

# Multidomain methods: using the data, all the data.

Susan Holmes

<http://webstat.stanford.edu/~susan/>

@SherlockpHolmes

Bio-X and Statistics, Stanford University

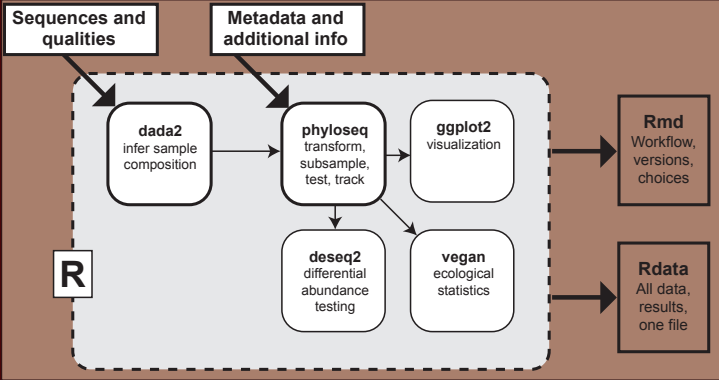
Pune, December 2019



## Solving some of the challenges when working on noisy biological data analyses.

- ▶ Tree and graph integration, uncertainty visualizations.
- ▶ Multi-table data integration.

# Reproducible Research Workflow



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<sup>1</sup>, Kris Sankaran<sup>1</sup>, Julia A. Fukuyama<sup>1</sup>, Paul J. McMurdie<sup>2</sup>,  Susan P. Holmes<sup>1</sup>

 [Author affiliations](#)

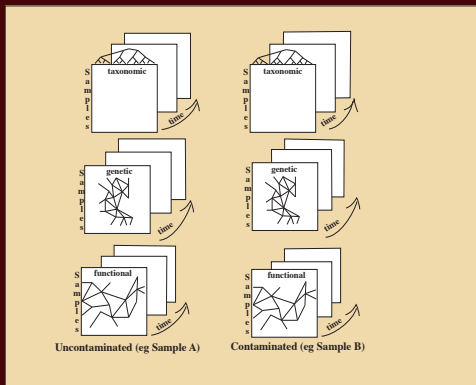
 [Grant information](#)



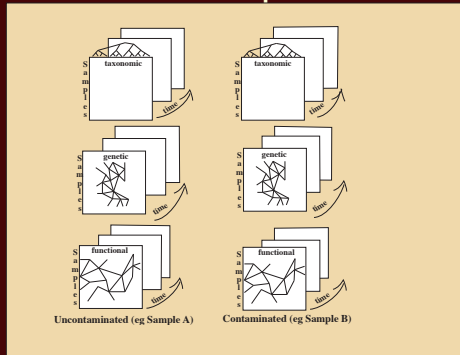
This article is included in the **Bioconductor** channel.

# Part I

## *Multidomain Data integration*



# Useful first order representation: Many Matrices

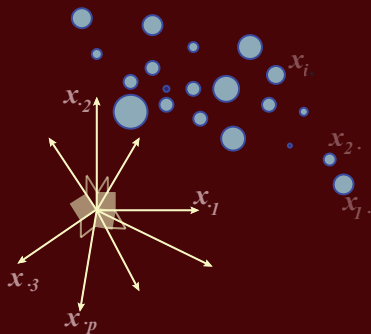


- ▶ Time series of abundance matrices.
- ▶ Bootstrap and Bayesian posterior analyses for many networks.
- ▶ Different types of data on same samples (taxa counts, clinical variates, spatial location).
- ▶ Networks in longitudinal studies.
- ▶ Explanatory (environmental) variables, Response variables.

# We can add information through choice of **distances**

Sample data can often be seen as points in a state space.

