Day 2 – Tabular data

Morning: lectures & demos

9-10 Lecture: analysis & visualization of tabular data

10-12 Demo & hands-on: univariate methods

12-13 Lunch break

Afternoon: hands-on

13-14 Demo: multivariate data analysis & visualization

14-17 Hands-on: multivariate data analysis & visualization

17-18 Presentations & Discussion

Anatomy of taxonomic profiling data

- Tabular data properties

- Special properties of taxonomic profiling data

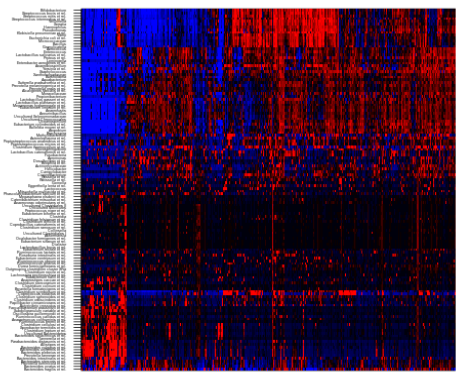
- Visualization and statistical summaries

”Omics” data

taxonomic abundance table

Omics in Oxford English Dictionary:
in cellular and molecular biology,
forming nouns with the sense
"all constituents considered collectively"

Features x samples



Genomics
Epigenomics
Microbiomics
Lipidomics
Proteomics
Glycomics
Foodomics
Transcriptomics
Metabolomics
Culturomics

Gut microbiota: 1000 western adults
(Lahti *et al.* Nature Comm. 2014)

Common study designs

Observational

Case-control

Intervention

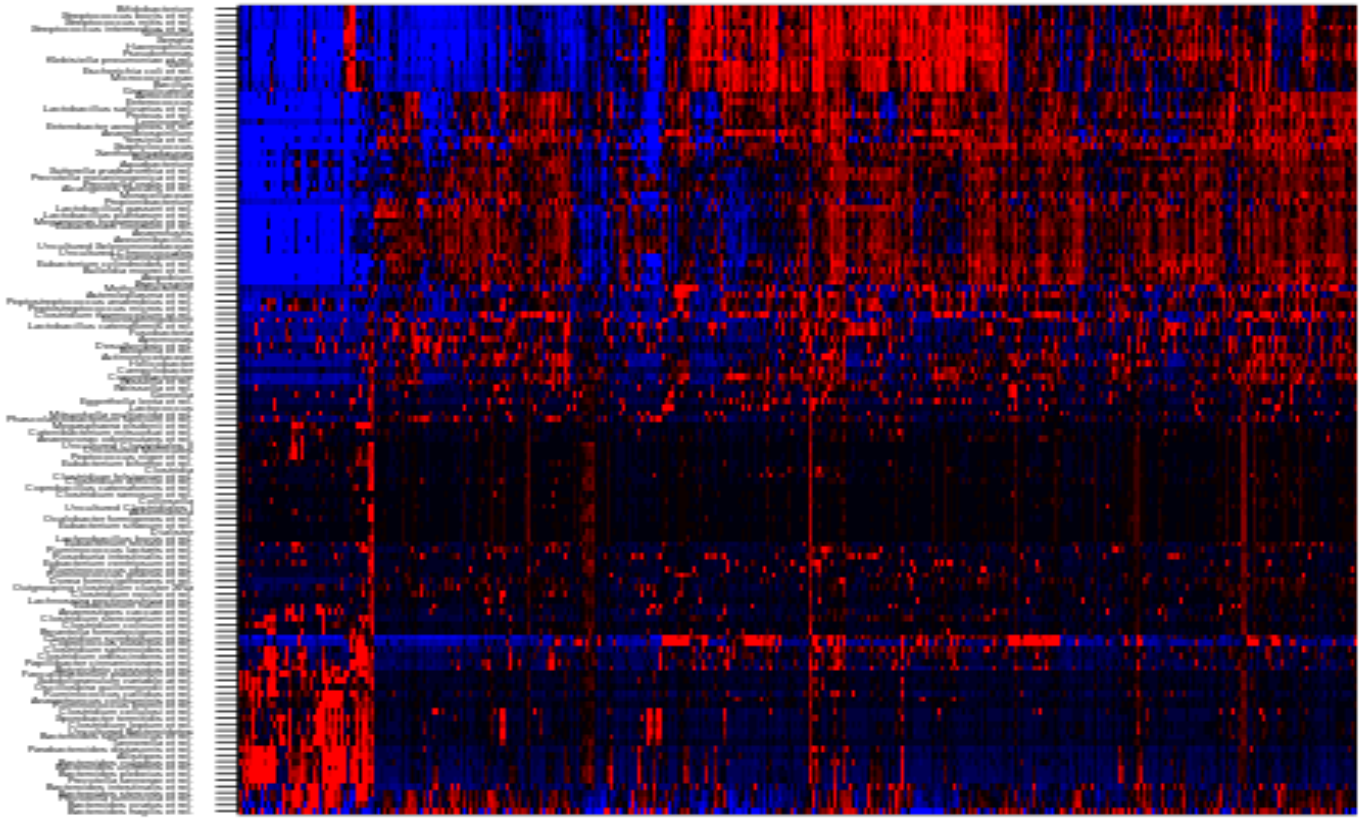
Cross-sectional

Prospective

Longitudinal

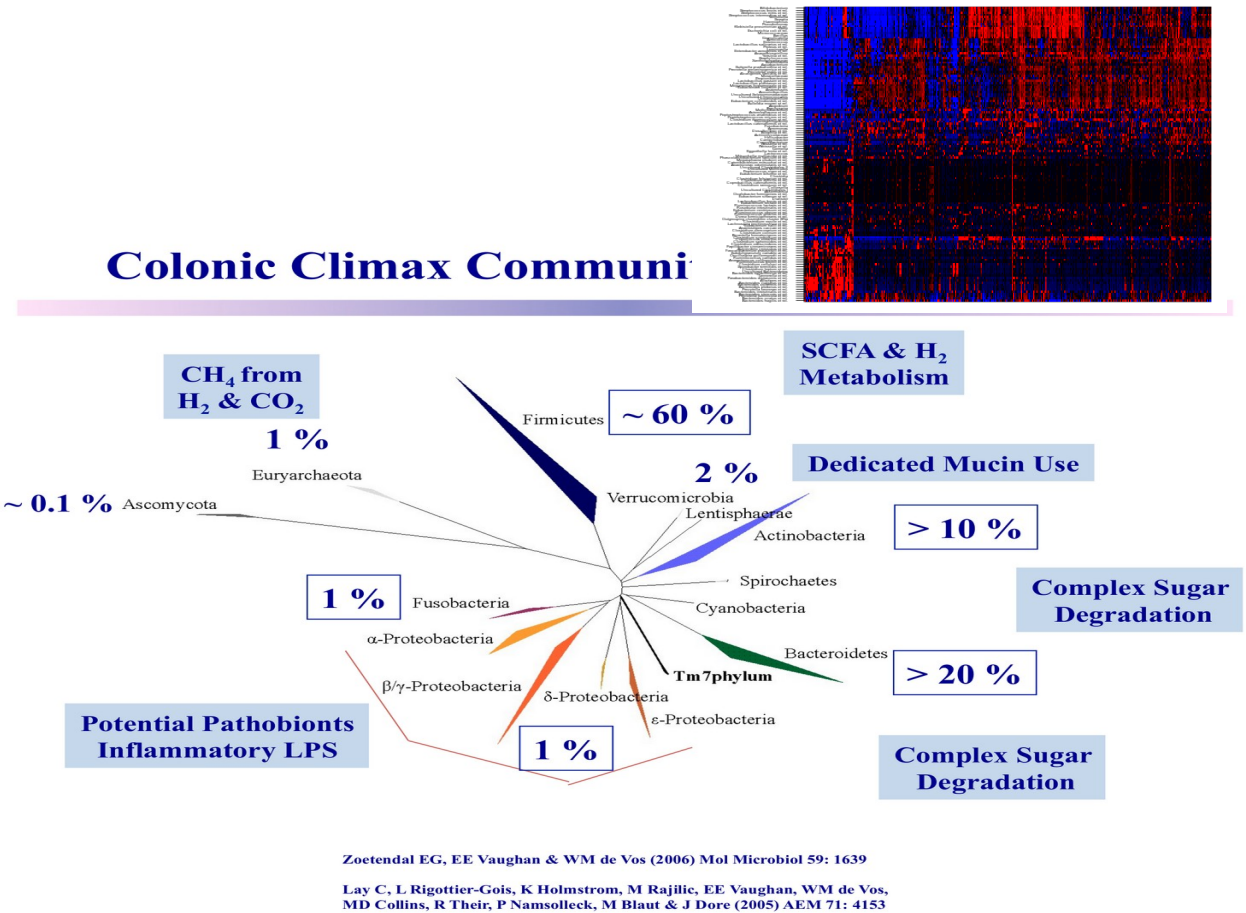
Organisms and samples are not independent

understanding & modeling the (latent) structure(s)



Special properties of microbiome data

- Sparse
- Compositional
- Non-Gaussian
- Overdispersed
- Discrete
- Complex
- Stochastic
- Multi-level



Zoetendal EG, EE Vaughan & WM de Vos (2006) Mol Microbiol 59: 1639

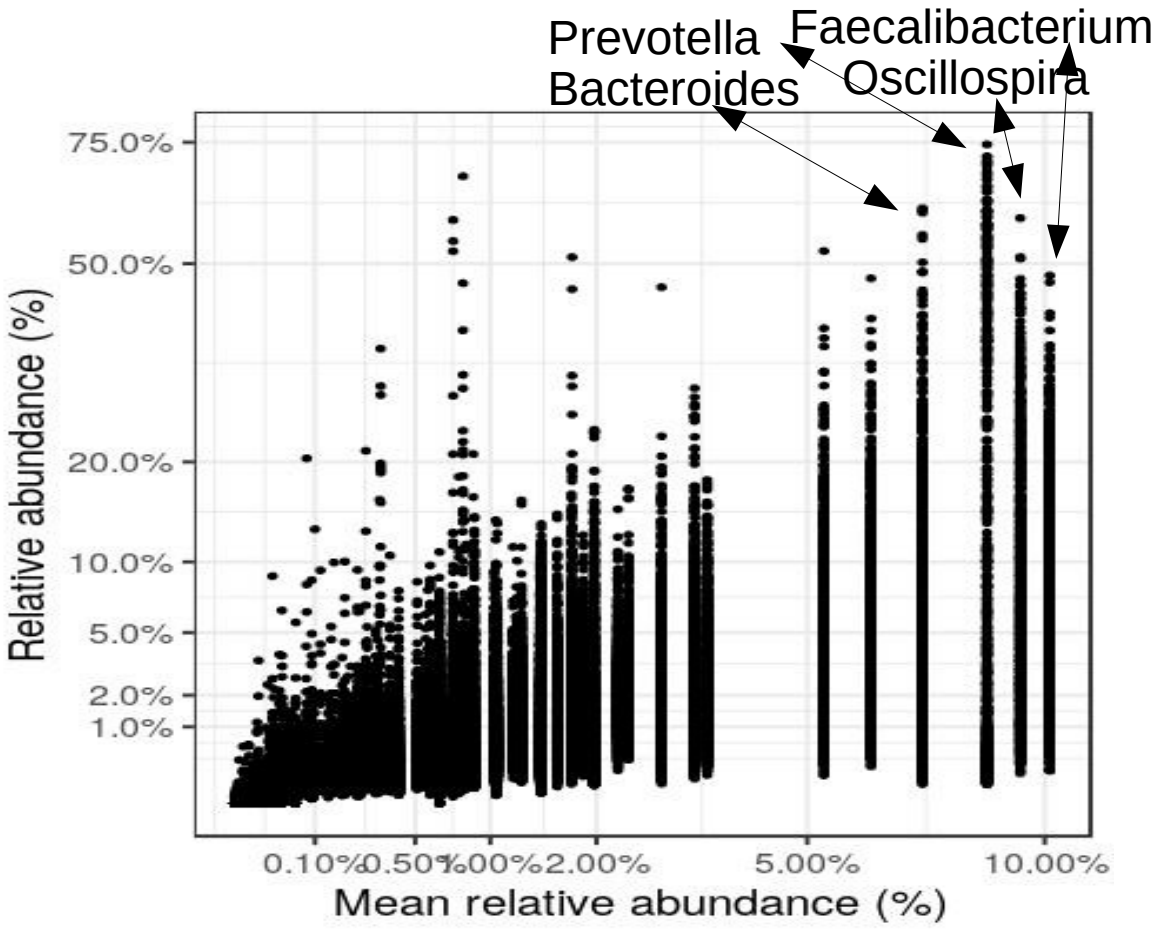
Lay C, L Rigottier-Gois, K Holmstrom, M Rajilic, EE Vaughan, WM de Vos, MD Collins, R Their, P Namsolleck, M Blaut & J Dore (2005) AEM 71: 4153

