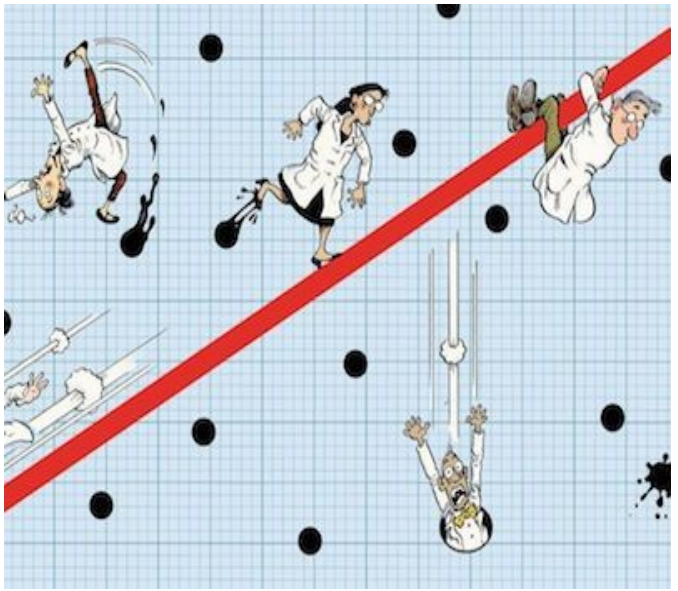
Introduction to statistical concepts & methods

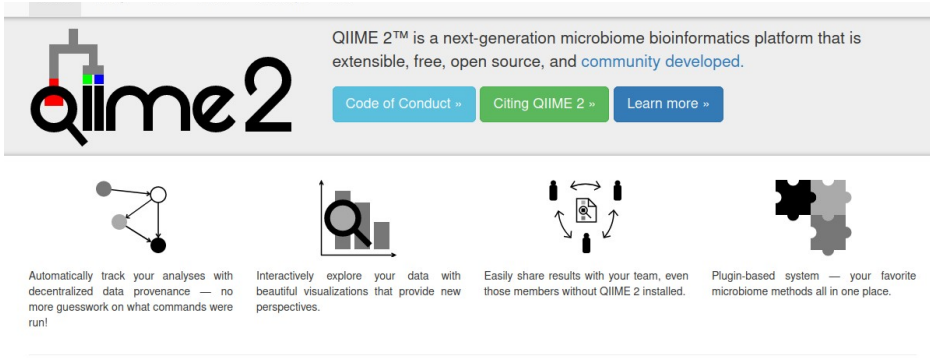


Picture: Nature Publishing Group

https://docs.google.com/document/d/1eBM4gevAKcOFw5gI6hbzgG_M_Ax4szFb0Nzq6yle338/edit?tab=t.0

HOLOGEN

open microbiome data science frameworks



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

PeerJ >

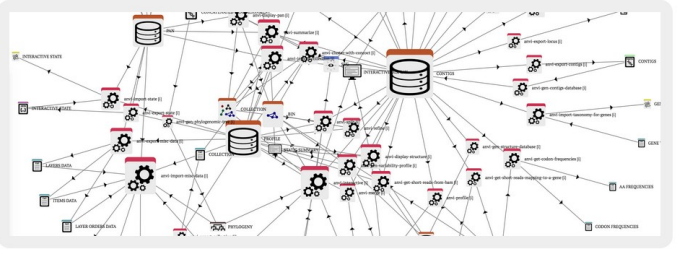
Anvi'o: an advanced analysis and visualization platform for 'omics data

Research article | Bioinformatics | Biotechnology | Computational Biology | Genomics | Microbiology

A. Murat Eren [1,2], Özcan C. Esen [1], Christopher Quince [3], Joseph H. Vineis [1], Hilary G. Morrison [1], Mitchell L. Sogin [1], Tom O. Delmont [1]

blished October 8, 2015

Anvi'o in a nutshell



Anvi'o is an open-source, community-driven analysis and visualization platform for 'omics data.

OPEN ACCESS

ESSAY

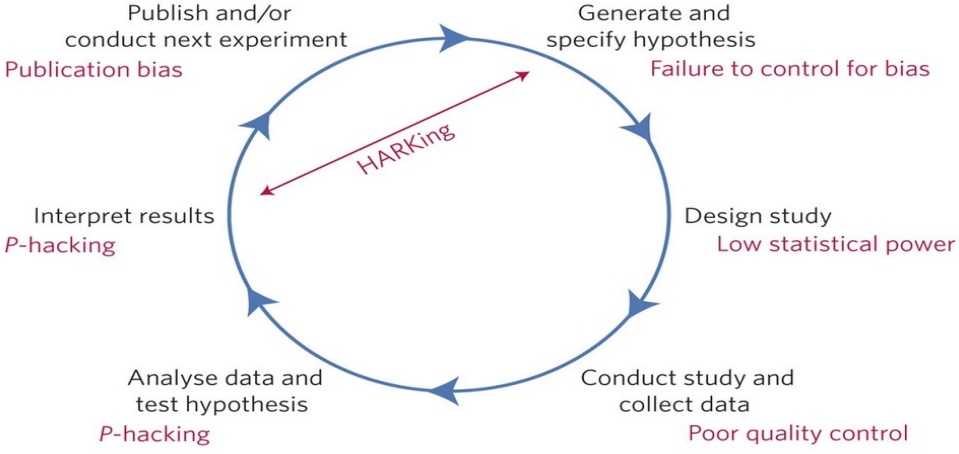
Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

Figure 1: Threats to reproducible science.

From: A manifesto for reproducible science



An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

A manifesto for reproducible science

Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

Learning goals - Statistical concepts

- Overview of common pitfalls in statistical analyses
- Typical ways to deal with these

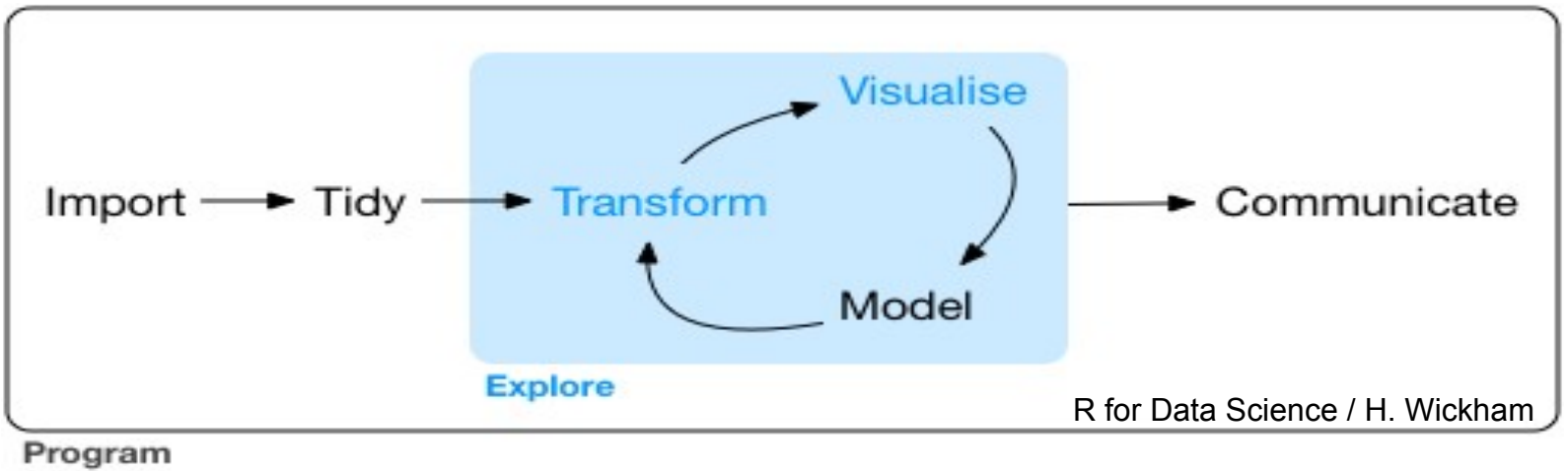
Contents

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.



R for Data Science / H. Wickham

Fundamental considerations in beta diversity analysis

Feature selection

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

Transformation

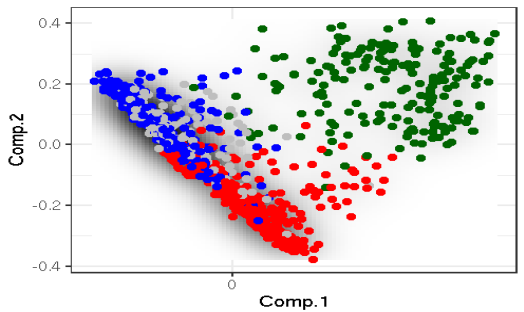
(absolute, compositional, CLR, Hellinger..?)

Dissimilarity measure

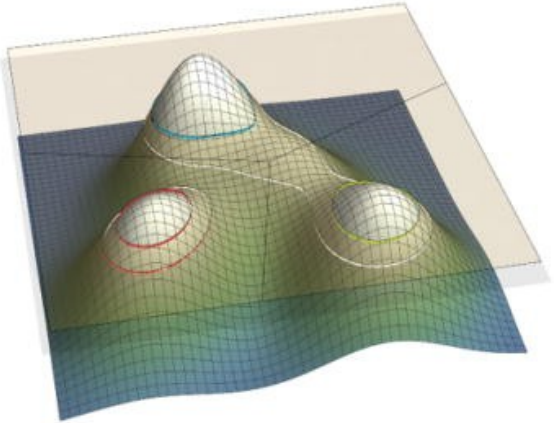
(Euclidean/L2, Bray-Curtis, Unifrac..?)

