Waste not, want not: using the data, all the data.

Susan Holmes
http://webstat.stanford.edu/˜susan/
@SherlockpHolmes

Bio-X and Statistics, Stanford University

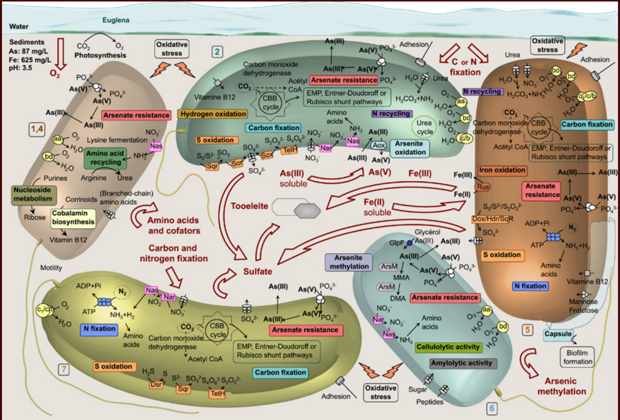
BonB, November 2017

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

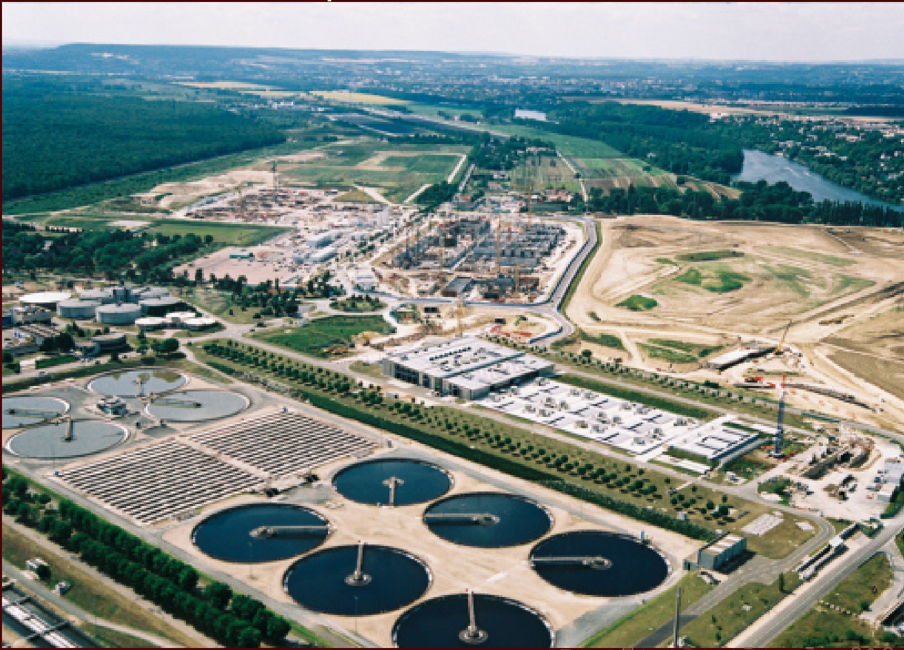
- Heteroscedasticity and sample depth inequality.
- Poor data quality, information leakage.
- Tree and graph integration, uncertainty visualizations.
- Multi-table data integration.

Part I

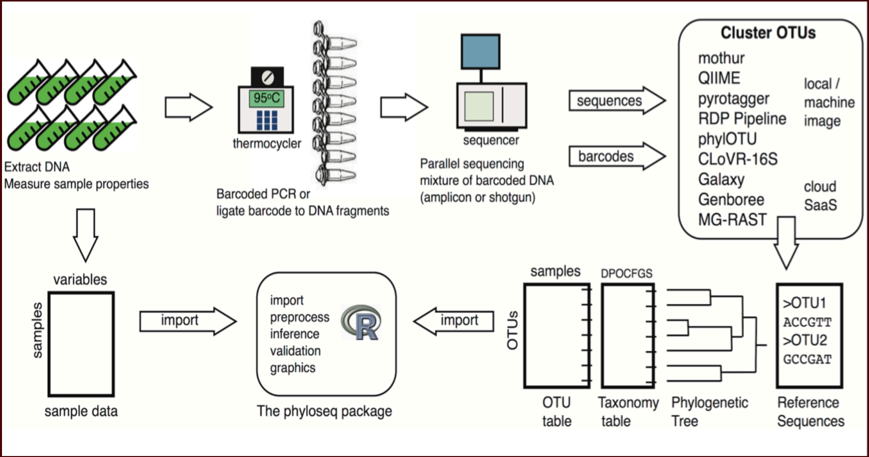
Microbiome



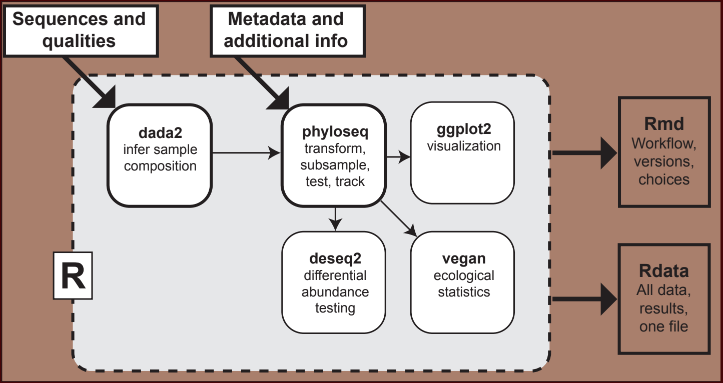
Waste water treatment plants



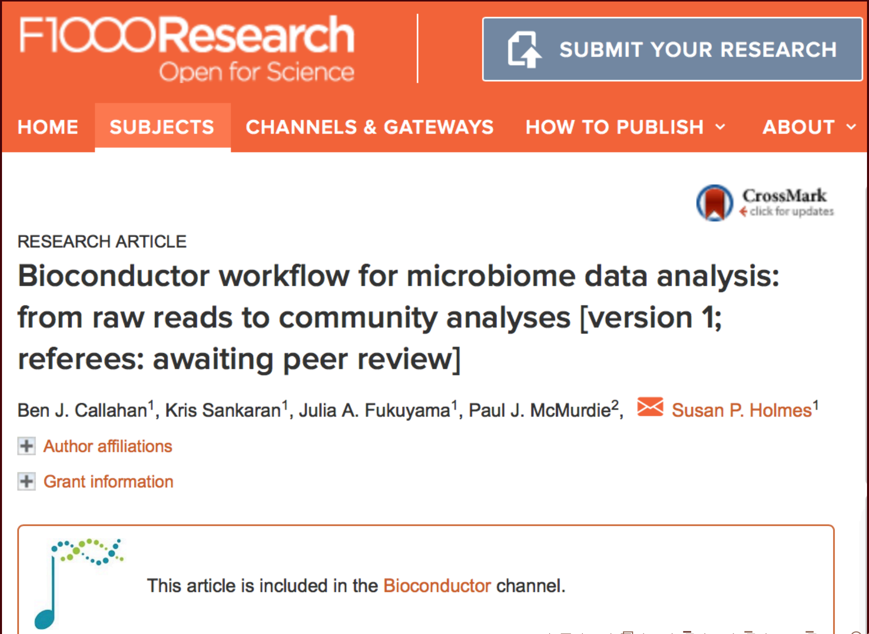
Heterogeneous Data Workflow with **phyloseq**



