Day 3: Community analyses

Leo Lahti

Overview of the week

Day 1 Basic data wrangling

Day 2 Assays, transformations & alpha diversity

Day 3 Taxonomic levels & beta diversity

Day 4 Further topics (time series, multi-omics integration)

Overview of Day 2

Time	Theme
9-10	Taxonomic levels, agglomeration, *altExp*
10-11	Beta diversity (community similarity)
11-12	Visualization techniques
12-	Summary, Q & A

Taxonomic levels & agglomeration

Basic data operations

- Extracting rowData, colData, assays
- Subsetting
- Transformations
- **Aggregation / agglomeration**

Aggregation

Aggregate & split by:

- taxonomic groupings (rows)
- sample groupings (cols)

Example

Let's check taxonomic mappings (rowData) for a demo data set.

```
library(mia)
data(GlobalPatterns)
tse <- GlobalPatterns
rowData(tse)[1:3, ] # First 3 rows
```

```
DataFrame with 3 rows and 7 columns
           Kingdom        Phylum        Class        Order        Family
       <character>   <character>  <character>  <character>   <character>
549322     Archaea Crenarchaeota Thermoprotei           NA            NA
522457     Archaea Crenarchaeota Thermoprotei           NA            NA
951        Archaea Crenarchaeota Thermoprotei Sulfolobales Sulfolobaceae
             Genus                 Species
       <character>            <character>
549322          NA                     NA
522457          NA                     NA
951     Sulfolobus Sulfolobusacidocalda..
```

Example

Unique elements per taxonomic level.

```
knitr::kable(
  sapply(rowData(tse), function (x) {length(unique(x))})
)
```

	x
Kingdom	2
Phylum	67
Class	140
Order	205
Family	335
Genus	984
Species	945

Aggregating taxonomic units

Task 18.6.1 / 1-3: Taxonomic levels & aggregation

- List the available taxonomic ranks in the data
- Merge the data to Phylum level
- Report dimensionality before and after aggomeration

Useful functions: taxonomyRanks, agglomerateByRank, mergeRows

Aggregating taxonomic units

18.6.1 / 4-5: Taxonomic levels & aggregation (optional)

- Perform transformations on the data; does this affect agglomeration?
- Calculate alpha diversities before and after aggregation
- List full taxonomic information for some given taxa (Hint: mapTaxonomy)
- Merge rows by other features (mergeRows)

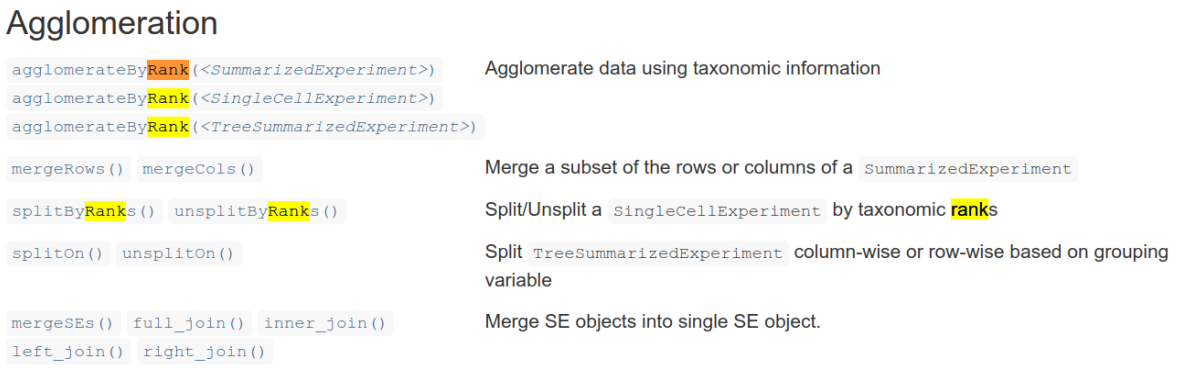
Splitting by samples

Analogous functions are available for sample aggregation and splits for essential use cases.

- Merge samples (columns): `mergeCols()`
- Split TreeSE object by col (or row) groupings: `splitOn()`

mia function reference page

mia homepage



Alternative experiments: *altExp*

Organizing the data

Agglomeration brings the data into a higher level, and changes its dimensionality (feature set size).

-> Can we keep track of the different, alternative levels in the same object?

Taxonomic ranks & *altExp*

The alternative experiments (*altExp*) mechanism allows us to include multiple abundance tables at different taxonomic levels.

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

Demo

Dimension of the original data:

```
dim(tse)
```

```
[1] 19216    26
```

Aggregate to Phylum level and check the new dimension:

```
tse.agg <- agglomerateByRank(tse, rank="Phylum")
dim(tse.agg)
```

```
[1] 67 26
```

Demo

Instead, add the new aggregated abundance table as alternative experiment. Check altExp names before and after.

```
tse <- GlobalPatterns
altExpNames(tse)
```

```
character(0)
```

```
altExp(tse, "Phylum") <- agglomerateByRank(tse, rank="Phylum")
altExpNames(tse)
```

```
[1] "Phylum"
```

We have now added a new alternative experiment in the TreeSE data container.

