Wednesday: Beta diversity

Demonstration

Community similarity

Practical

Beta diversity: estimation, analysis, and visualization

Key sources of microbial ecosystem variation

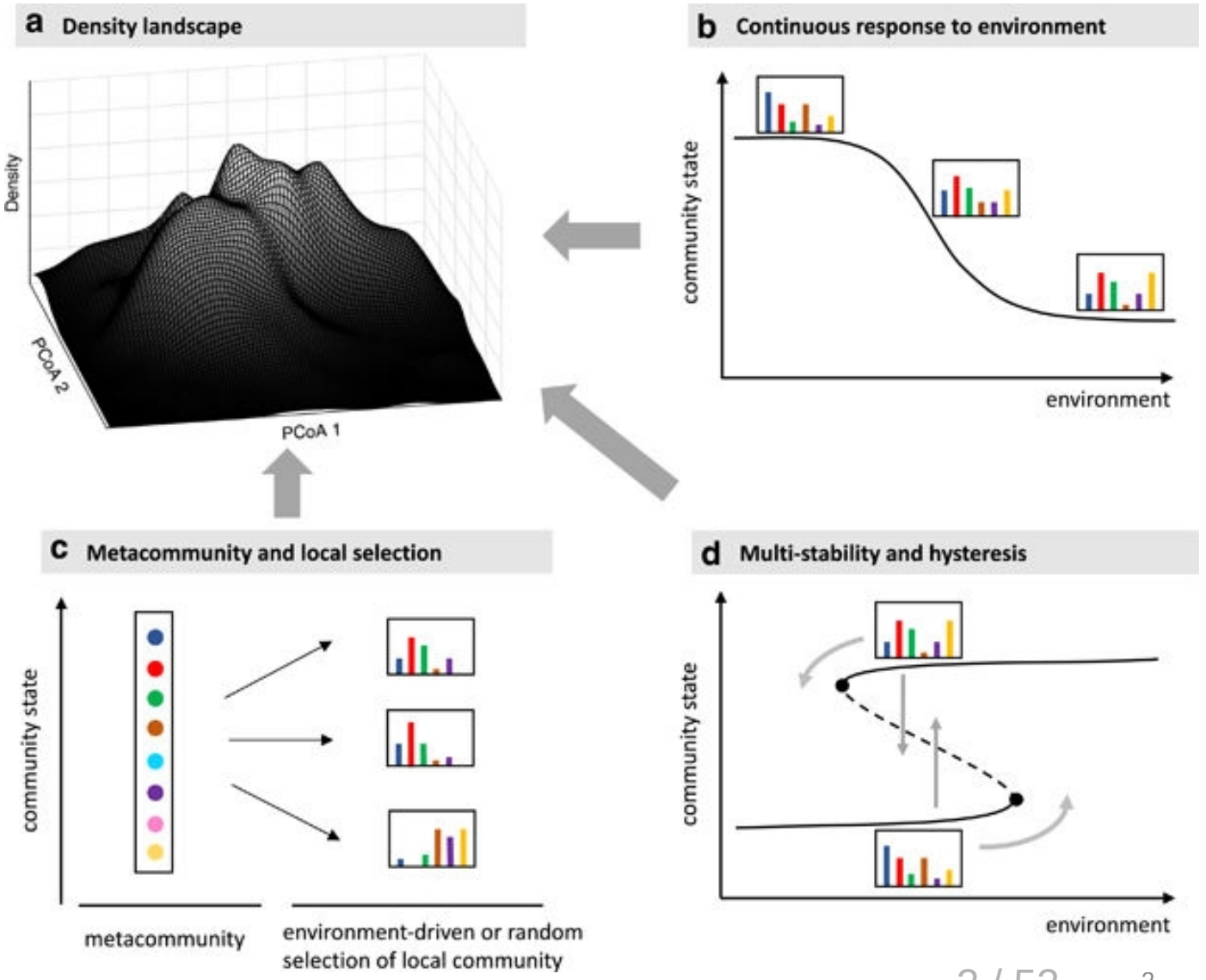
External perturbations (push & pulse)

Internal dynamics and multi-stability

Immigration

Stochasticity

Memory



Microbial communities as dynamical systems

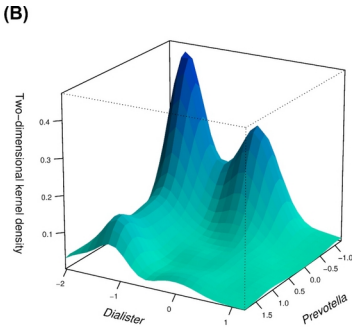
Didier Gonze [1, 2]✉, Katharine Z Coyte [3, 4], Leo Lahti [5, 6, 7], Karoline Faust [5]✉

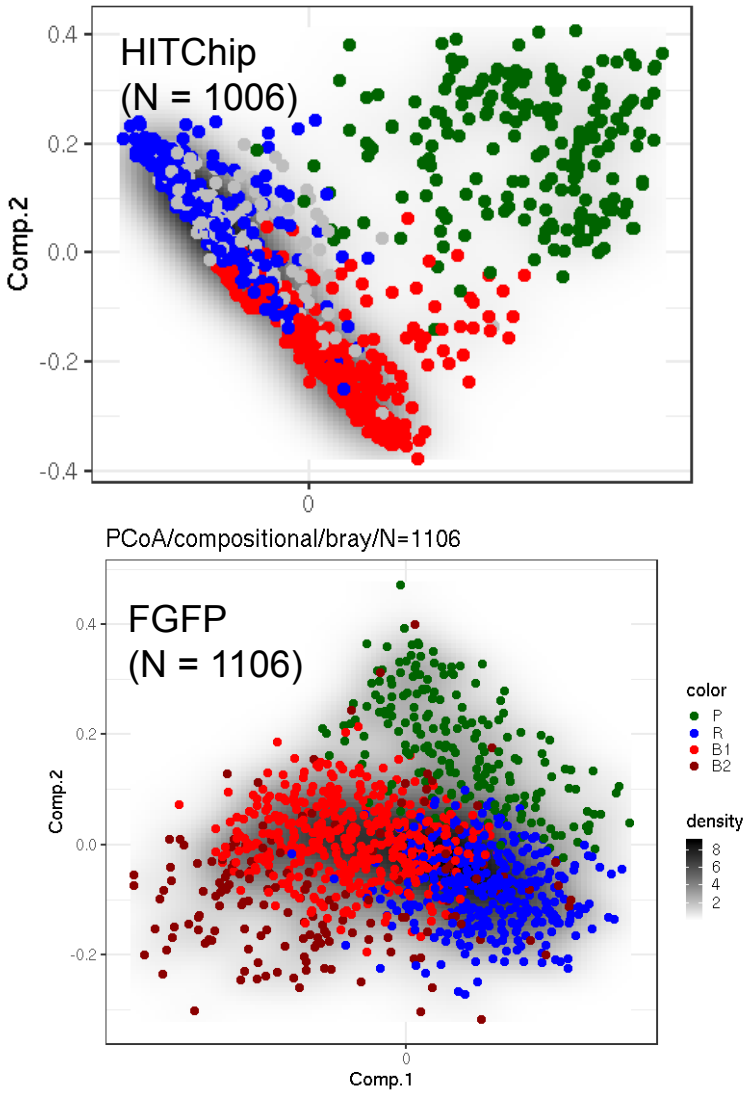
Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies FREE

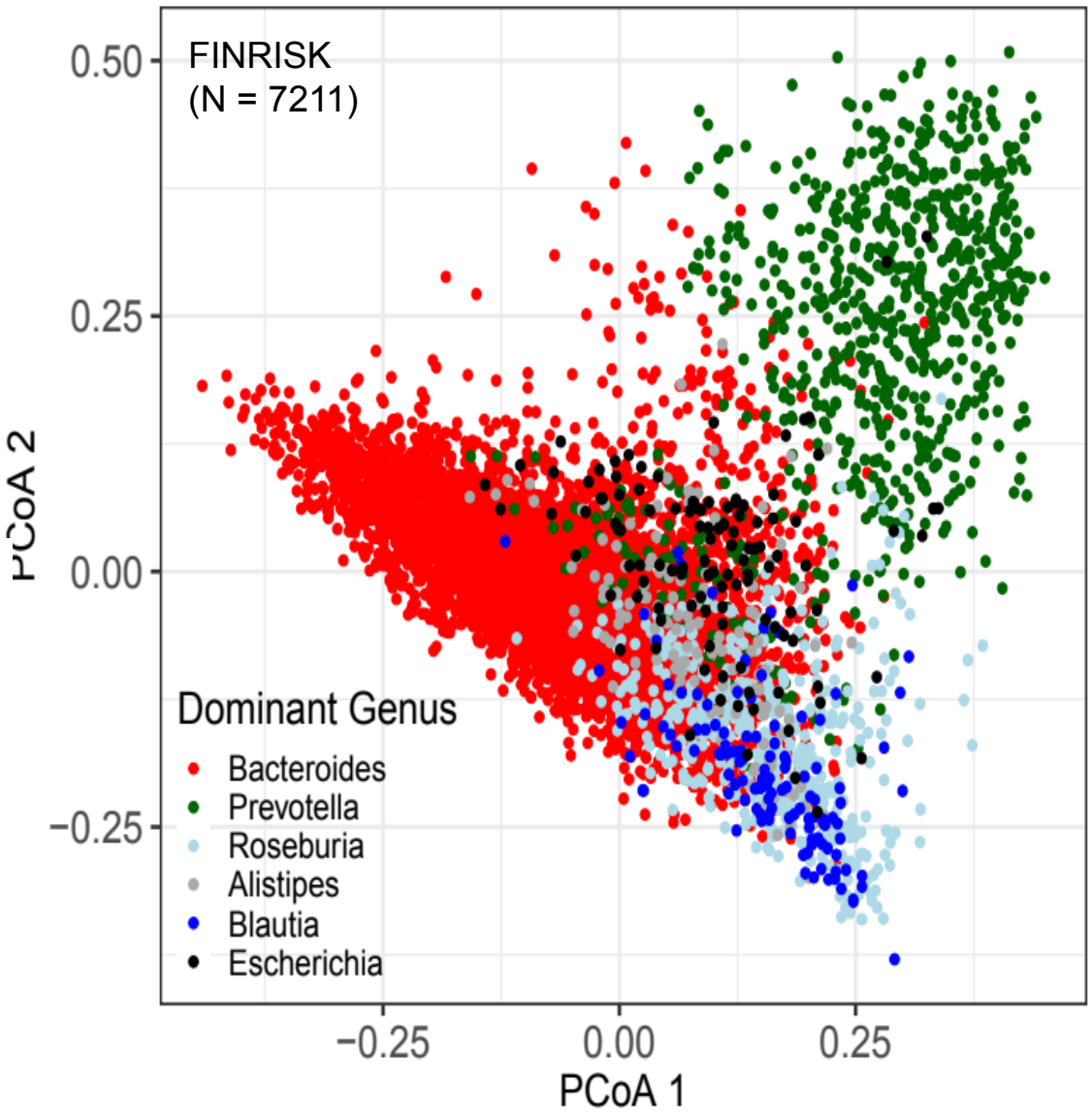
Sudarshan A. Shetty ✉, Floor Hugenholtz, Leo Lahti, Hauke Smidt, Willem M. de Vos

FEMS Microbiology Reviews, Volume 41, Issue 2, 1 March 2017, Pages 182–199, https://doi.org/10.1093/femsre/fuw045

Published: 09 February 2017 **Article history** ▾



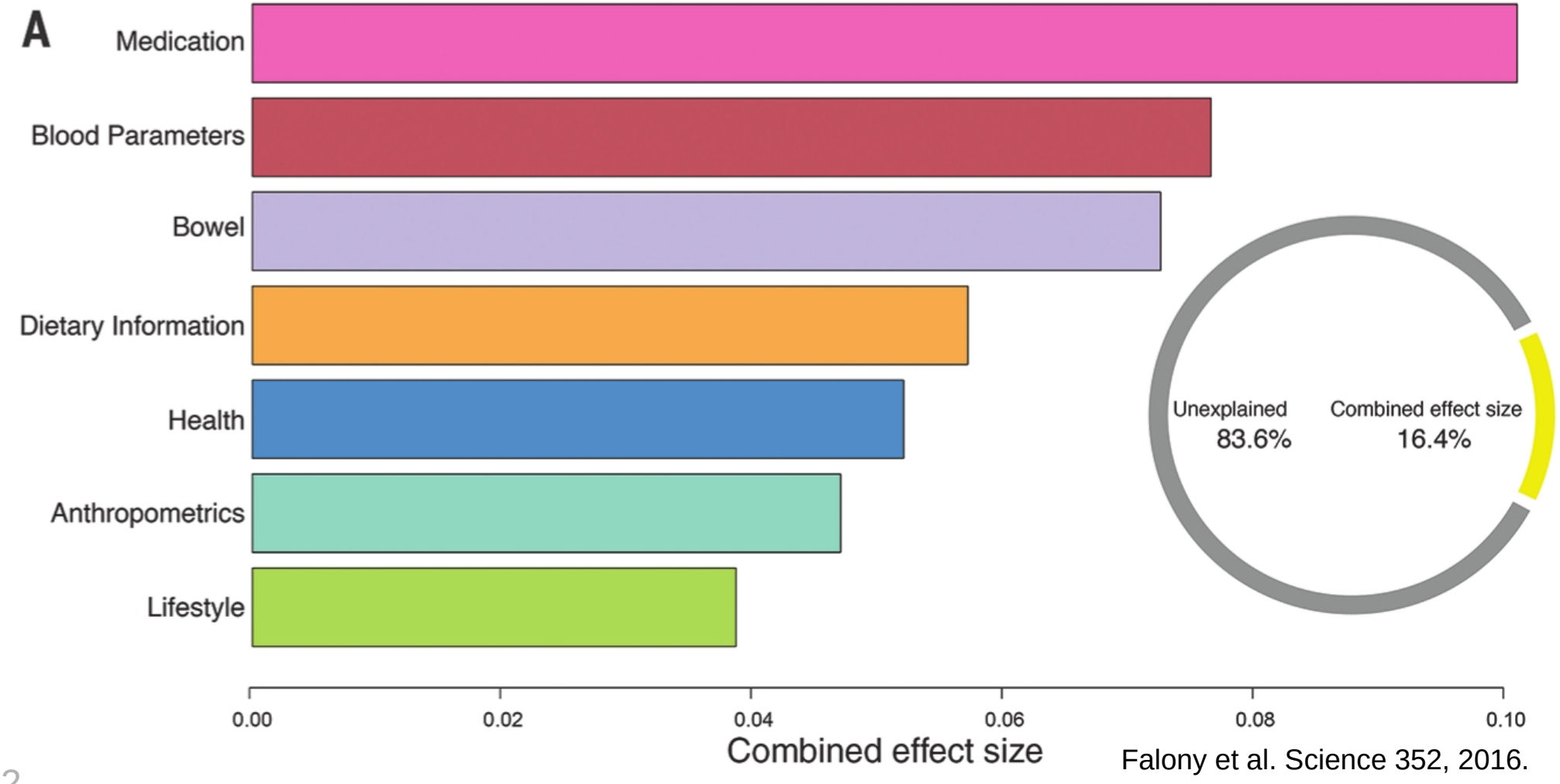




Total explained variation: 16.4%

(Flemish Gut Flora Project)

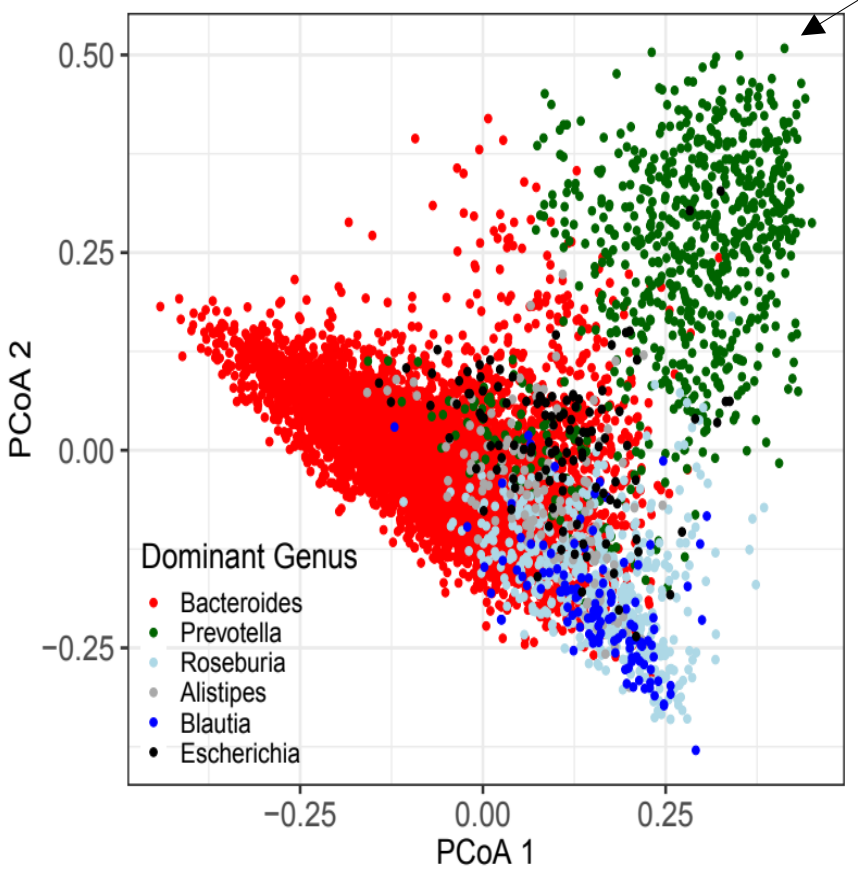
Proposed disease marker genera associated to host covariates and medication - inclusion in study design is essential !



Falony et al. Science 352, 2016.

Principal coordinates analysis (PcoA / MDS)

You?



