Statistical Challenges in the Analyses of the Human Microbiome.

Susan Holmes
http://webstat.stanford.edu/˜susan/
@SherlockpHolmes
CASBS fellow, 2017-2018

Bio-X and Statistics, Stanford University

Bayes on the Beach, November 13th, 2017

Challenges when working on microbiome analyses.

- Heterogeneity.
- Poor data quality, information leakage.
- Tree and graph integration, uncertainty visualizations.
- Propagation of uncertainty.

Part I

Heterogeneity

‘Homogeneous data are all alike;

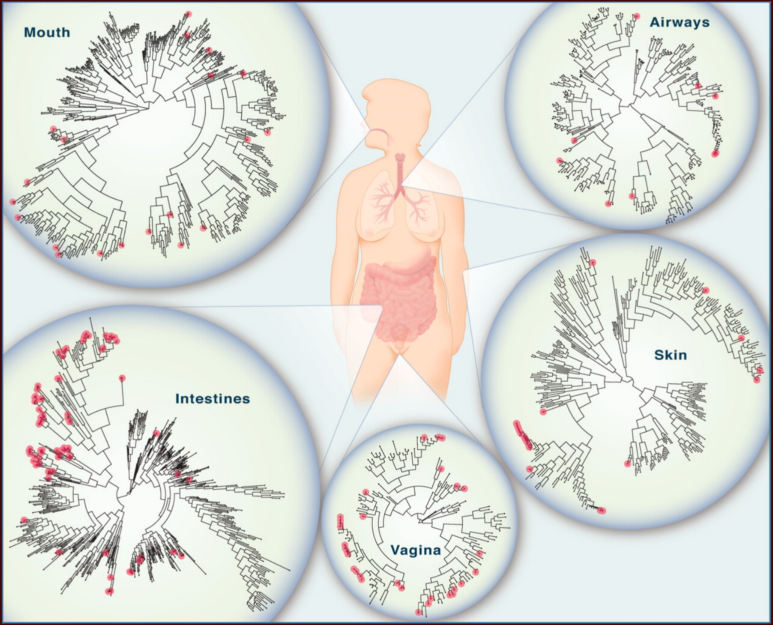
all heterogeneous data are heterogeneous

in their own way.’

Heterogeneity of Data

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

•



YK Lee and SK Mazmanian Science, 2010.

Human Microbiome: What are the data?

DNA The Genetic ‘signature‘ of the bacteria (16S rRNA-gene).

DNA Shotgun collection of genes present (metagenomes).

RNA What genes are being turned on (gene expression).

Proteomics Specific signatures of chemical compounds present.

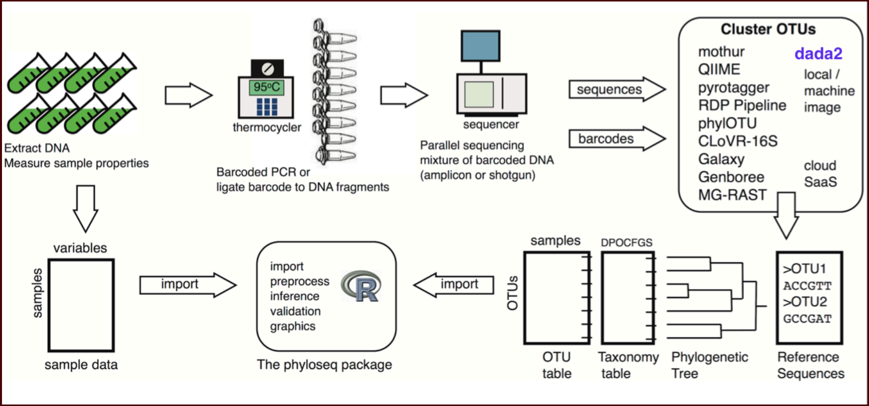
Clinical Multivariate information about patients’ clinical status, medication, weight.

Environmental Location, nutrition, time.

Domain Knowledge Metabolic networks, phylogenetic trees, gene ontologies.

Everything is data....

...... no metadata.

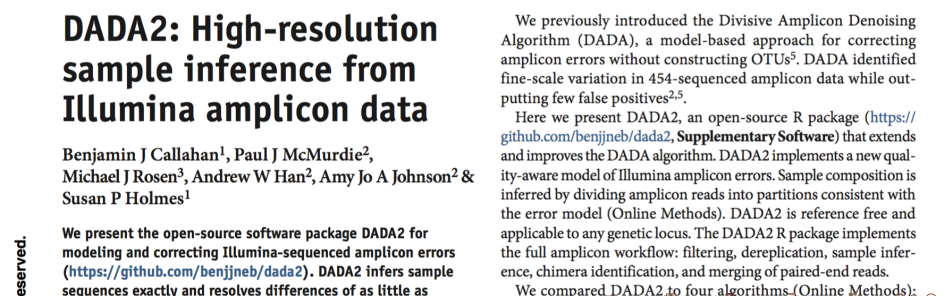


Heterogeneous Data Objects: S4 classes

Input and data manipulation with `phyloseq`
(McMurdie and Holmes, 2013, Plos ONE).

Part II

Improving data quality using probabilitistic denoising



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]

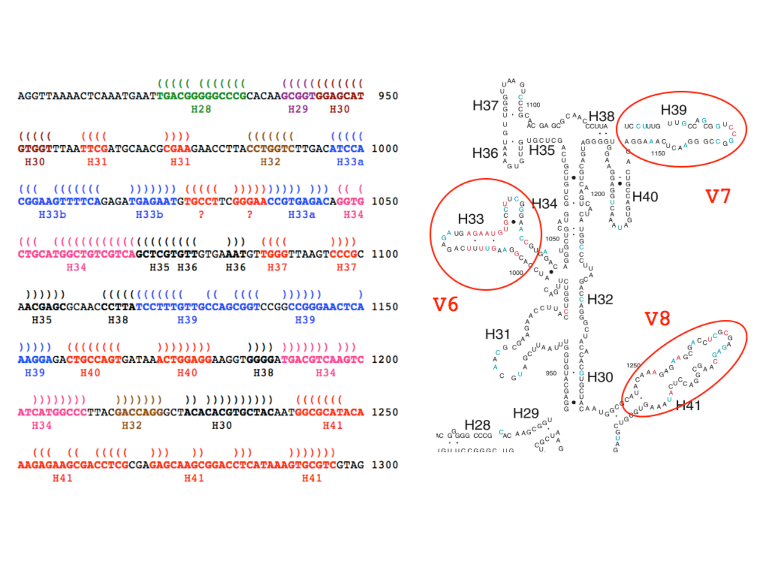
We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjjneb/dada2). DADA2 infers sample sequences exactly and resolves differences of as little as

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):