Taxonomic ranks & *altExp*

We saw how to aggregate data to a higher taxonomic level.

-> How to automate this for all levels?

Alternative experiments (*altExp*)

Task: 18.6.2 (1-2)

- Create taxonomic abundance tables for all different levels at once (splitByRanks)
- Check the available alternative experiment (altExp) names before and after splitByRanks
- Pick specific "experiment" (taxonomic rank) from specific altExp; and then a specific assay

Alternative experiments (*altExp*)

Task: 18.6.2 (3)

- Add a new transformed assay to one of the altExps..

Alternative experiments (*altExp*)

Task: 18.6.2 (4)

Optional:

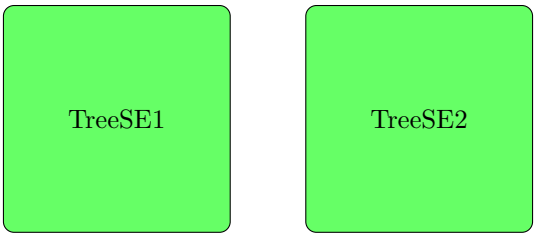
- Split / unsplit the data based on other (row) features: `(un)splitOn`
- Undo the split with `unsplitByRanks`

Aggregating TreeSE objects

We can aggregate TreeSE objects with different dimensionalities.

- mergeSEs()
- full_join()
- inner_join()
- left_join()
- right_join()

```
\usetikzlibrary{shapes.geometric, arrows}
\tikzstyle{greenbox} = [rectangle, rounded corners, minimum width=3cm, minimum height=3cm,
\begin{tikzpicture}[node distance=2cm]
\node (leftb) [greenbox] {TreeSE1};
\node (rightb) [greenbox, right of=leftb, xshift=2cm, yshift=0cm] {TreeSE2};
\end{tikzpicture}
```



mia function reference page

mia homepage

Agglomeration

`agglomerateByRank(<SummarizedExperiment>)` `agglomerateByRank(<SingleCellExperiment>)` `agglomerateByRank(<TreeSummarizedExperiment>)`	Agglomerate data using taxonomic information
`mergeRows()` `mergeCols()`	Merge a subset of the rows or columns of a `SummarizedExperiment`
`splitByRanks()` `unsplitByRanks()`	Split/Unsplit a `SingleCellExperiment` by taxonomic ranks
`splitOn()` `unsplitOn()`	Split `TreeSummarizedExperiment` column-wise or row-wise based on grouping variable
`mergeSEs()` `full_join()` `inner_join()` `left_join()` `right_join()`	Merge SE objects into single SE object.

Beta diversity (community similarity)

Key sources of microbial ecosystem variation

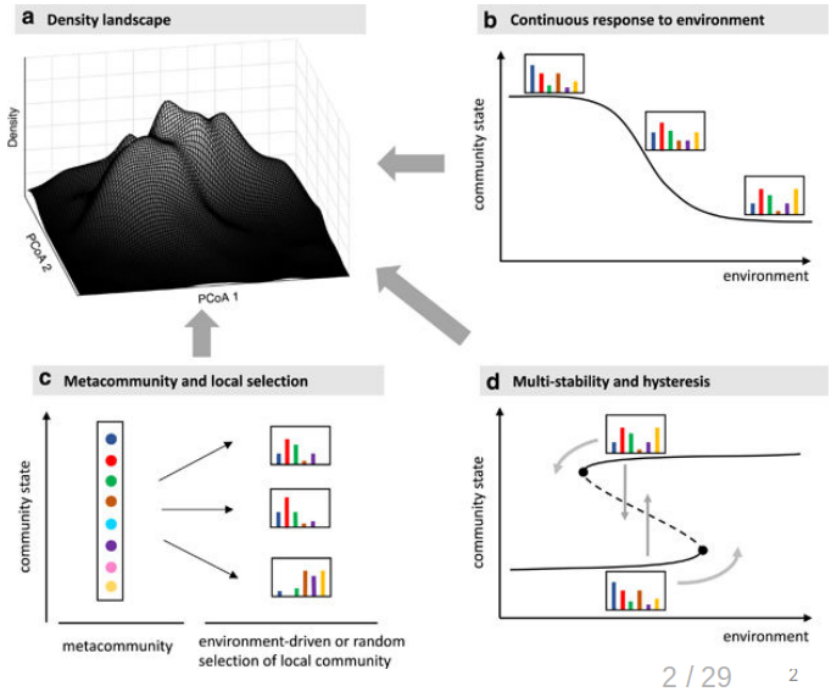
External perturbations (push & pulse)