PcoA Principal Coordinates Analysis (a.k.a MDS)

Transformation: compositional

Dissimilarity: Bray-Curtis

Method: **Preserves distances**

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

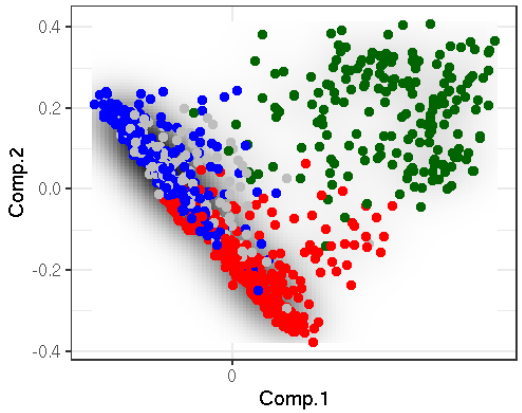
(absolute, compositional, CLR, Hellinger..?)

Dissimilarity measure

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

Analysis method

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



Fundamental considerations in beta diversity analysis

Feature selection

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

Transformation

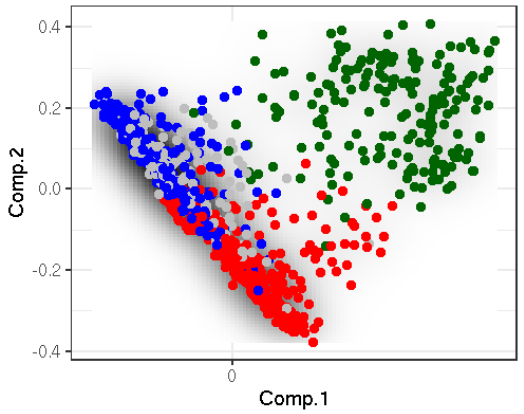
(absolute, compositional, CLR, Hellinger..?)

Dissimilarity measure

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

Analysis method

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



Fundamental considerations in beta diversity analysis

Feature selection

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

Transformation

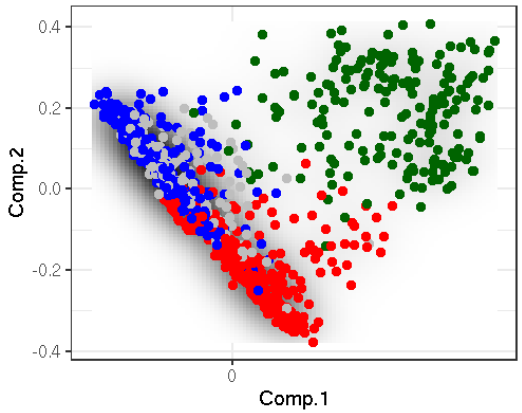
(absolute, compositional, CLR, Hellinger..?)

Dissimilarity measure

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

Analysis method

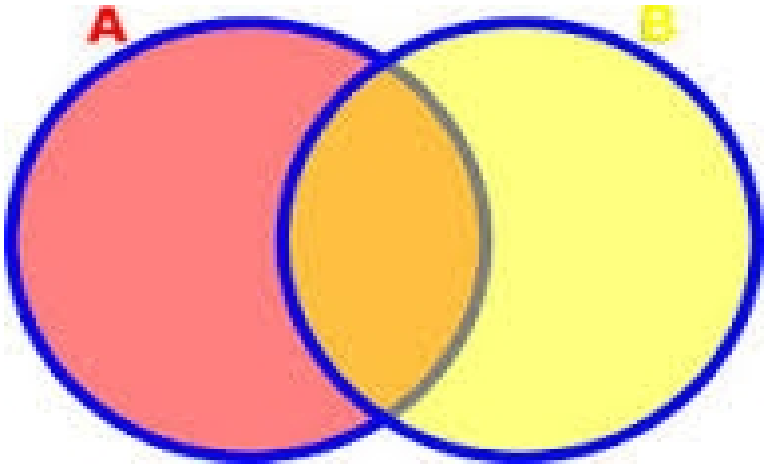
(PCA, PCoA, NMDS, t-SNE, UMAP..)



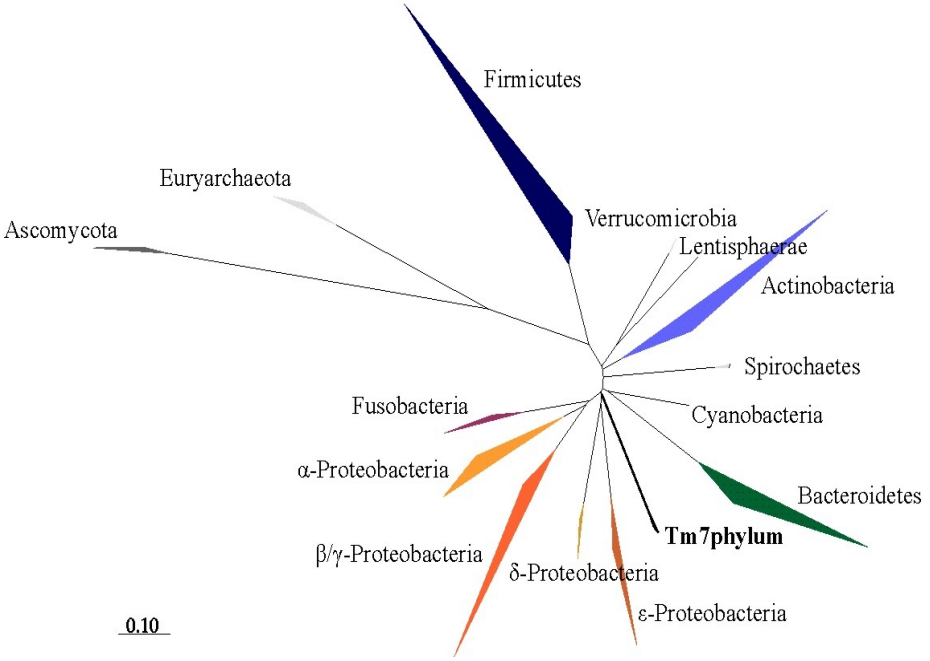
Beta diversity / dissimilarity / distance

- Euclidean
- Correlation
- Bray-Curtis
- Jaccard
- Unifrac (weighted & unweighted)

Jaccard index



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

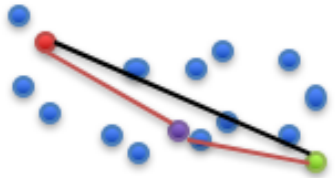


Bray-Curtis

$$BC_d = \frac{\sum \left| x_i - x_j \right|}{\sum \left(x_i + x_j \right)}$$

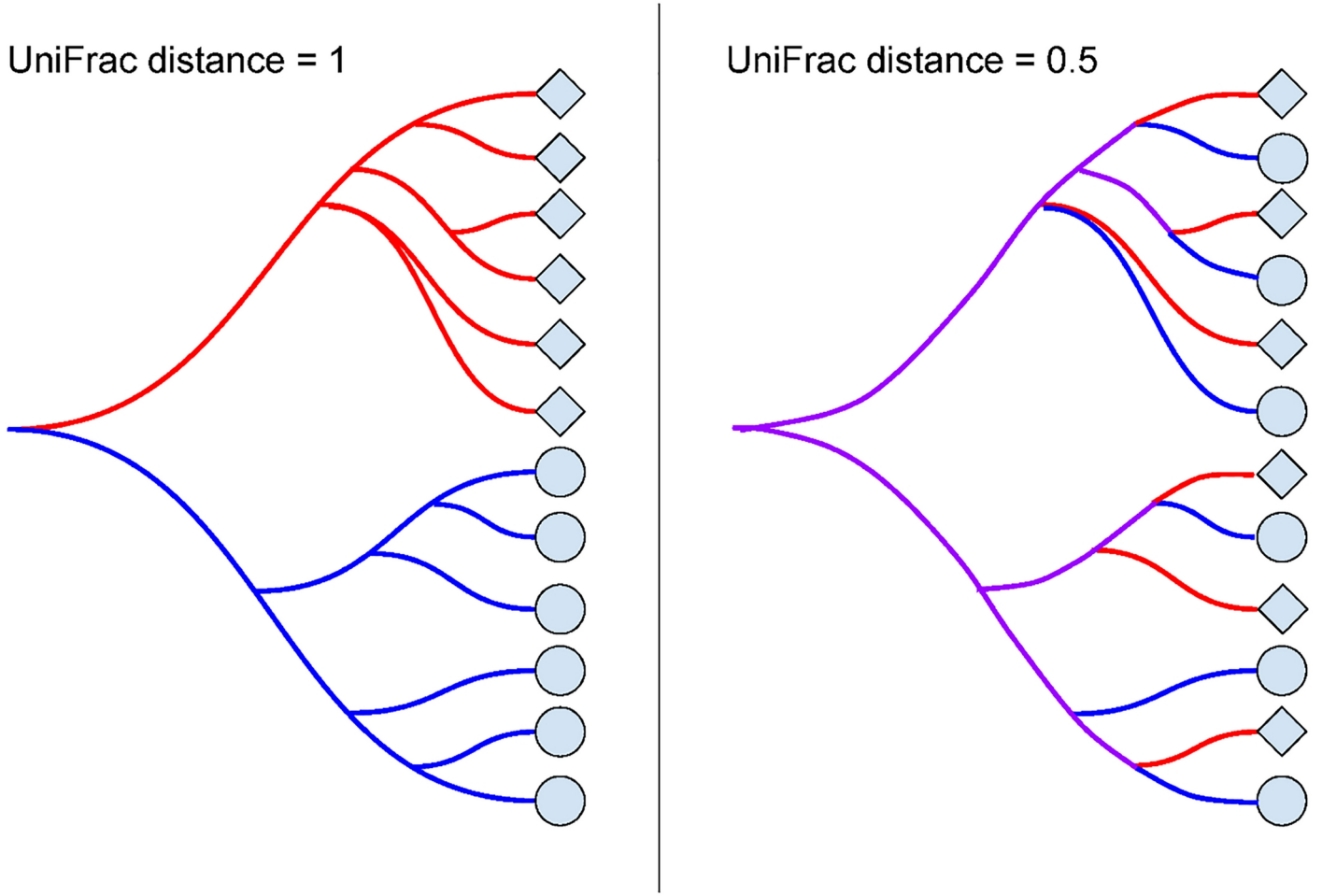
What is a distance metric?

- Scalar function d(.,.) of two arguments
- d(x, y) >= 0, always nonnegative;
- d(x, x) = 0, distance to self is 0;
- d(x, y) = d(y, x), distance is symmetric;
- d(x, y) < d(x, z) + d(z, y), triangle inequality.



5

$$\left(\frac{sum\ of\ unshared\ branch\ lengths}{sum\ of\ all\ tree\ branch\ lengths}\right) = fraction\ of\ total\ unshared\ branch\ lengths$$



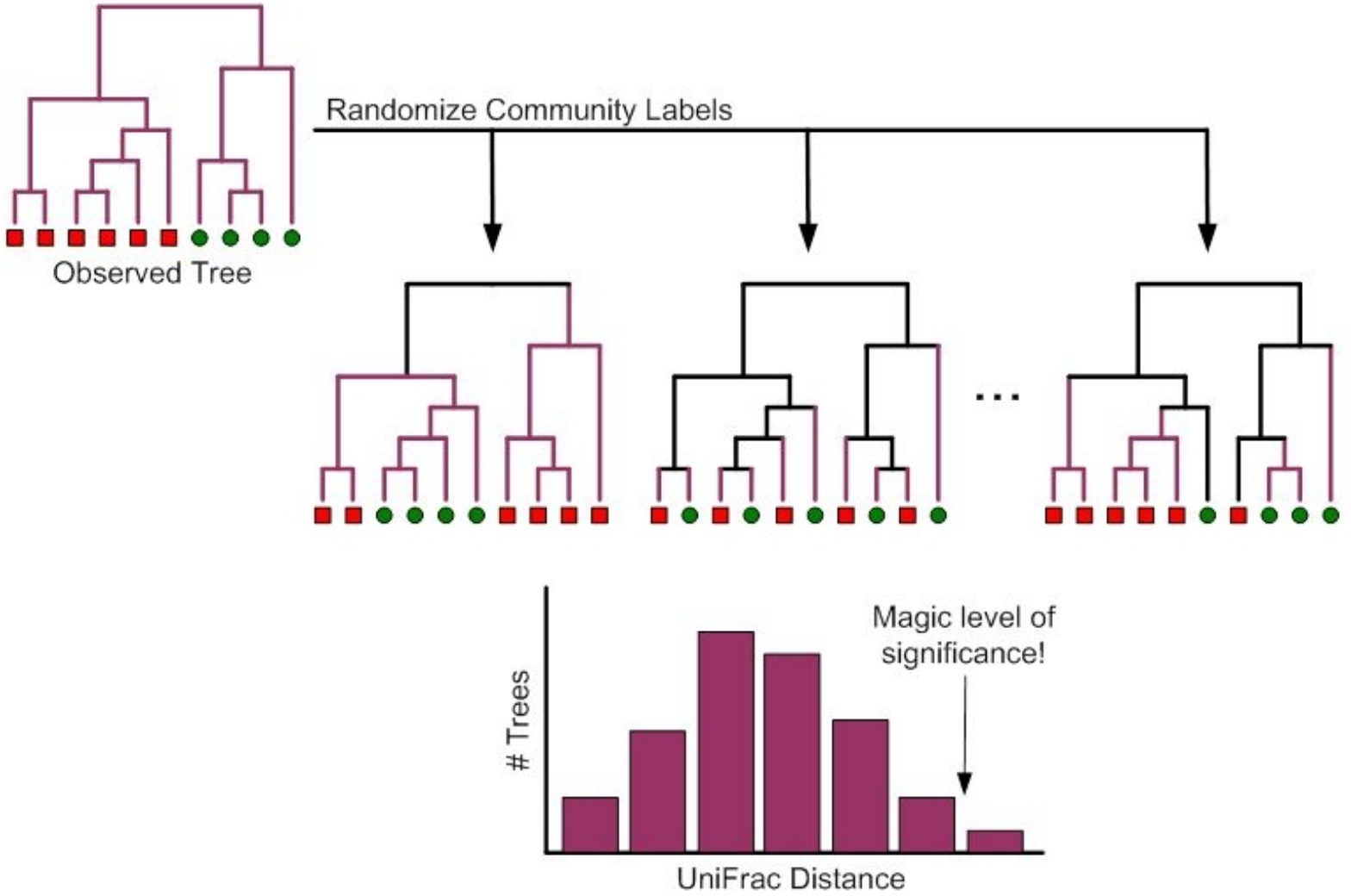
Expanding the UniFrac Toolbox

Ruth G. Wong, Jia R. Wu, Gregory B. Gloor

Published: September 15, 2016 • https://doi.org/10.1371/journal.pone.0161196

UniFrac: Significance Test

- Do two communities differ significantly?



Fundamental considerations in beta diversity analysis

Feature selection

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

Transformation

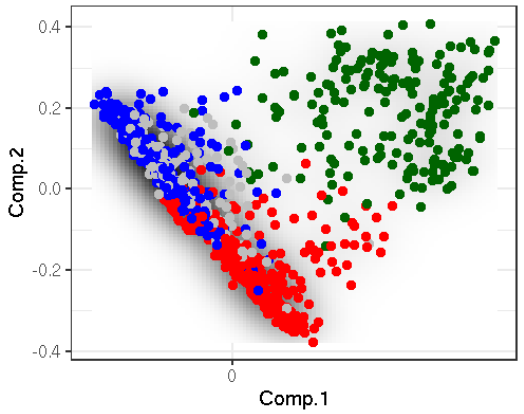
(absolute, compositional, CLR, Hellinger..?)

Dissimilarity measure

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

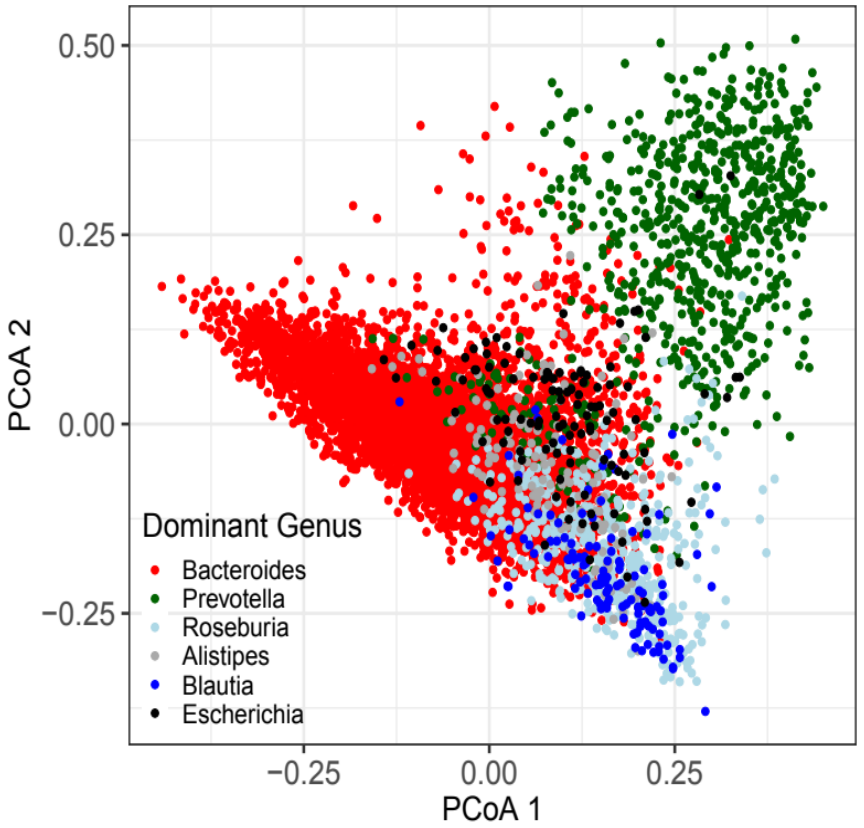
Analysis method

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



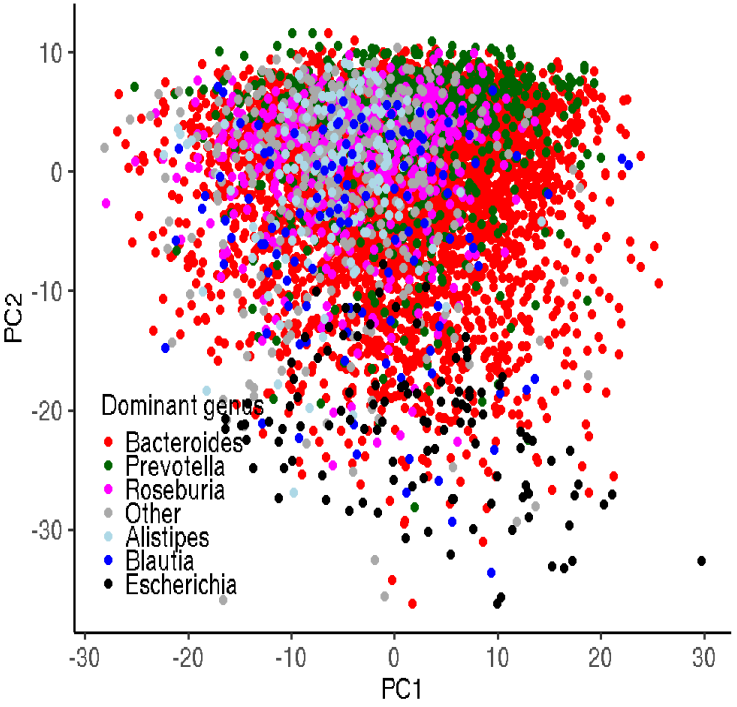
PCoA + Bray-Curtis

Preserves distances



PCA + Aitchison

Captures largest variation



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

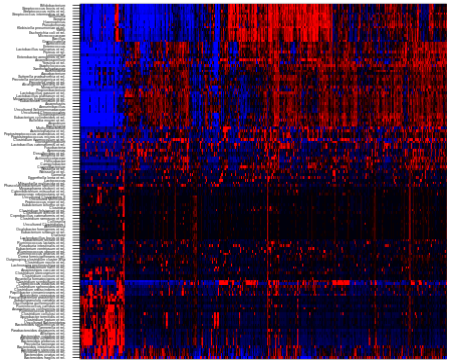
doi: https://doi.org/10.1101/2019.12.30.19015842

