A unified approach to unconstrained and constrained ordination of microbiome read count data

Stijn Hawinkel, Luc Bijnens, Frederiek-Maarten Kerckhof and Olivier Thas

June 18, 2019

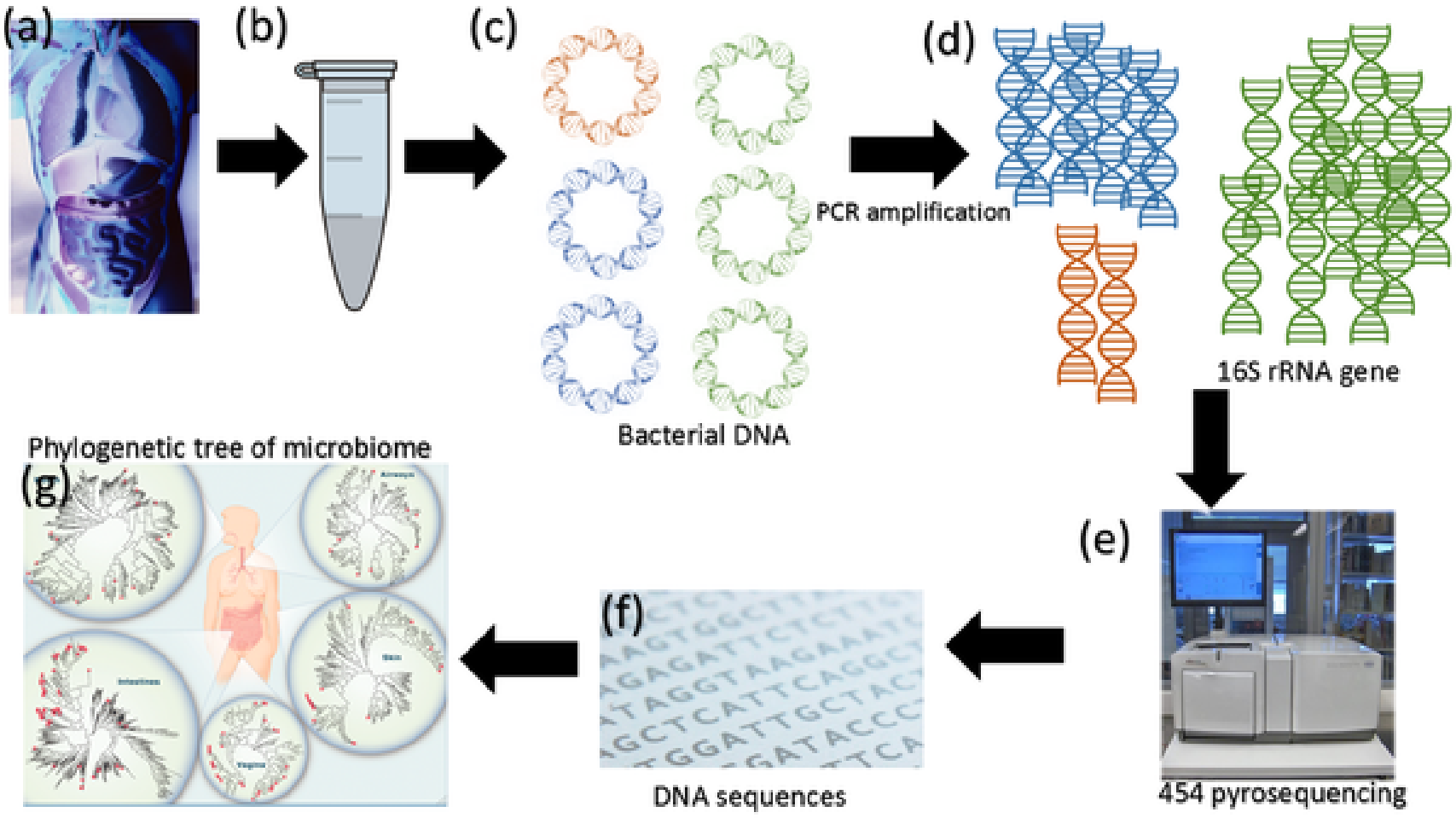
GHENT UNIVERSITY

2019 NCB Nonclinical Biostatistics Conference
RUTGERS
June 17-19, 2019

janssen
PHARMACEUTICAL COMPANIES OF Johnson&Johnson

The human microbiome

- The collection of **micro-organisms** living in and on our body
- Crucial to human **health and disease**