Analysis method

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



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

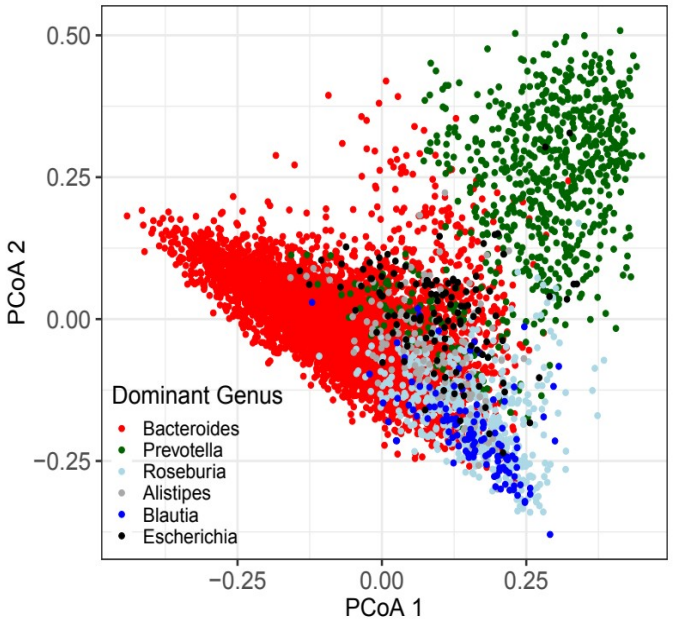
(Dis)similarity
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

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

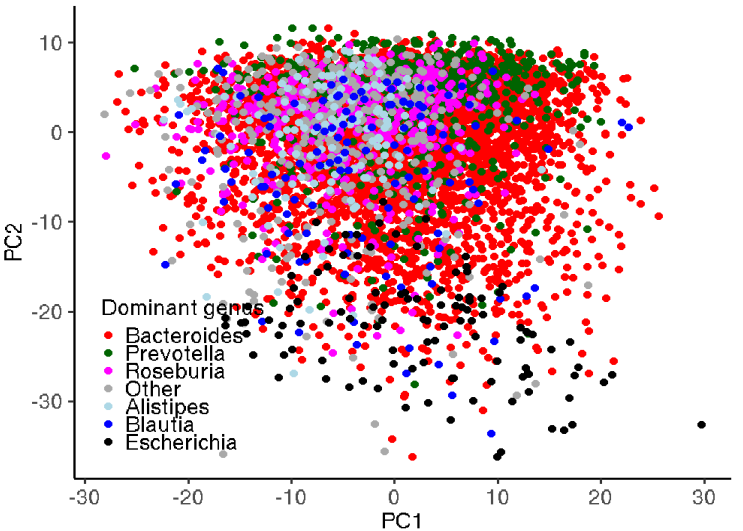
Regulation
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

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

PCoA + Bray-Curtis



PCA + Aitchison



Reproducible Research: Enterotype Example

Susan Holmes and Joey McMurdie

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

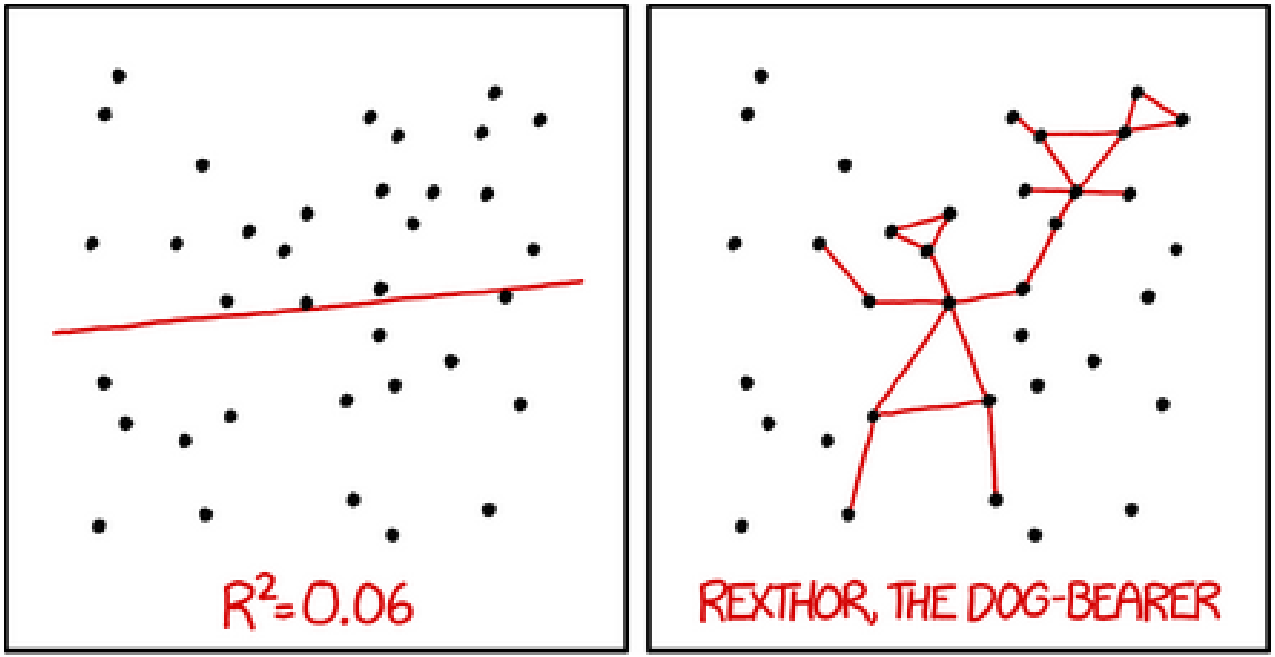
Comment on this paper

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

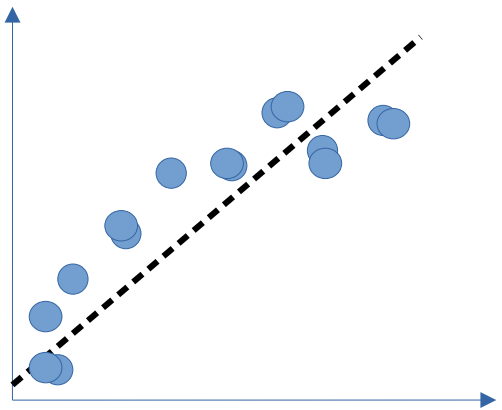
Aaro Salosensaari, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa,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

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

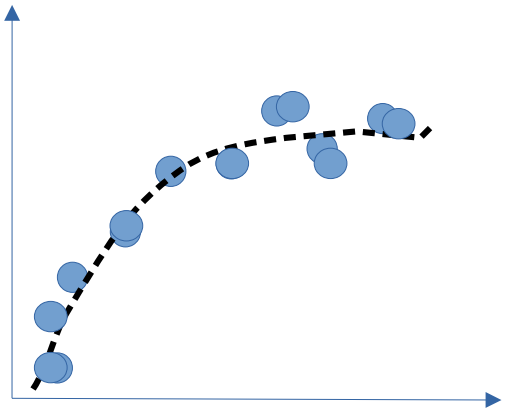
How we choose which model to apply?



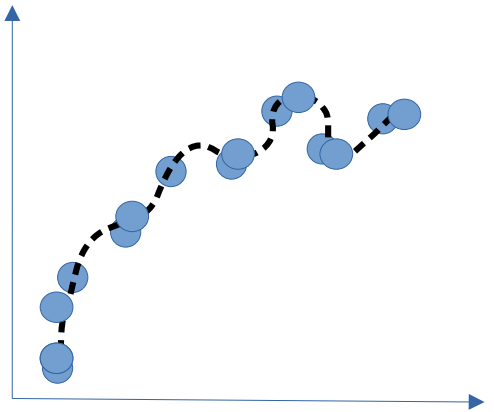
Tradeoff between robustness and sensitivity



Simple model: underfitting & high bias



Intermediate model – ”just right”

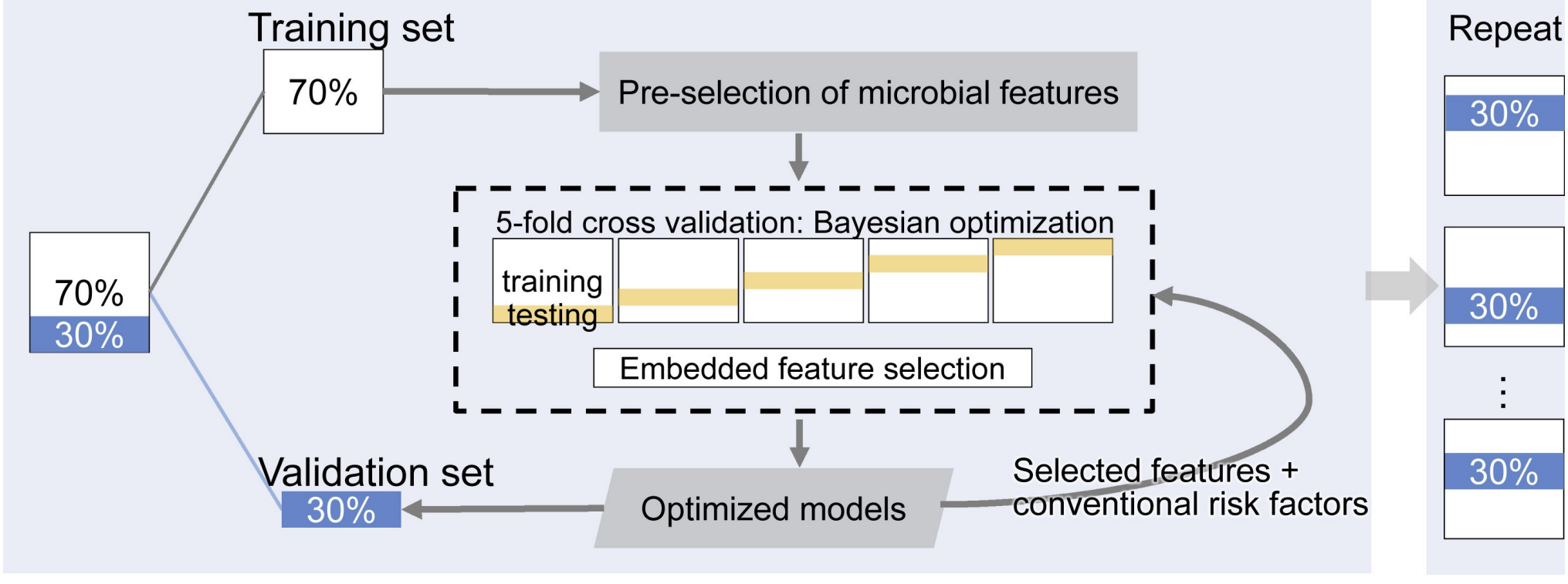


Complex model – overfitting & high variance

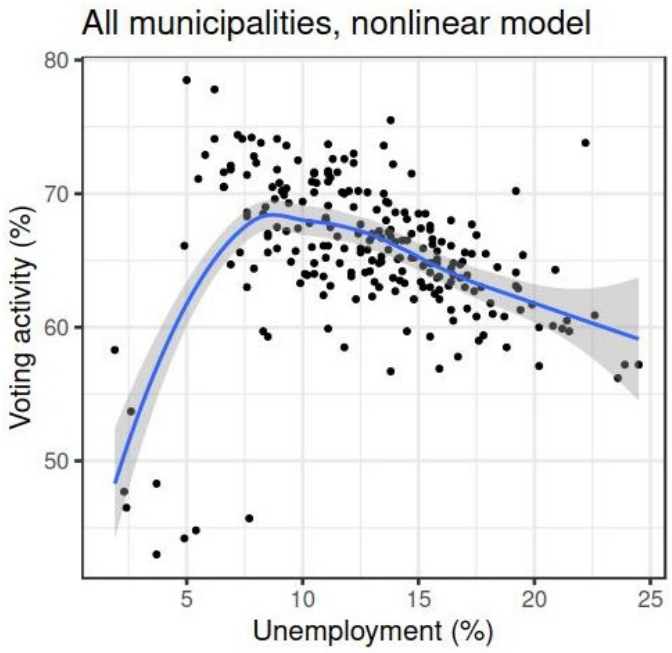
ARTICLE · Volume 34, Issue 5, P719-730.E4, May 03, 2022 · *Open Access* Download Full Issue