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]

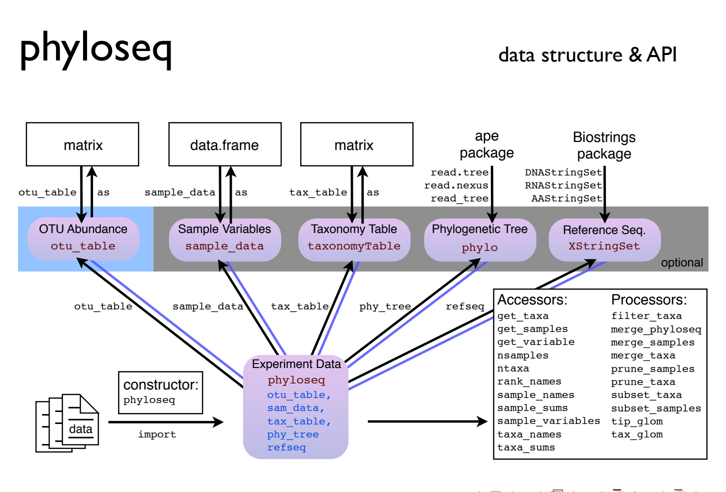
+ Author affiliations

+ Grant information

This article is included in the Bioconductor channel.

Heterogeneous Data Objects

Object oriented input and data manipulation with phyloseq



phyloseq

data structure & API

matrix

data.frame

matrix

ape package

Biostrings package

otu_table as

sample_data as

tax_table as

read.tree
read.nexus
read_tree

DNAStringSet
RNAStringSet
AAStringSet

OTU Abundance
otu_table

Sample Variables
sample_data

Taxonomy Table
taxonomyTable

Phylogenetic Tree
phylo

Reference Seq.
XStringSet

optional

otu_table

sample_data

tax_table

phy_tree

refseq

Experiment Data
phyloseq
otu_table,
sam_data,
tax_table,
phy_tree
refseq

data

constructor:
phyloseq

import

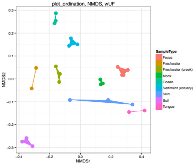
Accessors:
get_taxa
get_samples
get_variable
nsamples
ntaxa
rank_names
sample_names
sample_sums
sample_variables
taxa_names
taxa_sums

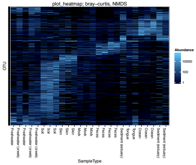
Processors:
filter_taxa
merge_phyloseq
merge_samples
merge_taxa
prune_samples
prune_taxa
subset_taxa
subset_samples
tip_glom
tax_glom

phyloseq

graphics

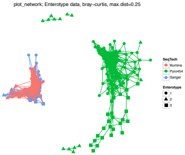
plot_ordination()

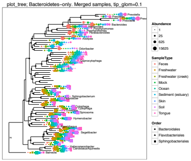




plot_heatmap()

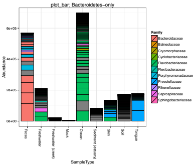
plot_network()

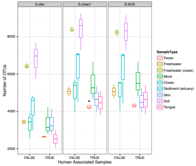




plot_tree()

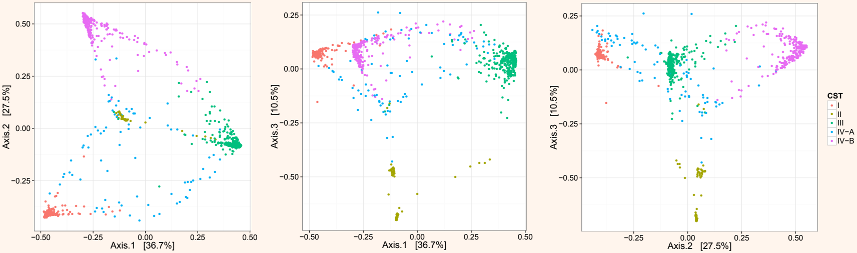
plot_bar()





plot_richness()

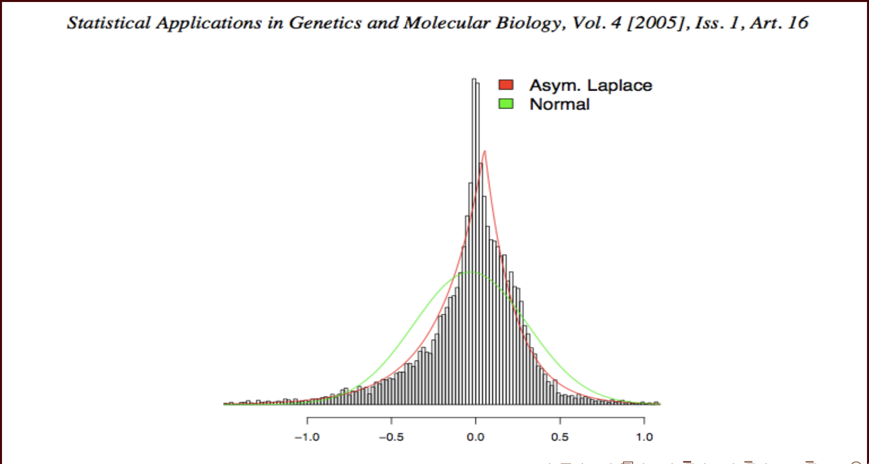
Output of phyloseq ordination



Microbial communities (Ravel, 2013)

Part II

Heteroscedasticity: Mixtures and how to Normalize them



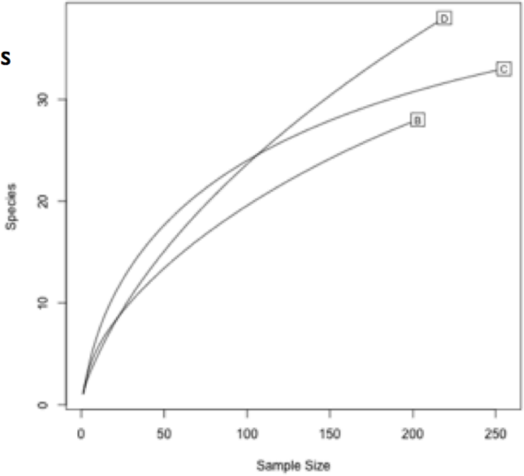
Statistical Applications in Genetics and Molecular Biology, Vol. 4 [2005], Iss. 1, Art. 16

How to deal with different numbers of reads?