The influence of hidden researcher decisions in applied microeconomics

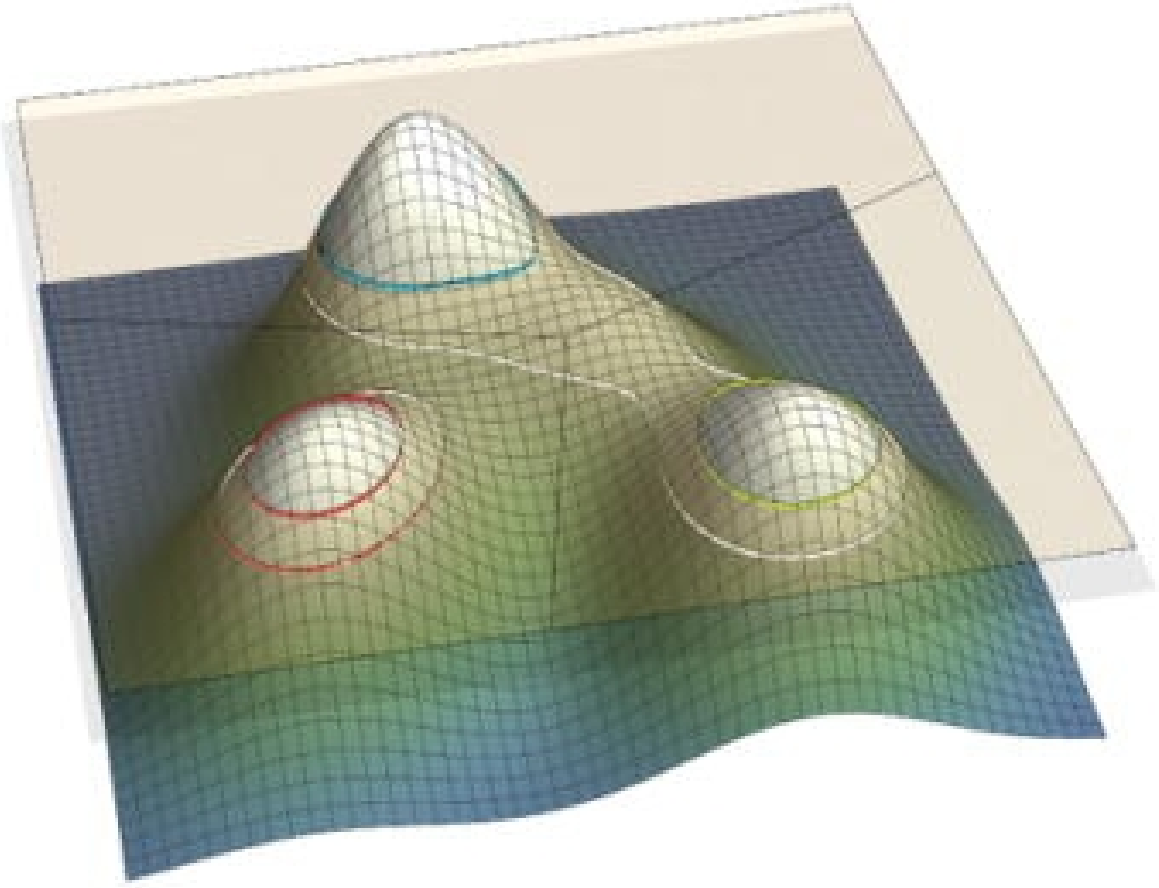
Nick Huntington-Klein ✉, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, Yaniv Stopnitzky

First published: 22 March 2021
https://doi.org/10.1111/ecin.12992

Researchers make hundreds of decisions about data collection, preparation, and analysis in their research. We use a many-analysts approach to measure the extent and impact of these decisions. Two published causal empirical results are replicated by seven replicators each. We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication. No two replicators reported the same sample size. Statistical significance varied across replications, and for one of the studies the effect's sign varied as well. The standard deviation of estimates across replications was 3–4 times the mean reported standard error.



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

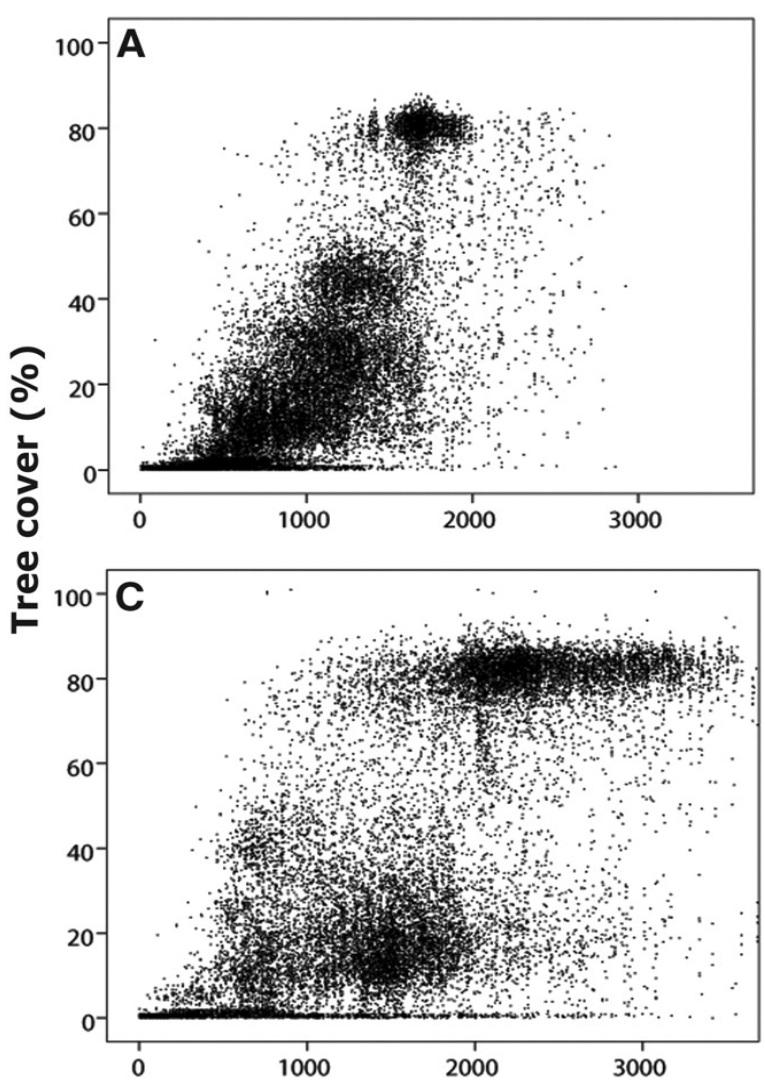
6840 Accesses | **253** Citations | **100** Altmetric | Metrics

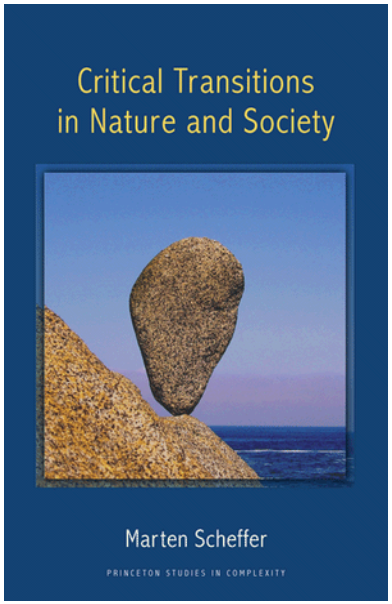
Science 14 October 2011:
Vol. 334 no. 6053 pp. 232-235
DOI: 10.1126/science.1210657

REPORT

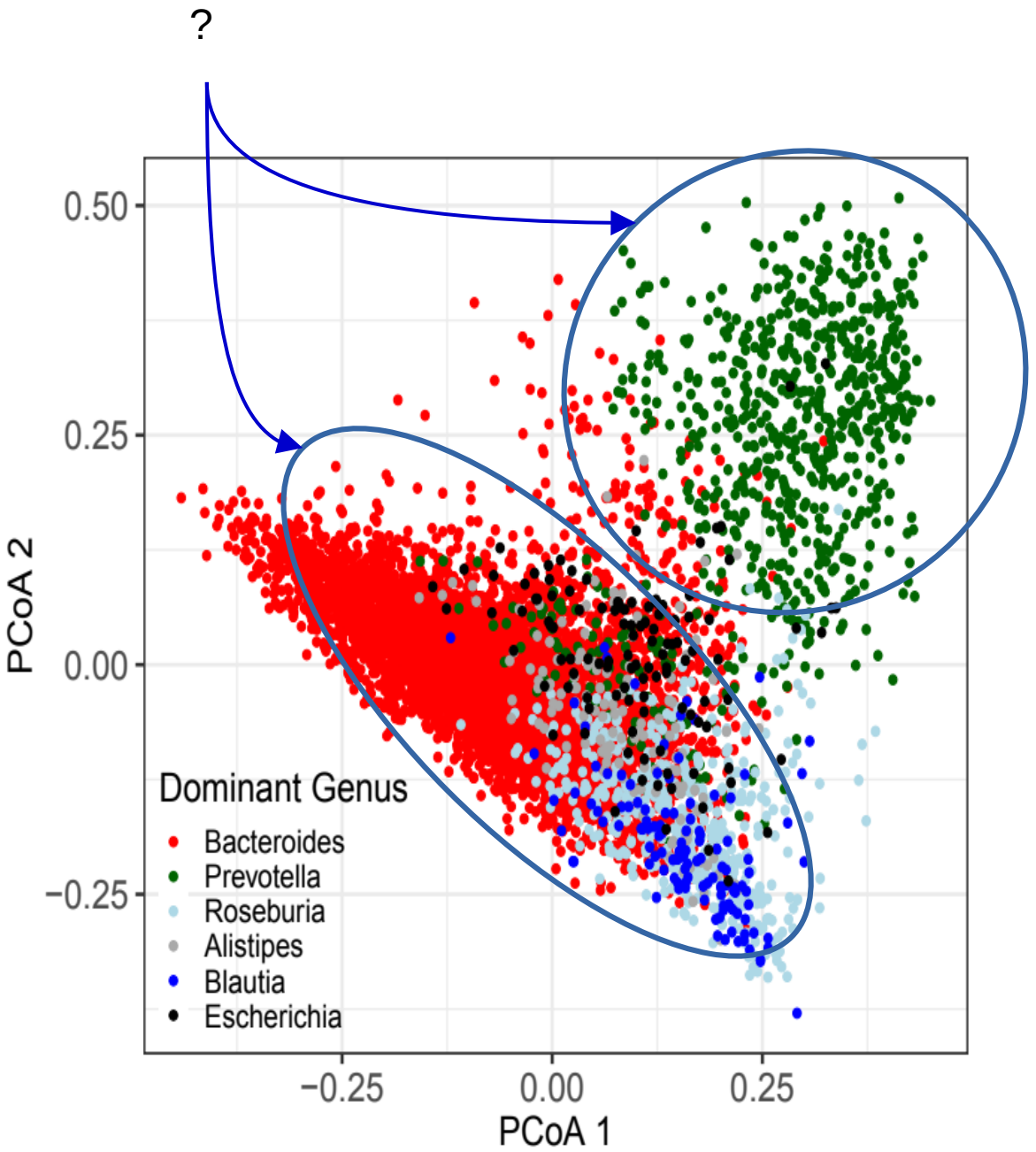
Global Resilience of Tropical Forest and Savanna to Critical Transitions

Marina Hirota[1], Milena Holmgren[2,*], Egbert H. Van Nes[1], Marten Scheffer[1]

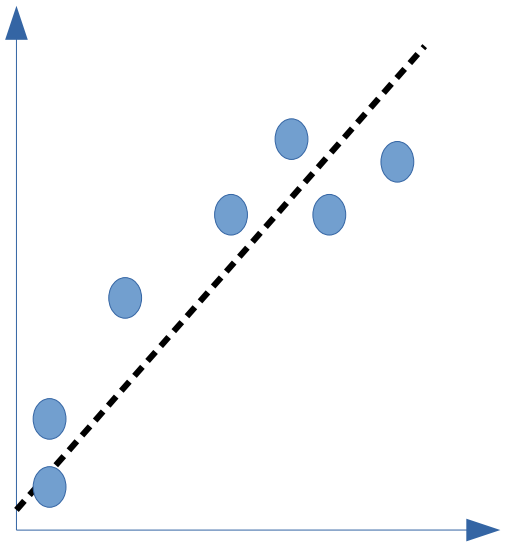




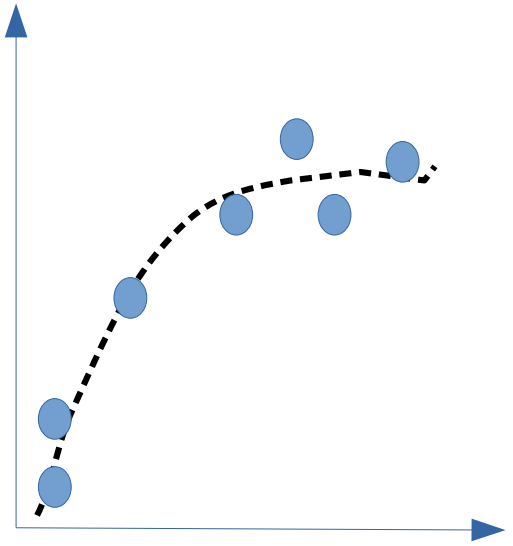
Aspects of model structure
- Cluster shape?
- Cluster number?
- Cluster assignments..?



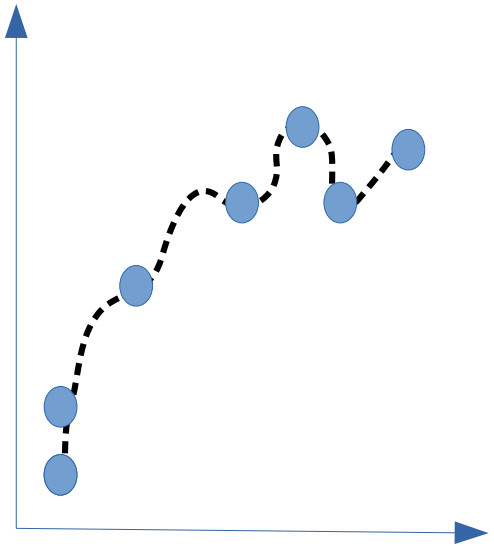
Tradeoff between robustness and sensitivity



Simple model: underfitting & high bias



Intermediate model – ”just right”