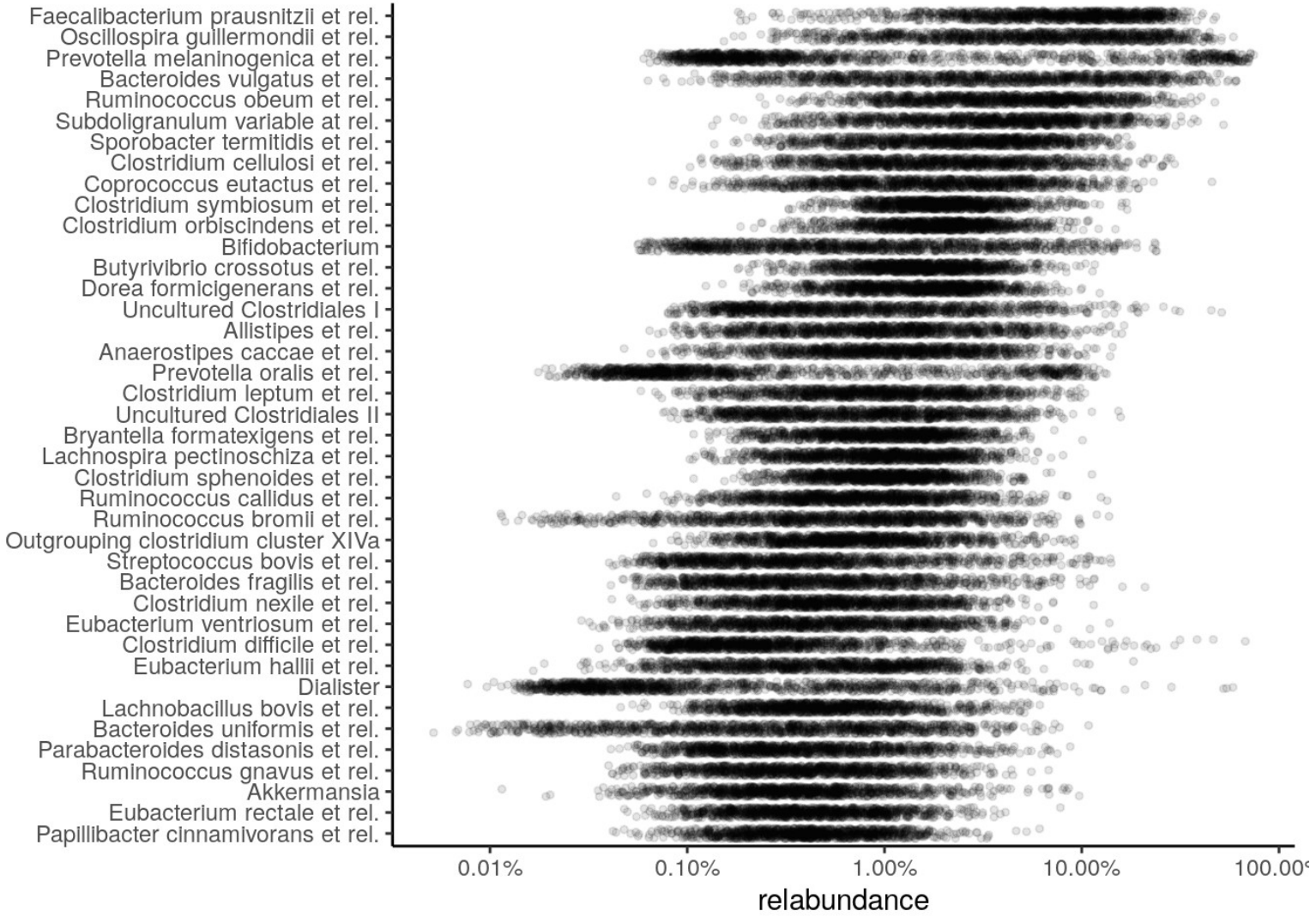
Taylor’s law (in HITChip Atlas)

Heteroschedasticity:
Variance increases with the mean

Overdispersion:
Variance increases faster than proposed by the model

Data: HITChip Atlas



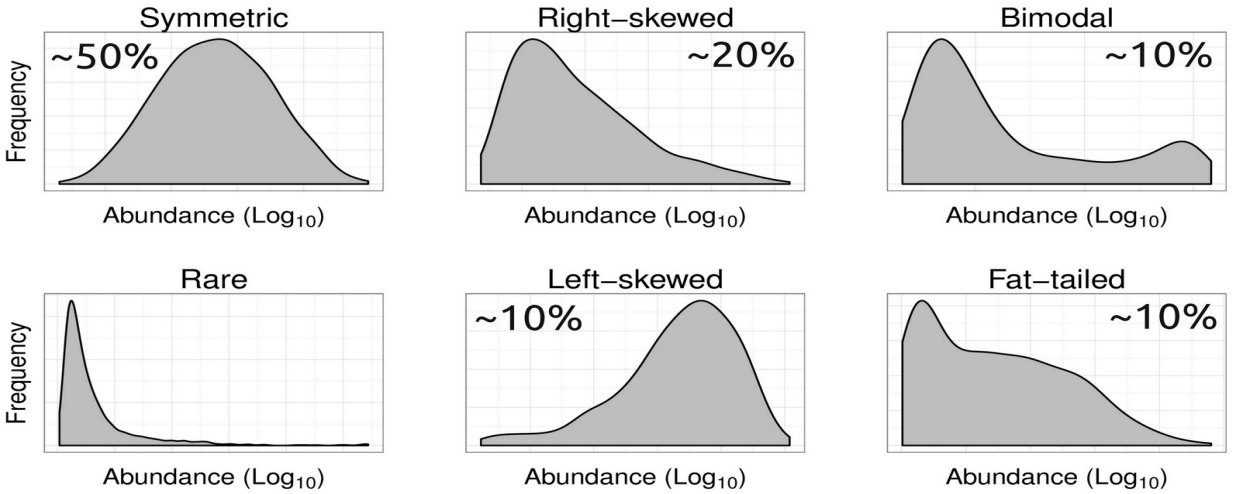

Faecalibacterium prausnitzii et rel.
Oscillospira guillermondii et rel.
Prevotella melaninogenica et rel.
Bacteroides vulgatus et rel.
Ruminococcus obeum et rel.
Subdoligranulum variable at rel.
Sporobacter termitidis et rel.
Clostridium cellulosi et rel.
Coprococcus eutactus et rel.
Clostridium symbiosum et rel.
Clostridium orbiscindens et rel.
Bifidobacterium
Butyrivibrio crossotus et rel.
Dorea formicigenerans et rel.
Uncultured Clostridiales I
Allistipes et rel.
Anaerostipes caccae et rel.
Prevotella oralis et rel.
Clostridium leptum et rel.
Uncultured Clostridiales II
Bryantella formatexigens et rel.
Lachnospira pectinoschiza et rel.
Clostridium sphenoides et rel.
Ruminococcus callidus et rel.
Ruminococcus bromii et rel.
Outgrouping clostridium cluster XIVa
Streptococcus bovis et rel.
Bacteroides fragilis et rel.
Clostridium nexile et rel.
Eubacterium ventriosum et rel.
Clostridium difficile et rel.
Eubacterium hallii et rel.
Dialister
Lachnobacillus bovis et rel.
Bacteroides uniformis et rel.
Parabacteroides distasonis et rel.
Ruminococcus gnavus et rel.
Akkermansia
Eubacterium rectale et rel.
Papillibacter cinnamivorans et rel.
0.01%
0.10%
1.00%
10.00%
relabundance

Differential abundance

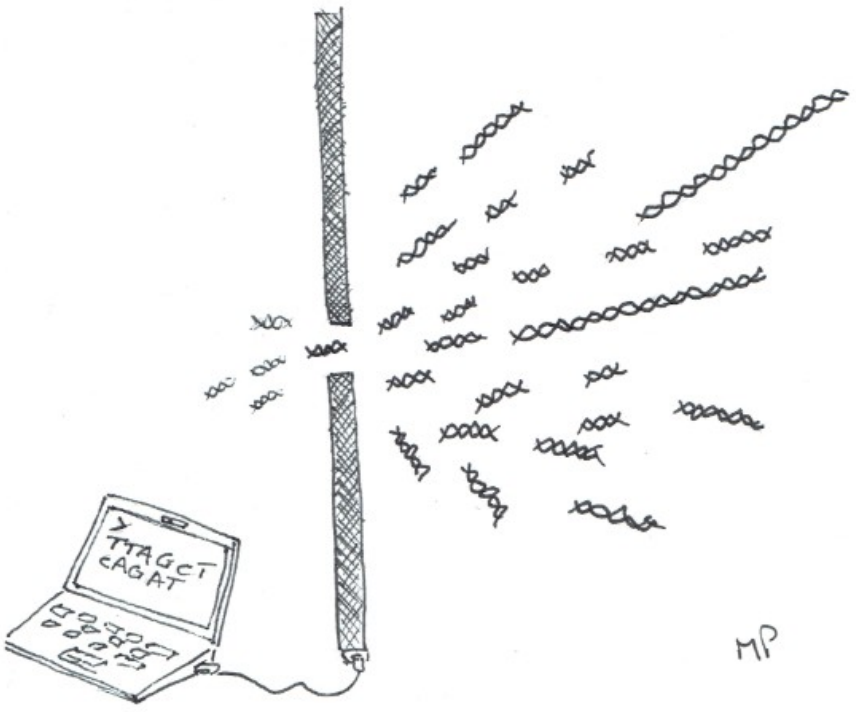
Standard t-test for two-group comparison?

Problems:

- Few replicates
- Non-gaussian, discrete, positive, skewed..
- Multiple testing



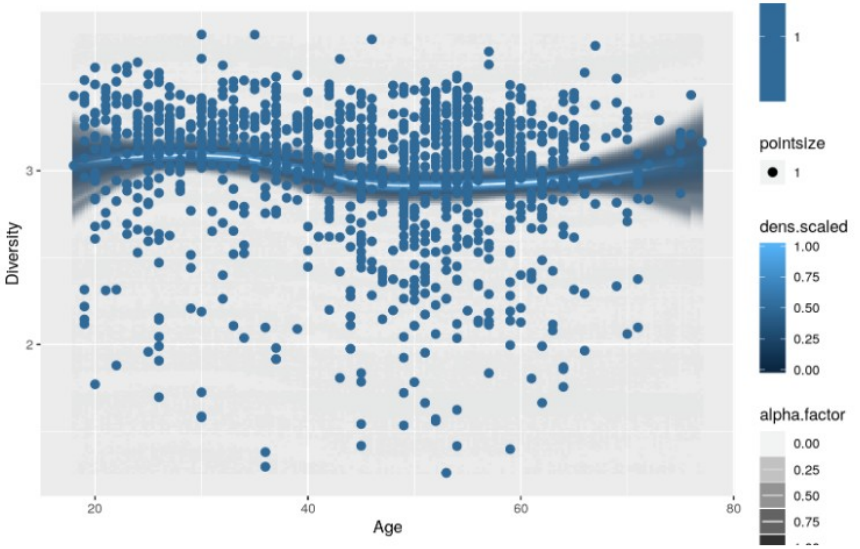
(barely) not statistically significant (p=0.052)
a barely detectable statistically significant difference (p=0.073)
a borderline significant trend (p=0.09)
a certain trend toward significance (p=0.08)
a clear tendency to significance (p=0.052)
a clear trend (p<0.09)
a clear, strong trend (p=0.09)
a considerable trend toward significance (p=0.069)
a decreasing trend (p=0.09)
a definite trend (p=0.08)
a distinct trend toward significance (p=0.07)
a favorable trend (p=0.09)
a favourable statistical trend (p=0.09)
a little significant (p<0.1)
a margin at the edge of significance (p=0.0608)
a marginal trend (p=0.09)
a marginal trend toward significance (p=0.052)
a marked trend (p=0.07)
a mild trend (p<0.09)
a moderate trend toward significance (p=0.068)
a near-significant trend (p=0.07)
a negative trend (p=0.09)
a nonsignificant trend (p<0.1)
a nonsignificant trend toward significance (p=0.1)

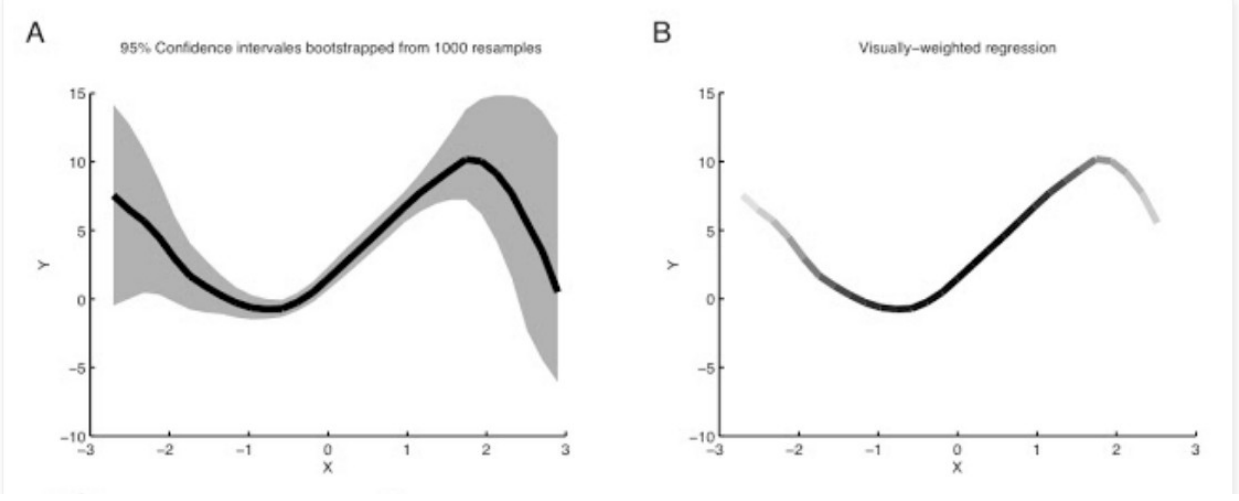


Visually-Weighted Regression

10 Pages • Posted: 17 May 2013

Solomon Hsiang
University of California, Berkeley; National Bureau of Economic Research





Altering the "visual weight" clearly communicates statistical confidence, even when readers are unfamiliar with the formal and abstract definitions of statistical uncertainty.

Color saturation, contrast of regression lines, and confidence intervals are parametrized by local measures of an estimate's variance.

Optimal container for microbiome data?