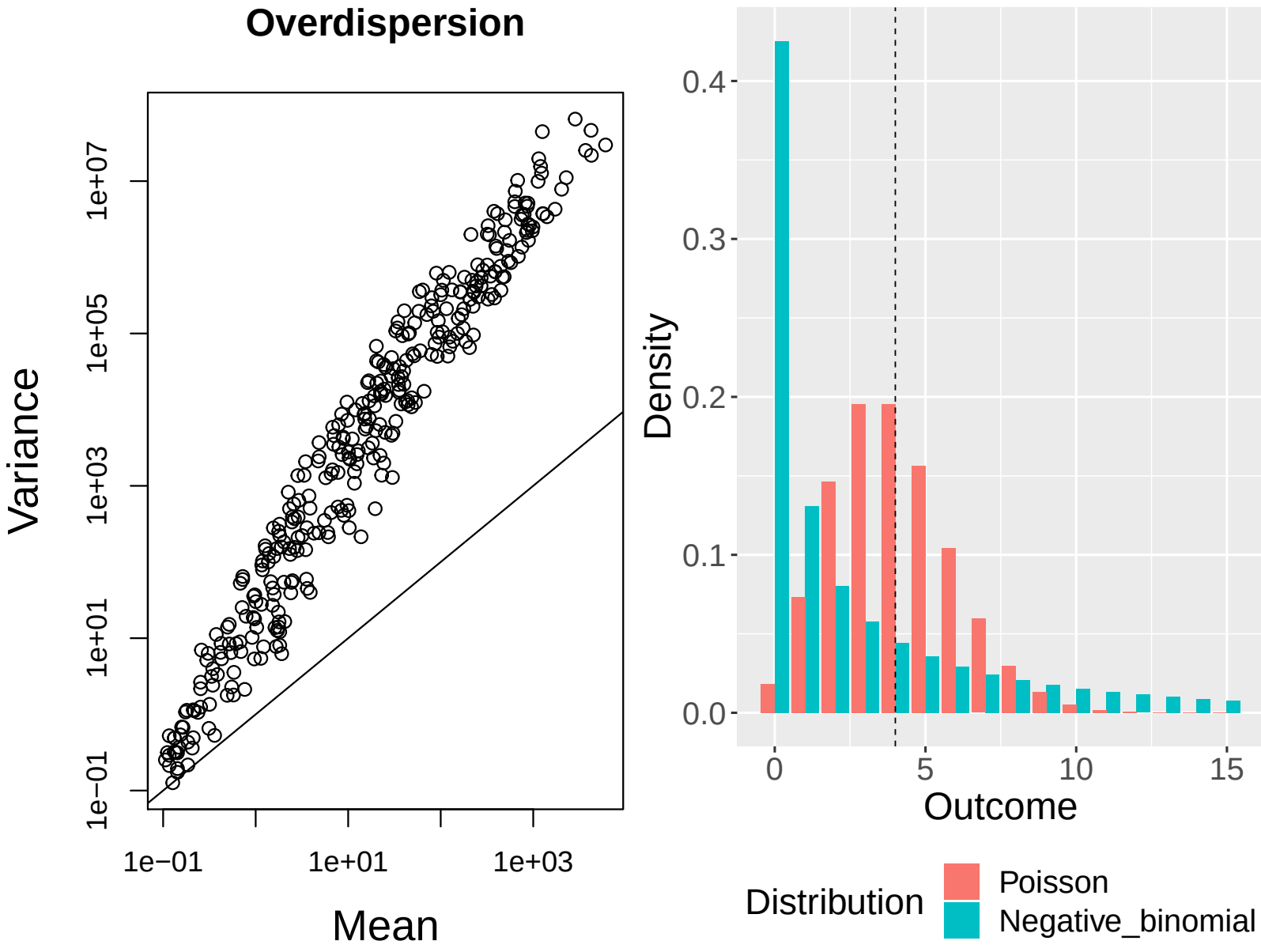


Microbiome data structure

Samples	species 1	...	species p	Library size	Covariates
sample 1	x_{11}	...	x_{1p}	$\sum_{j=1}^{p} x_{1j}$	$\mathbf{C_1}$
sample 2	x_{21}	...	x_{2p}	$\sum_{j=1}^{p} x_{2j}$	$\mathbf{C_2}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
sample n	x_{n1}	...	x_{np}	$\sum_{j=1}^{p} x_{nj}$	$\mathbf{C_n}$

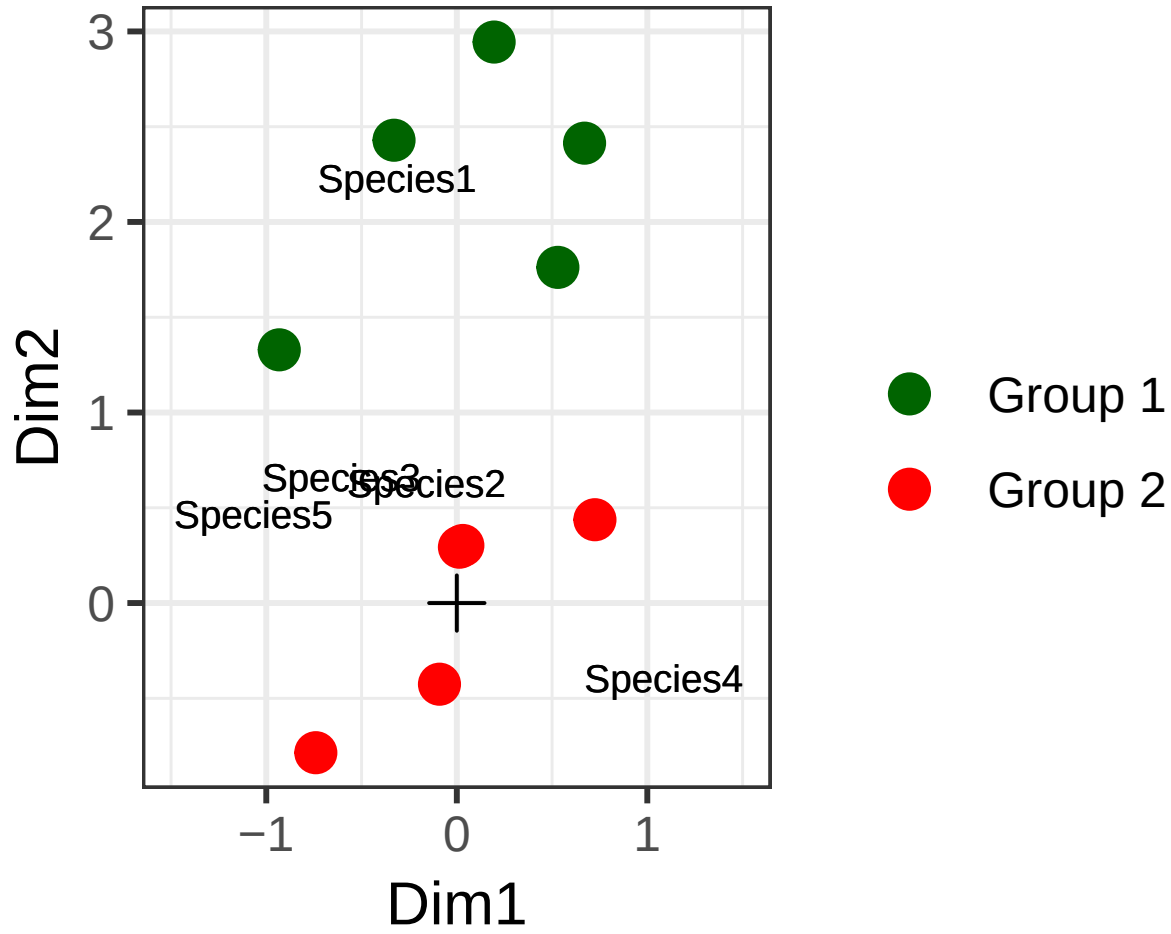
- ~75% zeroes
- Varying library sizes: total number of counts per sample
- Example datasets:
  - **Anterior nares** (nasal cavity) of healthy humans
  - Observational study of **colorectal cancer** patients and controls

Microbiome count data



Explorative visualization of high-dimensional datasets

- Requires a **dimension reduction**
  - **Biplots**: show species and samples in the same plot



- **Triplots**: add patient covariates

Outline

- Existing methods
  - Principal coordinates analysis (PCoA)
  - Compositional data analysis (CoDa)
- RC(M)-model
  - Unconstrained
  - Constrained: add patient covariates

Principal coordinates analysis (PCoA)

- Calculate **ecological distances** between all sample pairs
  - e.g. **Bray-Curtis** dissimilarities, **UniFrac** distance

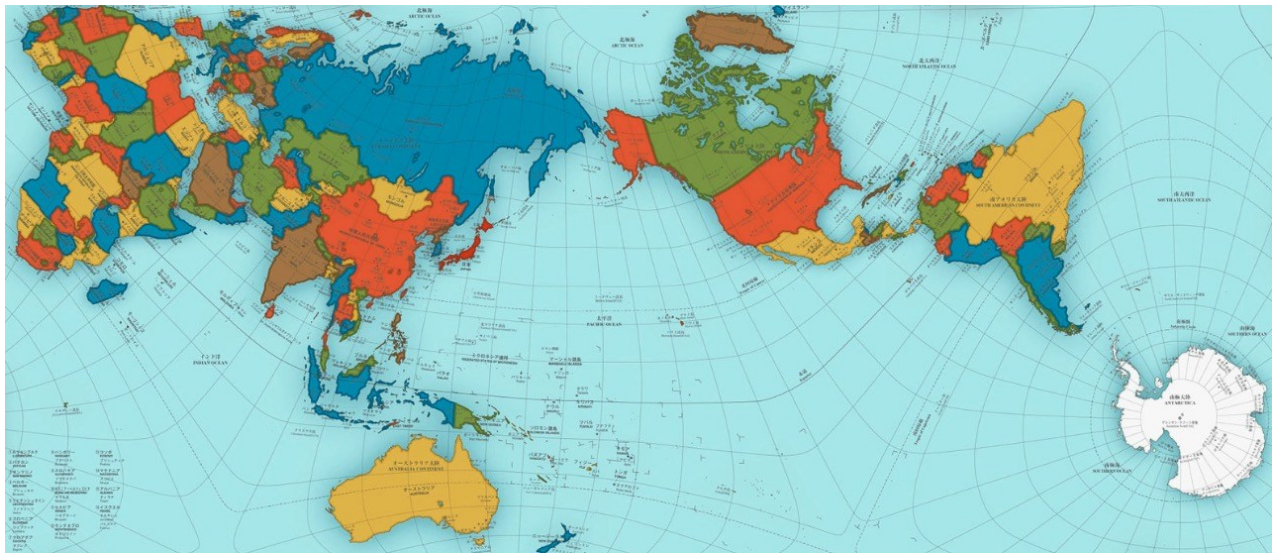
$$BC_{ij} = 1 - \frac{2S_{ij}}{S_i + S_j}$$

