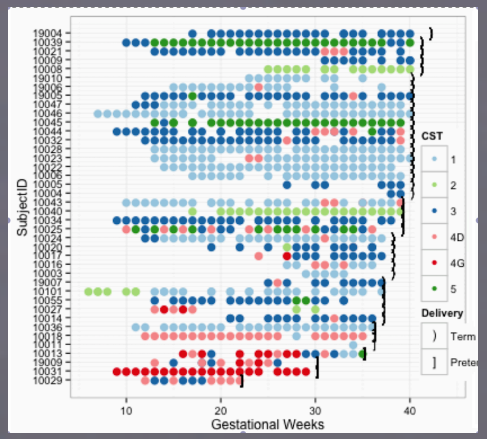


Illustration through Analyses





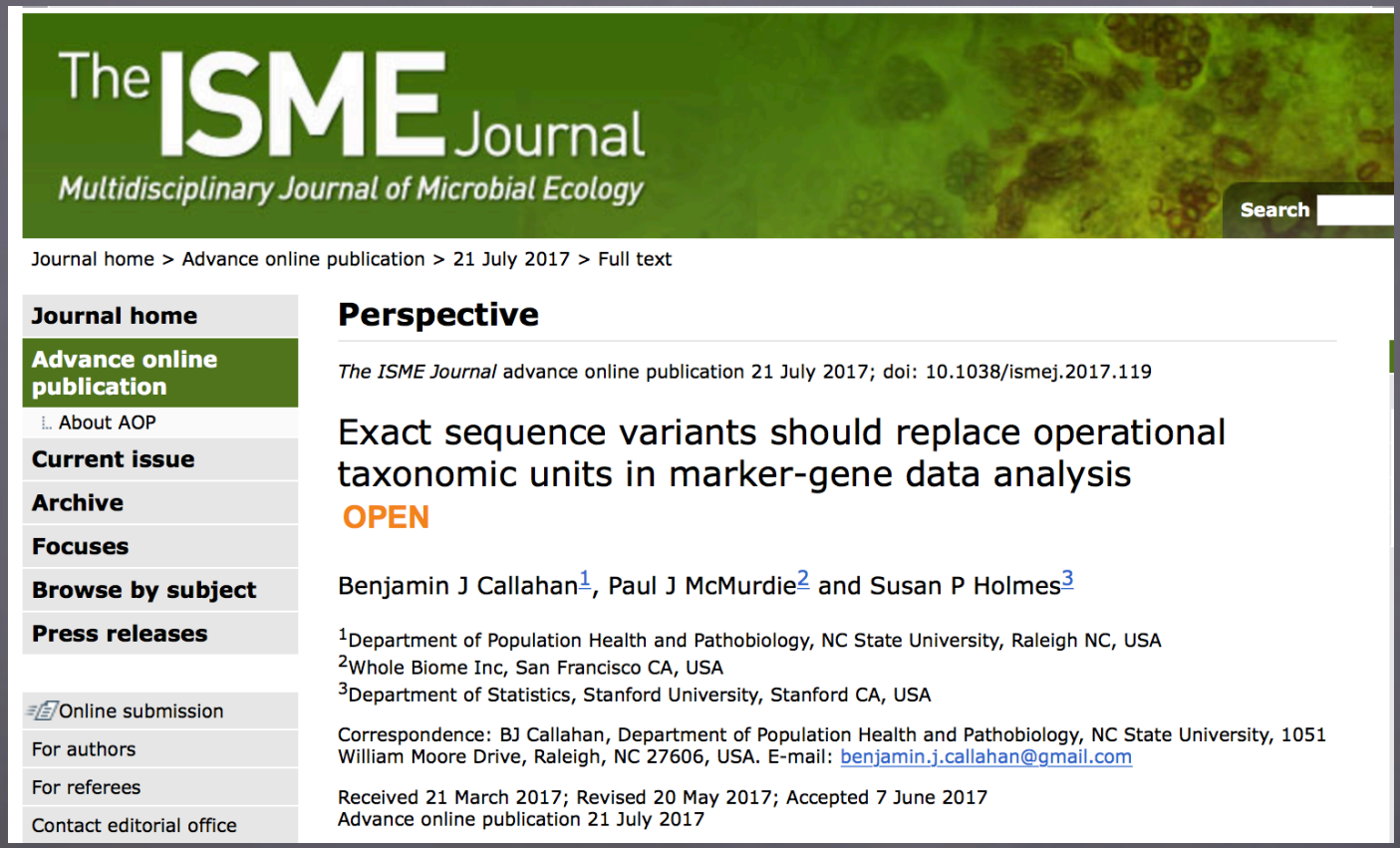
- Delivery Perturbation
- Preterm Prediction
- Stability

Conclusions for this study

- Microbiota community and diversity stable during pregnancy.
- Prevalence of a Lactobacillus-poor vaginal community state type (CST 4) was inversely correlated with gestational age at delivery (p=0.0039).
  Risk for preterm birth was more pronounced for subjects with CST 4 accompanied by elevated Gardnerella or Ureaplasma abundances.
- Finding validated with a separate diagnostic set of 246 vaginal specimens from nine women (four of whom delivered preterm).

Also : **BE LAZY**....

Be lazy: keep all the information



The ISME Journal
Multidisciplinary Journal of Microbial Ecology

Search

Journal home > Advance online publication > 21 July 2017 > Full text

Journal home
Advance online publication
About AOP
Current issue
Archive
Focuses
Browse by subject
Press releases

Online submission
For authors
For referees
Contact editorial office

Perspective

The ISME Journal advance online publication 21 July 2017; doi: 10.1038/ismej.2017.119

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

OPEN

Benjamin J Callahan[1], Paul J McMurdie[2] and Susan P Holmes[3]

[1]Department of Population Health and Pathobiology, NC State University, Raleigh NC, USA
[2]Whole Biome Inc, San Francisco CA, USA
[3]Department of Statistics, Stanford University, Stanford CA, USA

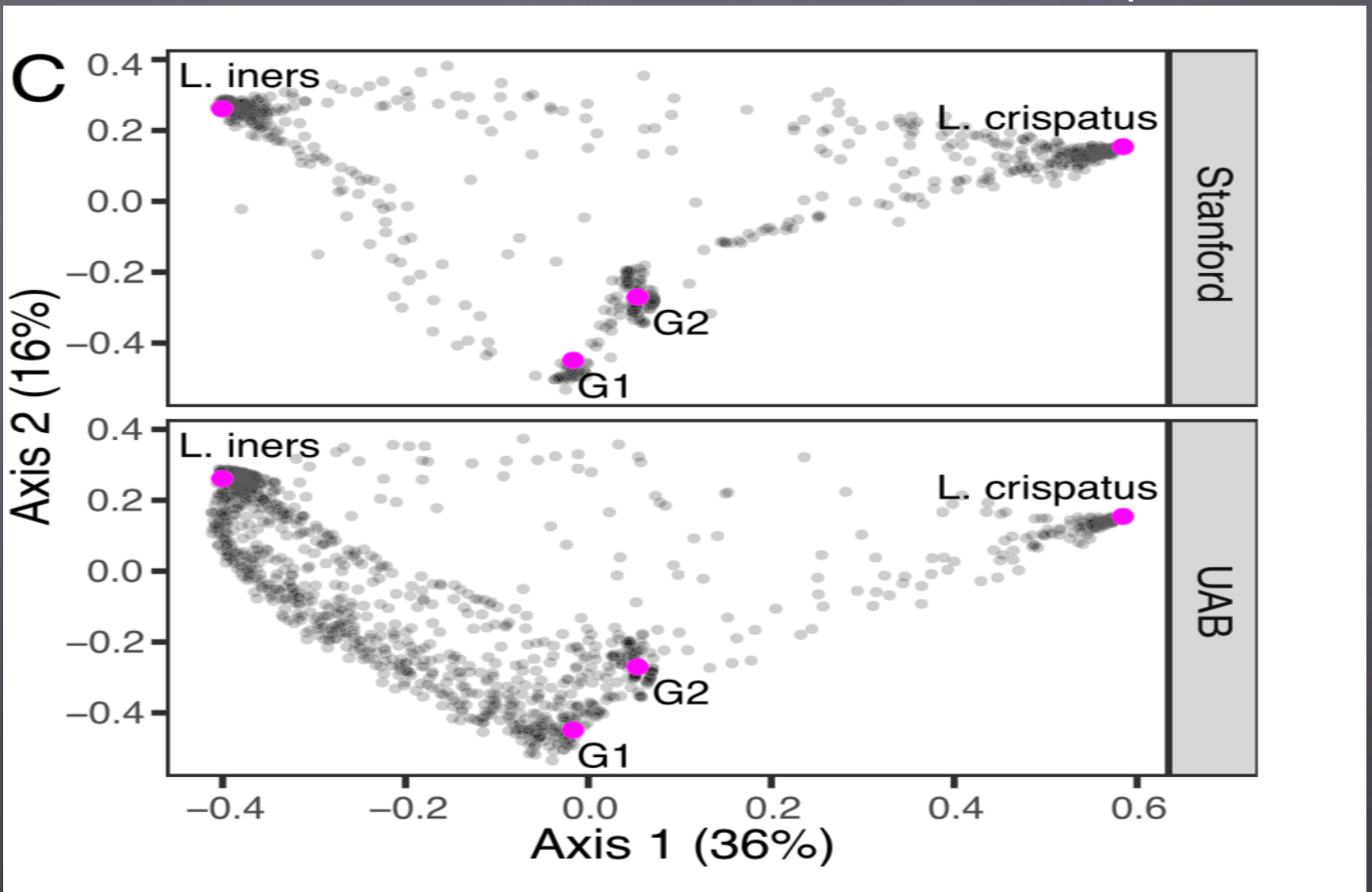
Correspondence: BJ Callahan, Department of Population Health and Pathobiology, NC State University, 1051 William Moore Drive, Raleigh, NC 27606, USA. E-mail: benjamin.j.callahan@gmail.com