- Sanders 1968
- non-parametric richness
- estimate coverage
- Normalize?

Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*

rarefaction curves



Current Method: Rarefying

Ad hoc library size normalization by random subsampling without replacement.

1. Select a minimum library size, $N_{L,min}$. This has also been called the *rarefaction level* though we will not use the term here.
2. Discard libraries (microbiome samples) that have fewer reads than $N_{L,min}$.
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,min}$.

Often $N_{L,min}$ is chosen to be equal to the size of the smallest library that is not considered *defective*, and the process of identifying defective samples comes with a risk of subjectivity and bias. In many cases researchers have also failed to repeat the random subsampling step (3) or record the pseudorandom number generation seed/process — both of which are essential for reproducibility.

Reduction of Data to Proportions

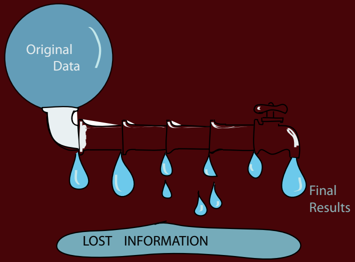
Many software programs automatically reduce the data to relative proportions, losing the information about library sizes or read counts. This makes comparisons very difficult.

Statistical Formulation: When making a (testing) decision, reducing results from a Binomial distribution into a proportion does not give an **admissible** procedure.

Definition:An admissible rule is an optimal rule for making a decision in the sense that there is no other rule that is always *better* than it.

How to compress the data?

.



...without losing too much information?

The proportion is not a sufficient statistic for the Binomial.

A statistic T(X) is called sufficient for θ if it contains all the information in X about θ.

Standard:

The joint probability distribution of the data conditional on the value of a sufficient statistic for a parameter, does not depend on that parameter: $P_\theta(X|T(X) = T)$ does not depend on θ. Wiki

Equivalent Definitions

Mutual Information:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} = K(P(x, y), P(x)P(y))$$

A function of the data $T(X)$ is a sufficient statistic for the distribution if

$$I(\theta, X) = I(\theta, T(X))$$

for all distributions on θ.

Note:

For a Bayesian, no matter what prior one uses, one only has to consider the sufficient statistic for making inference, because the posterior distribution given $T = T(x)$ is the same as the posterior given the data $X = x$.

Aim of the studies: Differential Abundance

Like differentially expressed genes, a species/OTU is considered differentially abundant if its mean proportion is significantly different between two or more sample classes in the experimental design.

Optimality Criteria:

Sensititivity or Power True Positive Rate.

Specificity True Negative Rate.

We have to control for many sources of error (blocking, modeling, etc..)

Rarefaction and Reduction to Proportions are Inadmissible

The following is a minimal example to explain why rarefying is statistically inadmissible, especially with regards to variance stabilization. Suppose we want to compare two different samples, called *A* and *B*, comprised of 100 and 1000 reads, respectively. In these hypothetical communities only two types of microbes have been observed, *OTU1* and *OTU2*

According to Table 1, Left.

Table: **A minimal example of the effect of rarefying on power.**

Original Abundance		
	A	B
OTU1	62	500
OTU2	38	500
Total	100	1000

Rarefied Abundance		
	A	B
OTU1	62	50
OTU2	38	50
	100	100

Standard Tests for Difference

P-value	χ^2	Prop	Fisher
Original	0.0290	0.0290	0.0272
Rarefied	0.1171	0.1171	0.1169

Hypothetical abundance data in its original (Top-Left) and rarefied (Top-Right) form, with corresponding formal test results for differentiation (Bottom).

Formally comparing the two proportions according to a standard test is done either using a χ^2 test (equivalent to a two sample proportion test here) or a Fisher exact test. This requires knowledge of the number of trials.

By rarefying (Table 1, top-right) so that both samples have the same number of counts, we are no longer able to differentiate between them. This loss of power is completely attributable to reducing the size of *B* by a factor of 10, which also increases the confidence intervals corresponding to each proportion such that they are no longer distinguishable from those in *A*, even though they are distinguishable in the original data.

The variance of the proportion's estimate $\hat{p}$ is multiplied by 10 when the total count is divided by 10.

Equalization of variances

In this binomial example the variance of the proportion estimate is $Var(\frac{X}{n}) = \frac{pq}{n} = \frac{q}{n}E(\frac{X}{n})$, a function of the mean. This is a common occurrence and one that is traditionally dealt with in statistics by applying variance-stabilizing transformations.

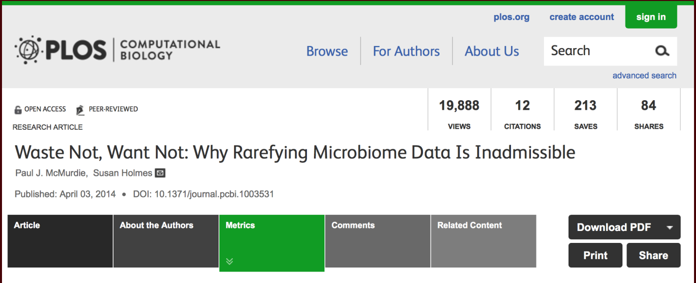
However, in order to find the right transformation, we need a good model for the error.

Mixture Modeling works Miracles

- Beta-Binomial (deepSNV).
- Zero inflated Poisson or Gaussian.
- Gamma-Poisson and ZINB.

Mixtures are ubiquitous because of a mathematical theorem
De Finnetti's Theorem

Gamma-Poisson mixture



RESEARCH ARTICLE

19,888	12	213	84
VIEWS	CITATIONS	SAVES	SHARES

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes

Published: April 03, 2014 • DOI: 10.1371/journal.pcbi.1003531

McMurdie and Holmes (2014) “Waste Not, Want Not: Why rarefying microbiome data is inadmissible” **Negative Binomial as a hierarchical mixture for read counts**

If technical replicates have same number of reads: s_j,

Poisson variation with mean $\mu = s_j u_i$.

Taxa i incidence proportion u_i.

Number of reads for the sample j and taxa i would be

$$K_{ij} \sim \text{Poisson}\,(s_j u_i)$$

Same as DESeq for RNA-seq

Variance Stabilization

Prefer to deal with errors across samples which are independent and identically distributed.