- $\mathbf{X}_{n \times p} \rightarrow \mathbf{D}_{n \times n}$
- *Contribution* of different species is lost

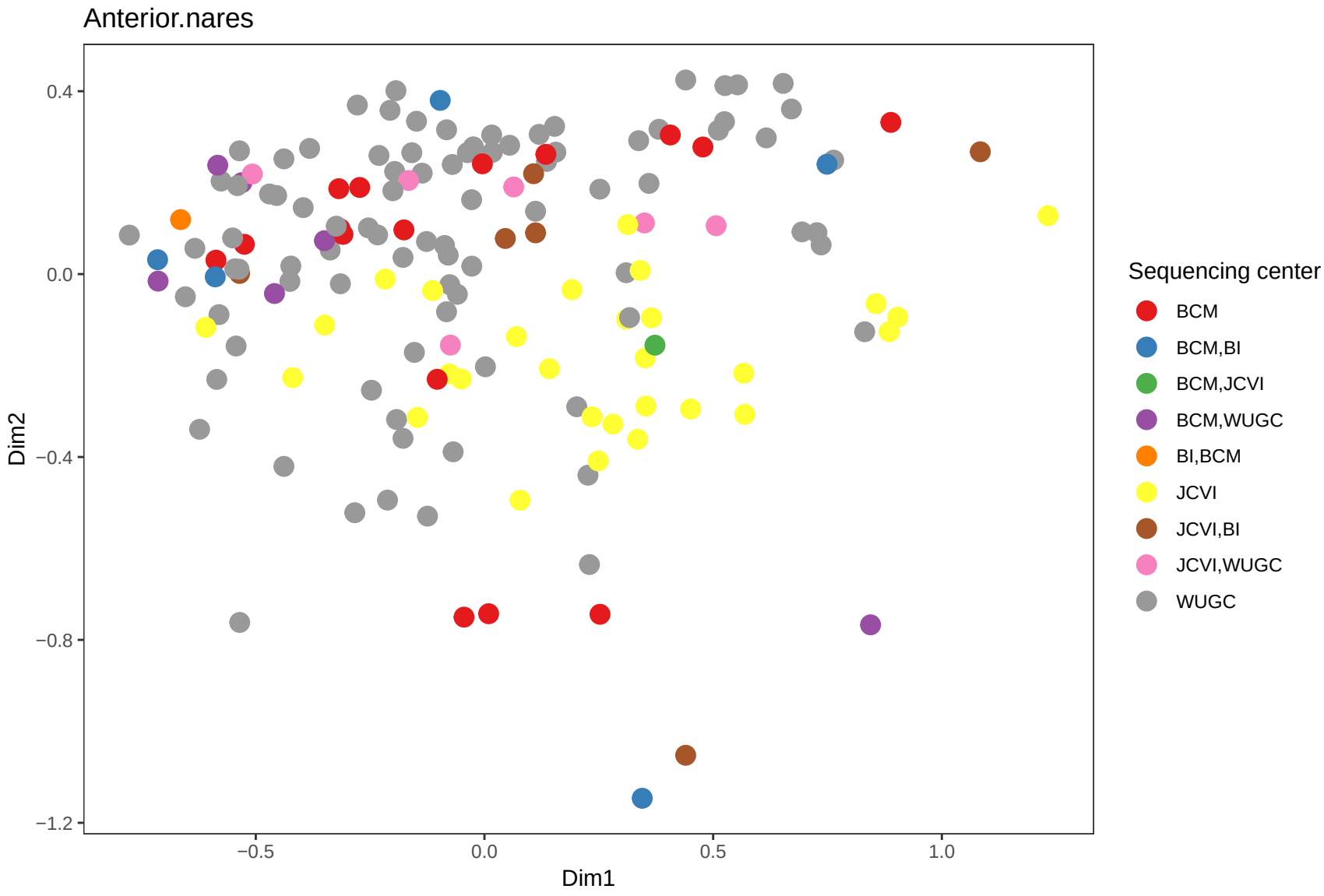
	A	B	C	D	E	F
A	0	16	47	72	77	79
B	16	0	37	57	65	66
C	47	37	0	40	30	35
D	72	57	40	0	31	23
E	77	65	30	31	0	10
F	79	66	35	23	10	0

Principal coordinates analysis (PCoA)

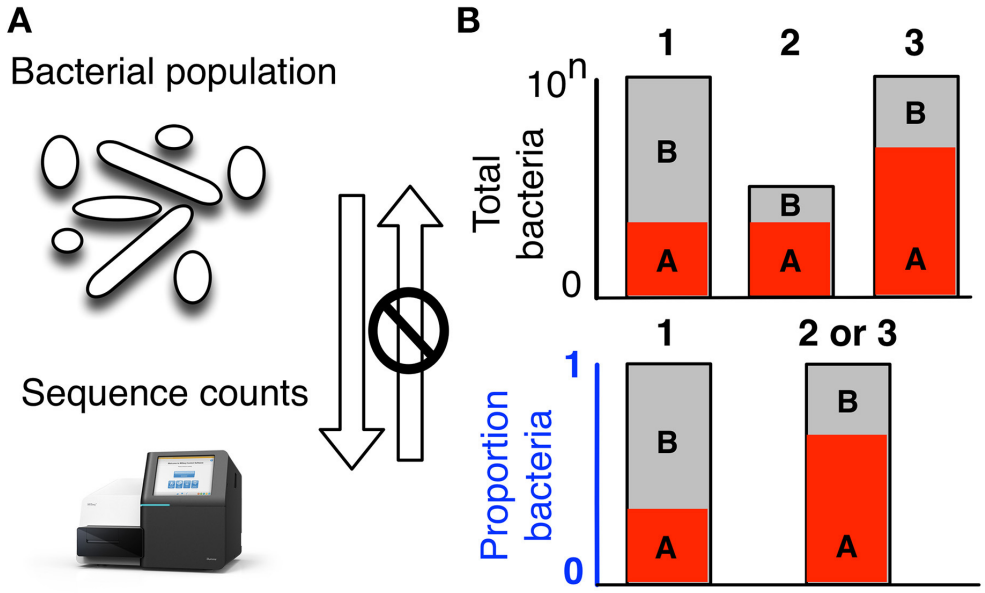
- Apply eigendecomposition to optimally to represent these distances **in 2D**