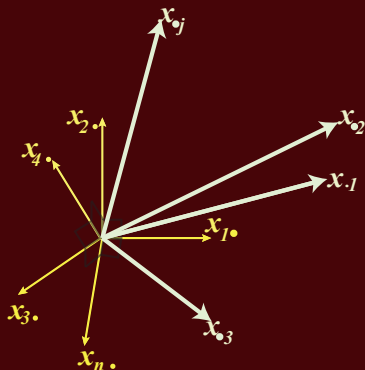
$\mathbb{R}^p$



$$x^t Q y = \langle x, y \rangle_Q$$

Variables are 'vectors' in data point space

$\mathbb{R}^n$



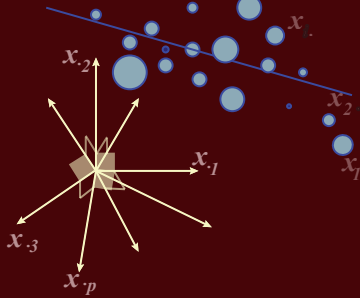
$$x^t D y = \langle x, y \rangle_D$$

Duality : Transposable data.

# Data Analysis: Geometrical Approach

- i. The data are  $p$  variables measured on  $n$  observations.
- ii.  $X$  with  $n$  rows (the observations) and  $p$  columns (the variables).
- iii.  $D$  is an  $n \times n$  matrix of weights on the “observations”, which is most often diagonal but not always.
- iv Symmetric definite positive matrix  $Q$ , weights on

variables, often  $Q = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & 0 & 0 & \dots \\ 0 & 0 & \ddots & 0 & \dots \\ \vdots & \dots & \dots & 0 & \frac{1}{\sigma_p^2} \end{pmatrix}.$





# Generalized Principal Component Analysis

gPCA seeks to replace the original (centered) matrix  $X$  by a matrix of lower rank, this can be solved using the singular value decomposition of  $X$ :

$$X = USV', \text{ with } U'DU = I_n \text{ and } V'QV = I_p \text{ and } S \text{ diagonal}$$

$$XX' = US^2U', \text{ with } U'DU = I_n \text{ and } S^2 = \Lambda$$

PCA is a linear nonparametric multivariate method for dimension reduction.  $D$  and  $Q$  are the relevant metrics on the dual row and column spaces of  $n$  samples and  $p$  variables.

# Discriminant Analysis is a special case

Case of a categorical response variable (group labels).

Let  $A$  be the  $g \times p$  matrix of group means in each of the  $p$  variables.

This satisfies

$$Y^tDX = \Delta_Y A \quad \text{where } \Delta_Y = Y^tDY = \text{diag}(w_1, w_2, \dots, w_g),$$

and  $w_k = \sum_{i: y_{ik}=1} d_i$ , the  $w_k$ 's are the group weights, as they are the sums of the weights as defined by  $D$  for all the elements in that group.

Call  $T$  the matrix  $T = X^tDX$ , a generalized between group variance-covariance is  $B = A^t\Delta_Y A$  and call the between group variance covariance the matrix  $W = (X - YA)^tD(X - YA)$ .

Huyghens' formula:

$$T = B + W$$

# Classical Dimension Reduction Algorithm: PCoA or MDS

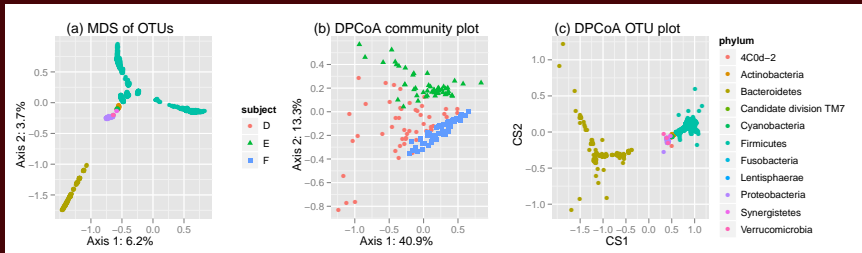
Given an  $n \times n$  matrix of squared interpoint distances  $D \bullet D$ , one can solve for points achieving these distances by:

1. Double centering the interpoint distance squared matrix:

$$B = -\frac{1}{2}HD \bullet DH.$$

2. Diagonalizing B:  $B = U\Lambda U^T$ .

3. Extracting  $\tilde{X}$ :  $\tilde{X} = U\Lambda^{1/2}$ .



(a) PCoA/MDS of the taxa based on the patristic distance, (b) community and (c) species points for DPCoA after removing two outlying species.