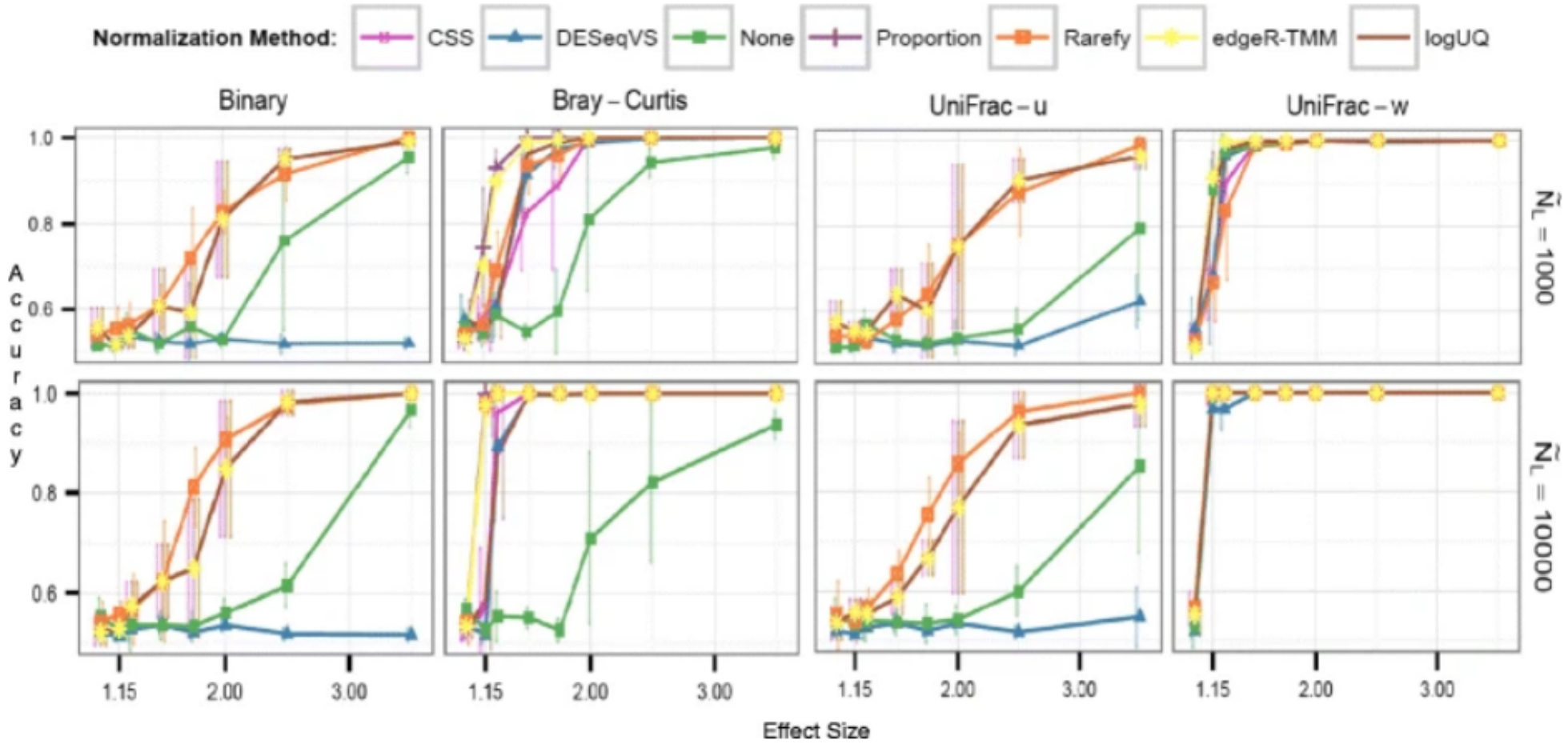


Complex model – overfitting & high variance

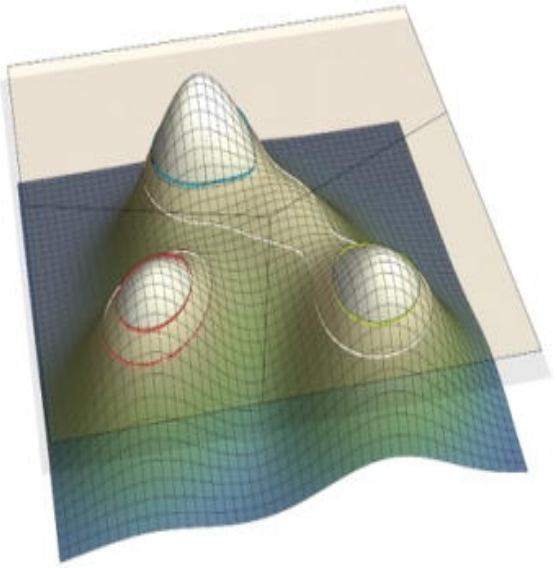
Fig. 2

From: Normalization and microbial differential abundance strategies depend upon data characteristics

Dissimilarity measure and normalization affect clustering accuracy



How to choose a correct model?
→ a community typing example



Enterotypes in the landscape of gut microbial community composition. Costea *et al*. Nature 2018.

$$2 \times 6^6 = 93312$$

Taxonomic level
- Phylum
- Family
- Order
- Genus
- Species
- Strain..

Filtering
- None
- Prevalent
- Core
- Excl. outliers
- High variance
- Custom

Normalization
- None
- TSS
- CSS
- ILR/ALR/CLR
- phILR
- Hellinger

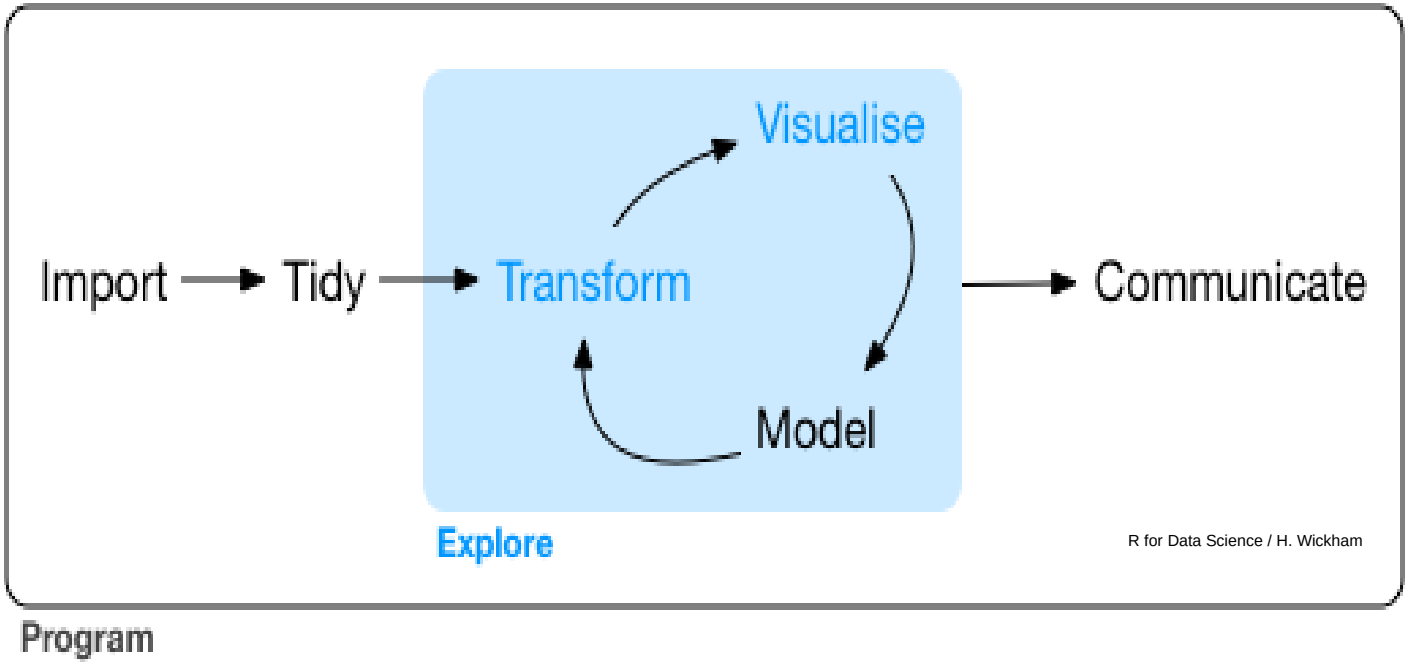
(Dis)similarity
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

Clustering method
- Hierarchical / Ward
- Hierarchical / Complete
- Gaussian mixture
- DMM
- PAMR
- K-means

Regulation
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

Walk-through example in R/Bioc by Holmes & McMurdie
http://statweb.stanford.edu/~susan/papers/EnterotypeRR.html

Reproducible workflows improve transparency and robustness



Taxonomic level?
- Phylum
- Family
- Order
- Genus
- Species
- Strain..

Normalization
- None
- TSS
- CSS
- ILR/ALR/CLR
- phILR
- Hellinger

(Dis)similarity?
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

Regulation
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

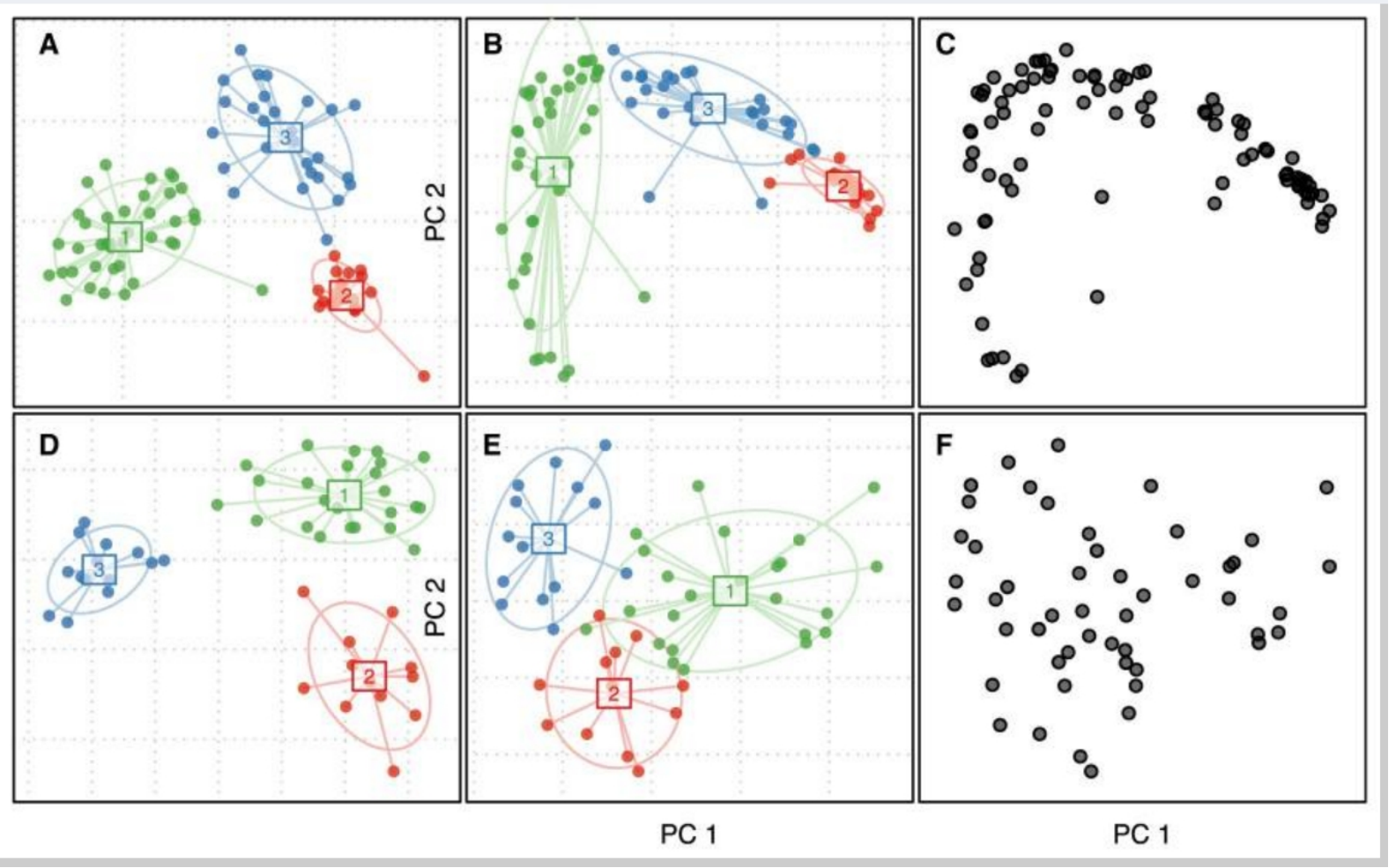
Clustering
- Hierarchical / Ward
- Hierarchical / Complete
- Gaussian mixture
- DMM
- PAMR
- K-means

Distinct clusters or extremes on a continuum?

Common Visualizations Can Support Different Conclusions

Soil samples with varying pH

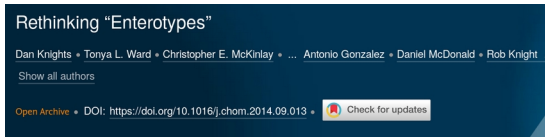
Simulated data with no cluster structure



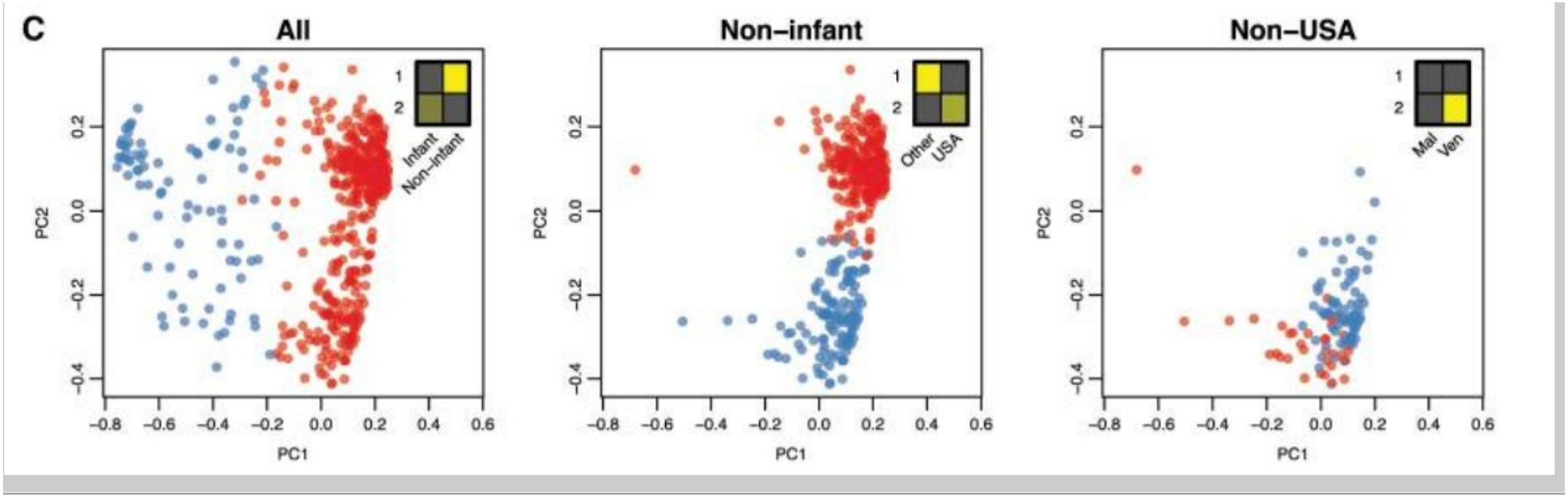
Supervised

Unsupervised with colors

Unsupervised without colors



External covariates can induce distinct clusters



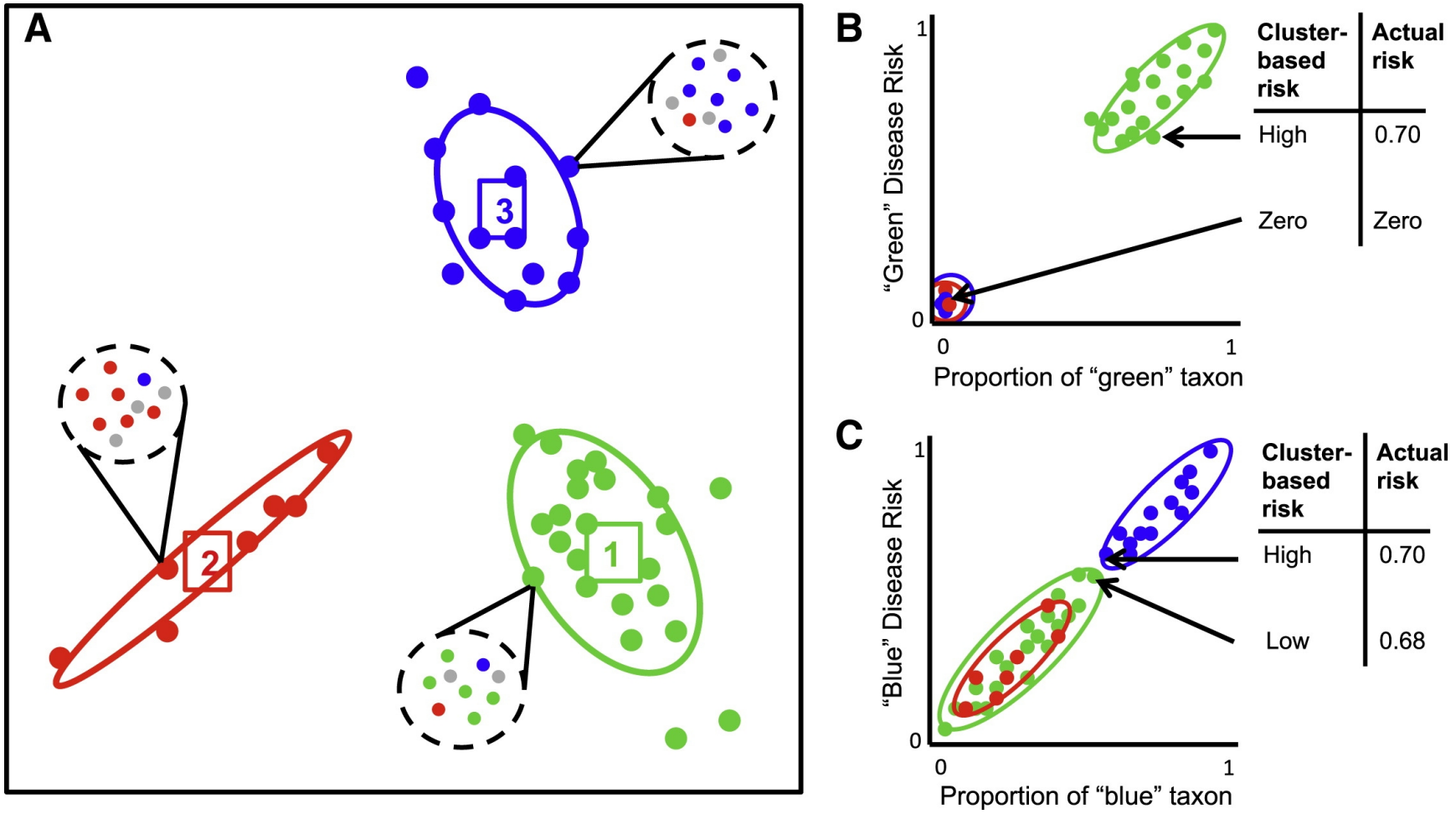
Rethinking “Enterotypes”

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

Clustering Continuous Data May Mask Within-Cluster Variation



Rethinking "Enterotypes"

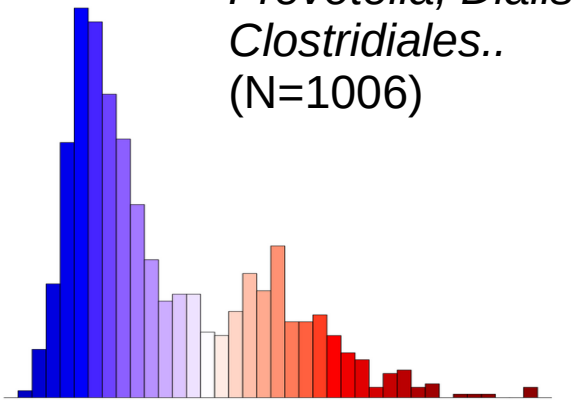
Dan Knights • Tonya L. Ward • Christopher E. McKinlay • ... Antonio Gonzalez • Daniel McDonald • Rob Knight