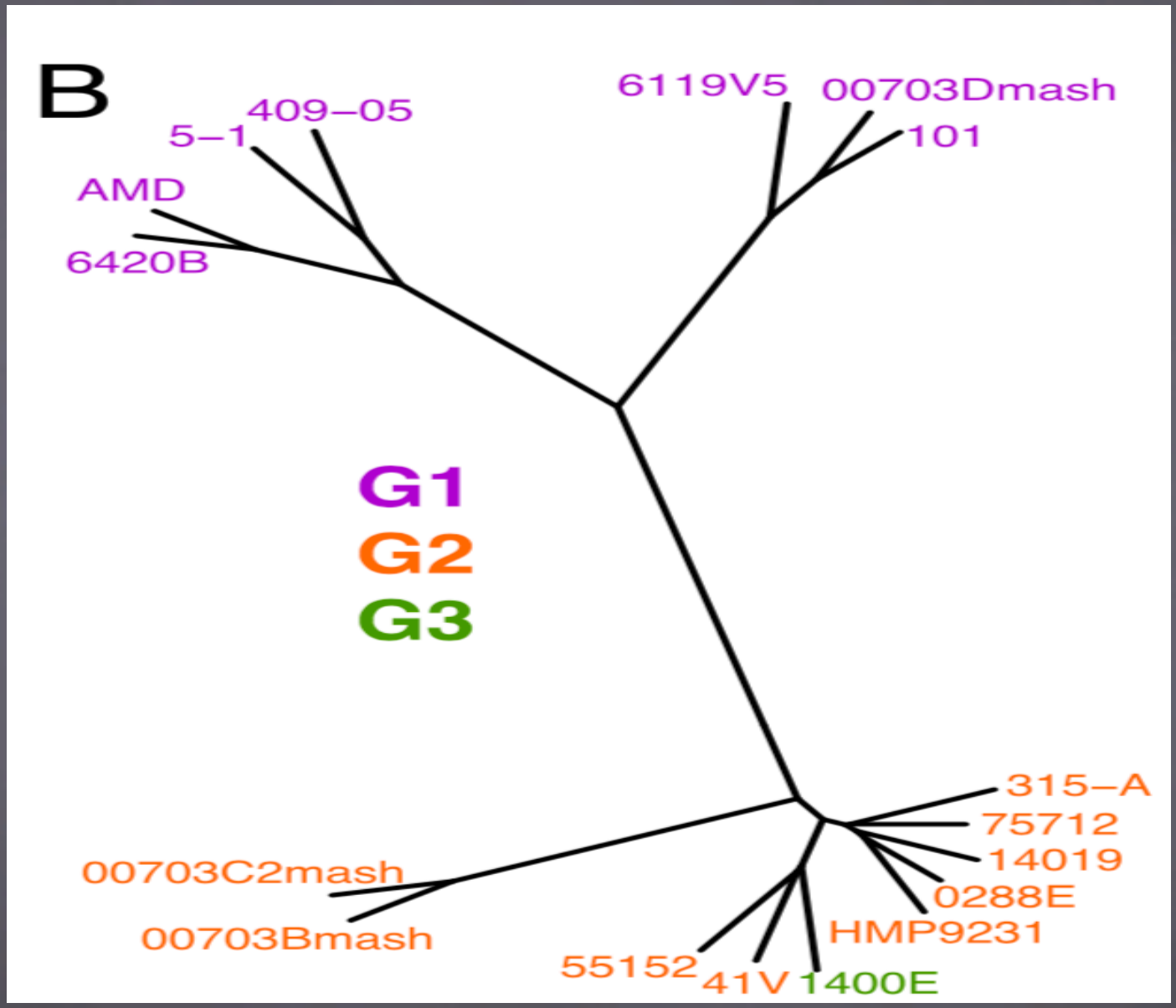
Received 21 March 2017; Revised 20 May 2017; Accepted 7 June 2017
Advance online publication 21 July 2017

but don’t rush to annotate everything.

Followup studies: Callahan et al., PNAS[?]

Study replication the study with a different cohort, showing that there were in fact several (3) strains of Gardnerella implicated.

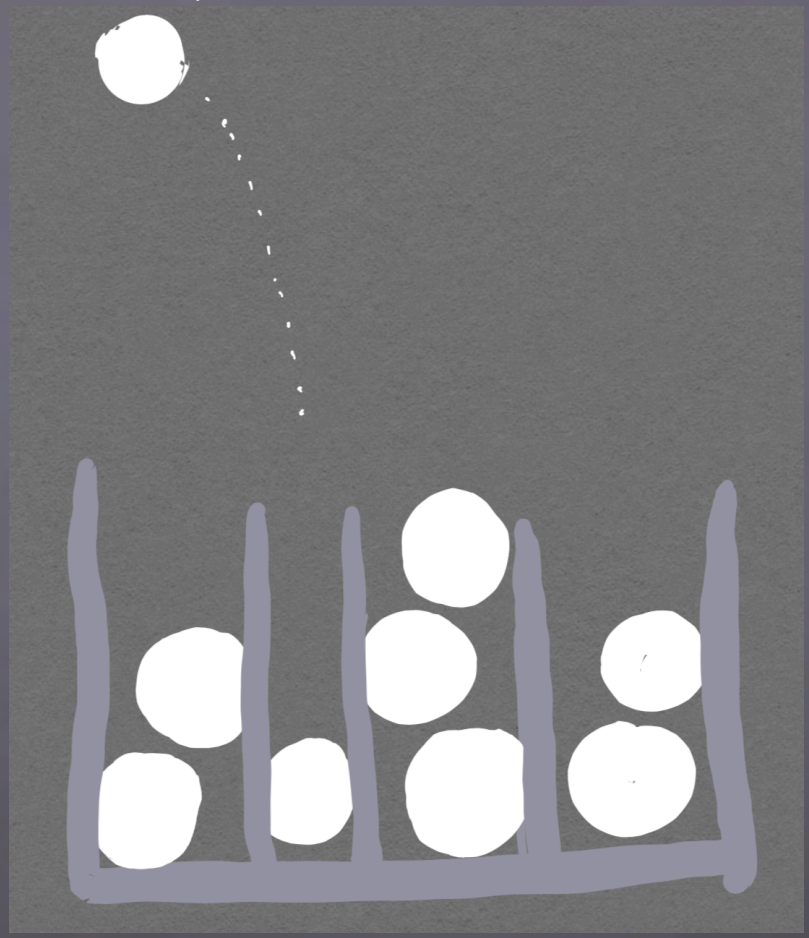




A complete genomic study of these Gardnerella strains to appear see Goltsman et al., 2018.

Part III

The Dirichlet for the multinomial



General Ideas about the multinomial

- Balls in boxes, not necessarily the same size.
- The number of balls is the number of reads, the boxes are the ASVs.
- Multinomial model gives the probability of seeing say (4,2,3,1) if the probabilities of the four boxes are $p_1 = 0.3, p_2 = 0.2, p_3 = 0.4, p_4 = 0.1$ this number is:

```
> dmultinom(c(4,2,3,1),prob=c(0.3,0.2,0.4,0.1))
[1] 0.02612736
```

- Apart from the fact that if a lot of balls fall in the first box there will be less balls for the other boxes, the boxes' contents are independent: that is BAD.