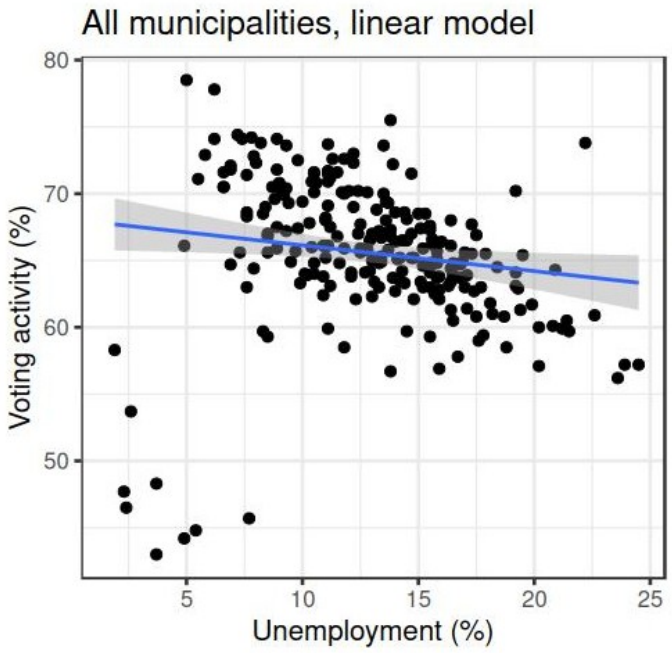
Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting

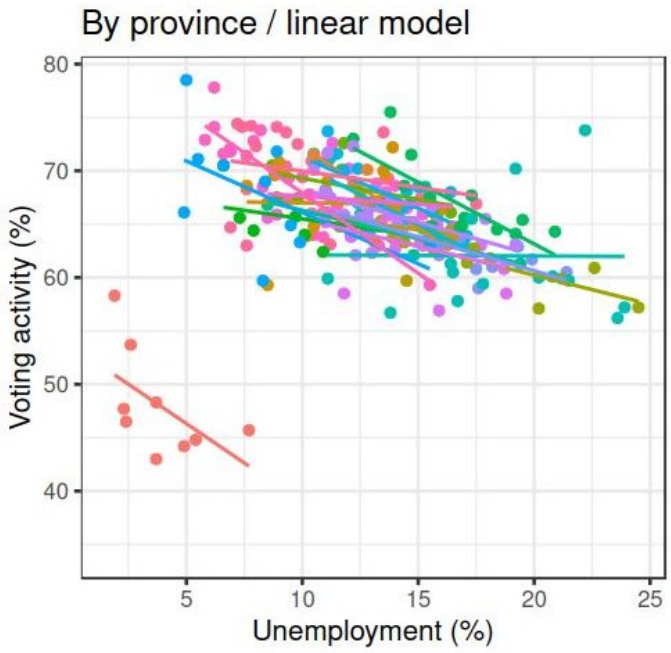
Yang Liu [1,2,24] · Guillaume Méric [1,2,3,4] · Aki S. Havulinna [5,6] · … · Veikko Salomaa [5] · Rob Knight [12,17,18] · Michael Inouye [1,2,3,7,20,21,22,23] … Show more



Simpson’s effect





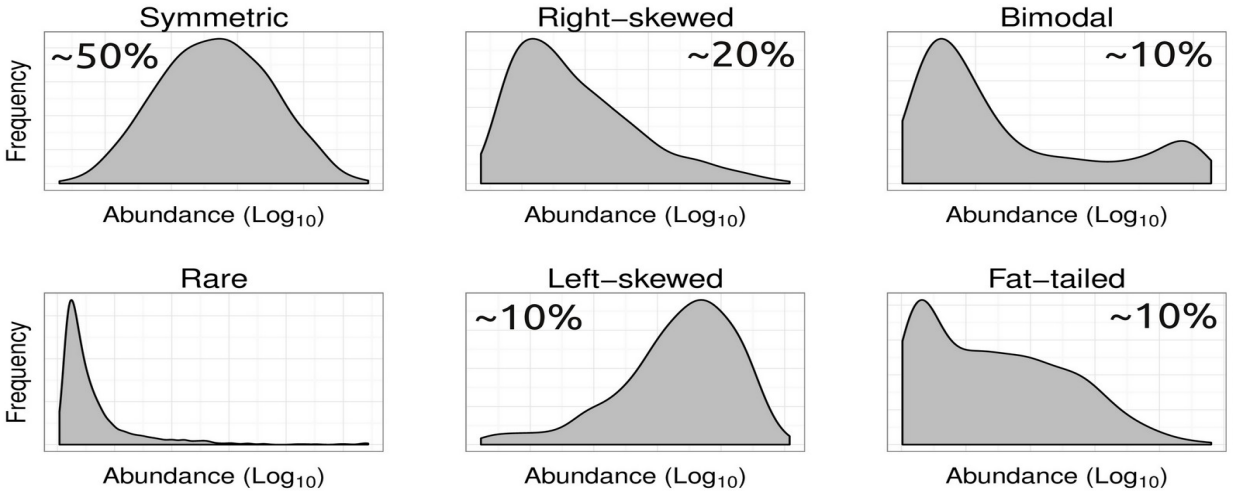


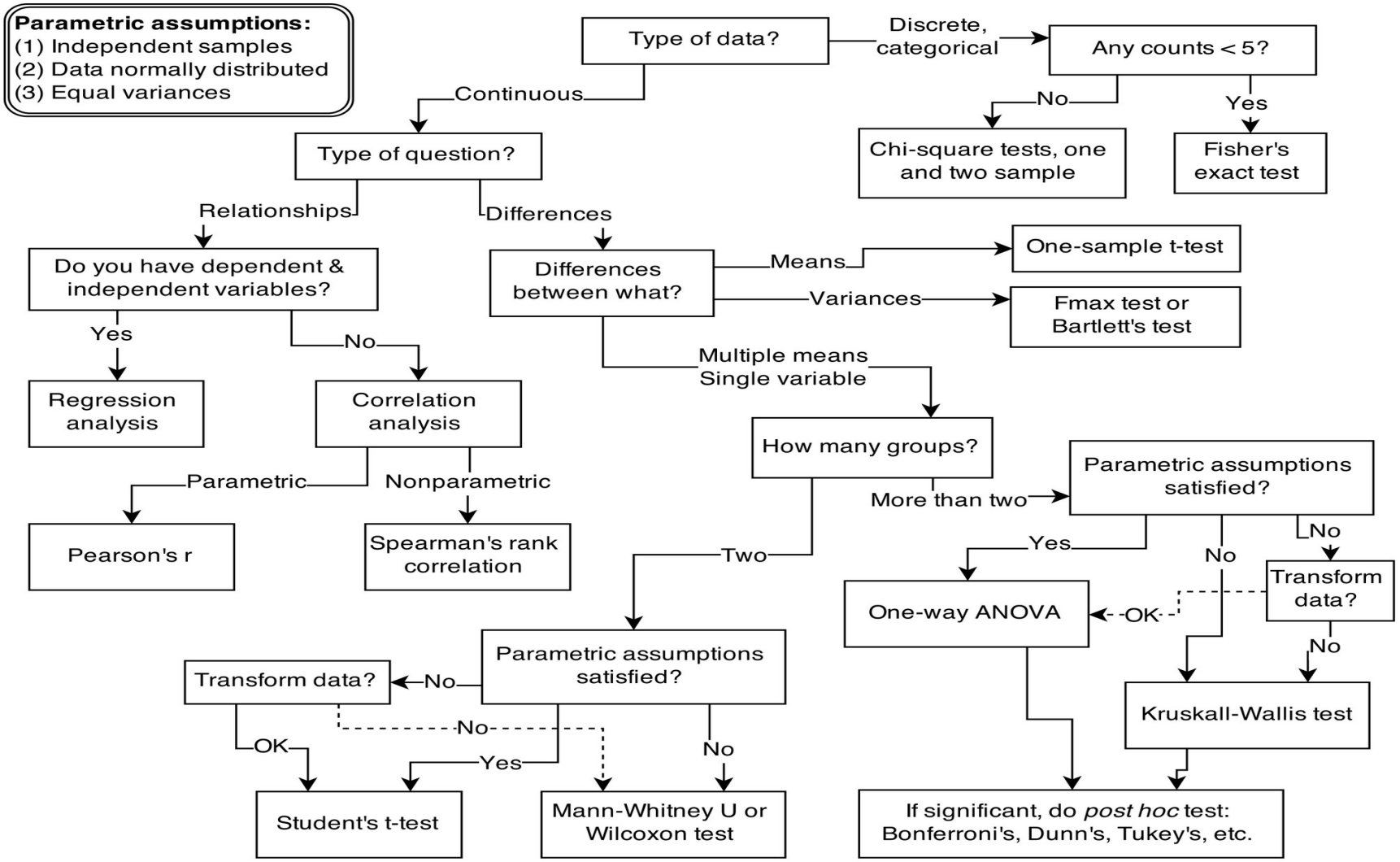
Differential abundance

Standard t-test for two-group comparison?

Problems:

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

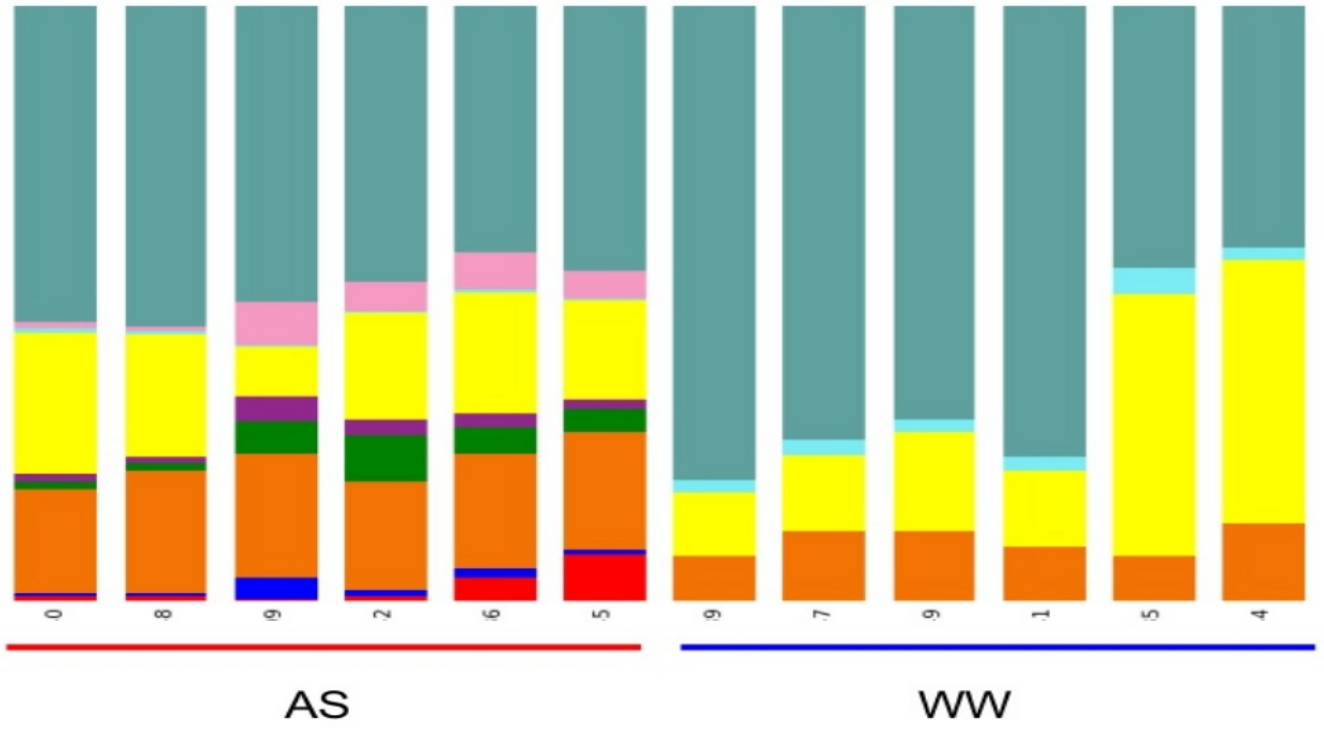




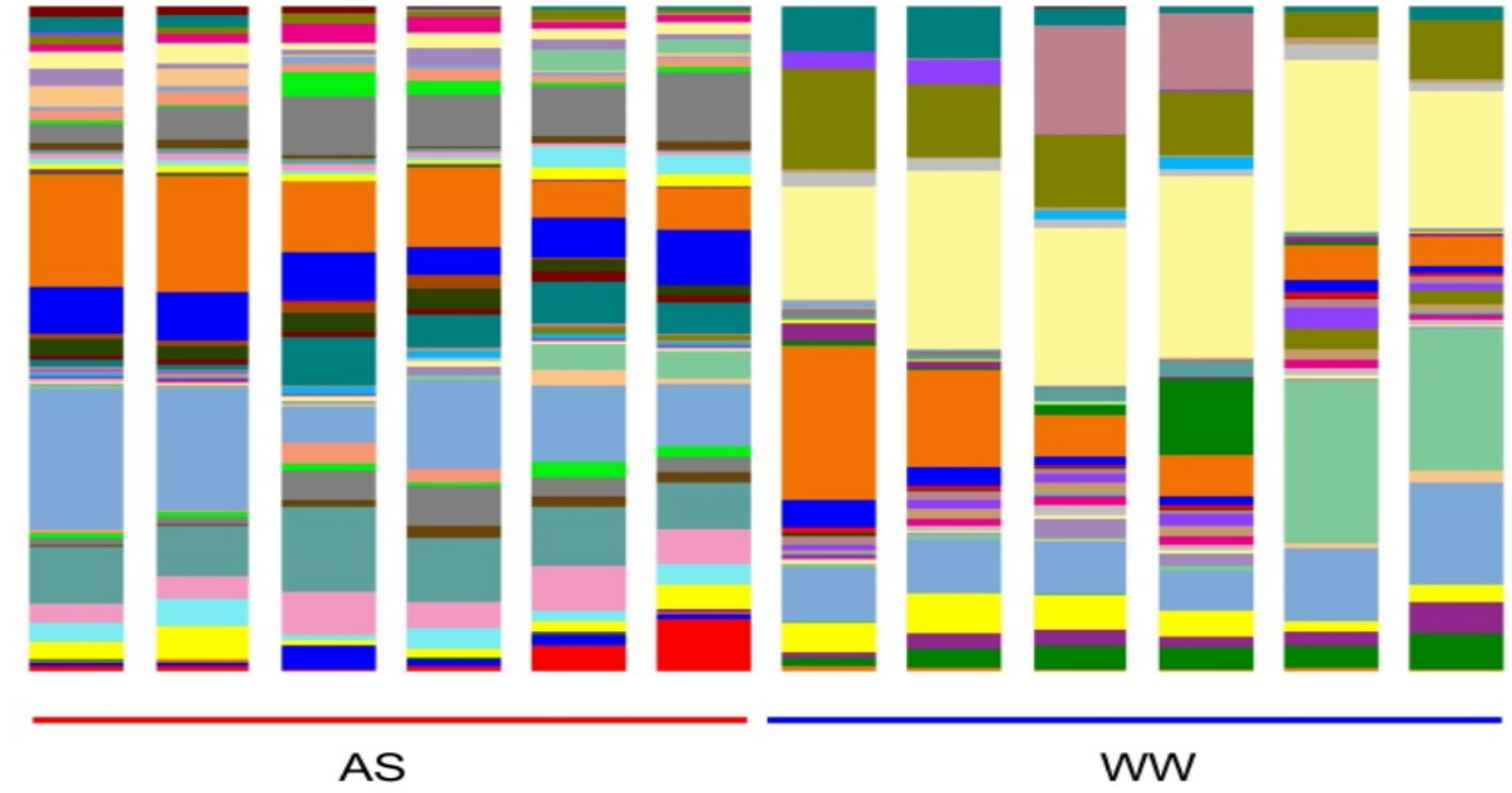
Richard McElreath HBES_plenary_2015_page3.png

(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)
a notable trend (p<0.1)
a numerical increasing trend (p=0.09)
a numerical trend (p=0.09)
a positive trend (p=0.09)
a possible trend (p=0.09)
a possible trend toward significance (p=0.052)
a pronounced trend (p=0.09)
a reliable trend (p=0.058)
a robust trend toward significance (p=0.0503)
a significant trend (p=0.09)
a slight slide towards significance (p<0.20)
a slight tendency toward significance(p<0.08)
a slight trend (p<0.09)
a slight trend toward significance (p=0.098)
a slightly increasing trend (p=0.09)
a small trend (p=0.09)
a statistical trend (p=0.09)
a statistical trend toward significance (p=0.09)
a strong tendency towards statistical significance (p=0.051)
a strong trend (p=0.077)
a strong trend toward significance (p=0.08)
a substantial trend toward significance (p=0.068)
a suggestive trend (p=0.06)
a trend close to significance (p=0.08)
a trend significance level (p=0.08)
a trend that approached significance (p<0.06)
a very slight trend toward significance (p=0.20)
a weak trend (p=0.09)
a weak trend toward significance (p=0.12)
an evident trend (p=0.13)
an expected trend (p=0.08)
an important trend (p=0.066)
an increasing trend (p<0.09)
an interesting trend (p=0.1)
an inverse trend toward significance (p=0.06)
an observed trend (p=0.06)
an obvious trend (p=0.06)
an overall trend (p=0.2)
an unexpected trend (p=0.09)
an unexplained trend (p=0.09)
an unfavorable trend (p<0.10)
appeared to be marginally significant (p<0.10)
approached acceptable levels of statistical significance (p=0.054)
approached but did not quite achieve significance (p>0.05)
approached but fell short of significance (p=0.07)
approached conventional levels of significance (p<0.10)
approached near significance (p=0.06)
approached our criterion of significance (p>0.08)
approached significant (p=0.11)
approached the borderline of significance (p=0.07)
approached the level of significance (p=0.09)
approached trend levels of significance (p0.05)
approached, but did reach, significance (p=0.065)
approaches but fails to achieve a customary level of statistical significance (p=0.154)
approaches statistical significance (p>0.06)
approaching a level of significance (p=0.089)
approaching an acceptable significance level (p=0.056)
approaching borderline significance (p=0.08)
approaching borderline statistical significance (p=0.07)
approaching but not reaching significance (p=0.53)
approaching clinical significance (p=0.07)
approaching close to significance (p<0.1)
approaching conventional significance levels (p=0.06)
approaching conventional statistical significance (p=0.06)
approaching formal significance (p=0.1052)
approaching independent prognostic significance (p=0.08)
approaching marginal levels of significance p<0.107)
at the cusp of significance (p=0.06)
at the edge of significance (p=0.055)
at the limit of significance (p=0.054)
at the limits of significance (p=0.053)
at the margin of significance (p=0.056)
at the margin of statistical significance (p<0.07)
at the verge of significance (p=0.058)
at the very edge of significance (p=0.053)
barely below the level of significance (p=0.06)
barely escaped statistical significance (p=0.07)
barely escapes being statistically significant at the 5% risk level (0.1>p>0.05)
barely failed to attain statistical significance (p=0.067)
barely fails to attain statistical significance at conventional levels (p<0.10
barely insignificant (p=0.075)
barely missed statistical significance (p=0.051)
barely missed the commonly acceptable significance level (p<0.053)
barely outside the range of significance (p=0.06)
barely significant (p=0.07)
below (but verging on) the statistical significant level (p>0.05)
better trends of improvement (p=0.056)
bordered on a statistically significant value (p=0.06)
bordered on being significant (p>0.07)
bordered on being statistically significant (p=0.0502)
bordered on but was not less than the accepted level of significance (p>0.05)
bordered on significant (p=0.09)
borderline conventional significance (p=0.051)
borderline level of statistical significance (p=0.053)
borderline significant (p=0.09)
borderline significant trends (p=0.099)
close to a marginally significant level (p=0.06)
close to being significant (p=0.06)
close to being statistically significant (p=0.055)
close to borderline significance (p=0.072)
close to the boundary of significance (p=0.06)
close to the level of significance (p=0.07)
close to the limit of significance (p=0.17)
close to the margin of significance (p=0.055)
close to the margin of statistical significance (p=0.075)
closely approaches the brink of significance (p=0.07)
closely approaches the statistical significance (p=0.0669)
closely approximating significance (p>0.05)
closely not significant (p=0.06)
closely significant (p=0.058)
close-to-significant (p=0.09)
did not achieve conventional threshold levels of statistical significance (p=0.08)
did not exceed the conventional level of statistical significance (p<0.08)
did not quite achieve acceptable levels of statistical significance (p=0.054)
did not quite achieve significance (p=0.076)
did not quite achieve the conventional levels of significance (p=0.052)
effectively significant (p=0.051)
equivocal significance (p=0.06)
essentially significant (p=0.10)
extremely close to significance (p=0.07)
failed to reach significance on this occasion (p=0.09)
failed to reach statistical significance (p=0.06)
fairly close to significance (p=0.065)
fairly significant (p=0.09)
falls just short of standard levels of statistical significance (p=0.06)
fell (just) short of significance (p=0.08)
fell barely short of significance (p=0.08)
fell just short of significance (p=0.07)
fell just short of statistical significance (p=0.12)
fell just short of the traditional definition of statistical significance (p=0.051)
fell marginally short of significance (p=0.07)
fell narrowly short of significance (p=0.0623)
fell only marginally short of significance (p=0.0879)
fell only short of significance (p=0.06)
fell short of significance (p=0.07)
fell slightly short of significance (p>0.0167)
fell somewhat short of significance (p=0.138)
felt short of significance (p=0.07)
flirting with conventional levels of significance (p>0.1)
heading towards significance (p=0.086)
highly significant (p=0.09)
hint of significance (p>0.05)
hovered around significance (p = 0.061)
hovered at nearly a significant level (p=0.058)
hovering closer to statistical significance (p=0.076)
hovers on the brink of significance (p=0.055)
in the edge of significance (p=0.059)
in the verge of significance (p=0.06)
inconclusively significant (p=0.070)
indeterminate significance (p=0.08)
indicative significance (p=0.08)
is just outside the conventional levels of significance
just about significant (p=0.051)
just above the arbitrary level of significance (p=0.07)
just above the margin of significance (p=0.053)
just at the conventional level of significance (p=0.05001)
just barely below the level of significance (p=0.06)
just barely failed to reach significance (p<0.06)
just barely insignificant (p=0.11)
just barely statistically significant (p=0.054)
just beyond significance (p=0.06)
just borderline significant (p=0.058)
just escaped significance (p=0.07)
just failed significance (p=0.057)
just failed to be significant (p=0.072)
just failed to reach statistical significance (p=0.06)
just failing to reach statistical significance (p=0.06)
just fails to reach conventional levels of statistical significance (p=0.07)
just very slightly missed the significance level (p=0.086)
leaning towards significance (p=0.15)
leaning towards statistical significance (p=0.06)
likely to be significant (p=0.054)
loosely significant (p=0.10)
marginal significance (p=0.07)
marginally and negatively significant (p=0.08)
marginally insignificant (p=0.08)
marginally nonsignificant (p=0.096)
marginally outside the level of significance
marginally significant (p>=0.1)
marginally significant tendency (p=0.08)
marginally statistically significant (p=0.08)
may not be significant (p=0.06)
medium level of significance (p=0.051)
mildly significant (p=0.07)
missed narrowly statistical significance (p=0.054)
moderately significant (p>0.11)
modestly significant (p=0.09)
narrowly avoided significance (p=0.052)
narrowly eluded statistical significance (p=0.0789)
narrowly escaped significance (p=0.08)
narrowly evaded statistical significance (p>0.05)
narrowly failed significance (p=0.054)
narrowly missed achieving significance (p=0.055)
narrowly missed overall significance (p=0.06)
narrowly missed significance (p=0.051)
narrowly missed standard significance levels (p<0.07)
narrowly missed the significance level (p=0.07)
narrowly missing conventional significance (p=0.054)
near limit significance (p=0.073)
near miss of statistical significance (p>0.1)
near nominal significance (p=0.064)
near significance (p=0.07)
near to statistical significance (p=0.056)
near/possible significance(p=0.0661)
near-borderline significance (p=0.10)
near-certain significance (p=0.07)
nearing significance (p<0.051)
nearly acceptable level of significance (p=0.06)
nearly approaches statistical significance (p=0.079)
nearly borderline significance (p=0.052)
nearly negatively significant (p<0.1)
nearly positively significant (p=0.063)
nearly reached a significant level (p=0.07)
nearly reaching the level of significance (p<0.06)
nearly significant (p=0.06)
nearly significant tendency (p=0.06)
nearly, but not quite significant (p>0.06)
near-marginal significance (p=0.18)
near-significant (p=0.09)
near-to-significance (p=0.093)
near-trend significance (p=0.11)
nominally significant (p=0.08)
non-insignificant result (p=0.500)
non-significant in the statistical sense (p>0.05
not absolutely significant but very probably so (p>0.05)
not as significant (p=0.06)
not clearly significant (p=0.08)
not completely significant (p=0.07)