Multiple assays
seamless interlinking

Hierarchical data
supporting samples & features

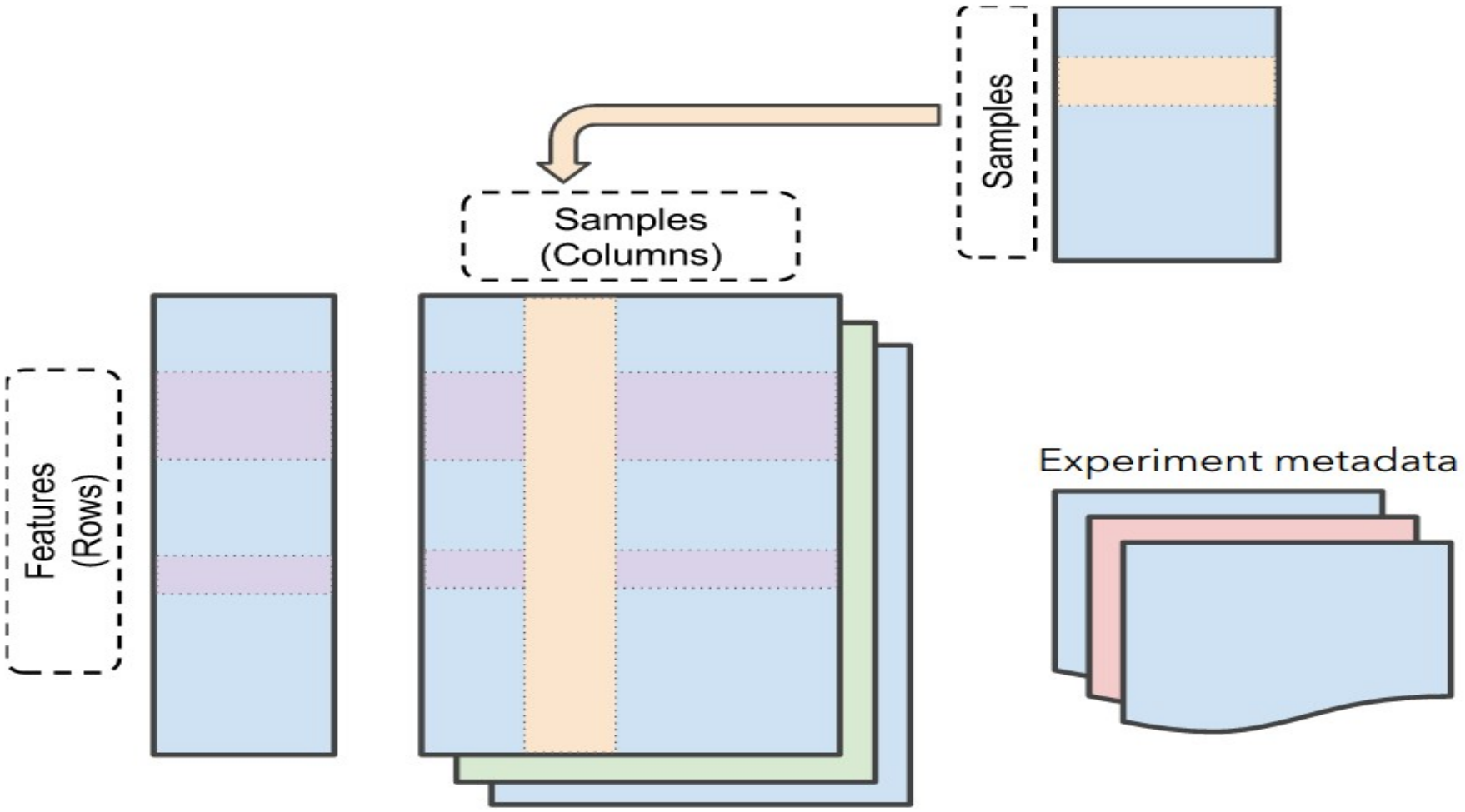
Side information
extended capabilities & data types

Optimized
for speed & memory

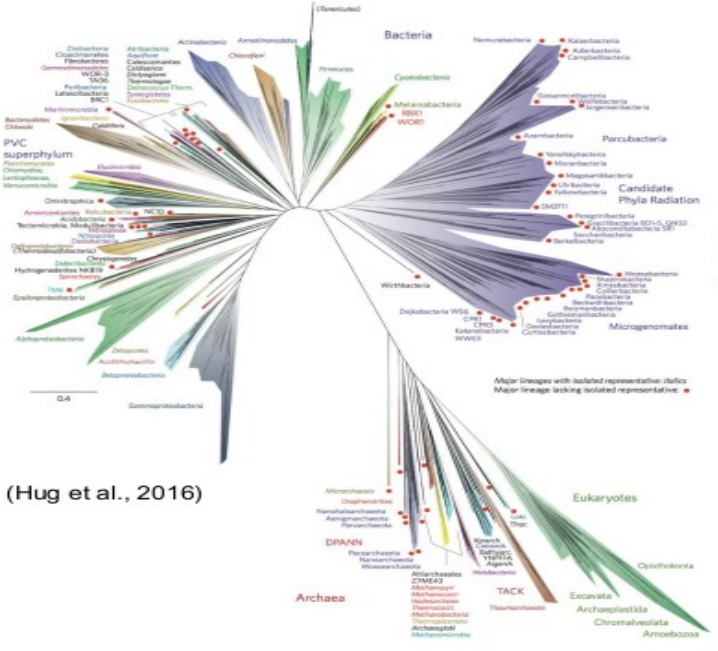
Integrated
with other applications & frameworks

Reduce overlapping efforts, improve interoperability, ensure sustainability.

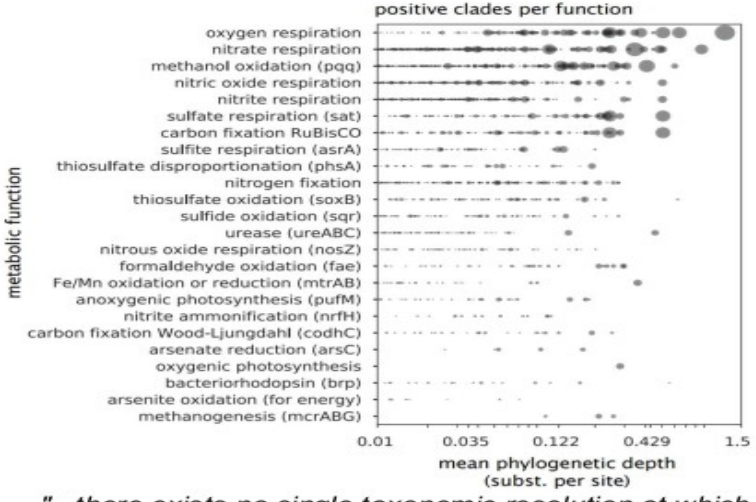
SummarizedExperiment

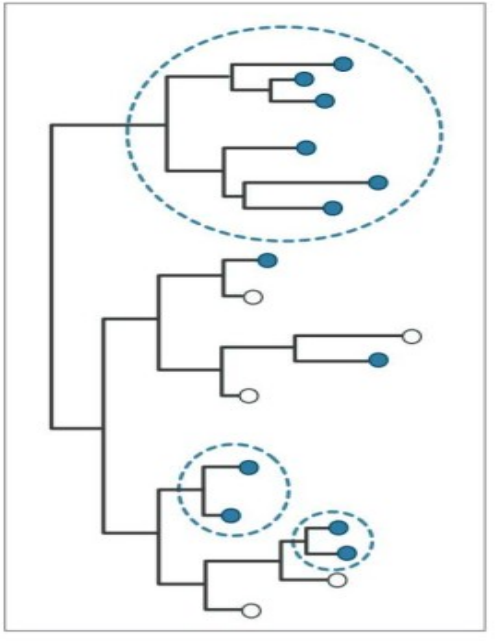


The use of phylogenetic information in metagenomics



(Hug et al., 2016)

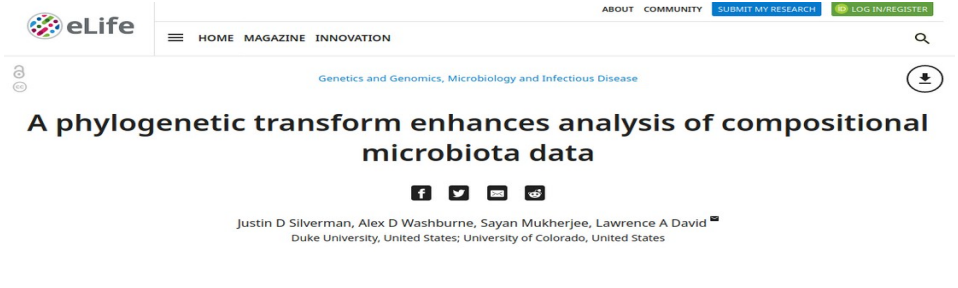




"...there exists no single taxonomic resolution at which taxonomic variation unambiguously reflects functional variation, and at which environmental selection of certain functions ... unambiguously translates to a selection of specific taxa."

(Louca et al., 2018)

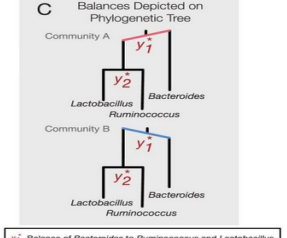
TURUN YLIOPISTO



Details for FLI and PhILR transform

$$\mathrm{FLI} = \left(\begin{array}{l} e^{0.953 \times \log_e(\mathrm{TG}) + 0.139 \times \mathrm{BMI} + 0.718 \times \log_e(\mathrm{GGT})} \\ + 0.053 \times \text{waist circumference} - 15.745 \end{array}\right) / \left(\begin{array}{l} 1 + e^{0.953 \times \log_e(\mathrm{TG}) + 0.139 \times \mathrm{BMI} + 0.718 \times \log_e(\mathrm{GGT})} \\ + 0.053 \times \text{waist circumference} - 15.745 \end{array}\right) \times 100$$

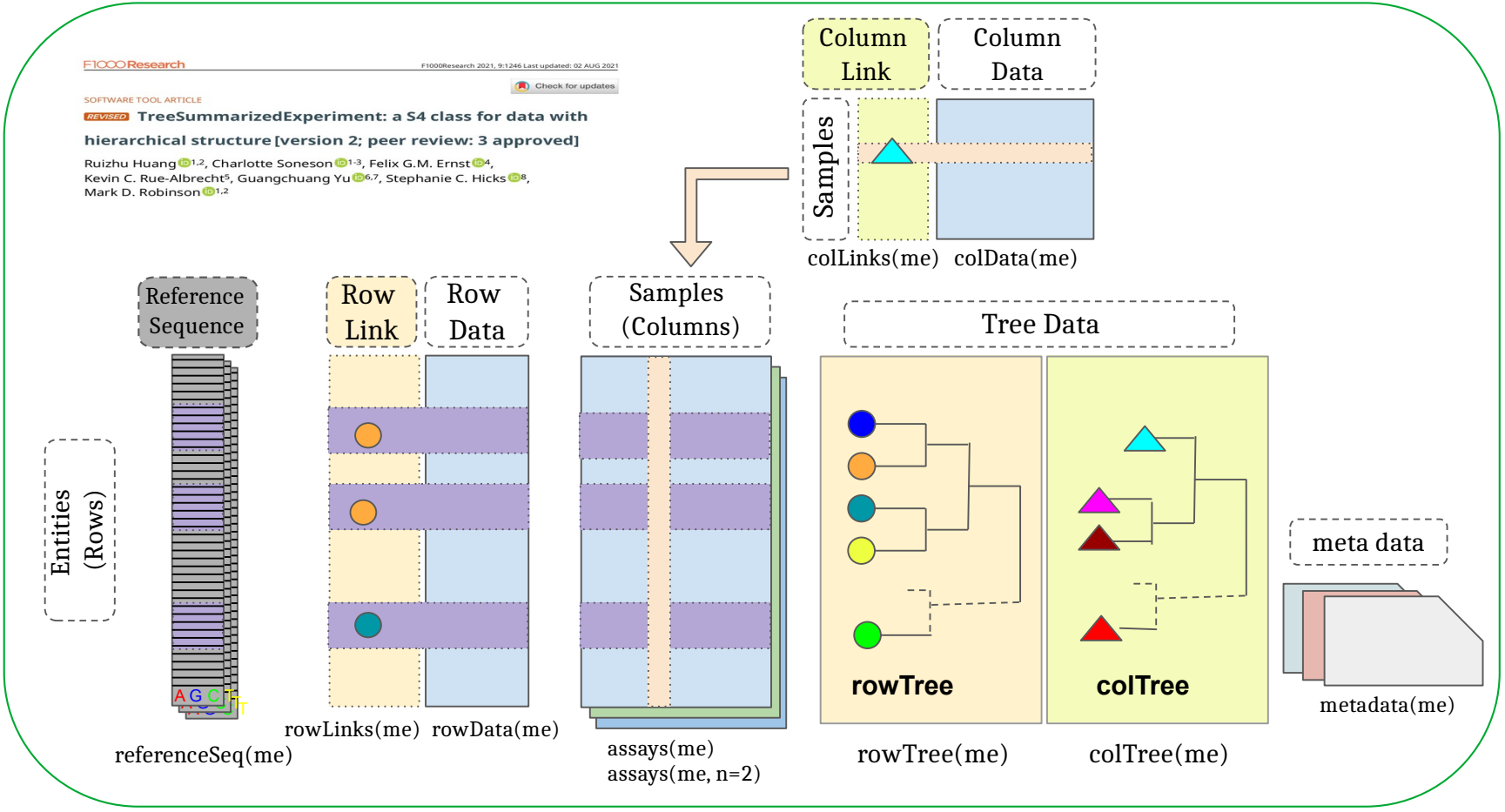
(Bedogni et al., 2006)



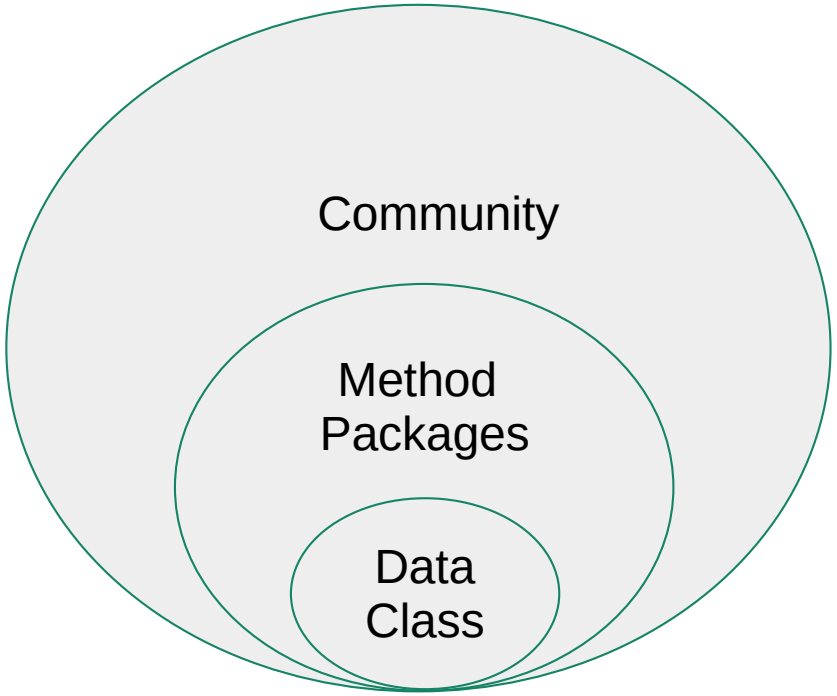
(Silverman et al., 2017)

TURUN YLIOPISTO

(Tree)SummarizedExperiment



Reduce overlapping efforts, improve interoperability, ensure sustainability.