Show all authors

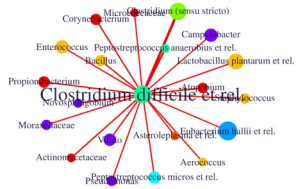
Open Archive • DOI: https://doi.org/10.1016/j.chom.2014.09.013 • Check for updates

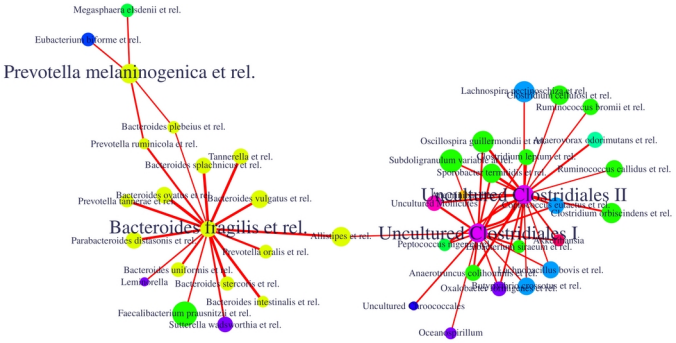
Mixture models bring flexibility in modeling

Bistable taxa:
Prevotella, Dialister, Clostridiales..
(N=1006)

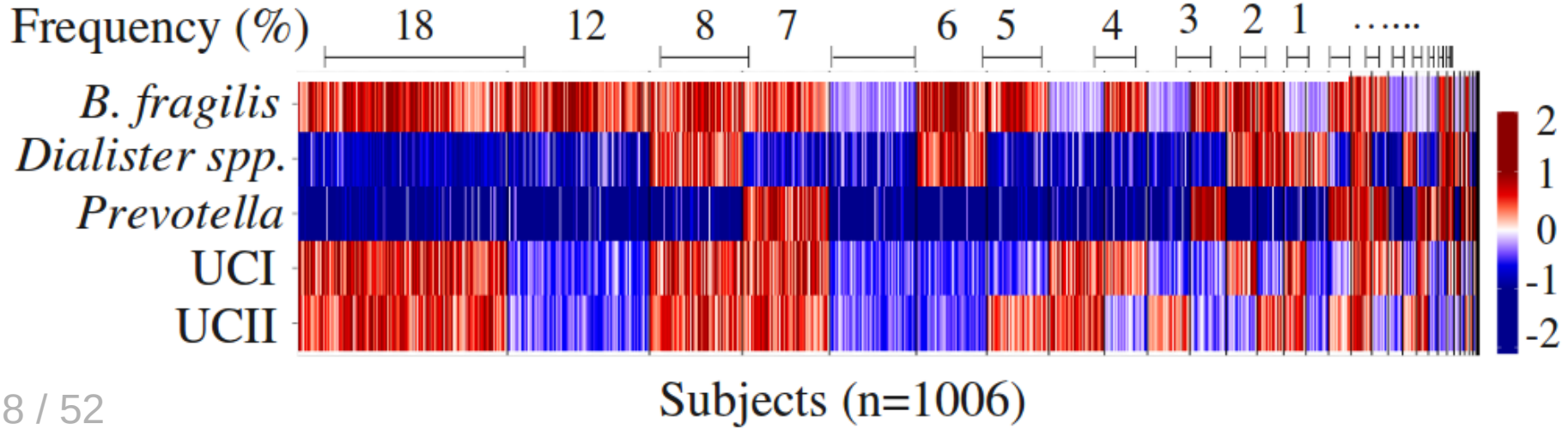


Independently varying sub-communities



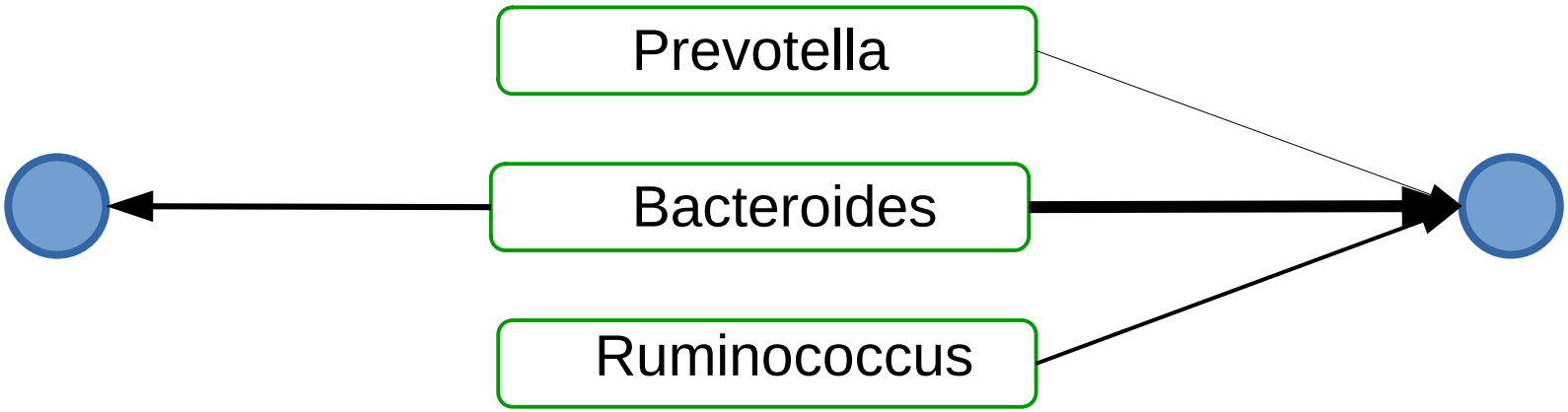


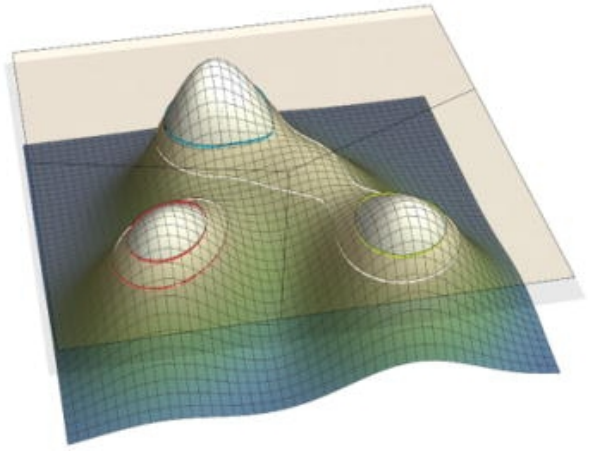
Ecosystem states are rich combinations of independent tipping elements ? Lahti et al. Nat. Comm. 2014

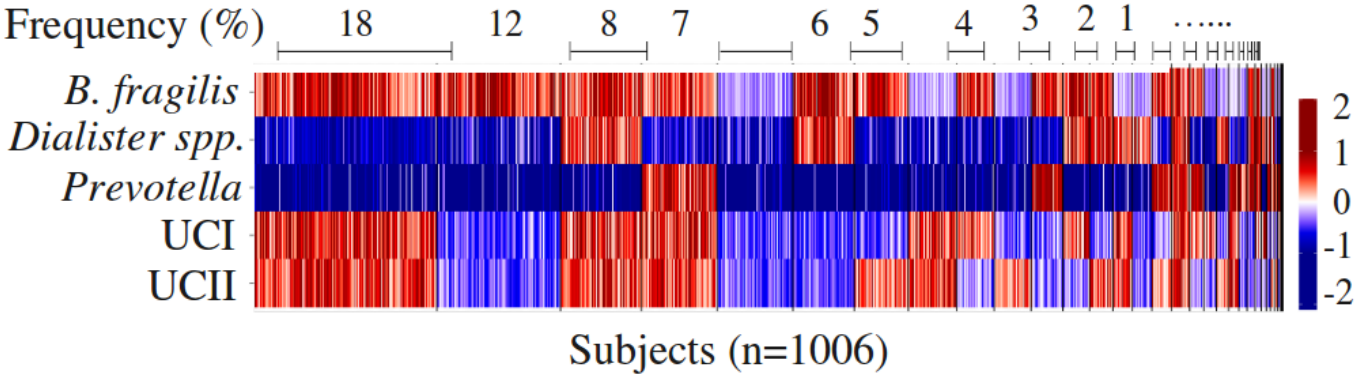


Clustering vs. Factorization

binary / continuous weights

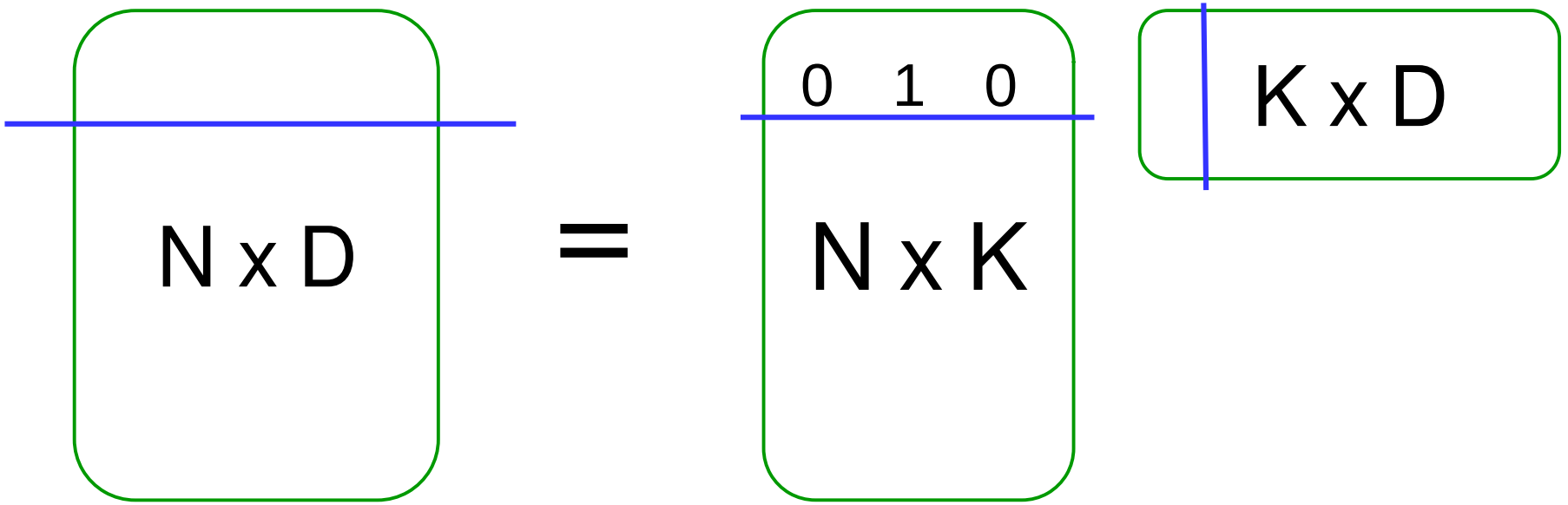






Clustering vs. Factorization

statistical formulation



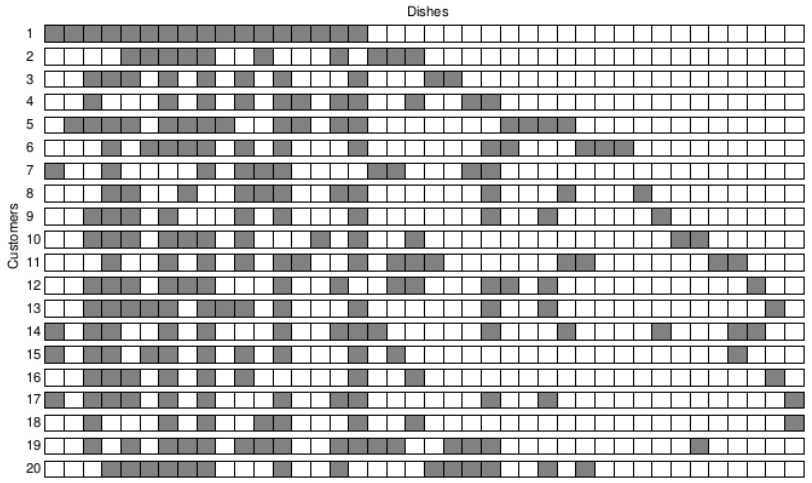
Observed data

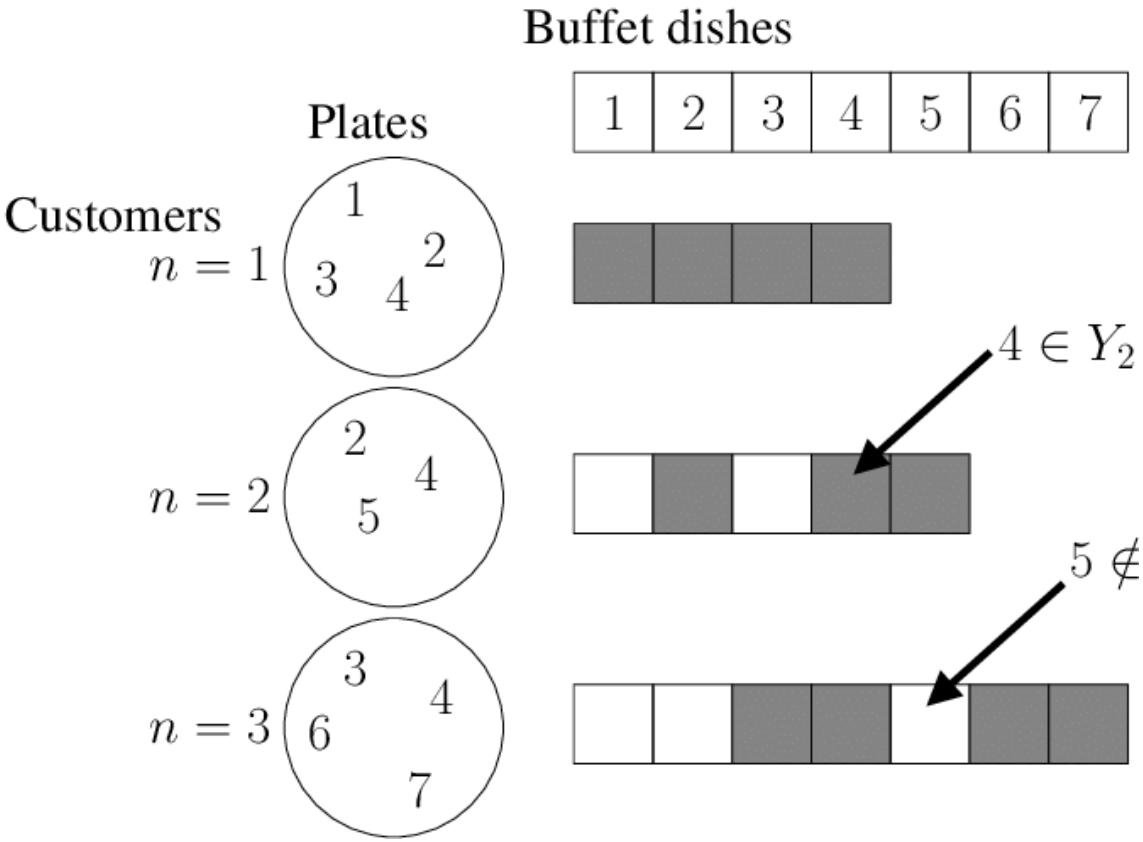
Component weights

Component profiles

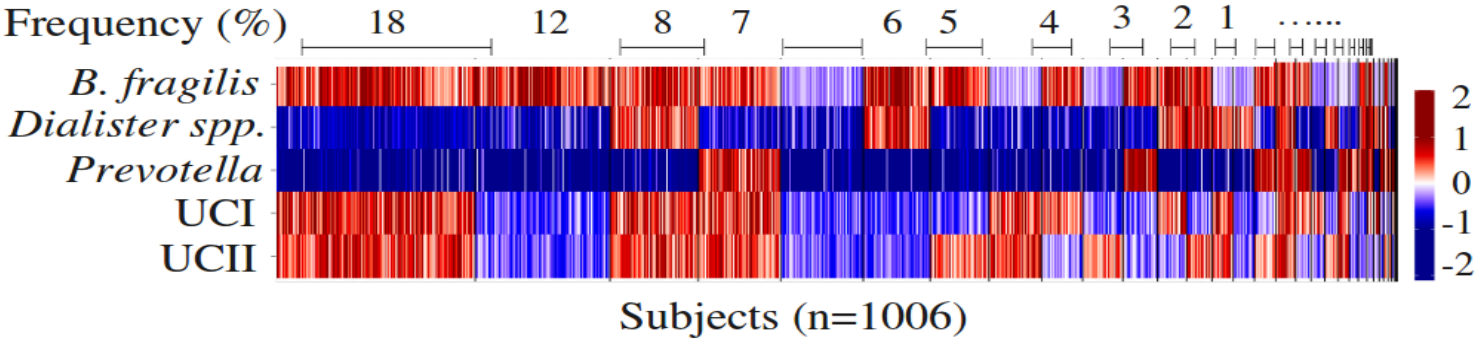
Indian Buffet Process: many *overlapping* clusters!

- Bayesian nonparametric model
- A prior on an infinite binary matrix.