Dirichlet

Make the p's vary randomly.
Hierarchical Model:
$ps \sim Dirichlet(\alpha, \alpha, \alpha, \alpha)$
Uniform on the simplex (four cornered pyramid).

```
x <- round(gtools::rdirichlet(5, c(1,1,1,1) ),2)
> x
     [,1] [,2] [,3] [,4]
[1,] 0.06 0.50 0.08 0.36
[2,] 0.20 0.57 0.18 0.05
[3,] 0.07 0.20 0.55 0.18
[4,] 0.57 0.04 0.00 0.39
[5,] 0.02 0.16 0.27 0.55
```

Birthday Problem with Dirichlet

What k required for a $50 - 50$ chance of a match when n=365:

- Uniform Prior, c=1 $k \doteq .83\sqrt{n}$, for $n = 365$, $k \doteq 16$
- Symmetric Prior, $a_i = c$

c	.5	1	2	5	20	∞
k_c	13.2	16.2	18.7	20.9	21.9	22.9

frame Construct a 2 “hyper”parameter family of Dirichlet priors writing $\alpha_i = A\pi_i$, with $\pi_1 + \pi_2 \cdots + \pi_n = 1$. Assign weekdays parameter $\pi_i = \alpha$, weekends $\pi_i = \gamma\alpha$, with $260\alpha + 104\gamma\alpha = 1$. Here γ is the parameter ‘ratio of weekends to weekdays’, (roughly we said $\gamma \doteq .7$) and A measures the strength of prior conviction. The table below shows how k varies as a function of A and γ. We have assumed the year has $7 \times 52 = 364$ days.

A \ γ	.5	.7	1
1	2.2	2.2	2.2
364	16.1	16.3	16.4
728	18.4	18.6	18.8
∞	22.2	22.4	22.6

Multinomial needs to be modified

Multivariate dependencies in bacterial communities

Data depart from a multinomial distribution within each row:

- Some taxa are quasi-exclusive (*Lactobacillus crispatus* and *Gardnerella*).
- Co-occurrence through syntrophy, in which a molecular hydrogen-consuming species (typically a methanogen, like *Methanobrevibacter smithii* in the human gut) enhances the growth of a molecular hydrogen-producing species (any of a number of secondary fermenters in the gut).
- In the mouth (subgingival crevice), where in cases of moderate to severe periodontitis, a methanogen (Methanobrevibacter oralis) is always found with a syntrophic partner.
- There are not a finite number of taxa a priori, taxa evolve, some are sample-specific.

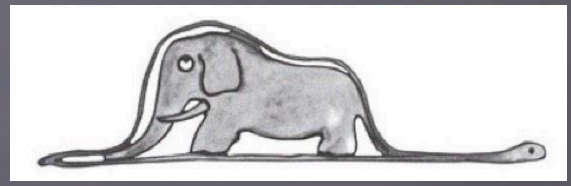
Part IV

Interpretability: latent variables and topic analysis

Discrete/disconnected Community state types are rare

Each sample is assigned to only one type of community.

Need a more nuanced model: mixtures.



Mixture models

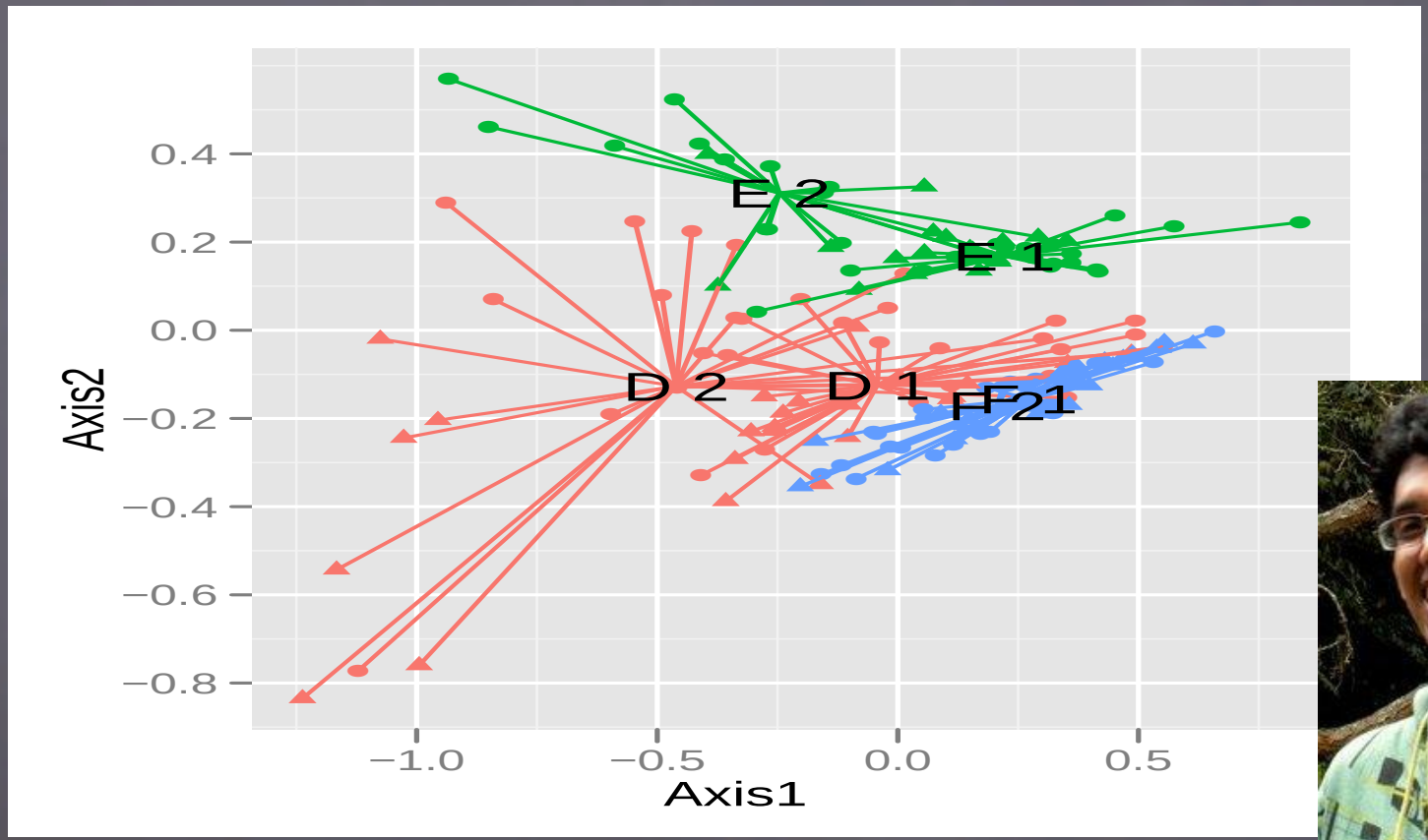
- In clustering and hidden discrete categorical variables, every sample belonged to a community state type.
- In a topic mixture model, every sample can be composed of several topics (**are these guilds?**).

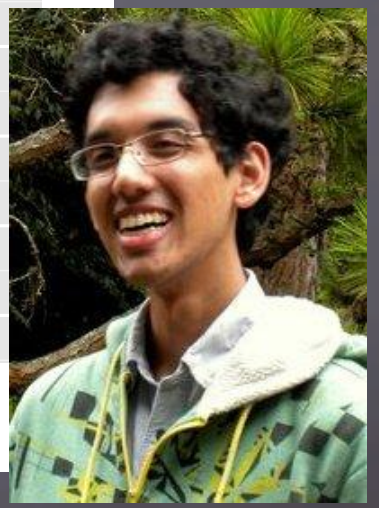
Most useful parallel: natural language processing.

Generative model

- Pick topics at random among a certain number of topics.
- Each topic corresponds to a probability distribution for many words.
- Pick a word at random according to the chosen topic/guild.

How to understand the the taxa involved in the perturbation?