Data packages

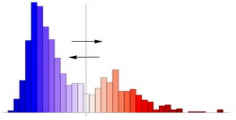
ExperimentHub

platforms all | rank 76 / 1974 | posts 2 / 1 / 2e+01 / 1 | in Bioc 4 years
build ok | updated before release | dependencies 72

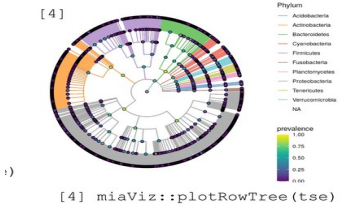
DOI: 10.18129/B9.bioc.ExperimentHub

mia – microbiome analysis

getDiversity(x)
calculateDMM(x)

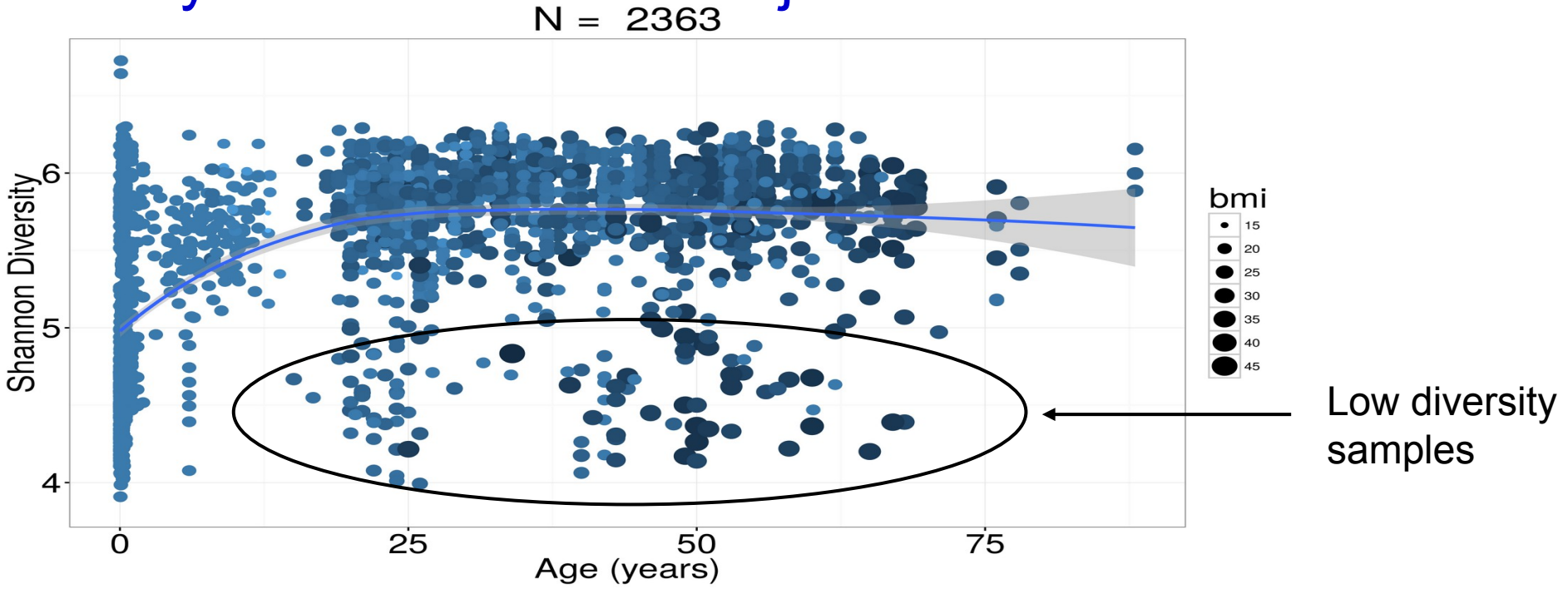


miaViz - Visualization

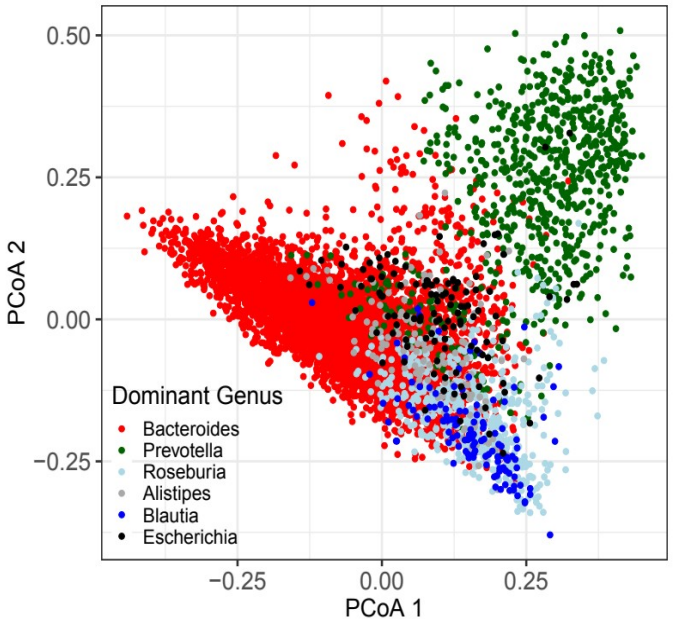


Package ecosystem

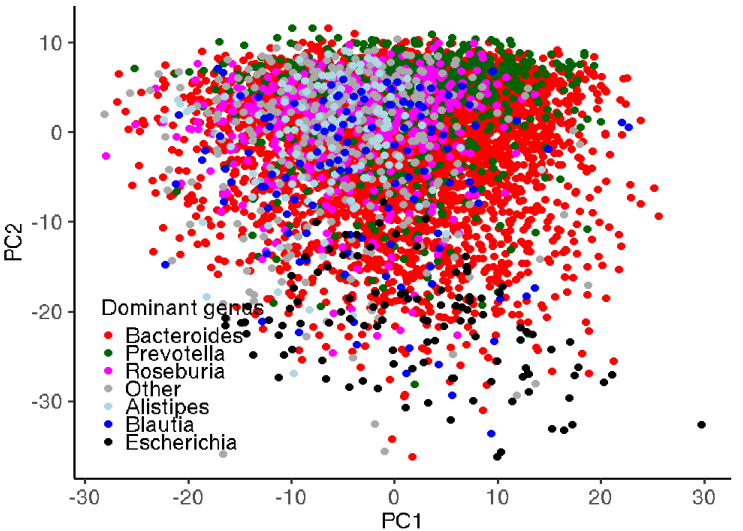
Alpha diversity & aging
healthy & normal obese subjects



PCoA + Bray-Curtis



PCA + Aitchison



Reproducible Research: Enterotype Example

Susan Holmes and Joey McMurdie

http://statweb.stanford.edu/~susan/papers/EnterotypeRR.html

Comment on this paper

Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota

Aaro Salosensaari, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa, Rob Knight, Leo Lahti, Teemu Niiranen

doi: https://doi.org/10.1101/2019.12.30.19015842

Fundamental considerations in beta diversity analysis

Feature selection

(all/core taxa; genus/strain level..?)

Transformation

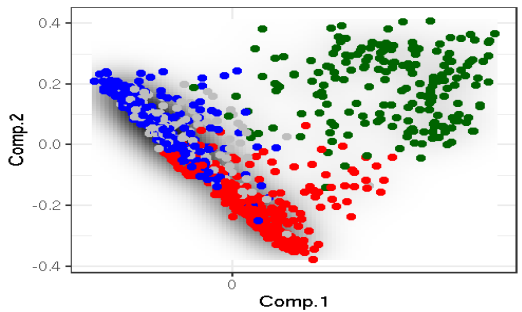
(absolute, compositional, CLR, Hellinger..?)

Dissimilarity measure

(Euclidean/L2, Bray-Curtis, Unifrac..?)

Analysis method

(PCA, PCoA, NMDS, t-SNE, UMAP..)



State diagnosis & manipulation: from specific targets to the overall ecosystem

Diet

Life style

Antibiotics

Probiotics

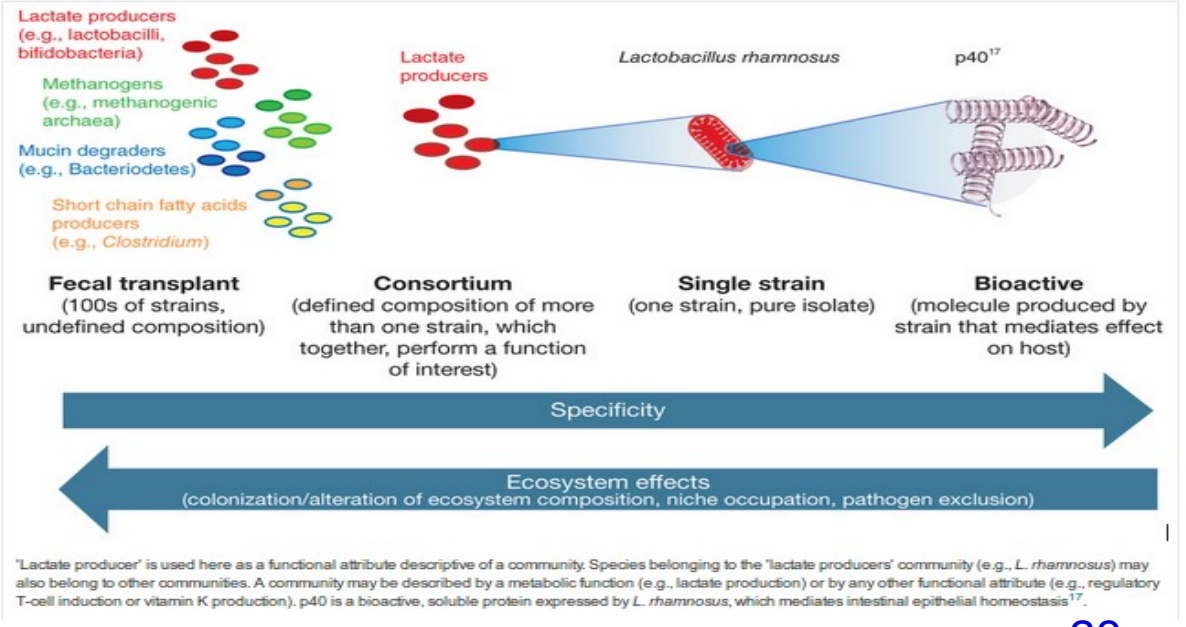
Prebiotics

Fecal transplants

Figure 3: Spectrum of microbiome-derived modulators being pursued by biotech companies, ranging from ecosystem-level interventions to single-target approaches.

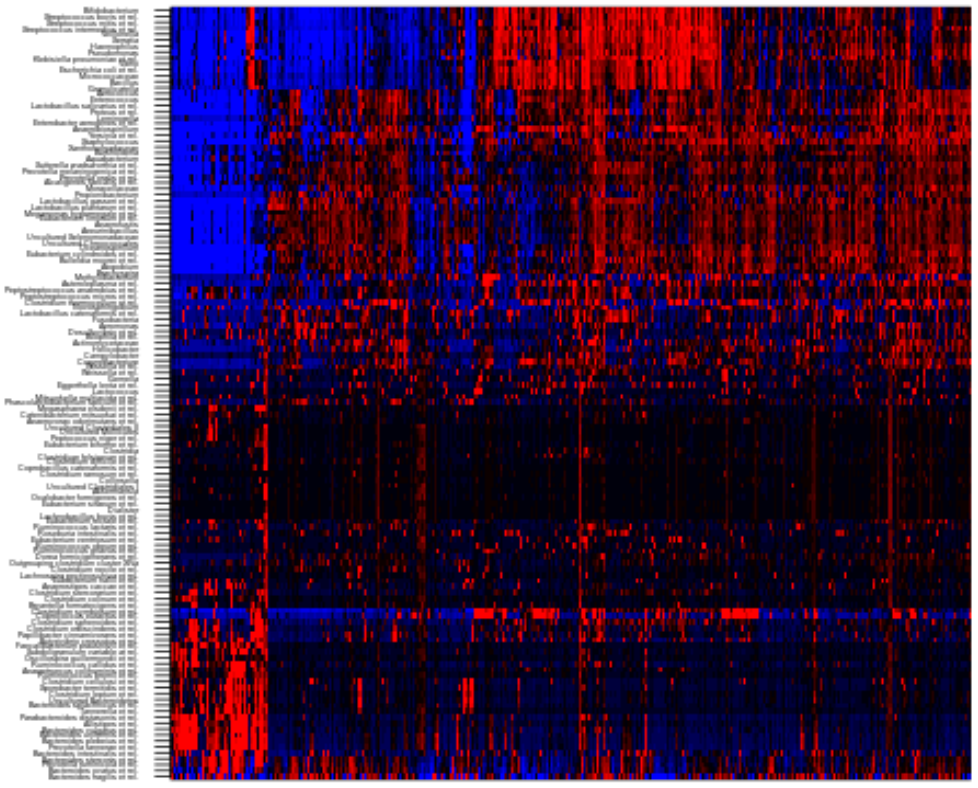
From
Medicines from microbiota
Bernat Olle
Nature Biotechnology **31**, 309–315 (2013) doi:10.1038/nbt.2548

Figure 3: Spectrum of microbiome-derived modulators being pursued by biotech companies, ranging from ecosystem-level interventions to single-target approaches.



