CODA 4 MICROBIOME

Malu Calle
Universitat de Vic - UCC

UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA

CODA 4 MICROBIOME

Compositional Data analysis

CoDA

Microbiome analysis

Metagenomics

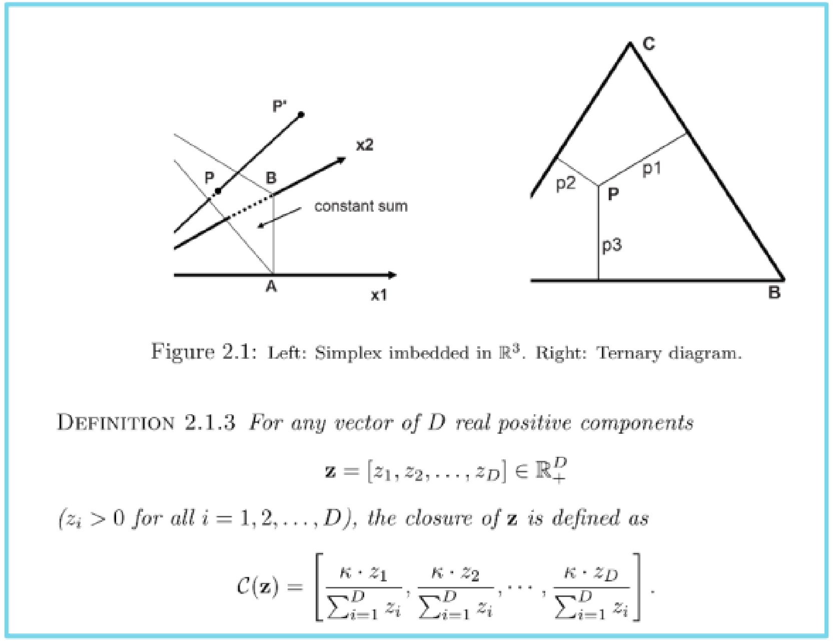

P'
x2
P
B
constant sum
A
x1
C
p2
p1
P
p3
B

Figure 2.1: Left: Simplex imbedded in $\mathbb{R}^3$. Right: Ternary diagram.

DEFINITION 2.1.3 *For any vector of D real positive components*

$$\mathbf{z} = [z_1, z_2, \dots, z_D] \in \mathbb{R}^D_+$$

($z_i > 0$ for all $i = 1, 2, \dots, D$), the closure of $\mathbf{z}$ is defined as

$$\mathcal{C}(\mathbf{z}) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^D z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^D z_i}, \cdots, \frac{\kappa \cdot z_D}{\sum_{i=1}^D z_i} \right].$$

Lecture Notes on Compositional Data Analysis,
Pawlowsky-Glahn, Egozcue and Tolosana-Delgado

CODA 4 MICROBIOME

https://malucalle.github.io/coda4microbiome/

Identification of microbial signatures

`coda_glmnet`: cross-sectional studies (Y binary or continuous)

`coda_glmnet_longitudinal`: longitudinal studies

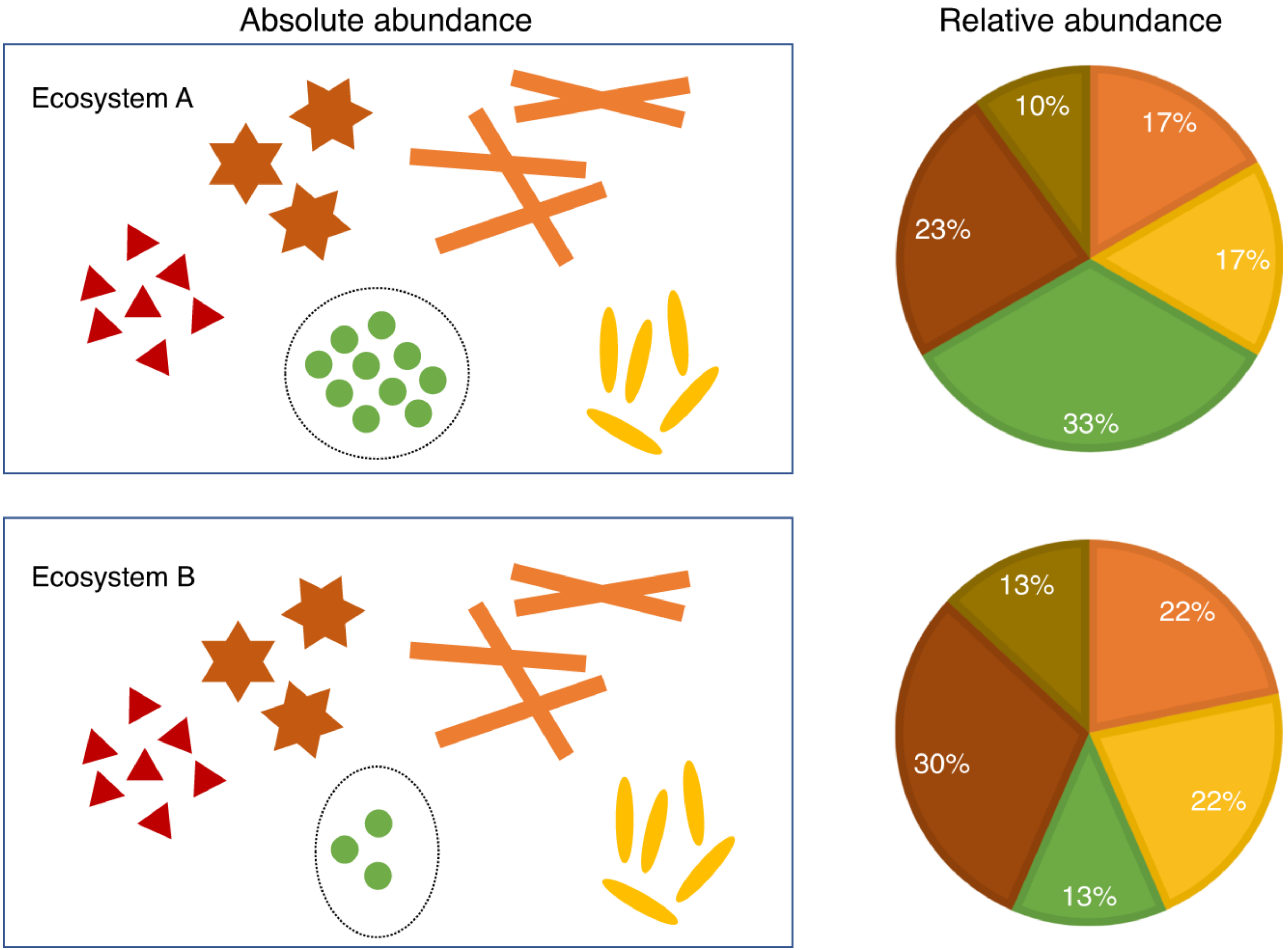
Log-ratio exploratory analysis

`explore_logratios`: association of each pairwise log-ratio with Y

`explore_lr_longitudinal`: association of a summary of each log-ratio trajectory with Y

Suplementary functions

`explore_zeros`, `impute_zeros`, `logratios_matrix`, `plot_prediction`, `plot_signature`, `coda_glmnet_null`, `filter_longitudinal`, `coda_glmnet_longitudinal_null`, `shannon`, `shannon_effnum`, `shannon_sim`



Lin and Peddada, Nature Communications 2020

Microbiome compositional data

Taxon1 ***Taxon2*** … ***TaxonM***

Y	OTU1	OTU2	OTU3	…	OTUk	TOTAL
Y_1	X_{11}	X_{12}	X_{13}	…	X_{1k}	N_1
Y_2	X_{21}	X_{22}	X_{23}	…	X_{2k}	N_2
…	…			…		
Y_n	X_{n1}	X_{n2}	X_{n3}	…	X_{nk}	N_n

"Absolute" abundances:

$$(300,\ \ 600\ \ , \dots ,\ \ \dots),\ \ \mathrm{N}_1 = 3000$$

They don't represent the total abundance (absolute) in the environment

Samples with different total, N_i,are not comparable

$$(100,\ \ 200,\ \ \dots ,\ \ \dots),\ \ \mathrm{N}_2 = 1000$$

Microbiome compositional data

	Taxon1		*Taxon2*	…	*TaxonM*	
Y	**OTU1**	**OTU2**	**OTU3**	**…**	**OTUk**	**TOTAL**
Y_1	X_{11}	X_{12}	X_{13}	…	X_{1k}	N_1
Y_2	X_{21}	X_{22}	X_{23}	…	X_{2k}	N_2
…	…			…		
Y_n	X_{n1}	X_{n2}	X_{n3}	…	X_{nk}	N_n

"Absolute" abundances:

$$(300,\ \ 600\ \ , \dots ,\ \ \dots),\ \ N_1 = 3000$$

"Solution": work with relative abundances (proportions):

$$\boldsymbol{p = (0.1, 0.2, \dots, \dots)}$$

The problem of working with proportions

$$p = (p_1, p_2, \ldots, p_K), \ \sum p_i = 1$$

○ ***Spurious correlations***

○ ***Subcompositionals incoherences***

○ ***False positives of univariate differential abundance tests***

Proportions and spurious correlations

*Working with proportions induces **spurious correlation** (Pearson 1896):*

- *Two or more variables will be **negatively correlated** simply because the data are transformed to have a constant sum*

$$x = \begin{bmatrix} 790 & 488 & 1174 & 1037 \\ 737 & 470 & 1052 & 1064 \\ 589 & 386 & 1112 & 772 \\ 634 & 344 & 741 & 870 \end{bmatrix} \qquad \pi_x = \begin{bmatrix} 0.226 & 0.139 & 0.336 & 0.297 \\ 0.221 & 0.141 & 0.316 & 0.320 \\ 0.206 & 0.135 & 0.388 & 0.270 \\ 0.244 & 0.132 & 0.286 & 0.336 \end{bmatrix}$$

$$\downarrow \qquad\qquad\qquad\qquad \downarrow$$

$$cor(x) = \begin{bmatrix} 1 & 0.89 & 0.43 & 0.94 \\ & 1 & 0.76 & 0.83 \\ & & 1 & 0.28 \\ & & & 1 \end{bmatrix} \qquad cor(\pi_x) = \begin{bmatrix} 1 & \mathbf{-0.28} & \mathbf{-0.93} & 0.88 \\ & 1 & 0.03 & \mathbf{-0.04} \\ & & 1 & \mathbf{-0.98} \\ & & & 1 \end{bmatrix}$$