16S rRNA gene for bacterial fingerprinting



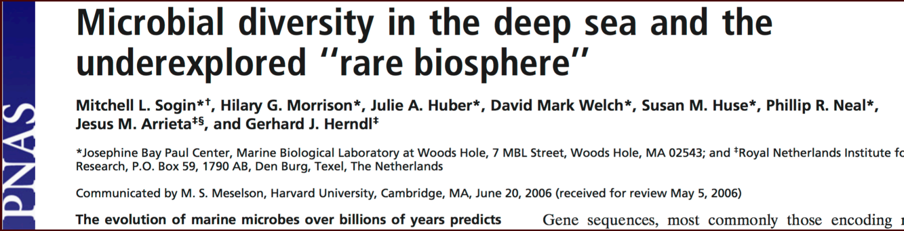
```
                    ((((( (((((((     ((((((((((((
AGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCAT  950
                         H28           H29   H30

(((((    ((((        ))))        (((((((     (((((
GTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTGGTCTTGACATCCA 1000
 H30      H31         H31          H32        H33a

(((  (((((((    ))))))) (((((  )))))))))) ))) (( (
CGGAAGTTTTCAGAGATGAGAATGTGCCTTCGGGAACCGTGAGACAGGTG 1050
   H33b          H33b     ?      ?     H33a    H34

(((  (((((((((((((((((((((    )))  ((((      ))))
CTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGC 1100
       H34         H35 H36    H36   H37       H37

 ))))))     ((((((((((((   ((  ((((   ))))))     )
AACGAGCGCAACCCTTATCCTTTGTTGCCAGCGGTCCGGCCGGGAACTCA 1150
  H35        H38       H39              H39

)))))  ((((((((     ))))))))     )))) ))))))) )))
AAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGACGTCAAGTC 1200
 H39     H40           H40        H38      H34

  ))))))))    )))))))   )) ))))))))))   (((((((
ATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCGCATACA 1250
  H34          H32          H30             H41

  (((  ((  (((((   )))   ))    )))     )))))))
AAGAGAAGCGACCTCGCGAGAGCAAGCGGACCTCATAAAGTGCGTCGTAG 1300
     H41                H41              H41
```

Diversities

- 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:



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 fo
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)

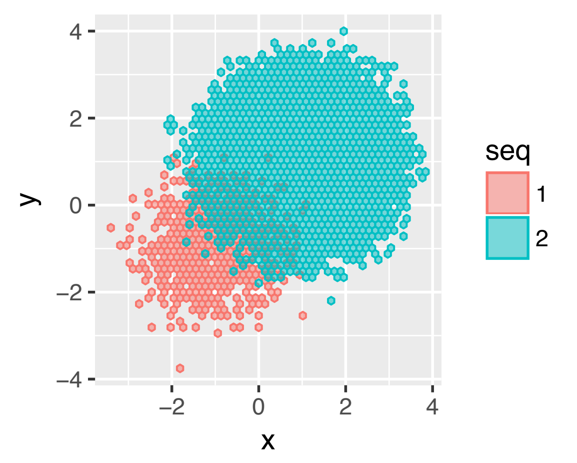
The evolution of marine microbes over billions of years predicts

Gene sequences, most commonly those encoding

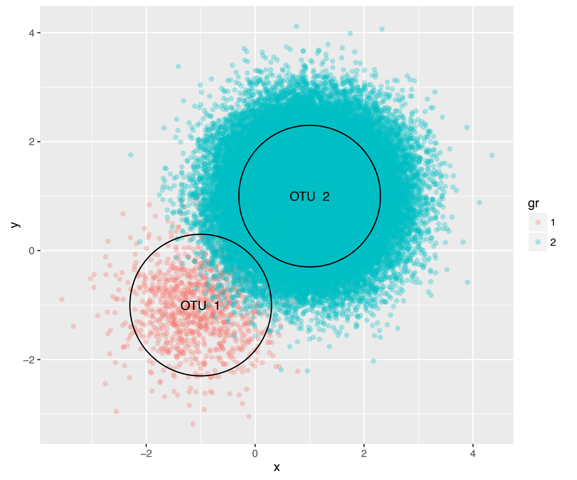
How many words does P. know?

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

From reads to Operational Taxonomic Units



From reads to Amplicon Sequencing Variances



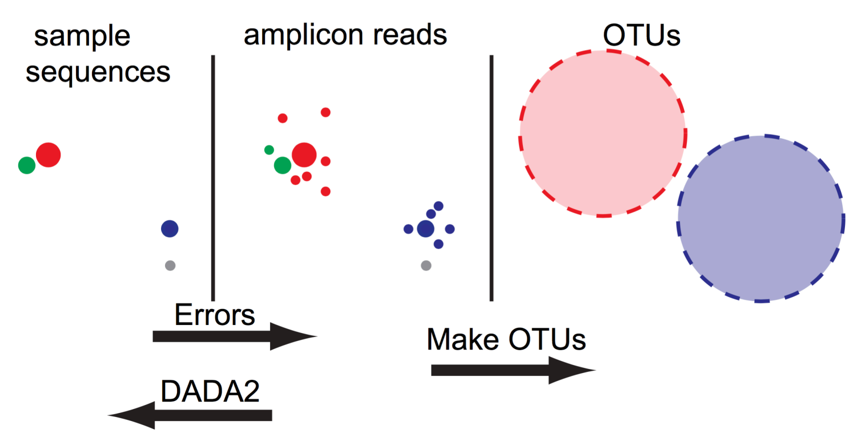
Curent practice (`qiime, mothur, rdp,...`): 97% similarity.

Problems involved in going from reads to ‘species’

Standard method: cluster within 97% similarity.

- Low resolution: 97% gives genus level at best
- High false positive rate: #(OTUs) $>>$ richness.
- Big data scaling: time scales super-linearly