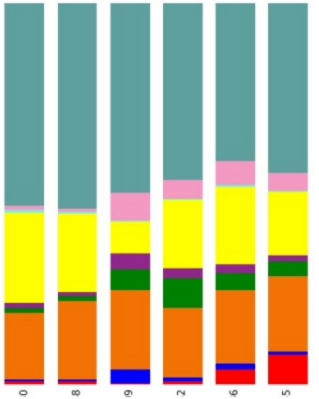
'Lactate producer' is used here as a functional attribute descriptive of a community. Species belonging to the 'lactate producers' community (e.g., *L. rhamnosus*) may also belong to other communities. A community may be described by a metabolic function (e.g., lactate production) or by any other functional attribute (e.g., regulatory T-cell induction or vitamin K production). p40 is a bioactive, soluble protein expressed by *L. rhamnosus*, which mediates intestinal epithelial homeostasis[17].

Aggregation & transformations

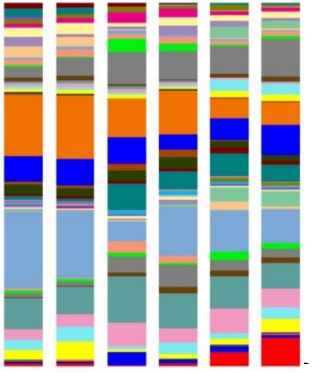


HITChip Atlas, WUR / Lahti & de Vos

Phylum level



Genus level



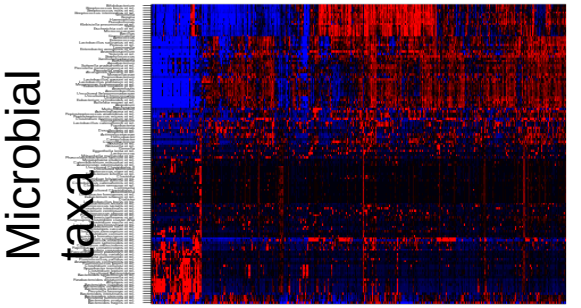
Core & prevalence

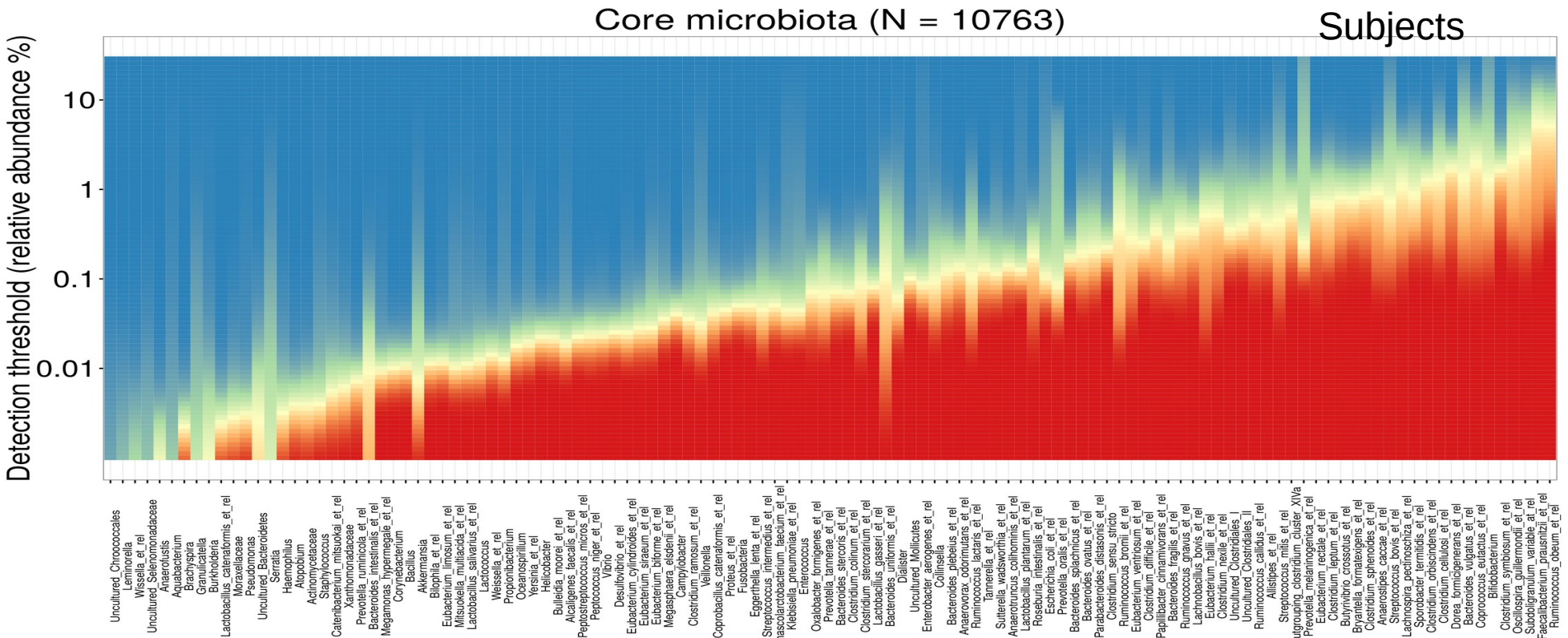
prevalentTaxa()

Core microbiota

only few species are prevalent (shared)

in population at a high abundance



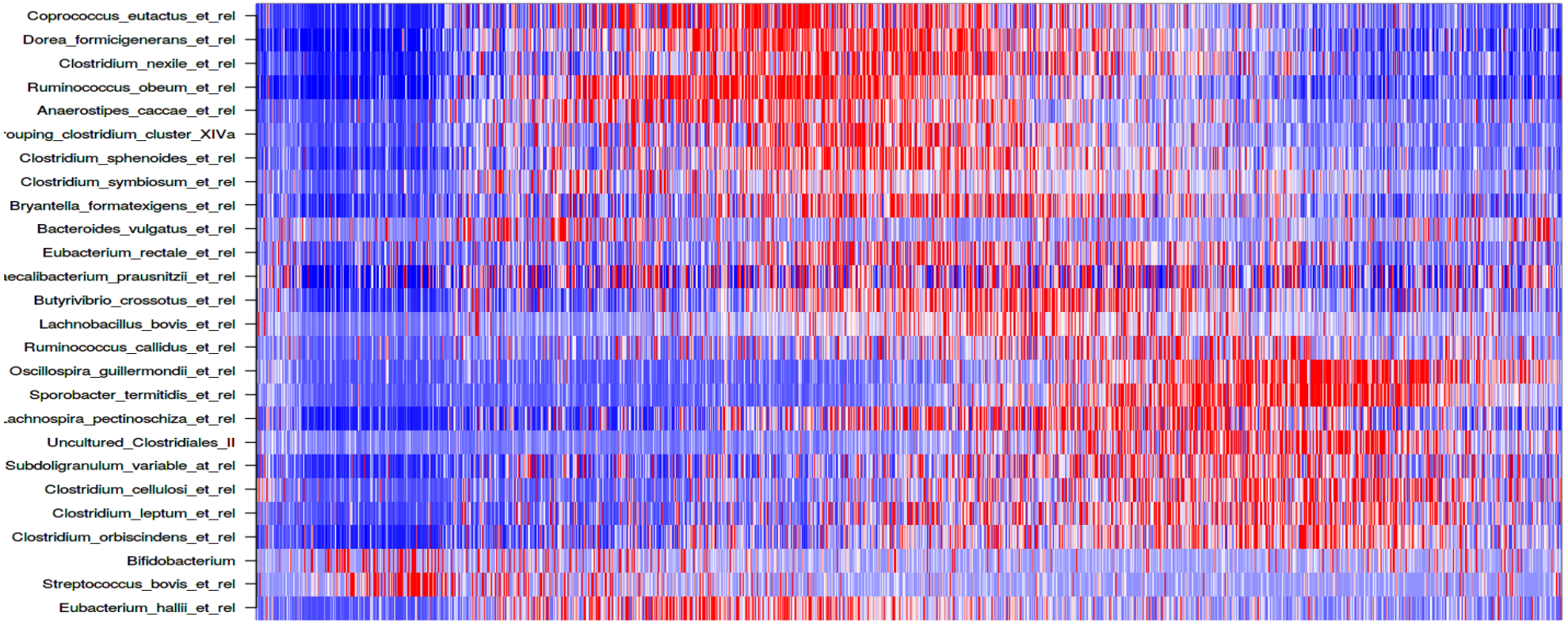


Data: HITChip Atlas

Core microbiota variation (N = 5005)

Z-score across subjects: red – high abundance & blue – low abundance

Core microbiota shows remarḱable variation across population.



Mining of Microbial Wealth and MetaGenomics pp 133-142 | Cite as

Rare Biosphere in Human Gut: A Less Explored Component of Human Gut Microbiota and Its Association with Human Health

Authors | Authors and affiliations

Shrikant S. Bhute, Saroj S. Ghaskadbi, Yogesh S. Shouche

The ISME Journal
Multidisciplinary Journal of Microbial Ecology

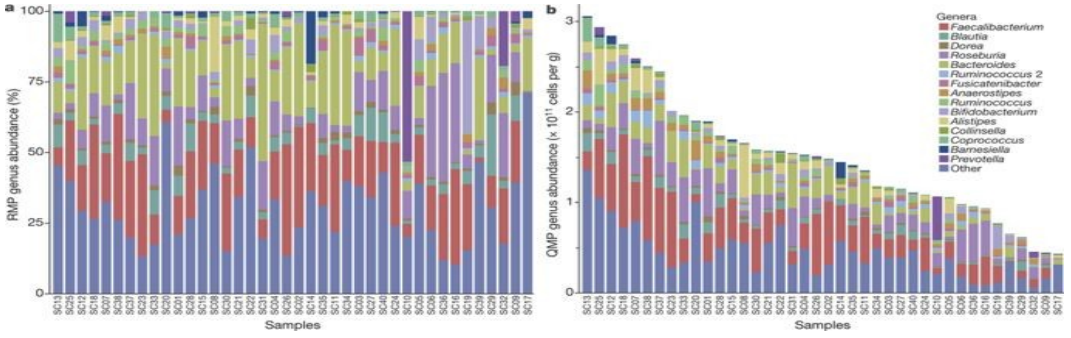
Mini Review | Open Access | Published: 10 January 2017

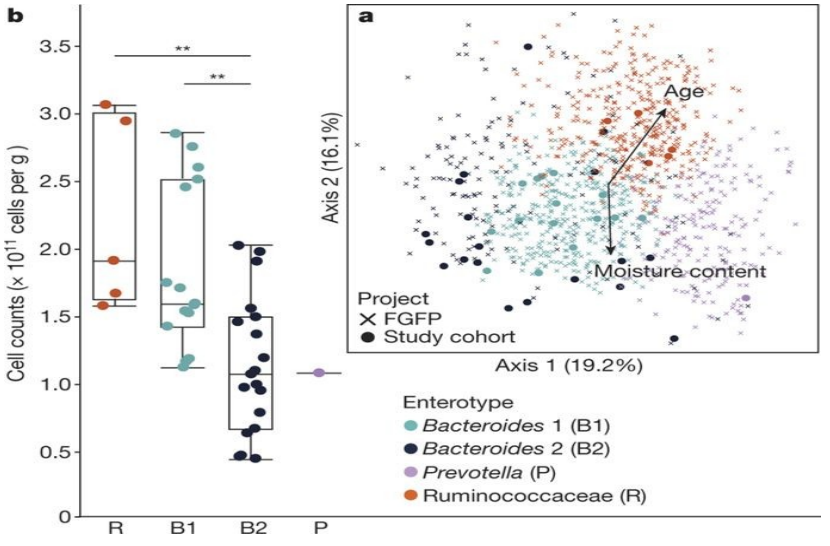
Where less may be more: how the rare biosphere pulls ecosystems strings

Alexandre Jousset, Christina Bienhold, Antonis Chatzinotas, Laure Gallien, Angélique Gobet, Viola Kurm, Kirsten Küsel, Matthias C Rillig, Damian W Rivett, Joana F Salles, Marcel G A van der Heijden, Noha H Youssef, Xiaowei Zhang, Zhong Wei & W H Gera Hol

The ISME Journal **11**, 853–862(2017) | Cite this article

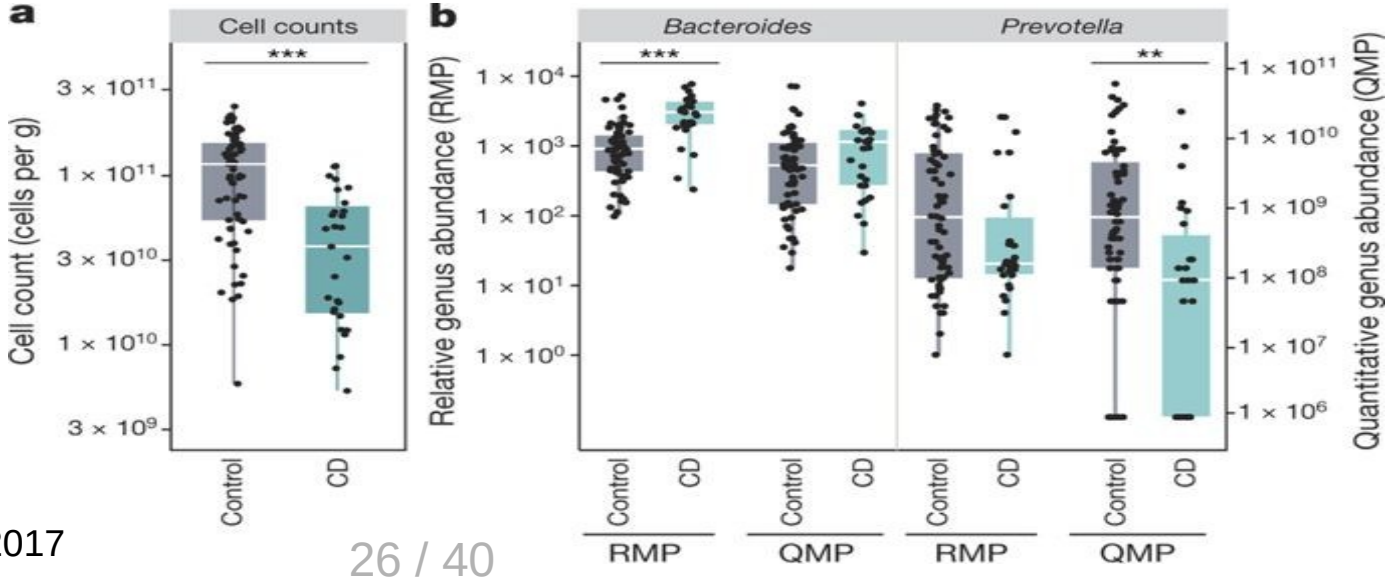
Relative versus absolute abundance: quantitative microbiome profiling





RMP vs. QMP:

drastic effect on conclusions!