Proportions and subcompositional incoherences

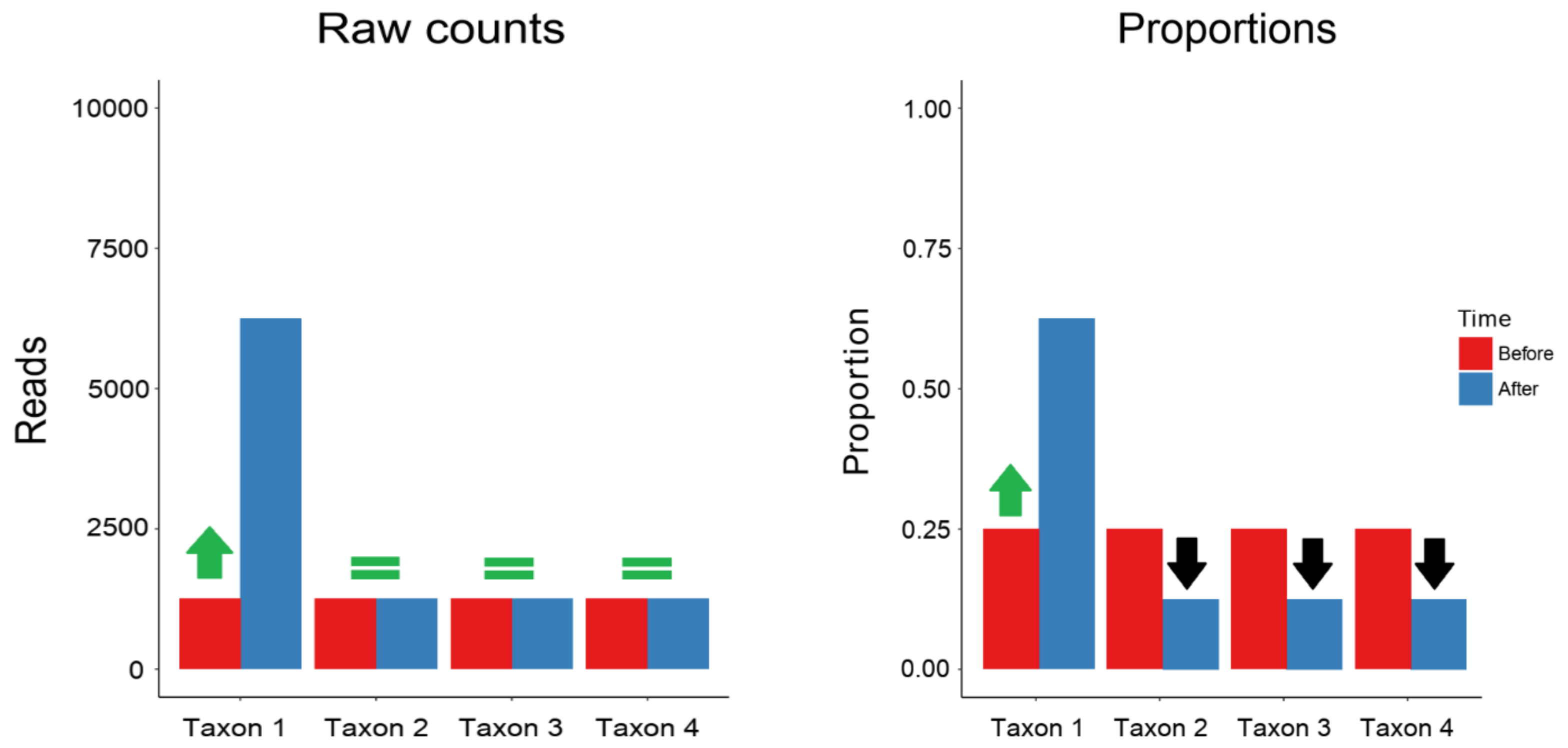
Working with proportions induces ***subcompositional incoherences****.*

$$x = \begin{bmatrix} 790 & 488 & 1174 & \xcancel{1037} \\ 737 & 470 & 1052 & \xcancel{1064} \\ 589 & 386 & 1112 & \xcancel{772} \\ 634 & 344 & 741 & \xcancel{870} \end{bmatrix}, \quad cor(\pi_x) = \begin{bmatrix} 1 & \mathbf{-0.28} & -093 & 0.88 \\ & 1 & \mathbf{0.03} & -0.04 \\ & & 1 & -0.98 \\ & & & 1 \end{bmatrix}$$

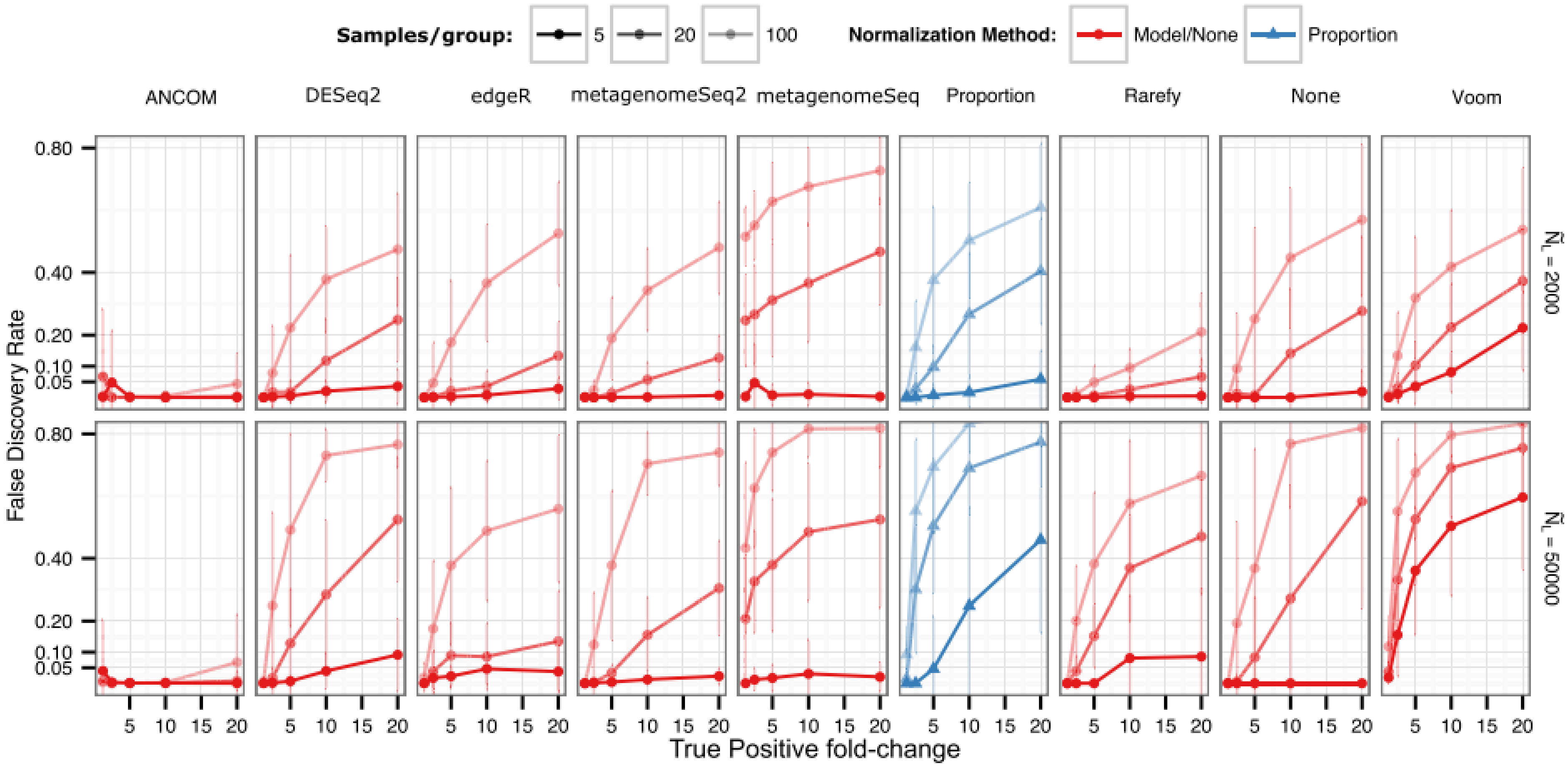
$$y = \begin{bmatrix} 790 & 488 & 1174 \\ 737 & 470 & 1052 \\ 589 & 386 & 1112 \\ 634 & 344 & 741 \end{bmatrix}, \quad cor(\pi_y) = \begin{bmatrix} 1 & \mathbf{0.64} & -0.98 \\ & 1 & \mathbf{-0.76} \\ & & 1 \end{bmatrix}$$