In particular homoscedasticity (equal variances) across all the noise levels.

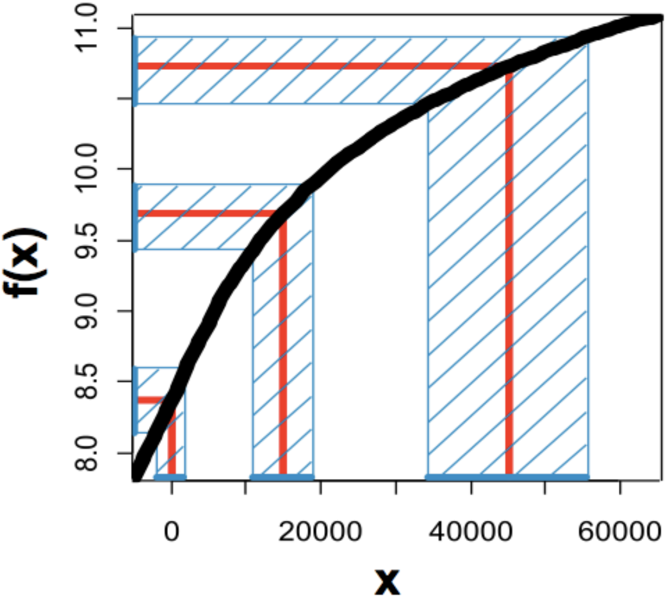
This is not the case when we have unequal sample sizes and variations in the accuracy across instruments.

A standard way of dealing with heteroscedastic noise is to try to decompose the sources of heterogeneity and apply transformations that make the noise variance almost constant.

These are called *variance stabilizing transformations*.

Take for instance different Poisson variables with mean μ_i. Their variances are all different if the μ_i are different. However, if the square root transformation is applied to each of the variables, then the transformed variables will have approximately constant variance. Actually if we take the transformation $x \longrightarrow 2\sqrt{x}$ we obtain a variance approximately equal to 1..

variance stabilizing transformation



Modeling read counts

If technical replicates have same number of reads: s_j,
Poisson variation with mean $\mu = s_j u_i$.
Taxa i incidence proportion u_i.
Number of reads for the sample j and taxa i would be

$$K_{ij} \sim \text{ Poisson } (s_j u_i)$$

Modeling Counts

For biological replicates within the same group – such as treatment or control groups or the same environments – the proportions u_i will be variable between samples.
Call the two parameters r_i and $\frac{p_i}{1-p_i}$.
So that U_{ij} the proportion of taxa i in sample j is distributed according to Gamma$(r_i, \frac{p_i}{1-p_i})$.
K_{ij} have a Poisson-Gamma mixture of different Poisson variables.
This gives the Negative Binomial with parameters $(m = u_i s_j)$ and ϕ_i as a satisfactory model of the variability.

Different Conditions

Samples belong to different conditions such as treatment and control or different environments.

Estimate the values of the parameters separately for each of the different biological replicate conditions/classes.

Use the index c for the different conditions, we then have the counts for the taxa i and sample j in condition c having a Negative Binomial distribution with $m_c = u_{ic}s_j$ and ϕ_{ic} so that the variance is written

$$u_{ic}s_j + \phi_{ic}s_j^2u_{ic}^2 \tag{1}$$

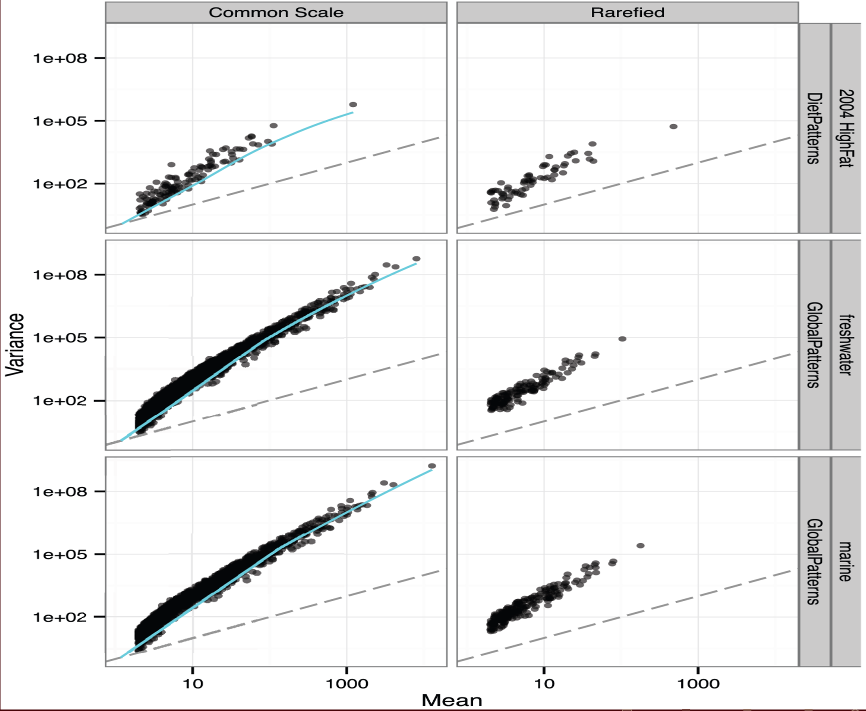
Estimate the parameters u_{ic} and ϕ_{ic} from the data for each OTU and sample condition.

The end result provides a variance stabilizing transformation of the data that allows a statistically efficient comparisons between conditions.

This application of a hierarchical mixture model is very similar to the random effects models used in the context of analysis of variance.

Overdispersion in 16S rRNA-seq Data

Common-Scale Variance versus Mean for Microbiome Data.
Each point in each panel represents a different OTU's mean/variance estimate for a biological replicate and study.
The data in this figure come from the *Global Patterns* surveyand the *Long-Term Dietary Patterns* study(Right) Variance versus mean abundance for rarefied counts.
(Left) Common-scale variances and common-scale means, estimated according to the DESeq2 package.
The dashed gray line denotes the $\sigma^2 = \mu$ case (Poisson; $\phi = 0$). The cyan curve denotes the fitted variance estimate using DESeq.