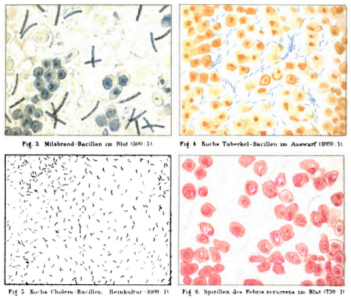


Biostatistics, 2018,
Latent Variable Modeling for the Microbiome.
Kris Sankaran's Topic Page

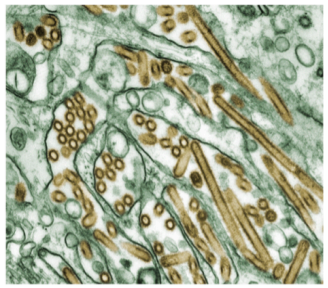
Kris Sankaran

Parallel between document topics and community analyses

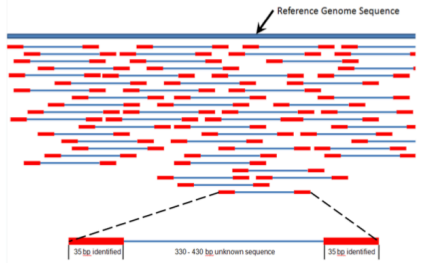
Sample



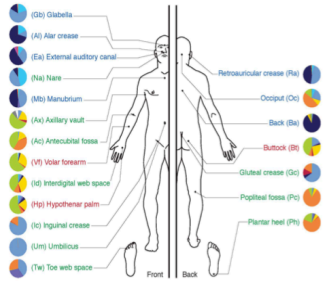
Species



Community

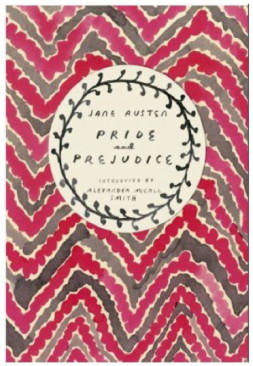


Reads

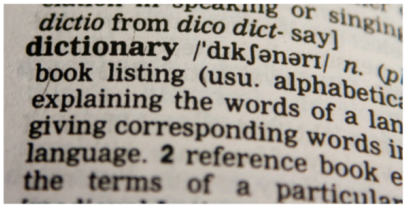


Environment

Document

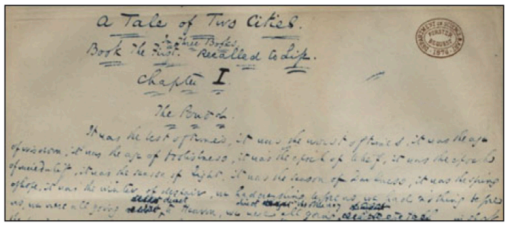


Term



Topic

Words



Corpus

Credit: Kris Sankaran

Parallel between topic and community analyses

index	book	elizabeth	darcy	bennet	miss	jane	bingley	time
0	P & P	0	0	4	0	1	3	0
1	P & P	1	0	5	0	1	4	0
2	P & P	0	0	6	0	0	5	1
3	P & P	1	4	5	1	0	9	1
4	P & P	3	3	5	4	4	5	3
5	P & P	3	0	0	2	1	6	1
6	P & P	0	6	6	7	1	5	1

time	subject	Unc06grq	Unc09fy6	Unc06bhm	Unc06g1h	Unc06af7
0	D	791	0	79	108	11
1	D	1616	0	1413	192	31
2	D	1323	0	915	165	23
3	D	1846	0	1366	170	31
4	D	2314	0	689	135	26
5	D	2244	0	776	310	175
6	D	1652	0	609	235	181

Statistical Model

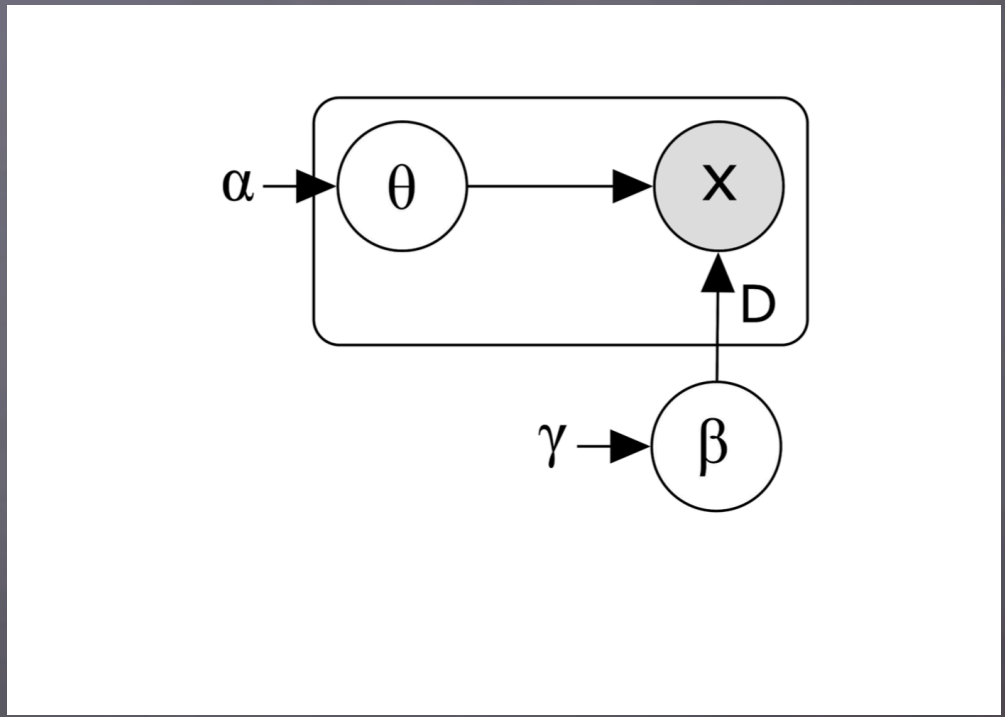
Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling.

It assumes samples have mixed memberships across topics.
(See Pritchard et. al 2000, Blei et. al. 2003)

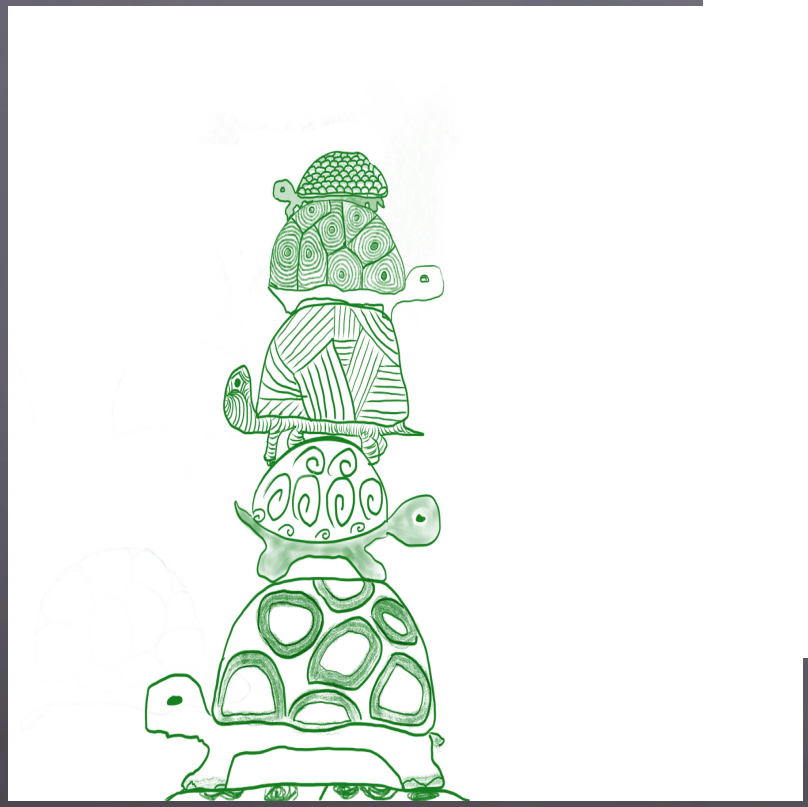
Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling.

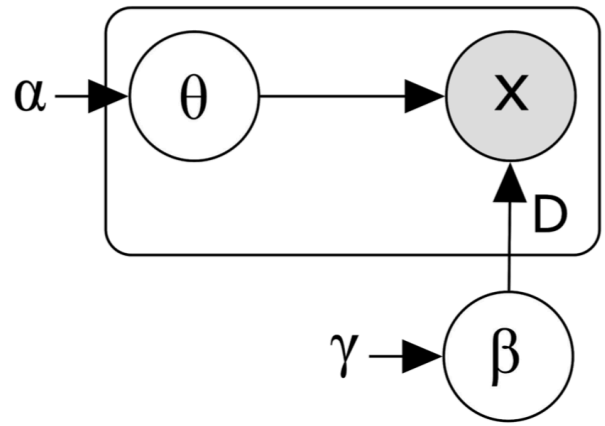
Observed microbiomes $\sim$ mixtures of underlying community types.