PCoA example

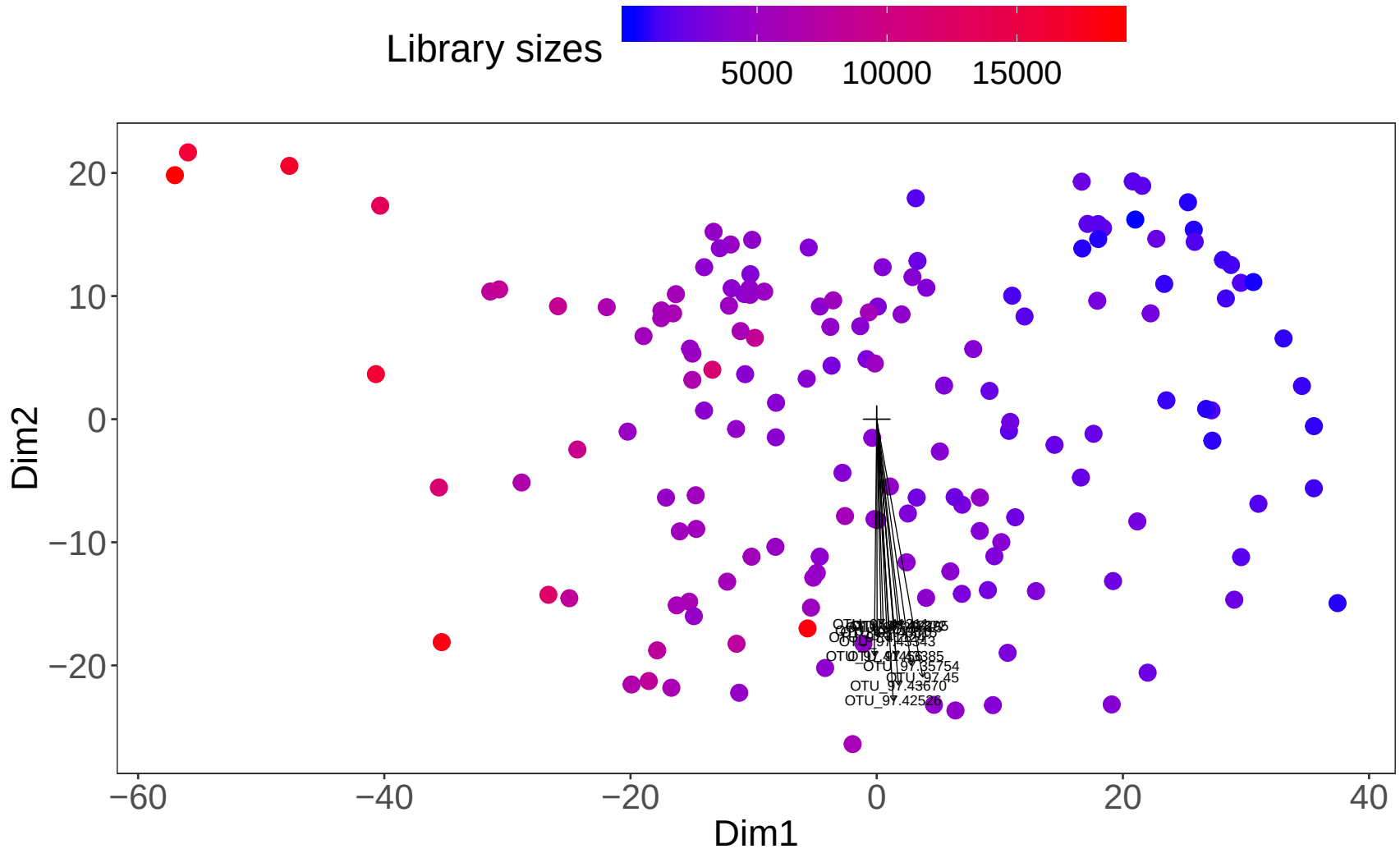


Compositional data



- Account for compositionality by working with **log-ratios**
  - Addition of *pseudocounts* needed because of the many **zeroes**
  - Ratios discard information on **variance**

Compositional data biplot



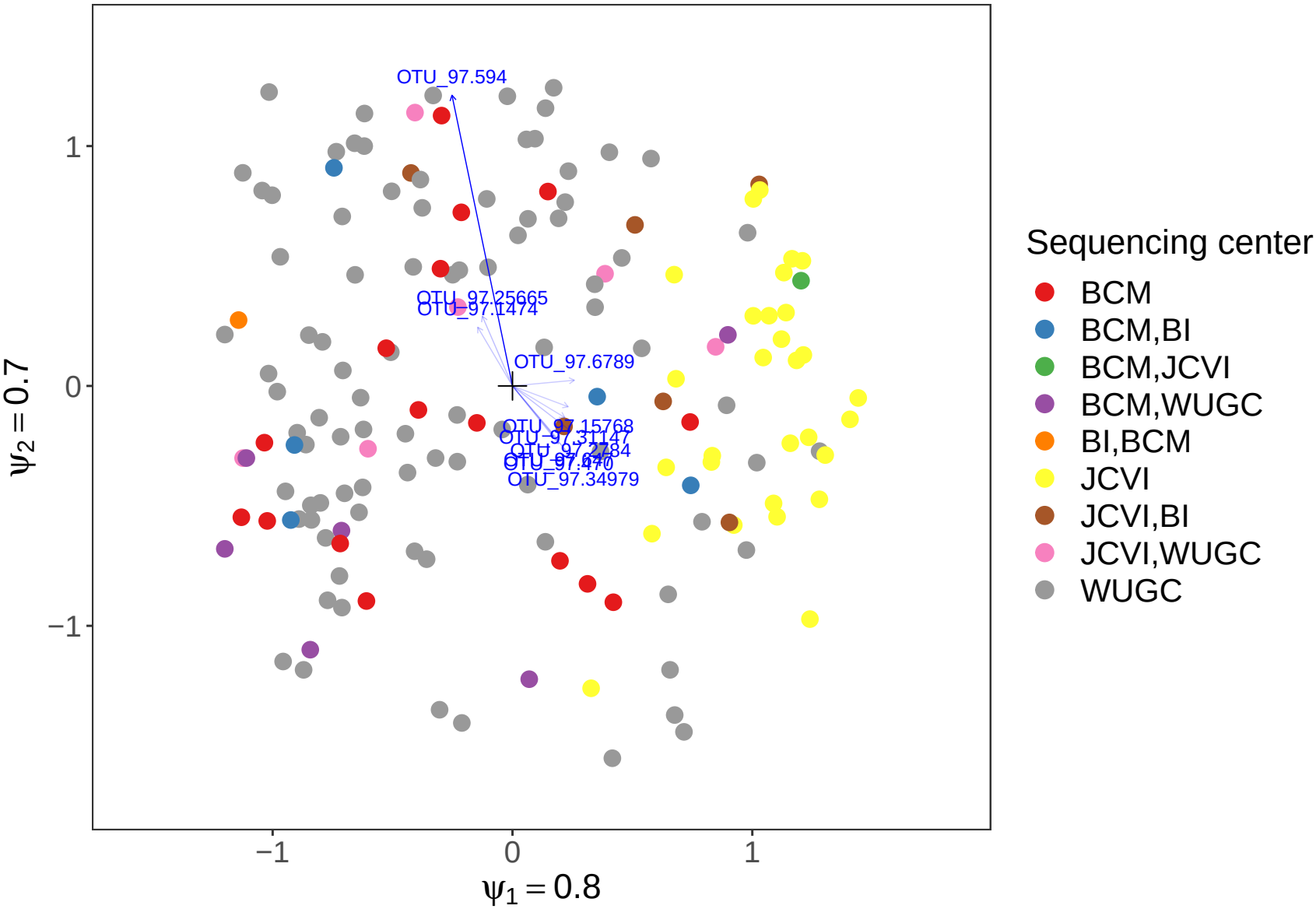
RC(M) model

- Signal in the data = Departure from independence

$$\log\left(E(X_{ij})\right) = \overbrace{u_i + u_j}^{\text{Independence model}} + \sum_{k=1}^{K} \psi_k r_{ik} s_{jk}$$

- r_{ik}: sample scores, s_{jk}: species scores
- ψ_k: **strength** of the departure
- Augment with **any** error distribution!
  - **Negative binomial** captures skewness and overdispersion