# Double Principal Coordinate Analysis

Pavoine, Dufour and Chessel (2004), Purdom (2010) and Fukuyama et al. (2011). . Suppose we have  $n$  species in  $p$  locations and a matrix  $\Delta$  giving the squares of the pairwise distances between the species on the tree (patristic). Then we can

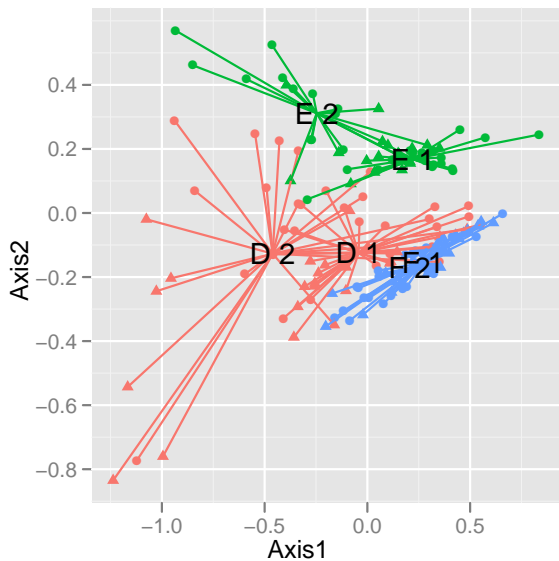
- ▶ Use the distances between species to find an embedding in  $n - 1$ -dimensional space such that the euclidean distances between the species is the same as the distances between the species defined in  $\Delta$ .
- ▶ Place each of the  $p$  locations at the barycenter of its species profile. The euclidean distances between the locations will be the same as the square root of the Rao dissimilarity between them.
- ▶ Use PCA to find a lower-dimensional representation of the locations.

Give the species and communities coordinates such that the inertia decomposes the same way the diversity does.

## Antibiotic Stress

We next want to visualize the effect of the antibiotic. Ordinations of the communities due to DPCoA with information about whether the community was stressed or not stressed (pre cipro, interim, and post cipro were considered “not stressed”, while first cipro, first week post cipro, second cipro, and second week post cipro were considered “stressed”).

DPCoA separates out the stressed communities along the first axis (in the direction associated with *Bacteroidetes*), although only for subjects D and E.



Community points as represented by DPCoA. The labels represent subject plus antibiotic condition.

## Conclusions for Antibiotic Stress

DPCoA also separates the subjects and the stressed versus non-stressed communities, and examining the community and taxa ordinations can tell us about the differences in the compositions of these communities. Much larger study under way with 100 patients and more than 8,000 samples.



## treelapse (Kris Sankaran):Key elements

<https://github.com/krisrs1128/treelapse/>

Enable a rapid change of focus and brushing on the tree and the time series.








treelapse currently supports four kinds displays

- ▶ DOI Trees: Navigate large trees according to the Degree-of-Interest (DOI) defined by clicking on different nodes.
- ▶ DOI Sankeys: Create a DOI Tree where abundances are split across several groups.
- ▶ Timeboxes: Visually query a (tree-structured) collection of time series, and see which nodes are associated with selected series.
- ▶ Treeboxes: The converse of timeboxes – select nodes and see which series are associated.

<http://statweb.stanford.edu/~kriss1/antibiotic.html>.