Statistical Model



Statistical Model





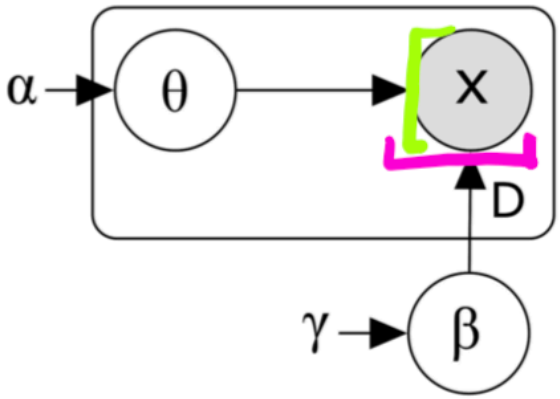
Statistical Model

samples
layer
observa.

sample 1 ⋯
⋮
sample n ⋯

rows (X)

hidden layer for taxa

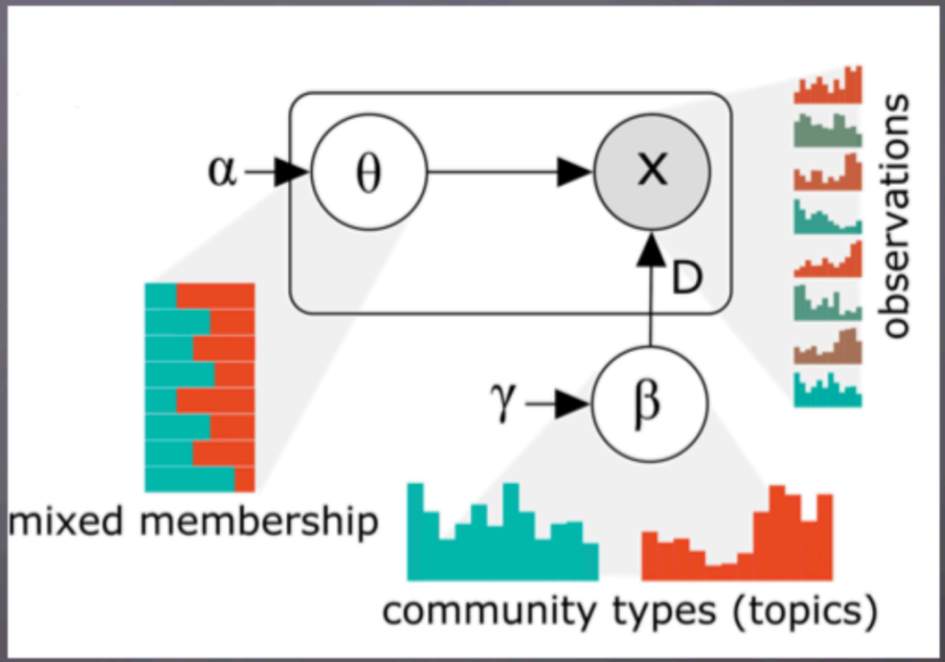


D Documents
K communities or topics

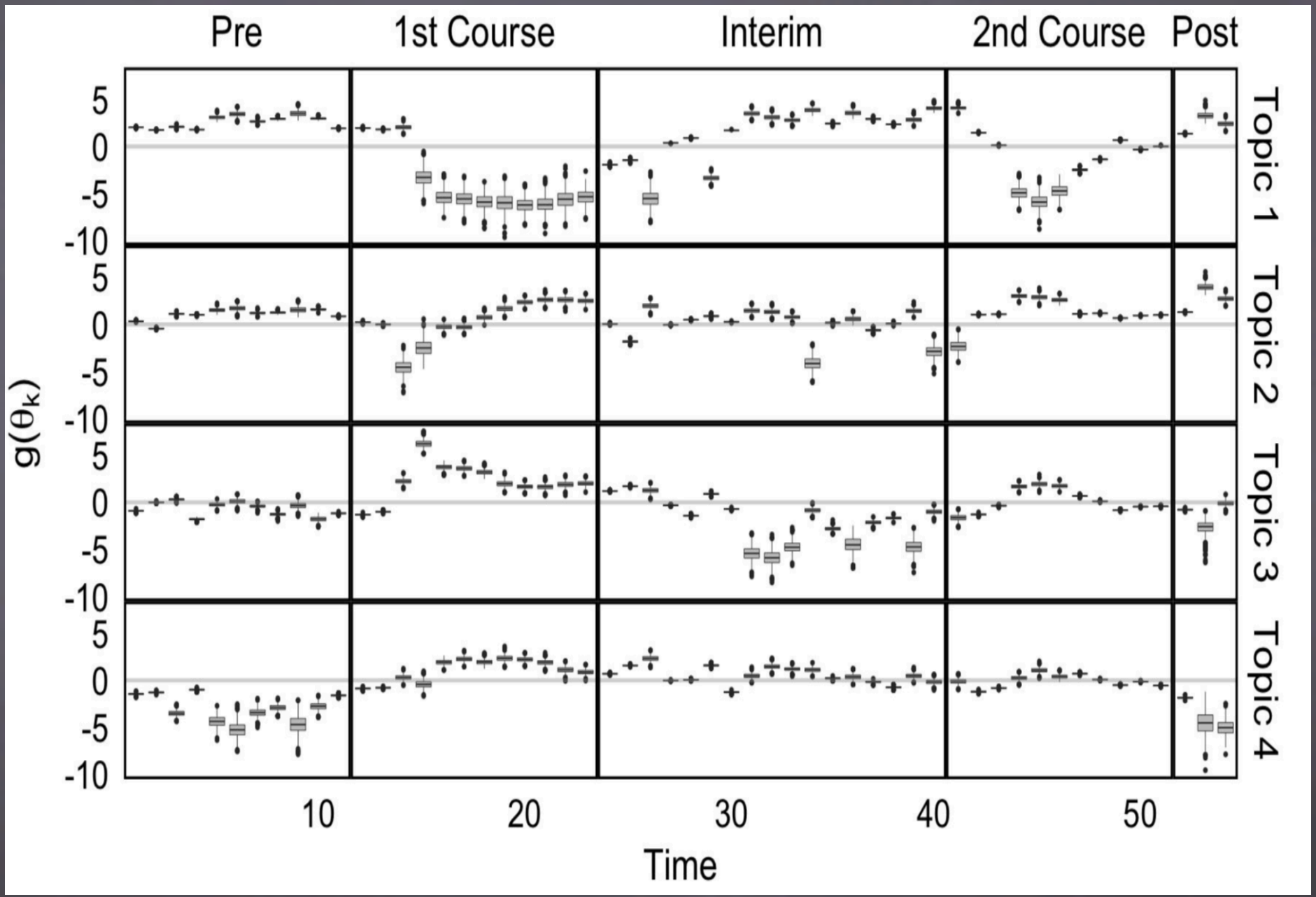
Statistical Model

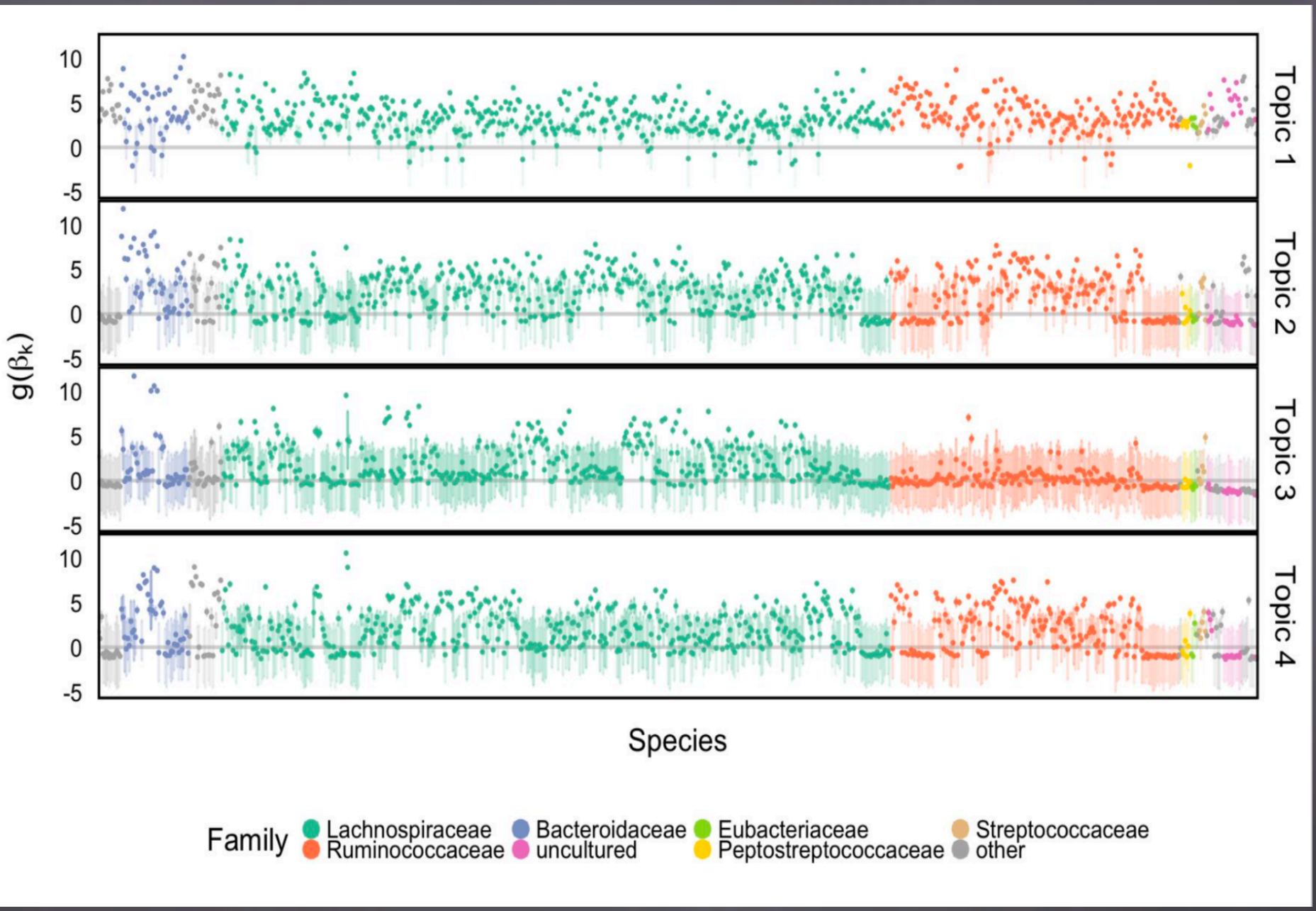
$$x_{d\cdot} \mid \beta \sim \text{Mult}(N_d, B\theta_d)$$