Internal dynamics and multi-stability

Immigration

Stochasticity

Memory



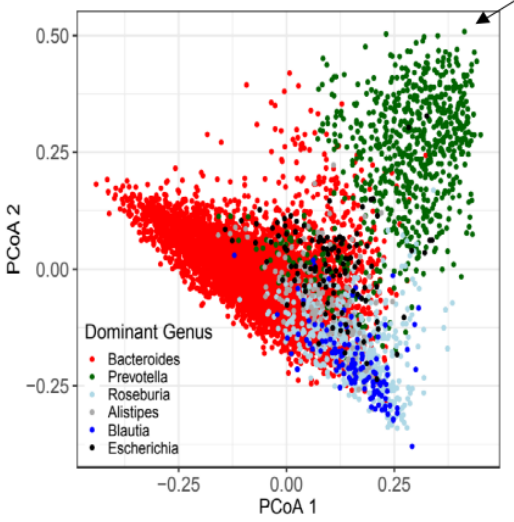
Microbial communities as dynamical systems

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

2 / 29 2

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

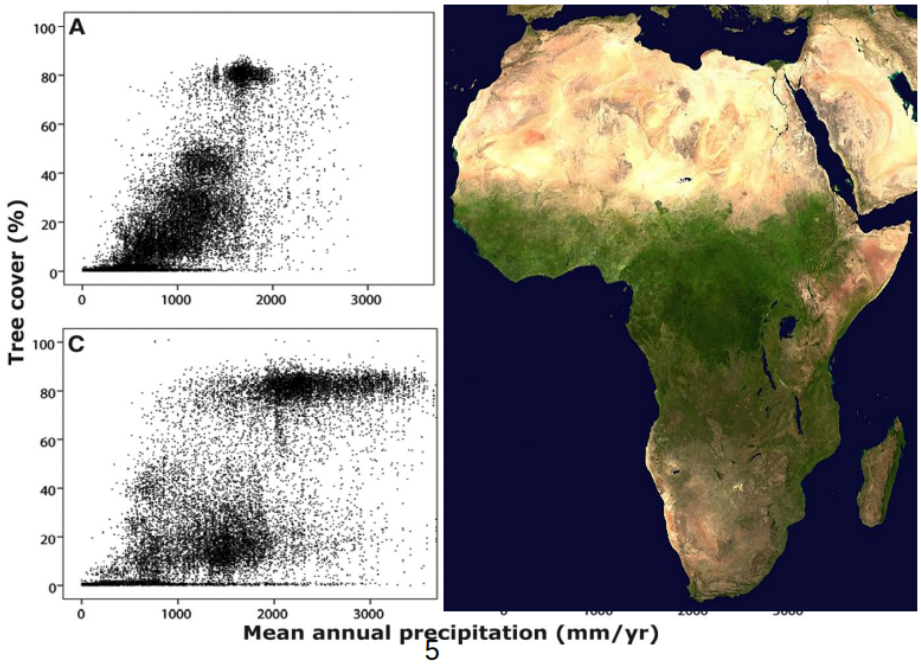
Science 14 October 2011:
Vol. 334 no. 6053 pp. 232-235
DOI: 10.1126/science.1210657

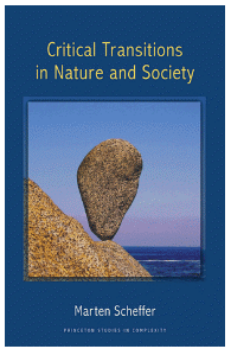
< Prev | Table of Contents | Next >

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]

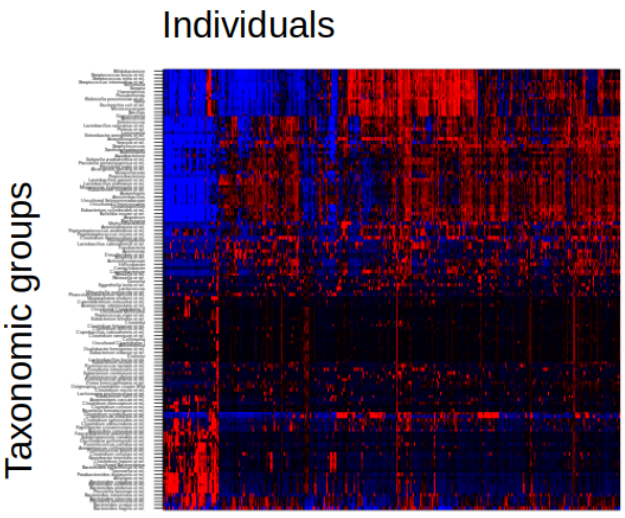




5 / 29

5

Taxonomic abundance table



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

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

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.