RC(M) plot

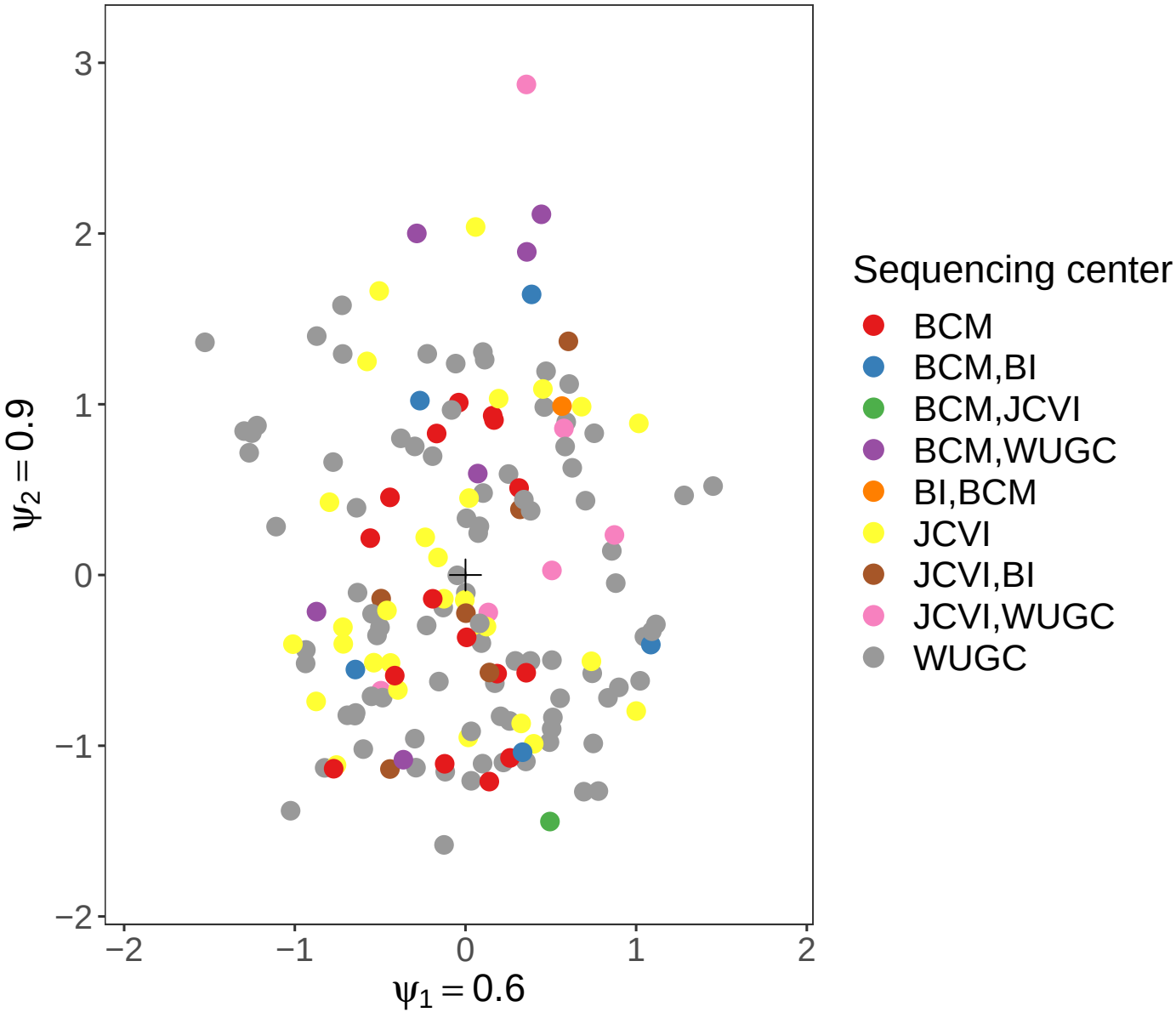

OTU_97.594
OTU_97.25665
OTU_97.1474
OTU_97.6789
OTU_97.15768
OTU_97.31147
OTU_97.27284
OTU_97.470
OTU_97.34979
1
0
−1
−1
0
1
$\psi_2 = 0.7$
$\psi_1 = 0.8$
Sequencing center
BCM
BCM,BI
BCM,JCVI
BCM,WUGC
BI,BCM
JCVI
JCVI,BI
JCVI,WUGC
WUGC

Conditioning: Looking past the obvious

- Biologically insignificant variables affect the sequencing data
  - E.g. sequencing center or technology
- Condition on confounder matrix **G**

$$\log\left(E(X_{ij})\right) = \underbrace{\overbrace{u_i + u_j}^{\text{Independence model}} + \sum_{l=1}^{L} \zeta_{jl} g_{il}}_{\text{Extended null model}} + \overbrace{\sum_{k=1}^{K} \psi_k r_{ik} s_{jk}}^{\text{Biological signal}}$$

Conditioning: Looking past the obvious



Incorporating covariates

$$\log\Big(E(X_{ij})\Big) = u_i + u_j + \sum_{k=1}^{K} \psi_k f_j(\boldsymbol{\alpha}_k^t \mathbf{C}_i)$$

- $\boldsymbol{\alpha}_k^t$ an **environmental gradient**: reveals which variables shape the environment
- $\boldsymbol{\alpha}_k^t \mathbf{C}_i = r_{ik}$ the **environmental score**: a linear combination of environmental variables
- f_j a species specific **response function**

