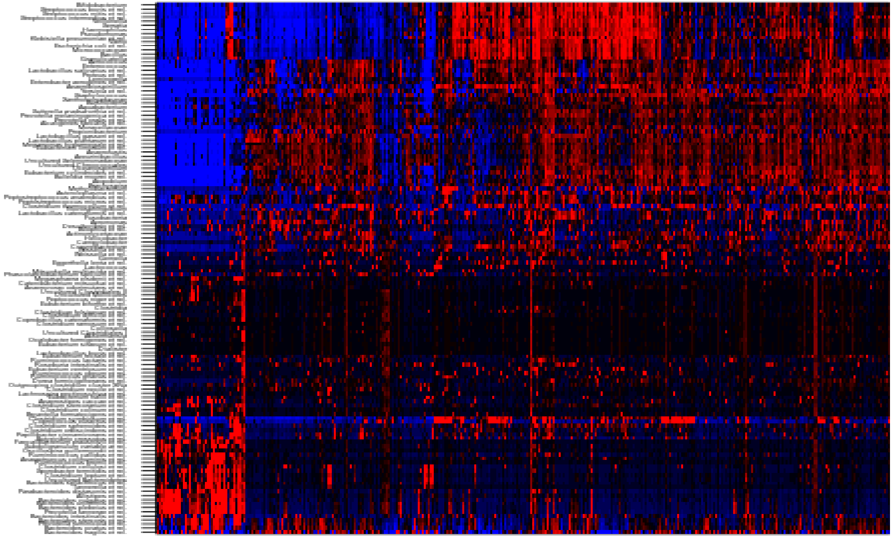
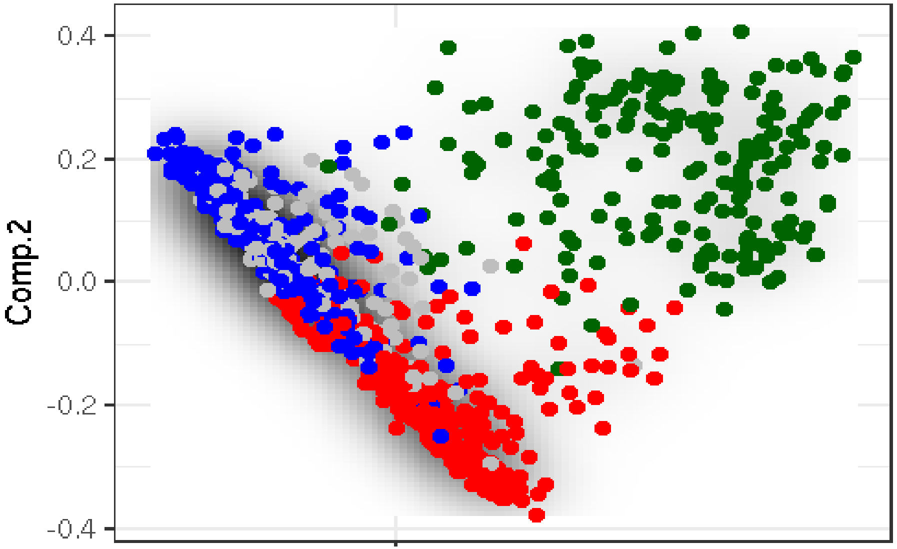
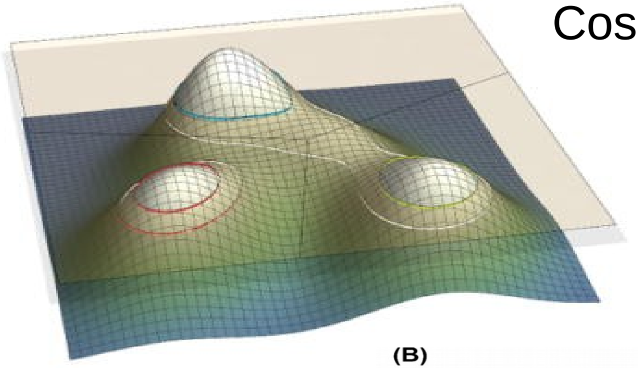
Individuality and population variation



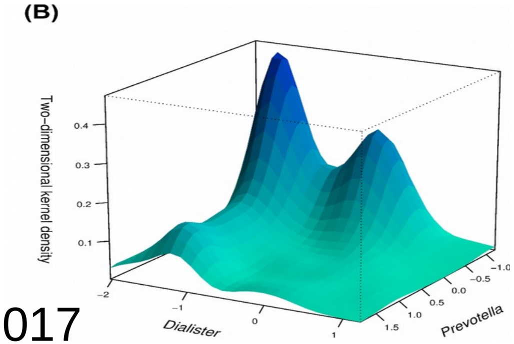


Landscape model

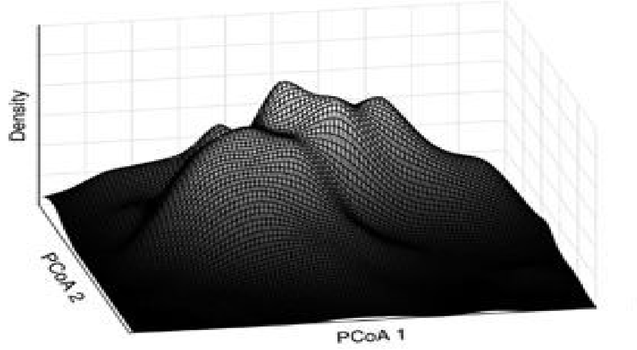
- Concept
- Relevance
- Criticism
- Alternatives
- Modeling & visualization



Costea et al. 2018



Shetty et al. 2017



Falony et al. 2016

Common study designs

Cross-sectional

population (cohort) studies

Prospective

long-term follow-ups

Longitudinal

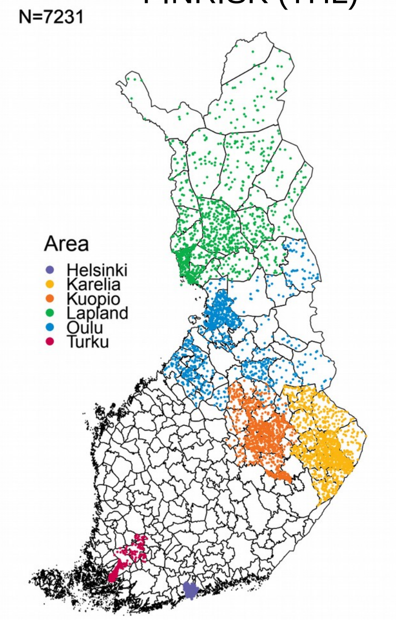
ecosystem dynamics

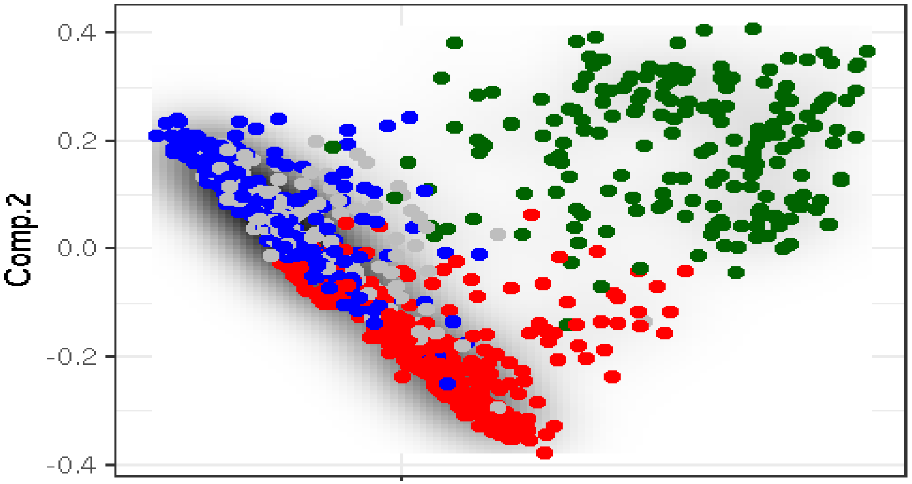
Case-control & Intervention

targeted experimental testing

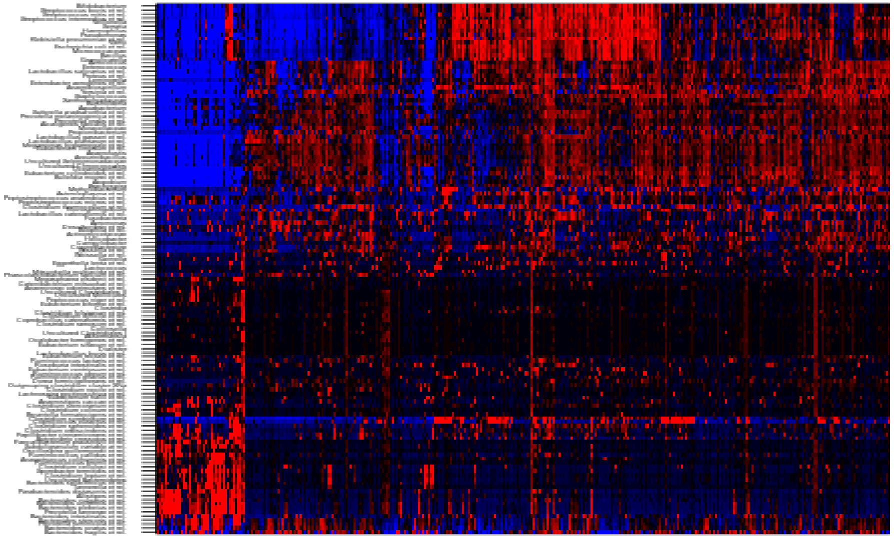
Cross-sectional

FINRISK (THL)

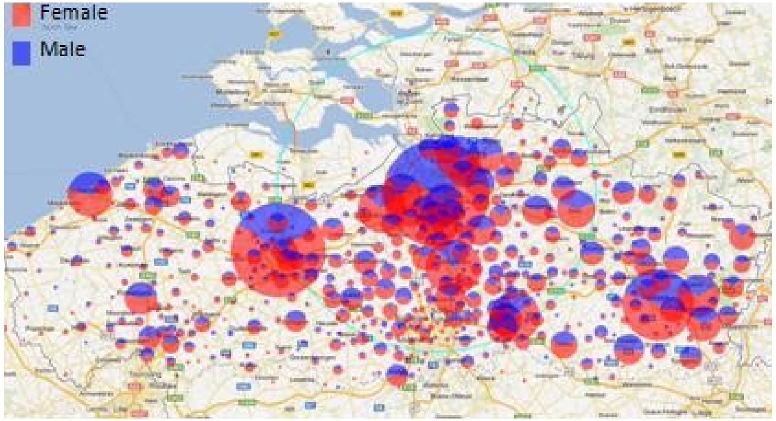


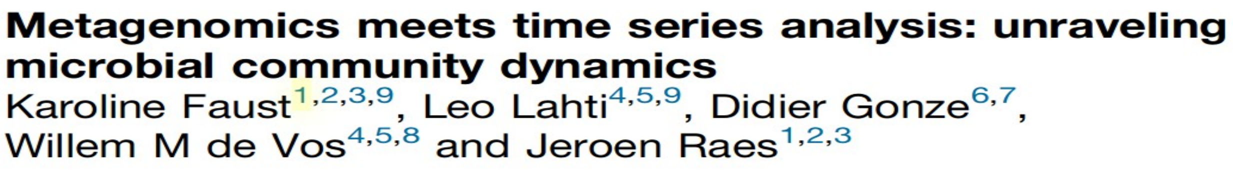


HITChip Atlas (Wageningen



Flemish Gut Flora (Raes Lab)

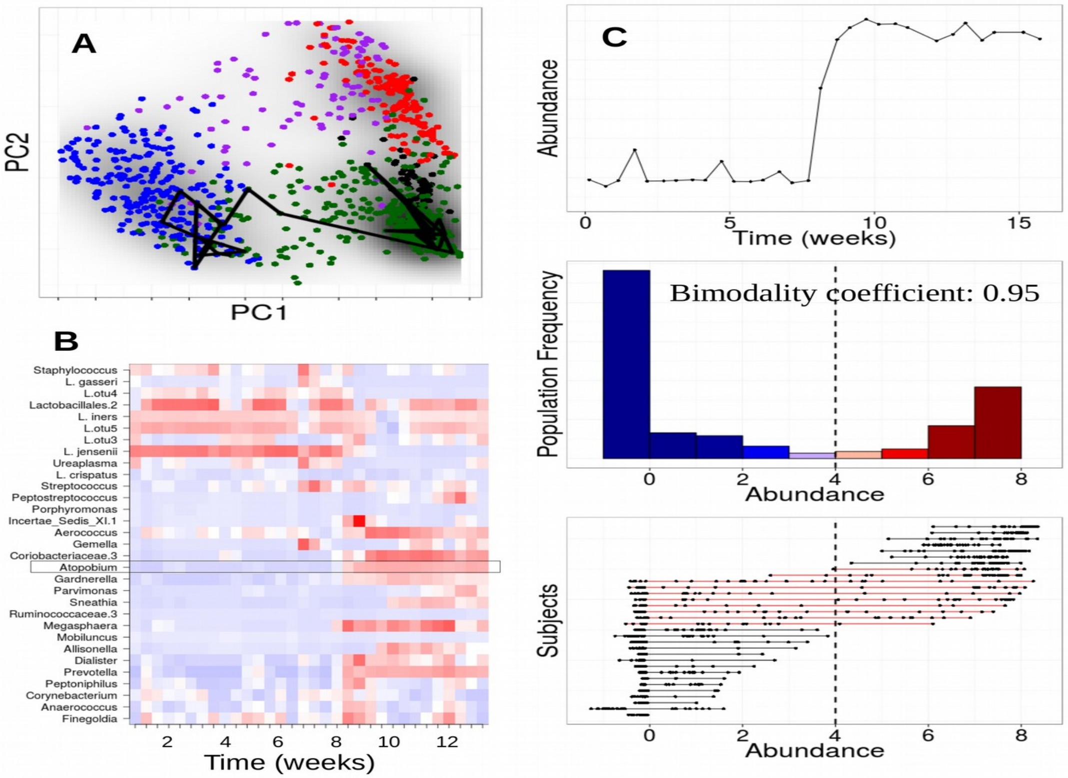




Metagenomics meets time series analysis: unraveling microbial community dynamics

Karoline Faust[1,2,3,9], Leo Lahti[4,5,9], Didier Gonze[6,7], Willem M de Vos[4,5,8] and Jeroen Raes[1,2,3]

Longitudinal

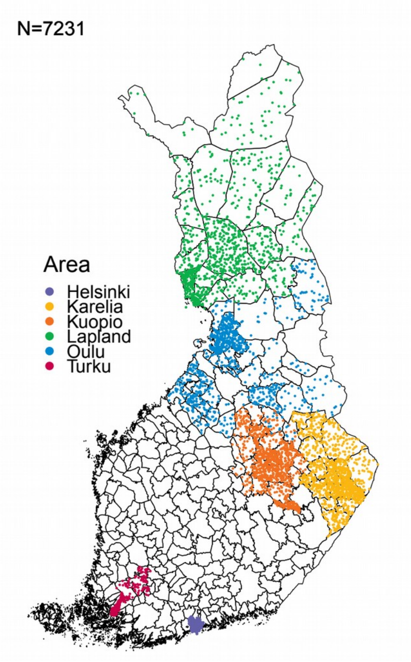


Data: Vaginal microbiota time series from 32 women. Gajer *et al. 2012*

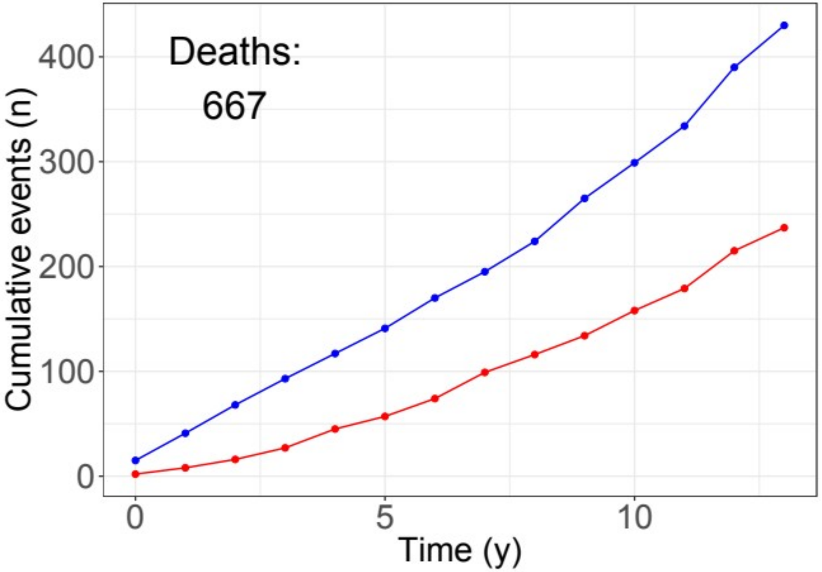
15 year **prospective** view (Finland / FINRISK2002)

2002 ~7000+ stool samples: omics and clinical measurements.

2017 comprehensive health information from Finnish registers



15+ year follow up



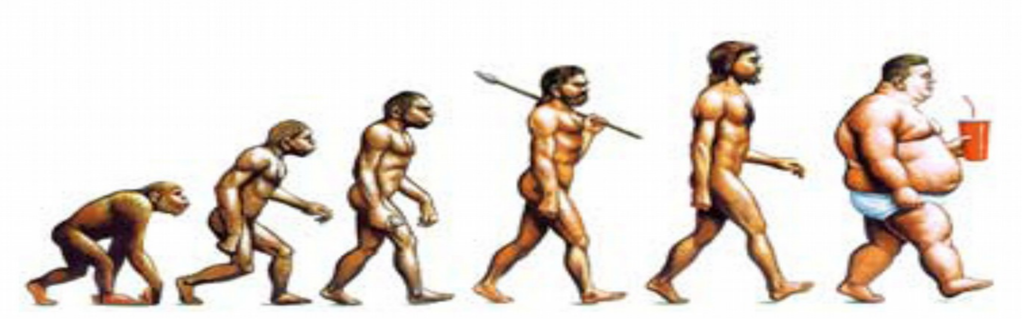
With: Teemu Niiranen, Veikko Salomaa, Rob Knight, Mohit Jain, Mike Inouye, Aaro Salosensaari, Ville Laitinen, and others

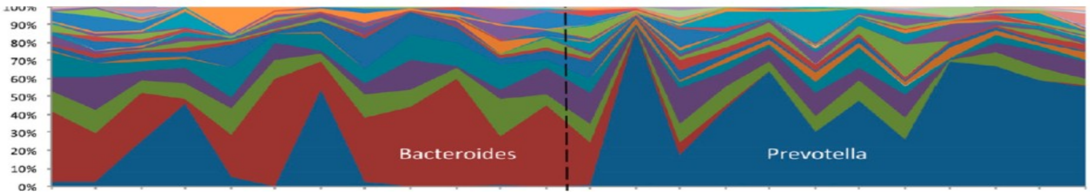
Medical research

Bowel cancer risk may be reduced by rural African diet, study finds

Tests on subjects who swapped a fatty, meat-heavy diet for foods rich in beans and vegetables found a drop in biological markers for cancer in just two weeks

Case-control & Intervention





NATURE COMMUNICATIONS | ARTICLE

Fat, fibre and cancer risk in African Americans and rural Africans

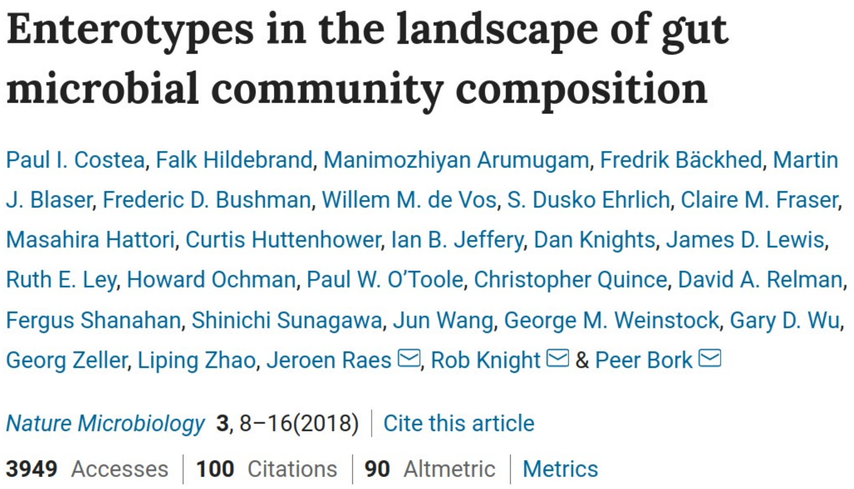
Stephen J. D. O'Keefe, Jia V. Li, Leo Lahti, Junhai Ou, Franck Carbonero, Khaled Mohammed, Joram M. Posma, James Kinross, Elaine Wahl, Elizabeth Ruder, Kishore Vipperla, Vasudevan Naidoo, Lungile Mtshali, Sebastian Tims, Philippe G. B. Puylaert, James DeLany, Alyssa Krasinskas, Ann C. Benefiel, Hatem O. Kaseb, Keith Newton + *et al.*