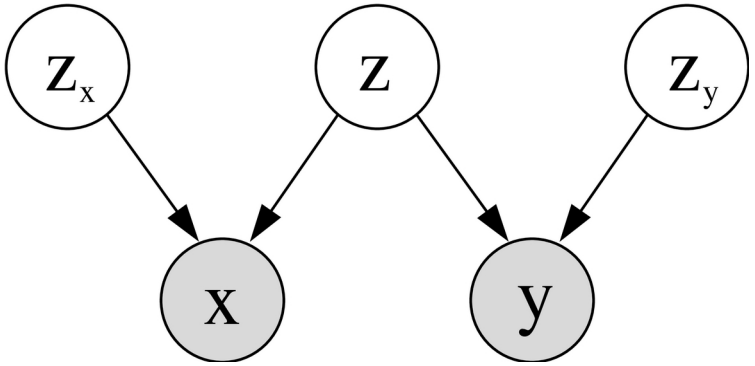


Michael Jordan



Data integration with latent variable models

probabilistic canonical correlation analysis:
a generative model with shared latent structure



$$\begin{cases} X = W_x \mathbf{z} + \varepsilon_x \\ Y = W_y \mathbf{z} + \varepsilon_y \end{cases}$$

Latent variable modeling for the microbiome FREE

Kris Sankaran ✉, Susan P Holmes

Biostatistics, kxy018, https://doi.org/10.1093/biostatistics/kxy018
Published: 03 June 2018 **Article history** ▾

Latent variable modeling for the microbiome FREE

Kris Sankaran ✉, Susan P Holmes

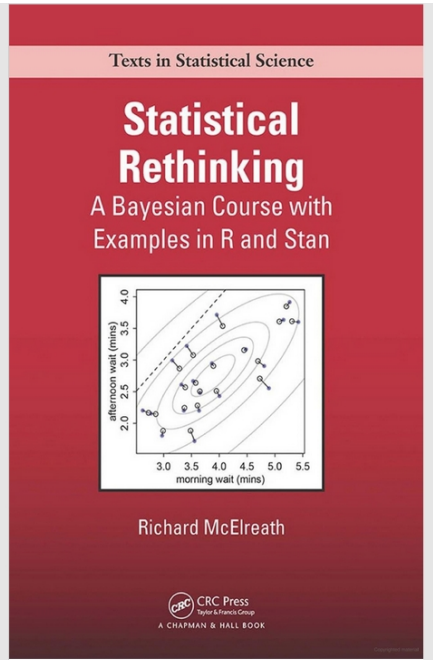
Biostatistics, kxy018, https://doi.org/10.1093/biostatistics/kxy018

Published: 03 June 2018 **Article history** ▾

SUMMARY

The human microbiome is a complex ecological system, and describing its structure and function under different environmental conditions is important from both basic scientific and medical perspectives. Viewed through a biostatistical lens, many microbiome analysis goals can be formulated as latent variable modeling problems. However, although probabilistic latent variable models are a cornerstone of modern unsupervised learning, they are rarely applied in the context of microbiome data analysis, in spite of the evolutionary, temporal, and count structure that could be directly incorporated through such models. We explore the application of probabilistic latent variable models to microbiome data, with a focus on Latent Dirichlet allocation, Non-negative matrix factorization, and Dynamic Unigram models. To develop guidelines for when different methods are appropriate, we perform a simulation study. We further illustrate and compare these techniques using the data of Dethlefsen and Relman (2011, Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences***108**, 4554–4561), a study on the effects of antibiotics on bacterial community composition. Code and data for all simulations and case studies are available publicly.

Figure 5: The online version provides the text in HTML, data files and up-to-date code.



Take-home messages

observations are always incomplete and noisy

generative models describe how the data is (presumably) generated

probabilistic programming can support applications

Latent variable modeling for the microbiome FREE

Kris Sankaran ✉, Susan P Holmes

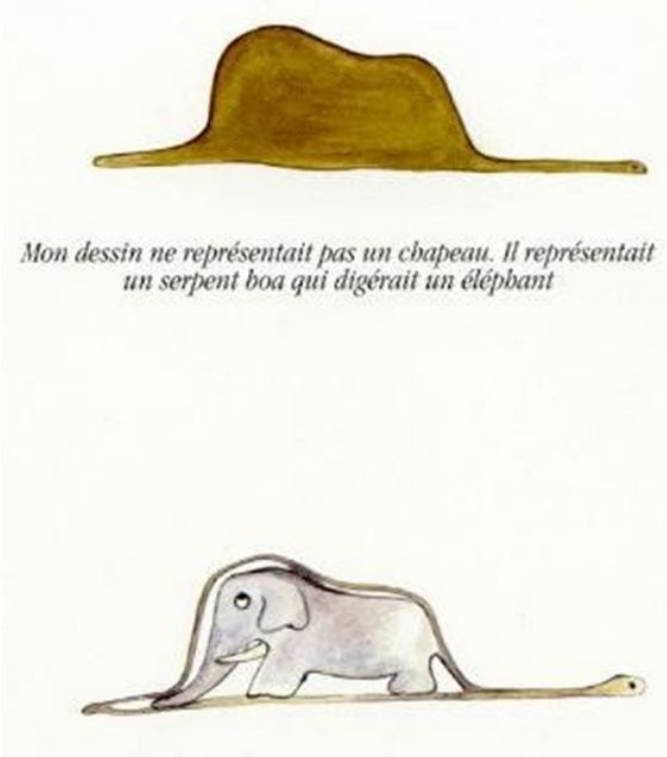
Biostatistics, kxy018, https://doi.org/10.1093/biostatistics/kxy018
Published: 03 June 2018 **Article history** ▾

PDF Split View Cite Permissions Share ▾

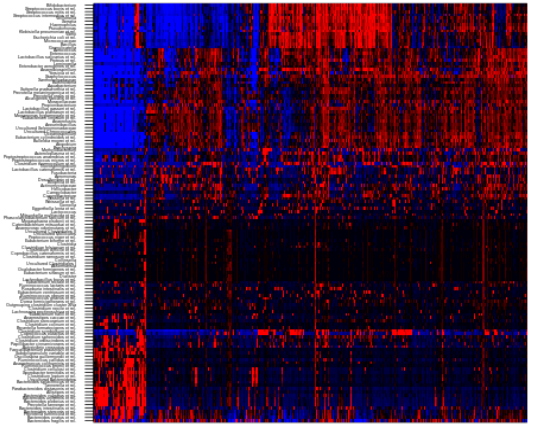
SUMMARY

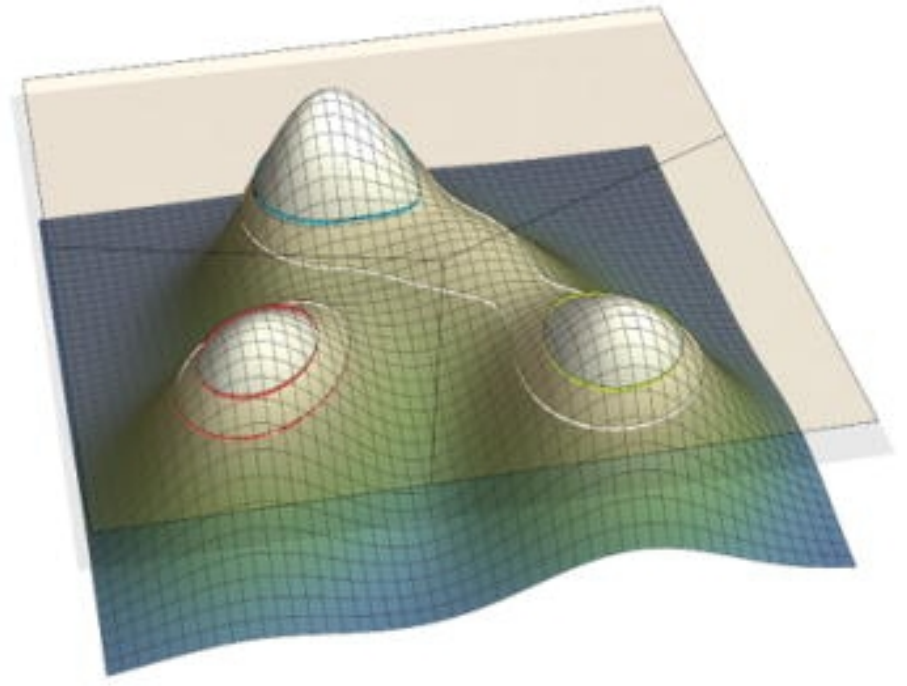
The human microbiome is a complex ecological system, and describing its structure and function under different environmental conditions is important from both basic scientific and medical perspectives. Viewed through a biostatistical lens, many microbiome analysis goals can be formulated as latent variable modeling problems. However, although probabilistic latent variable models are a cornerstone of modern unsupervised learning, they are rarely applied in the context of microbiome data analysis, in spite of the evolutionary, temporal, and count structure that could be directly incorporated through such models. We explore the application of probabilistic latent variable models to microbiome data, with a focus on Latent Dirichlet allocation, Non-negative matrix factorization, and Dynamic Unigram models. To develop guidelines for when different methods are appropriate, we perform a simulation study. We further illustrate and compare these techniques using the data of Dethlefsen and Relman (2011, Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences***108**, 4554–4561), a study on the effects of antibiotics on bacterial community composition. Code and data for all simulations and case studies are available publicly.

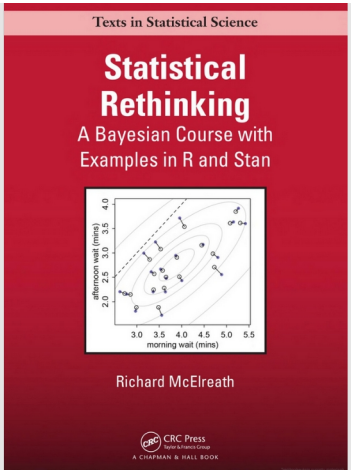
Figure 5: The online version provides the text in HTML, data files and up-to-date code.



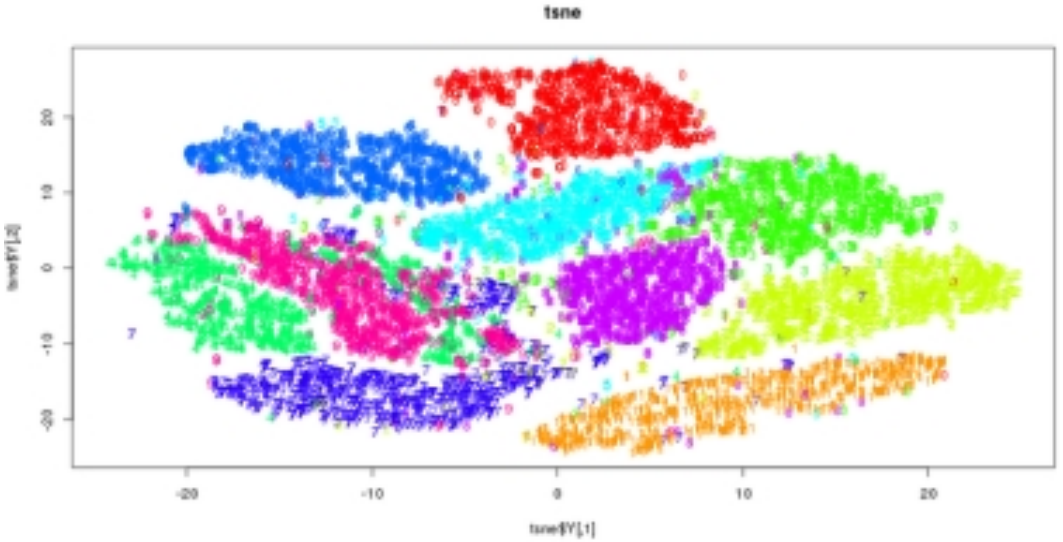
build explicit models for the data,
and compare *degree of evidence*
between models



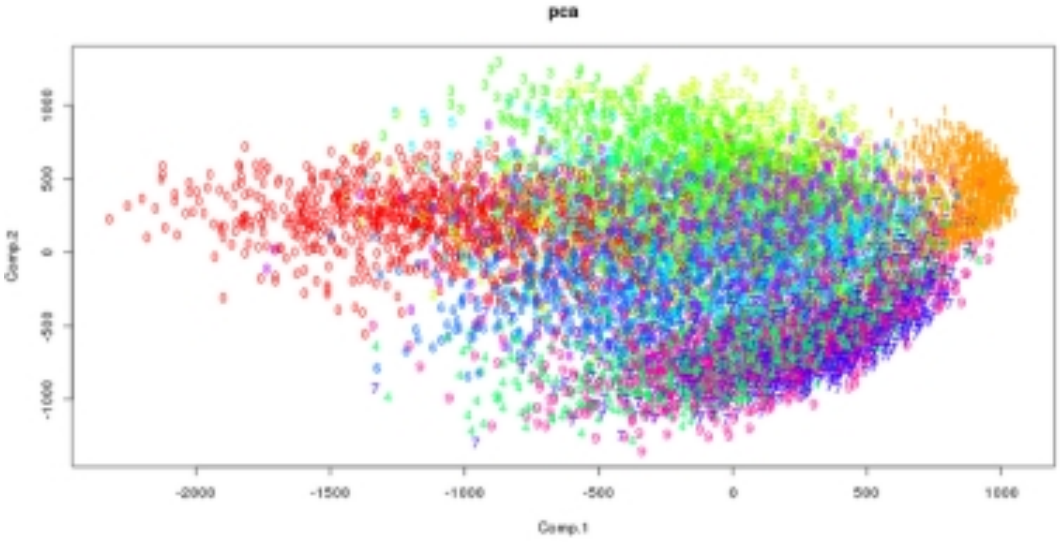




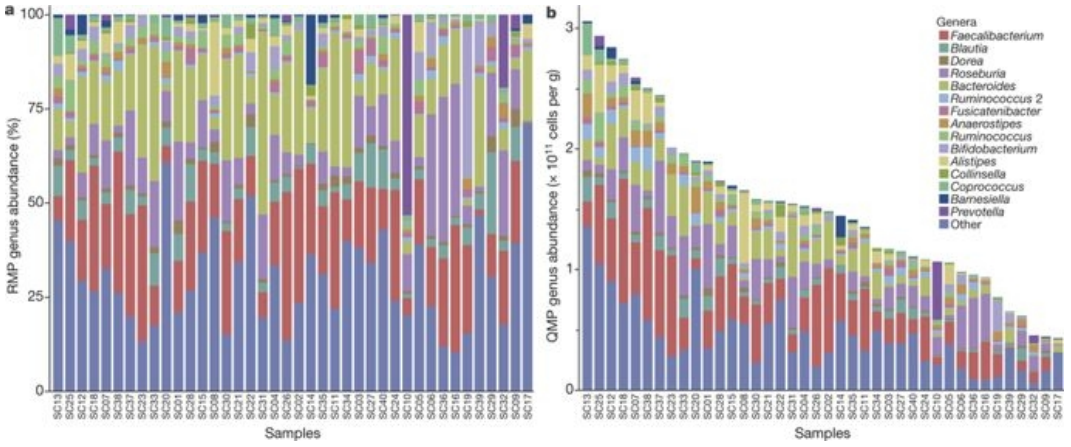
Stochastic Neighbor Embedding (t-SNE)

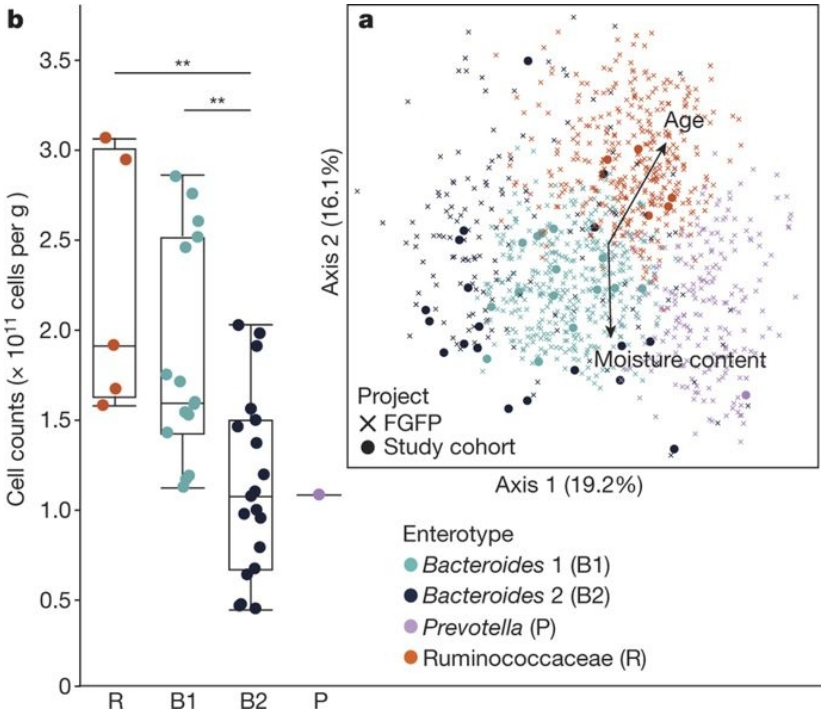


https://www.r-bloggers.com/playing-with-dimensions-from-clustering-pca-t-sne-to-carl-sagan/



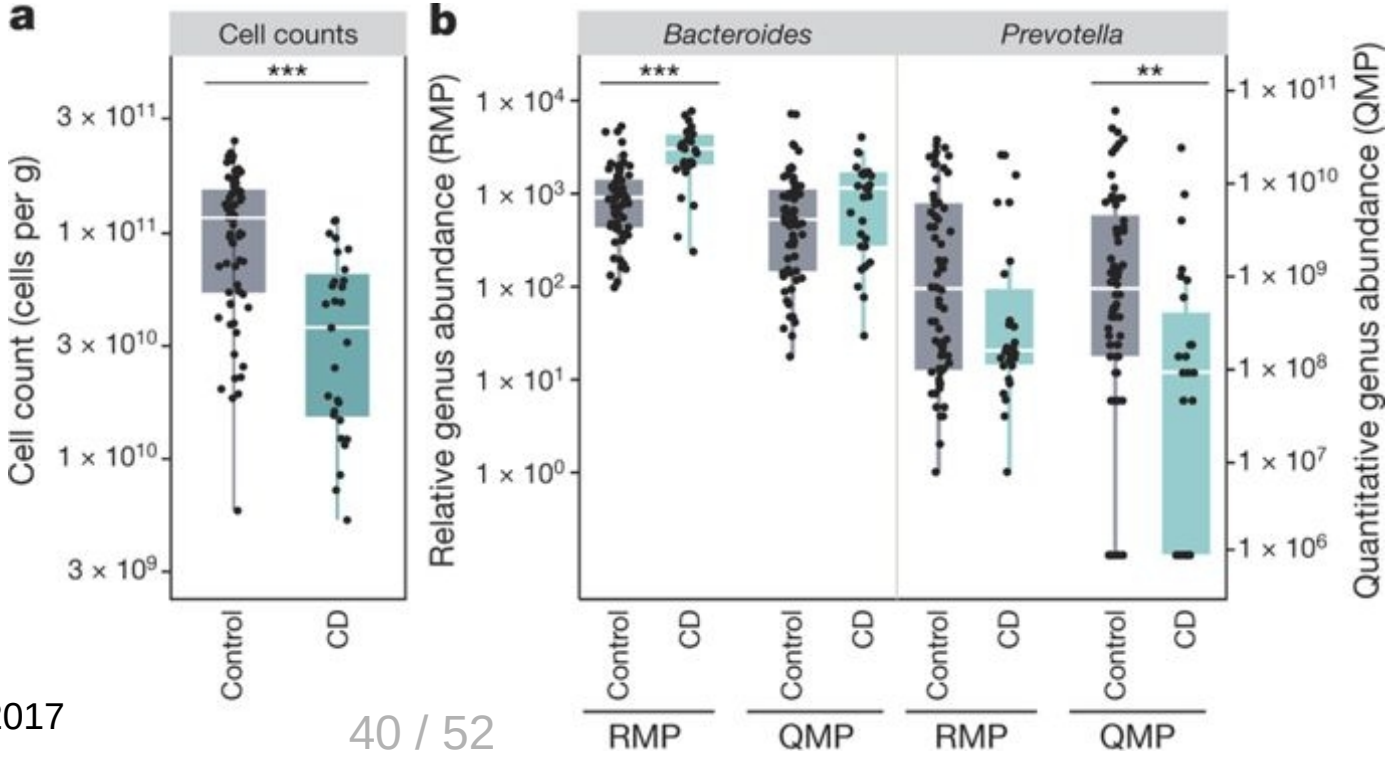
Relative versus absolute abundance: quantitative microbiome profiling





RMP vs. QMP:

drastic effect on conclusions!