Probabilistic Model



Error Model

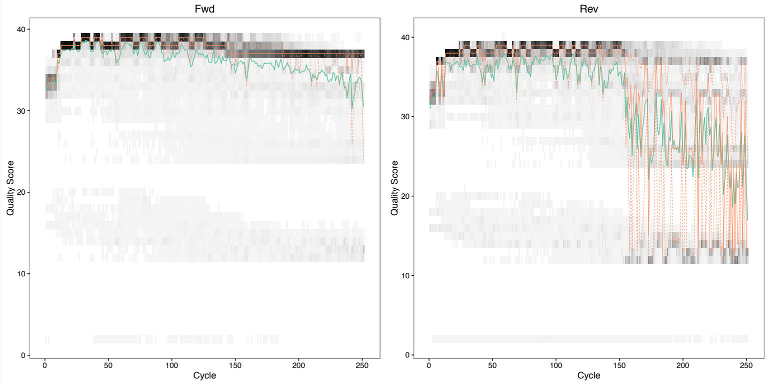
```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |              |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

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

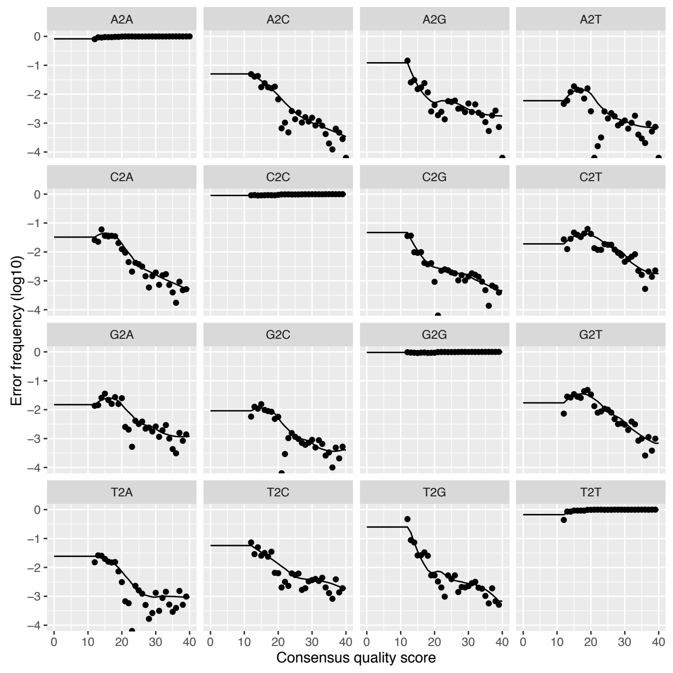
P probabilities of substitutions ($A- > C$)

q Quality score (Q=30)

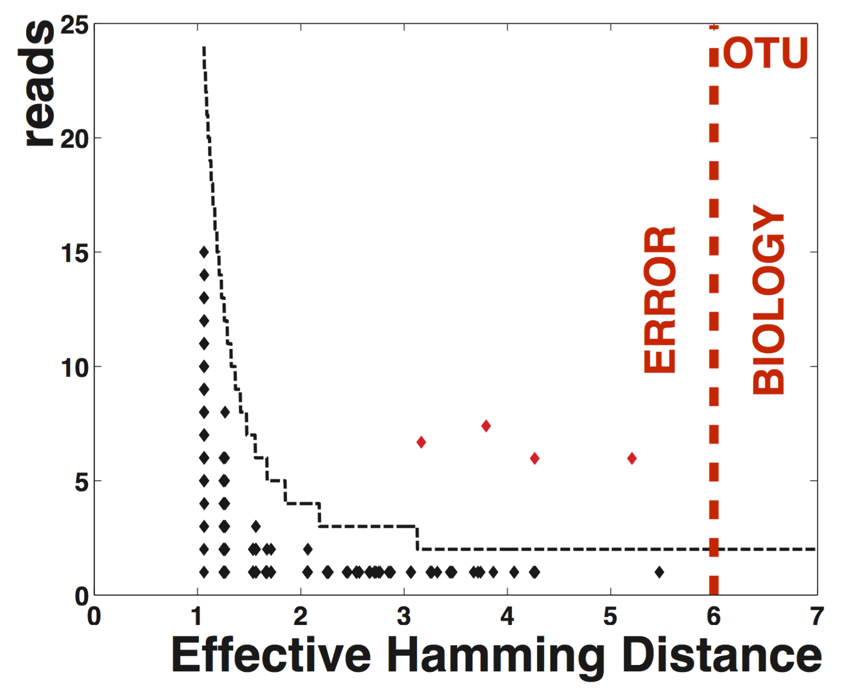
Batch effect (run)

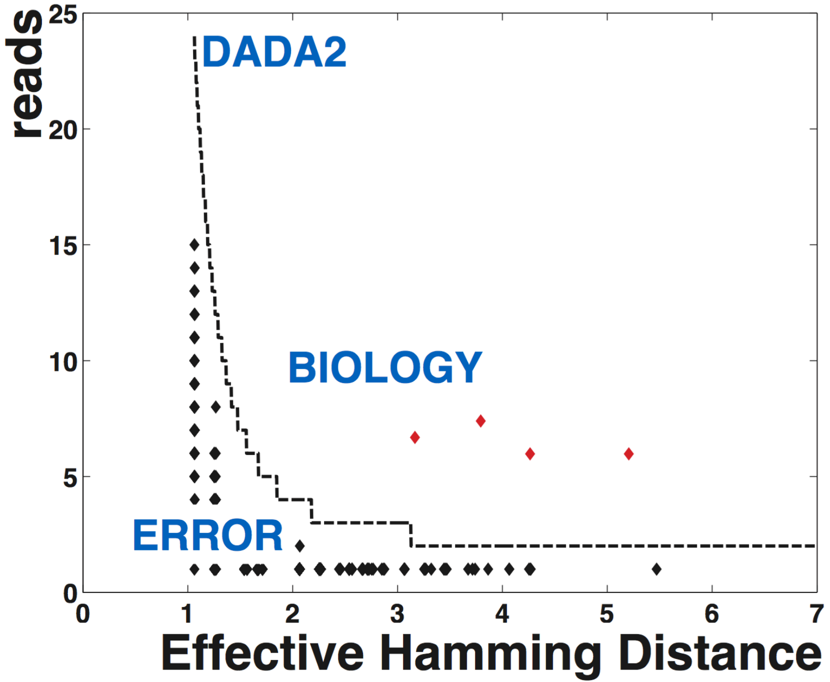


Forward and Reverse quality profiles along the reads.

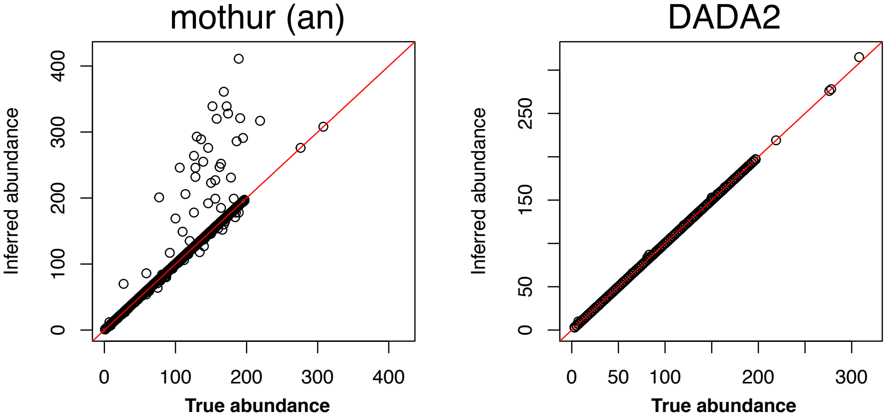


Frequencies of each type of nucleotide transition as a function of quality


reads
25
20
15
10
5
0
0
1
2
3
4
5
6
7
Effective Hamming Distance
OTU
ERROR
BIOLOGY


reads
25
20
15
10
5
0
DADA2
BIOLOGY
ERROR
0
1
2
3
4
5
6
7
Effective Hamming Distance

Accuracy: Simulated data

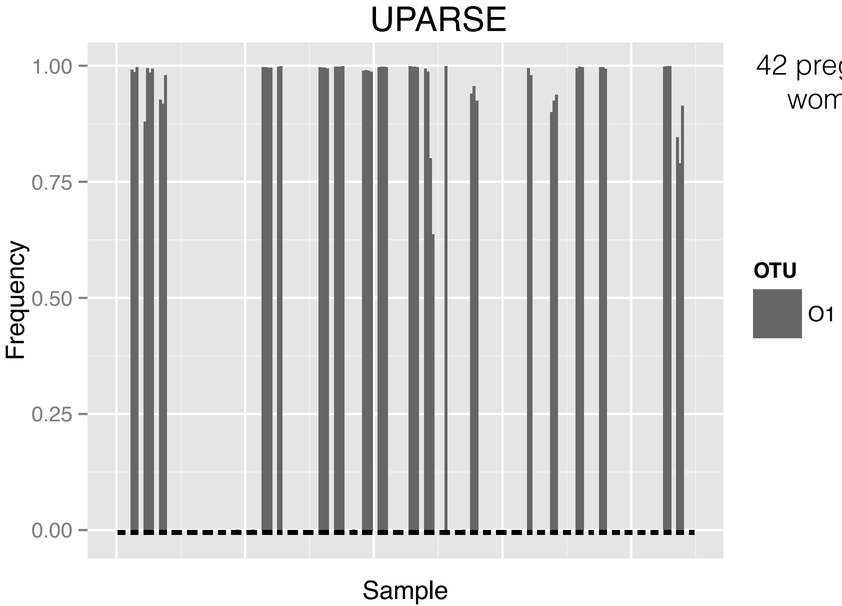


TP:	978
FP:	272
FN:	77
cor:	0.935

TP:	1042
FP:	0
FN:	13
cor:	0.999

Data: Kopylova, et al. mSystems, 2016.

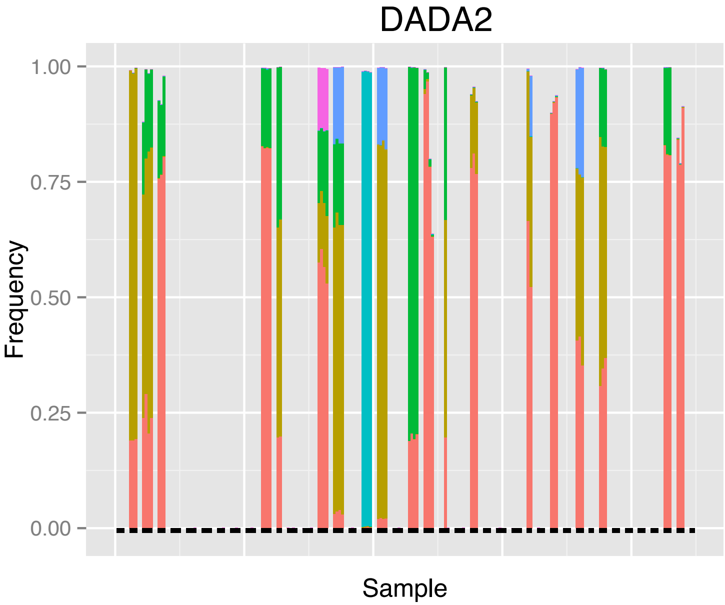
Resolution: L. crispatus



42 pregnant women

Data: MacIntyre et al. Scientific Reports, 2015.

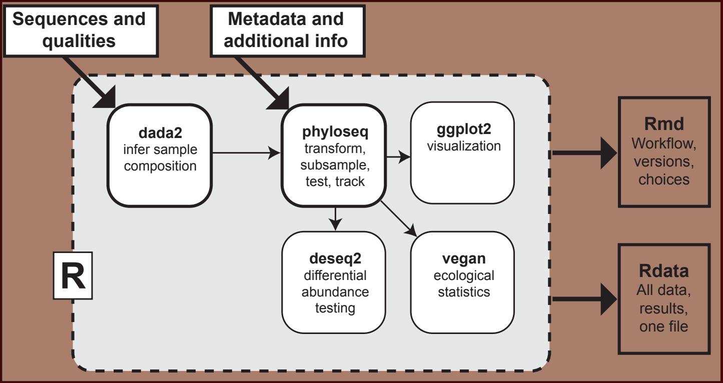
Resolution: L. crispatus



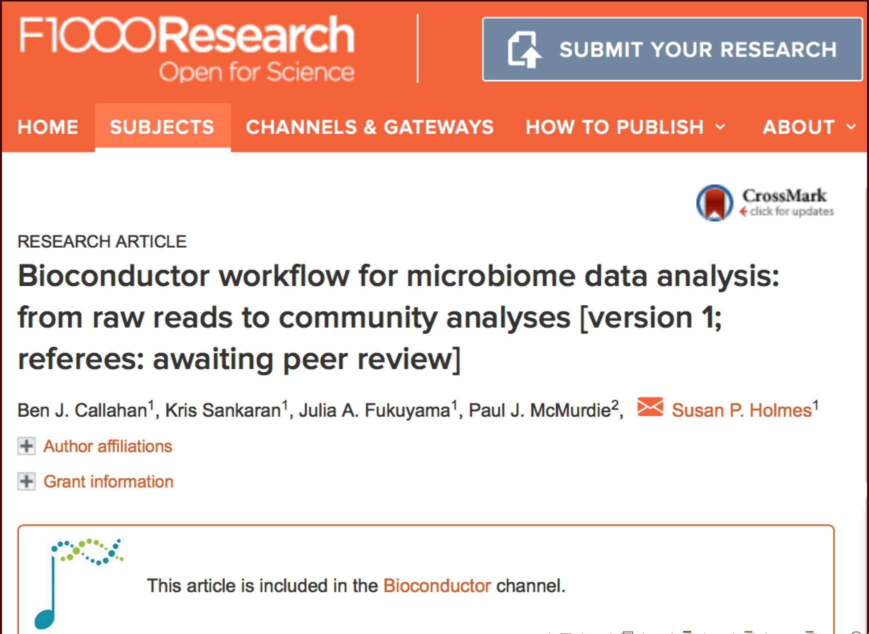
42 pregnant women

Data: MacIntyre et al. Scientific Reports, 2015.

Reproducible Research Workflow


Sequences and qualities
Metadata and additional info
dada2
infer sample composition
phyloseq
transform, subsample, test, track
ggplot2
visualization
deseq2
differential abundance testing
vegan
ecological statistics
R
Rmd
Workflow, versions, choices
Rdata
All data, results, one file

See complete workflow on Bioconductor channel of F1000:
http://f1000research.com/articles/5-1492/v1



RESEARCH ARTICLE

Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 1; referees: awaiting peer review]

Ben J. Callahan[1], Kris Sankaran[1], Julia A. Fukuyama[1], Paul J. McMurdie[2], Susan P. Holmes[1]

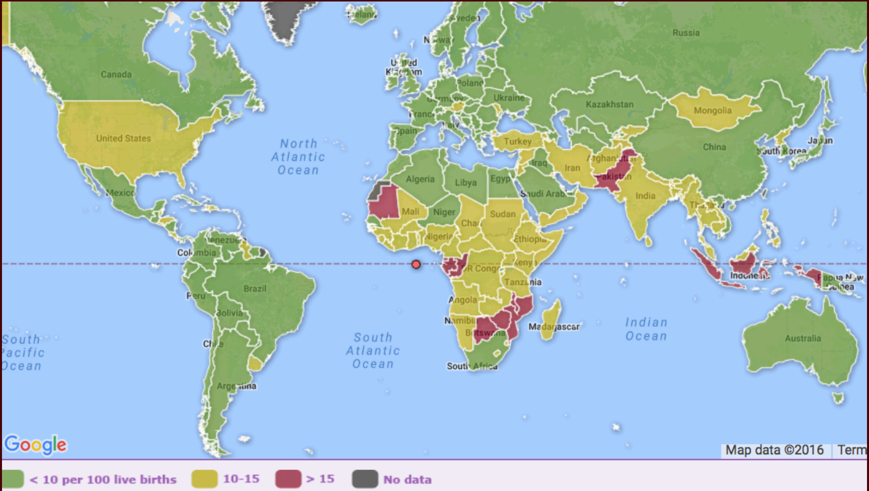
+ Author affiliations

+ Grant information

This article is included in the Bioconductor channel.

Part III

Preterm Birth Prediction



Pregnancy data: perturbation, stability and preterm

Study 1: 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.

Attention to detail

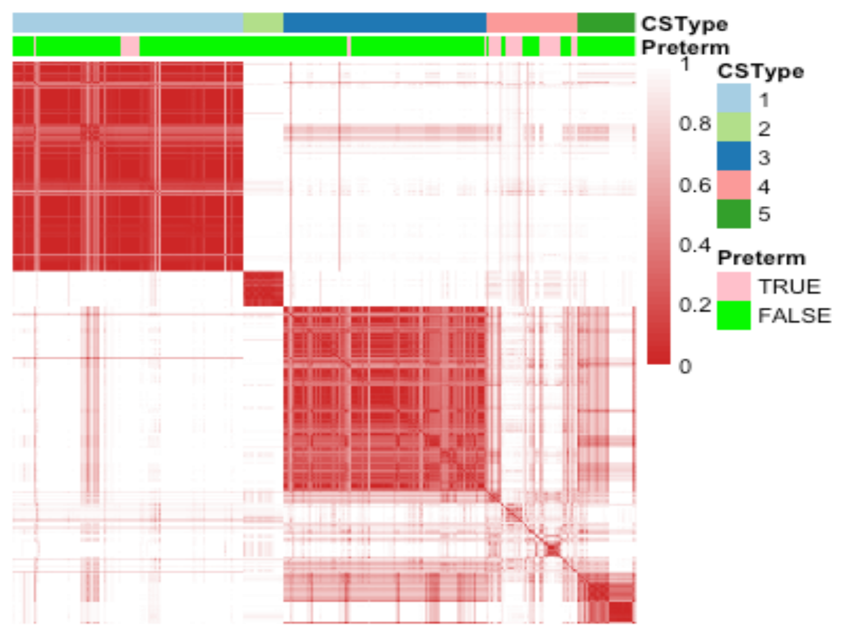
- Careful probabilistic noise model (`dada2`) and variance stabilization (`arcsinh`).