Vandeputte et al. Nature 551:507-511, 2017

$$\mathrm{clr}(\mathrm{x}) = \left[\ln \frac{x_i}{g(\mathrm{x})}; \ldots; \ln \frac{x_D}{g(\mathrm{x})}\right]$$

$$\mathrm{alr}(\mathrm{x}) = \left[\ln \frac{x_i}{x_D}; \ldots; \ln \frac{x_{D-1}}{x_D}\right]$$

Open Access | Published: 05 May 2017

Multi-stability and the origin of microbial community types

Didier Gonze, Leo Lahti, Jeroen Raes & Karoline Faust

The ISME Journal **11**, 2159–2166(2017) | Cite this article

Many mechanisms underlay community assembly

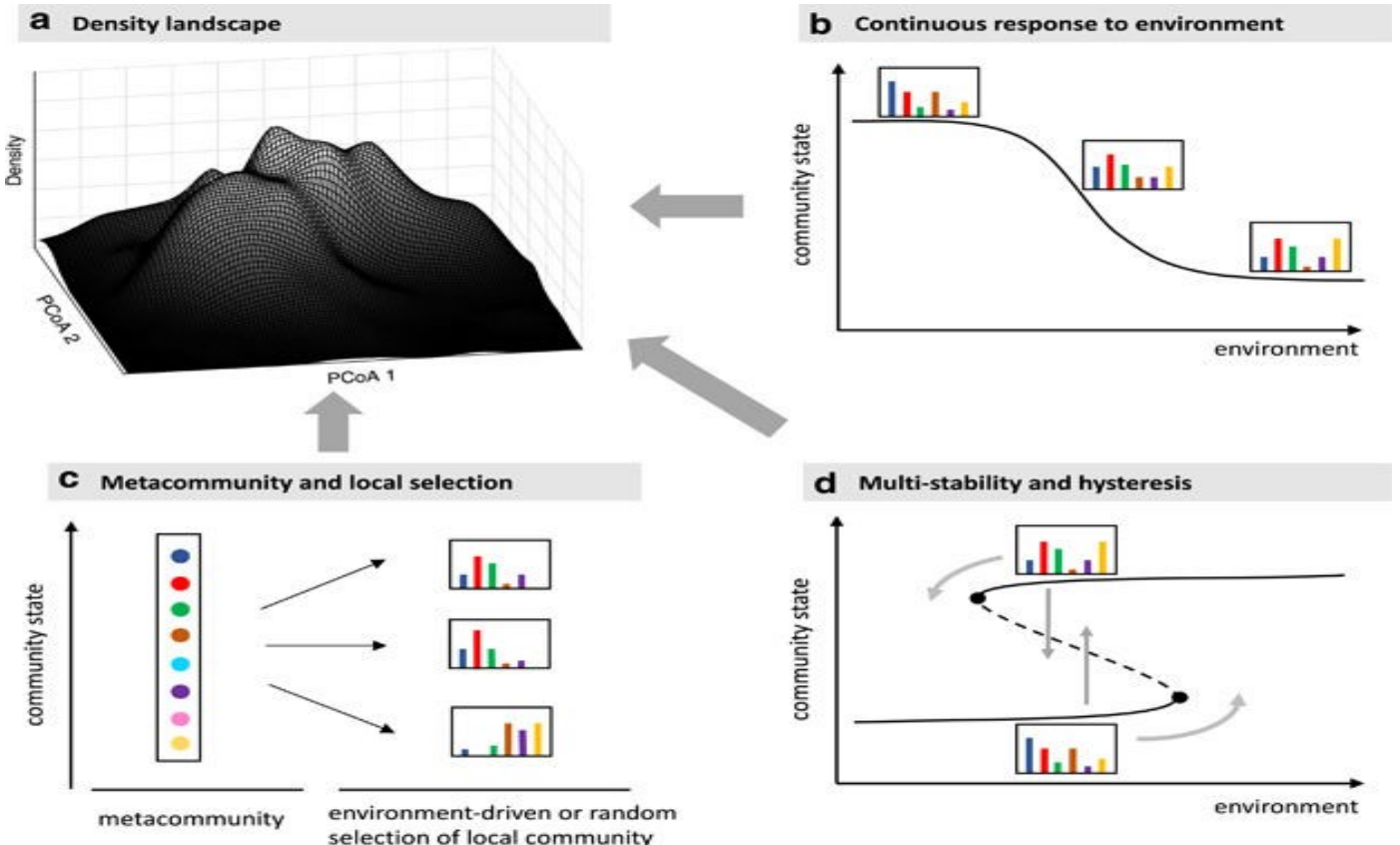
External perturbations (push & pulse)

Internal dynamics and multi-stability

Immigration

Stochasticity

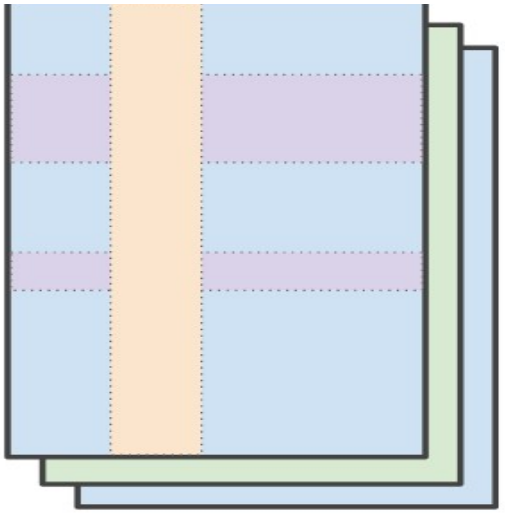
Memory



Feature selection task

Try ordination (PCA/PcoA) with..
- subset of taxa → e.g. tse[1:10,]
- higher taxonomic levels.. (after agglomerateByRanks or ***splitByRanks***..)
- for prevalent taxa (see prevalentTaxa()..)
- for dominant taxa (see dominantTaxa()..)
- with transformations (counts, relabundance, clr..)
- different (dis)similarities (Euclid, Bray-Curtis, Jaccard, other..?)

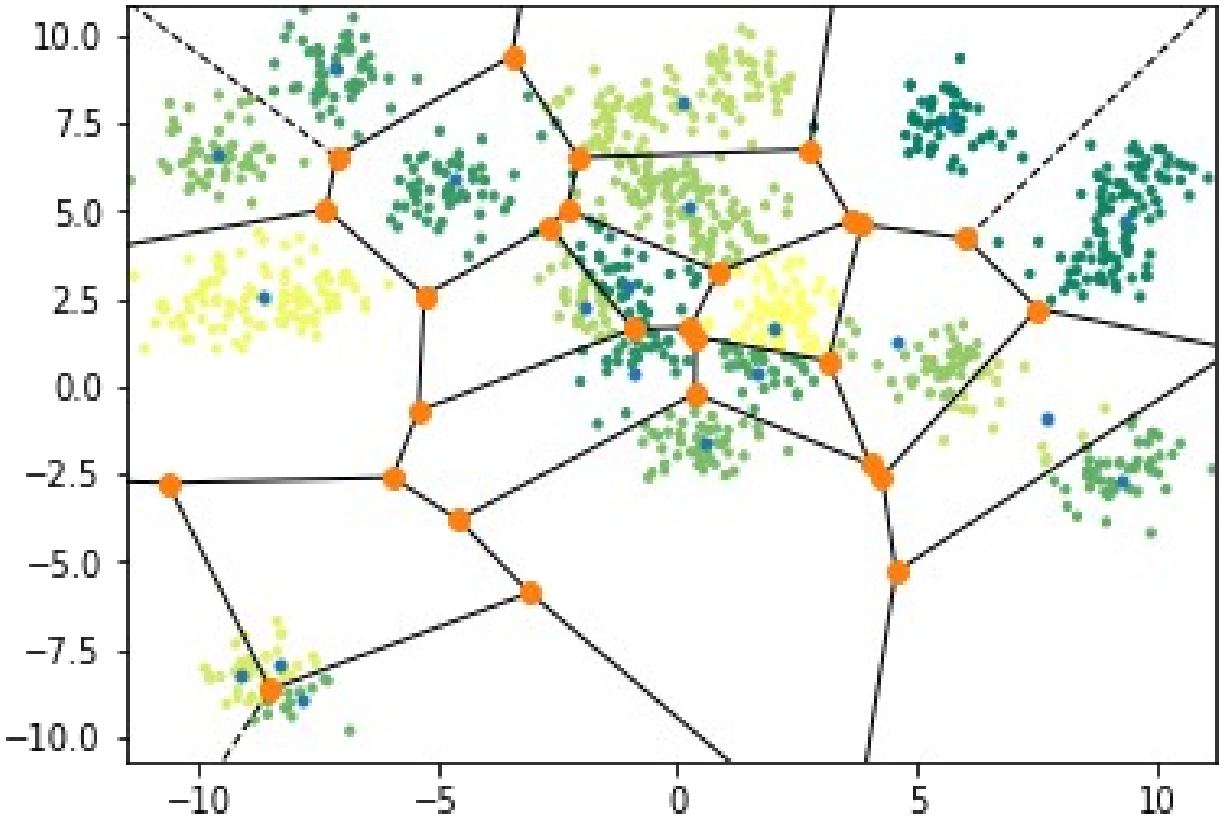
altExps(tse) <- splitByRanks(tse)



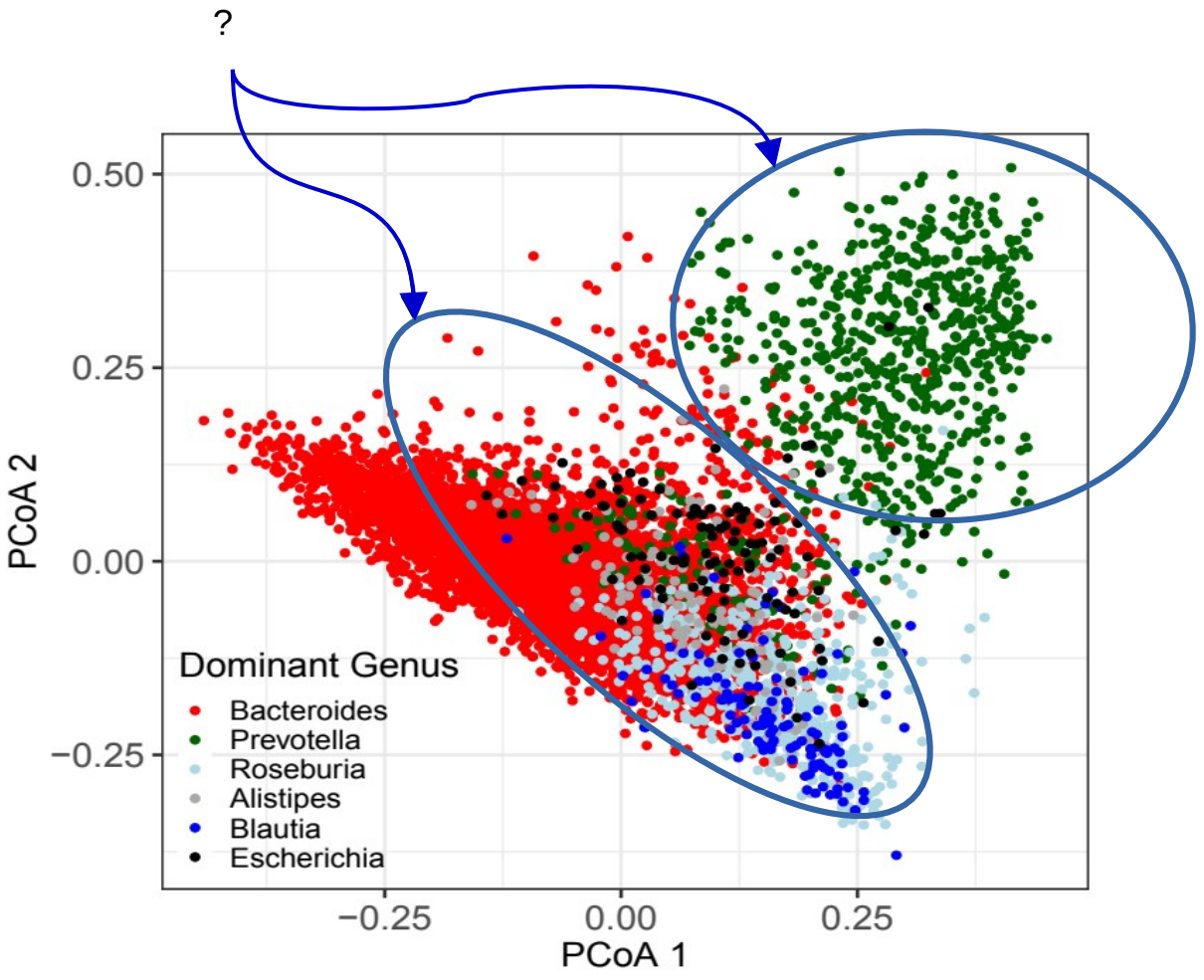
Option	Rows (features)	Cols (samples)	Recommended
assays	match	match	Data transformations
altExp	free	match	Alternative experiments
MultiAssay	free	free (mapping)	Multi-omic experiments