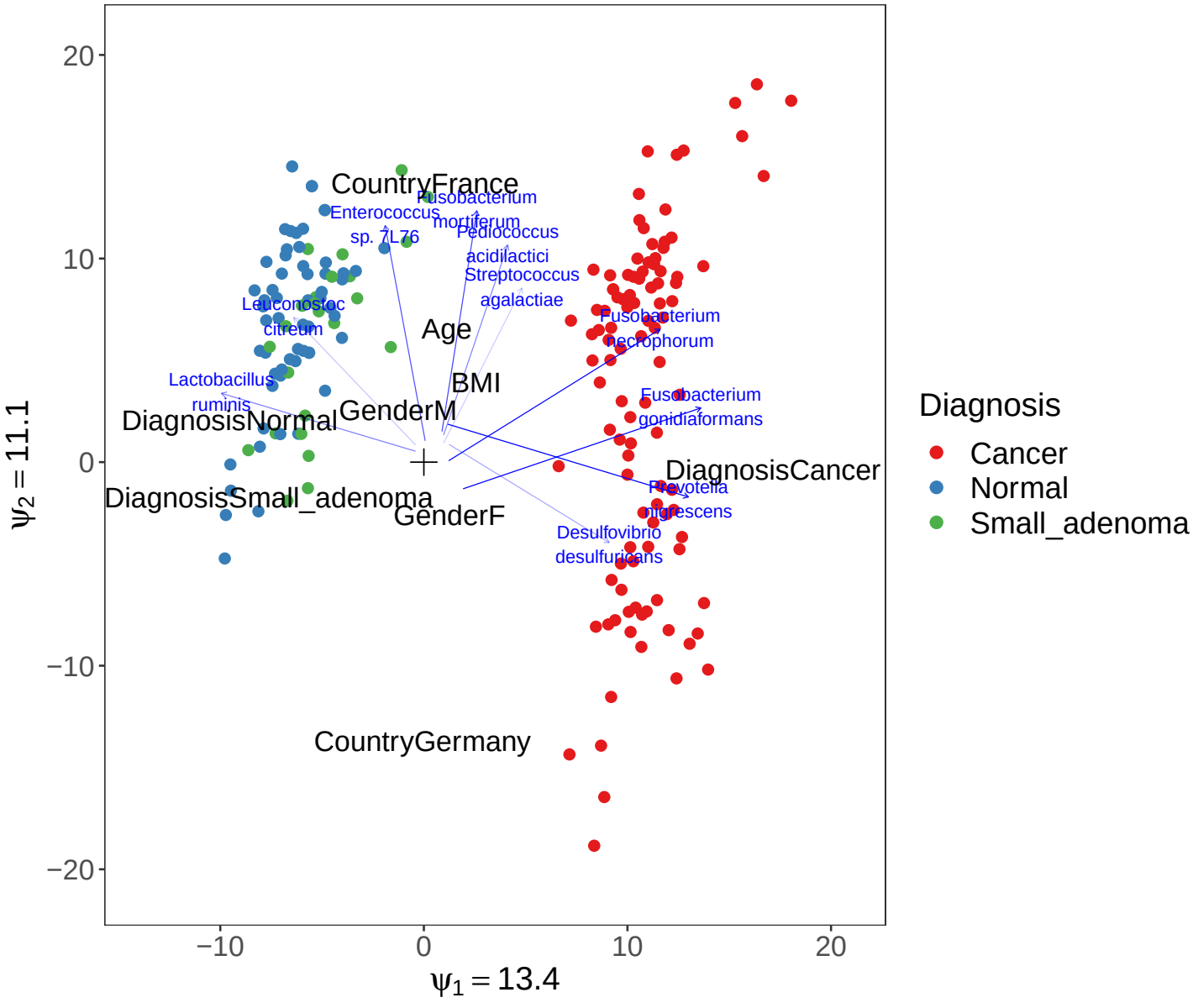
Estimating the environmental gradient

- α_k^t is estimated by comparing two different models

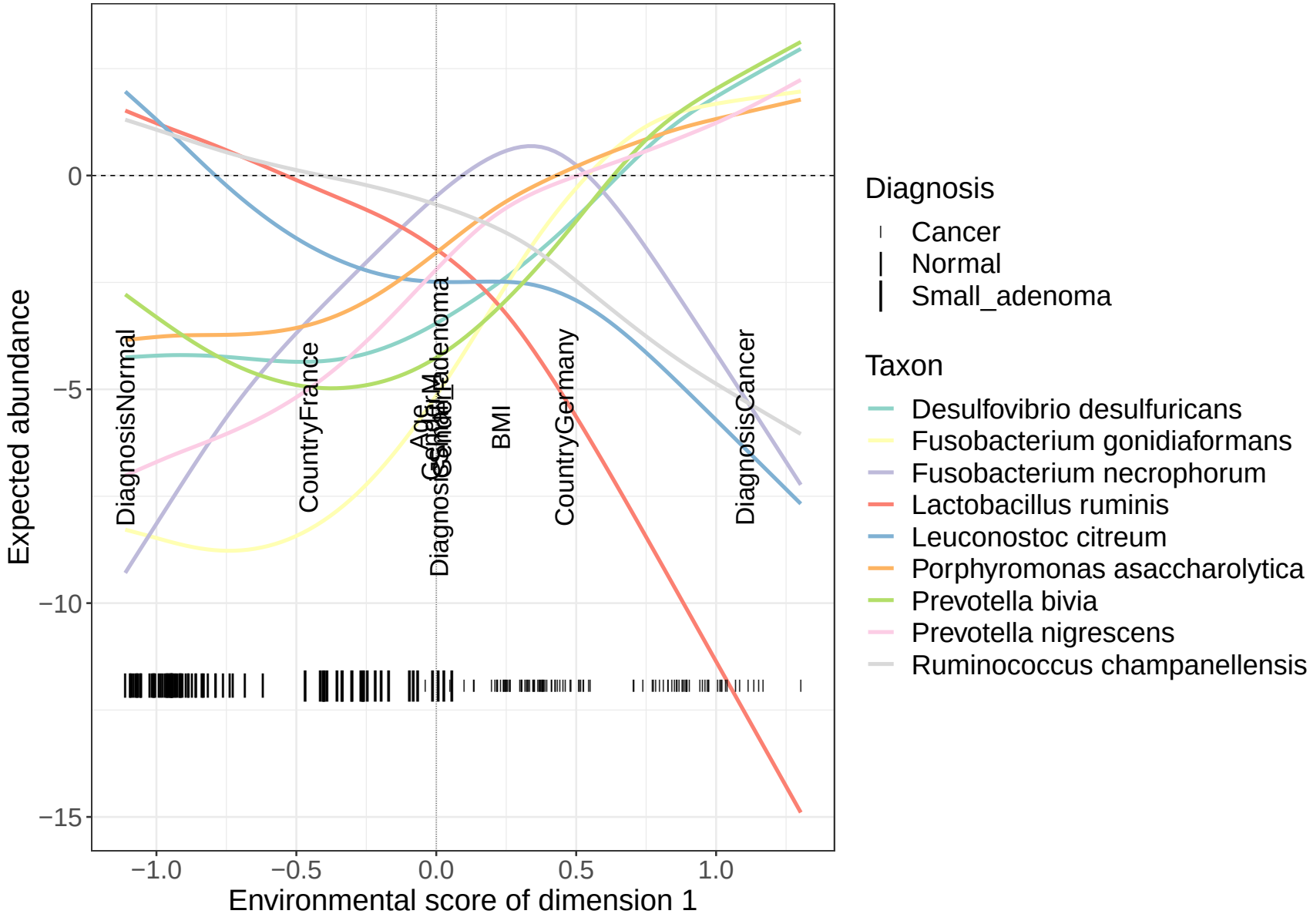
$$LR(\boldsymbol{\alpha}) = \frac{\overbrace{\prod_{i=1}^{n}\prod_{j=1}^{p} l(X_{ij}|u_i, u_j, \mathbf{C}_i, \boldsymbol{\alpha}, \psi, f_j)}^{\text{Every species reacts differently to the environment}}}{\underbrace{\prod_{i=1}^{n}\prod_{j=1}^{p} l(X_{ij}|u_i, u_j, \mathbf{C}_i, \boldsymbol{\alpha}, \psi, f)}_{\text{All species react equally}}}$$

- Optimize $LR(\boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\alpha}$
- Encourage competition and **differential niche use** between species