Common Scale
Rarefied
2004 HighFat
DietPatterns
freshwater
GlobalPatterns
marine
GlobalPatterns
Variance
1e+08
1e+05
1e+02
Mean
10
1000

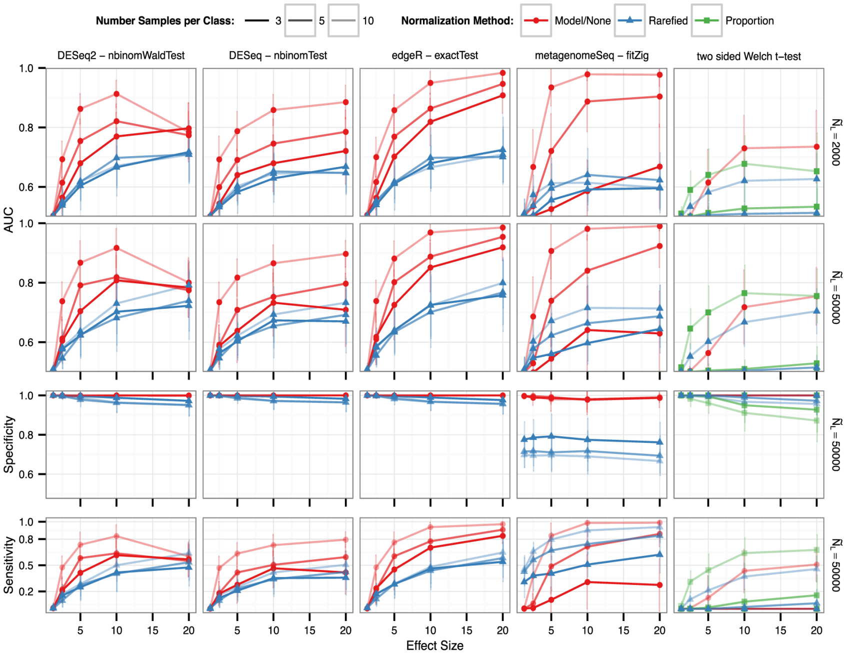
Improvement in Power and FDR

Performance of differential abundance detection with and without rarefying summarized by “Area Under the Curve” (AUC) metric of a Receiver Operator Curve (ROC) (vertical axis).
Briefly, the AUC value varies from 0.5 (random) to 1.0 (perfect).
The horizontal axis indicates the effect size, shown as the factor applied to OTU counts to simulate a differential abundance.
Each curve traces the respective normalization method’s mean performance of that panel, with a vertical bar indicating a standard deviation in performance across all replicates and microbiome templates.

The right-hand side of the panel rows indicates the median library size, N, while the darkness of line shading indicates the number of samples per simulated experiment.

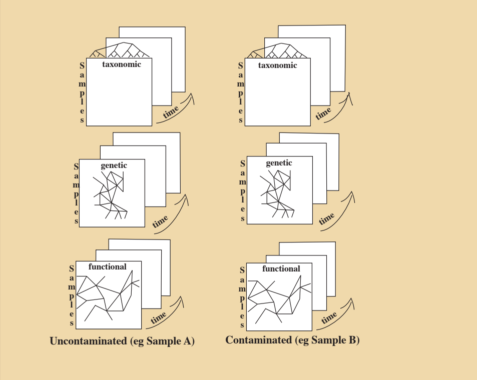
Color shade and shape indicate the normalization method.

Detection among multiple tests was defined using a False Discovery Rate (Benjamini-Hochberg) significance threshold of 0.05.

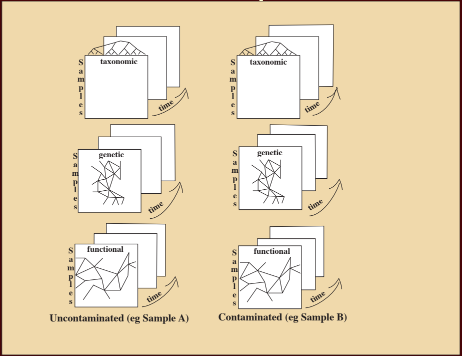

Number Samples per Class:
3
5
10
Normalization Method:
Model/None
Rarefied
Proportion
DESeq2 – nbinomWaldTest
DESeq – nbinomTest
edgeR – exactTest
metagenomeSeq – fitZig
two sided Welch t-test
AUC
Specificity
Sensitivity
1.0
0.8
0.6
0.5
0.2
5
10
15
20
Effect Size
Ñ_L = 2000
Ñ_L = 50000

Part III

Multidomain Data integration



Useful first order representation: Many Matrices



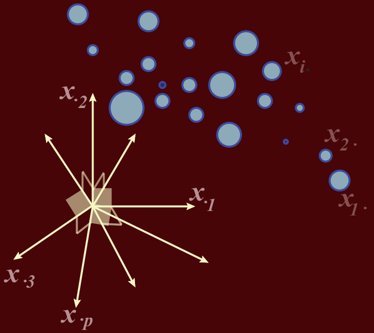
- Time series of abundance matrices.
- Bootstrap and Bayesian posterior analyses for many networks.
- Different types of data on same samples (taxa counts, clinical variates, spatial location).
- Networks in longitudinal studies.
- Explanatory (environmental) variables, Response variables.

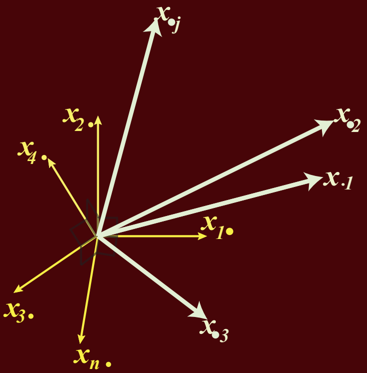
Holmes (2005), Duality Diagrams.

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

Sample data can often be seen as points in a state space.
$\mathbb{R}^p$

Variables are ‘vectors’ in data point space
$\mathbb{R}^n$





$x^t Q y =< x, y >_Q$

$x^t D y =< x, y >_D$

Duality : Transposable data.

Data Analysis: Geometrical Approach

i. The data are p variables measured on n observations.

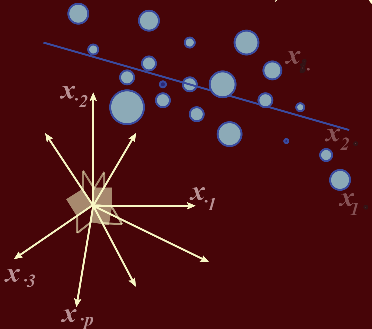
ii. X with n rows (the observations) and p columns (the variables).