# Part II

## Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment

Julia Fukuyama , Laurie Rumker , Kris Sankaran , Pratheepa Jeganathan, Les Dethlefsen, David A. Relman  , Susan P. Holmes  

Version 2



Published: August 18, 2017 - <https://doi.org/10.1371/journal.pcbi.1005706>

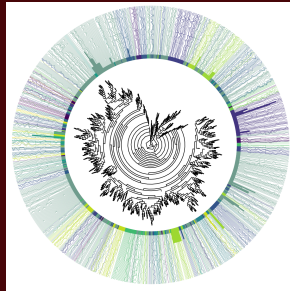
Article

Authors

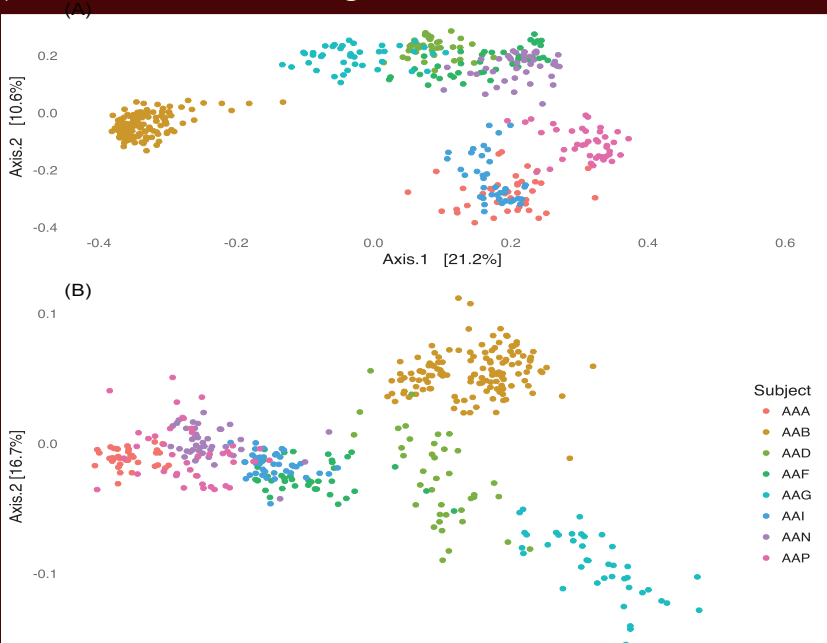
Metrics

Comments

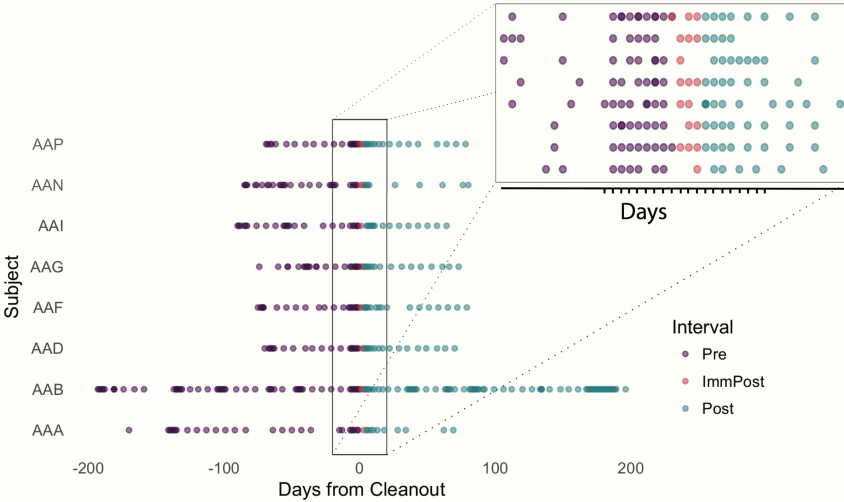
Related Content



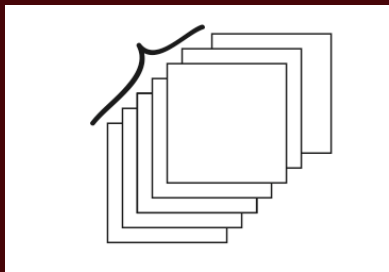
# Subject effect is the strongest



# Experimental Design



## Multidomain data: multiple table methods



In PCA we compute the variance-covariance matrix, in multiple table methods we can take a cube of tables and compute the RV coefficient of their characterizing operators.