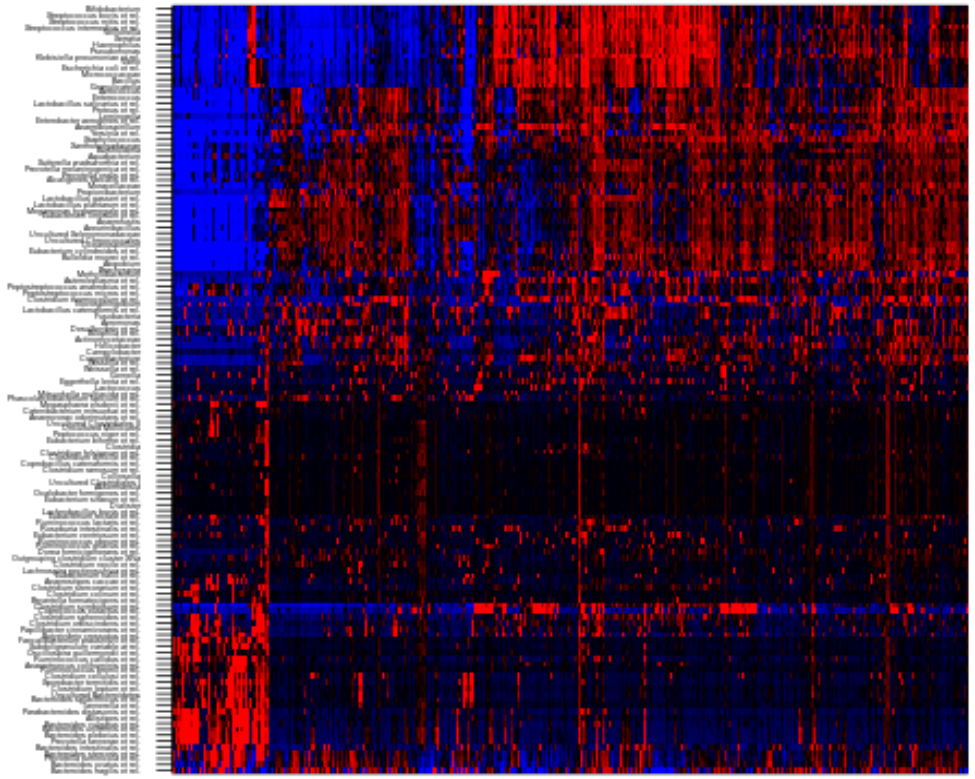


Vandeputte et al. Nature 551:507-511, 2017

Taxonomic abundance table

Individuals

Taxonomic groups



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

Latent variable modeling for the microbiome

Kris Sankaran, Susan P Holmes

Biostatistics, kxy018, https://doi.org/10.1093/biostatistics/kxy018
Published: 03 June 2018 **Article history** ▾

PDF Split View Cite Permissions Share ▾

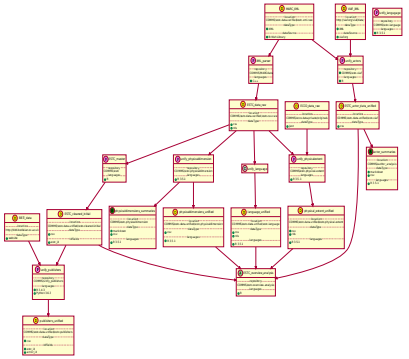
SUMMARY

The human microbiome is a complex ecological system, and describing its structure and function under different environmental conditions is important from both basic scientific and medical perspectives. Viewed through a biostatistical lens, many microbiome analysis goals can be formulated as latent variable modeling problems. However, although probabilistic latent variable models are a cornerstone of modern unsupervised learning, they are rarely applied in the context of microbiome data analysis, in spite of the evolutionary, temporal, and count structure that could be directly incorporated through such models. We explore the application of probabilistic latent variable models to microbiome data, with a focus on Latent Dirichlet allocation, Non-negative matrix factorization, and Dynamic Unigram models. To develop guidelines for when different methods are appropriate, we perform a simulation study. We further illustrate and compare these techniques using the data of Dethlefsen and Relman (2011, Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences***108**, 4554–4561), a study on the effects of antibiotics on bacterial community composition. Code and data for all simulations and case studies are available publicly.

Algorithmic bias: an extended view

Sources of bias:

- data
- algorithms
- users

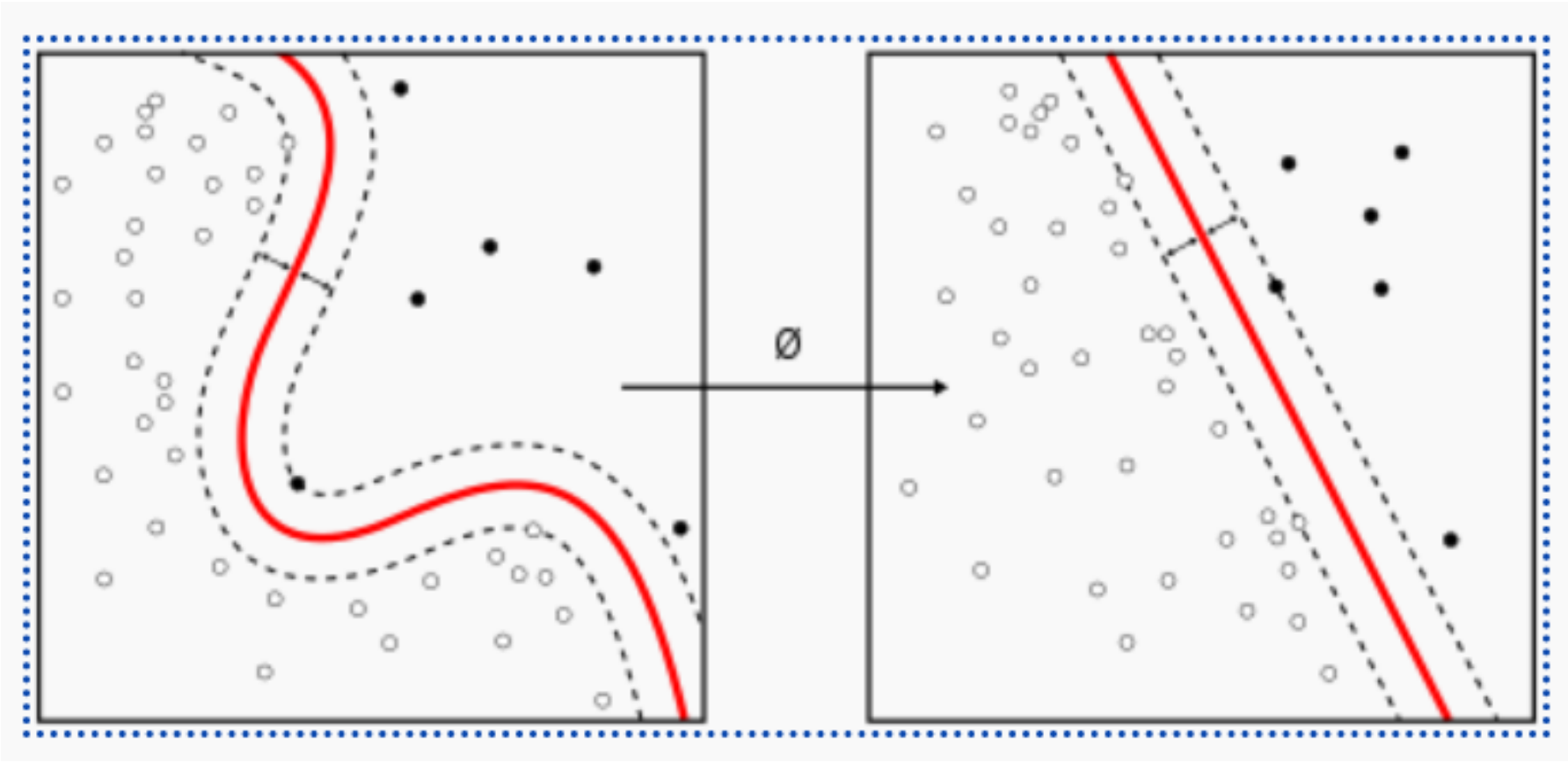


01000111
00010100
00110000
00111100

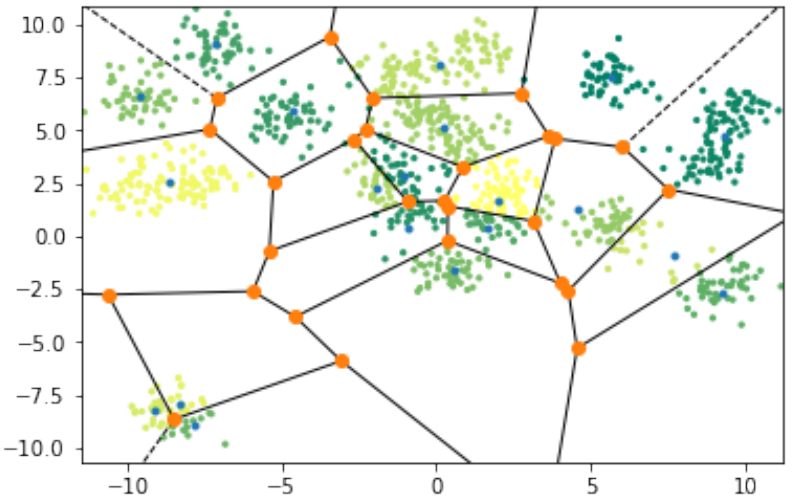
http://wiki.biomine.skelleftea.se/wiki/index.php/Wiki

(Some) types of bias:

- Selection bias
- Confirmation bias
- Over/underfitting
- Confounder bias

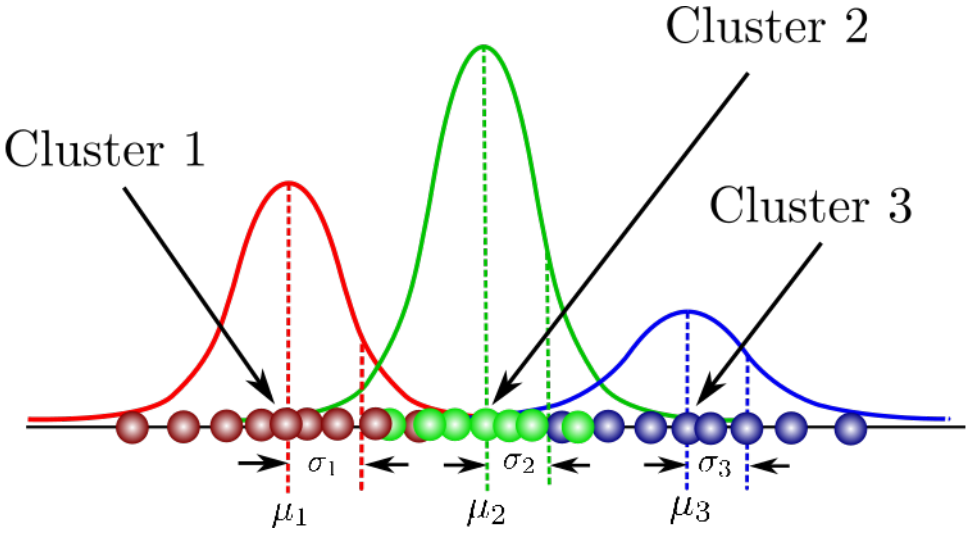

Ø

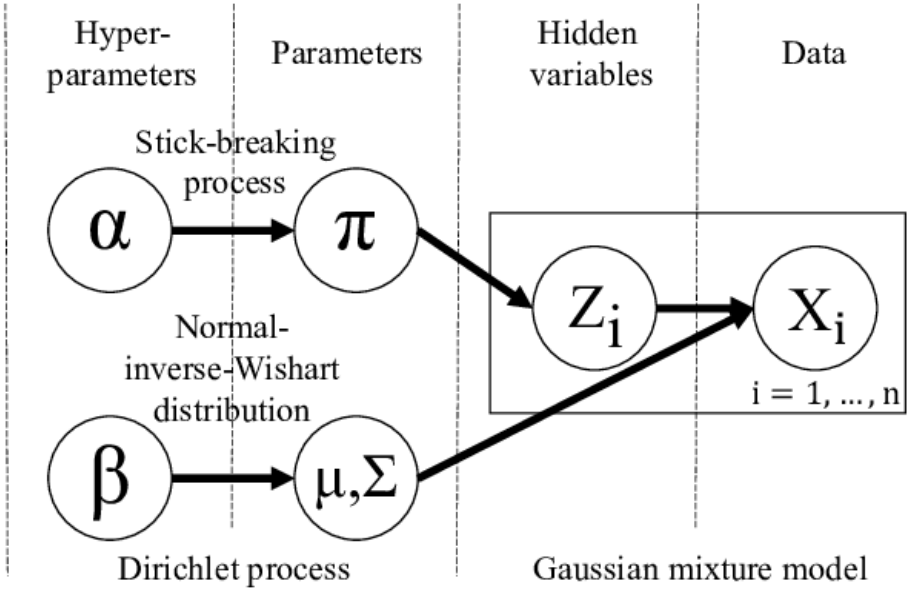
Non-parametric clustering:
Voronoi regions



Parametric clustering
(*latent class analysis*):

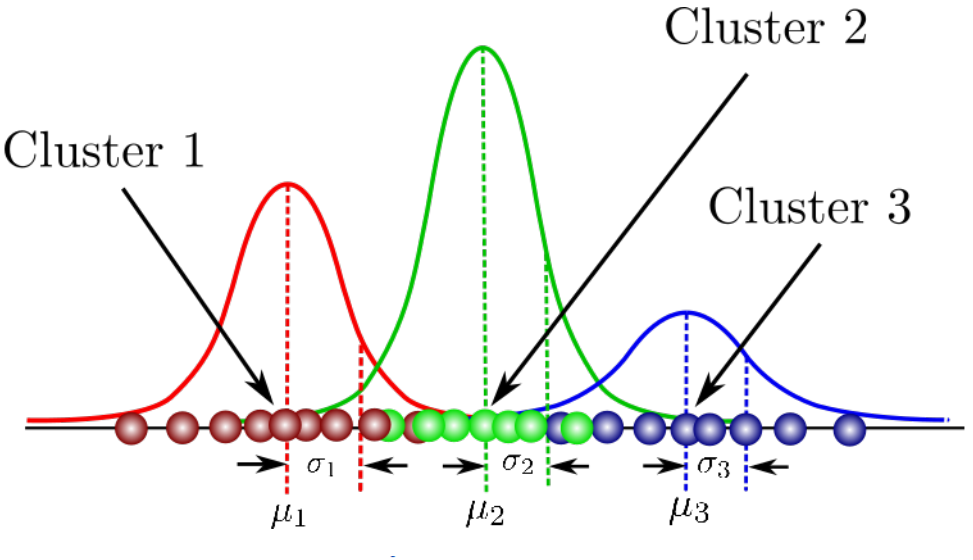
(Finite) Gaussian mixture





From learning *parameters* to learning the *model structure*: hierarchical generative models

B. Wu, S. Sakti, J. Zhang and S. Nakamura, "Tackling Perception Bias in Unsupervised Phoneme Discovery Using DPGMM-RNN Hybrid Model and Functional Load," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 348-362, 2021, doi: 10.1109/TASLP.2020.3042016.