- Random effects, mixed models.
- Finite State Markov chains.
- Differential abundance testing provides biomarkers for preterm birth.

This work involves data collected by David Relman's group.
Statistical analyses done jointly with Ben Callahan.

Questions asked?

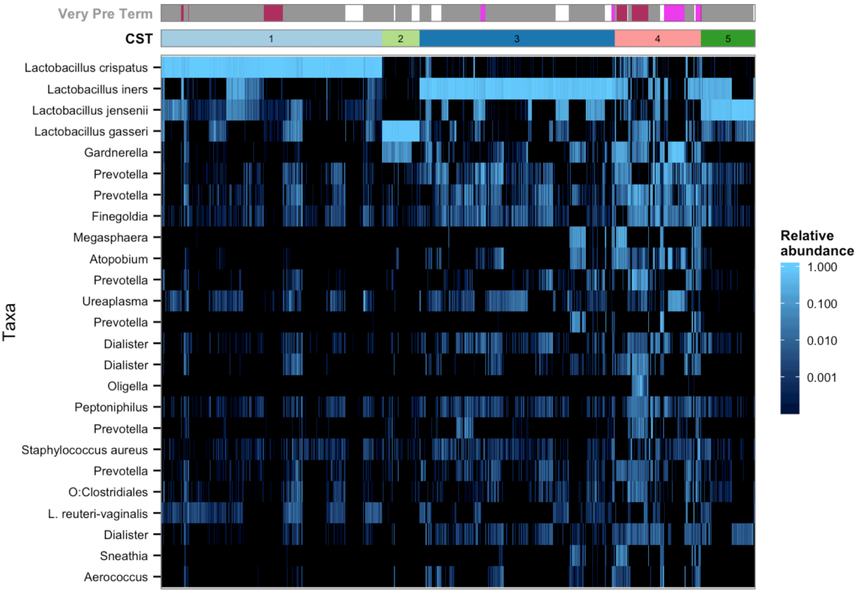
- Are the community state types the same as seen in previous studies?
- How stable are the communities within each individual during pregnancy?
- What alterations of the vaginal microbiome predict preterm birth?
- How early do these alterations occur?
- What changes in the microbiome occur at delivery?

Communities of bacteria organized into 5 different types



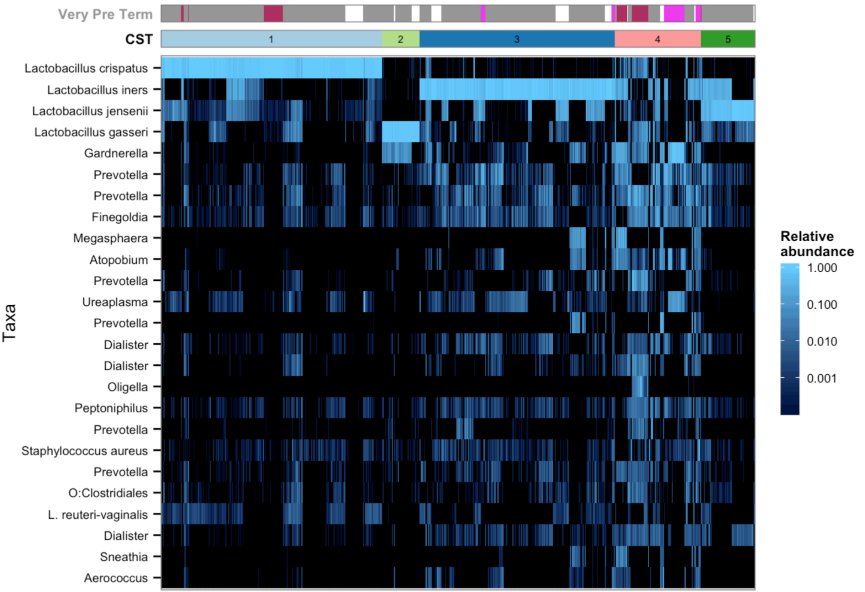
Previously known Community State Types

Checked clustering of samples into community types.

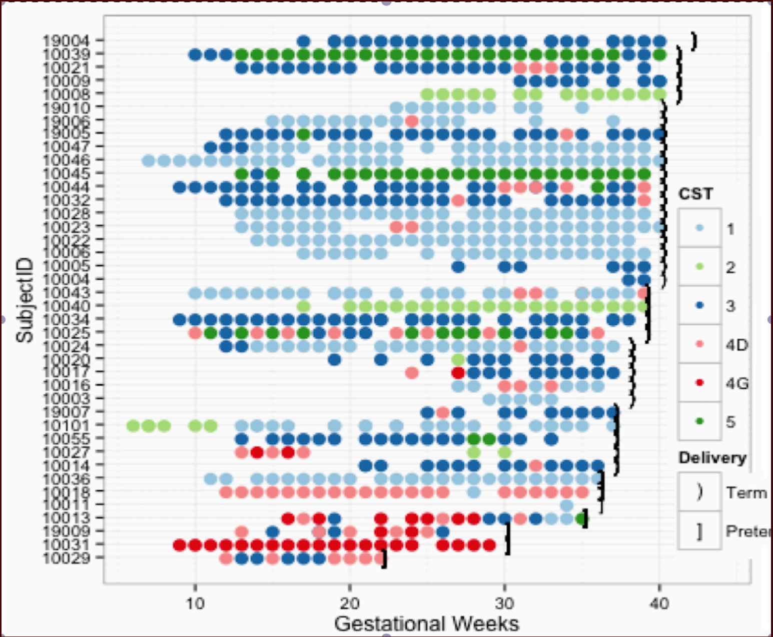


Previously known Microbial Community State Types

Samples into community types and species patterns associated.

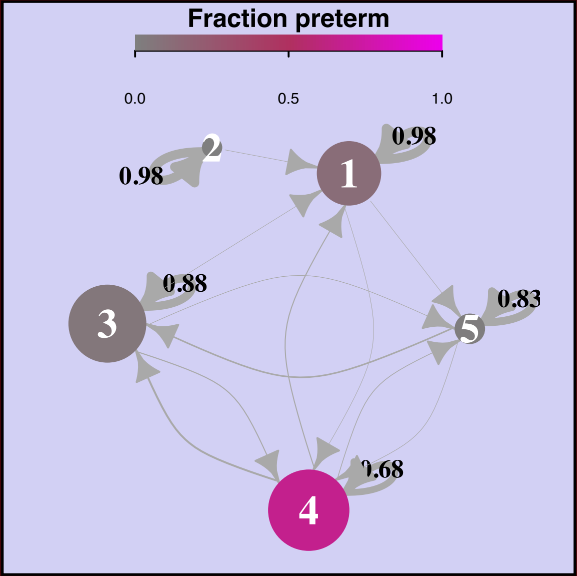


Longitudinal Analyses



Markov Chain Model

Transitions between states, as in simple ecological models.

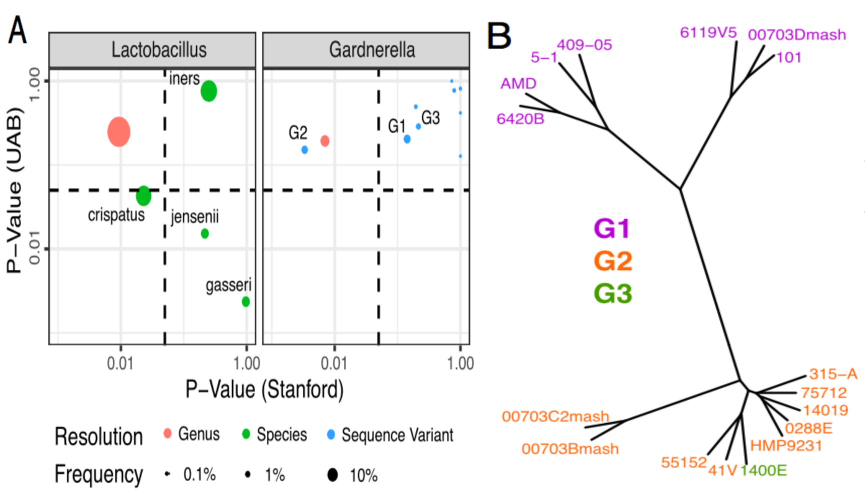


Conclusions for Vaginal Microbiome

- Prevalence of a Lactobacillus-poor vaginal community state type (CST 4) was inversely correlated with gestational age at delivery (p=0.0039).

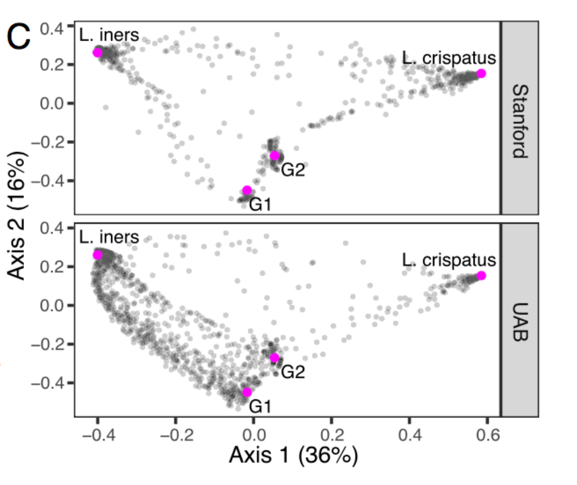
  Risk for preterm birth was more pronounced for subjects with CST 4 accompanied by elevated Gardnerella or Ureaplasma abundances.
- Finding validated with a separate diagnostic set of 246 vaginal specimens from nine women (four of whom delivered preterm).
- Post-delivery vaginal community disturbance with a decrease in Lactobacillus species and an increase in diverse anaerobes such as Peptoniphilus, Prevotella and Anaerococcus species.
- Reproducible research full record:
  `http://statweb.stanford.edu/~susan/papers/PNASRR.html`

Confirmation and Replication



Callahan BJ, DiGiulio DB, ... & Holmes, SP and Relman, DA Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. PNAS, 2017, Aug 28:201705899.

Confirmation and Replication



Callahan BJ, DiGiulio DB, ... & Holmes, SP and Relman, DA Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. PNAS, 2017, Aug 28:201705899.

Part IV

Uncertainty propagation - putting the noise back.