We then diagonalize this and find the best weighted 'ensemble'.

This is called the 'compromise' and all the individual tables can be projected onto it.

# Multi-table - multidomain methods

## Inertia, Co-Inertia

We generalize "covariation" in several directions through the idea of inertia.

In physics: inertia is a weighted sum of distances of weighted points. This enables us to use abundance data in a contingency table and compute its inertia which in this case will be the weighted sum of the squares of distances between observed and expected frequencies, such as is used in computing the chis-square statistic.

Another generalization of variance-inertia is the useful Phylogenetic diversity index. (computing the sum of distances between a subset of taxa through the tree).

We also have such generalizations that cover variability of points on a graph taken from standard spatial statistics.

# Co-Inertia

When studying two variables measured at the same locations, for instance PH and humidity the standard quantification of covariation is the *covariance*.

$$\text{sum}(x1 * y1 + x2 * y2 + x3 * y3)$$

if  $x$  and  $y$  co-vary -in the same direction this will be big.

A simple generalization to this when the variability is more complicated to measure as above is done through Co-Inertia analysis (CIA).

Co-inertia analysis (CIA) is a multivariate method that identifies trends or co-relationships in multiple datasets which contain the same samples or the same time points. That is the rows or columns of the matrix have to be weighted similarly and thus must be matchable.

## RV coefficient

The global measure of similarity of two data tables as opposed to two vectors can be done by a generalization of covariance provided by an inner product between tables that gives the RV coefficient, a number between 0 and 1, like a correlation coefficient, but for tables.

$$RV(A, B) = \frac{\text{Tr}(A'B)}{\sqrt{\text{Tr}(A'A)}\sqrt{\text{Tr}(B'B)}}$$

Survey on RV: Josse, Holmes (2016)..



# Sparse CCA, then PCA

CCA: Canonical Correlation Analysis.

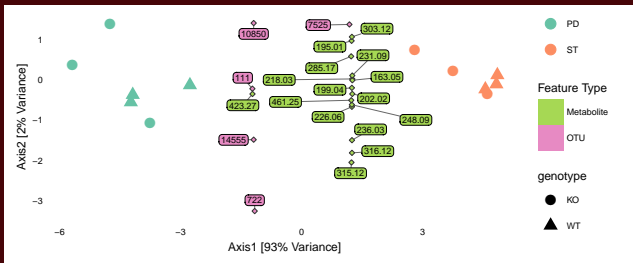
PCA: Principal Components Analysis.

- ▶ There are two tables in the study presented here, one for microbes and another with metabolites. 12 samples were obtained, each with measurements at 637 m/z values and 20,609 OTUs; however, about 96% of the entries of the microbial abundance table are exactly zero.
- ▶ CCA chooses a subset of available features that capture the most **co-Inertia**.
- ▶ We then apply PCA to this selected subset of features. In this sense, we use sparse CCA as a screening procedure, rather than as an ordination method.

```
## Call: CCA(x = t(X), z = t(metab), penaltyx = 0.15,  
##                                           penaltyz = 0.15)  
##  
## Num non-zeros u's: 5  
## Num non-zeros v's: 15  
## Type of x: standard  
## Type of z: standard  
## Penalty for x: L1 bound is 0.15  
## Penalty for z: L1 bound is 0.15  
## Cor(Xu,Zv): 0.974
```

With these parameters, 5 microbes and 15 metabolites have been selected, based on their ability to explain covariation between tables. Further, these 20 features result in a correlation of 0.974 between the two tables.

The microbial and metabolomic data reflect similar underlying signals. To relate the recovered metabolites and OTUs to characteristics of the samples on which they were measured, we use them as input to an ordinary PCA.