3 / 29

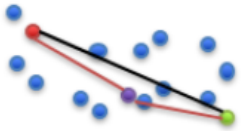
Standard (dis)similarity indices

Phylogenetically neutral beta diversity measures

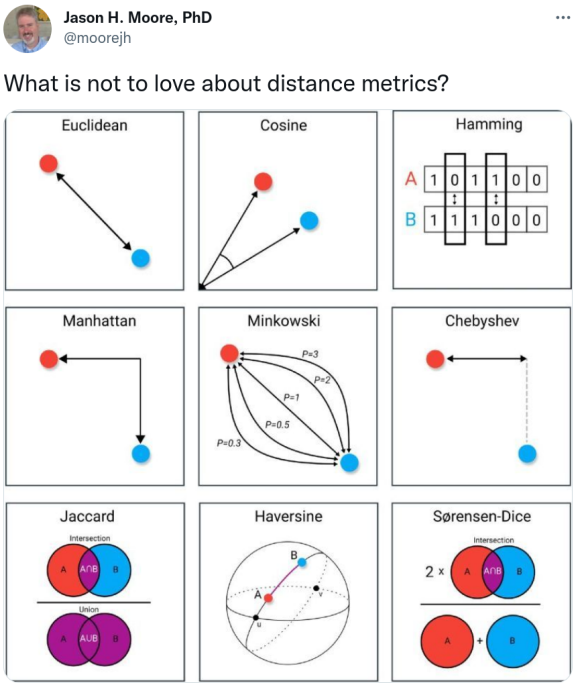
- Euclidean
- Correlation
- Bray-Curtis
- Aitchison (CLR transformation + Euclidean distance)
- Jaccard
- Jensen-Shannon (JS)

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

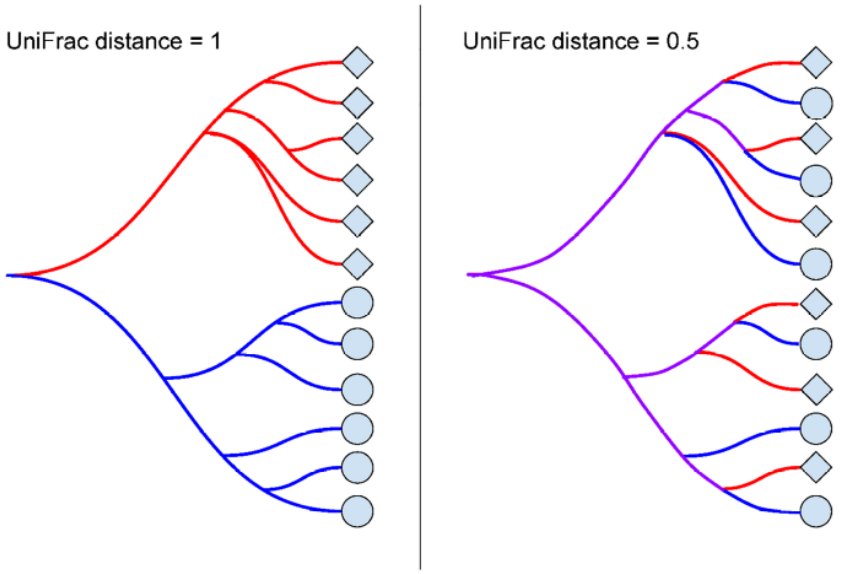


Standard (dis)similarity indices

Phylogenetic beta diversity measures

- Unifrac (weighted)
- Unifrac (unweighted)

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



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

Fundamental considerations in beta diversity analysis

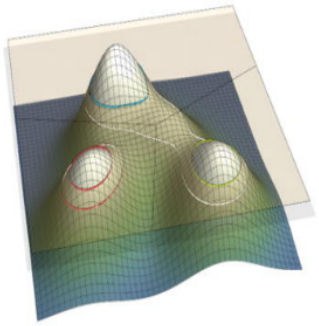
Feature selection (all/core taxa; genus/strain level..?)

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

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

Ordination (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

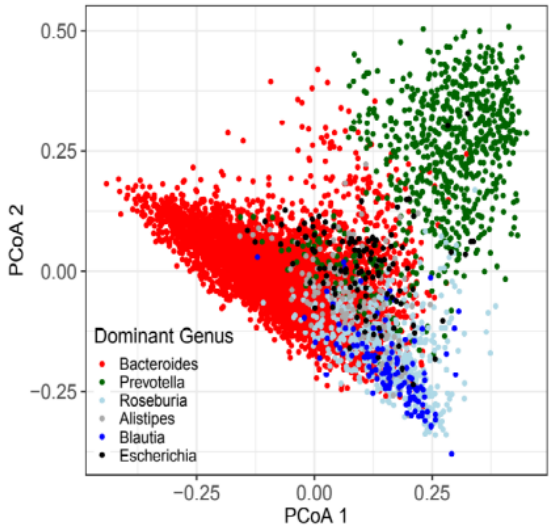
Clustering method
- Hierarchical / Ward
- Hierarchical / Complete
- Gaussian mixture
- DMM
- PAMR
- K-means