Unreliable univariate differential abundance testing

***The** change in abundance of one species induces changes in the **observed** abundances of the other species*

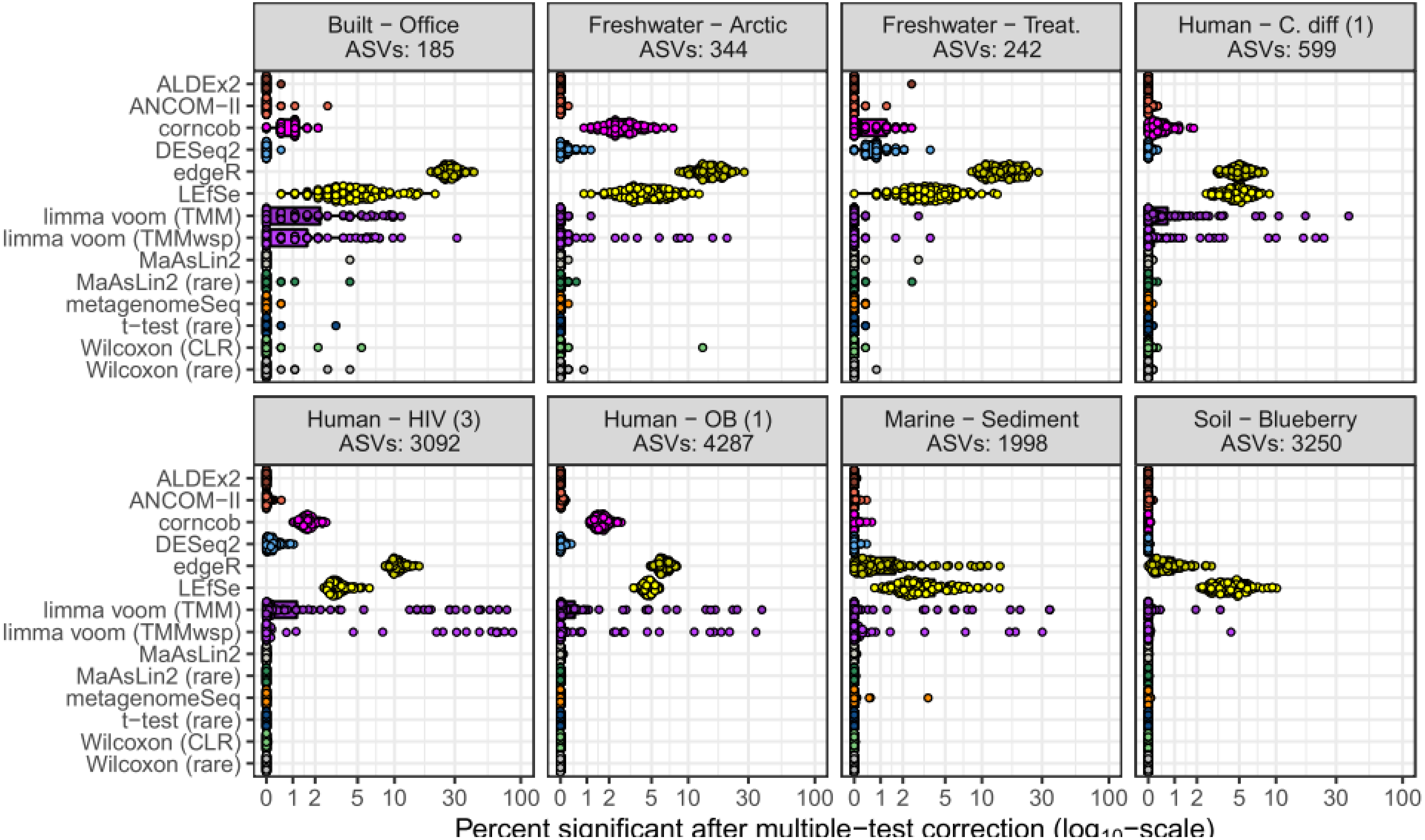


Bias microbiome diff. abundance testing



Weiss et al. *Microbiome* 2017

Bias microbiome diff. abundance testing



Nearing et al. *Nat.Comm* 2022

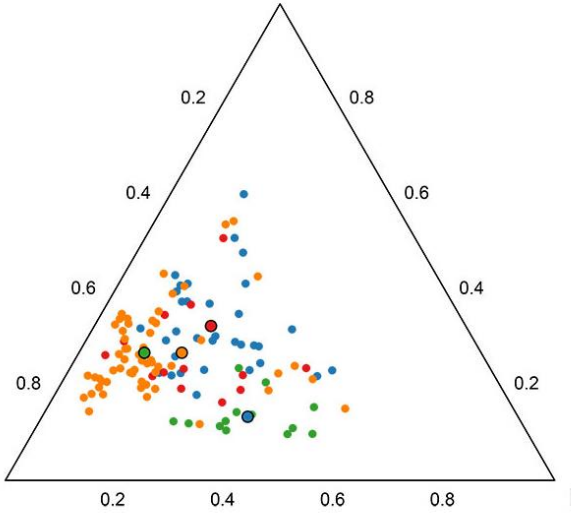
Compositional data analysis

Aitchison, 1986:

*A **composition** is defined as a vector of **positive** real numbers, $x = (x_1, \dots, x_k)$, x_i>0, that contains **relative information**.*

$$S^k = \{ x = (x_1, \dots, x_k), \quad x_i > 0 , \quad \sum_{i=1}^{k} x_i = 1\}$$

*The **simplex:** the sample space of compositional data*



https://doi.org/10.3832/ifor2187-010

The log-ratio approach

CoDA principles:

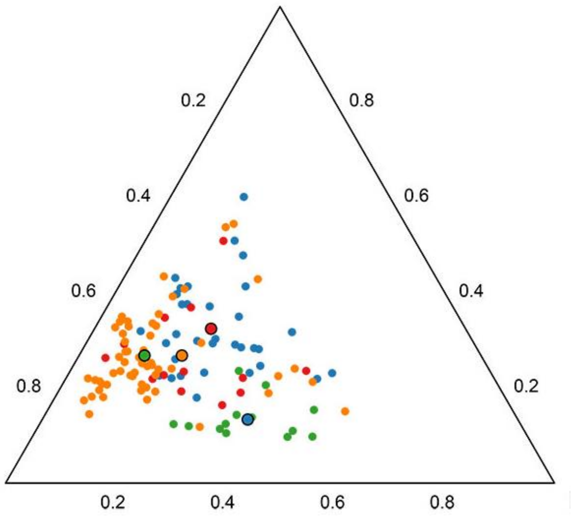
- *permutation invariance*
- *scale invariance:* $f(x) = f(\alpha \cdot x)$
- *sub-compositional coherence*

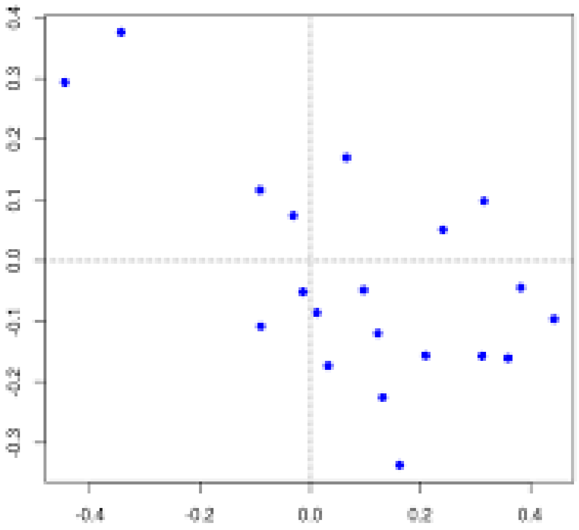
The simplest invariant function is the ***log-ratio*** *between two components:*

$$f(x) = log\left(\frac{x_i}{x_j}\right) = log(x_i) - log(x_j),\ i,j \in \{1, \dots, k\}.$$

Coda transformations

Scale-invariant transformations from the simplex S^k to the real space $\mathbb{R}^{k-1}$





https://doi.org/10.3832/ifor2187-010ali

Coda transformations

- *The **additive log-ratio transformation** (alr):*

$$\text{alr}(x) = alr(x_1, \dots, x_k) = (log(x_1/x_k), \dots, log(x_{k-1}/x_k))$$

- *The **centered log-ratio transformation** (clr):*

$$clr(x) = clr(x_1, \dots, x_k) = (log(x_1/g(x)), \dots, log(x_k/g(x))) =$$

$$where\ g(x) = \prod x_j^{1/k}\ \ geometric\ mean$$

$$= (log(x_1) - M, \dots, log(x_k) - M)\ where\ M = \log(g(x)) = \frac{1}{k}\sum_j log\left(x_j\right).$$

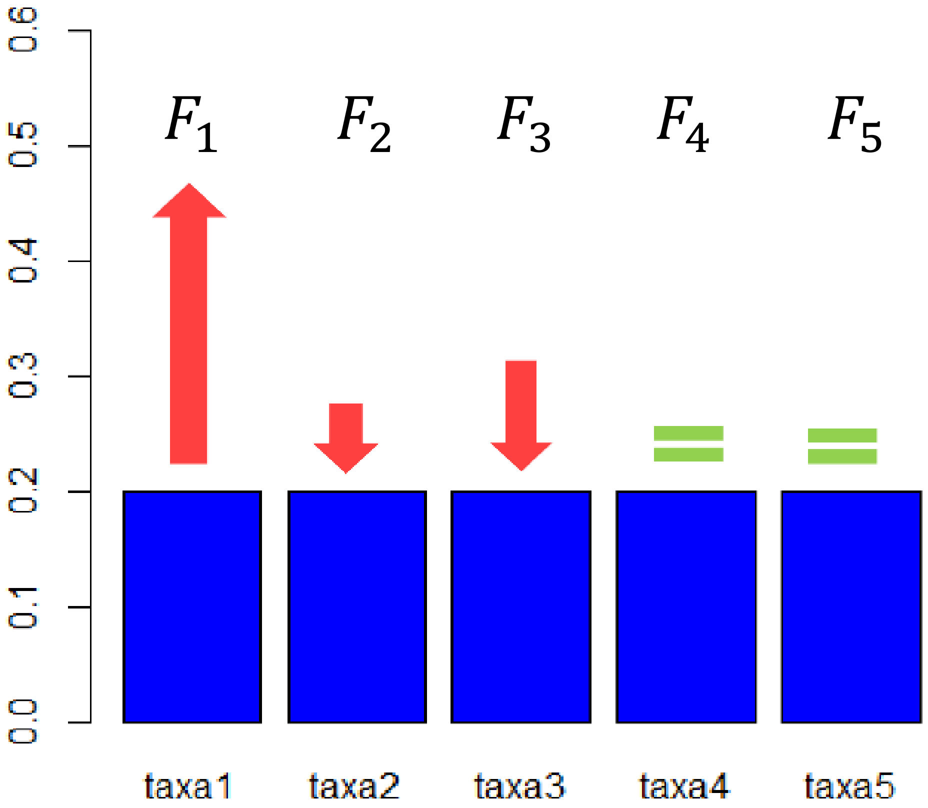
Variable selection of microbial species

Univariate differential abundance testing results in ***large proportion of FP***

Reason: *The change in abundance of one species induces changes in the* ***observed abundances*** *of the other species*

- *Quantification of the* **bias** *when working with proportions.*
- *Is the clr-transformation* ***unbiased****?*

Variable selection of microbial species



Univariate differential abundance testing:

$$H_0\colon\ F_i\ = 1$$

$$H_1\colon\ F_i\ \neq 1$$

Testing (log) proportions

*Composition in **environment 1**:* $p = (p_1, p_2, \dots, p_K),\ \sum p_i = 1$

Fold-change effects: $F = (F_1, F_2, \dots, F_K), \qquad F_i > 0$

*Observed proportions in **environment 2**:*

$$p^* = \left(\frac{F_1 \cdot p_1}{C}, \frac{F_2 \cdot p_2}{C}, \dots, \frac{F_K \cdot p_K}{C}\right), \qquad C = \sum F_i p_i$$