Affiliations | Contributions | Corresponding author

Nature Communications **6**, Article number: 6342 | doi:10.1038/ncomms7342
Received 23 May 2014 | Accepted 20 January 2015 | Published 28 April 2015

The importance of research question

- Questions guide experimental design
- Exploratory analysis of earlier data can be useful for hypothesis generation



Enterotypes in the landscape of gut microbial community composition

Paul I. Costea, Falk Hildebrand, Manimozhiyan Arumugam, Fredrik Bäckhed, Martin J. Blaser, Frederic D. Bushman, Willem M. de Vos, S. Dusko Ehrlich, Claire M. Fraser, Masahira Hattori, Curtis Huttenhower, Ian B. Jeffery, Dan Knights, James D. Lewis, Ruth E. Ley, Howard Ochman, Paul W. O'Toole, Christopher Quince, David A. Relman, Fergus Shanahan, Shinichi Sunagawa, Jun Wang, George M. Weinstock, Gary D. Wu, Georg Zeller, Liping Zhao, Jeroen Raes ✉, Rob Knight ✉ & Peer Bork ✉

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

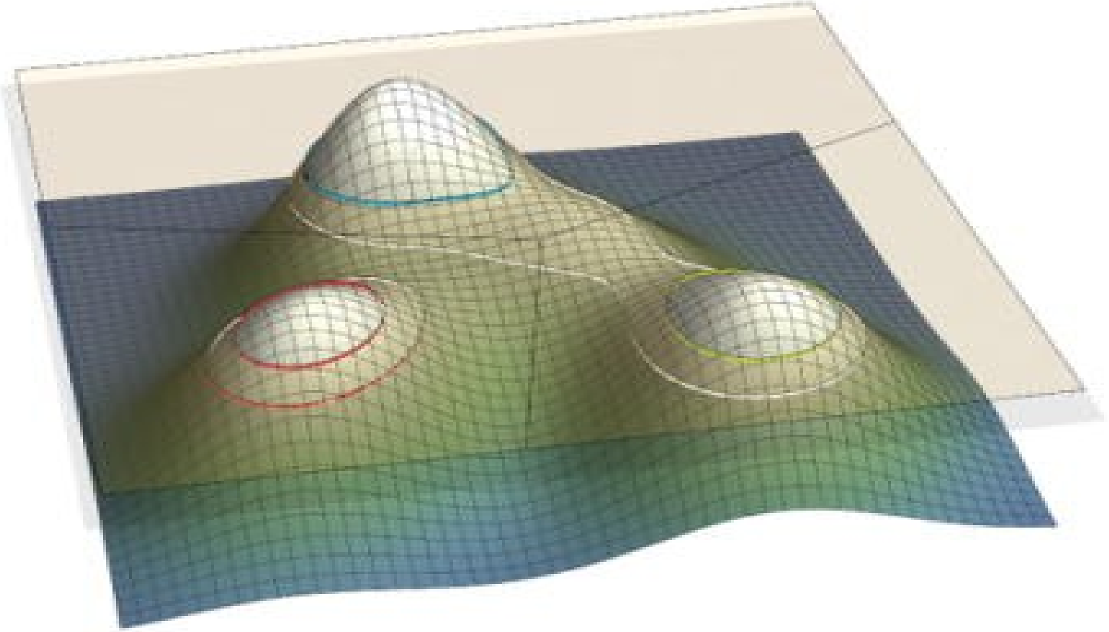
3949 Accesses | **100** Citations | **90** Altmetric | Metrics

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

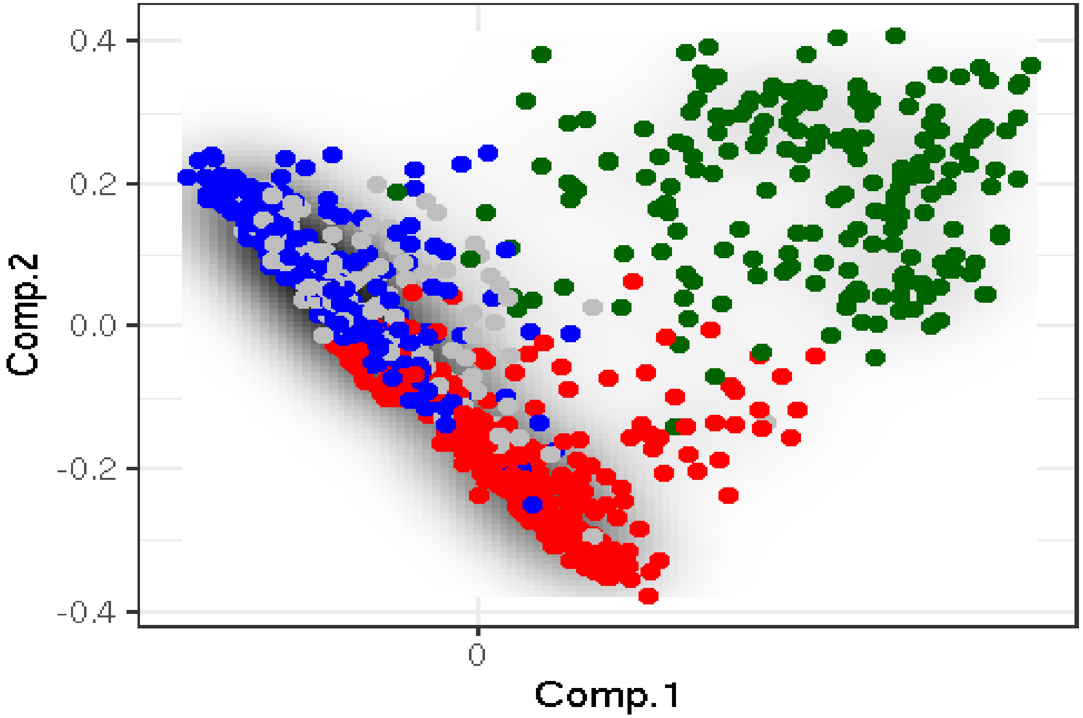
FEMS Microbiology Reviews, Volume 41, Issue 2, 1 March 2017, Pages 182–199, https://doi.org/10.1093/femsre/fuw045

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

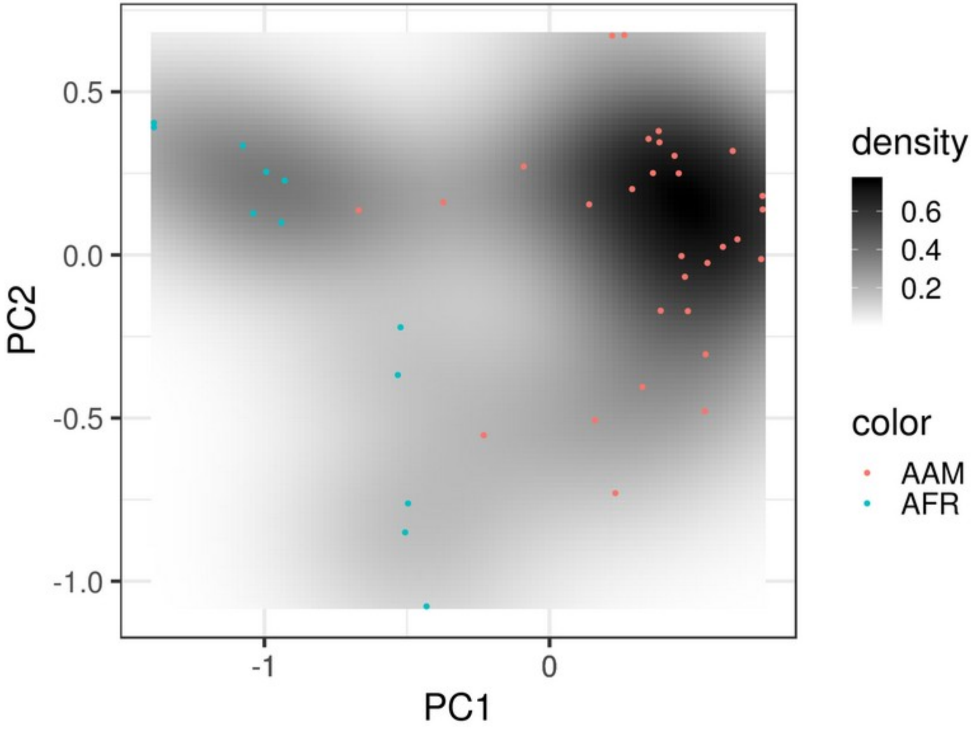


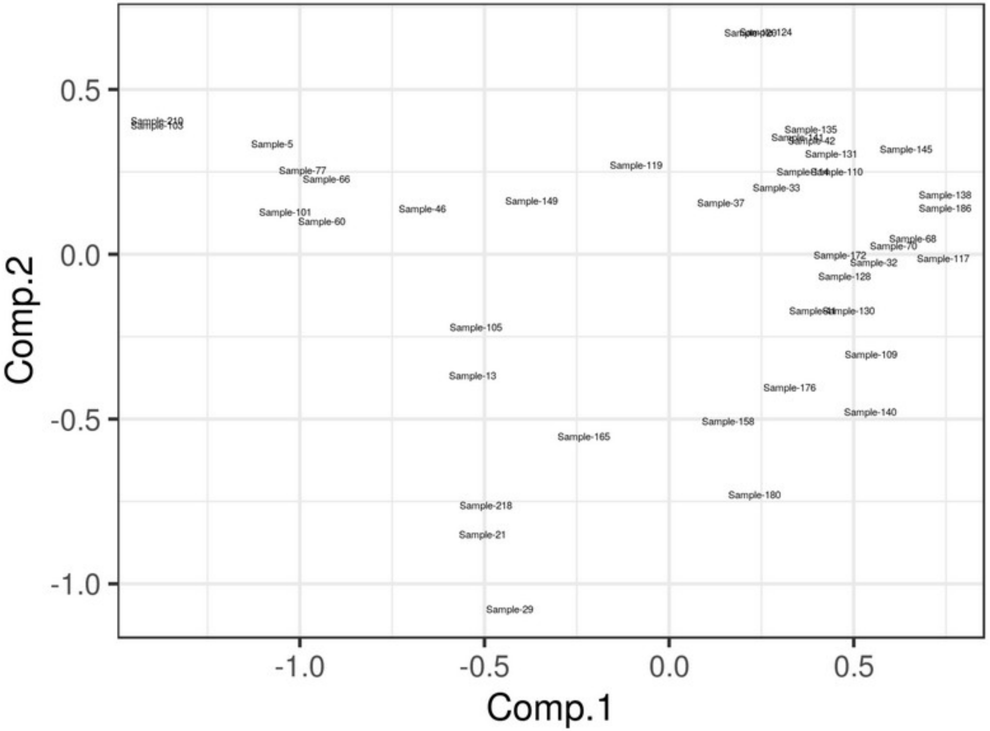
Gut microbiota landscape

HITChip
(N = 1006)



Kernel density estimates & Parzen windows



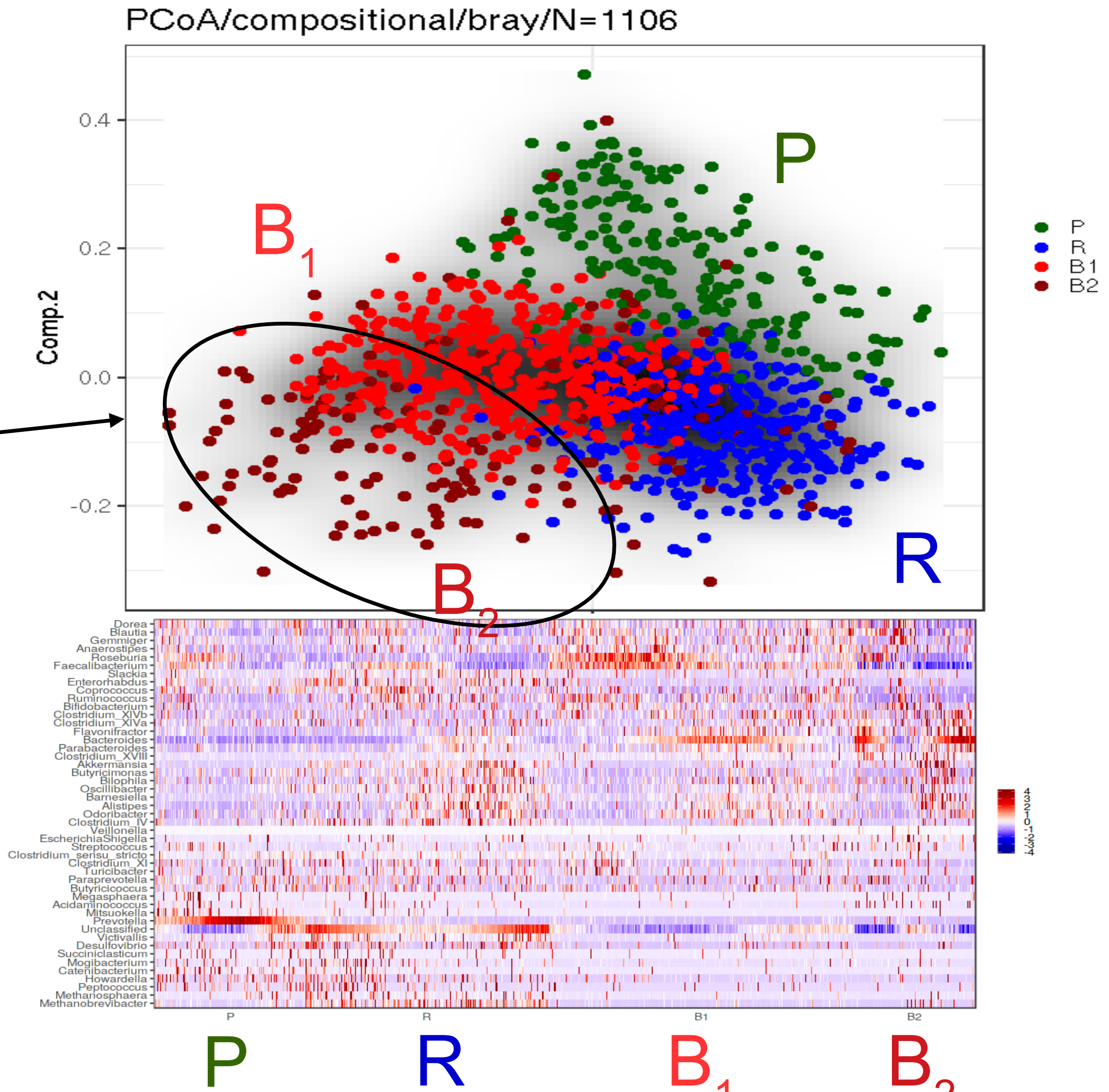


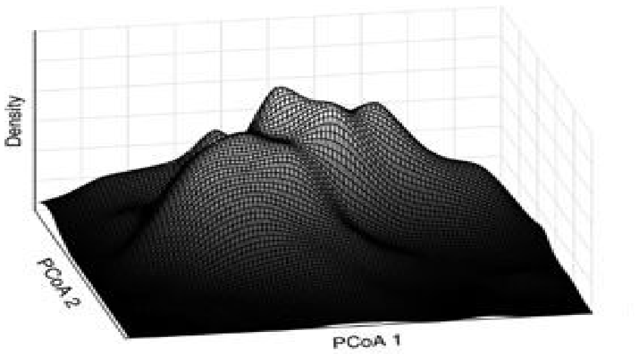
Broad community types (Flemish Gut Flora)

Bacteroides splits in two subtypes

B_2 subtype associated with IBD & low cell count

(Vandeputte et al. Nature 2017)

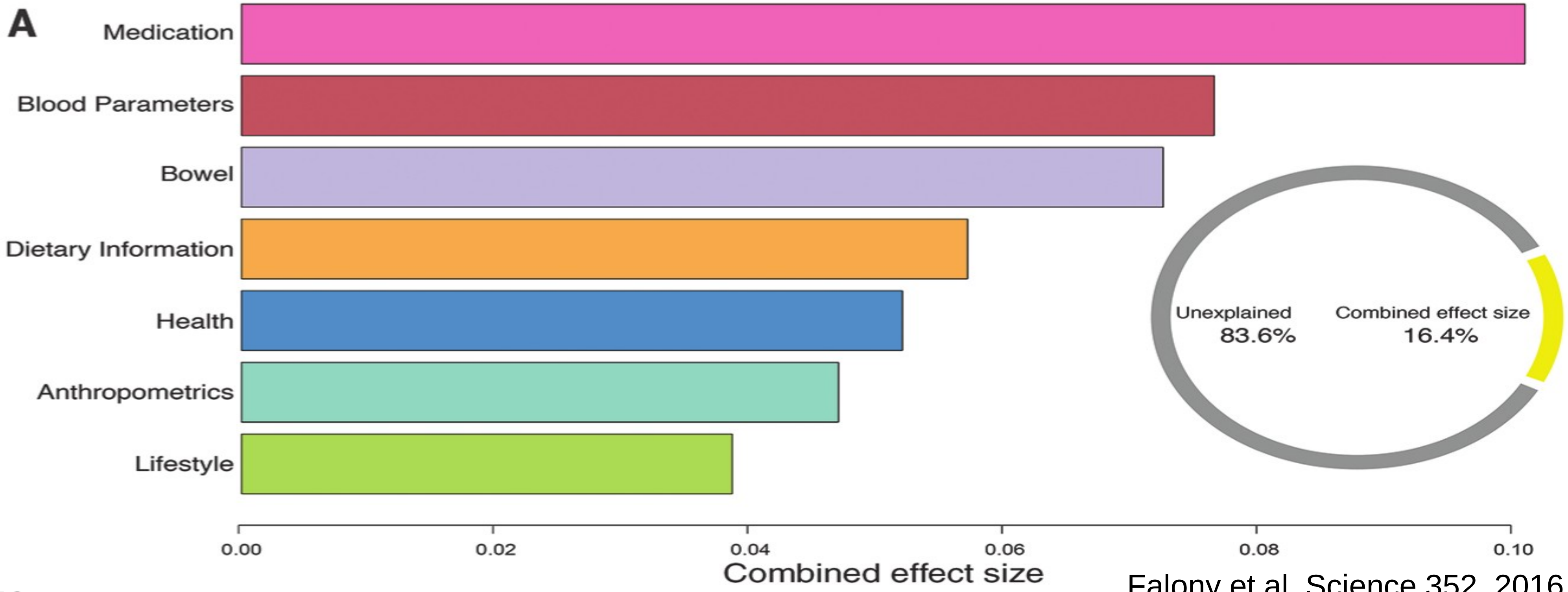




Total explained variation: 16.4%

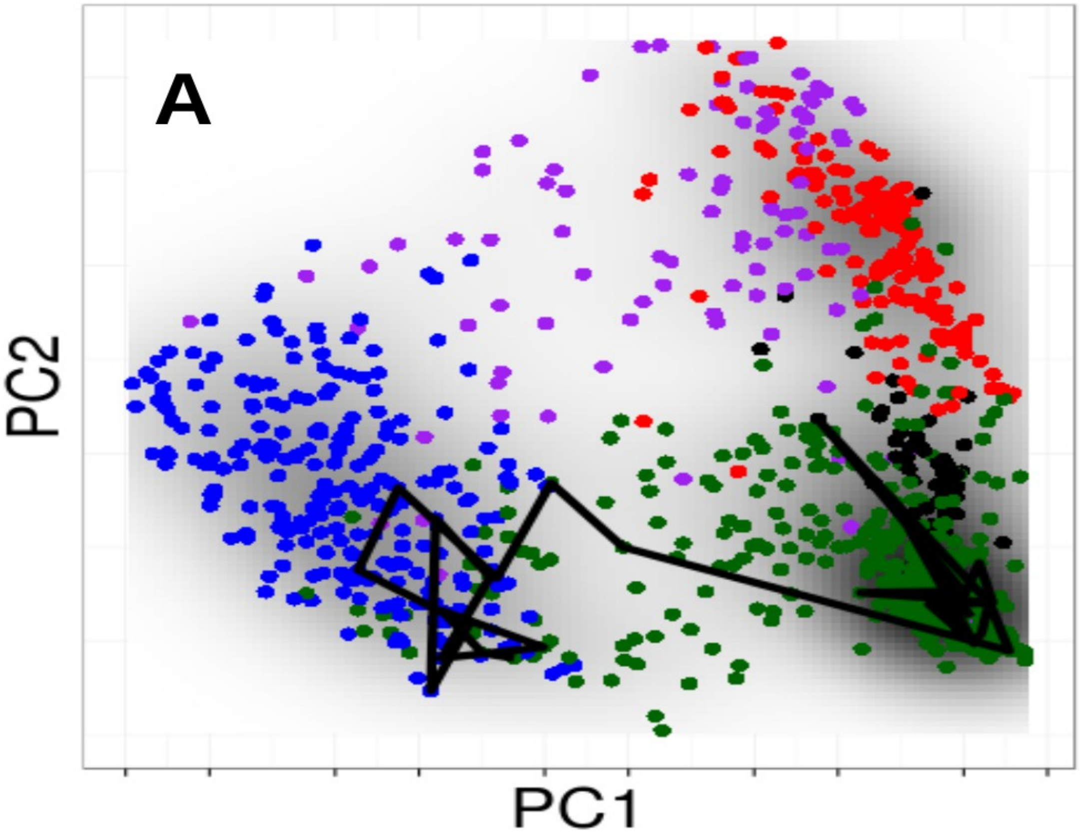
(Flemish Gut Flora Project)

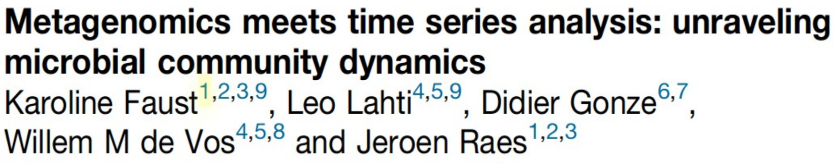
Proposed disease marker genera associated to host covariates and medication - inclusion in study design is essential !