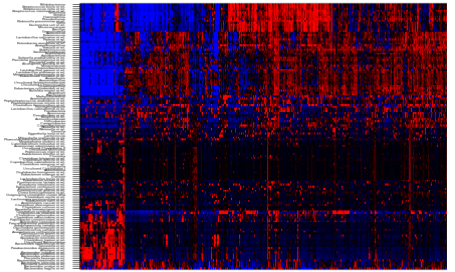
Non-parametric clustering:
Voronoi regions

The task:
find *centroids that describe the data*

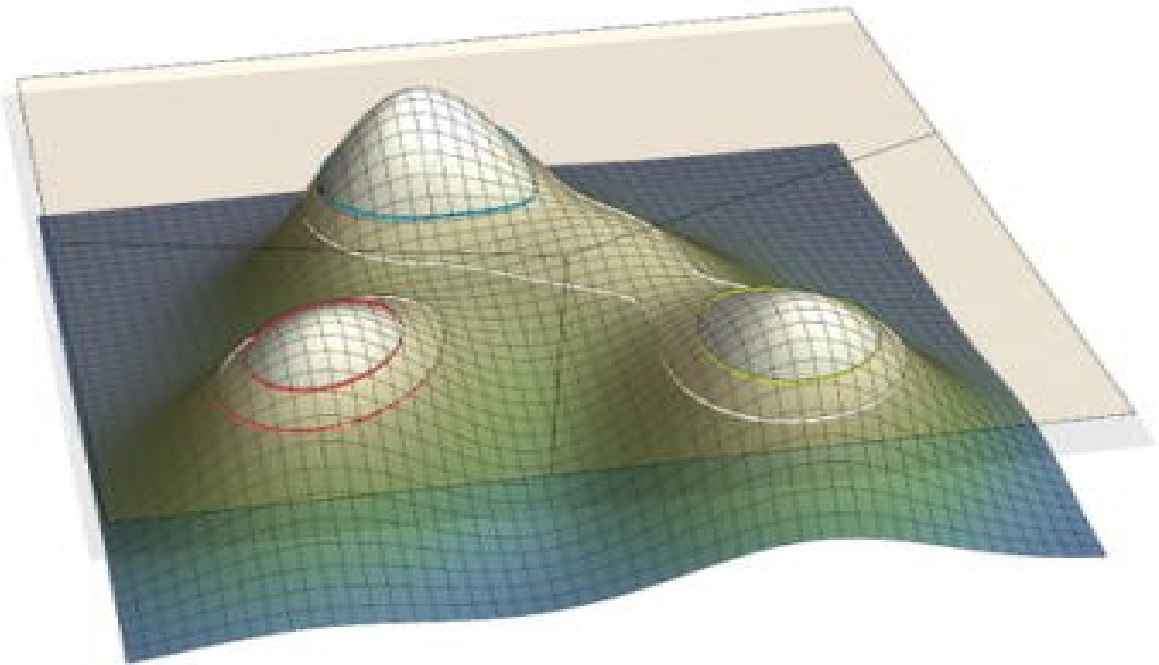


- How many clusters?
- What are the correct shapes?
- Which points go to which cluster?
- How do we evaluate the answers?





Community typing



Perspective | Published: 18 December 2017

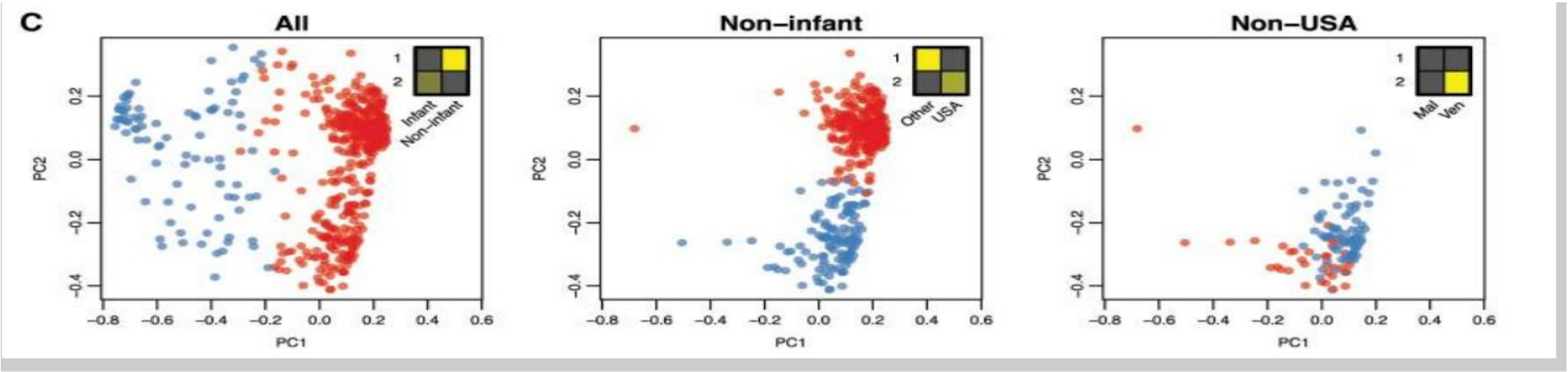
Enterotypes in the landscape of gut microbial community composition

Paul I. Costea, Falk Hildebrand, [...] Peer Bork

Nature Microbiology **3**, 8–16(2018) | Cite this article

6840 Accesses | **253** Citations | **100** Altmetric | Metrics

External covariates can induce distinct clusters



Rethinking “Enterotypes”

Dan Knights • Tonya L. Ward • Christopher E. McKinlay • ... Antonio Gonzalez • Daniel McDonald • Rob Knight

Show all authors

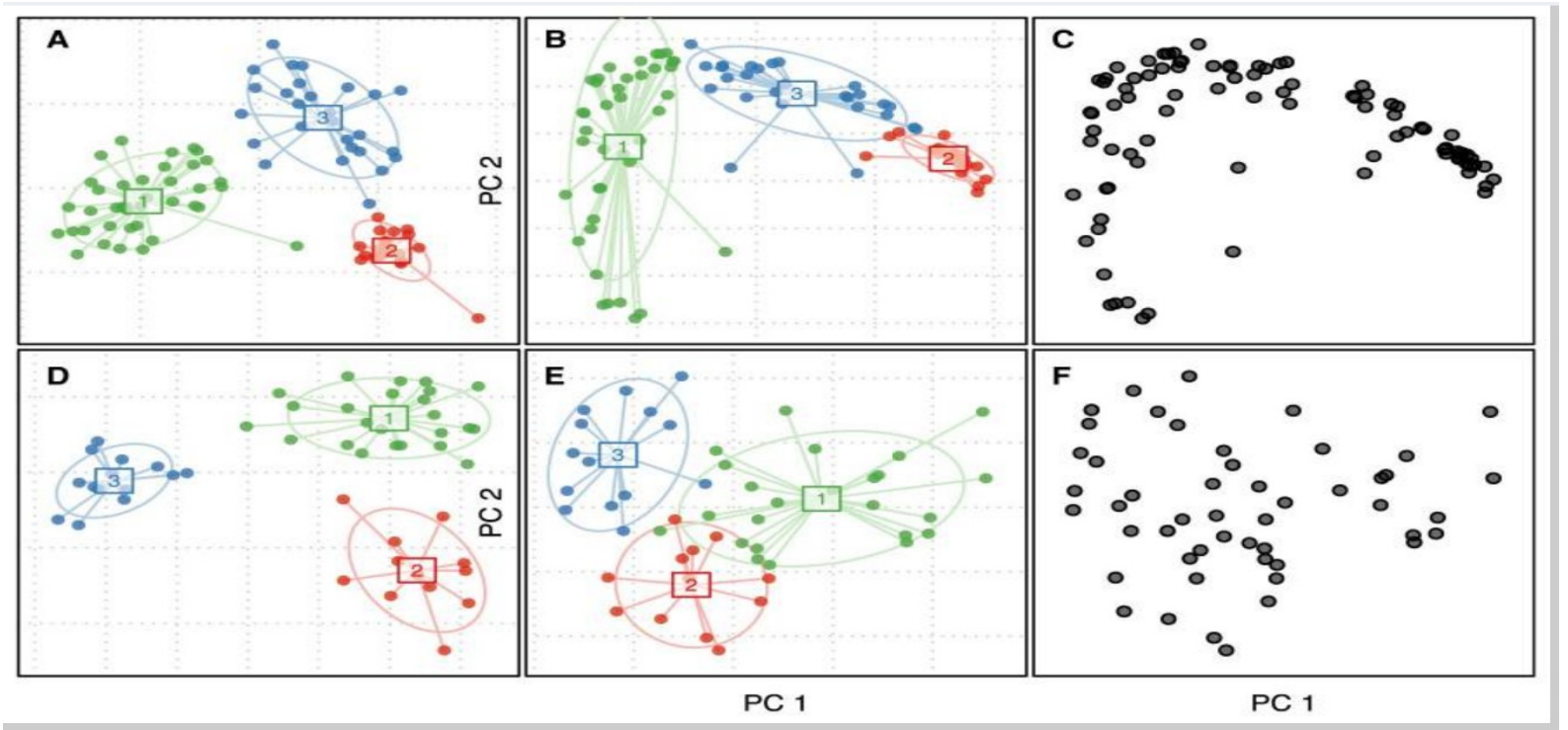
Open Archive • DOI: https://doi.org/10.1016/j.chom.2014.09.013 • Check for updates

Distinct clusters or extremes on a continuum?

Common Visualizations Can Support Different Conclusions

Soil samples with varying pH

Simulated data with no cluster structure



Supervised

Unsupervised with colors

Unsupervised without colors

Multinomial as a model for compositional microbiome data

$$f(x_1, \ldots, x_k; p_1, \ldots, p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i}$$

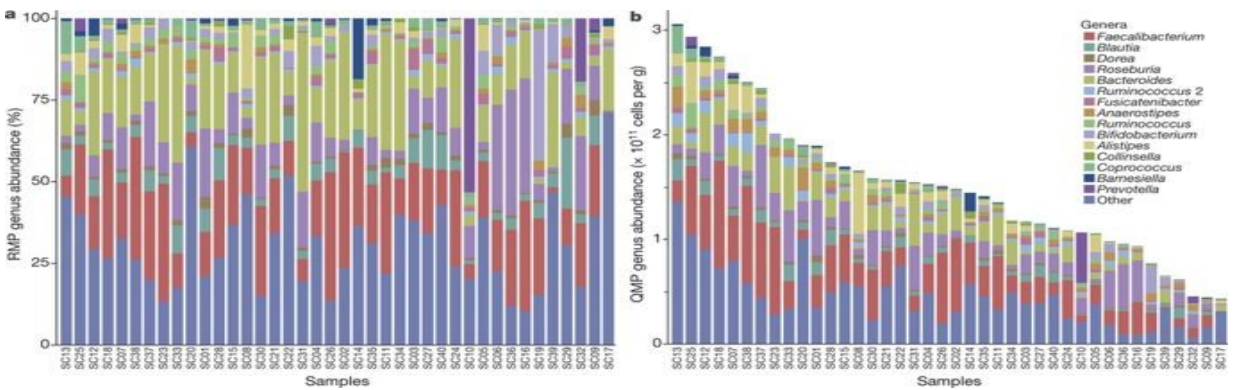


Image: Vandeputte et al. Nature 551:507-511, 2017

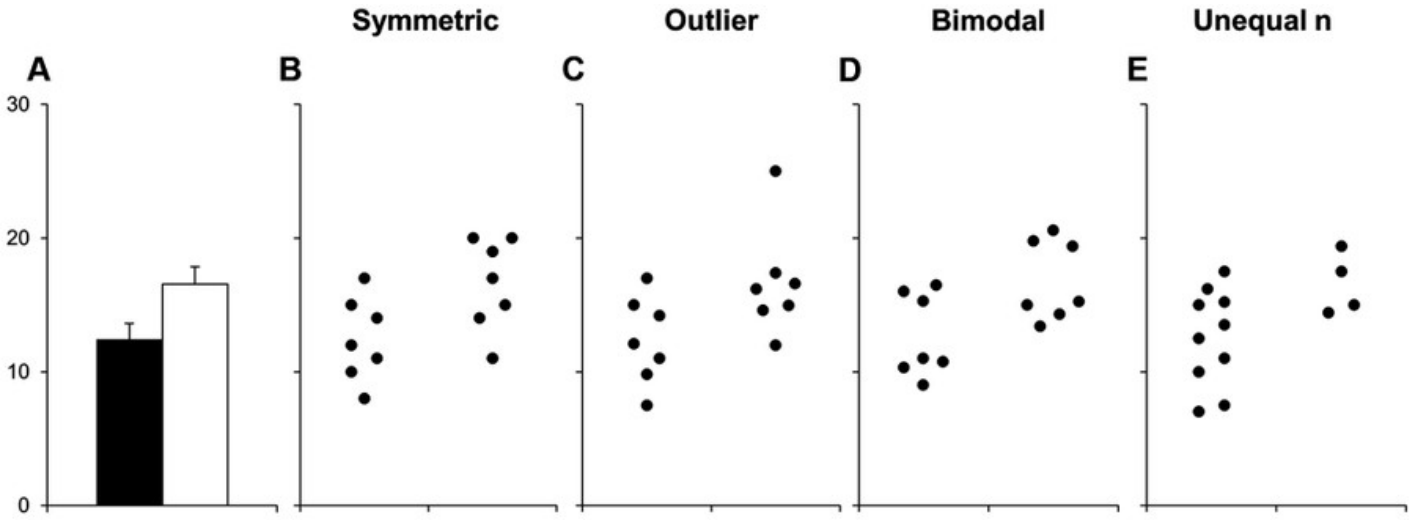
P < 0.04

Effect?