What are the data ?

A contingency table:

Table: An example of species table.

Taxa	Ctrl1	Ctrl2	Ctrl3	Ctrl4	Ctrl5	IBD1	IBD2	IBD3	IBD4
Bacteroides	1822	913	147	2988	4616	172	3516	657	550
Bifidobacterium	0	162	0	0	84	0	85	1927	0
Collinsella	1359	0	0	206	0	327	0	0	160
Enterococcus	621	0	0	3	40	0	0	0	0
Streptococcus	75	139	2161	110	97	1820	85	58	5

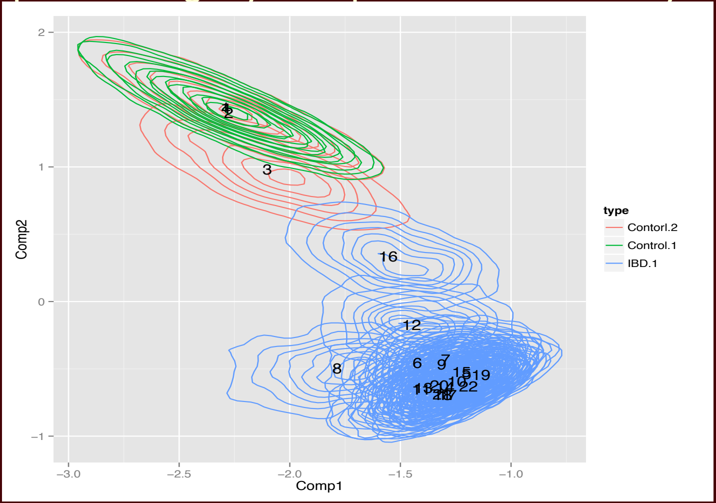
Models for taxa: dependent and ‘infinite‘

- Contingency tables with Taxa counts across biological samples.
- Idea 1: C. Quince, 2012: Dirichlet-Multinomial,

Models for taxa: dependent and ‘infinite‘

- Contingency tables with Taxa counts across biological samples.
- Idea 1: C. Quince, 2012: Dirichlet-Multinomial,
  .... however taxa ar not known in advance (∞).
- Need: Latent factors to describe variations of Taxa counts across biological samples.
- Bayesian analysis for dependent distributions to endow ordinations with estimates of uncertainty.

Output showing Bayesian posterior uncertainty measures



A Bayesian nonparametric prior for dependent normalized random measures is constructed, which is marginally equivalent to a normalized generalized Gamma process.

Ren, Bacallado, Favaro, Holmes, Trippa (2017, JASA)

The contingency table and the sample ‘distributions‘

Contingency table $(n_{i,j})_{i\leq I, j\leq J}$,

- Where $I = \#$ observed taxa.
- And $J = \#$ biological samples.
- $n_{i,j}$ be the observed frequency of species Z_i in biological sample j.
- Considered multinomial with $P^j\{Z_i\}$ probability of seeing species i in sample j.
- All samples have the same (infinite) set of taxa $Z_1, Z_2, \ldots \in \mathcal{Z}$.
- We expect the variation in the respective P^j's to be explained by specific characteristics of the samples (low dimensional latent factors).

The latent factors we don't know

$$
\begin{aligned}
P^j(A) &= M^j(A)/M^j(\mathcal{Z}), \\
M^j(A) &= \sum_{i=1}^{\infty} \mathbb{I}(Z_i \in A)\sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2},
\end{aligned}
\tag{1}
$$

where $\sigma_i \in (0, 1)$, $\mathbf{X}_i, \mathbf{Y}^j \in \mathbb{R}^m$, $\mathbb{I}(\cdot)$ is the indicator function, and $x^+ = x \times \mathbb{I}(x > 0)$. In addition, $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^m$.

$$
\begin{aligned}
P^j(A) &= M^j(A)/M^j(\mathcal{Z}), \\