***Observed effect** on the log-scale:*

$$\log(p_i^*) - \log(p_i) = \log(F_i) \mathbf{- \log(\boldsymbol{C})}$$

When $F_i = 1, \quad \log(p_i^*) - \log(p_i) = \mathbf{- \log(\boldsymbol{C})}$

Testing clr-transformed data

*Composition in **environment 1**:* $p = (p_1, p_2, \ldots, p_K)$, $\sum p_i = 1$

*Observed proportions in **environment 2** :*

$$p^* = \left(\frac{F_1 \cdot p_1}{C}, \frac{F_2 \cdot p_2}{C}, \ldots, \frac{F_K \cdot p_K}{C}\right), \qquad C{=}\sum F_i p_i$$

$$clr(p) = (\log(p_1) - M, \ldots, \log(p_K) - M), \;\; M = \frac{1}{k}\sum_j log\left(p_j\right)$$

$$clr(p^*) = (\log(p_1^*) - M^*, \ldots, \log(p_k^*) - M^*), \;\; M^* = \frac{1}{k}\sum_j log\left(p_j^*\right)$$

Observed effect on clr:

$$clr(p_i^*) - clr(p_i) = \log(F_i) \boldsymbol{- \log\left(g(F)\right)}, \qquad g(F)\text{: geometric mean of } F$$

When $F_i = 1$, $\qquad clr(p_i^*) - clr(p_i) = \boldsymbol{-\log\left(g(F)\right)}$

Univariate testing bias

Log-proportions bias:

$$B_{\log(p)} = -\mathbf{\log(C) =} \qquad C{=}\sum F_i p_i$$

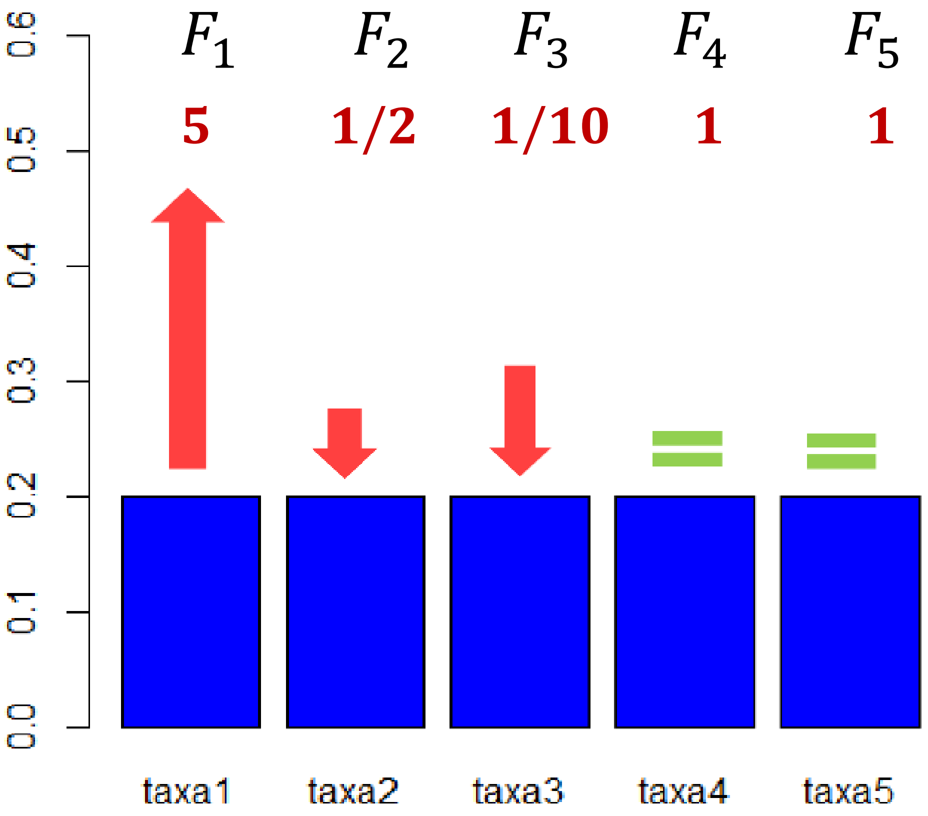
$$= -\log\left(1 + \sum_{j;F_j \neq 1} (F_j - 1) \cdot p_j\right)$$

Clr bias:

$$B_{clr} = -\log\big(g(F)\big) =$$

$$= -\frac{1}{K} \sum_{j;F_j \neq 1} log(F_j)$$

Testing proportions

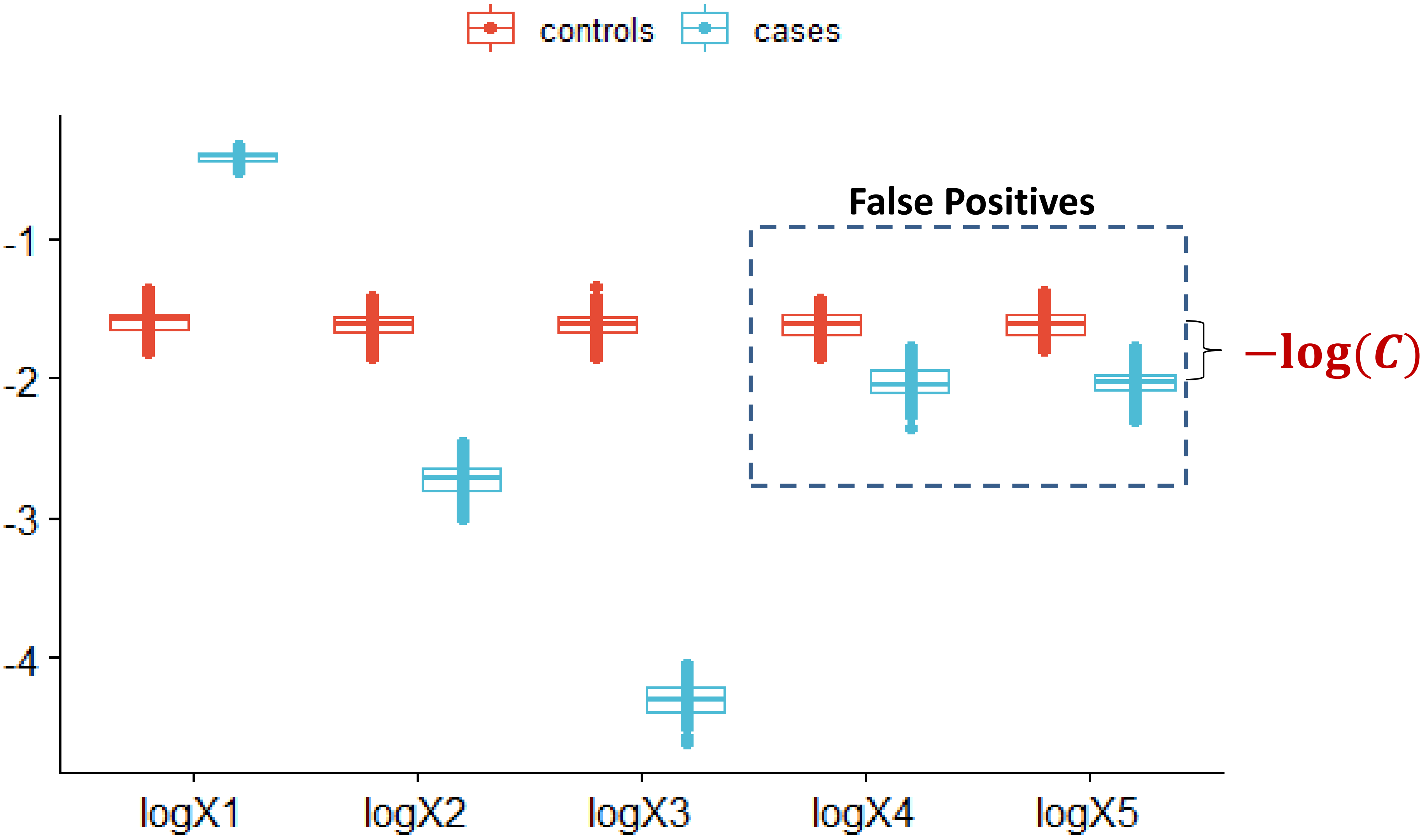


$$p = (p_1 = 0.2, p_2 = 0.2, p_3 = 0.2, p_4 = 0.2, p_5 = 0.2)$$

$$p = (p_1 = 0.02, p_2 = 0.02, p_3 = 0.02, p_4 = 0.47, p_5 = 0.47)$$

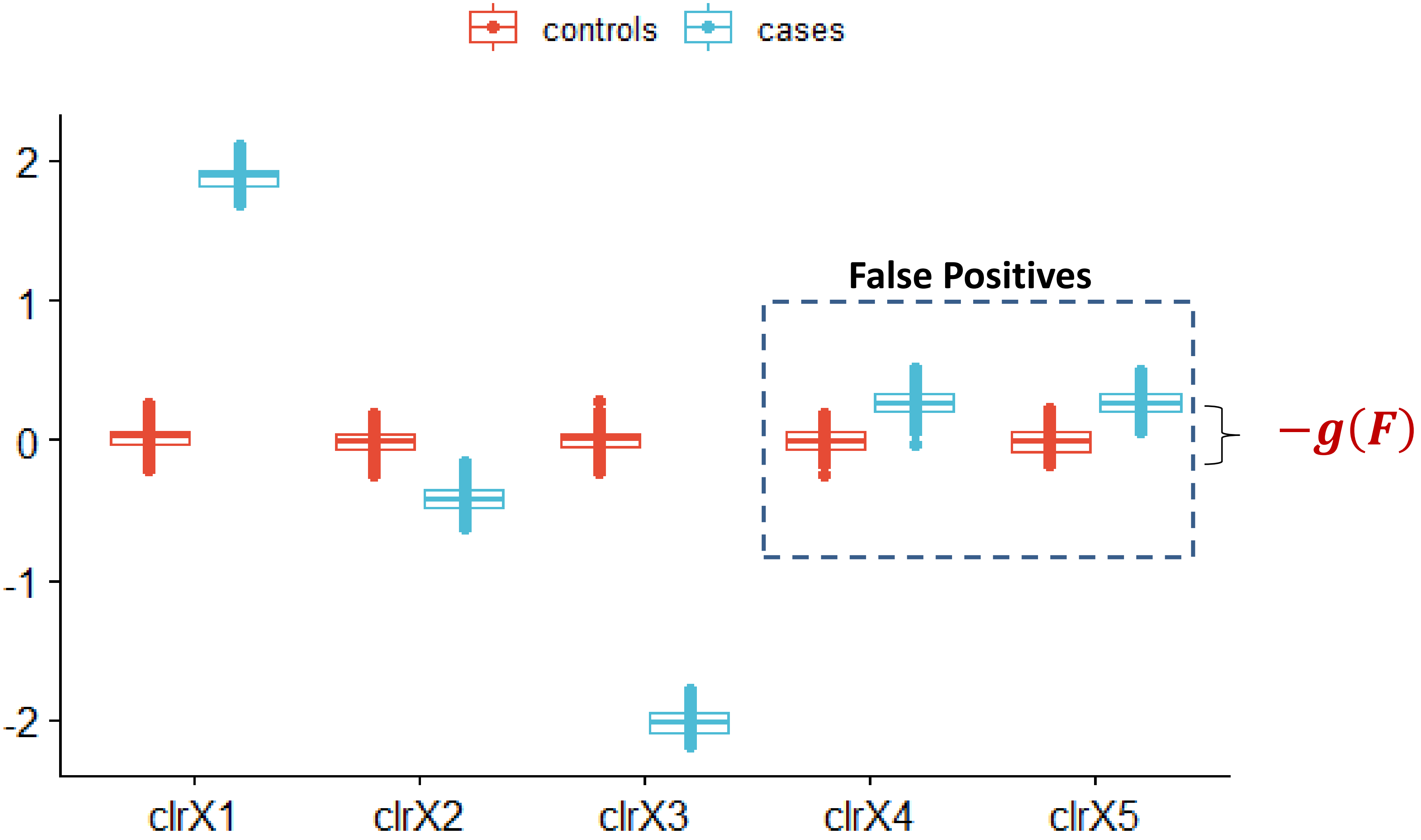
Testing (log)proportions

$p = (p_1 = 0.2, p_2 = 0.2, p_3 = 0.2, p_4 = 0.2, p_5 = 0.2)$



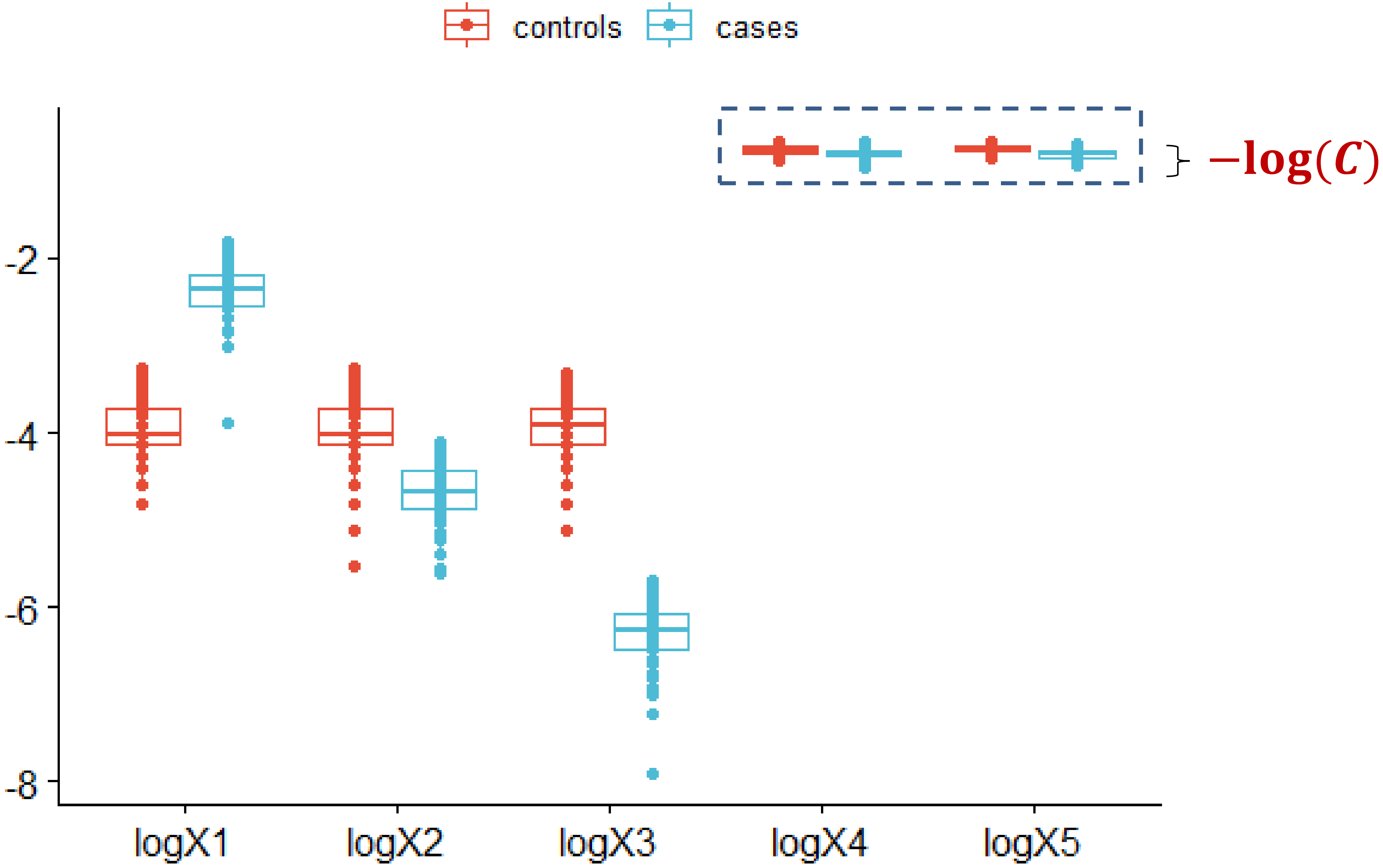
Testing clr-transformed data

$p = (p_1 = 0.2, p_2 = 0.2, p_3 = 0.2, p_4 = 0.2, p_5 = 0.2)$



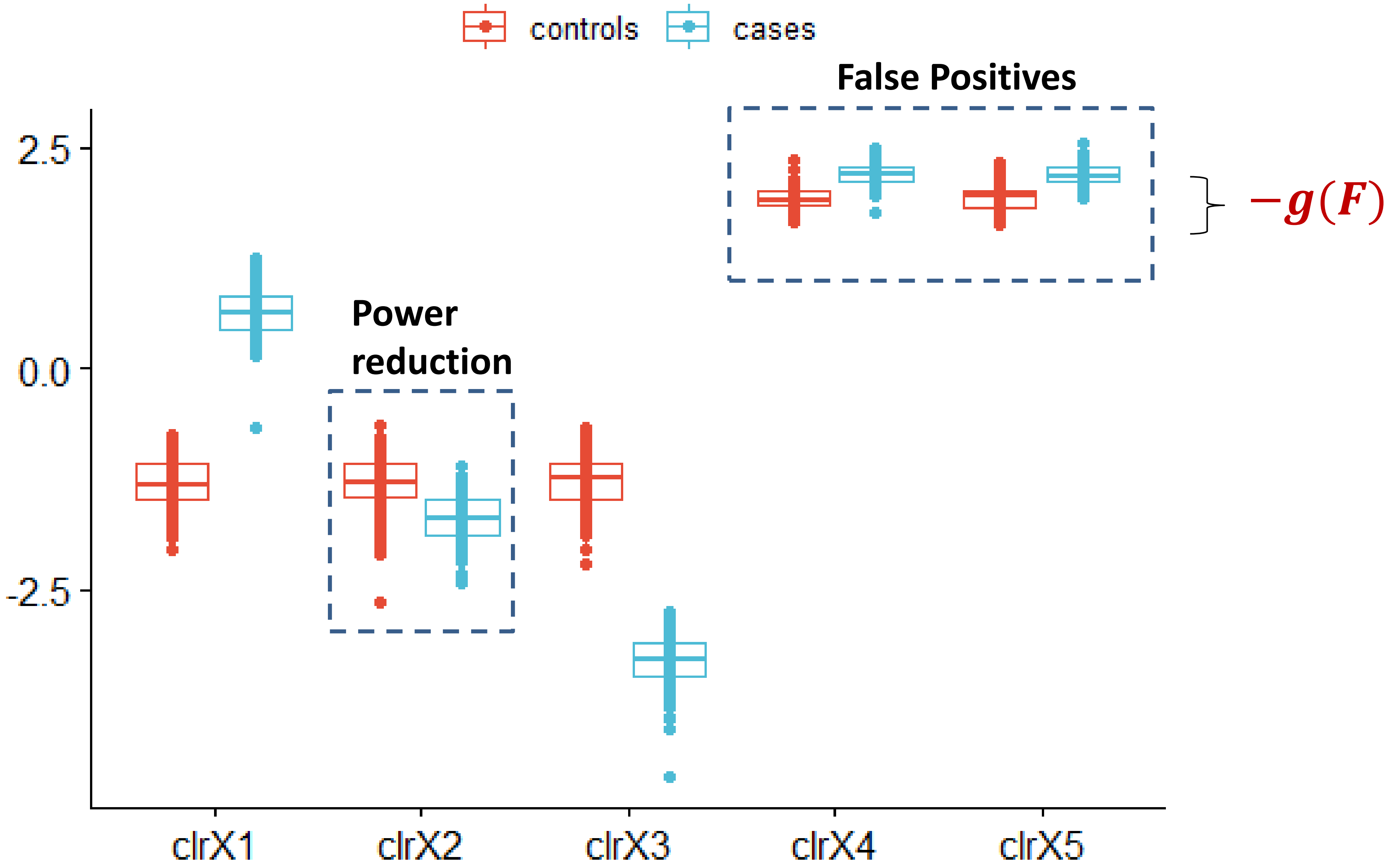
Testing (log)proportions

$p = (p_1 = 0.02, p_2 = 0.02, p_3 = 0.02, p_4 = 0.47, p_5 = 0.47)$



Testing clr-transformed data

$$p = (p_1 = 0.02, p_2 = 0.02, p_3 = 0.02, p_4 = 0.47, p_5 = 0.47)$$

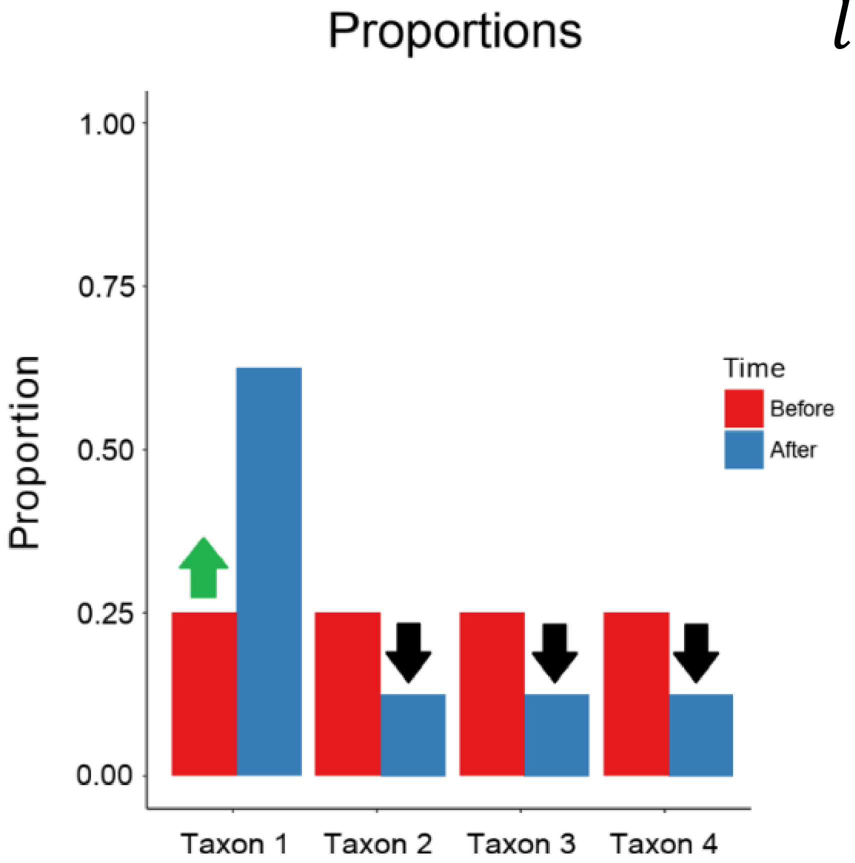


False Positives

Power reduction

$-\boldsymbol{g(F)}$

The log-ratio approach



$$log(x_i/x_j)$$

i	j	log-ratio before	log-ratio after
1	2	log(0.25/0.25)=0	log(0.7/0.1)=log(7)
1	3	log(0.25/0.25)=0	log(0.7/0.1)=log(7)
1	4	log(0.25/0.25)=0	log(0.7/0.1)=log(7)
2	3	log(0.25/0.25)=0	log(0.1/0.1)=0
2	4	log(0.25/0.25)=0	log(0.1/0.1)=0
3	4	log(0.25/0.25)=0	log(0.1/0.1)=0

ANCOM

Mandal et al. (2015)

- *The log-ratio of all pairs of variables is tested,*
- *The number of significant results involving each variable is used to determine its significance*

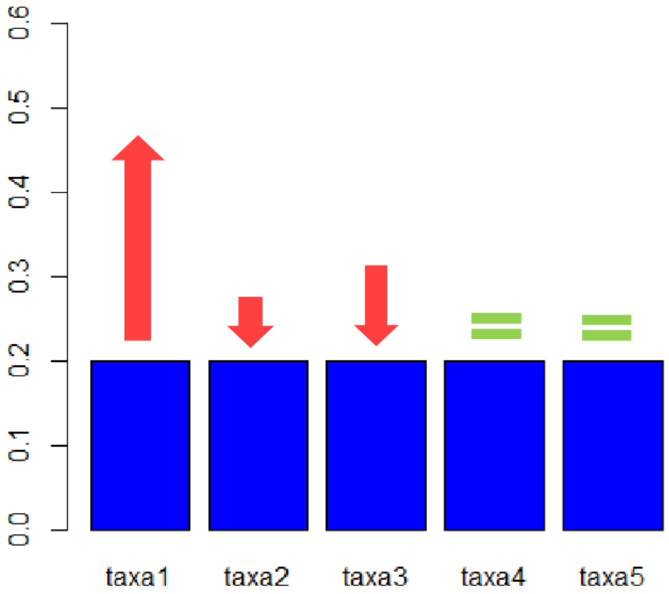
The log-ratio approach

The simplest invariant function is the ***log-ratio*** *between two components:*

$$f(x) = log\left(\frac{x_i}{x_j}\right) = \ log(x_i) - log(x_j),\ \ i,j\ \in\ \{1,\ldots,k\}.$$

Log-ratio extensions

- ***Compositional Balances ->*** ***Selbal***