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

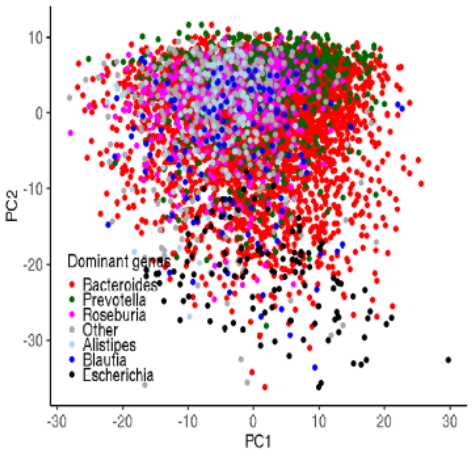
PCoA + Bray-Curtis

Preserves distances



PCA + Aitchison

Captures largest variation



Comment on this paper

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

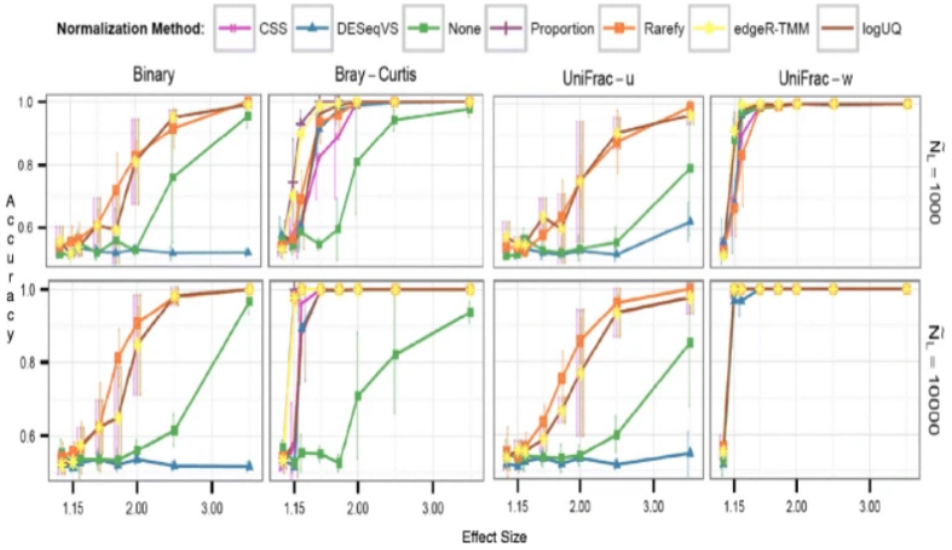
Aaro Salosensaari, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa,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

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

Fig. 2

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

Dissimilarity measure and normalization affect clustering accuracy



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.

Beta diversity task

OMA, Chapter 18

Task 18.9.1: Multivariate ordination

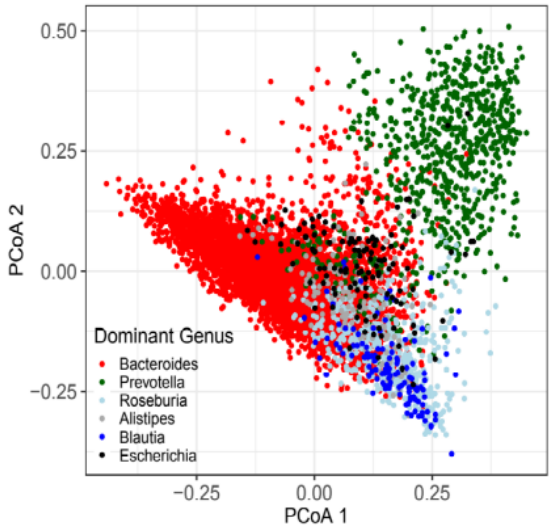
1. Load experimental dataset from mia.
2. Create PCoA with Bray-Curtis dissimilarities
3. Create PCA with Aitchison dissimilarities
4. Visualize and compare both
5. Test other feature sets, transformations, altExps, dissimilarities, and ordination methods
6. Color the dominant taxonomic groups?

Useful functions: runMDS, runNMDS, transformSamples, ggplot, plotReducedDim

Also check function reference pages for the **mia** and **miaViz** homepage.