M^j(A) &= \sum_{i=1}^{\infty} \mathbb{I}(Z_i \in A)\sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2},
\end{aligned}
\tag{2}
$$

- Assume m = number of latent characteristics.
- σ_i is related to the average abundance of taxa i across all biological samples (large if taxa i is prevalent: Z_i will also be large).
- $\mathbf{X}_i$ and $\mathbf{Y}^j$ as taxa vector and biological sample vectors.
- The variation of the P^j's is determined by the latent characteristics in vectors $\mathbf{Y}^j$,
- The vector $\mathbf{X}_i$ denotes the effects of each of the m latent characteristics on the abundance of the taxa Z_i. ($\mathbf{X}_i$ has m entries).

$$P^{j}(A) = M^{j}(A)/M^{j}(\mathcal{Z}),$$
$$M^{j}(A) = \sum_{i=1}^{\infty} \mathbb{I}(Z_i \in A)\sigma_i \langle \mathbf{X}_i, \mathbf{Y}^{j}\rangle^{+2},$$

For Z_i, ... a sequences of independent random variables $\sim$ G.
P^{j} is a Dirichlet process with base measure G.
The prior on $\sigma = (\sigma_1, \sigma_2, \ldots)$ is the distribution of ordered points ($\sigma_i > \sigma_{i+1}$) in a Poisson process on $(0, 1)$ with intensity

$$\nu(\sigma) = \alpha\sigma^{-1}(1 - \sigma)^{-1/2}, \tag{3}$$

where $\alpha > 0$ is a concentration parameter.

Similarity between P^{j}'s

The degree of similarity between the discrete distributions $\{\mathrm{P}^{\mathrm{j}}; \mathrm{j} \in \mathcal{J}\}$ is summarized by a Gram matrix $(\phi(\mathrm{j}, \mathrm{j}') = \langle \mathbf{Y}^{\mathrm{j}}, \mathbf{Y}^{\mathrm{j}'} \rangle; \mathrm{j}, \mathrm{j}' \in \mathcal{J})$.
Parameters for samples

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

Define a joint prior on these factors through the Gram matrix

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

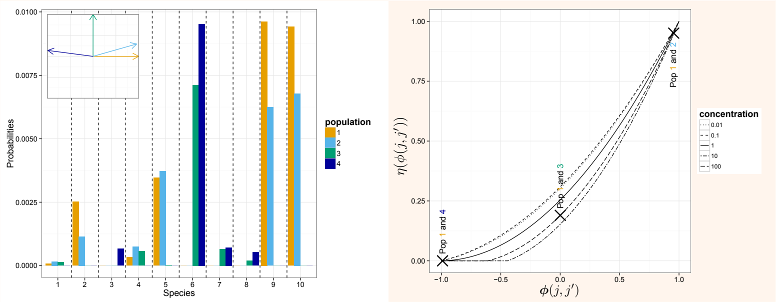
The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of Taxa s.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \tag{4}$$

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

Correlation of $P^j(A)$ and $P^{j'}(A)$

There exists a real function $\eta : [0,1] \rightarrow [0,1]$ such that the correlation between $P^j(A)$ and $P^{j'}(A)$ is equal to $\eta\left(\phi(j,j')\right)$ for every A that satisfies $G(A) > 0$. In different words, the correlation between $P^j(A)$ and $P^{j'}(A)$ does not depend on the specific measurable set A, it is a function of the angle defined by $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$.