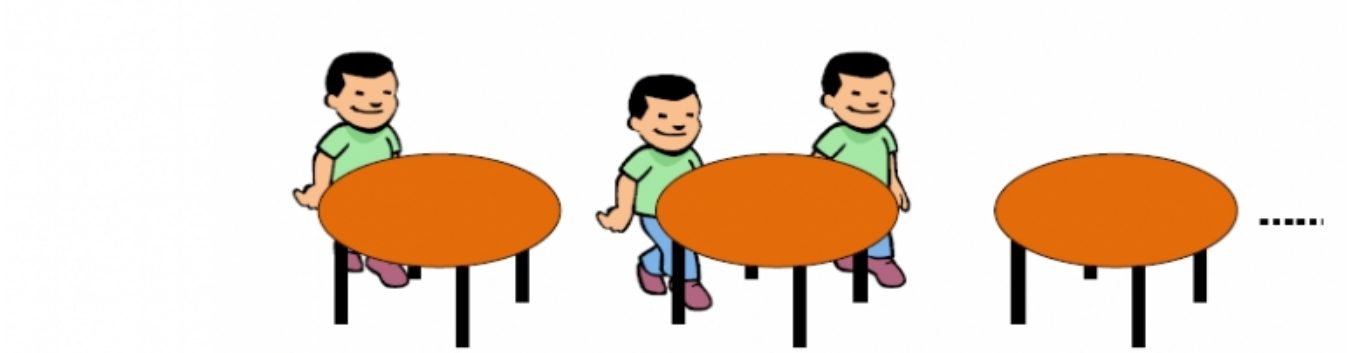


$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$

Dirichlet Process aka. Chinese Restaurant Process (CRP)

probabilistic infinite mixture model to find optimal number of clusters

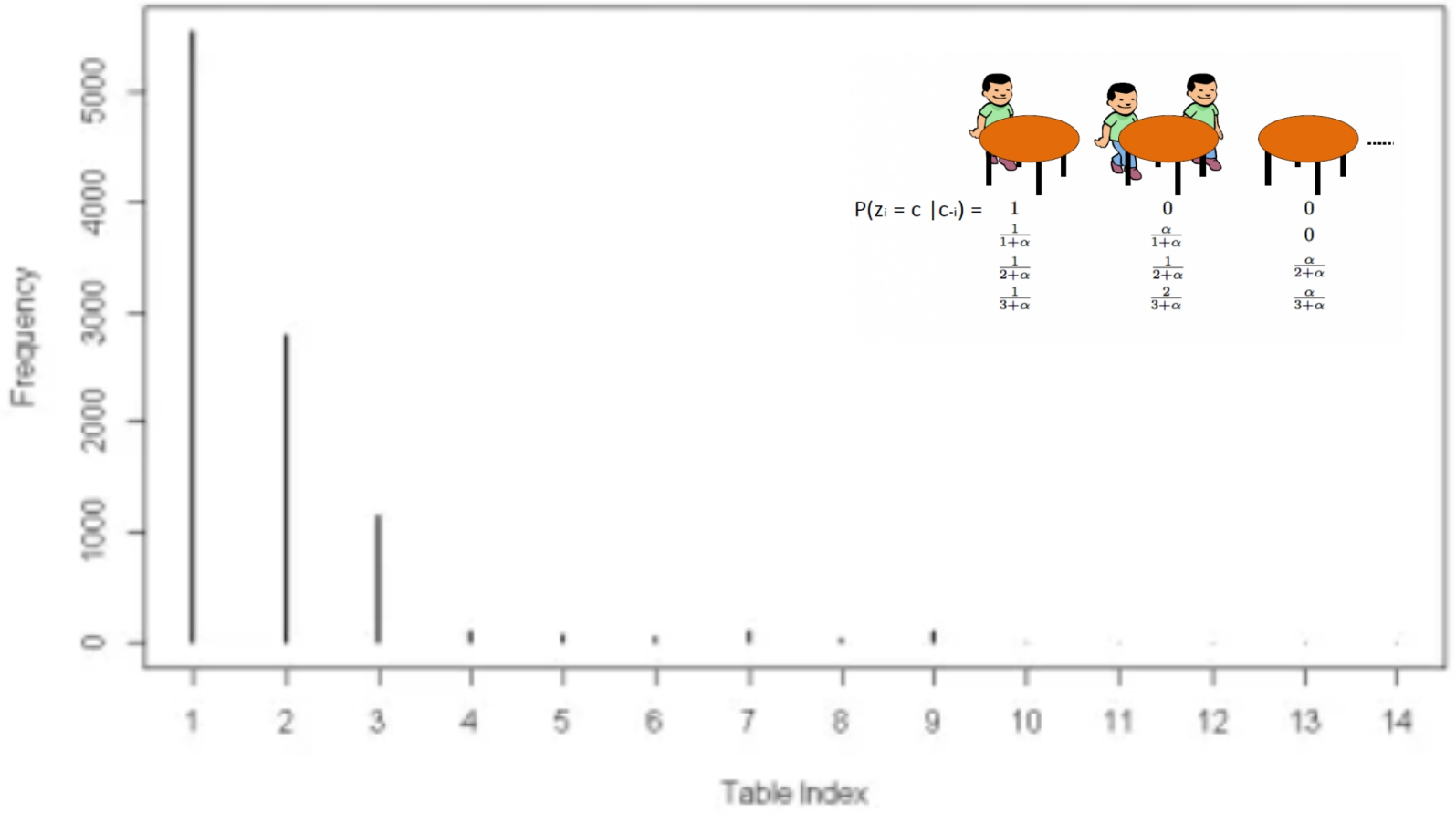
Fast detection of optimal cluster number based on explicit modeling assumptions and potentially infinitely many clusters



$$P(z_i = c \mid c_{-i}) = \begin{matrix} 1 & 0 & 0 \\ \frac{1}{1+\alpha} & \frac{\alpha}{1+\alpha} & 0 \\ \frac{1}{2+\alpha} & \frac{1}{2+\alpha} & \frac{\alpha}{2+\alpha} \\ \frac{1}{3+\alpha} & \frac{2}{3+\alpha} & \frac{\alpha}{3+\alpha} \end{matrix}$$

blog.datumbox.com/the-dirichlet-process-the-chinese-restaurant-process-and-other-representations/

Higher cluster numbers are possible but unlikely



https://www.r-bloggers.com/2013/04/dirichlet-process-infinite-mixture-models-and-clustering/

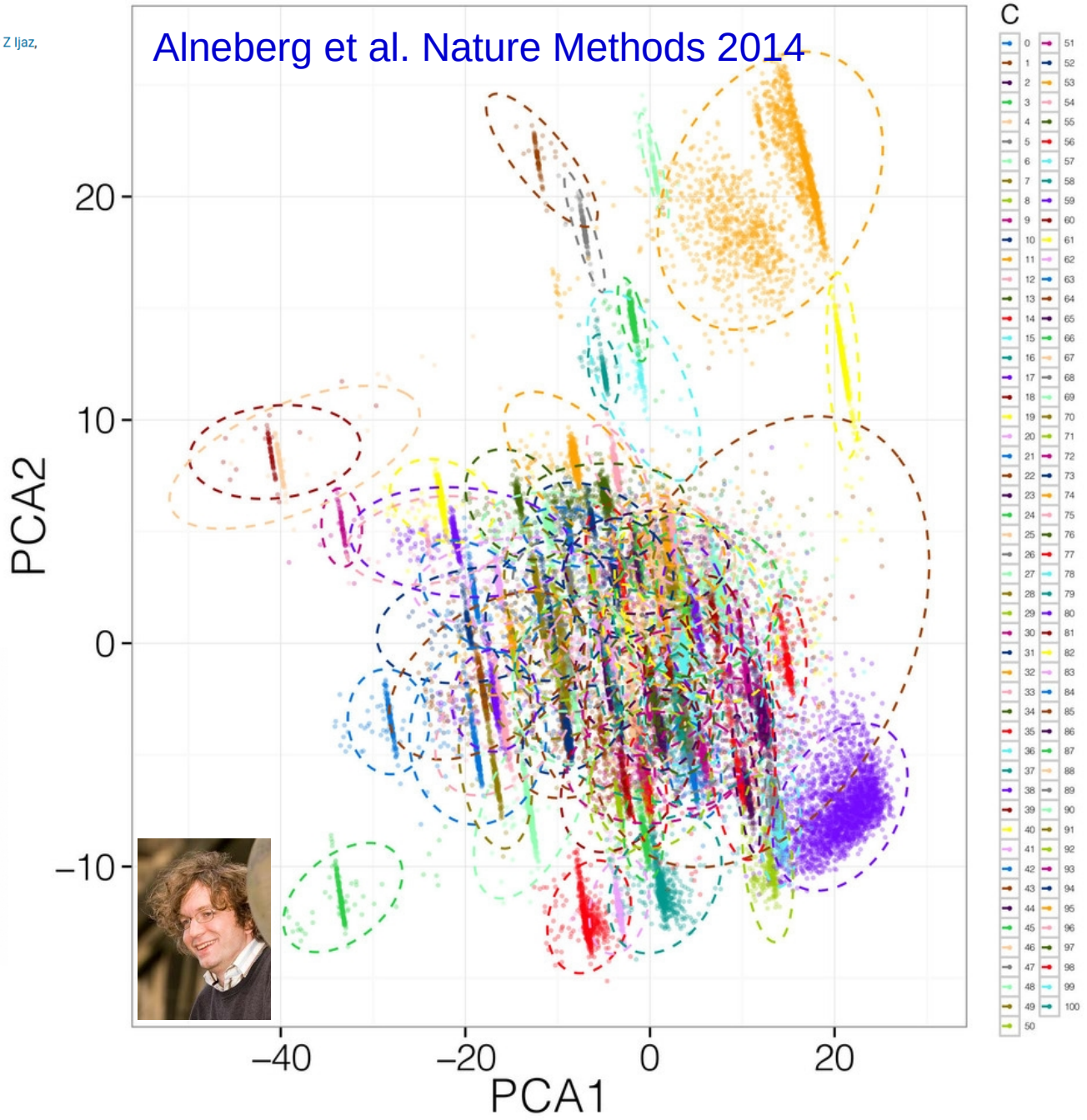
Published: 14 September 2014

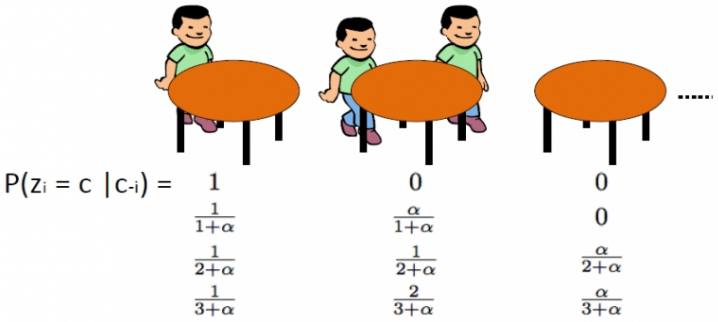
Binning metagenomic contigs by coverage and composition

Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson ✉ & Christopher Quince ✉

Nature Methods **11**, 1144–1146(2014) | Cite this article

Mock community: genomes by color; ellipses indicate identified clusters.



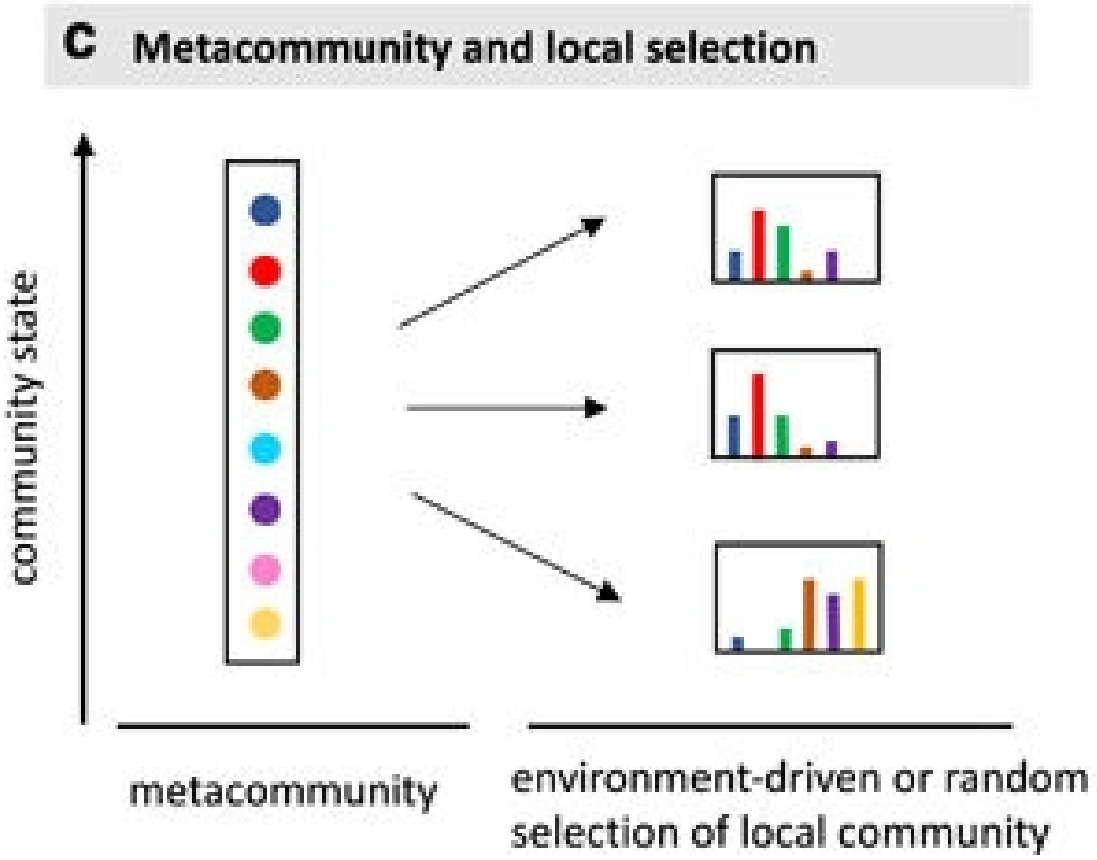


Open Access | Published: 05 May 2017

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



Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process [1]

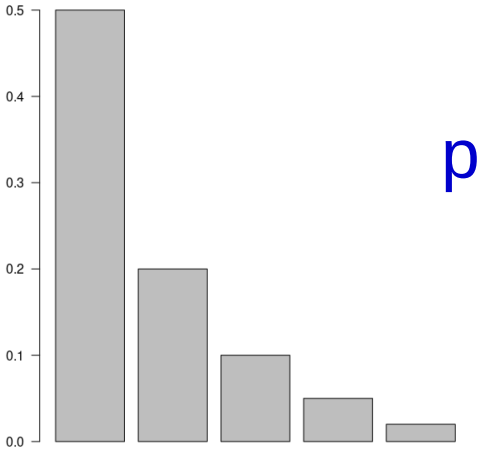
Keith Harris[1], Todd L Parsons[2], Umer Z Ijaz[3], Leo Lahti[4], Ian Holmes[5], Christopher Quince[6,*]

$$X_i \mid N_i, p_i \sim MN(N_i, p_i)$$
$$p_i \mid I_i, \beta \sim DP(I_i, \beta)$$
$$\beta \mid \theta \sim Stick(\theta) \sim DP(\theta, 1)$$

Sampling from the multinomial

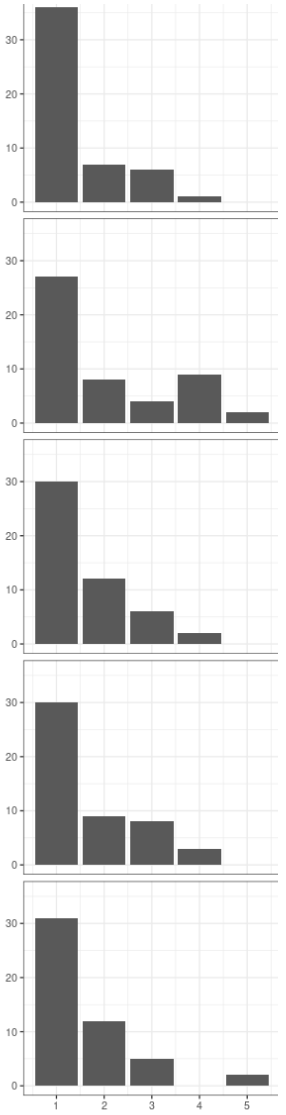
Stochastic (observed) realizations x (50 reads)

"Real" latent distribution p (of species abundances)

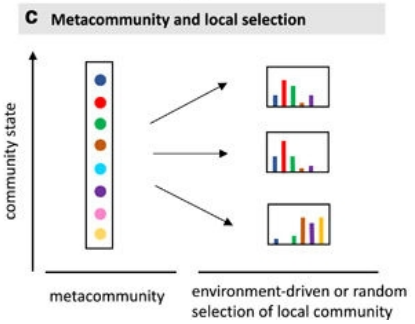


p → x

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




```
ggplot(melt(
  rmultinom(n = 5, size = 50,
      prob = c(0.5, 0.2, 0.1, 0.05, 0.02))),
  aes(x = X1, y = value)) +
  geom_col() + facet_grid(X2 ~ 1)
```



Dirichlet-Multinomial as the observation model;
alpha gives prior for species distribution
(“biodiversity” parameter).

observed species read counts

“true” species abundance p

uncertainty of p

$$\Pr(\mathbf{x} \mid \boldsymbol{\alpha}) = \int_{\mathbf{p}} \Pr(\mathbf{x} \mid \mathbf{p}) \Pr(\mathbf{p} \mid \boldsymbol{\alpha}) \mathrm{d}\mathbf{p}$$

Multinomial

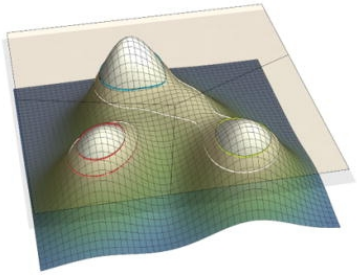
Dirichlet

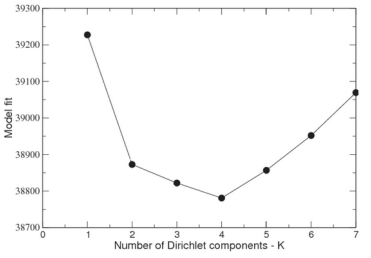
Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics

Ian Holmes, Keith Harris, Christopher Quince

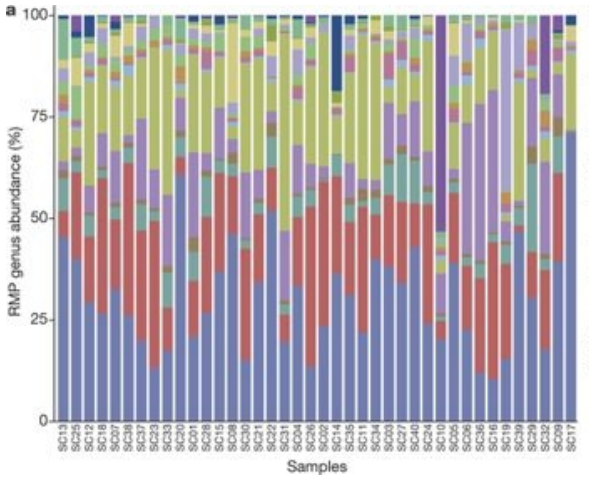
Published: February 3, 2012 • https://doi.org/10.1371/journal.pone.0030126

$$P(\bar{p}_i|Q) = \sum_{k=1}^{K} \mathrm{Dir}(\bar{p}_i|\bar{\alpha}_k)\pi_k,$$





Replacing the single Dirichlet prior with a mixture of K Dirichlet components



Problems: syntheny, Niche neutrality, ecosystem-level clusters, K depends on sample size..