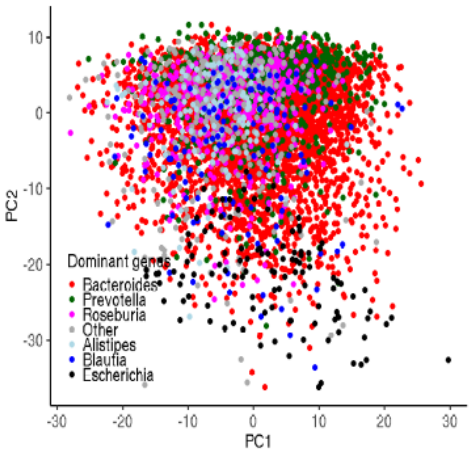
PCoA + Bray-Curtis

Preserves distances



PCA + Aitchison

Captures largest variation

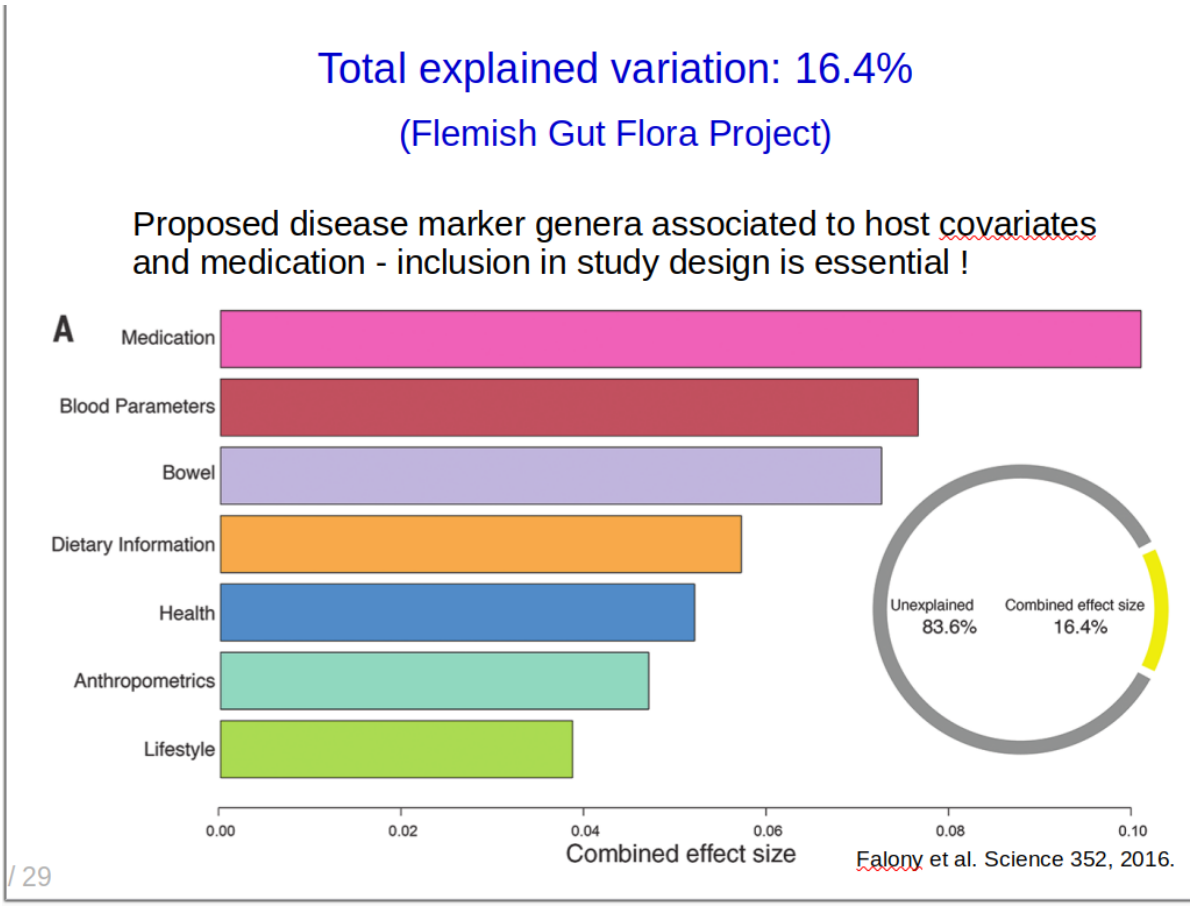


Comment on this paper

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



Statistical community comparisons

- PERMANOVA
- dbRDA

OMA, Chapter 18

- Task: 18.8.1 Beta diversity basics
- Task: 18.8.2 Beta diversity extra

Clustering microbiome samples

Dirichlet Multinomial Mixtures (DMM)

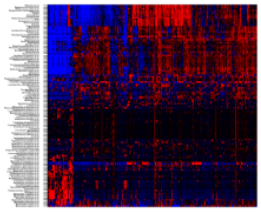
Perspective | Published: 18 December 2017