- ***log-contrast ->*** ***coda4microbiome***

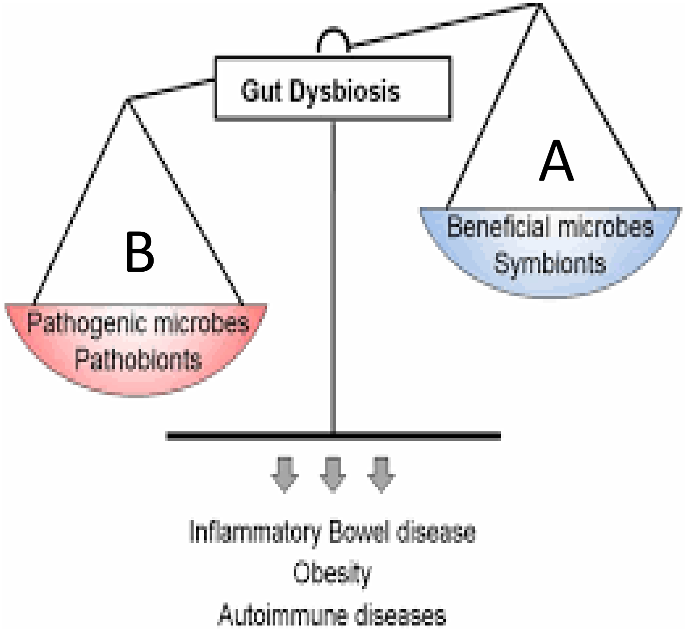


Compositional balance

- ***Compositional Balances***

The balance between two ***sub-compositions A*** *and* ***B*** *of a composition* $X = (X_1, X_2, \dots, X_k)$ *:*

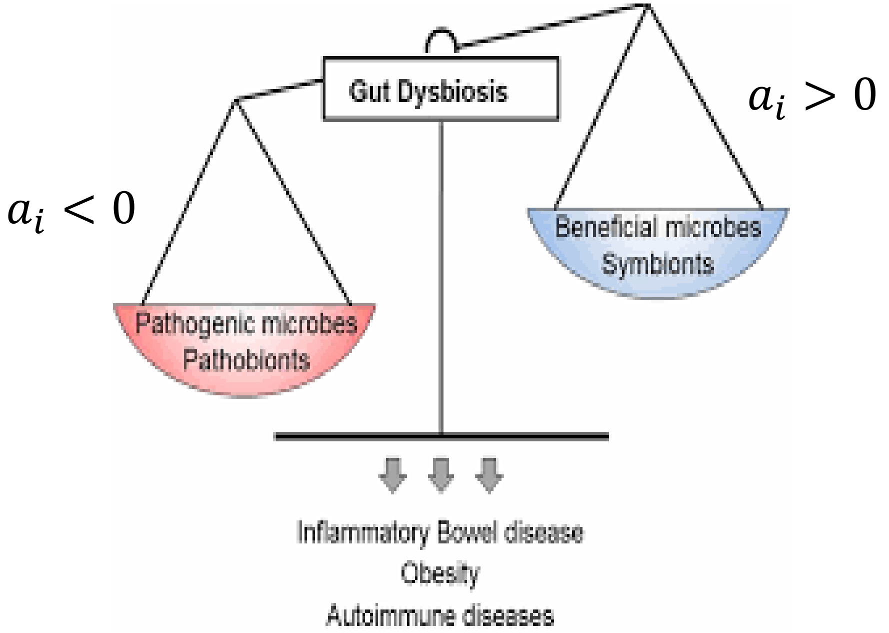
$$\mathcal{B}(A,B) = log\frac{g(X_A)}{g(X_B)} = \frac{1}{n_A}\sum_{X_j \in A} log(x_j) - \frac{1}{n_B}\sum_{X_j \in B} log(x_j)$$



Log-contrast function

- ***Log-contrast function***

$$f(x) = \sum_{i=1}^{k} a_i \, log(x_i)\,; \quad with \quad \sum_{i=1}^{k} a_i = \; 0.$$



Invariant functions

Both, compositional balances and log-contrasts are invariant functions

Consider $\alpha X = (\alpha X_1, \alpha X_2, \dots, \alpha X_k)$ *:*

$$\mathcal{B}_{\alpha X}(A,B) = \frac{1}{n_A}\sum_{X_j \in A} log\left(\alpha x_j\right) - \frac{1}{n_B}\sum_{X_j \in B} log\left(\alpha x_j\right) =$$

$$log(\alpha) + \frac{1}{n_A}\sum_{X_j \in A} log\left(x_j\right) - log(\alpha) - \frac{1}{n_B}\sum_{X_j \in B} log\left(x_j\right) = \mathcal{B}_X(A,B)$$

$$f(\alpha x) = \sum_{i=1}^{k} a_i \, log(\alpha x_i) = log(\alpha)\sum_{i=1}^{k} a_i + \sum_{i=1}^{k} a_i \, log(x_i) =$$

$$= \sum_{i=1}^{k} a_i \, log(x_i) = f(x) \qquad \textit{since} \qquad \sum_{i=1}^{k} a_i = 0.$$

Selbal algorithm

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

Balances: a New Perspective for Microbiome Analysis

J. Rivera-Pinto,[a,b] J. J. Egozcue,[c] V. Pawlowsky-Glahn,[d] R. Paredes,[a,b,e,f] M. Noguera-Julian,[a,b,e] M. L. Calle[b]

- ***Compositional ilr Balances***

$$\mathcal{B}(A,B) = log\frac{g(X_A)}{g(X_B)} = \frac{1}{n_A}\sum_{X_j \in A} log(x_j) - \frac{1}{n_B}\sum_{X_j \in B} log(x_j)$$