Falony et al. Science 352, 2016.

Vaginal microbiota landscape



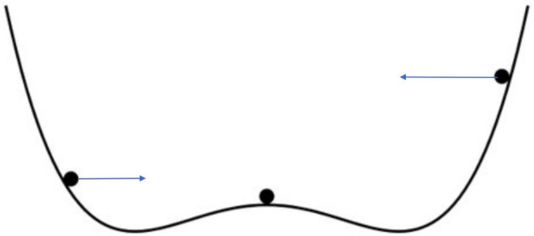

Metagenomics meets time series analysis: unraveling microbial community dynamics

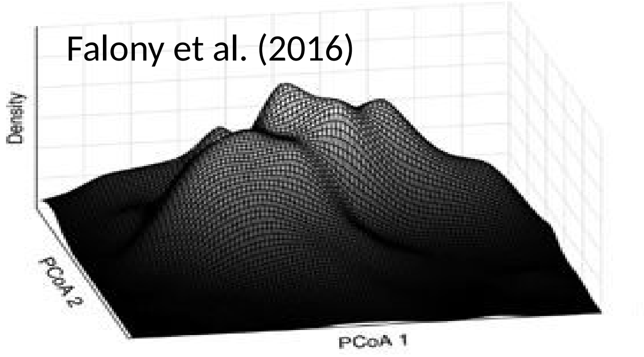
Karoline Faust[1,2,3,9], Leo Lahti[4,5,9], Didier Gonze[6,7], Willem M de Vos[4,5,8] and Jeroen Raes[1,2,3]

Data: Gajer *et al. 2012*

Towards a dynamic landscape model of the microbiome

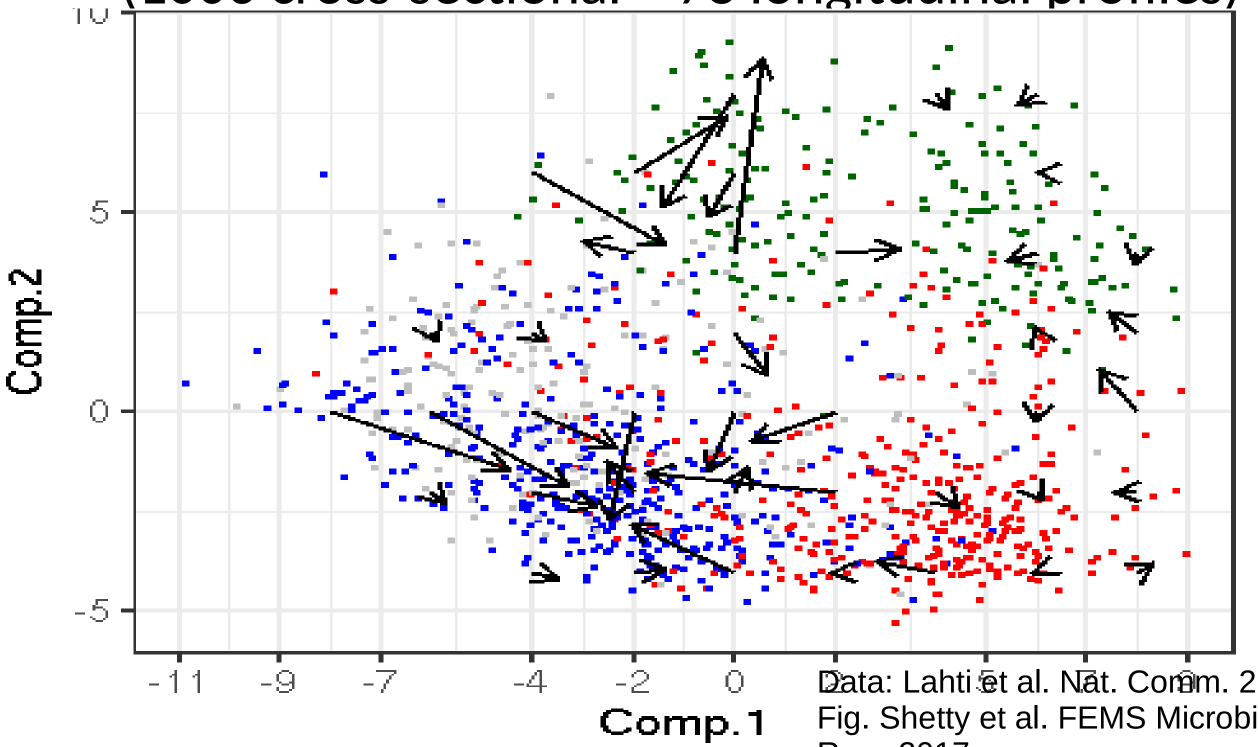
Prior info from background cohorts & pooling evidence across individuals





Falony et al. (2016)

Gradients of change on HITChip PCA landscape (1006 cross-sectional + 78 longitudinal profiles)



Data: Lahti et al. Nat. Comm. 2014
Fig. Shetty et al. FEMS Microbiol Rev.. 2017

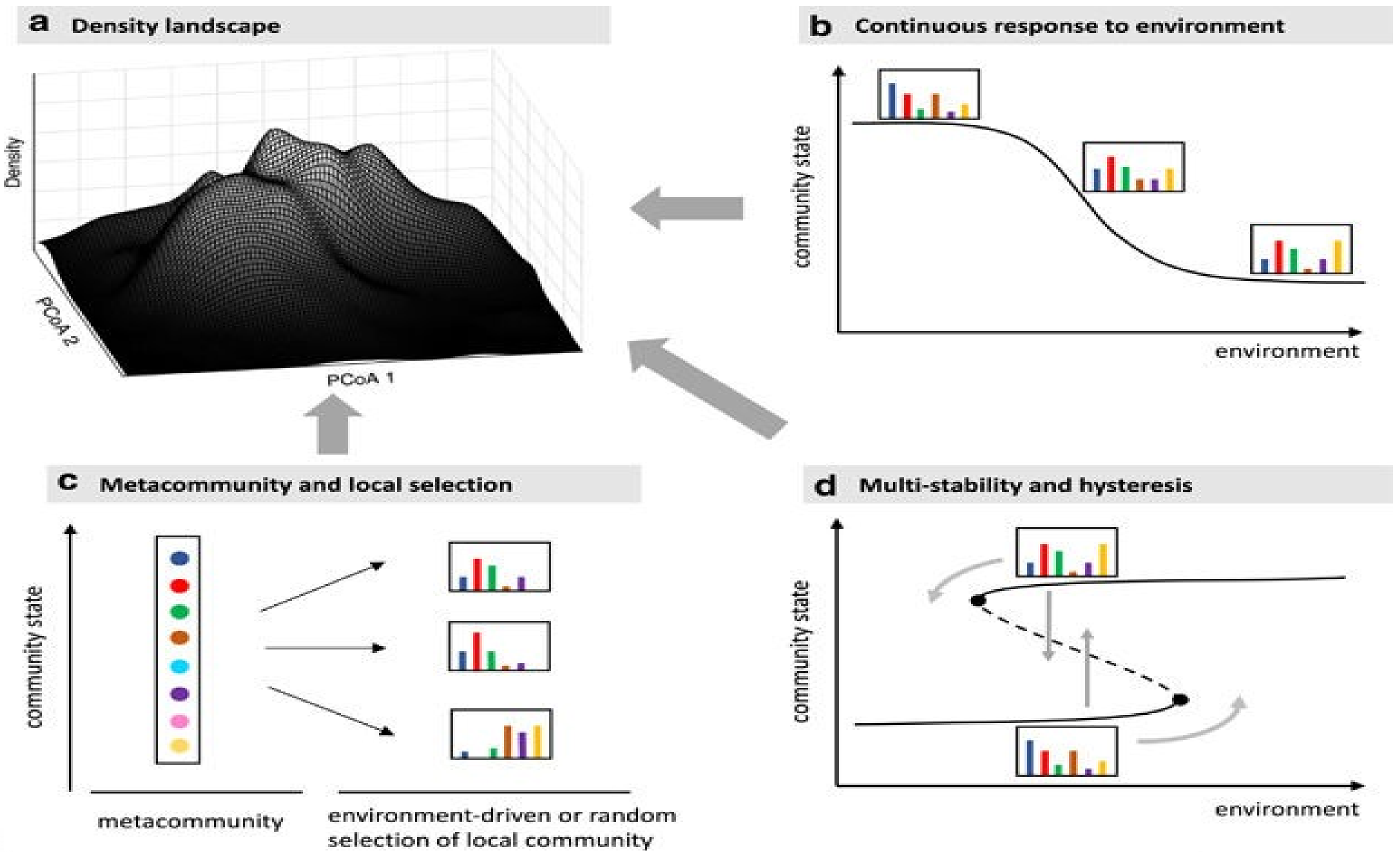
Key sources of microbial ecosystem variation

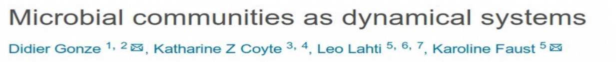
External perturbations (push & pulse)

Internal dynamics and multi-stability

Immigration

Stochasticity

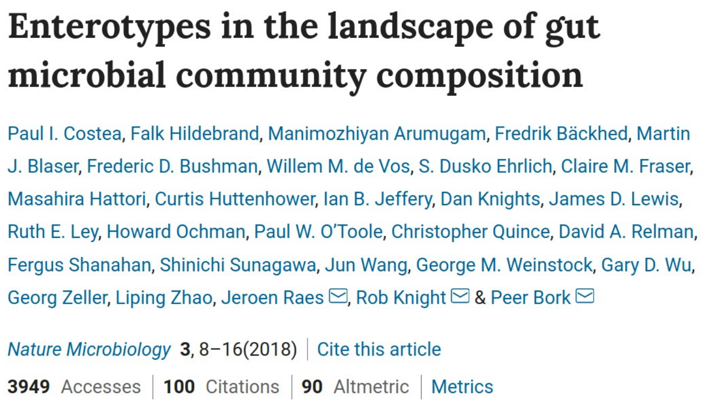



Microbial communities as dynamical systems

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

Community types could stratify individuals:

densely populated areas of the community landscape

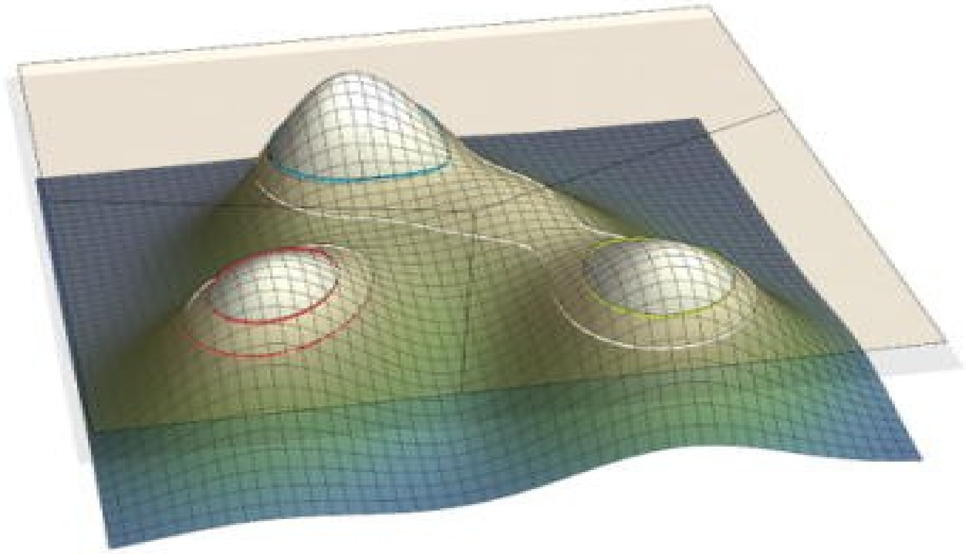


Enterotypes in the landscape of gut microbial community composition

Paul I. Costea, Falk Hildebrand, Manimozhiyan Arumugam, Fredrik Bäckhed, Martin J. Blaser, Frederic D. Bushman, Willem M. de Vos, S. Dusko Ehrlich, Claire M. Fraser, Masahira Hattori, Curtis Huttenhower, Ian B. Jeffery, Dan Knights, James D. Lewis, Ruth E. Ley, Howard Ochman, Paul W. O'Toole, Christopher Quince, David A. Relman, Fergus Shanahan, Shinichi Sunagawa, Jun Wang, George M. Weinstock, Gary D. Wu, Georg Zeller, Liping Zhao, Jeroen Raes ✉, Rob Knight ✉ & Peer Bork ✉

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

3949 Accesses | **100** Citations | **90** Altmetric | Metrics



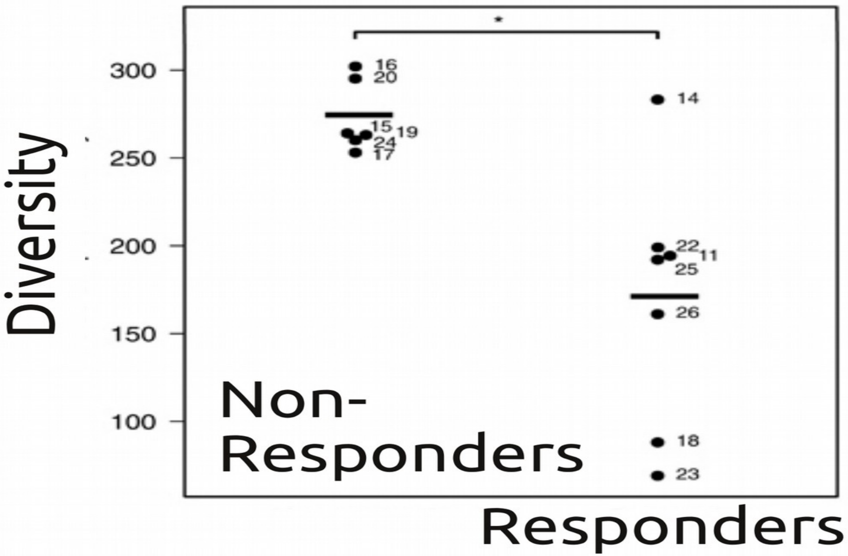
The ISME Journal , (24 April 2014) | doi:10.1038/ismej.2014.63

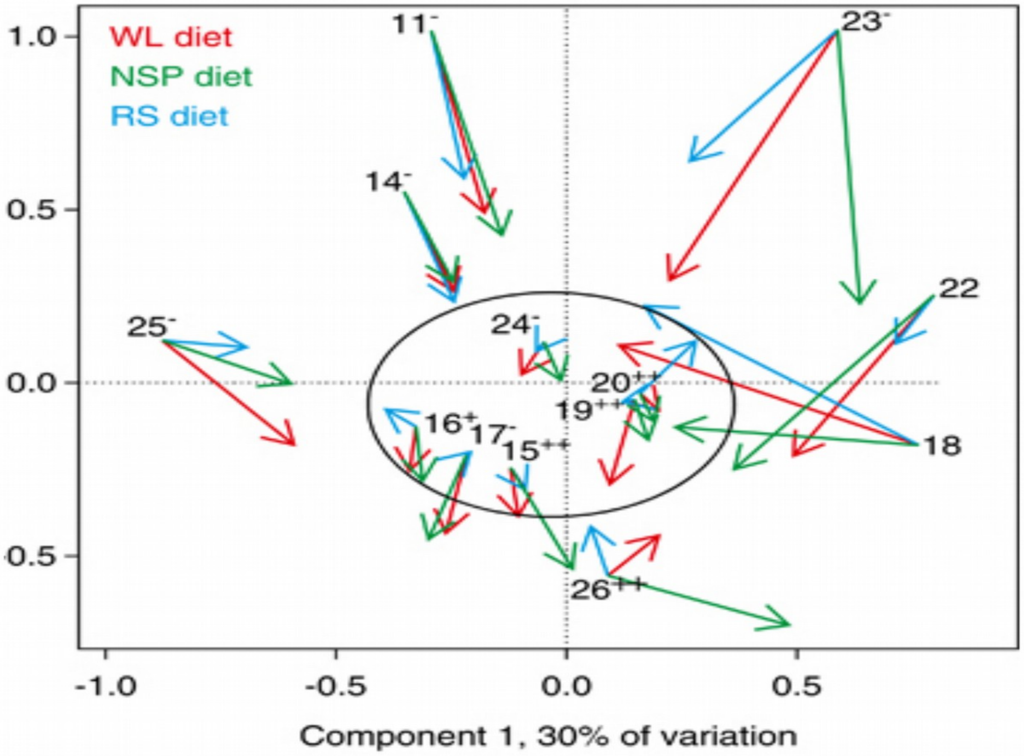
Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men

Anne Salonen, Leo Lahti, Jarkko Salojärvi, Grietje Holtrop, Katri Korpela, Sylvia H Duncan, Priya Date, Freda Farquharson, Alexandra M Johnstone, Gerald E Lobley, Petra Louis, Harry J Flint and Willem M de Vos

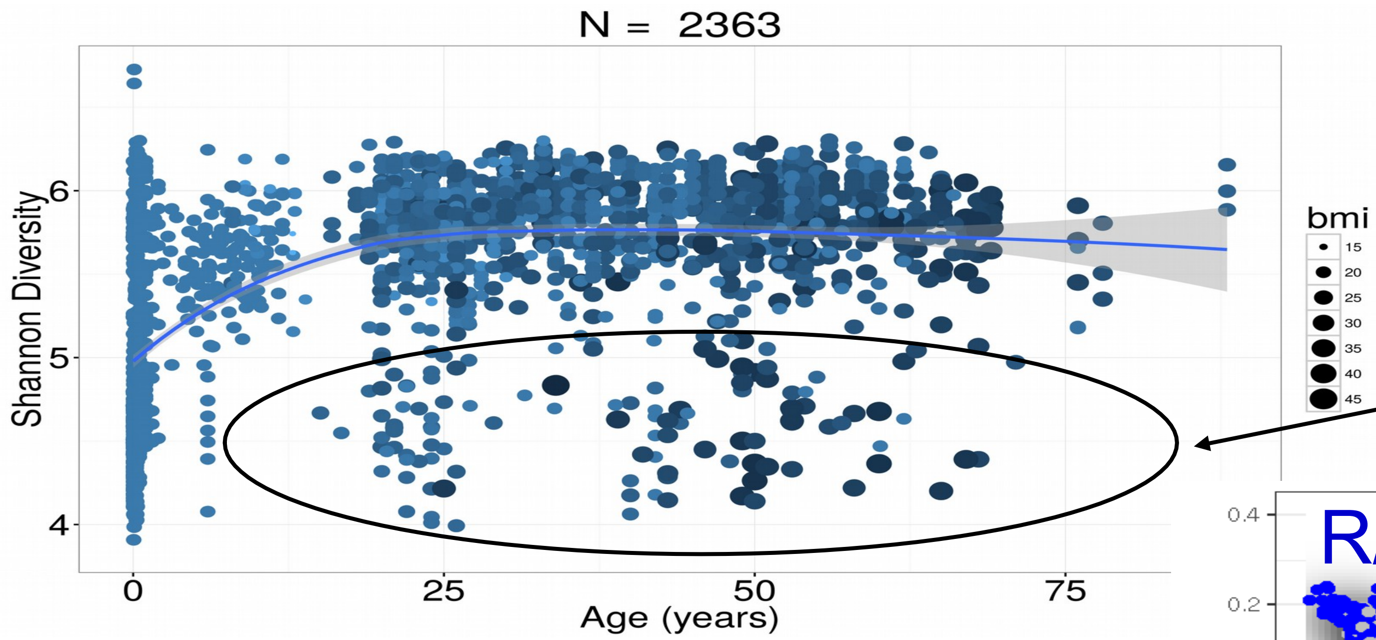
Lahti et al. Nat. Comm. 5:4344, 2014

Korpela et al. PLoS ONE 9(3): e90702, 2014

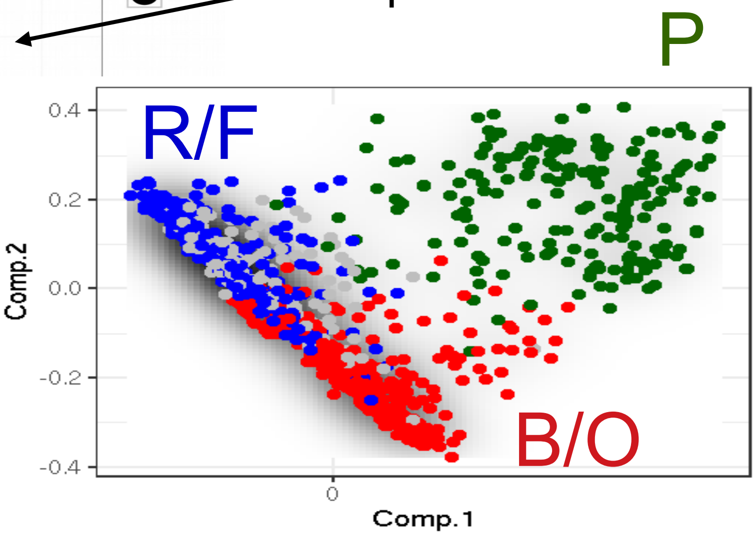




Microbiome diversity and age: healthy & normal obese subjects (HITChip Atlas)