Enterotypes in the landscape of gut microbial community composition

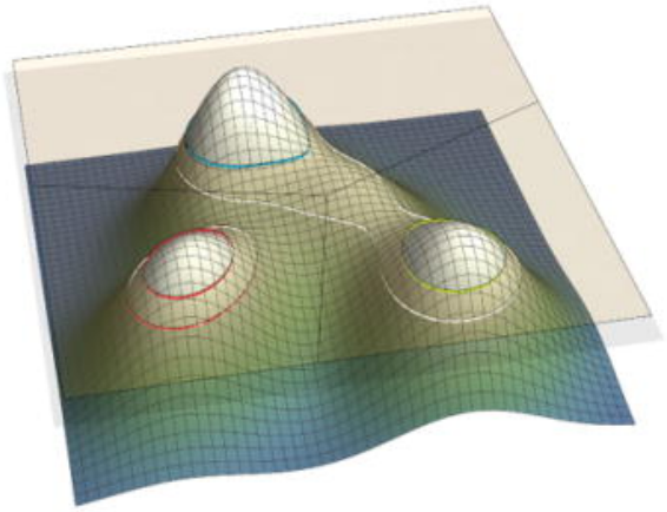
Paul I. Costea, Falk Hildebrand, [...] Peer Bork

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

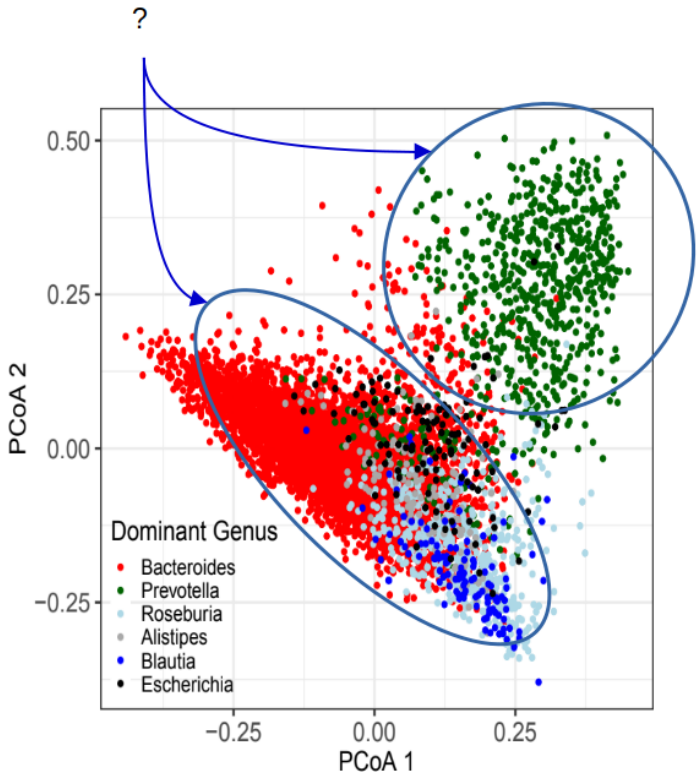
6840 Accesses | 253 Citations | 100 Altmetric | Metrics



Community typing



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

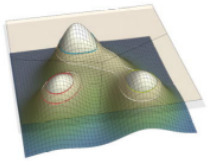


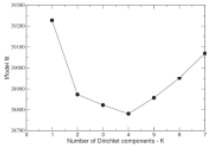
Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics

Ian Holmes, Keith Harris, Christopher Quince

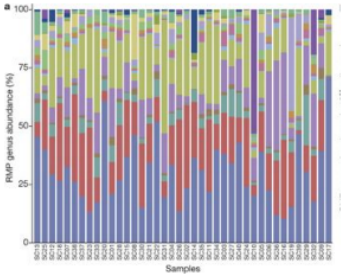
Published: February 3, 2012 • https://doi.org/10.1371/journal.pone.0030126

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





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



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

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

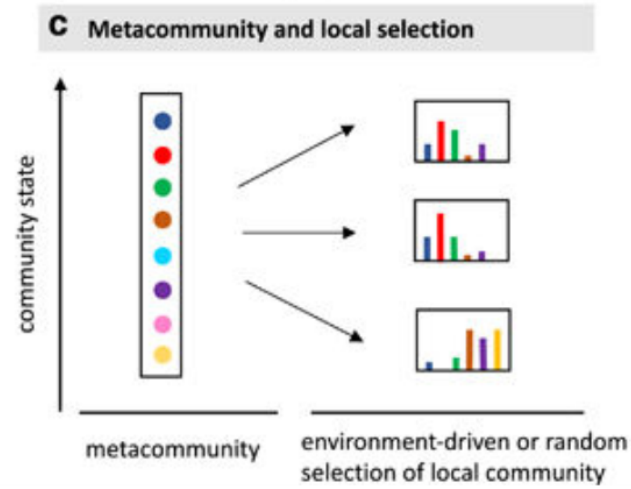
Dirichlet

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,*]