Phylum level Classification



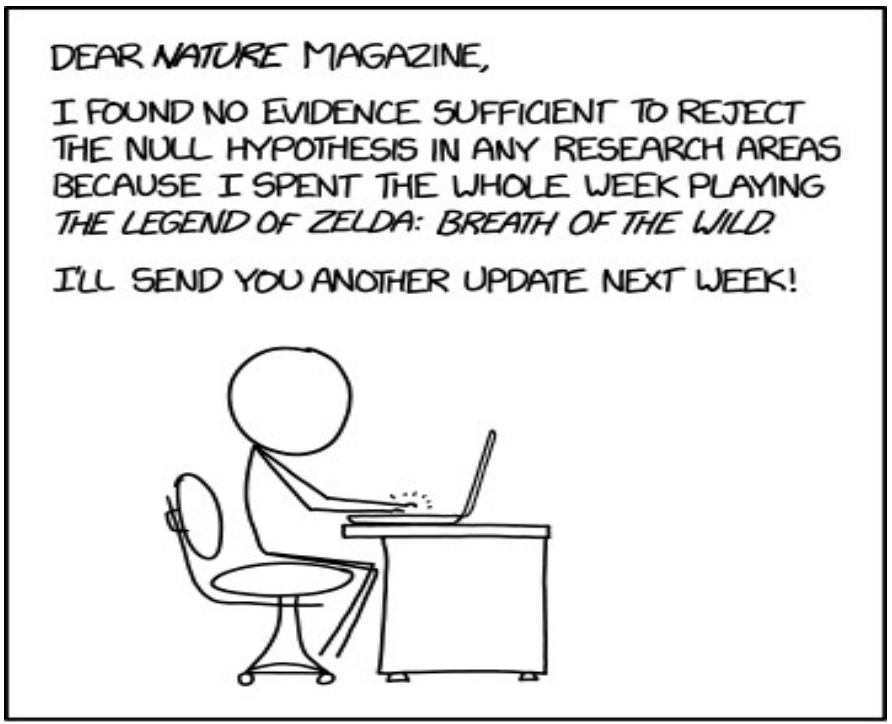
Genus level Classification



absence of evidence

is not

evidence of absence



http://web.stanford.edu/class/cs109l/unrestricted/images/

Learning goals - Statistical methods

Overview of common methods for statistical analyses

1) Data cleaning
- Normalization, denoising & preprocessing

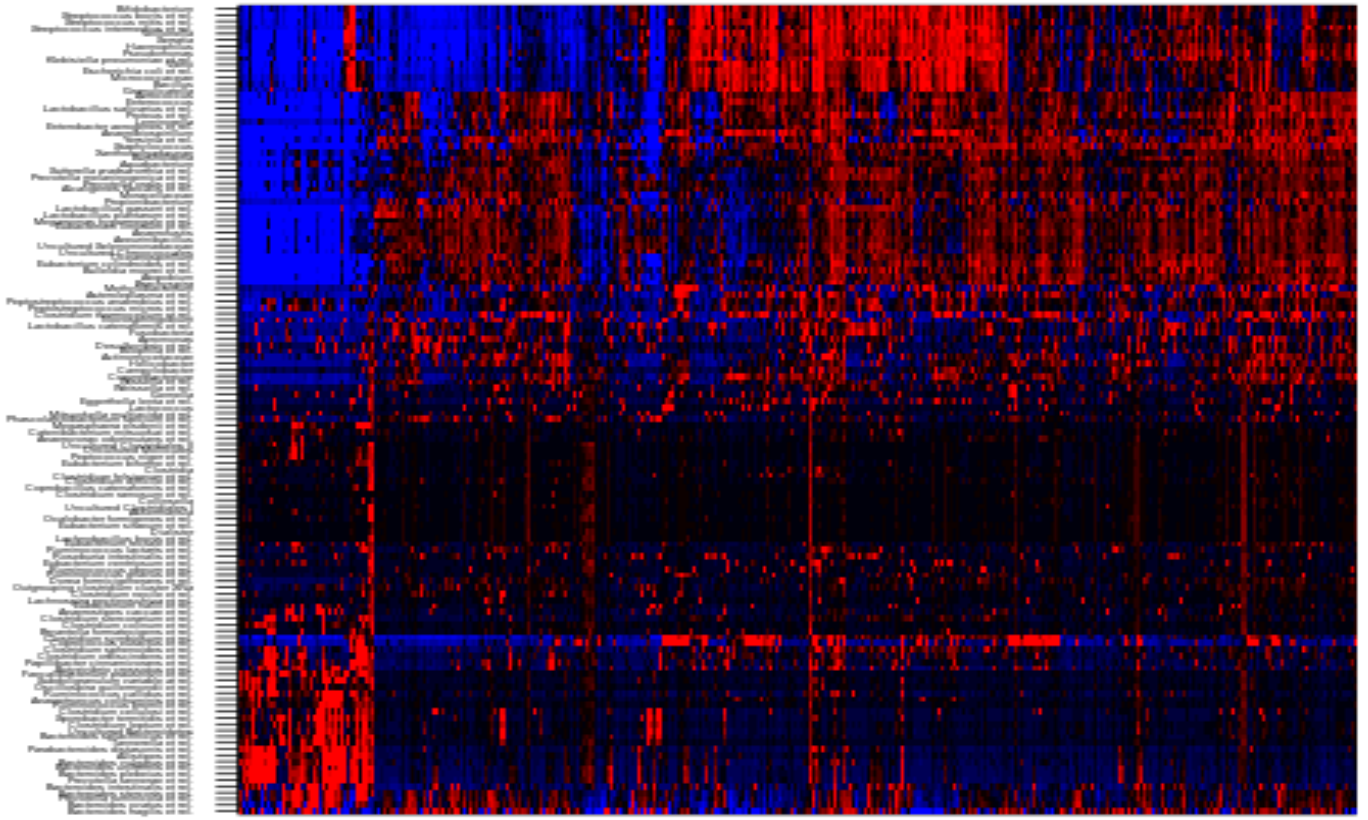
2) Data analysis
- Linear model vs. nonlinear models
- Parametric vs. non-parametric tests
- Bias & confounders: modeling the covariates
- Power calculations

3) Model validation
- Overfitting & cross-validation (internal, external)
- Garden of the forking paths: model comparisons & information criteria
- P-hacking: sensitivity analyses

4) Data visualization

Organisms and samples are not independent

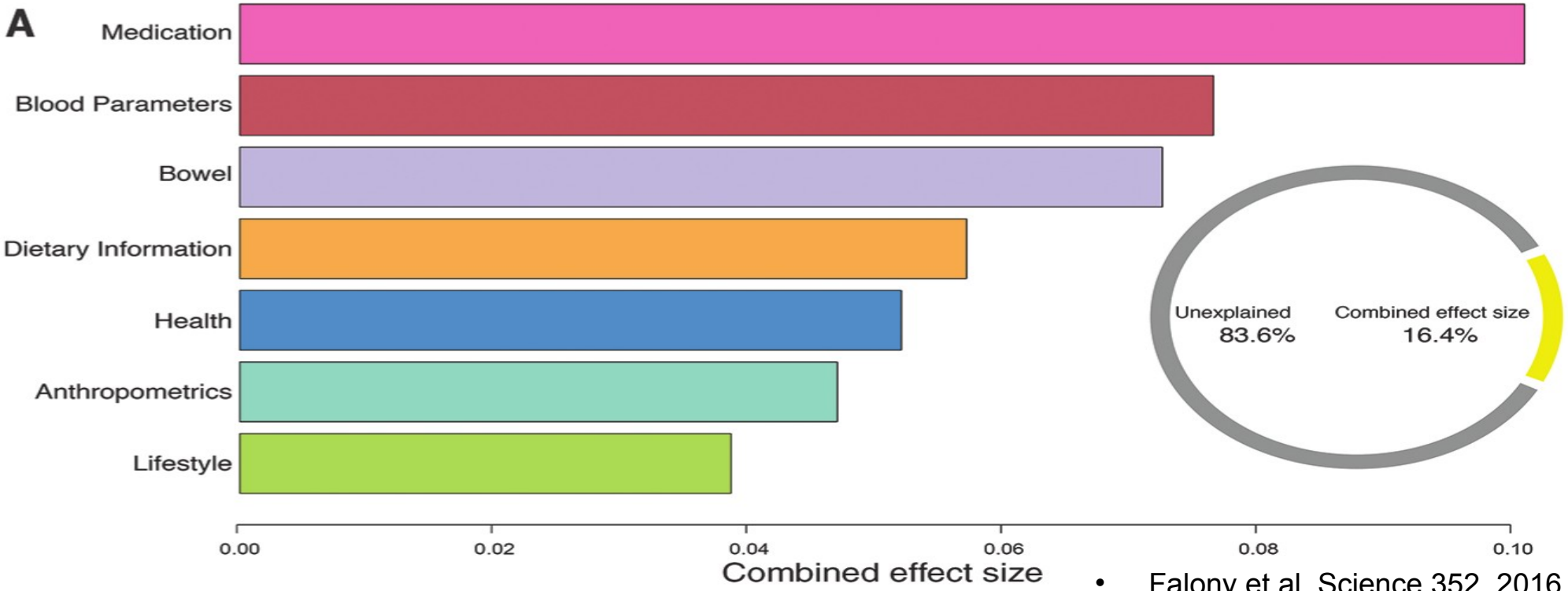
understanding & modeling the (latent) structure(s)



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 !