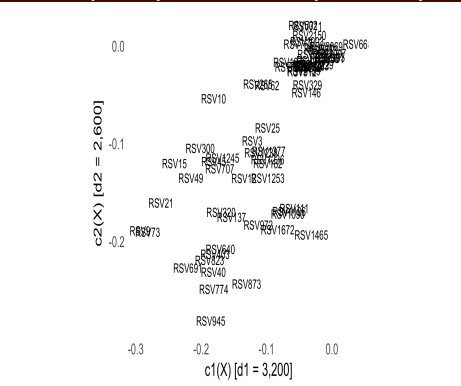
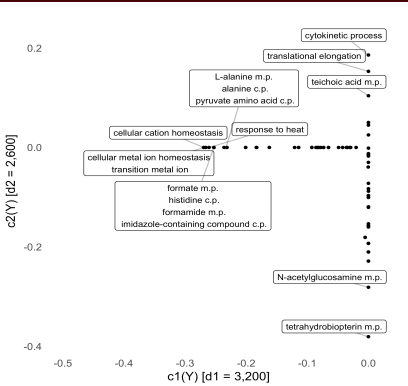


A PCA triplot produced from the CCA selected features in from multiple data types (metabolites and OTUs). Triangles for Knockout and circles for wild type. The main variation in the data is across PD and ST samples (different diets).

Kashyap PC, et al.: Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. Proc Natl Acad Sci U S A. 2013; 110(42): 17059-17064.

# Sparse CCA method for the CC perturbation data.

Create multi-table correlations with sparsity: more interpretability.



# Tree-informed prior modulating deep branches

DPCoA emphasize the deep branches.

- ▶ Q Kernel :  $Q_{ij}$  represents shared ancestral branch length between species  $i$  and  $j$ .
- ▶ Covariance of a Brownian motion run along the branches of the tree.

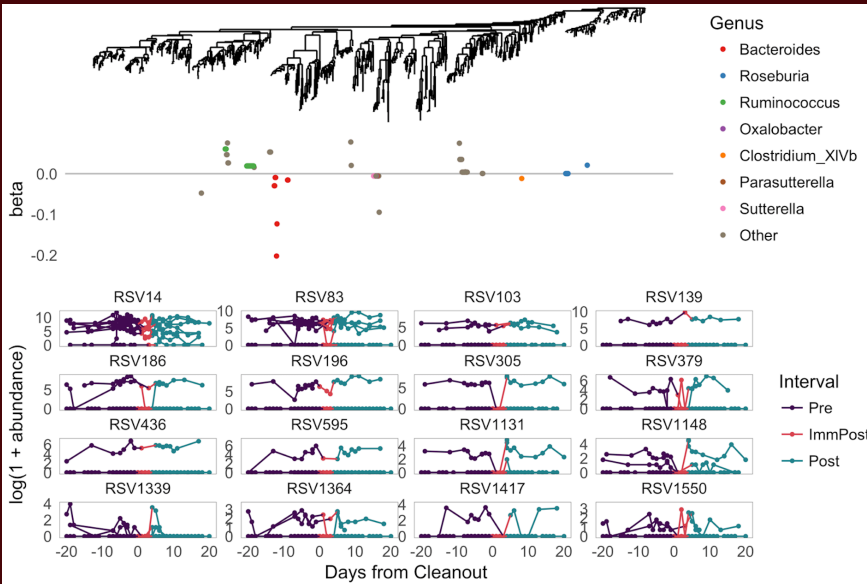
$$x_i \sim N(\mu_i, \sigma_1^2 \mathbb{I}) \quad \mu_i \sim N(0, \sigma_1^2 Q), i = 1, \dots, n.$$

- ▶ Inference using this prior regularizes towards this structure.

$$\mu_i | x_i = x \sim N(\sigma_2^{-1} S x, S) \quad S = (\sigma_1^{-2} Q^{-1} + \sigma_2^{-2} \mathbb{I})^{-1}$$