Low diversity samples

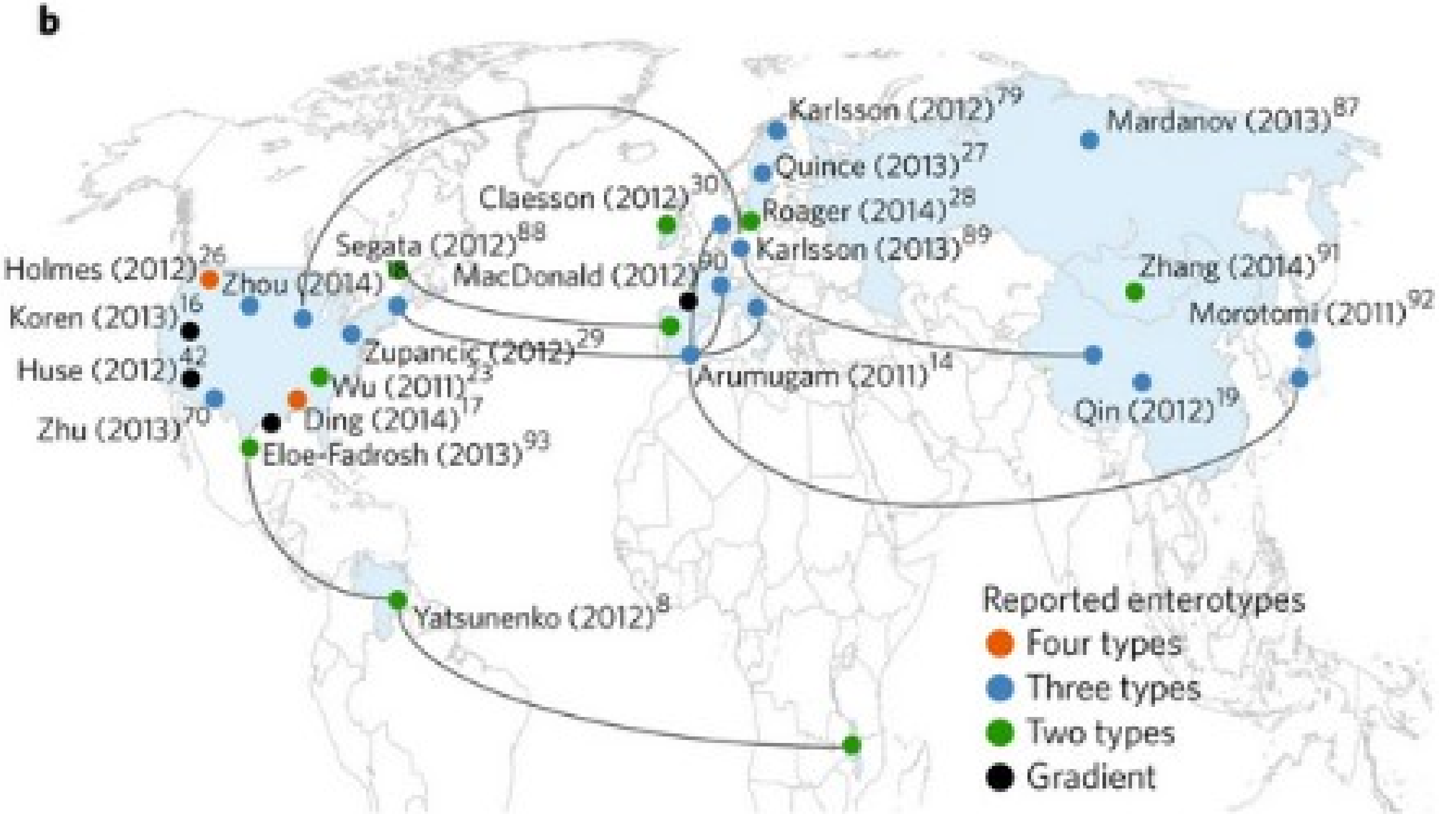


Dominant genera:

Prevotella
Ruminococcus / Faecalibacterium
Bacteroides / Oscillospira

Fig. 2: Stratification of the microbial composition landscape of the human gut microbiome.

From: Enterotypes in the landscape of gut microbial community composition



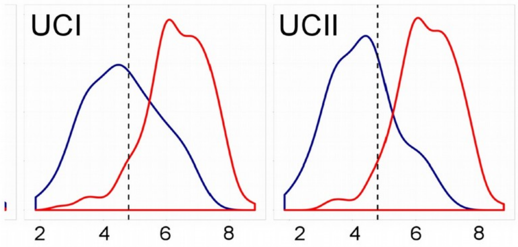
Zooming in: sub-ecosystem variation

Effect on the overall ecosystem

Prevotella states seen at ecosystem level

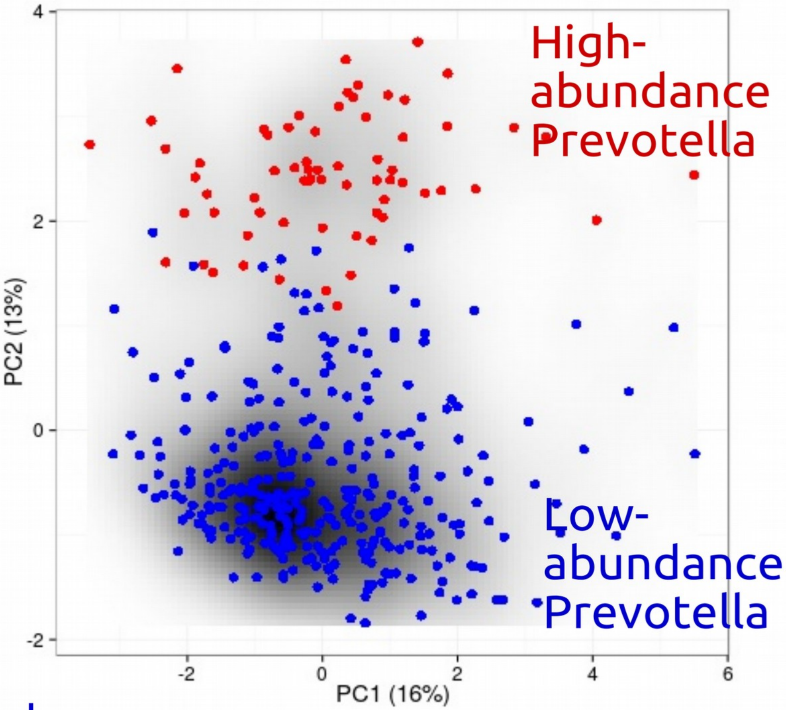
High ~10% relative abundance !

Other bi-stable taxa:
<2% relative abundance, no ecosystem-level switches are visible



Uncultured Clostridiales associated with overall metagenomic richness

PCA +
Correlation (Spearman)

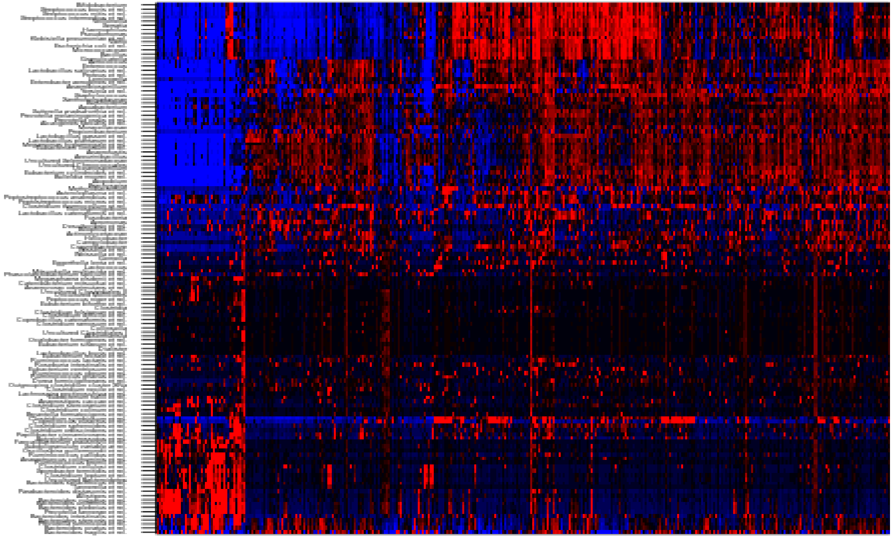


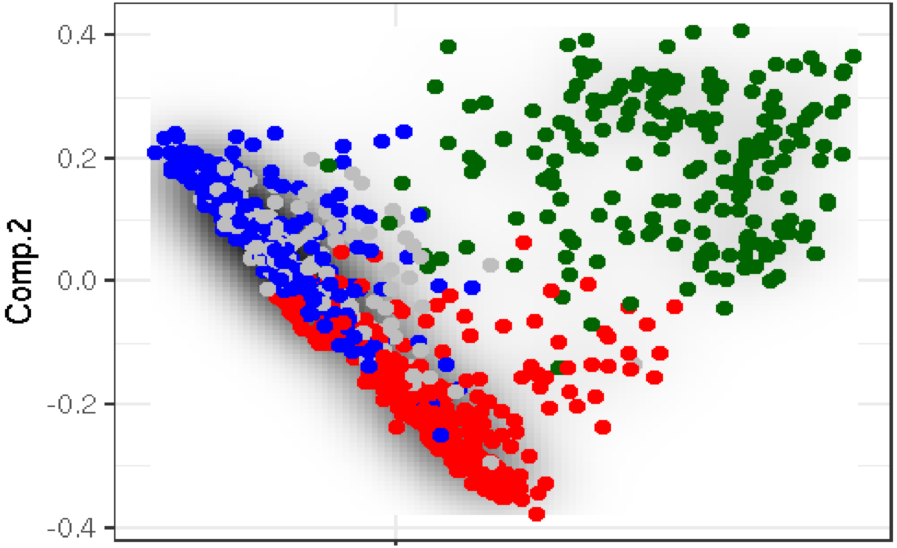
N = 401 western adults

Lahti et al. Nat. Comm. 5:4344,

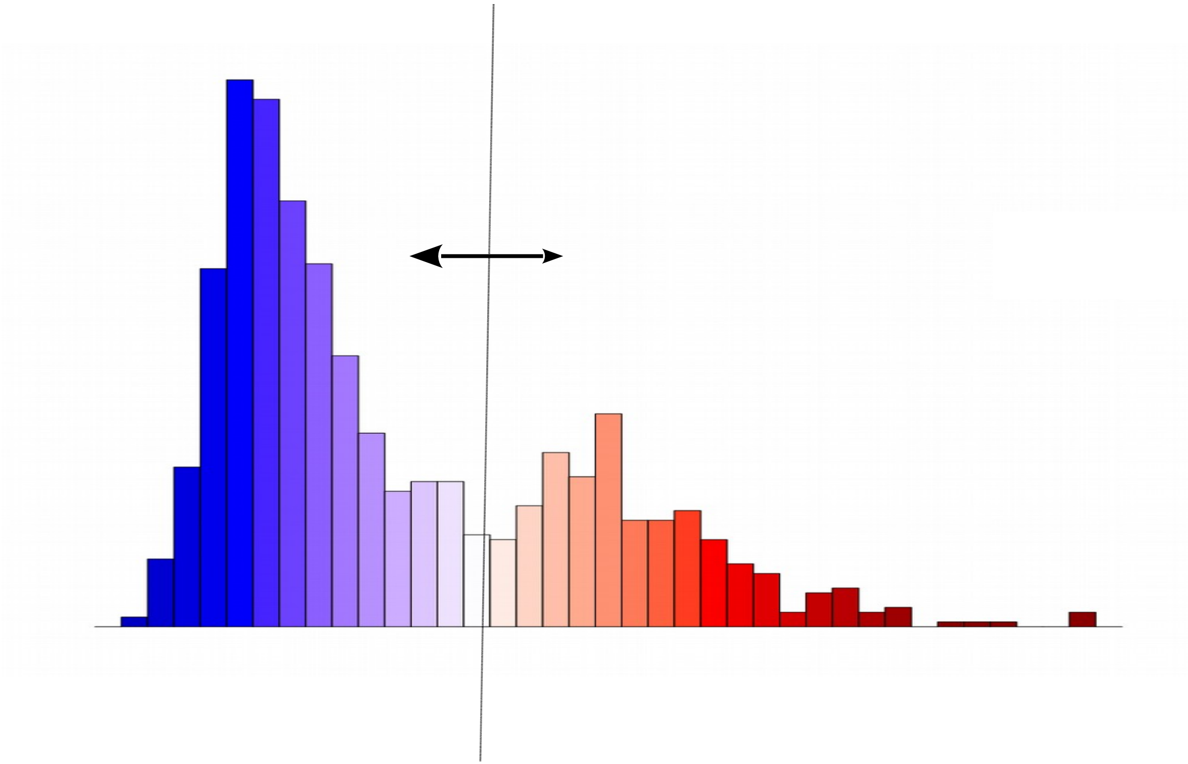
Lahti et al. Nat. Comm. 5:4344, 2014

Sub-ecosystem variation?





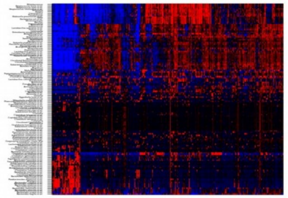
Dialister *spp.*



Received 23 Jan 2014 | Accepted 9 Jun 2014 | Published 8 Jul 2014 DOI: 10.1038/ncomms5344 OPEN

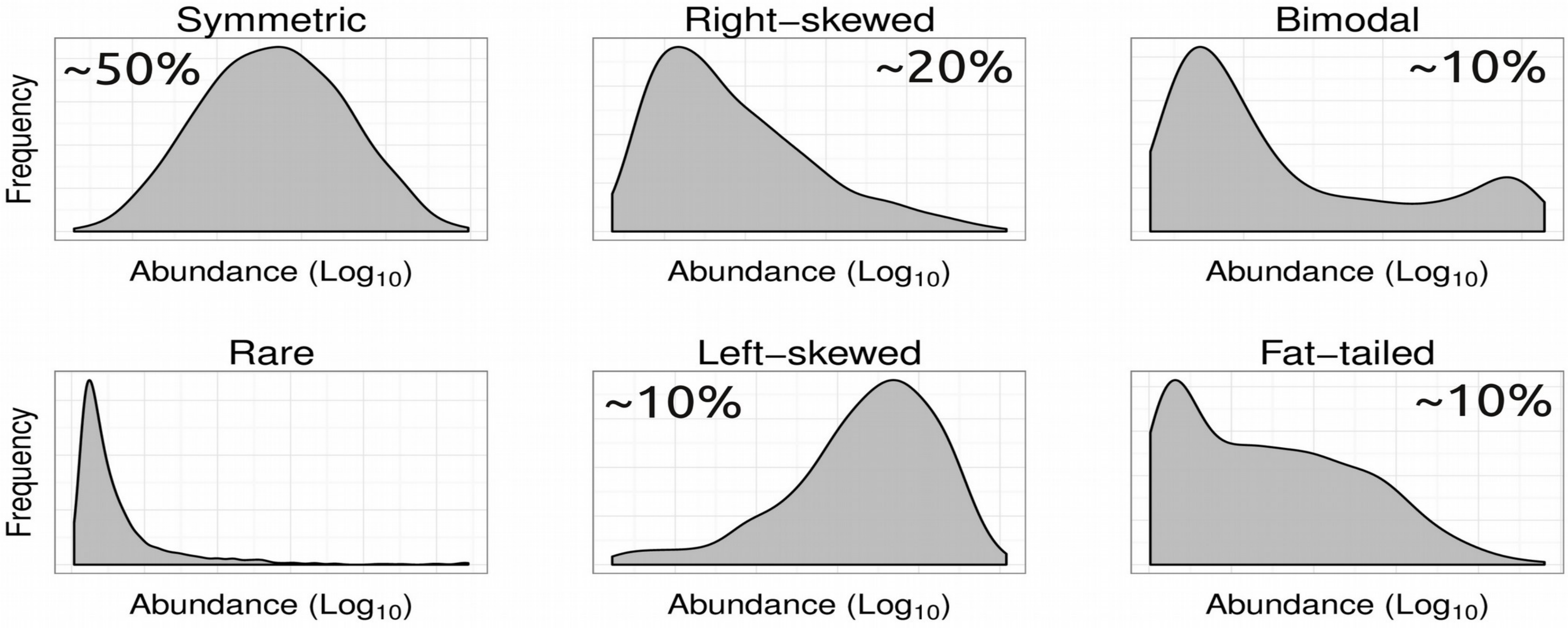
Tipping elements in the human intestinal ecosystem

Leo Lahti[1,2], Jarkko Salojärvi[1,*], Anne Salonen[3,*], Marten Scheffer[4] & Willem M. de Vos[1,2,3]



Bacterial 'abundance types' in 1000 western adults:

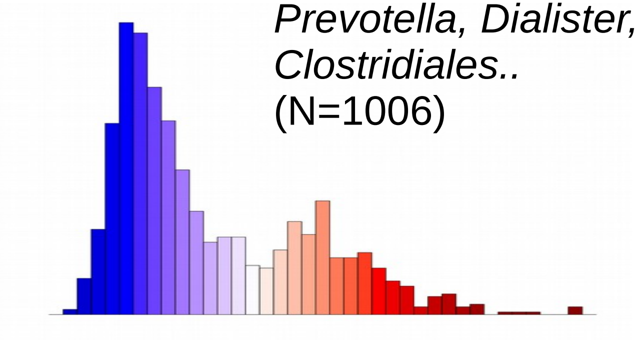
~% indicates proportion among prevalent taxa