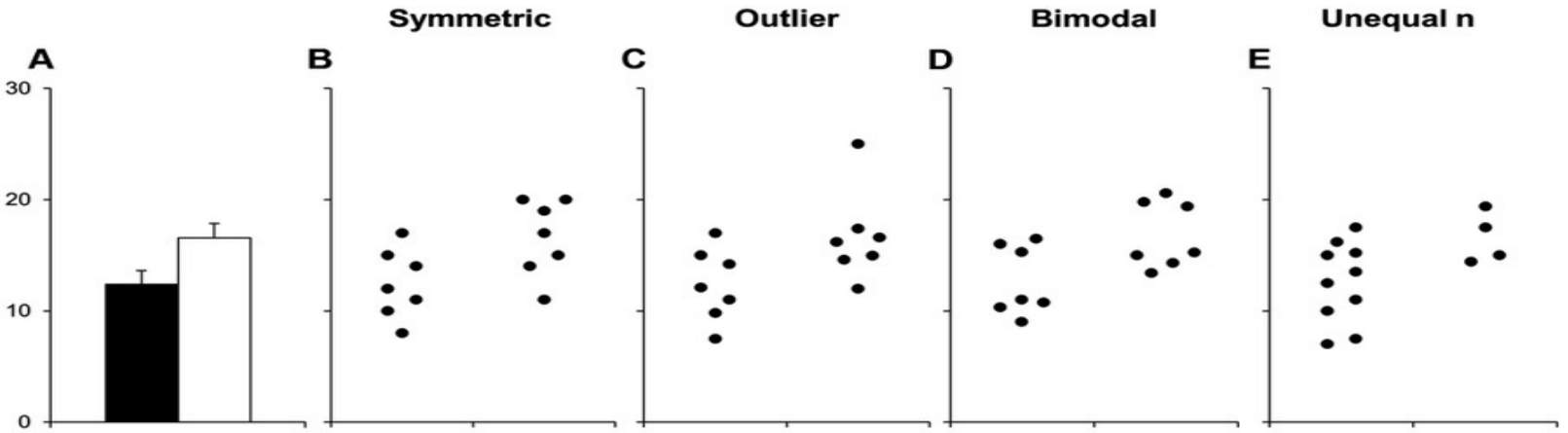


- Falony et al. Science 352, 2016.

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

Editorial

The ASA Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

Pages 129-133 | Accepted author version posted online: 07 Mar 2016, Published online:09 Jun 2016

Download citation https://doi.org/10.1080/00031305.2016.1154108 Check for updates

ASA advises researchers to avoid drawing scientific conclusions or making policy decisions based on P values alone. Researchers should describe not only the data analyses that produced statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust. "the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation"

1) P-values can indicate how incompatible the data are
2) with a specified statistical model.
3) P-values do not measure the probability that the studied
4) hypothesis is true, or the probability that the data were
5) produced by random chance alone.
6) Scientific conclusions and business or policy decisions
7) should not be based only on whether a p-value passes
8) a specific threshold.
9) Proper inference requires full reporting and transparency.
10) A p-value, or statistical significance, does not measure the
11) size of an effect or the importance of a result.
12) By itself, a p-value does not provide a good measure of
13) evidence regarding a model or hypothesis.

"Statisticians often supplement or even replace p-values with other approaches. These include methods "that emphasize estimation over testing such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence such as likelihood ratios or Bayes factors; and other approaches such as decision-theoretic modeling and false discovery rates."

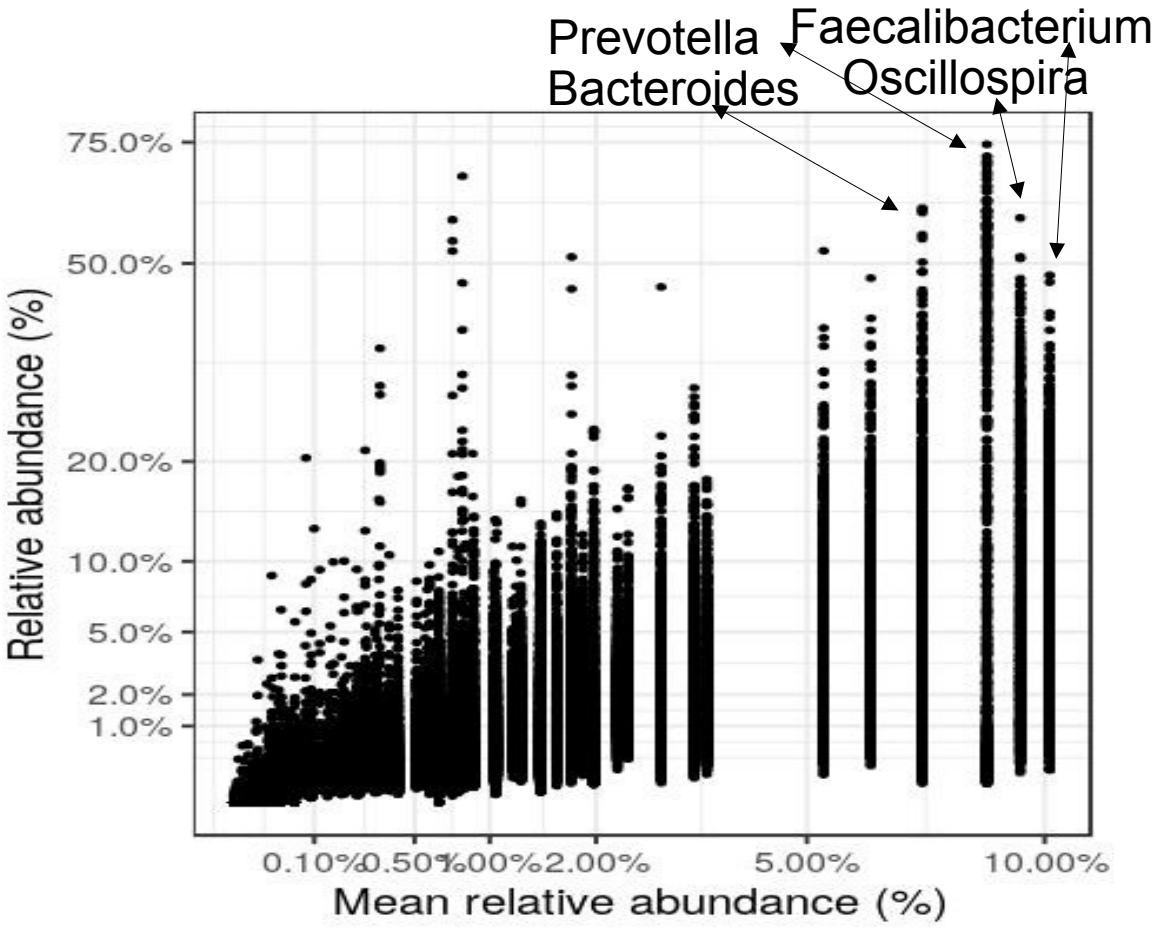
"The p-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'"

Taylor's law (in HITChip Atlas)

Heteroschedasticity:
Variance increases with the mean

Overdispersion:
Variance increases faster than proposed by the model

Data: HITChip Atlas

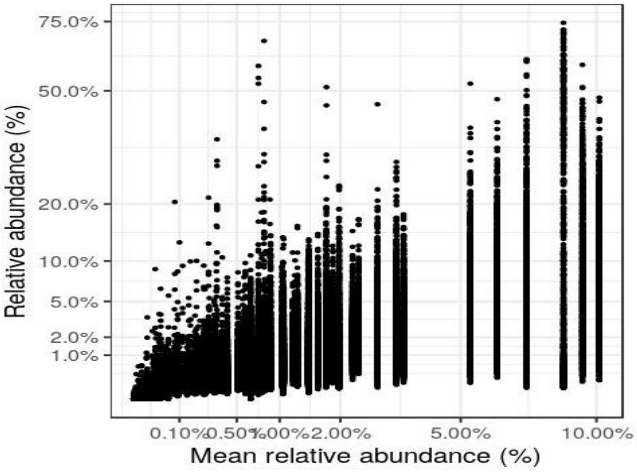


Microbiome data properties

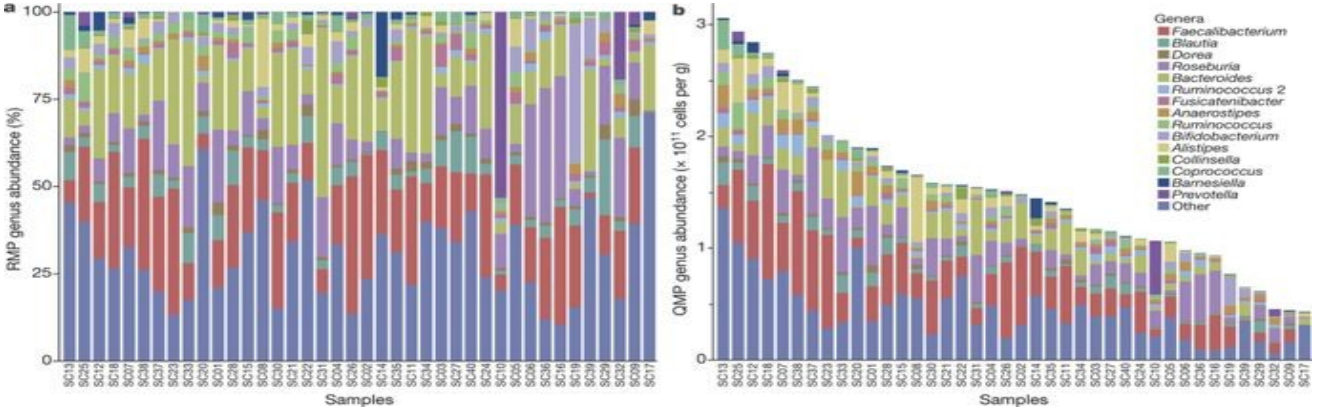
- Sparse
- Non-Gaussian
- Overdispersed
- Compositional
- Complex
- Stochastic
- Hierarchical

Heteroschedasticity

Data: HITChip Atlas (Lahti et al. Nature Communications 2014)