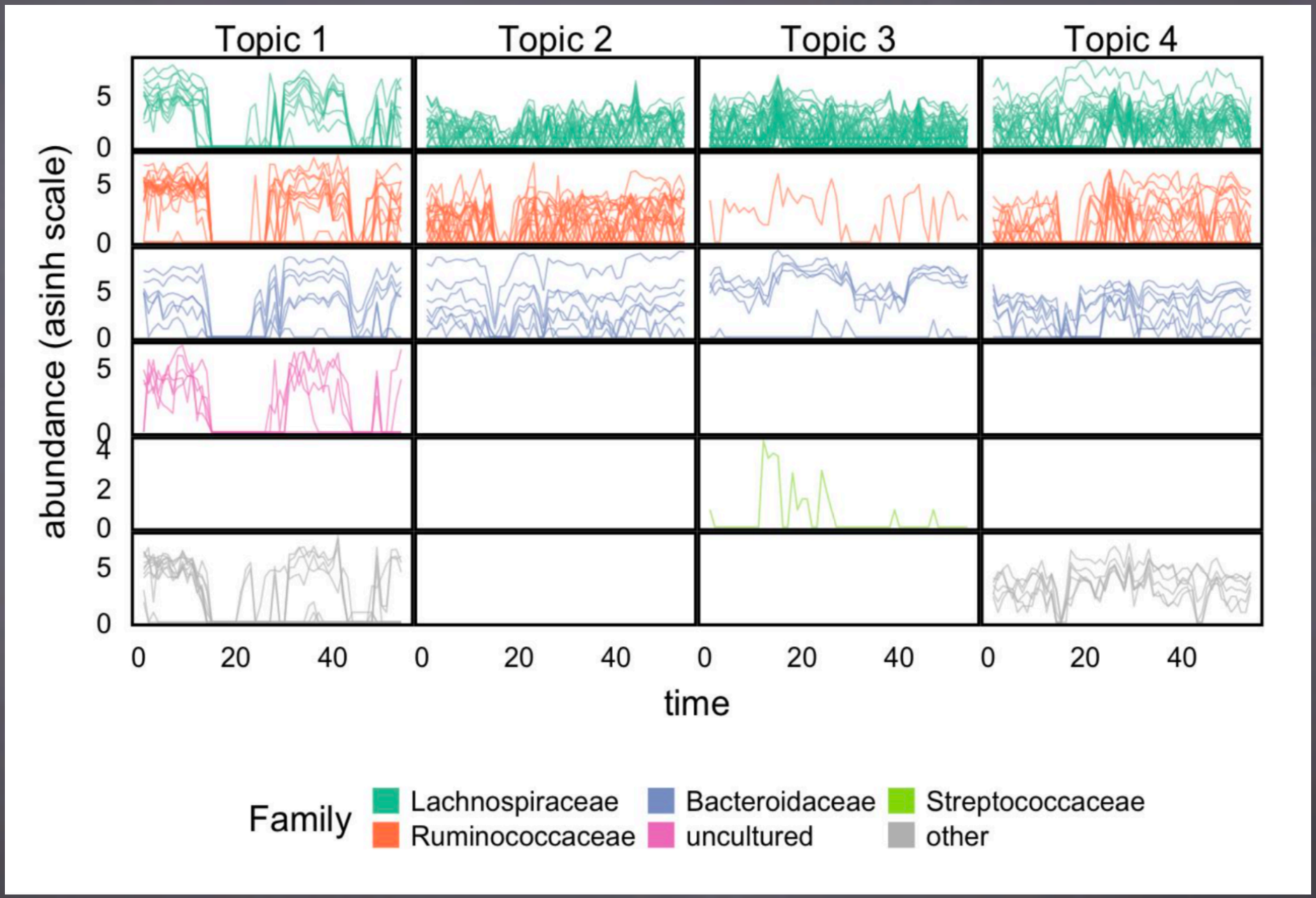
$$\theta_d \sim \text{Dir}(\alpha)$$
$$d = 1 \ldots D$$



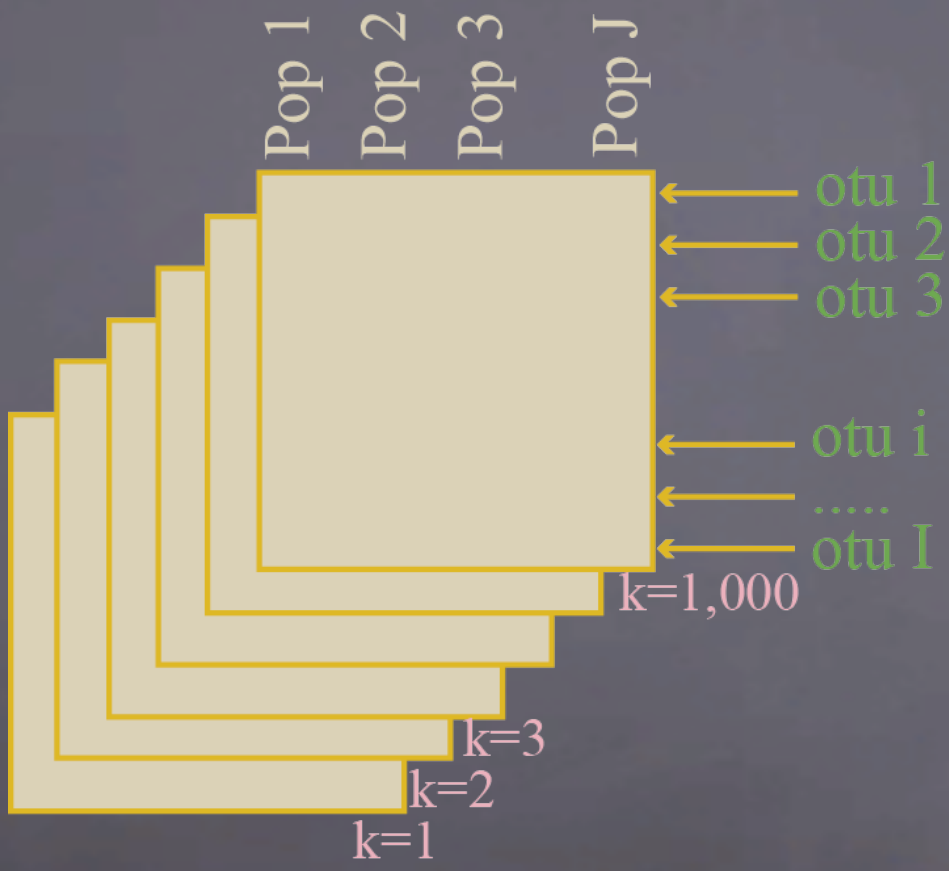
$$\beta_k \sim \text{Dir}(\gamma), \quad k = 1 \ldots K$$


Pre
1st Course
Interim
2nd Course
Post
Topic 1
Topic 2
Topic 3
Topic 4
5
0
-5
-10
10
20
30
40
50
Time
g(θk)


Topic 1
Topic 2
Topic 3
Topic 4
g(βk)
10
5
0
-5
Species
Family
Lachnospiraceae
Ruminococcaceae
Bacteroidaceae
uncultured
Eubacteriaceae
Peptostreptococcaceae
Streptococcaceae
other


Topic 1
Topic 2
Topic 3
Topic 4
abundance (asinh scale)
5
0
5
0
5
0
5
0
4
2
0
5
0
0
20
40
time
Family
Lachnospiraceae
Ruminococcaceae
Bacteroidaceae
uncultured
Streptococcaceae
other

Generalization: Bayesian posterior uncertainty measures



Parameters for samples $\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$

Define a joint prior on these factors through the Gram matrix $(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$

The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of ASVs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j},$$

$\epsilon_{i,j}$ $\quad$ iid Normal

Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren, Bacallado, Favaro, Holmes, Trippa (2017, JASA).