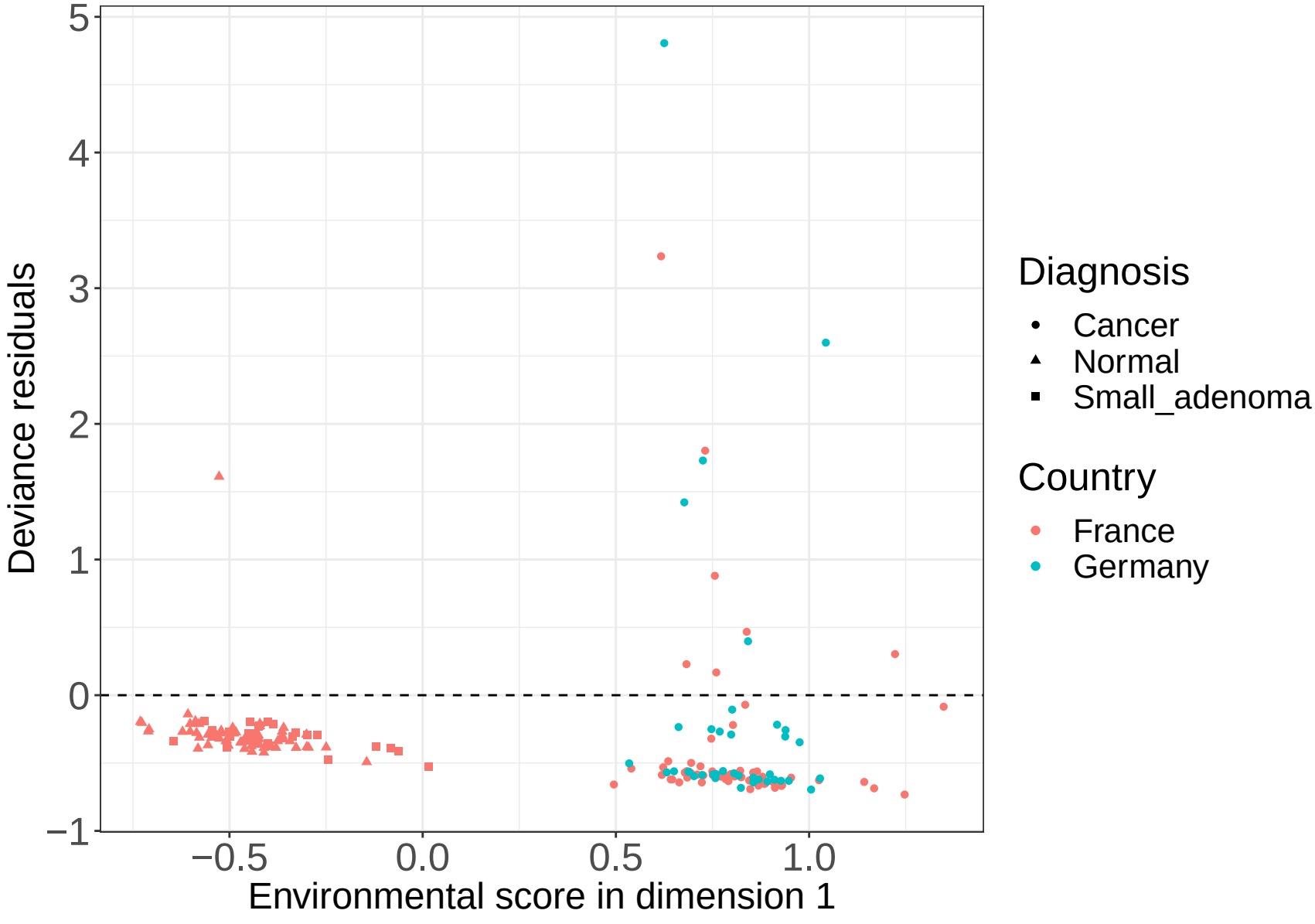
Constrained RC(M): linear response function



Constrained RC(M): Non-parametric response function



Checking the assumptions



Pros and cons of the RC(M)-method

Weaknesses

- **Parametric assumption**

Strengths

- Intuitive **interpretation**
- Flexible in dealing with covariates: **conditioning** and **constrained** analysis
- Naturally deals with **missing values**
- Assumptions are made **explicit** and can be checked

Published!

PLOS ONE

PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A unified framework for unconstrained and constrained ordination of microbiome read count data

Stijn Hawinkel, Frederiek-Maarten Kerckhof, Luc Bijnens, Olivier Thas

Published: February 13, 2019 • https://doi.org/10.1371/journal.pone.0205474 • >> See the preprint

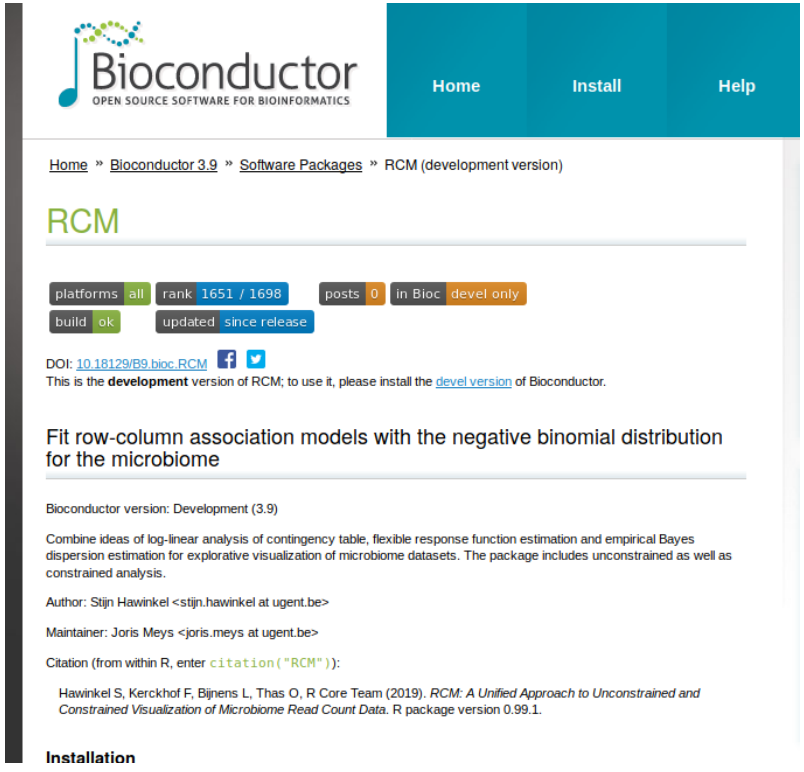
Article | Authors | Metrics | Comments | Media Coverage

Abstract
Introduction
Materials and methods
Results
Discussion
Supporting information
Acknowledgments
References

Reader Comments (0)
Media Coverage (0)
Figures

Abstract

Explorative visualization techniques provide a first summary of microbiome read count datasets through dimension reduction. A plethora of dimension reduction methods exists, but many of them focus primarily on sample ordination, failing to elucidate the role of the bacterial species. Moreover, implicit but often unrealistic assumptions underlying these methods fail to account for overdispersion and differences in sequencing depth, which are two typical characteristics of sequencing data. We combine log-linear models with a dispersion estimation algorithm and flexible response function modelling into a framework for unconstrained and constrained ordination. The method is able to cope with differences in dispersion between taxa and varying sequencing depths, to yield meaningful biological patterns. Moreover, it can correct for observed technical confounders, whereas other methods are adversely affected by these artefacts. Unlike distance-based ordination methods, the assumptions underlying our method are stated explicitly and can be verified using simple diagnostics. The combination of unconstrained and constrained ordination in the same framework is unique in the field and facilitates microbiome data exploration. We illustrate the advantages of our method on simulated and real datasets, while pointing out flaws in existing methods. The algorithms for fitting and plotting are available in the R-package *RCM*.



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help

Home » Bioconductor 3.9 » Software Packages » RCM (development version)

RCM

platforms all | rank 1651 / 1698 | posts 0 | in Bioc devel only
build ok | updated since release

DOI: 10.18129/B9.bioc.RCM
This is the **development** version of RCM; to use it, please install the devel version of Bioconductor.

Fit row-column association models with the negative binomial distribution for the microbiome

Bioconductor version: Development (3.9)

Combine ideas of log-linear analysis of contingency table, flexible response function estimation and empirical Bayes dispersion estimation for explorative visualization of microbiome datasets. The package includes unconstrained as well as constrained analysis.