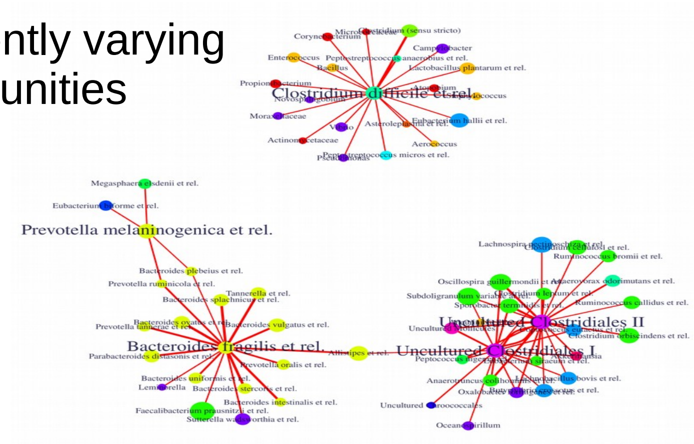
Lahti et al. Nat. Comm. 5:4344, 2014

Tipping elements

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

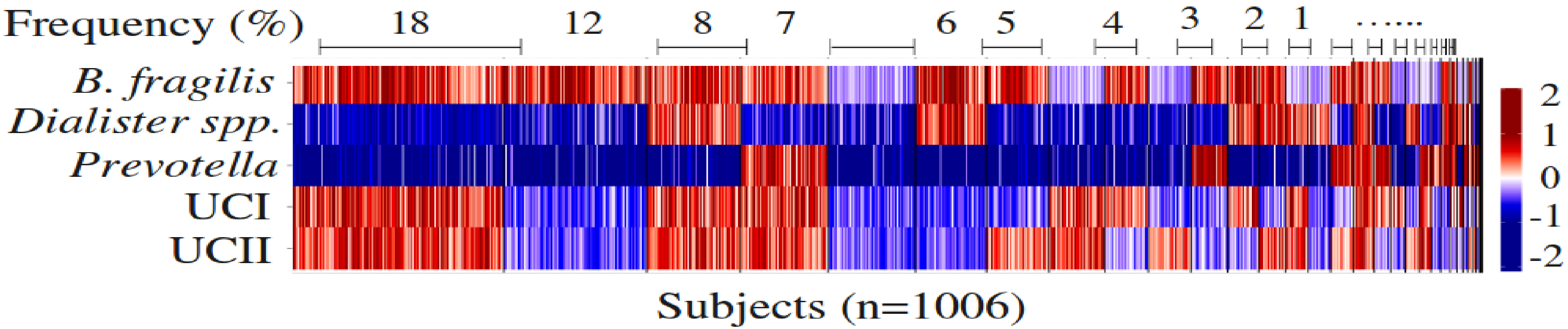


Independently varying
sub-communities

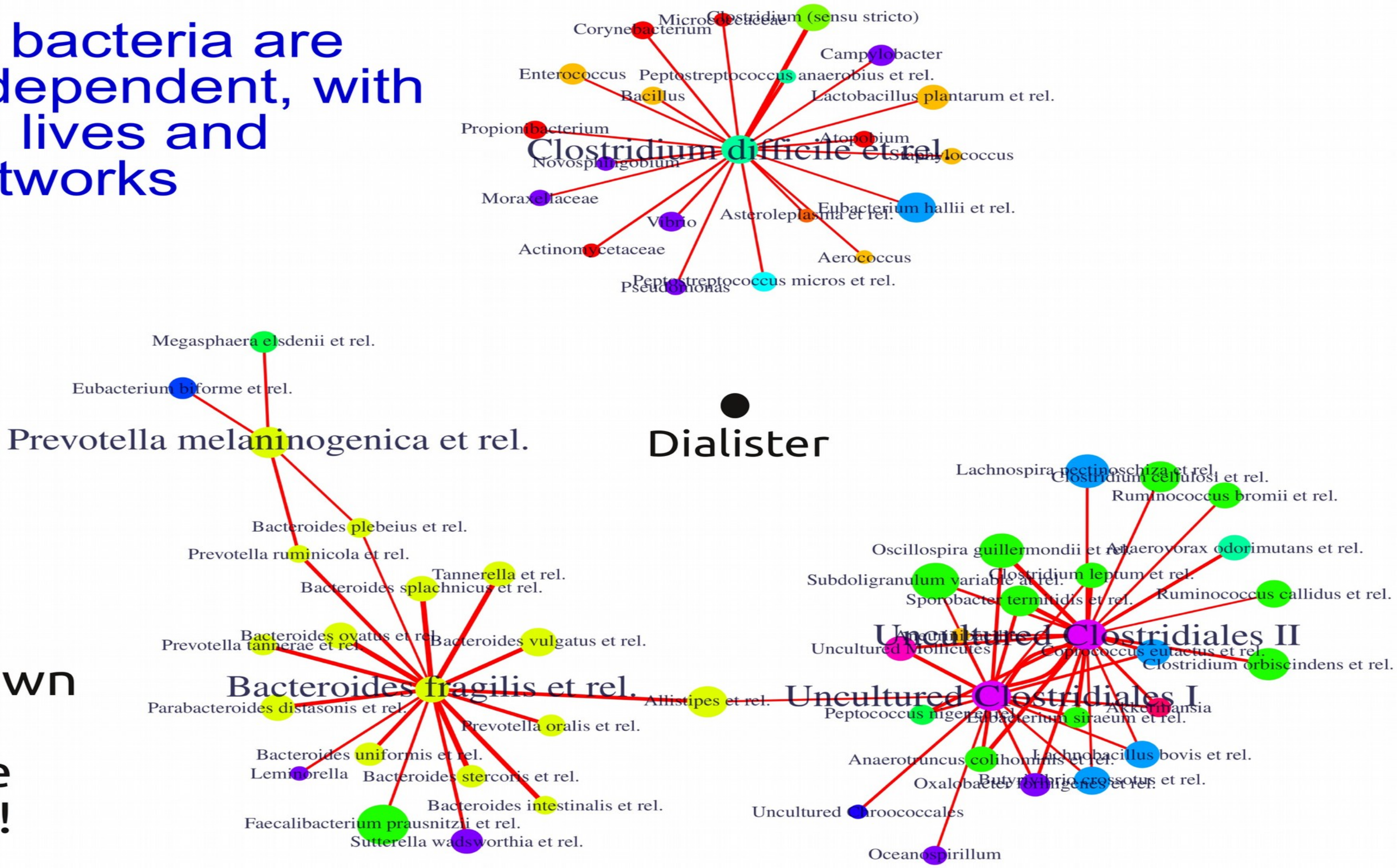


Ecosystem states are rich combinations
of independent **tipping elements** ?

Lahti et al. Nat. Comm. 2014



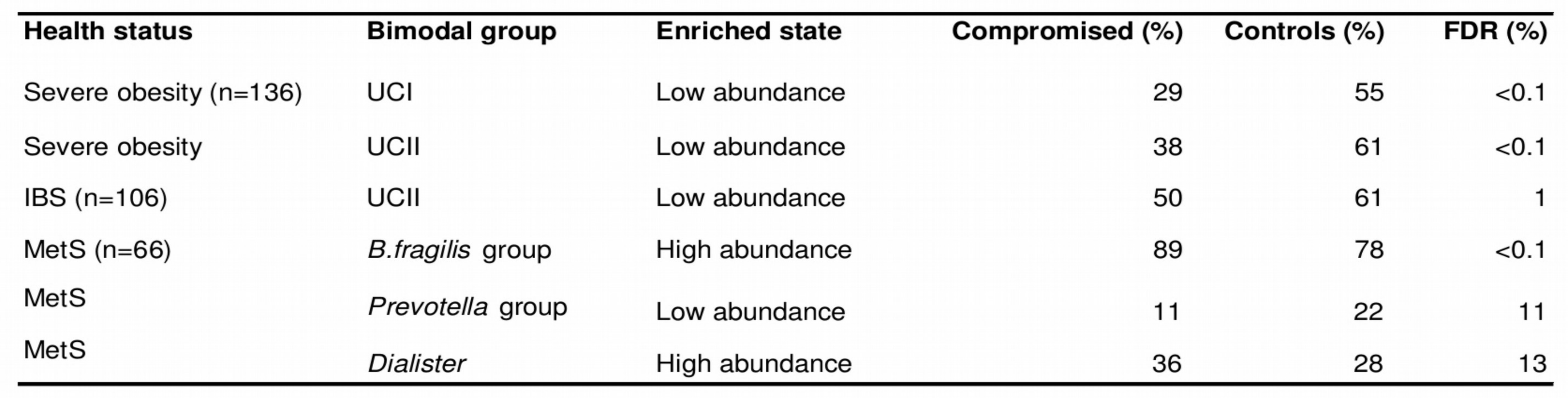
Bi-stable bacteria are rather independent, with their own lives and social networks



|r| > 0.33 shown

Only positive correlations !

Health associations of bi-stable tipping elements



Health status	Bimodal group	Enriched state	Compromised (%)	Controls (%)	FDR (%)
Severe obesity (n=136)	UCI	Low abundance	29	55	<0.1
Severe obesity	UCII	Low abundance	38	61	<0.1
IBS (n=106)	UCII	Low abundance	50	61	1
MetS (n=66)	*B.fragilis* group	High abundance	89	78	<0.1
MetS	*Prevotella* group	Low abundance	11	22	11
MetS	*Dialister*	High abundance	36	28	13

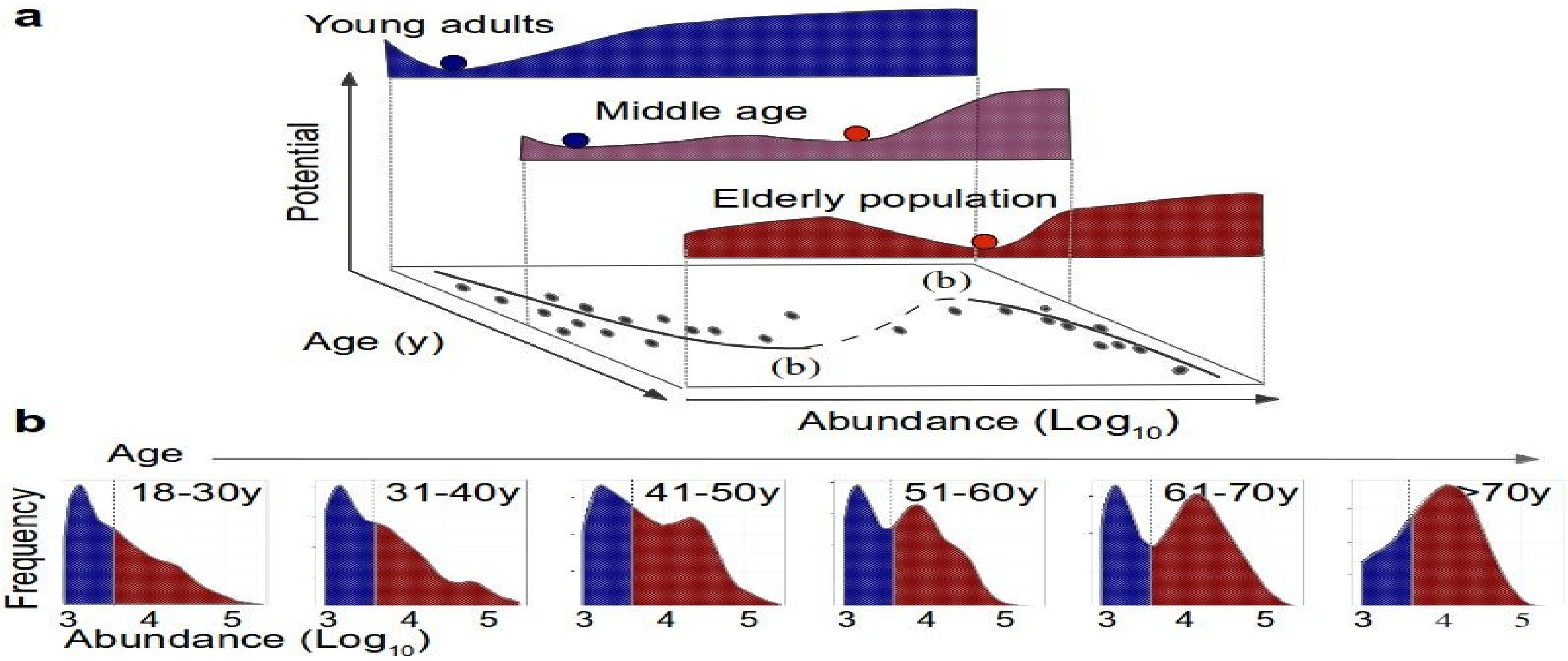
Received 23 Jan 2014 | Accepted 9 Jun 2014 | Published 8 Jul 2014 DOI: 10.1038/ncomms5344 OPEN

Tipping elements in the human intestinal ecosystem

Leo Lahti[1,2], Jarkko Salojärvi[1,*], Anne Salonen[3,*], Marten Scheffer[4] & Willem M. de Vos[1,2,3]

Tipping elements of the human gut microbiota: Uncultured Clostridiales I

Lahti et al. Nat. Comm. 5:4344, 2014



Bias is everywhere

- Absolute abundances cannot be determined
- Relative abundances are biased
- Each step from experiments to analysis adds noise

→ Experimental design & validation!

Dealing with confounding effects

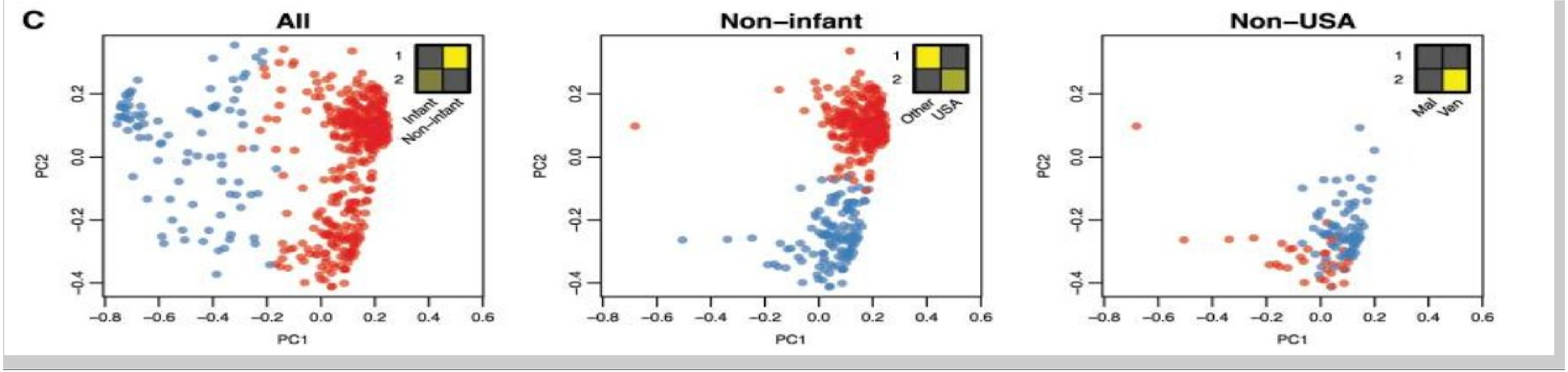
- Any observation may be *confounded*
- Are there technical explanations for observed patterns? Uneven sequencing / Batch effects / Edge effects..
- Excluding alternative explanations is critical

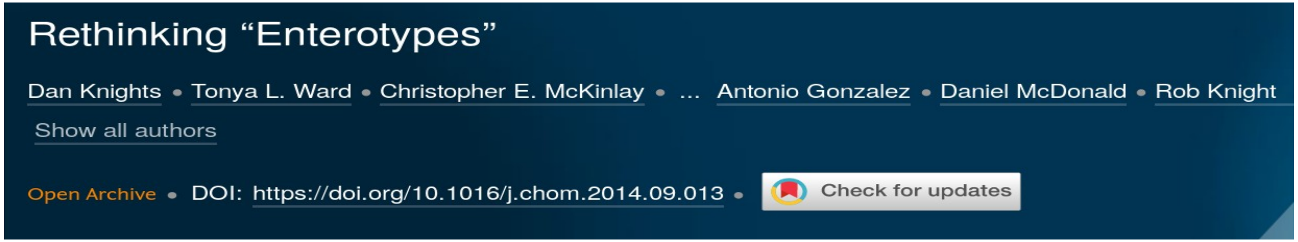
  - Example: negative correlation between two ASVs. Biological or technical explanation?

    ASV1: 0 1 0 0 1 1 0 0 1

    ASV2: 1 0 0 1 0 0 1 0 1

External covariates can induce distinct clusters





Rethinking “Enterotypes”

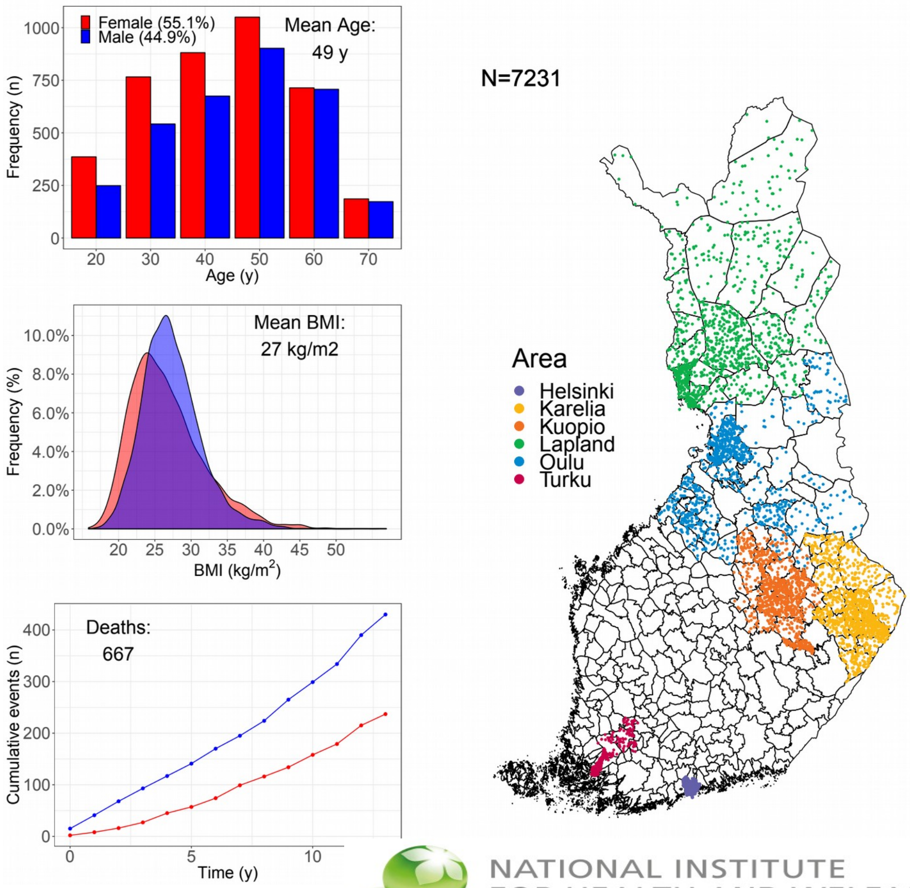
Dan Knights • Tonya L. Ward • Christopher E. McKinlay • ... Antonio Gonzalez • Daniel McDonald • Rob Knight

Show all authors

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

Ways to handle covariates

- Stratify and analyze homogeneous subsets
- Model the variation



Why (nonlinear) PCoA and not (linear) PCA?

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

Transformation: compositional

Dissimilarity: Bray-Curtis

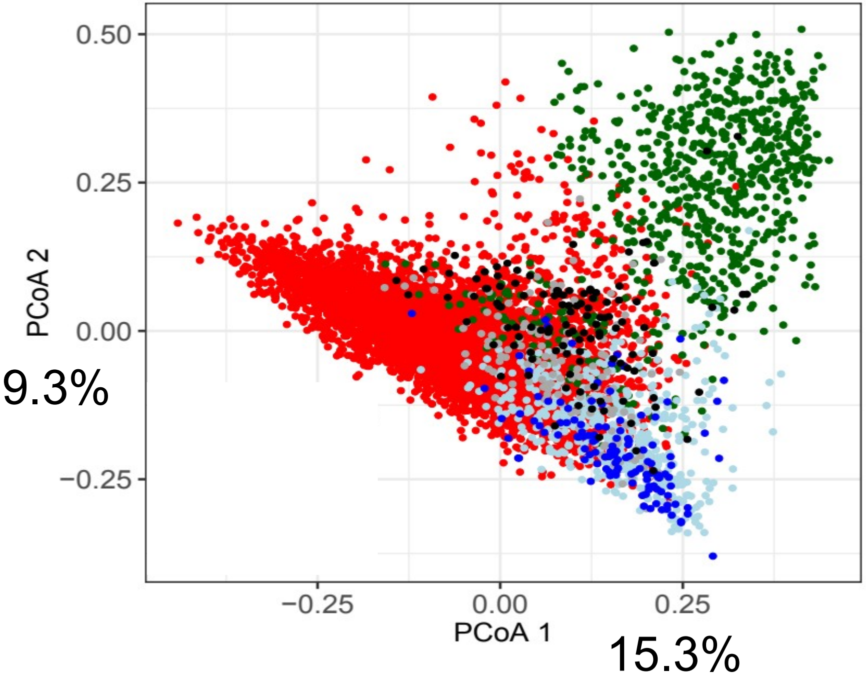
Method: **Preserves distances**

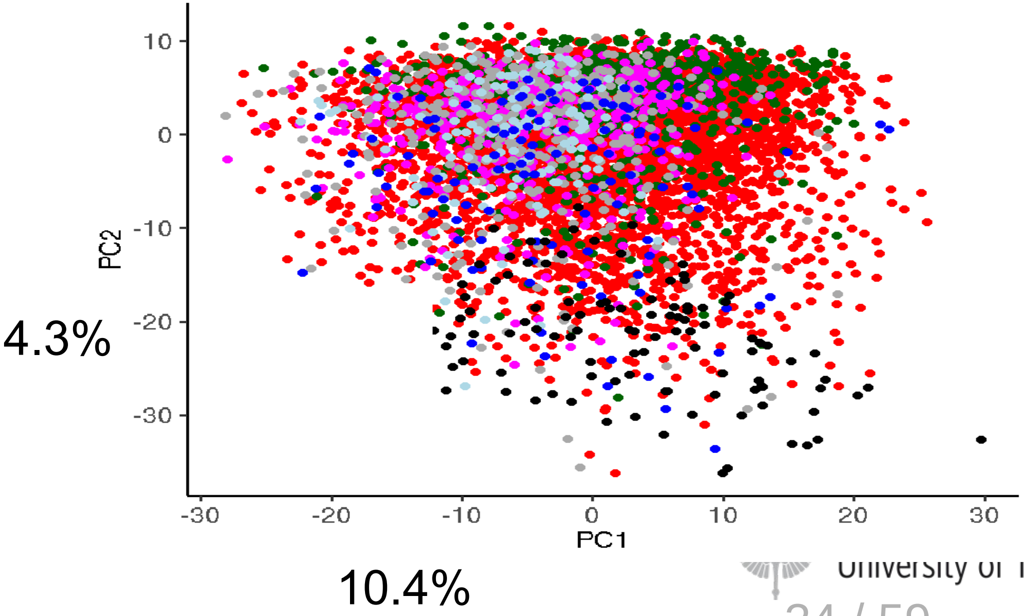
PCA Principal Component Analysis

Transformation: CLR

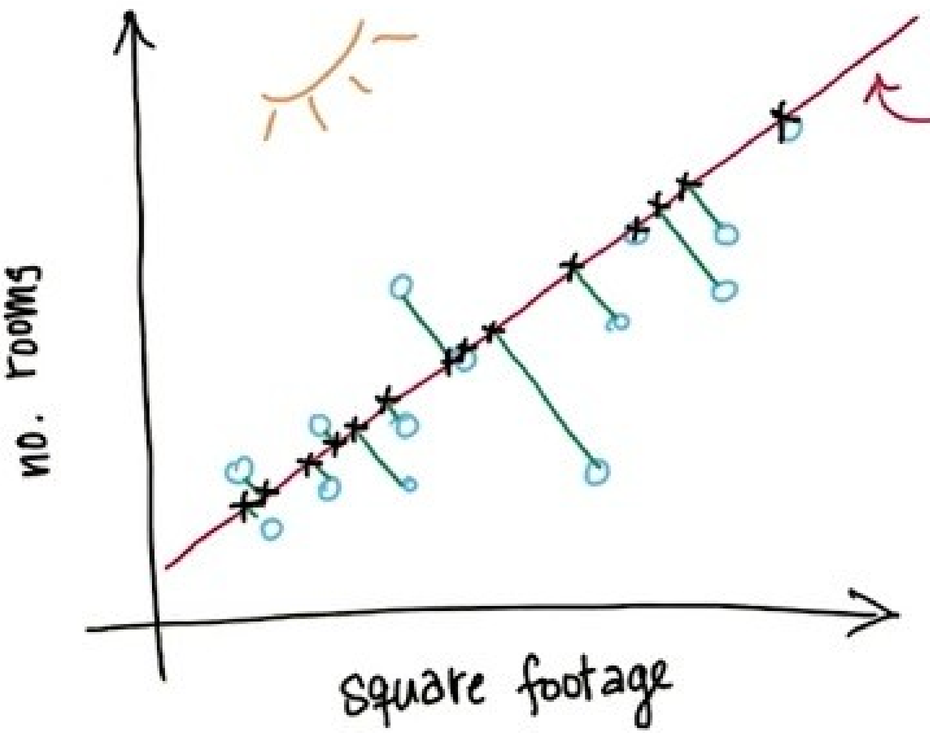
Dissimilarity: Euclidean

Method: **Captures largest variation**





Example: Square Footage + No. Rooms → Size



principal component
(NOT a regression!)