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

$$Y^t DX = \Delta_Y A \qquad \text{where } \Delta_Y = Y^t DY = diag(w_1, w_2, \ldots, w_g),$$

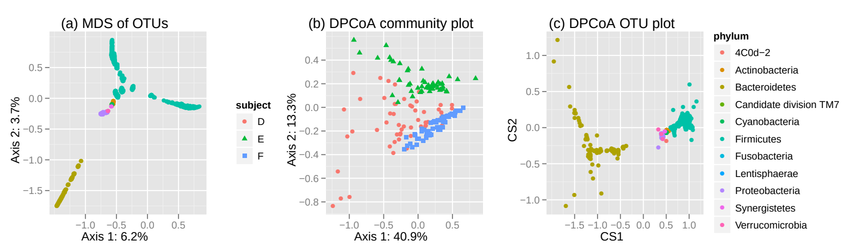
and $w_k = \sum_{i:y_{ik}=1} d_i$, the w_k's are the group weights, as they are the sums of the weights as defined by D for all the elements in that group.
Call T the matrix $T = X^t DX$, a generalized between group variance-covariance is $B = A^t \Delta_Y A$ and call the between group variance covariance the matrix $W = (X - YA)^t D(X - YA)$.
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: $B = -\frac{1}{2} H D \bullet D H$.
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 Δ giving the squares of the pairwise distances between the species on the tree (patristic). Then we can

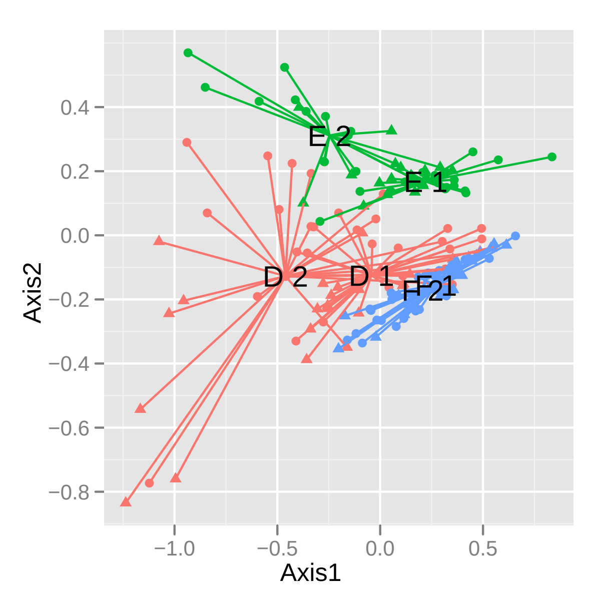
- Use the distances between species to find an embedding in $n - 1$ -dimensional space such that the euclidean distances between the species is the same as the distances between the species defined in Δ.
- Place each of the p locations at the barycenter of its species profile. The euclidean distances 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

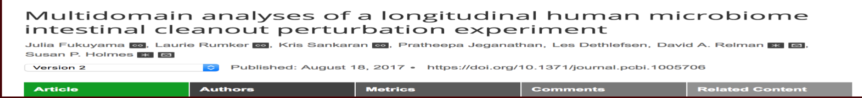
- DOI Trees: Navigate large trees according to the Degree-of-Interest (DOI) defined by clicking on different nodes.
- DOI Sankeys: Create a DOI Tree where abundances are split across several groups.
- Timeboxes: Visually query a (tree-structured) collection of time series, and see which nodes are associated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

`http://statweb.stanford.edu/~kriss1/antibiotic.html`.

Part IV

Adaptive gPCA using a Tree on Perturbation Data,

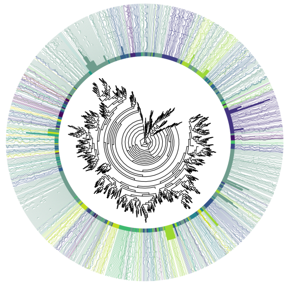
Fukuyama,Rumker,Sanakaran, et al., PLOS Comp Bio., 2017



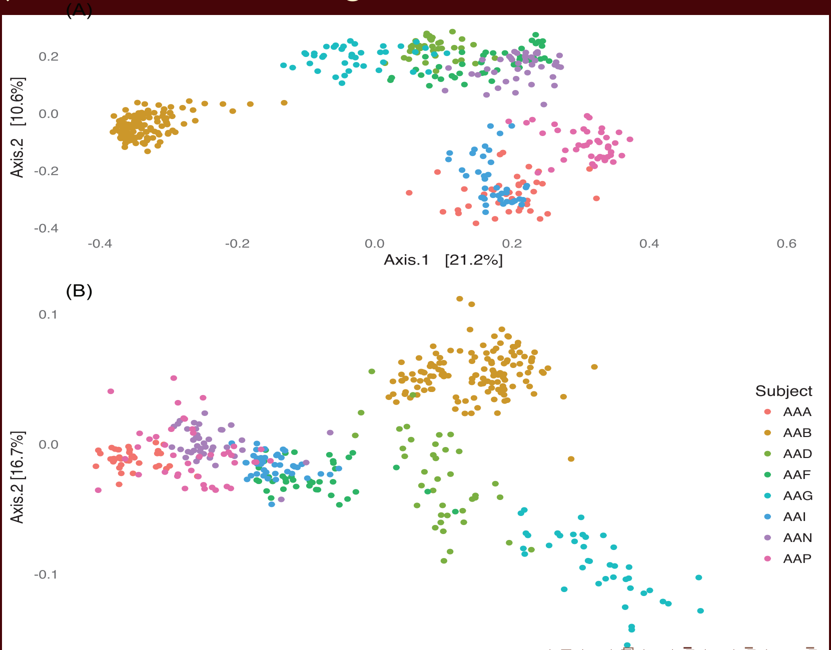
Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment

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

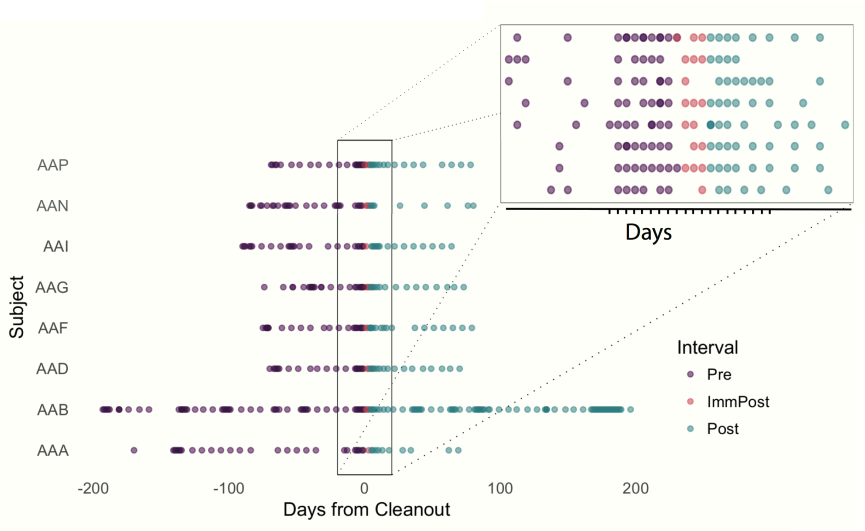
Version 2 | Published: August 18, 2017 • https://doi.org/10.1371/journal.pcbi.1005706



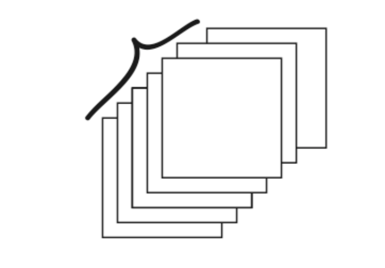
Subject effect is the strongest



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-quare 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*.

$$sum(x1 * y1 + x2 * y2 + x3 * y3)$$

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.