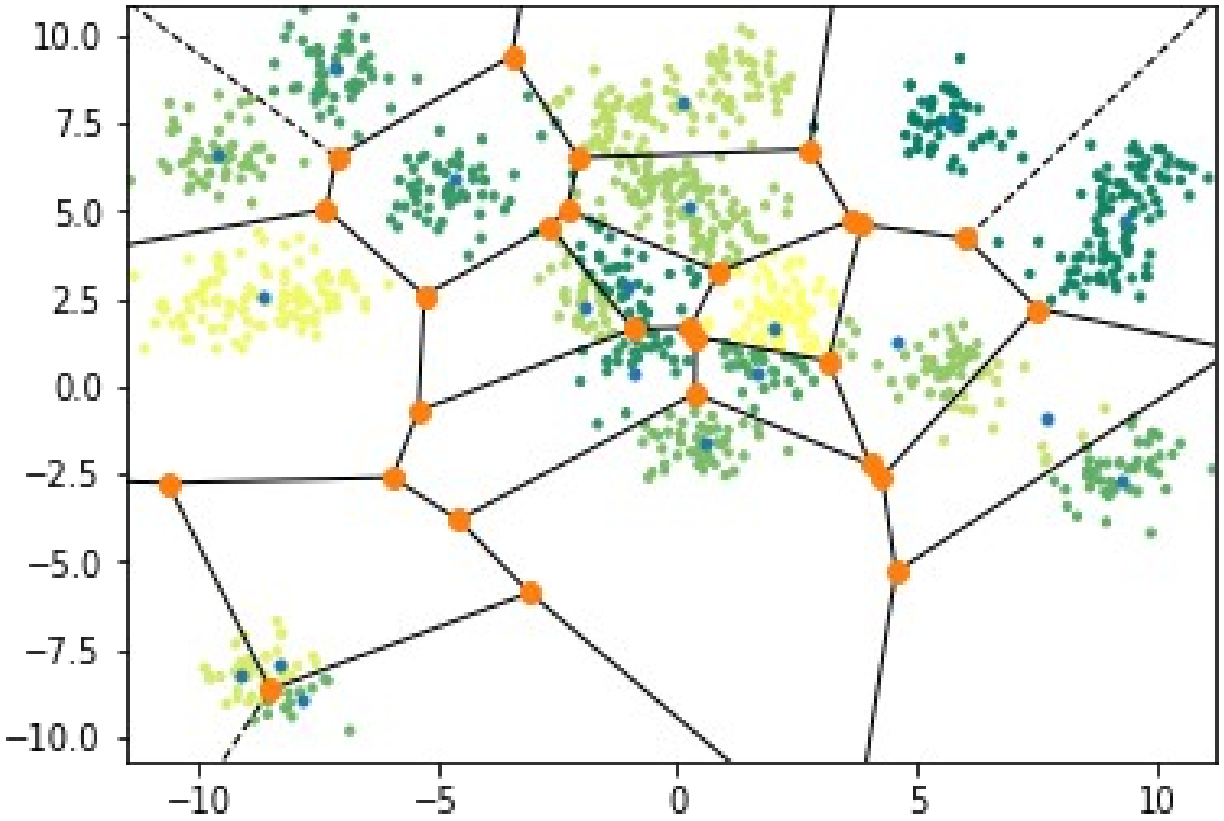
Compositionality



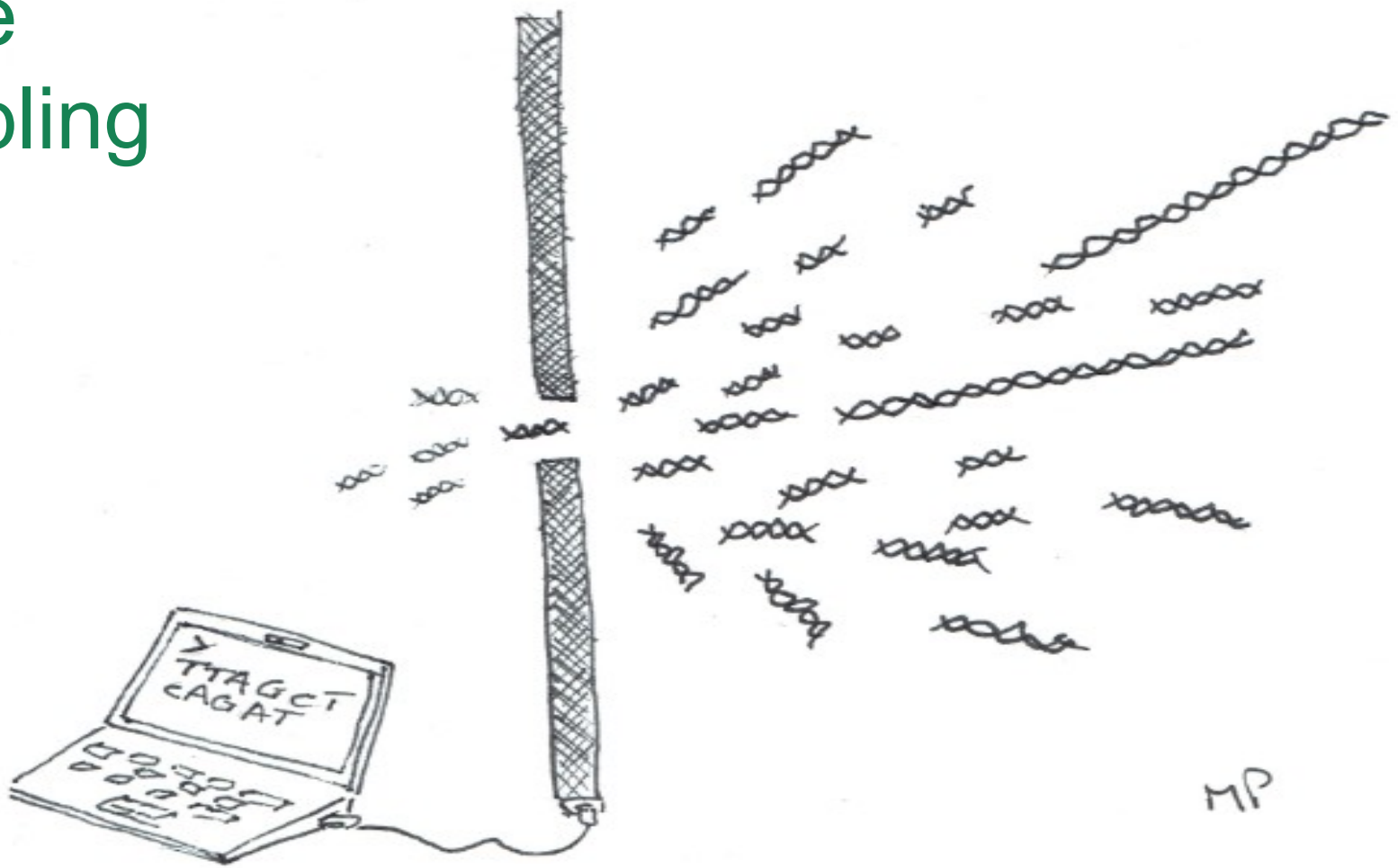
Vandeputte et al. Nature 551:507-511, 2017

Non-parametric clustering:
Voronoi regions

The task:
find *centroids that describe the data*

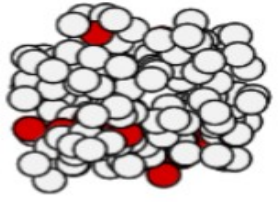


Finite sampling

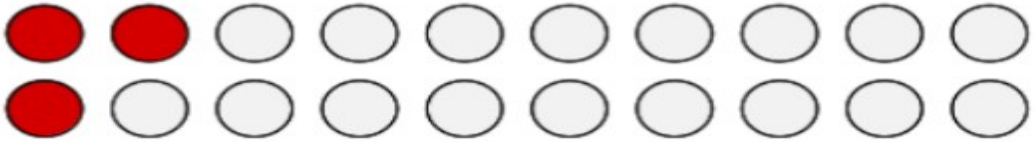


https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf

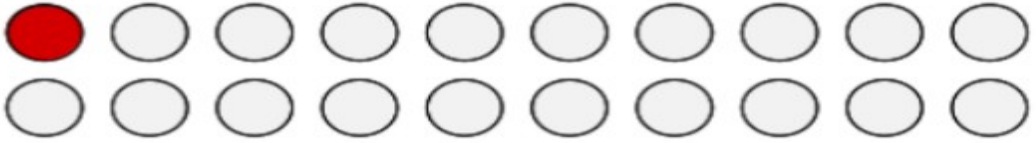
The Poisson distribution



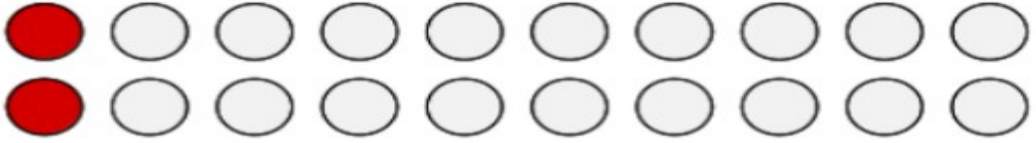
- This bag contains very many small balls, 10% of which are red.
- Several experimenters are tasked with determining the percentage of red balls.
- Each of them is permitted to draw 20 balls out of the bag, without looking.