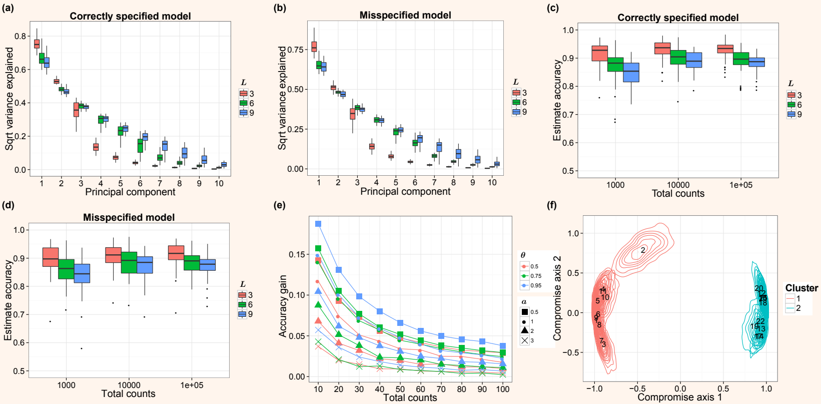


Left panel: realization of 4 microbial distributions from a dependent Dirichlet processes with 10 Taxa s .

Right panel: correlation of two random probability measures when the cosine $\phi(j,j')$ between $\mathbf{Y}^{j}$ and $\mathbf{Y}^{j'}$ varies from -1 to 1.

(Ren et al, JASA, 2017).

Simulations

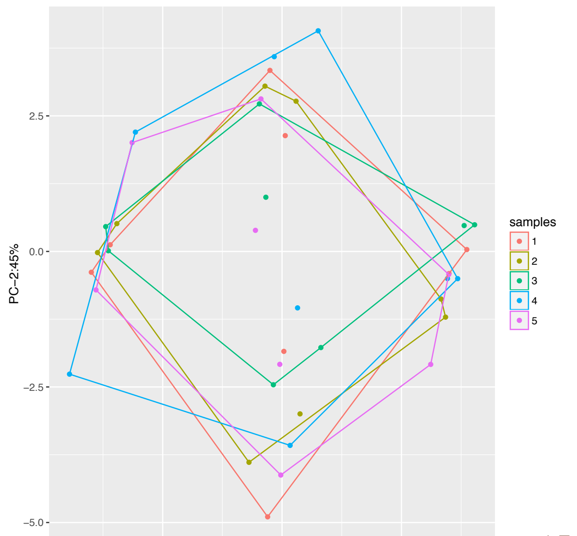


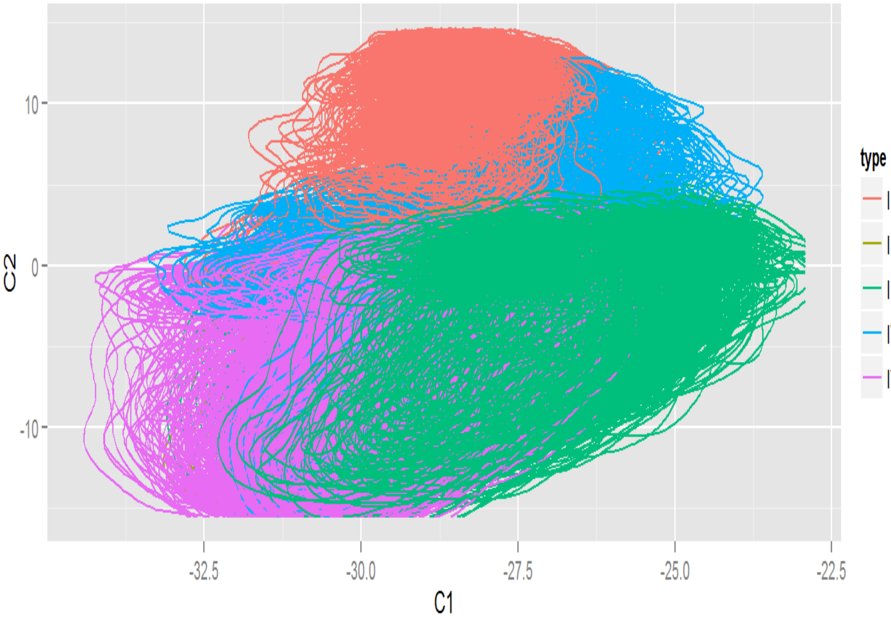
PCA-type projections

Use the normalized Gram matrix **S** between biological samples.
S is the correlation matrix of $(Q_{i,1}, \ldots, Q_{i,J})$.
Based on a single posterior instance of **S**: visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once.
Many instances of **S**.

A projection approach for all points?

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
I
I
I
I

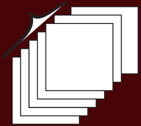
Why?

- Principal coordinate directions are only defined up to a sign.
- Principal coordinates, 1 and 2 or 2 and 3 can be permuted.
- We need to do registration first.

Alternatively

We identify a consensus lower dimensional space for all posterior samples using STATIS (Escoufier, 1980, see also Holmes, 2005). We list the three main steps used to visualize the variability of $\mathbf{S}$.

Registration: Find $\mathbf{S}_0$



Identify a Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_K$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$.

We prefer to choose $\mathbf{S}_0$, the Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_K$.

We use the **RV** similarity metric between two symmetric square matrices **A** and **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:

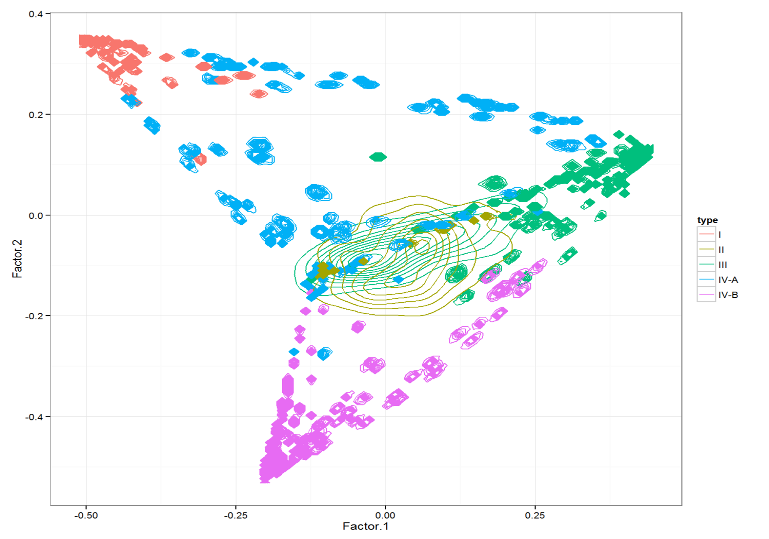
$$(\psi_1^0, \psi_2^0) = \mathbf{S}_0(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2}).$$

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.

Posterior distribution of ordination projections



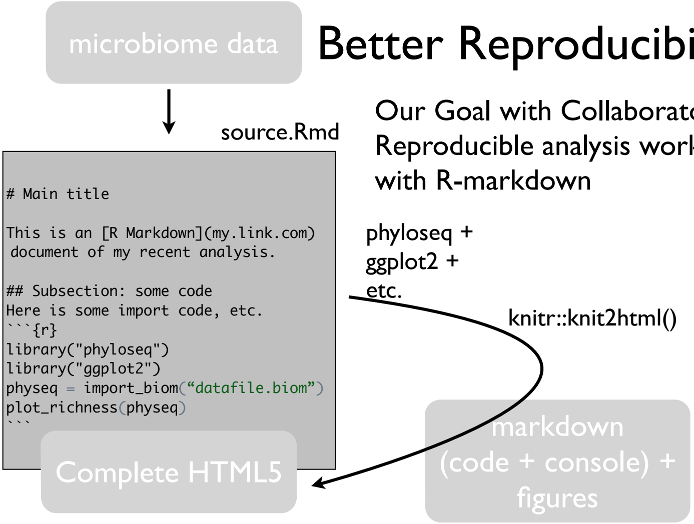
Given posterior samples of the model parameters, we use a procedure to plot credible regions in visualizations.

Availability: currently with Gibbs sampler

Rpackage: `https://github.com/boyuren158/DirFactor`

Better Reproducibility

Our Goal with Collaborators:
Reproducible analysis workflow
with R-markdown



microbiome data

source.Rmd