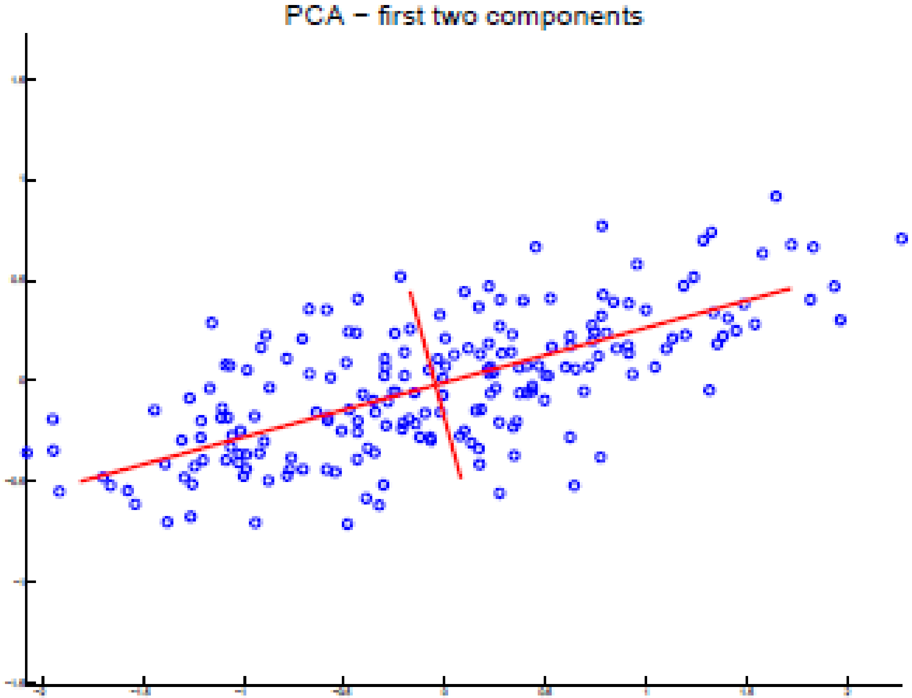
What is the distribution of the new feature?

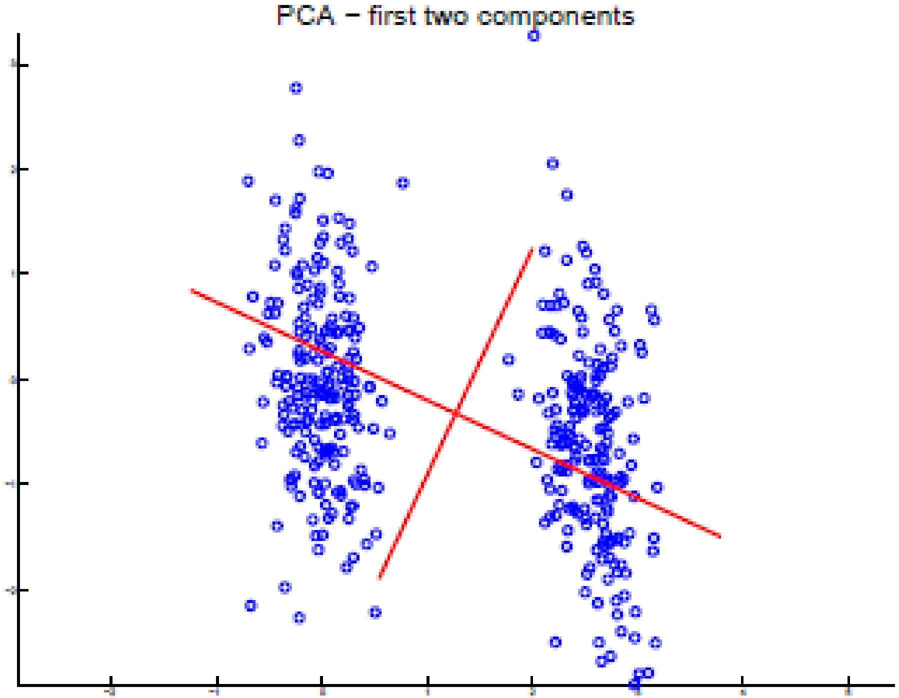
○ 

○ 

○ 

○ 


PCA − first two components


PCA − first two components

https://ai-pool.com/a/s/dimensionality-reduction--pca-intro

Fundamental considerations in beta diversity analysis

Feature selection

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

Transformation

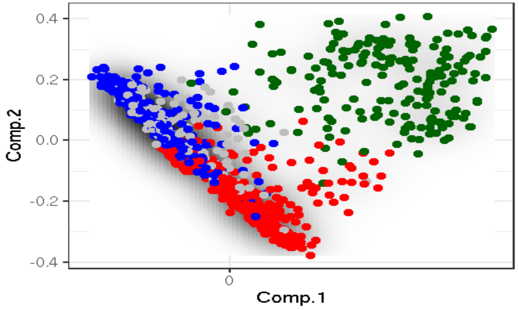
(absolute, compositional, CLR, Hellinger..?)

Dissimilarity measure

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

Analysis method

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



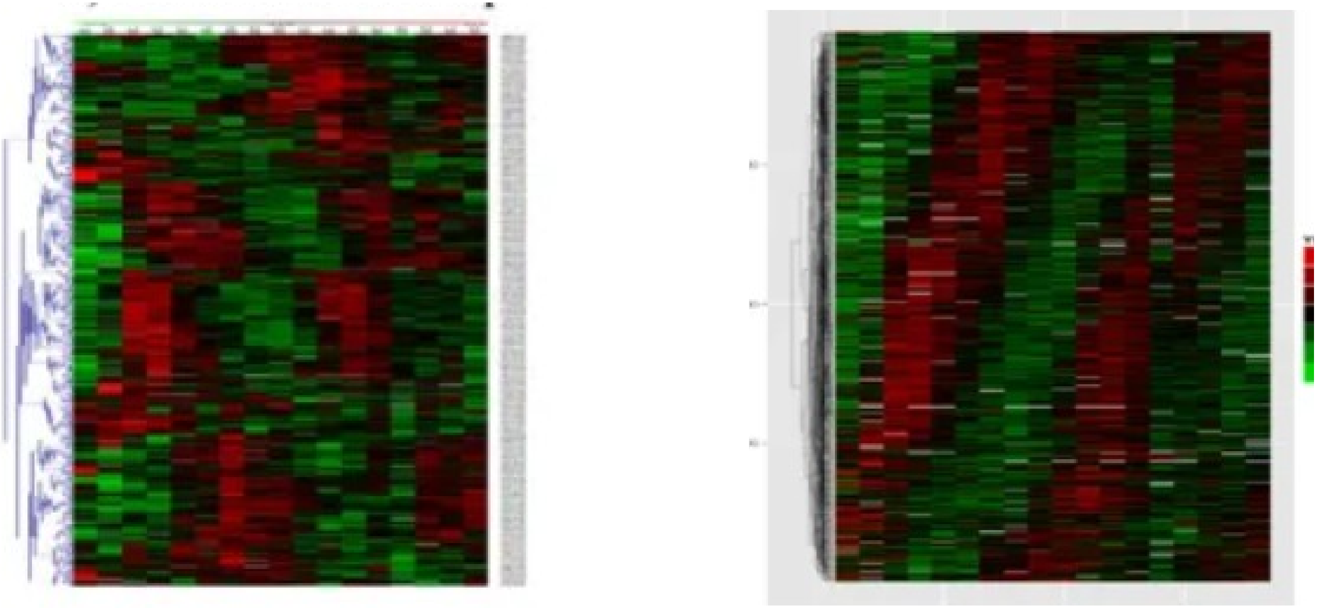
Software | Open Access | Published: 22 January 2010

NeatMap - non-clustering heat map alternatives in R

Satwik Rajaram & Yoshi Oono

BMC Bioinformatics **11**, Article number: 45 (2010) | Cite this article

26k Accesses | **47** Citations | **2** Altmetric | Metrics

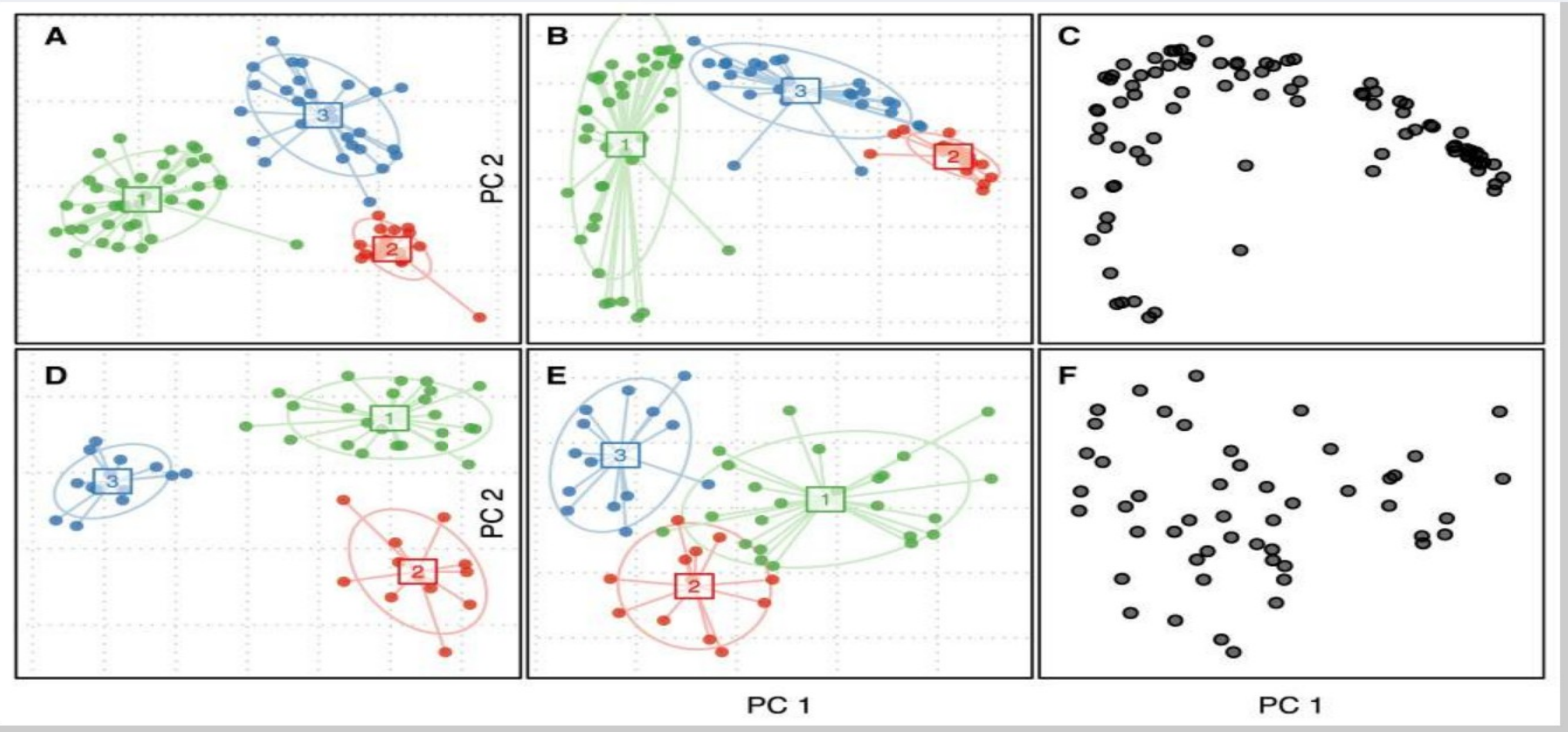


Distinct clusters or extremes on a continuum?

Common Visualizations Can Support Different Conclusions

Soil samples with varying pH

Simulated data with no cluster structure



Supervised

Unsupervised with colors

Unsupervised without colors


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

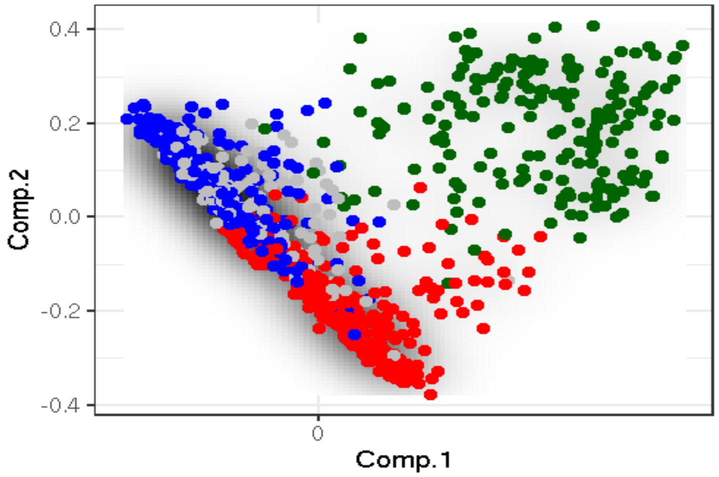
Dirichlet multinomial mixture (DMM)

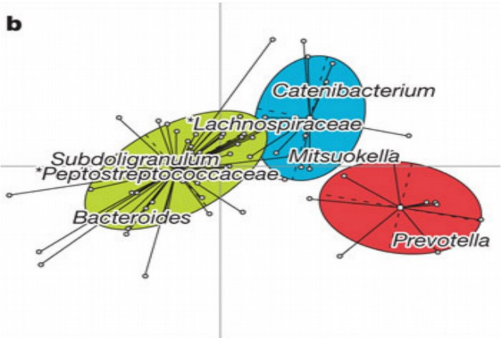
Probabilistic cluster analysis

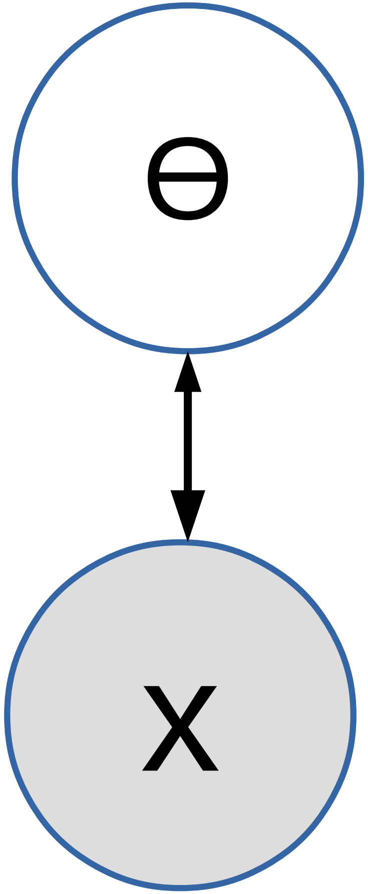
How many
Clusters?

Which samples
in which clusters?

How the clusters
are shaped?

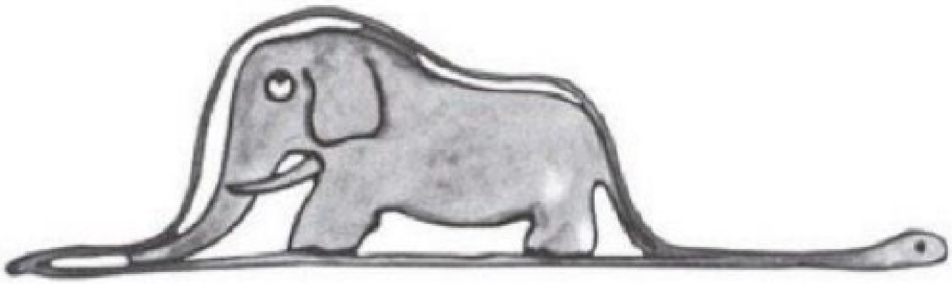




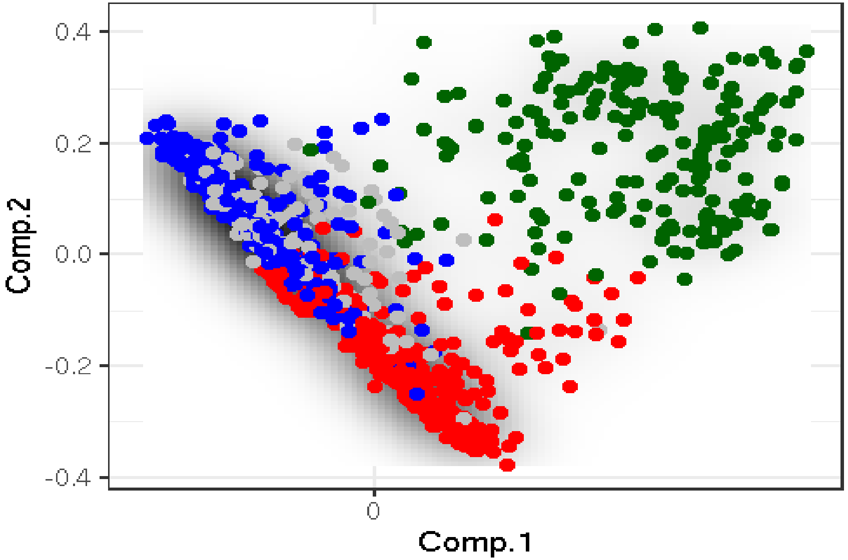


Model

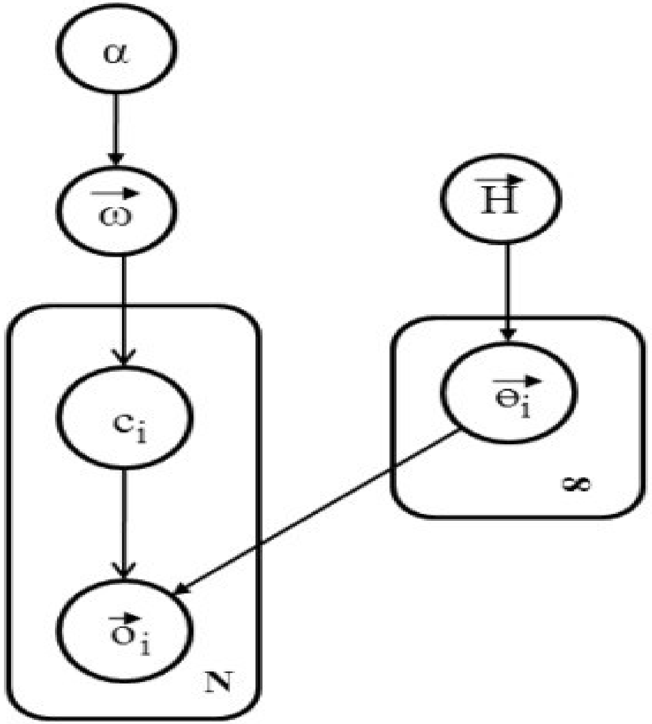
Observations
(Data)



hidden variables.