Author: Stijn Hawinkel <stijn.hawinkel at ugent.be>

Maintainer: Joris Meys <joris.meys at ugent.be>

Citation (from within R, enter `citation("RCM")`):

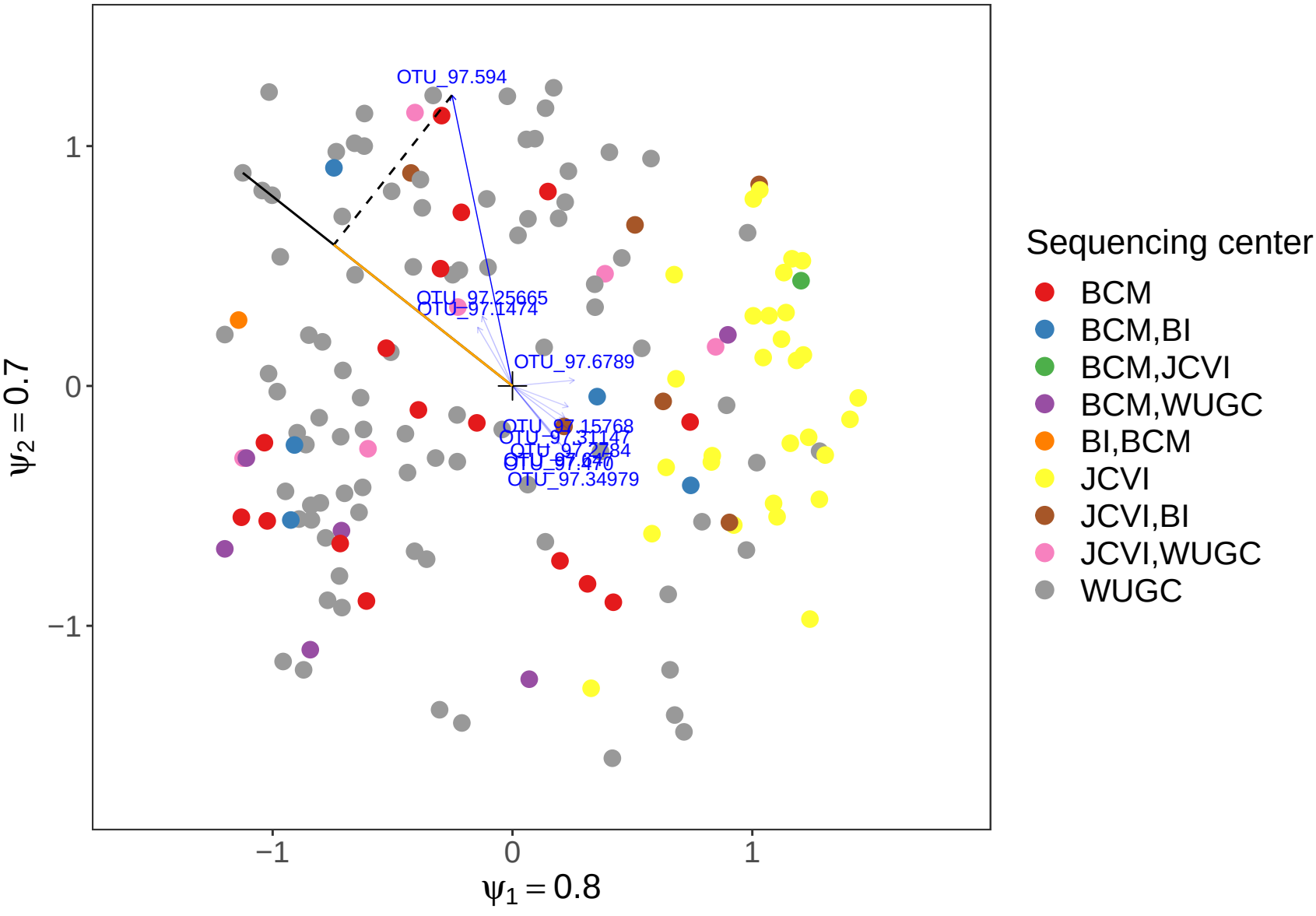
Hawinkel S, Kerckhof F, Bijnens L, Thas O, R Core Team (2019). *RCM: A Unified Approach to Unconstrained and Constrained Visualization of Microbiome Read Count Data*. R package version 0.99.1.

Installation

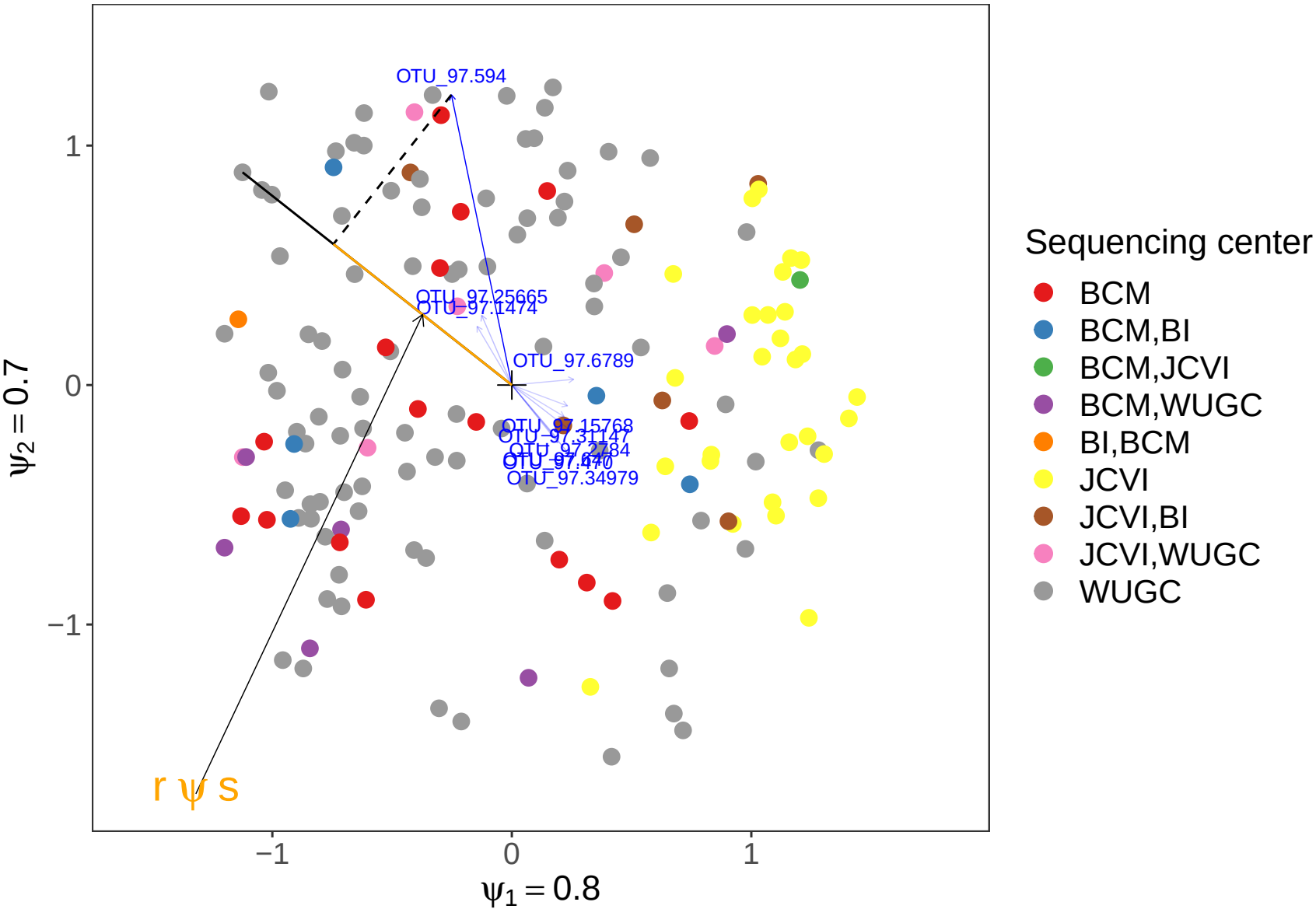
Thanks for your attention!

Any questions?

RC(M) plot: orthogonal projection



RC(M) plot: orthogonal projection



Shape of the Response function