Goal:

*Identify two **sub-compositions A** and **B** whose balance $\mathcal{B}(A,B)$ is associated with Y after adjustment for covariates Z:*

Selbal algorithm

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

Balances: a New Perspective for Microbiome Analysis

J. Rivera-Pinto,[a,b] J. J. Egozcue,[c] V. Pawlowsky-Glahn,[d] R. Paredes,[a,b,e,f] M. Noguera-Julian,[a,b,e] M. L. Calle[b]

- ***Compositional ilr Balances***

$$\mathcal{B}(A,B) = log\frac{g(X_A)}{g(X_B)} = \frac{1}{n_A}\sum_{X_j\in A} log(x_j) - \frac{1}{n_B}\sum_{X_j\in B} log(x_j)$$

Generalized linear model:

$$E(g(Y)) = \beta_0 + \beta_1 \cdot \boldsymbol{\mathcal{B}(A,B)} + \gamma' Z$$

Linear regression (Y continuous): $g(Y) = Y$

Logistic regression (Y binary): $g(Y) = logit(Y)$

Selbal forward selection

STEP 0: *Zero replacement*

STEP 1: *Optimal balance between **two** components,* $\mathcal{B}^{(1)}$

The algorithm evaluates all possible balances between two components:

$$\mathcal{B}(X_i, X_j) = \left(log(X_i) - log(X_j)\right) \text{ for } i,j \in \{1, \ . \ . \ . \ , \ k\} \quad i \neq j.$$

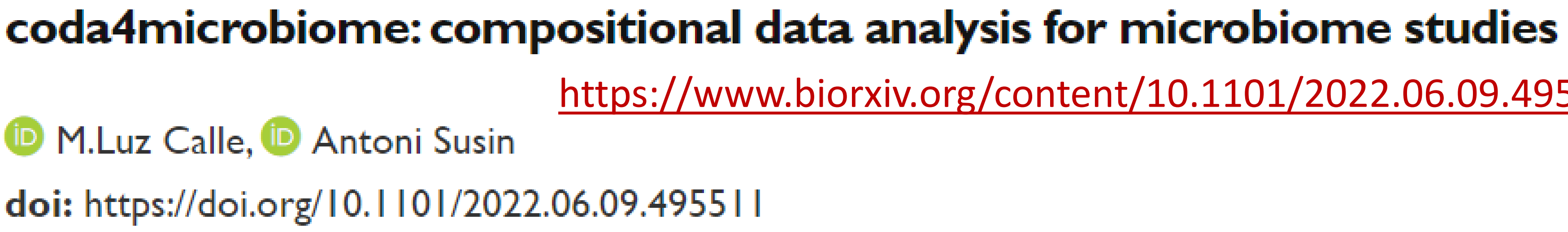
STEP s>1: *Optimal balance adding a new component:*

- *Evaluate the balances obtained by adding* $log\left(X_p\right)$ *to* $\mathcal{B}^{(s-1)}$*, for each remaining variable* X_p
- *Select* $\mathcal{B}^{(s)}$ *that maximizes the optimization criterion (*R^2*, AUC).*

STOP criterion*: cross-validation*

coda4microbiome

CODA 4 MICROBIOME



coda4microbiome: compositional data analysis for microbiome studies

M.Luz Calle, Antoni Susin

doi: https://doi.org/10.1101/2022.06.09.495511

https://www.biorxiv.org/content/10.1101/2022.06.09.495511v1

Considers the "all pairswise log-ratio model":

$$g\left(E\left(Y\right)\right) = \beta_0 + \sum_{1 \le j < k \le K} \beta_{jk} \cdot \log\left(X_j / X_k\right)$$

Coda4microbiome penalized regression

Penalized regression

$$g\left(E(Y)\right) = \beta_0 + \sum_{1\le j<k\le K} \beta_{jk} \cdot \log\left(X_j/X_k\right)$$

with $\|\beta\|_2^2 + \|\beta\|_1 < t$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{L(\beta) + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1\}$$

Linear regression: $\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{\|Y - M\beta\|_2^2 + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1\}$,

M *is the matrix of all pairwise log-ratios*

Coda4microbiome penalized regression

Penalized regression

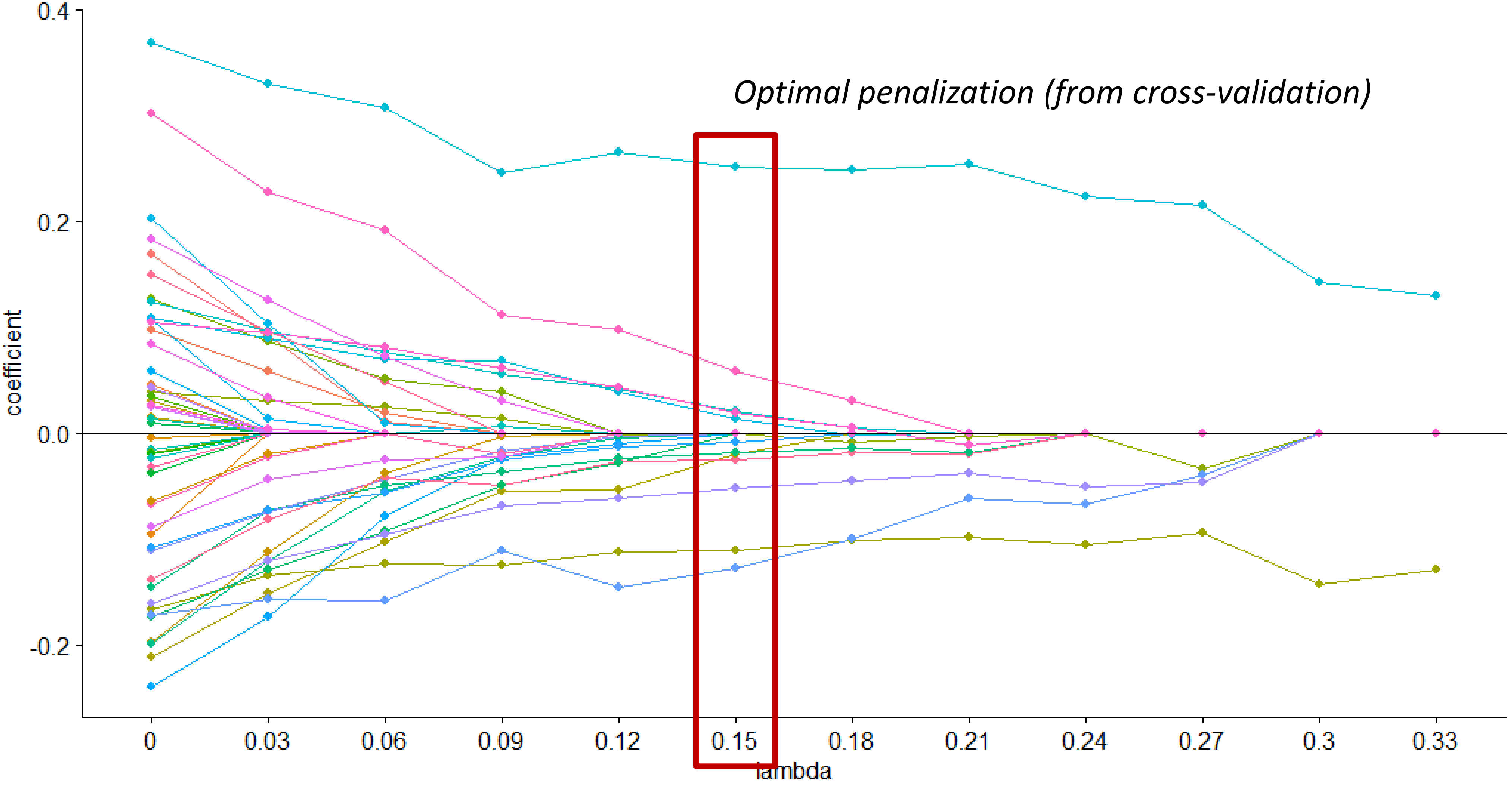
$$g\left(E(Y)\right) = \beta_0 + \sum_{1 \le j < k \le K} \beta_{jk} \cdot \log\left(X_j / X_k\right)$$

$$\text{with } \|\beta\|_2^2 + \|\beta\|_1 < t$$

$\lambda_1 = \lambda(1 - \alpha)/2$ *and* $\lambda_2 = \lambda\alpha$

λ controls the amount of penalization and α the mixing between the two norms (default $\alpha = 0.9$).

Coda4microbiome penalized regression



Optimal penalization (from cross-validation)

coda4microbiome

CODA 4 MICROBIOME

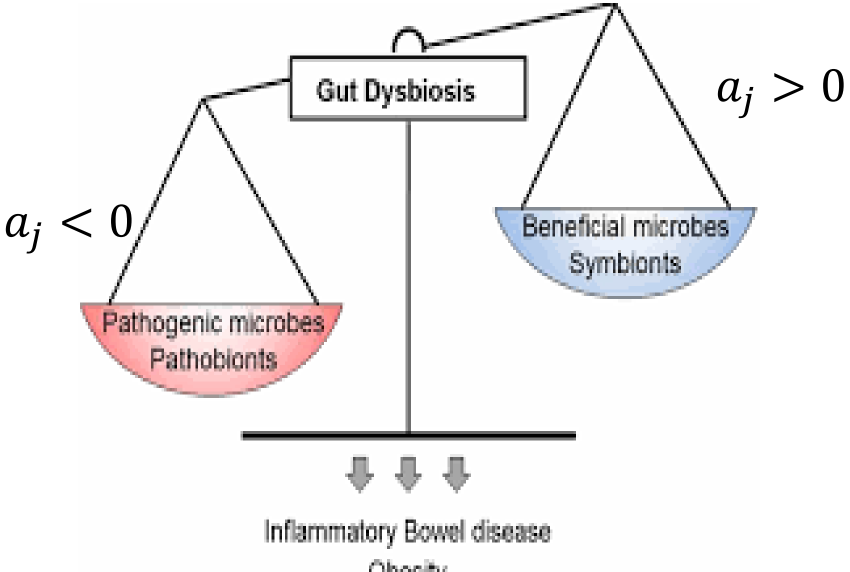
Reparametrization:

$$g\left(E\left(Y\right)\right) = \beta_0 + \sum_{1 \le j < k \le K} \beta_{jk} \cdot \log\left(X_j / X_k\right)$$