From learning the model <u>parameters</u> to learning the model <u>structure</u>



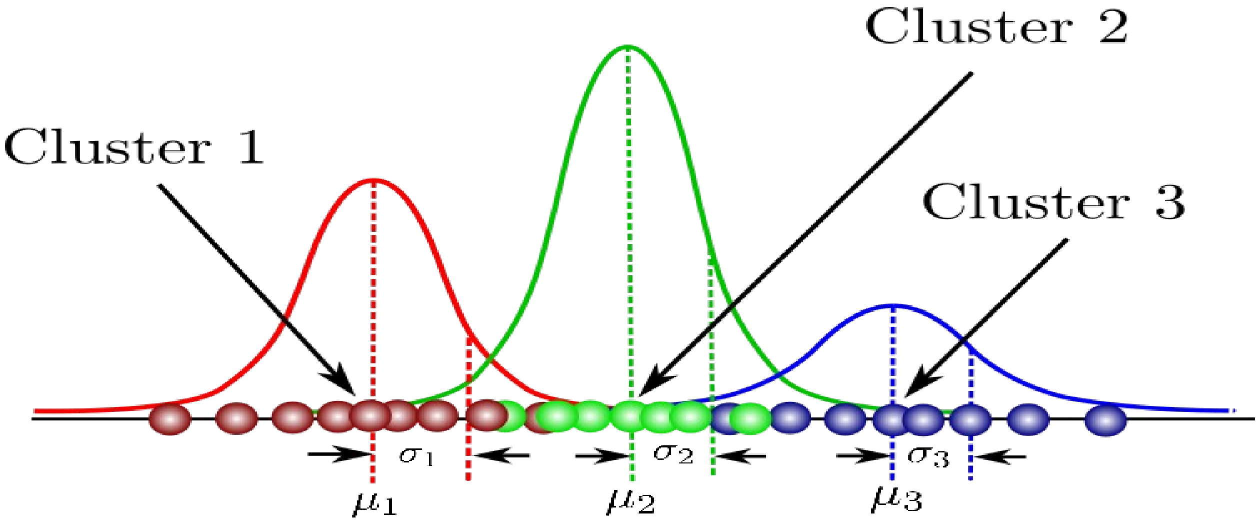
(b) Infinite Gaussian Mixture Model (IGMM)

Balancing model complexity & fit: Bayesian Information Criterion

$$\mathrm{BIC} = \ln(n)k - 2\ln(\widehat{L}).$$

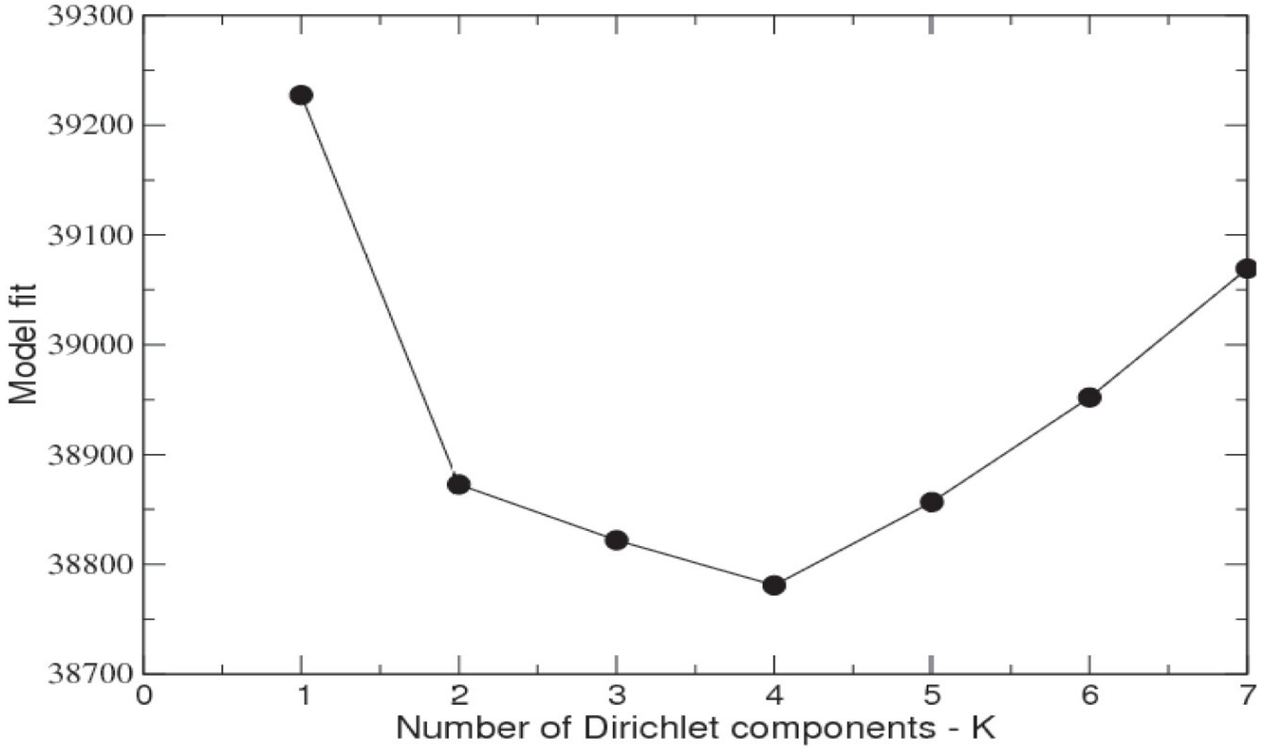
Model
complexity

Model
fit



Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics

Ian Holmes, Keith Harris, Christopher Quince ✉



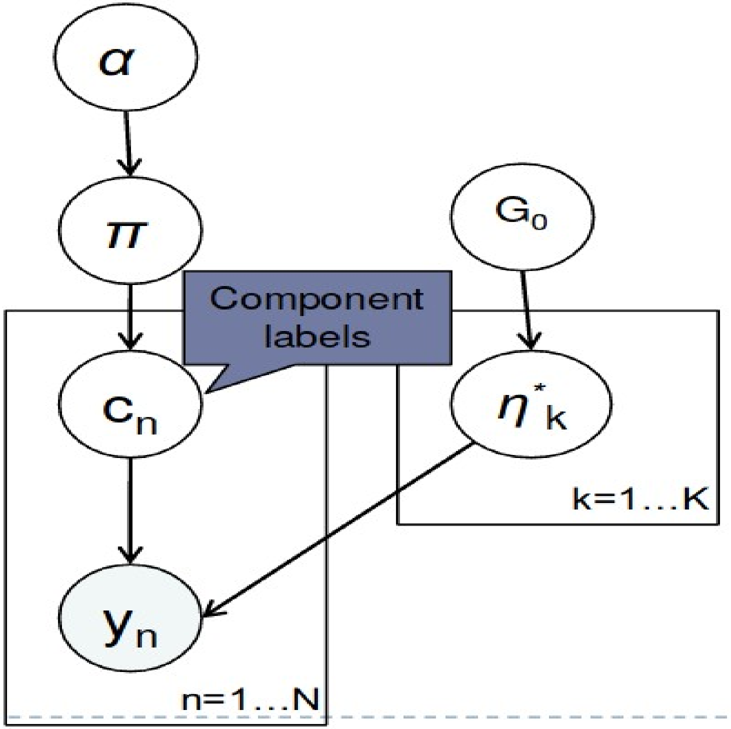
The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data

John D. O’Brien, Nicholas R. Record, Peter Countway

doi: https://doi.org/10.1101/045468

Finite Mixture Models

- A finite mixture model assumes that the data come from a mixture of a finite number of distributions.



$\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$

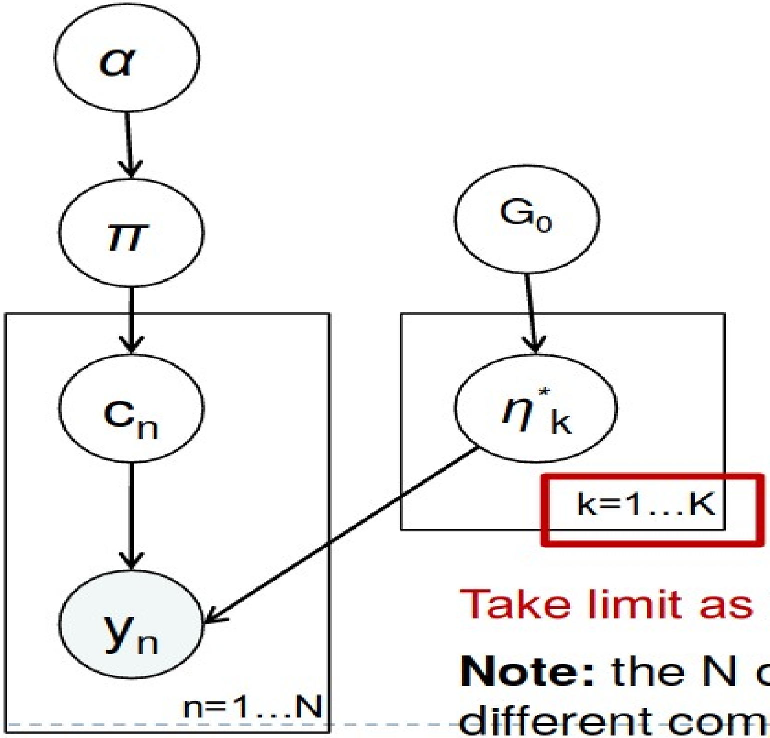
$c_n \sim \text{Multinomial}(\pi)$

$\eta_k \sim G_0$

$y_n \mid c_n, \eta_1, \ldots \eta_K \sim F(\,\cdot \mid \eta_{c_n})$

Infinite Mixture Models

- An infinite mixture model assumes that the data come from a mixture of an *infinite* number of distributions



$\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$

$c_n \sim \text{Multinomial}(\pi)$

$\eta_k \sim G_0$

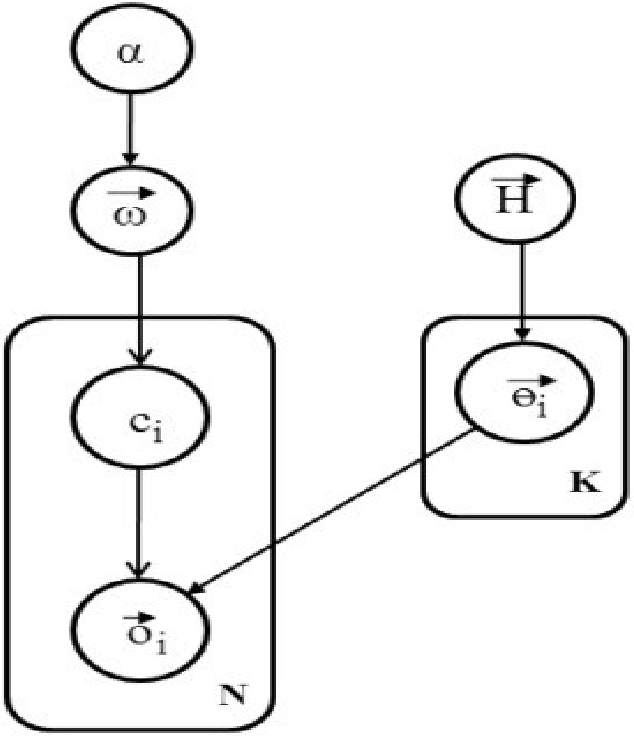
$y_n \mid c_n, \eta_1, \ldots \eta_K \sim F(\,\cdot \mid \eta_{c_n})$

Take limit as K goes to ∞

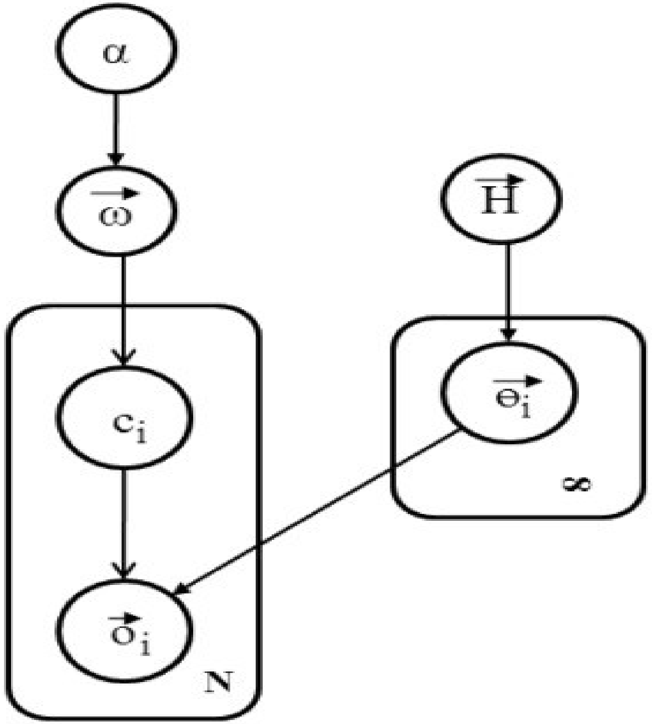
Note: the N data points still come from at most N different components

[Rasmussen 2000]

Non-parametric models bring flexibility in the model structure



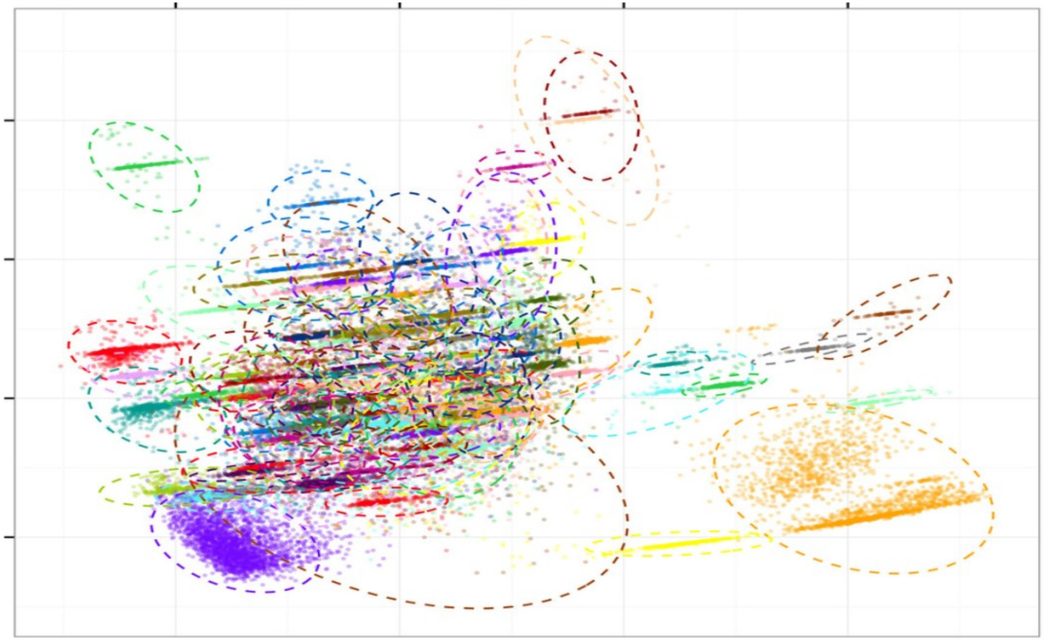
(a) Finite Gaussian Mixture Model (FGMM)



(b) Infinite Gaussian Mixture Model (IGMM)

Multivariate Gaussian mixture model: how many clusters?

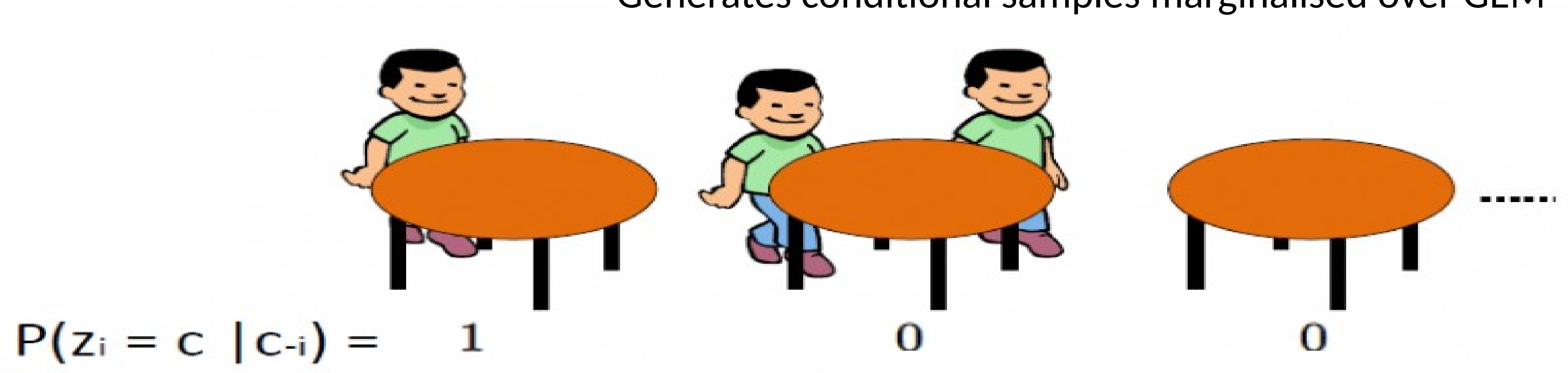
→ Infinite Gaussian mixtures!



DP as Chinese Restaurant Process (CRP)

Future samples derive from new species (table) with prob. proportional to θ and join existing species (table) with prob. proportional to their size

Generates conditional samples marginalised over GEM



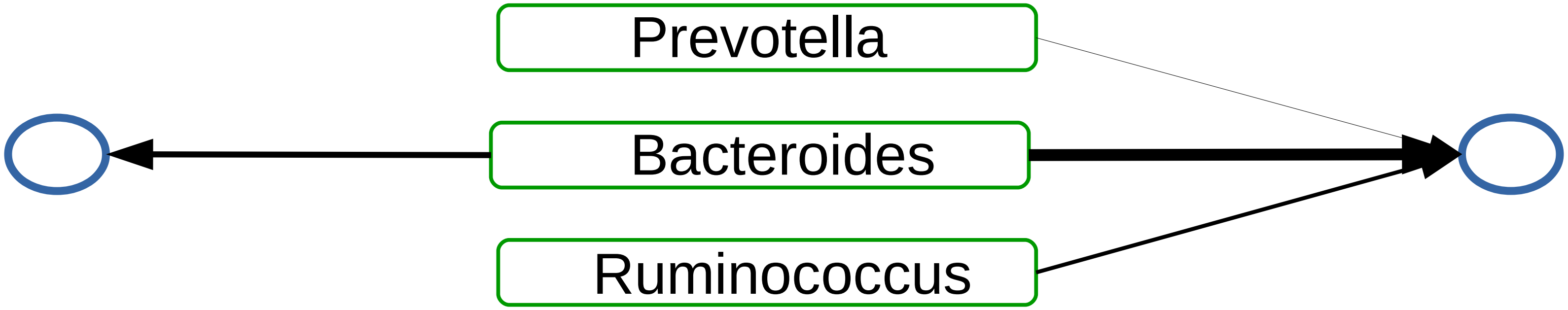
$P(z_i = c \mid c_{-i}) =$

1	0	0
$\frac{1}{1+\theta}$	$\frac{\theta}{1+\theta}$	0
$\frac{1}{2+\theta}$	$\frac{1}{2+\theta}$	$\frac{\theta}{2+\theta}$
$\frac{1}{3+\theta}$	$\frac{2}{3+\theta}$	$\frac{\theta}{3+\theta}$

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

Clustering vs. Factorization

binary / continuous weights



Prevotella

Bacteroides

Ruminococcus

Individual has
one community type

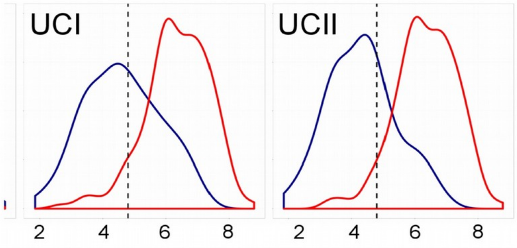
Individual is associated with
multiple community types

Effect on the overall ecosystem

Prevotella states seen at ecosystem level

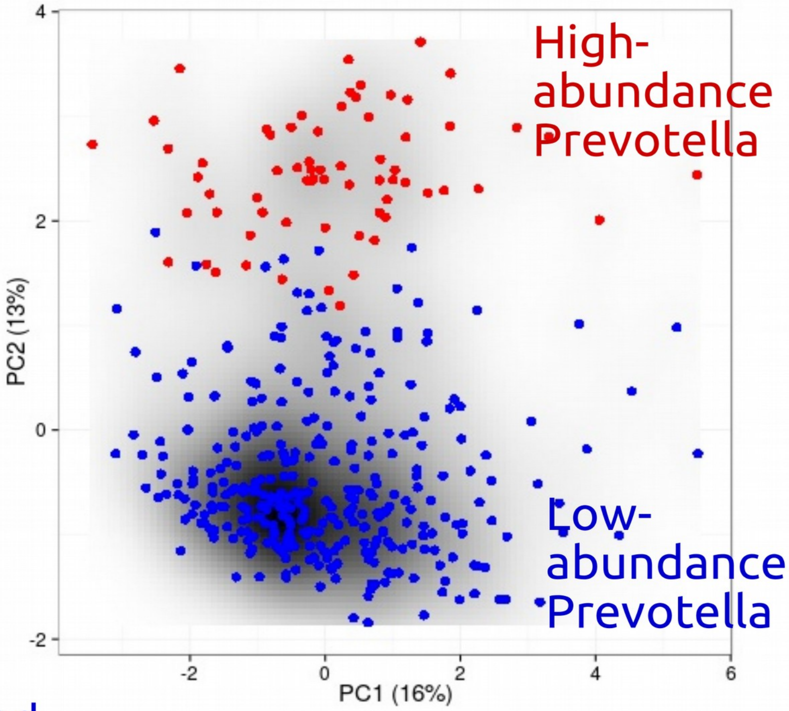
High ~10% relative abundance !