Parameters for samples

$$\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$$

Define a joint prior on these factors through the Gram matrix

$$(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$$

The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of OTUs.

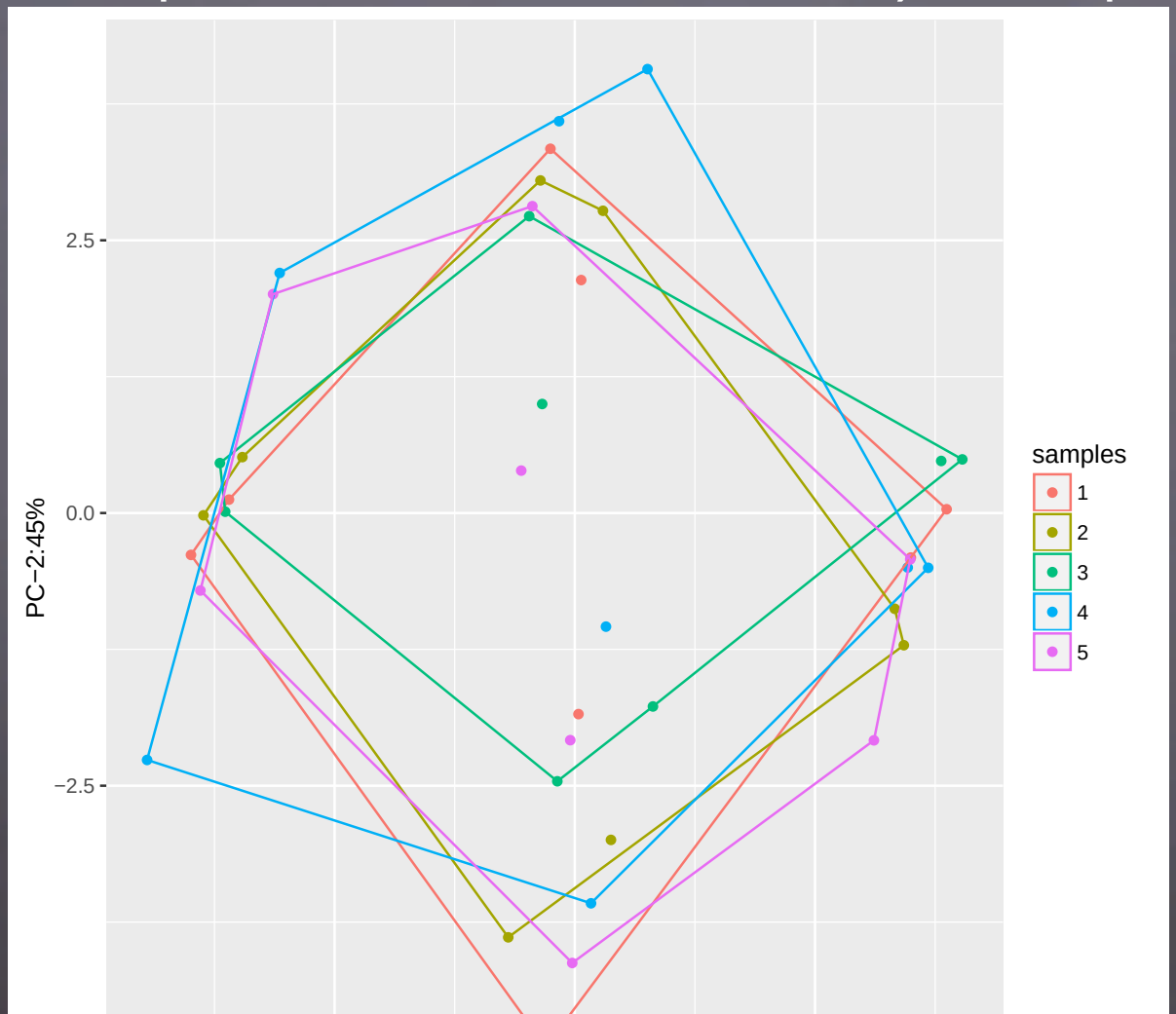
$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \tag{1}$$

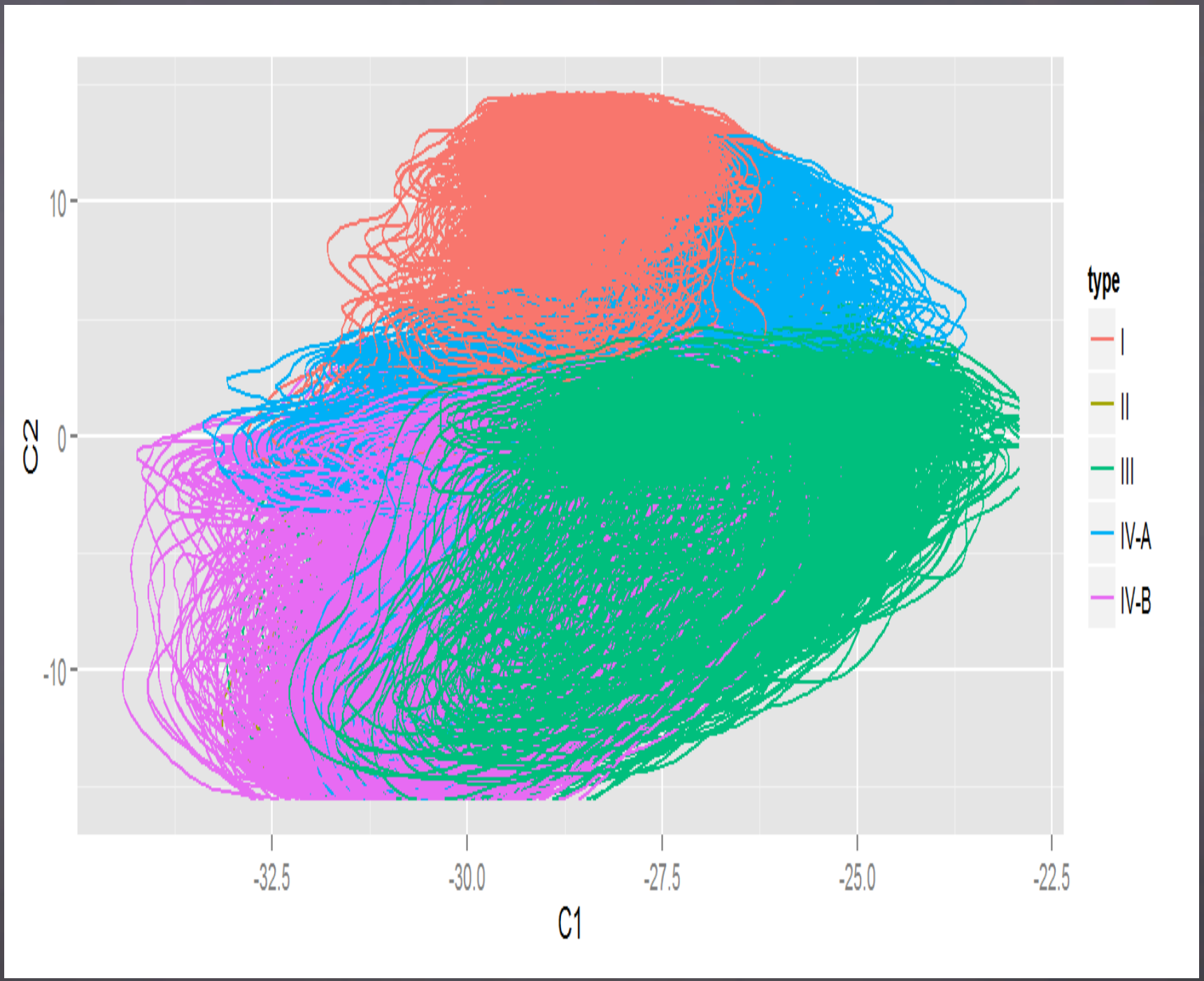
where the $\epsilon_{i,j}$ are independent Normal variables.

The methods that we consider here are all related to PCA and use the normalized Gram matrix **S** between biological samples. **S** is the correlation matrix of $(\mathrm{Q}_{\mathrm{i},1}, \ldots, \mathrm{Q}_{\mathrm{i},\mathrm{J}})$. Based on a single posterior instance of **S**, we can visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once.

A projection approach

Naively overlaying projections of the principal coordinate loadings generated from different posterior samples of **S** on the same plot *could* show the variability of the projections.