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

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

Sparse CCA, then PCA

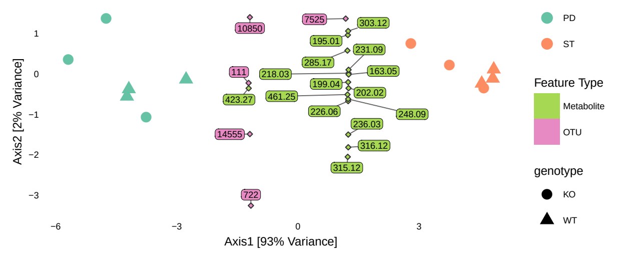
CCA: Canonical Correlation Analysis.
PCA: Principal Components Analysis.

- 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.

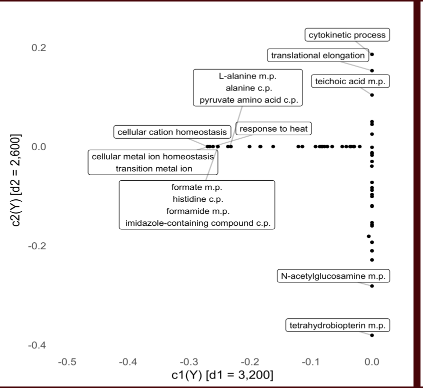


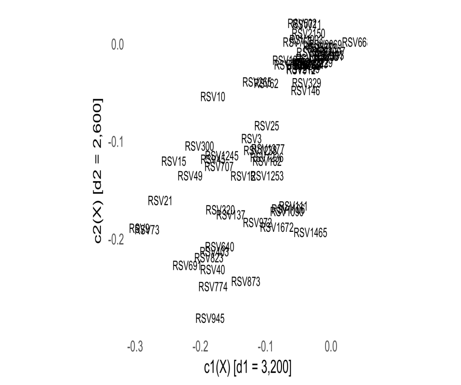
A PCA triplot produced from the CCA selected features in from multiple data types (metabolites and OTUs). Triangles for Knockout and circles for wild type. The main variation in the data is across PD and ST samples (different diets).

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.

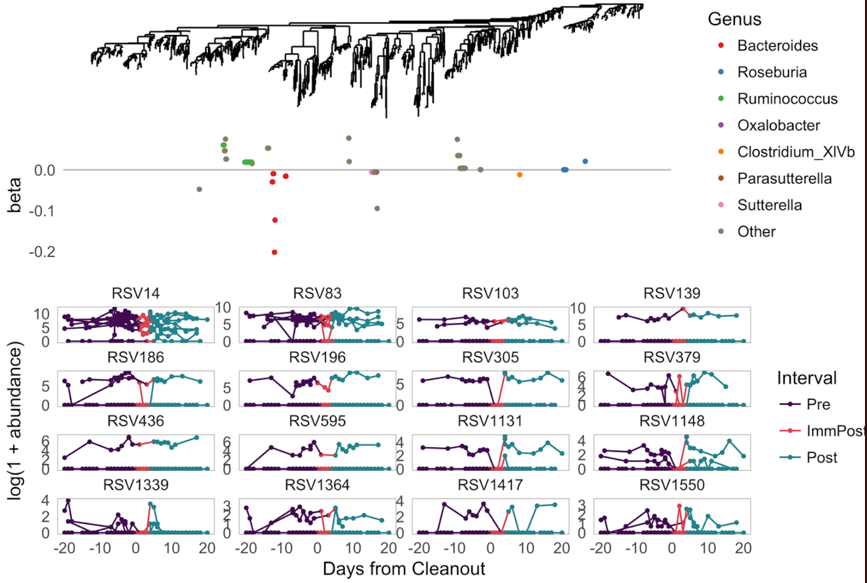
- Q Kernel : Q_{ij} represents shared ancestral branch length between species i and j.
- Covariance of a Brownian motion run along the branches of the tree.

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

- Inference using this prior regularizes towards this structure.

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

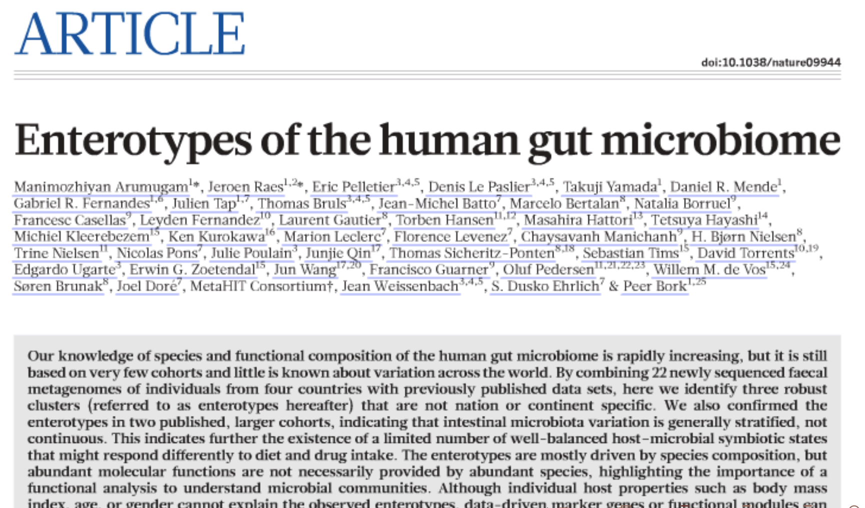
gPCA on $(X, S, \mathbb{I}_n)$
$\sigma_1/\sigma_2 \longrightarrow 0$ then PCA. $\sigma_2/\sigma_1 \longrightarrow 0$ then DPCoA.