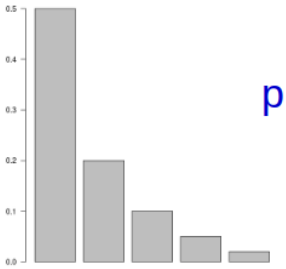
24 / 29

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

Sampling from the multinomial

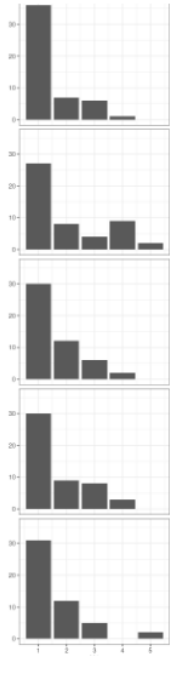
Stochastic (observed) realizations x (50 reads)

"Real" latent distribution p (of species abundances)

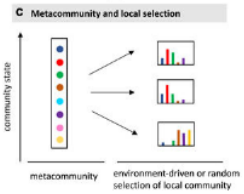


p → x

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

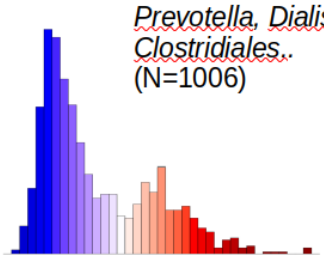


```
ggplot(melt(
  rmultinom(n = 5, size = 50,
      prob = c(0.5, 0.2, 0.1, 0.05, 0.02))),
  aes(x = X1, y = value)) +
  geom_col() + facet_grid(X2 ~ 1)
```

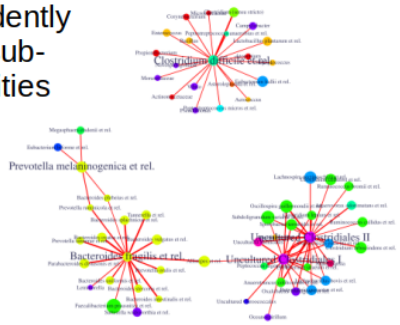


Mixture models bring flexibility in modeling

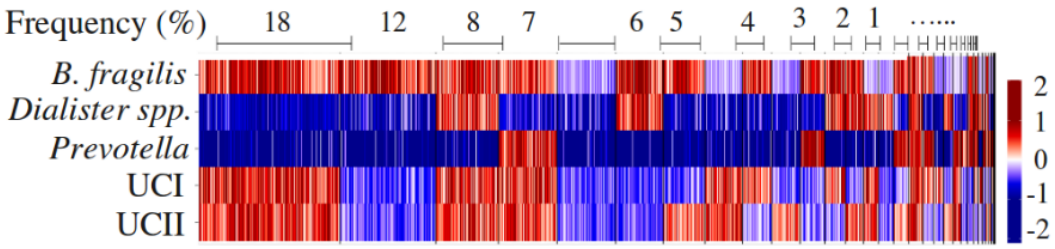
Bistable taxa:
Prevotella, *Dialister*, *Clostridiales*..
(N=1006)



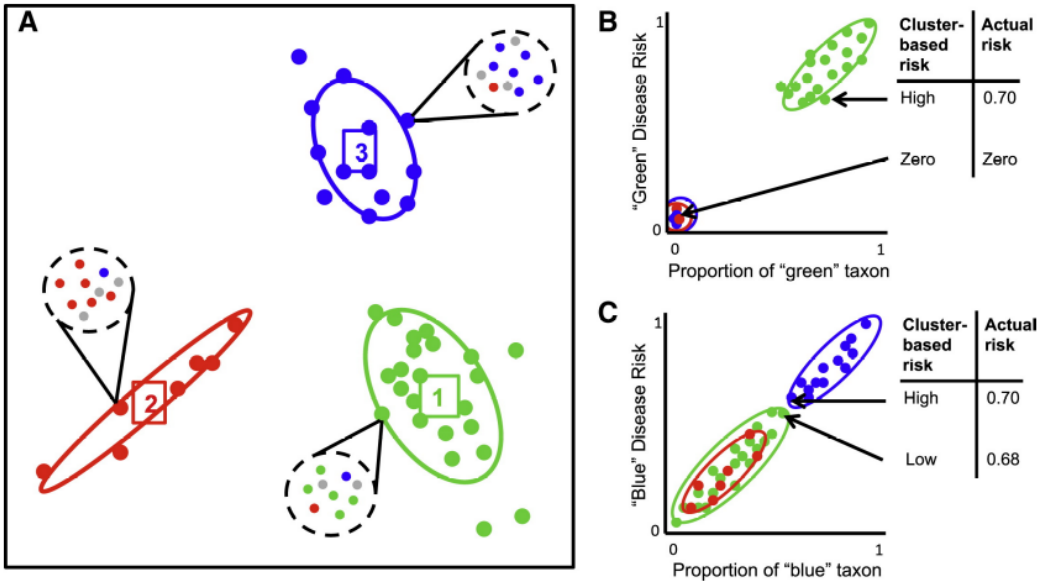
Independently varying sub-communities



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



Clustering 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

/ 29