10
0
-10
C2
-32.5
-30.0
-27.5
-25.0
-22.5
C1
type
I
II
III
IV-A
IV-B

Registration: Find $\mathbf{S}_0$

Identify a Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_\mathrm{K}$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/\mathrm{K}$.
We prefer to choose $\mathbf{S}_0$, the Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_\mathrm{K}$.
We use the **RV** similarity metric between two symmetric square matrices $\mathbf{A}$ and $\mathbf{B}$

$$\mathrm{RV}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{AB})/\sqrt{\mathrm{Tr}(\mathbf{AA})\mathrm{Tr}(\mathbf{BB})}$$

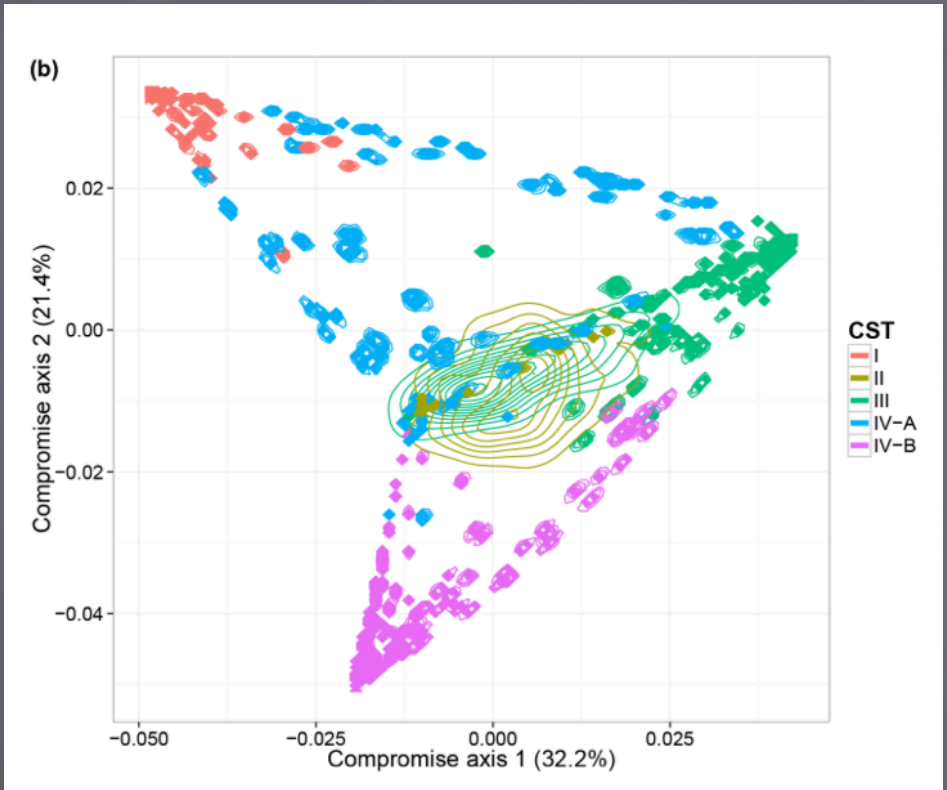
We diagonalize the **RV** matrix to obtain $\mathbf{S}_0$.

Find lower dimensional consensus space V

For dim 2, $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\mathbf{S}_0$ corresponding to the largest eigenvalues λ_1 and λ_2. All biological samples in V are visualized by projecting rows of $\mathbf{S}_0$ onto V: $(\psi_1^0, \psi_2^0) = \mathbf{S}_0(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$.

Project the rows of posterior sample $\mathbf{S}_k$ onto V by $(\psi_1^k, \psi_2^k) = \mathbf{S}_k(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$. Overlaying all the ψ^k displays uncertainty of $\mathbf{S}$ in the same linear subspace. Posterior variability of the biological samples' projections is visualized in V by plotting each row of the matrices (ψ_1^k, ψ_2^k), $k = 1, \ldots, K$, in the same figure.

We can see the uncertainties



Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren et al, 2017 (JASA).
A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space $\mathcal{V}$.

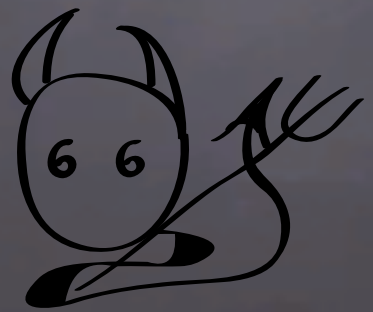
The Yoda of Silicon Valley

“premature optimization is the root of all evil in coding”

In Statistics

"premature summarization is the root of all evil in statistics"

In Statistics



R packages and resources

phyloseq: http://bioconductor.org/packages/stats/bioc/phyloseq/

dada2: http://bioconductor.org/packages/stats/bioc/dada2/

treelapse: https://krisrs1128.github.io/treelapse/

treelapse antibiotics http://statweb.stanford.edu/~kriss1/antibiotic.html

microbiome_pvlm: https://github.com/krisrs1128/microbiome_plvm

decontam: https://github.com/benjjneb/decontam/

adaptiveGPCA: https://cran.r-project.org/web/packages/adaptiveGPCA/index.html

bootLong: https://github.com/PratheepaJ/bootLong/blob/master/vignettes/Workflow.Rmd

Modern Statistics for Modern Biology

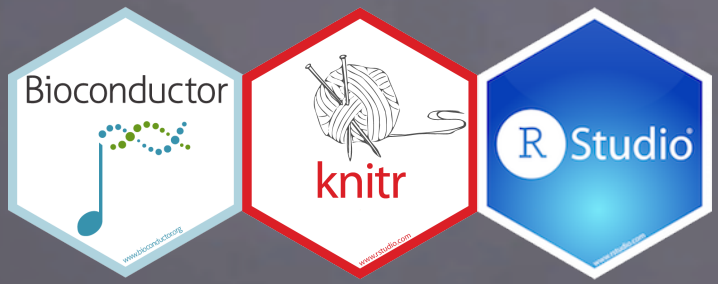
http://bios221.stanford.edu/book/

Solutions for microbiome analyses: respect the data.

- Poor data quality, information $\longrightarrow$ quality scores & probability.
- Maintain all information $\longrightarrow$ sequences are names.
- Interpretation $\longrightarrow$ latent variables (gradients or clusters).
- Reproducibility $\longrightarrow$ complete code source.
- Heterogeneity $\longrightarrow$ multicomponent objects:phyloseq.
- Training and collaboration $\longrightarrow$ Rmd and html.

Benefitting from the tools and schools of Statisticians.....

Thanks to the `R` and `Bioconductor` community and to co-authors.



Wolfgang Huber, Martin Morgan, Joey McMurdie, Ben Callahan, JJ Allaire and Rob Gentleman.

Thank you to the organizers for inviting me.

Lab Group and David Relman

Postdoctoral Fellows Paul (Joey) McMurdie, Ben Callahan, Christof Seiler, Pratheepa Jeganathan. Students: John Cherian, Diana Proctor, Daniel Sprockett, Lan Huong Nguyen, Julia Fukuyama, Kris Sankaran, Claire Donnat. Funding from NIH TR01 and NSF-DMS.

phyloseq

Joey McMurdie (joey711 on github).
Available in Bioconductor.
How can I (my students, my postdocs...) learn more?
Ask me.
http://www-stat.stanford.edu/~susan/

[1] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.M. Batto, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[2] Benjamin Callahan, Daniel DiGiulio, Daniela Goltsman, Christine Sun, Elizabeth Costello, Pratheepa Jeganathan, Joseph Biggio, Ronald Wong, Maurice Druzin, Gary Shaw, et al. Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. *PNAS*, page 201705899, 2017.

[3] Benjamin J. Callahan, Paul J. McMurdie, and Susan P. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, (10.1038/ismej.2017.119):1–5, 2017.

[4] Daniel Chessel, Anne Dufour, and Jean Thioulouse. The ade4 package - i: One-table methods. *R News*, 4(1):5–10, 2004.

[5] P. Diaconis, S. Goel, and S. Holmes. Horseshoes in

multidimensional scaling and kernel methods. *Annals of Applied Statistics*, 2007.

[6] Susan Holmes. Multivariate analysis: The French way. In D. Nolan and T. P. Speed, editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 56 of *IMS Lecture Notes–Monograph Series*. IMS, Beachwood, OH, 2006.

[7] S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26(11):1463–1464, 2010.

[8] K. Mardia, J. Kent, and J. Bibby. *Multiariate Analysis*. Academic Press, NY., 1979.

[9] P. J. McMurdie and S. Holmes. Phyloseq: A bioconductor package for handling and analysis of high-throughput phylogenetic sequence data.

[10] P. J. McMurdie and S. Holmes. Phyloseq: Reproduible research platform for bacterial census data. *PlosONE*, 2013. April 22,.

[11] P. J. McMurdie and S. Holmes. Waste not, want not: Why rarefying microbiome data is inadmissible. *Plos Computational Biology*, 2014. April 03.

[12] C. R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya A*, 26:329–359., 1964.