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

Part V

An Example



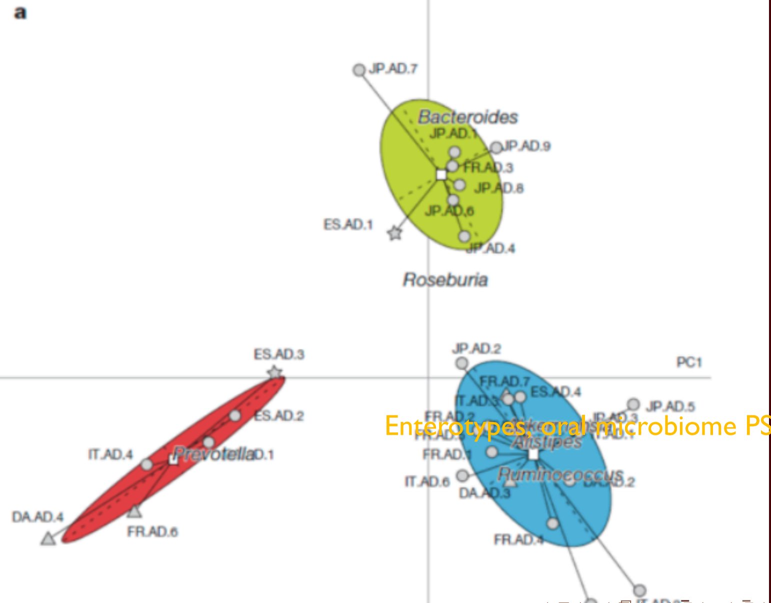
ARTICLE

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1*], Jeroen Raes[1,2*], Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]

Our knowledge of species and functional composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about variation across the world. By combining 22 newly sequenced faecal metagenomes of individuals from four countries with previously published data sets, here we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific. We also confirmed the enterotypes in two published, larger cohorts, indicating that intestinal microbiota variation is generally stratified, not continuous. This indicates further the existence of a limited number of well-balanced host–microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but abundant molecular functions are not necessarily provided by abundant species, highlighting the importance of a functional analysis to understand microbial communities. Although individual host properties such as body mass index, age, or gender cannot explain the observed enterotypes, data-driven marker genes or functional modules can



Enterotypes, oral microbiome PSB 2016.

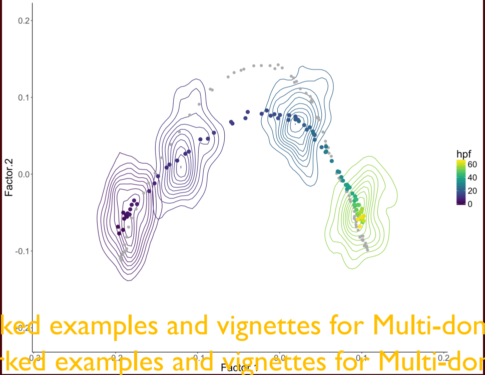
Summary of the study

- Choose the data transformation (here proportions replaced the original counts).
  ... log, rlog, subsample, prop, orig.
- Take a subset of the data, some samples declared as outliers. ... leave out 0, 1, 2 ,..,9, + criteria (10)......
- Filter out certain taxa (unknown labels, rare, etc...)
  ... remove rare taxa (threshold at 0.01%, 1%, 2%,...)
- Choose a distance.
  ... 40 choices in vegan/phyloseq.
- Choose an ordination method and number of coordinates.
  ... MDS, NMDS, k=2,3,4,5..
- Choose a clustering method, choose a number of clusters.
  ... PAM, KNN, density based, hclust ...
- Choose an underlying continuous variable (gradient or group of variables: manifold).
- Choose a graphical representation.

There are thus more than 200 million possible ways of analyzing this data:

$$5 \times 100 \times 10 \times 40 \times 8 \times 16 \times 2 \times 4 = 204800000$$

Resources and Workflows



http link 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

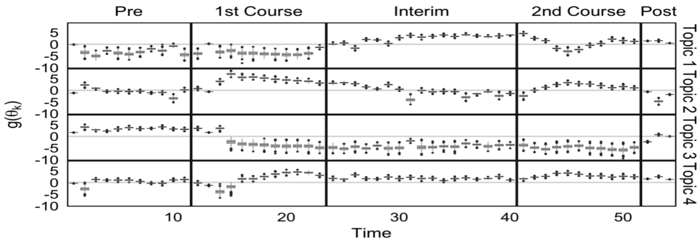


Figure 1: Boxplots represent approximate posteriors for estimated mixture memberships θ_d, and their evolution over time. Each row of panels provides a different sequence of θ_{dk} for a single k, and different columns distinguish different phases of sampling. Note that the y-axis is on the g-scale, which is defined as a translated logit, $g(\mathbf{p}) := \left(\log p_1 - \overline{\log \mathbf{p}}, \ldots, \log p_K - \overline{\log \mathbf{p}}\right)$. 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.

Kris Sankaran's Topic Page

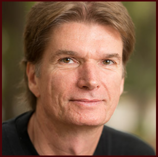
Benefitting from the tools and schools of Statisticians……

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

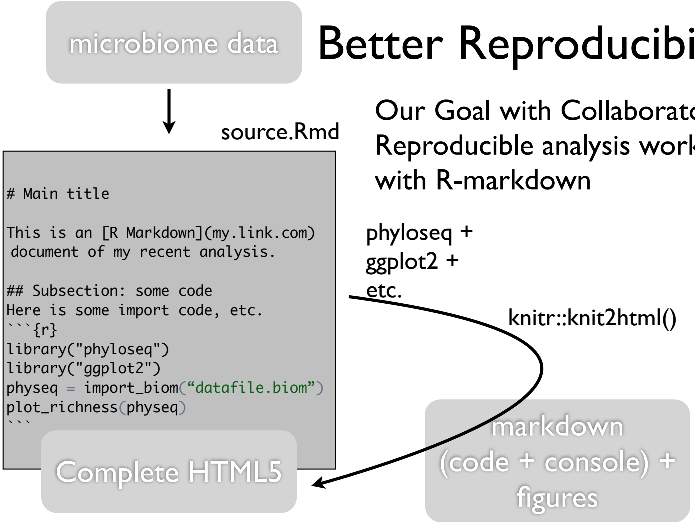
Postdoctoral Fellows Paul (Joey) McMurdie, Ben Callahan, Christof Seiler, Pratheepa Jeganathan

Students: John Cherian, Diana Proctor, Daniel Sprockett, Lan Huong Nguyen, Julia Fukuyama, Kris Sankaran, Claire Donnat.

Funding from NIH TR01 and NSF-DMS.

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)
```
~~~

phyloseq +
ggplot2 +
etc.

knitr::knit2html()

markdown
(code + console) +
figures

Complete HTML5

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

Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment.
Plos Computational Biology, 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: 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.

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