Other bi-stable taxa:
<2% relative abundance,
no ecosystem-level switches are visible



Uncultured Clostridiales associated with overall metagenomic richness

PCA +
Correlation (Spearman)

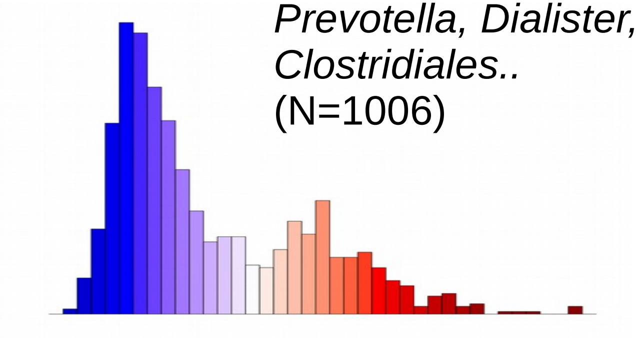


N = 401 western adults

Lahti et al. Nat. Comm. 5:4344,

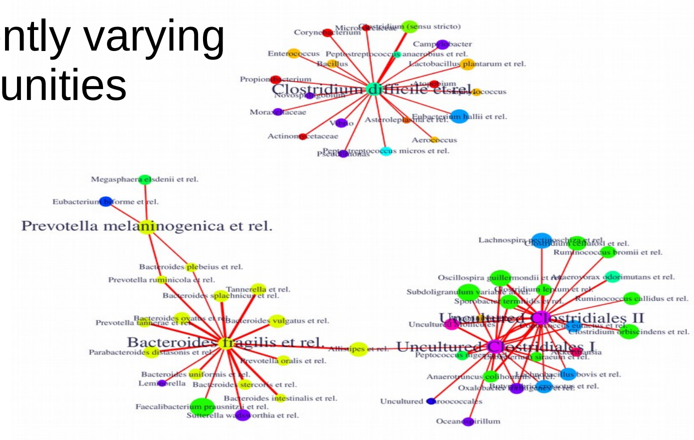
Lahti et al. Nat. Comm. 5:4344, 2014

Mixture models bring flexibility in modeling



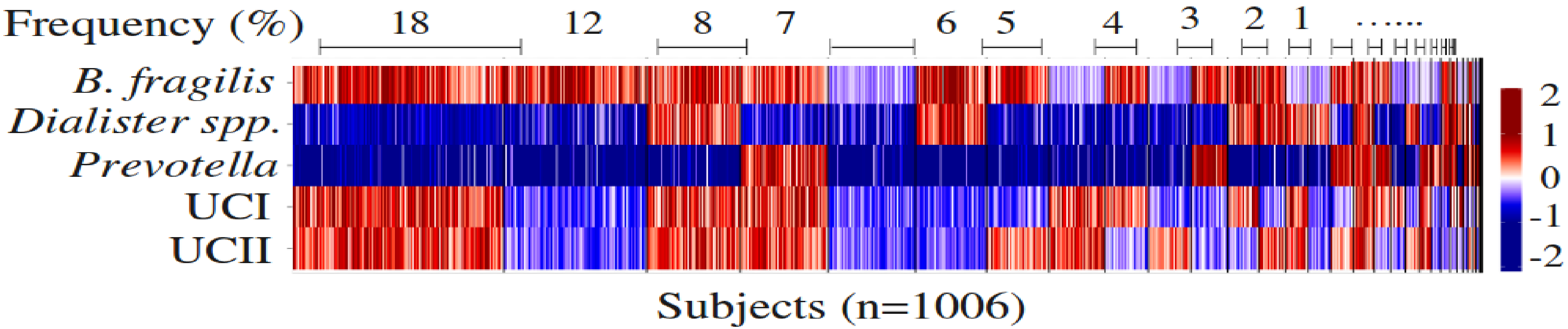
Bistable taxa:
Prevotella, Dialister,
Clostridiales..
(N=1006)

Independently varying
sub-communities



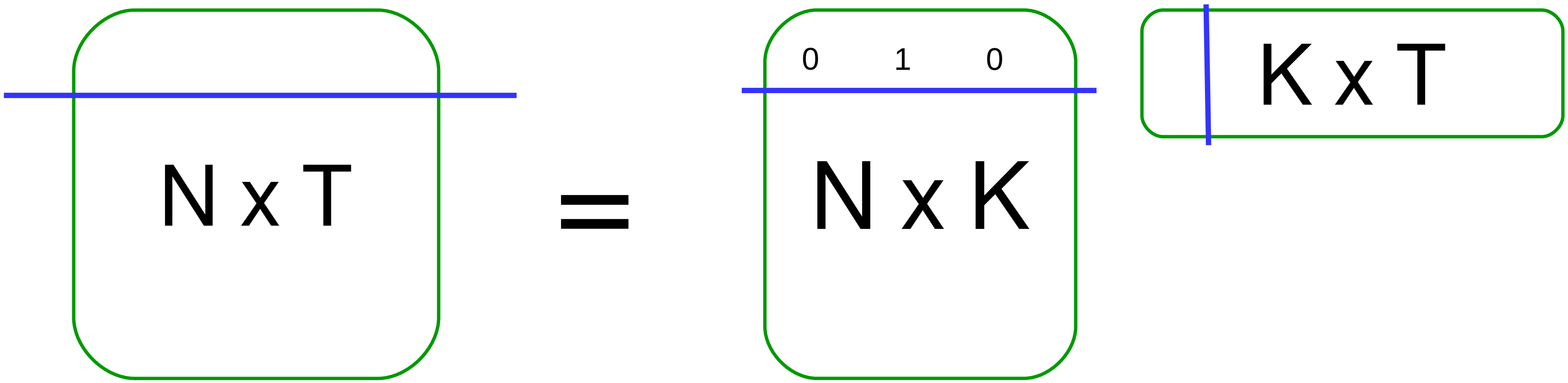
Ecosystem states are rich combinations
of independent tipping elements ?

Lahti et al. Nat. Comm. 2014



Clustering vs. Factorization

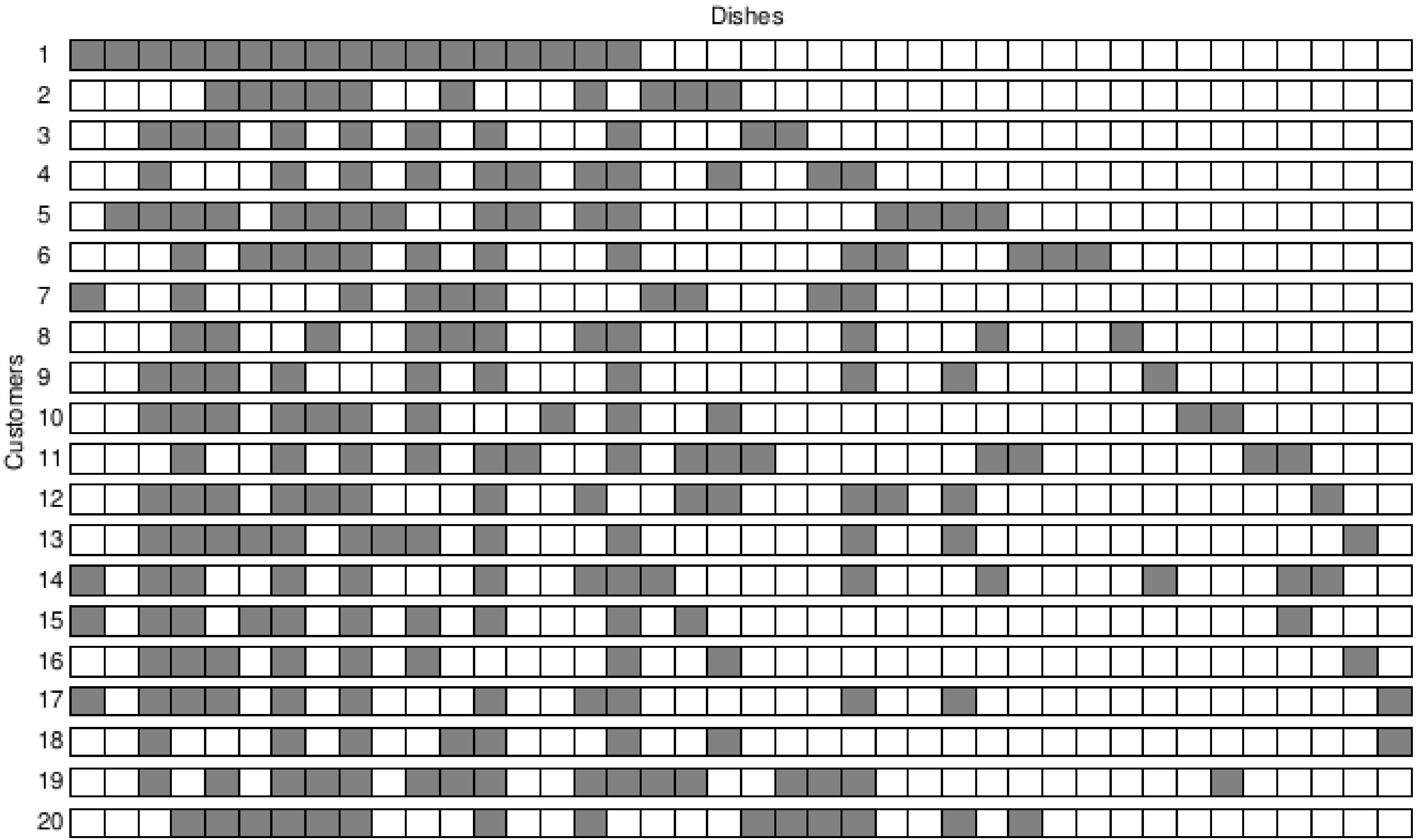
statistical formulation



Observed data

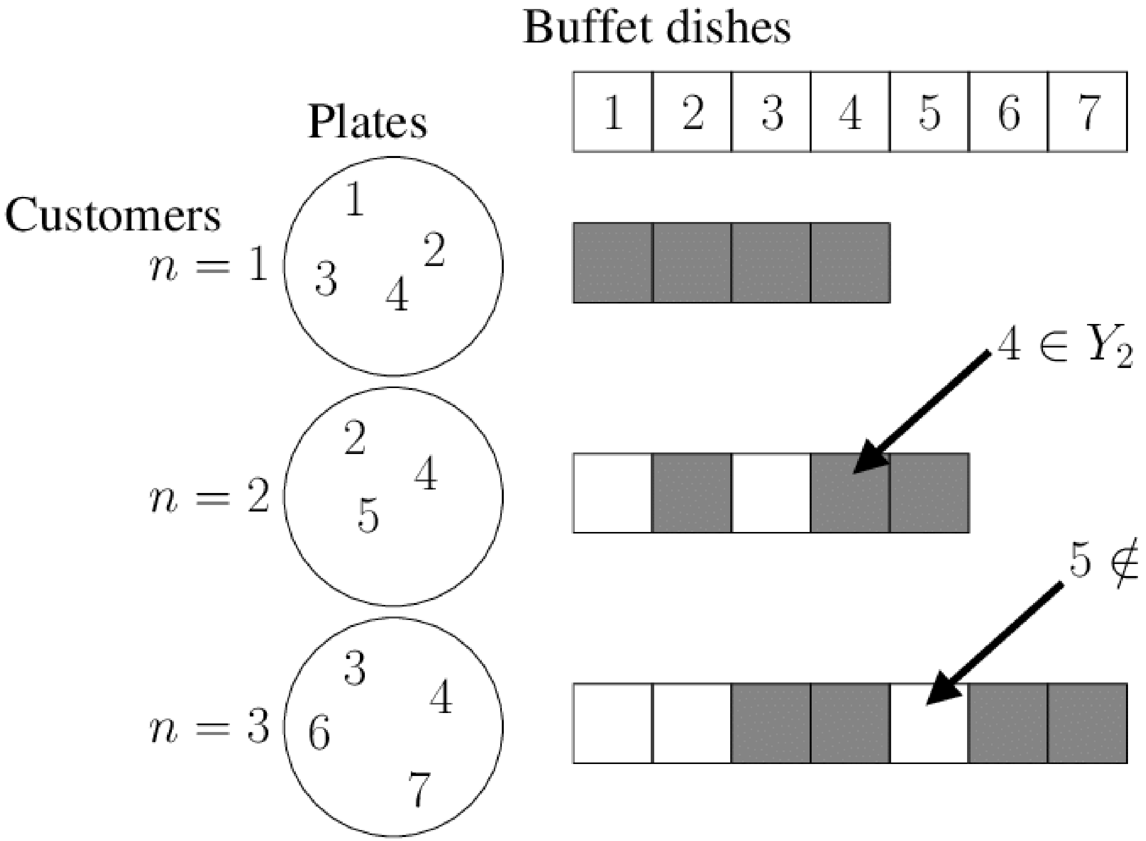
Component weights

Component profiles


Dishes
Customers
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

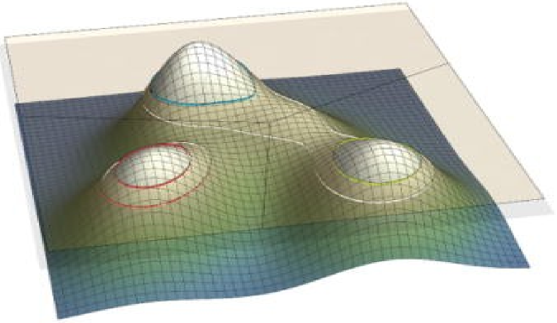
Indian Buffet Process

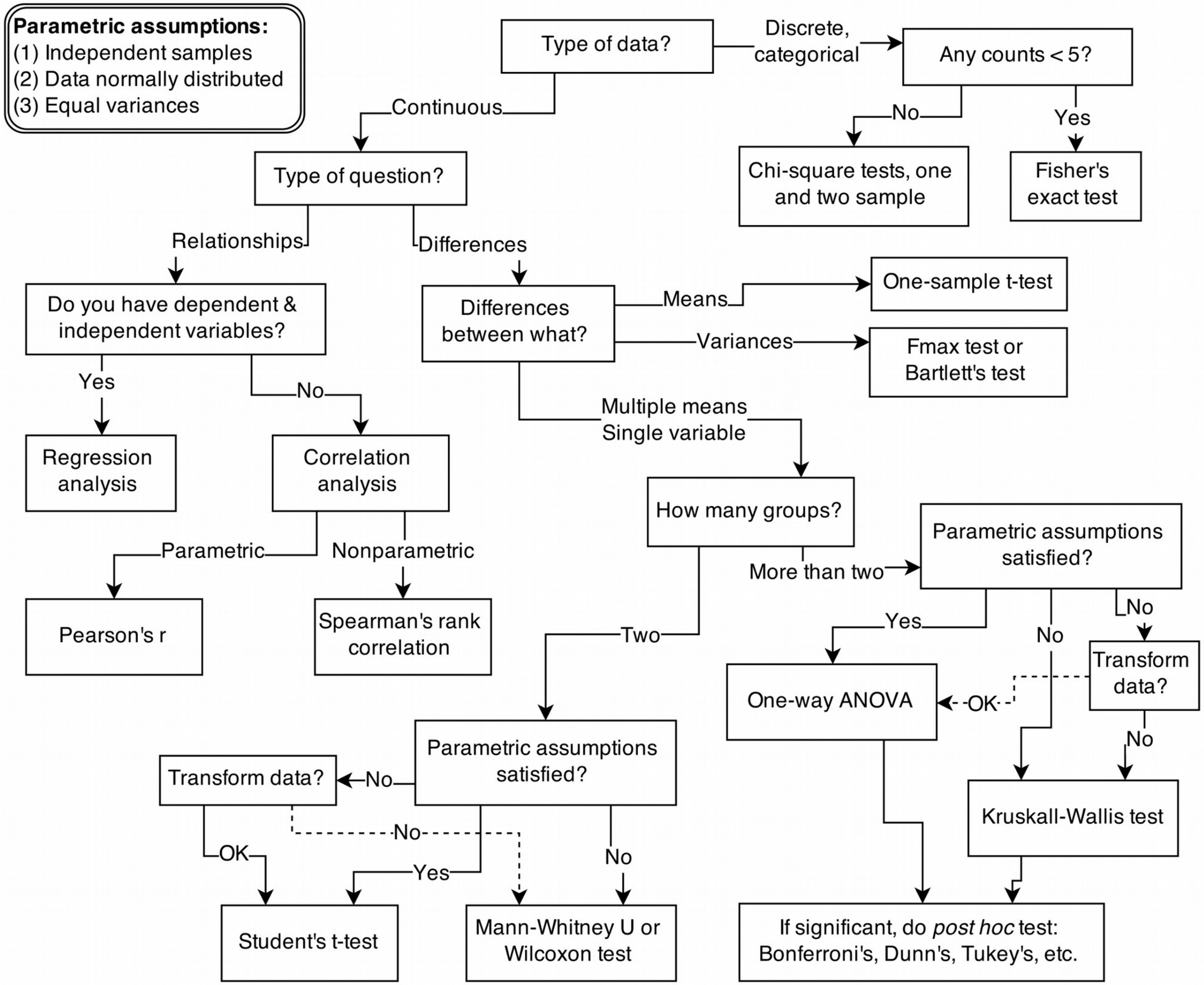
- Bayesian nonparametric model
- A prior on an infinite binary matrix.



Michael Jordan

How to choose a correct model or test for my data?





Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?

Discrete, categorical

Any counts < 5?

No

Chi-square tests, one and two sample

Yes

Fisher's exact test

Continuous

Type of question?

Relationships

Do you have dependent & independent variables?

Yes

Regression analysis

No

Correlation analysis

Parametric

Pearson's r

Nonparametric

Spearman's rank correlation

Differences

Differences between what?

Means

One-sample t-test

Variances

Fmax test or Bartlett's test

Multiple means Single variable

How many groups?

Two

Parametric assumptions satisfied?

Yes

Student's t-test

No

Transform data?

OK

No

Mann-Whitney U or Wilcoxon test

More than two

Parametric assumptions satisfied?

Yes

One-way ANOVA

No

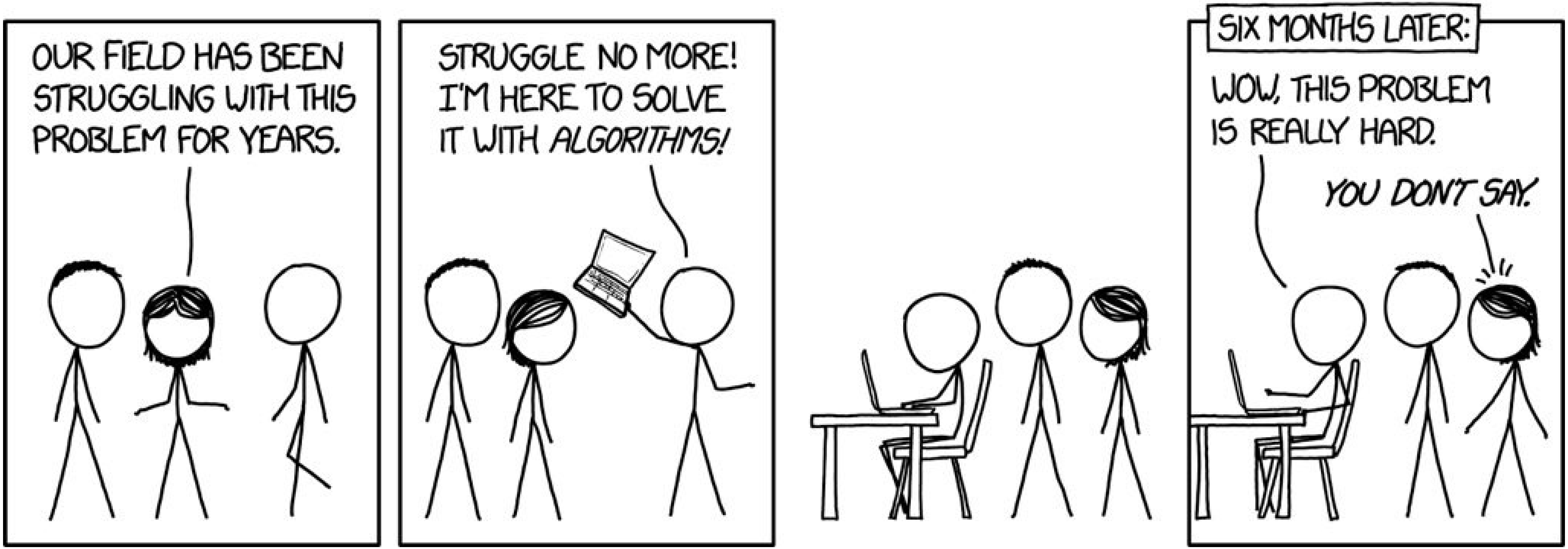
Transform data?

OK

No

Kruskall-Wallis test

If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH *ALGORITHMS!*
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.