~~~
# Main title

This is an [R Markdown](my.link.com)
 document of my recent analysis.

## Subsection: some code
Here is some import code, etc.
```{r}
library("phyloseq")
library("ggplot2")
physeq = import_biom("datafile.biom")
plot_richness(physeq)
```
~~~

Reproduce our research

- Complete workflow from reads to community networks, F1000Research. F1000Research paper
- Pregnancy study, PNAS 2015 Delivery Perturbation
- Enterotypes, oral microbiome PSB 2016.
- Waste not, want not paper, Plos Comp Bio. supplemental: Waste not, want not

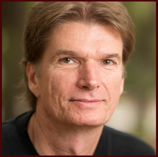
Benefitting from the tools and schools of Statisticians......

Thanks to the `R` and `Bioconductor` community:
Chessel and team for `ade4` , Wolfgang Huber and his team for `DESeq2`, and Emmanuel Paradis for `ape`.

RStudio

Collaborators:

David Relman



Alfred Spormann

Elisabeth Purdom

Josh Elias

Justin Sonnenburg

Sergio Bacallado

Lab Group

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.

Goals already attained:

- Data quality through more NGS denoising (DADA implementation)[1].
- Data integration `phyloseq`.
- Data normalization **Gamma-Poisson** noise model (tutorial).
- High quality graphics, easy to make and change.
- Conjoint analyses of trees, networks and count data.
- Threshold, sensitivity tests and modeling simulations.
- Interactive graph visualizations: Shiny-phyloseq.
- Reproducibility: open source standards, publication of source code and data. (`R`).

Current work in progress

- Longitudinal analyses : antibiotic dynamics and perturbations.
- NMR, Mass spec, proteomic multi-table integration within `phyloseq`.
- Spatial studies: oral microbiome.

References

BJ Callahan, PJ McMurdie, MJ Rosen, AW Han, AJ Johnson, and SP Holmes.
Dada2: High resolution sample inference from amplicon data.
Nature Methods, 2016.

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

S. Holmes, A. Alekseyenko, A. Timme, T. Nelson, P.J. Pasricha, and A. Spormann.
Visualization and statistical comparisons of microbial communities using R packages on Phylochip data.
In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 142, 2011.

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.

Purna C Kashyap, Angela Marcobal, Luke K Ursell, Samuel A Smits, Erica D Sonnenburg, Elizabeth K Costello, Steven K Higginbottom, Steven E Domino, Susan P Holmes, David A Relman, et al.
Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota.
Proceedings of the National Academy of Sciences, page 201306070, 2013.

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

P. J. McMurdie and S. Holmes.
Waste not, want not: Why rarefying microbiome data is inadmissible.
Plos Computational Biology, 2014.

April 03.

Sandrine Pavoine, Anne-Béatrice Dufour, and Daniel Chessel.
From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis.
Journal of Theoretical Biology, 228(4):523–537, 2004.

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