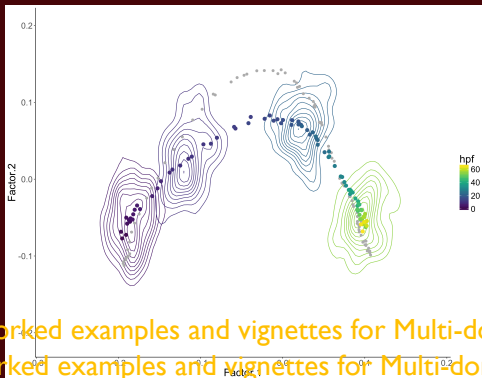
gPCA on  $(X, S, \mathbb{I}_n)$

$\sigma_1/\sigma_2 \rightarrow 0$  then PCA.  $\sigma_2/\sigma_1 \rightarrow 0$  then DPCoA.



Results from tree-based **sparse** discriminant analysis.

# Resources and Workflows



link ([http](#)) to worked examples and vignettes for Multi-domain analyses  
local link to worked examples and vignettes for Multi-domain analyses

# Useful parallel between word-topic modeling and bacteria-communities

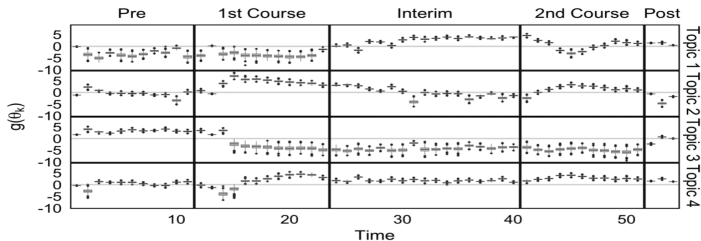


Figure 1: Boxplots represent approximate posteriors for estimated mixture memberships  $\theta_d$ , and their evolution over time. Each row of panels provides a different sequence of  $\theta_{dk}$  for a single  $k$ , and different columns distinguish different phases of sampling. Note that the  $y$ -axis is on the  $g$ -scale, which is defined as a translated logit,  $g(\mathbf{p}) := (\log p_1 - \overline{\log \mathbf{p}}, \dots, \log p_K - \overline{\log \mathbf{p}})$ . The first and second antibiotic time courses result in meaningful shifts in these sequences, and that there appear to be long-term effects of treatment among bacteria in Topic 3.



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

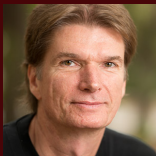
Thanks to the R and Bioconductor community:

Chessel and team for `ade4` , Wolfgang Huber and his team for `DESeq2`,  
and Emmanuel Paradis for `ape`.





David Relman

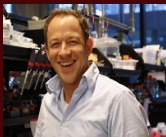


Alfred Spormann



Elisabeth Purdom

Collaborators:



Josh Elias



Justin Sonnenburg



Sergio Bacallado

# Lab Group



**Postdoctoral Fellows** Past: Paul (Joey) McMurdie, Ben Callahan, Christof Seiler, Diana Proctor, Current: Pratheepa Jegathan, Laura Symul, Ann-Maude Ferreira.

**Students:** Past: Daniel Sprockett, Lan Huong Nguyen, John Cherian, Julia Fukuyama, Kris Sankaran.

Current: Claire Donnat.

**Funding from** NIH TR01 and NSF-DMS.

microbiome data

# Better Reproducibility

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)
```
```

Complete HTML5

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

phyloseq +  
ggplot2 +  
etc.

knitr::knit2html()


markdown  
(code + console) +  
figures


## 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](#)


## References

-  Julia Fukuyama, Laurie Rumker, Kris Sankaran, Pratheepa Jeganathan, Les Dethlefsen, David A. Relman, and Susan P. Holmes.  
Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment.  
*PLOS Computational Biology*, (10.1371/journal.pcbi.1005706), 2017.  
August 16.
-  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.

 P. J. McMurdie and S. Holmes.  
Phyloseq: Reproducible 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.

 Elizabeth Purdom.  
Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree.  
*Annals of Applied Statistics*, Jul 2010.