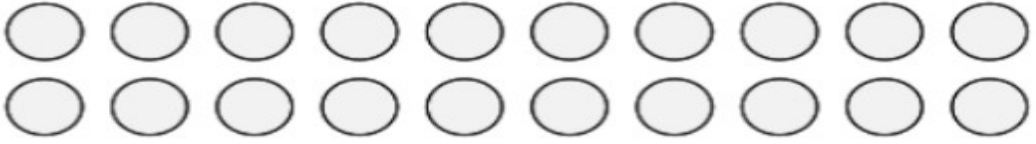
3 / 20 = 15%



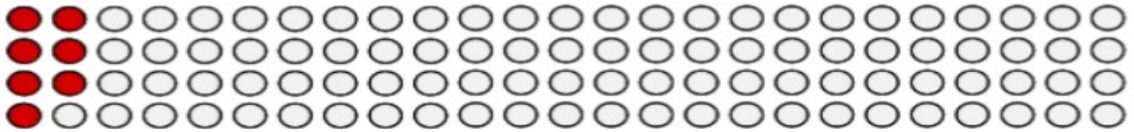
1 / 20 = 5%



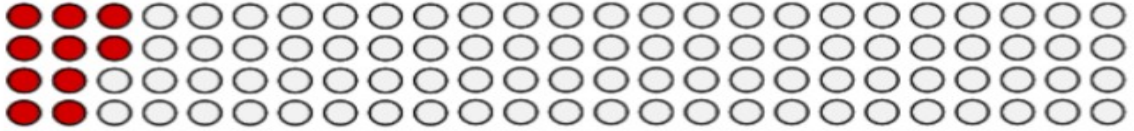
2 / 20 = 10%



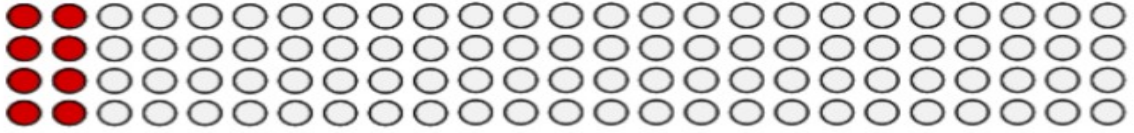
0 / 20 = 0%



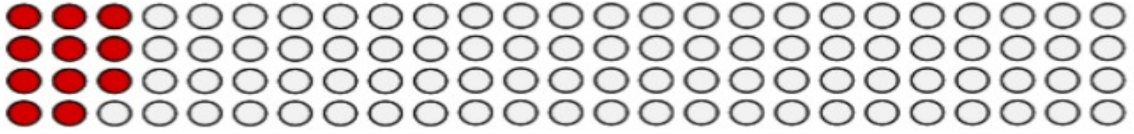
7 / 100 = 7%



10 / 100 = 10%



8 / 100 = 8%



11 / 100 = 11%

Poisson distribution: Counting uncertainty

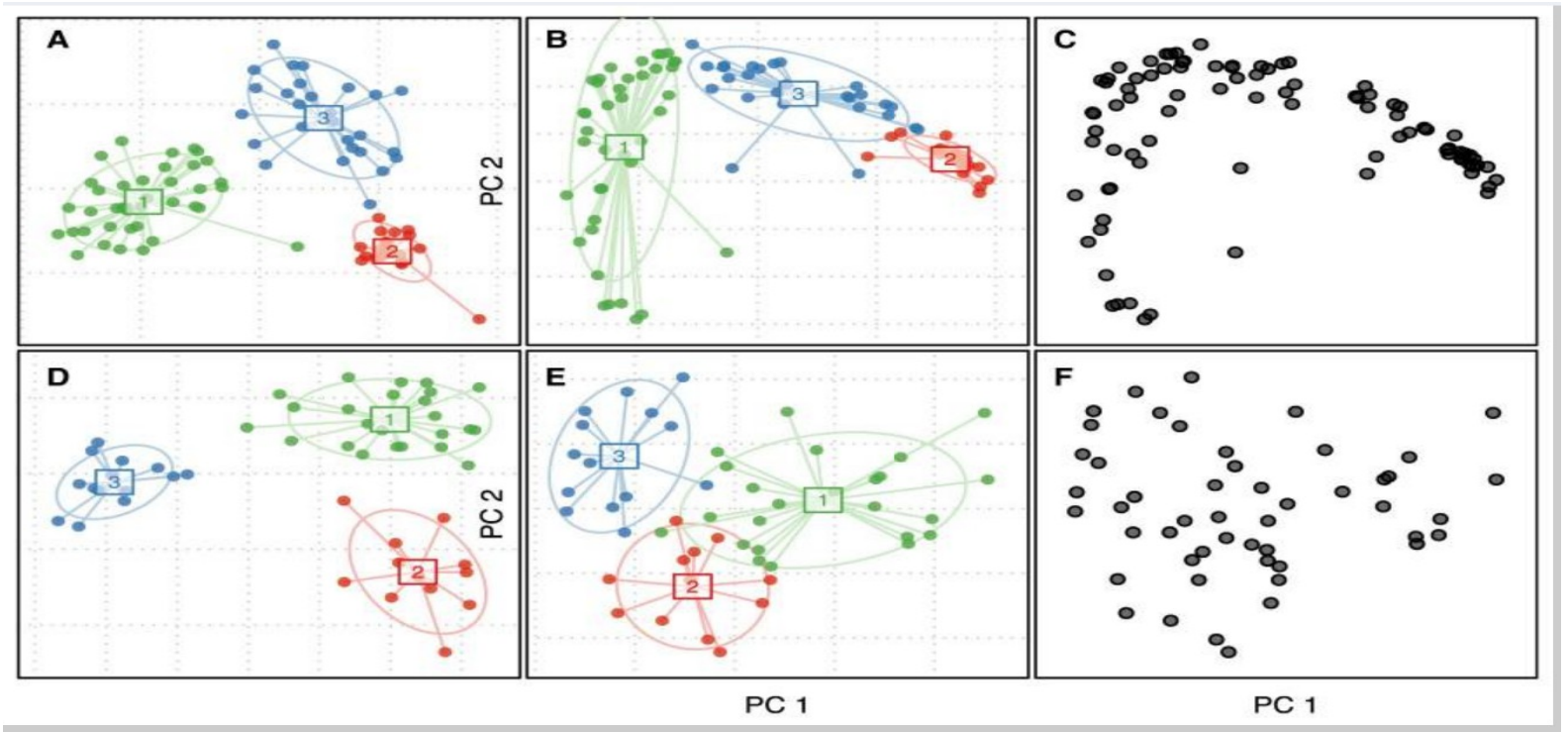
expected number of red balls	standard deviation of number of red balls	relative error in estimate for the fraction of red balls
10	$\sqrt{10} = 3$	$1 / \sqrt{10} = 31.6\%$
100	$\sqrt{100} = 10$	$1 / \sqrt{100} = 10.0\%$
1,000	$\sqrt{1{,}000} = 32$	$1 / \sqrt{1000} = 3.2\%$
10,000	$\sqrt{10{,}000} = 100$	$1 / \sqrt{10000} = 1.0\%$

Distinct clusters or extremes on a continuum?

Common Visualizations Can Support Different Conclusions

Soil samples with varying pH

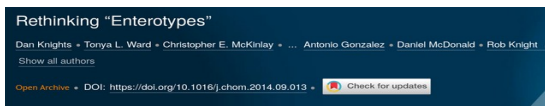
Simulated data with no cluster structure



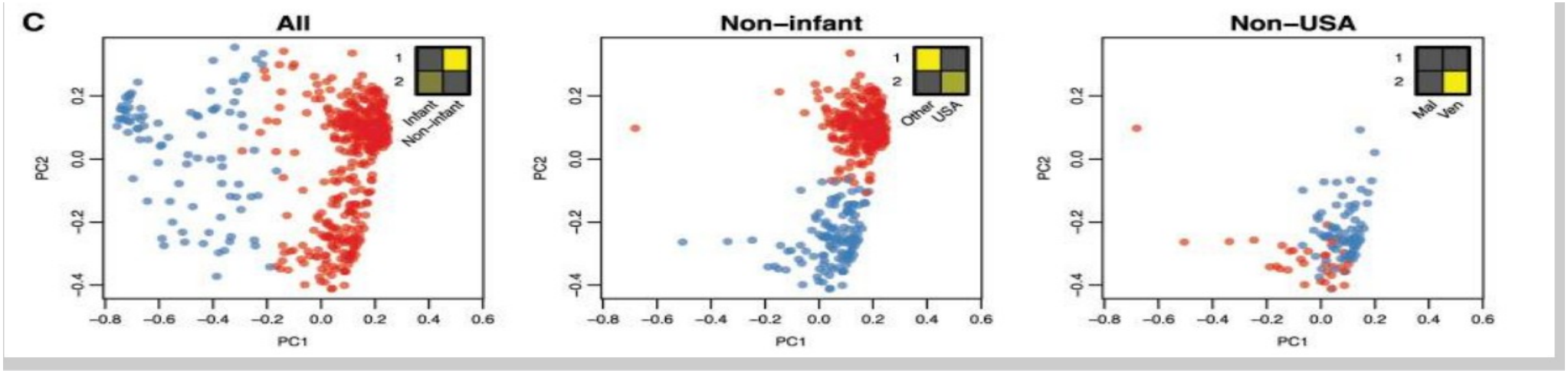
Supervised

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

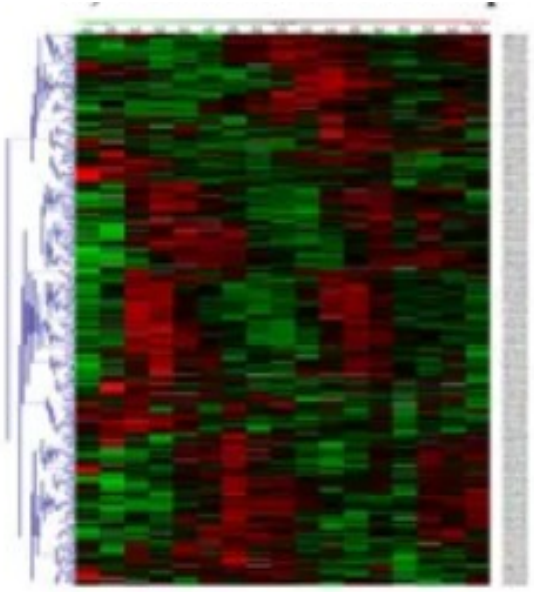
Open Archive • DOI: https://doi.org/10.1016/j.chom.2014.09.013 • Check for updates

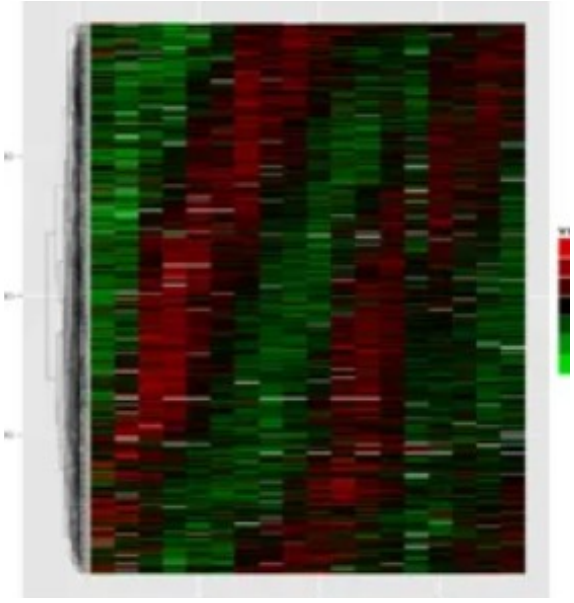
Software | Open Access | Published: 22 January 2010

NeatMap - non-clustering heat map alternatives in R

Satwik Rajaram & Yoshi Oono

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





Typical study designs

Case-control studies

Interventions

Cross-sectional population cohorts

Prospective follow-ups

Longitudinal time series

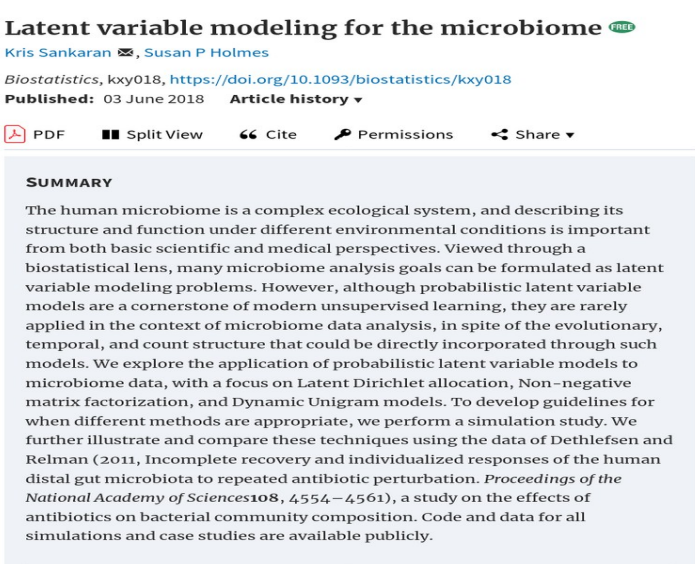
Multi-omics

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.