# P < 0.05

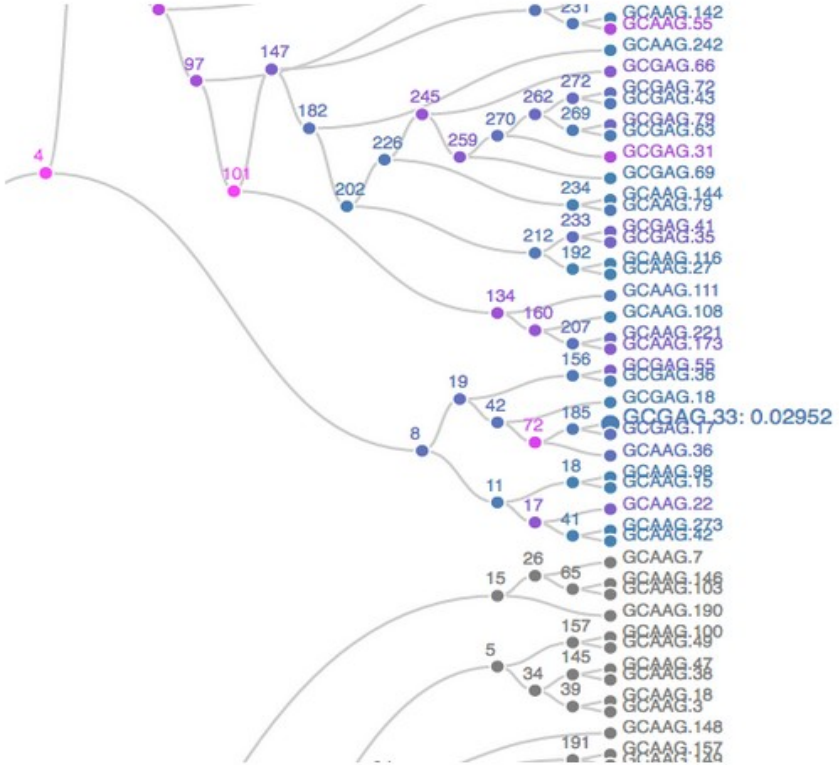
P < 0.06

No effect?



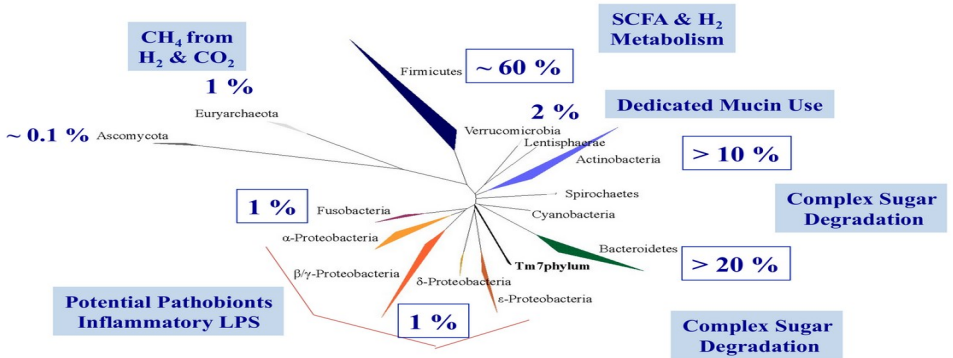
Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Hierarchical testing (Kris Sankaran)



Taking account of the phylogenetic tree when testing:

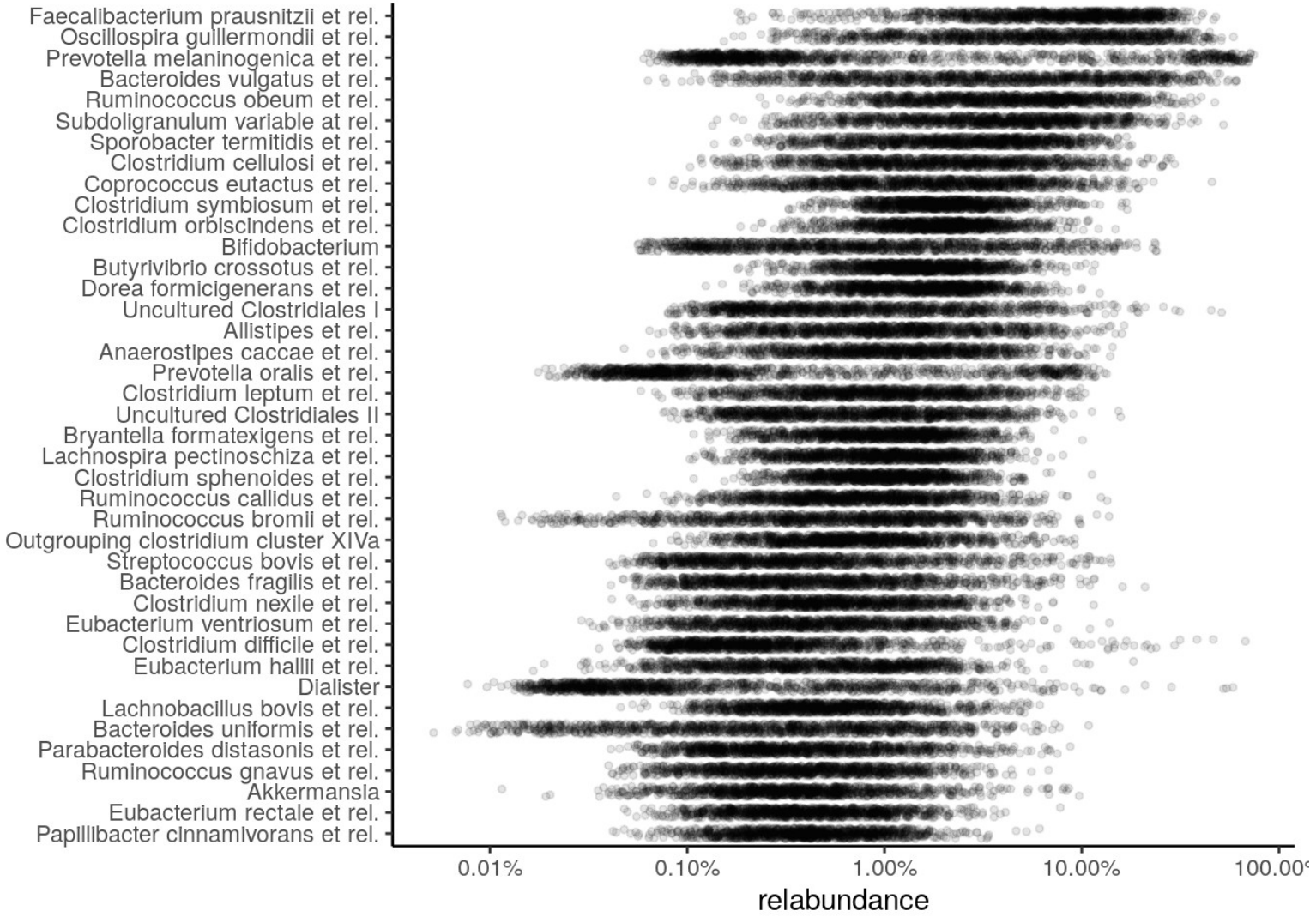
- CRAN package: structSSI
- Journ. Stat. Software paper JSS link



Tree-based methods

- StructSSI
- phylofactor
- tree-PCA
- UniFrac

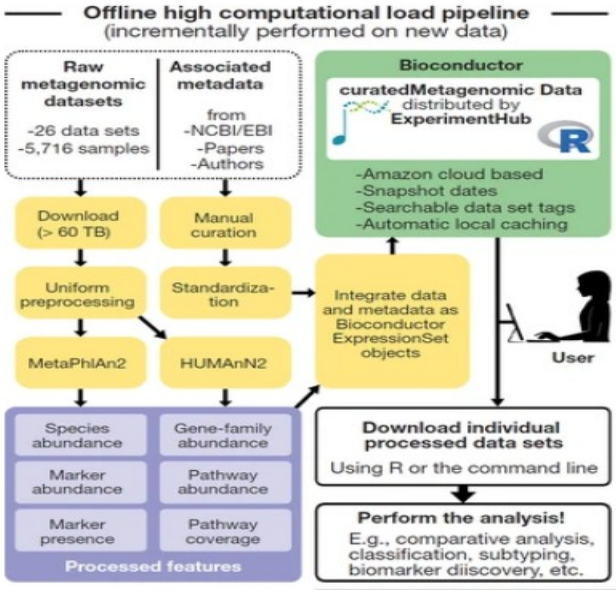
Source: Susan Holmes | http://web.stanford.edu/class/bios221/Short-Phyloseq-Resources.html

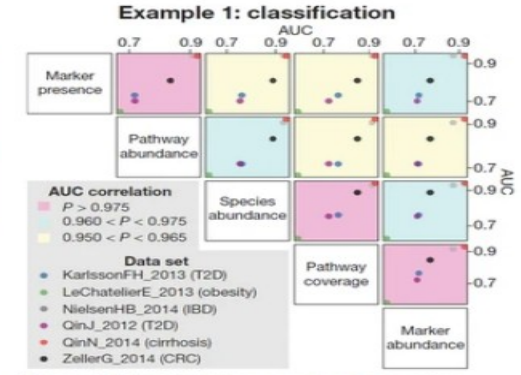

Faecalibacterium prausnitzii et rel.
Oscillospira guillermondii et rel.
Prevotella melaninogenica et rel.
Bacteroides vulgatus et rel.
Ruminococcus obeum et rel.
Subdoligranulum variable at rel.
Sporobacter termitidis et rel.
Clostridium cellulosi et rel.
Coprococcus eutactus et rel.
Clostridium symbiosum et rel.
Clostridium orbiscindens et rel.
Bifidobacterium
Butyrivibrio crossotus et rel.
Dorea formicigenerans et rel.
Uncultured Clostridiales I
Allistipes et rel.
Anaerostipes caccae et rel.
Prevotella oralis et rel.
Clostridium leptum et rel.
Uncultured Clostridiales II
Bryantella formatexigens et rel.
Lachnospira pectinoschiza et rel.
Clostridium sphenoides et rel.
Ruminococcus callidus et rel.
Ruminococcus bromii et rel.
Outgrouping clostridium cluster XIVa
Streptococcus bovis et rel.
Bacteroides fragilis et rel.
Clostridium nexile et rel.
Eubacterium ventriosum et rel.
Clostridium difficile et rel.
Eubacterium hallii et rel.
Dialister
Lachnobacillus bovis et rel.
Bacteroides uniformis et rel.
Parabacteroides distasonis et rel.
Ruminococcus gnavus et rel.
Akkermansia
Eubacterium rectale et rel.
Papillibacter cinnamivorans et rel.
0.01%
0.10%
1.00%
10.00%
relabundance

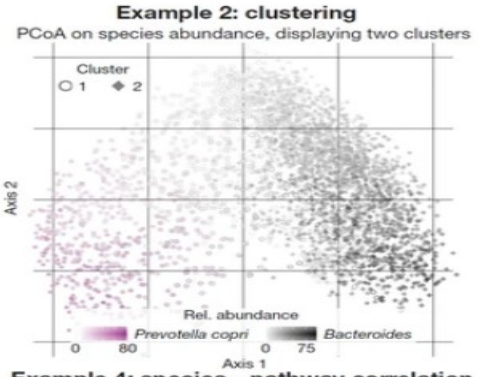
Accessible, curated metagenomic data through ExperimentHub

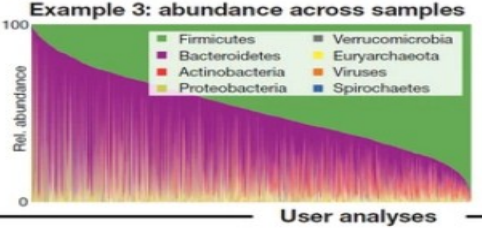
Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata ✉ & Levi Waldron ✉

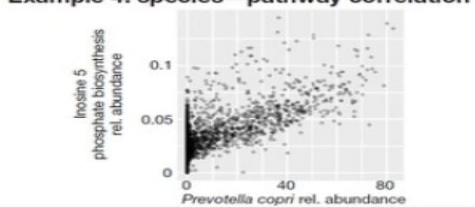
Nature Methods **14**, 1023–1024 (2017) | Cite this article











User analyses

curatedMetagenomicData

platforms all | rank 30 / 408 | support 1 / 1 | build ok
updated < 1 month | dependencies 155

DOI: 10.18129/B9.bioc.curatedMetagenomicData

Curated Metagenomic Data of the Human Microbiome

microbiomeDataSets

platforms all | rank 99 / 408 | support 0 / 0 | build ok
updated before release | dependencies 113

DOI: 10.18129/B9.bioc.microbiomeDataSets

Experiment Hub based microbiome datasets