Log-contrast model

$$= a_0 + \sum_{j=1}^{K} a_j \cdot \log\left(X_j\right) \quad \text{with} \quad \sum_{j=1}^{K} a_j = 0$$

The results are given as a ***weighted balance*** *between two groups of taxa.*

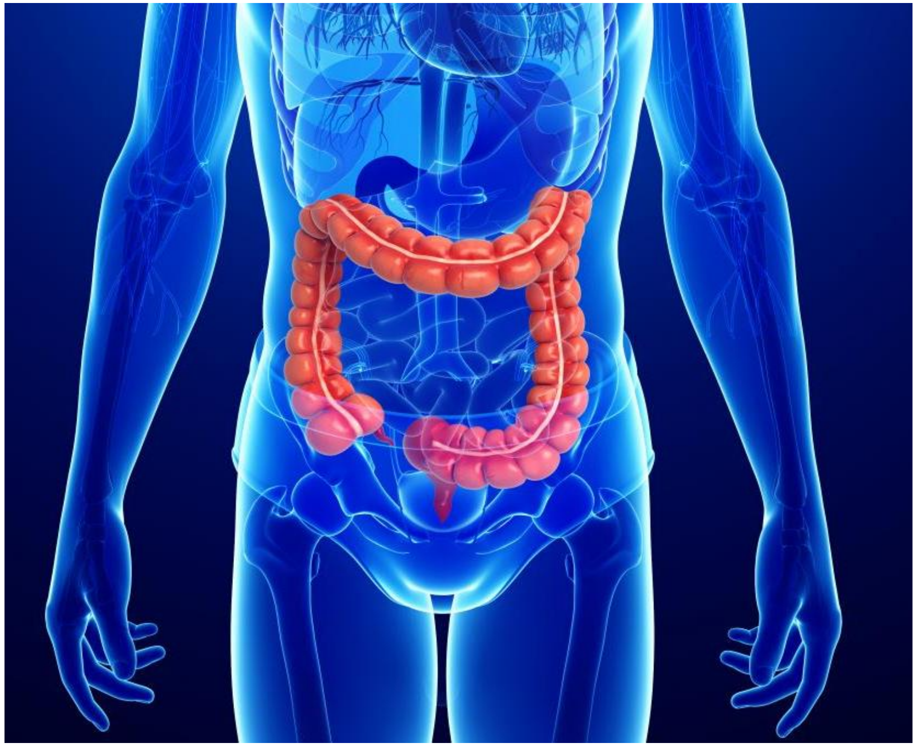


Crohn's disease (CD) study

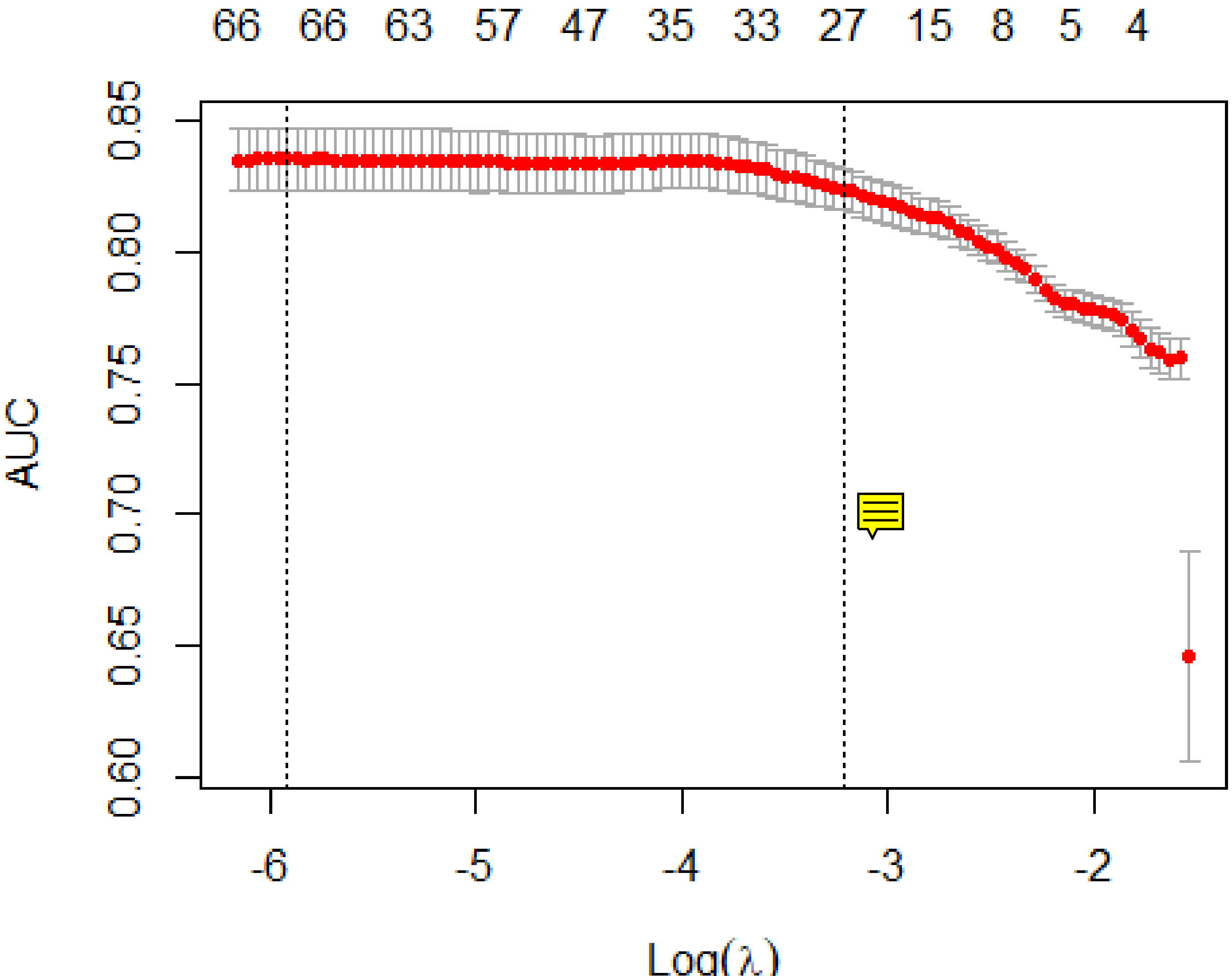
- ***Coda4microbiome with binary outcome*** *Y = disease status (CD or not)*

  n=975 individuals (662 CD and 313 without any symptoms)

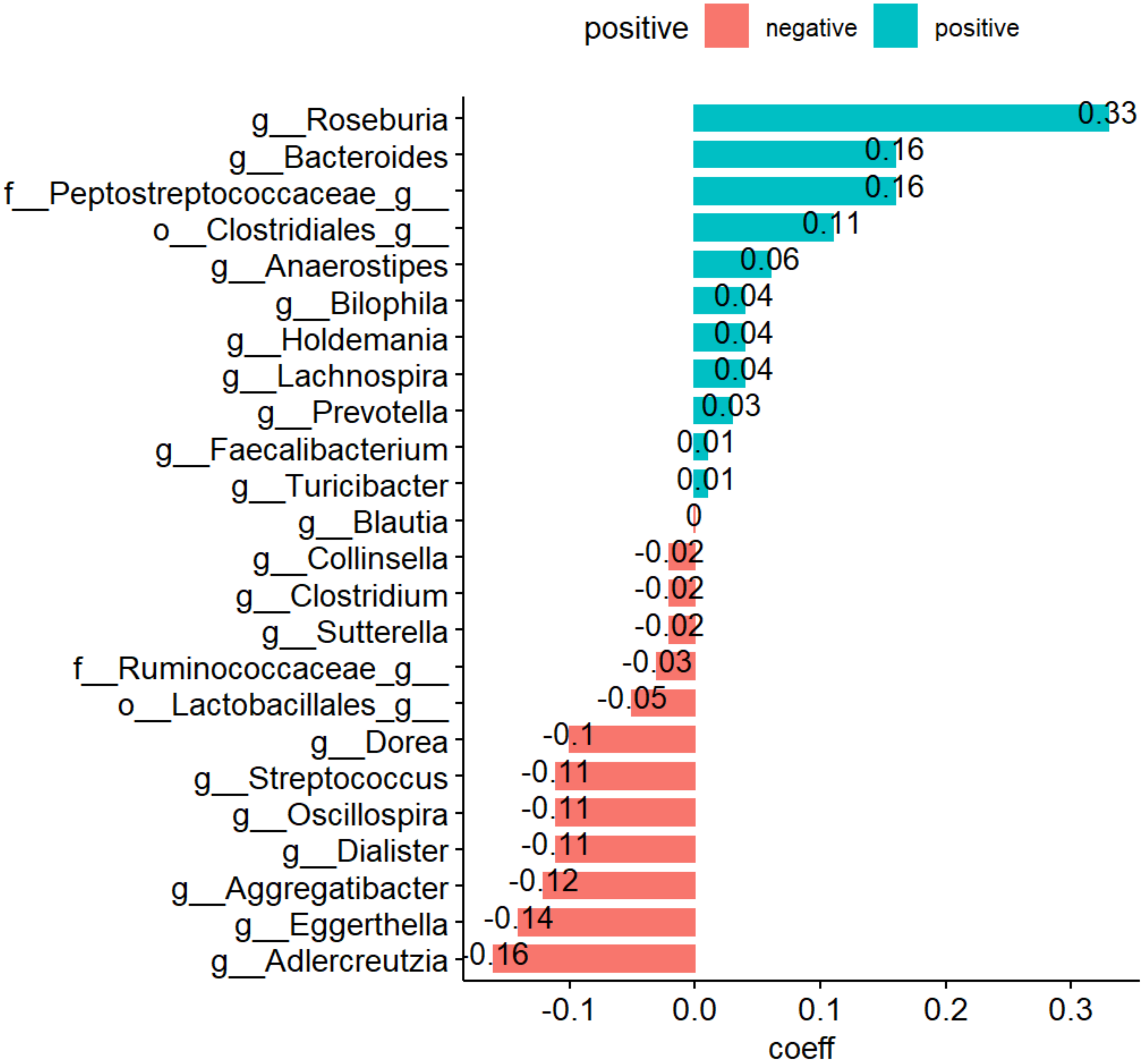
  Microbiome data at genus level: k=48 genera



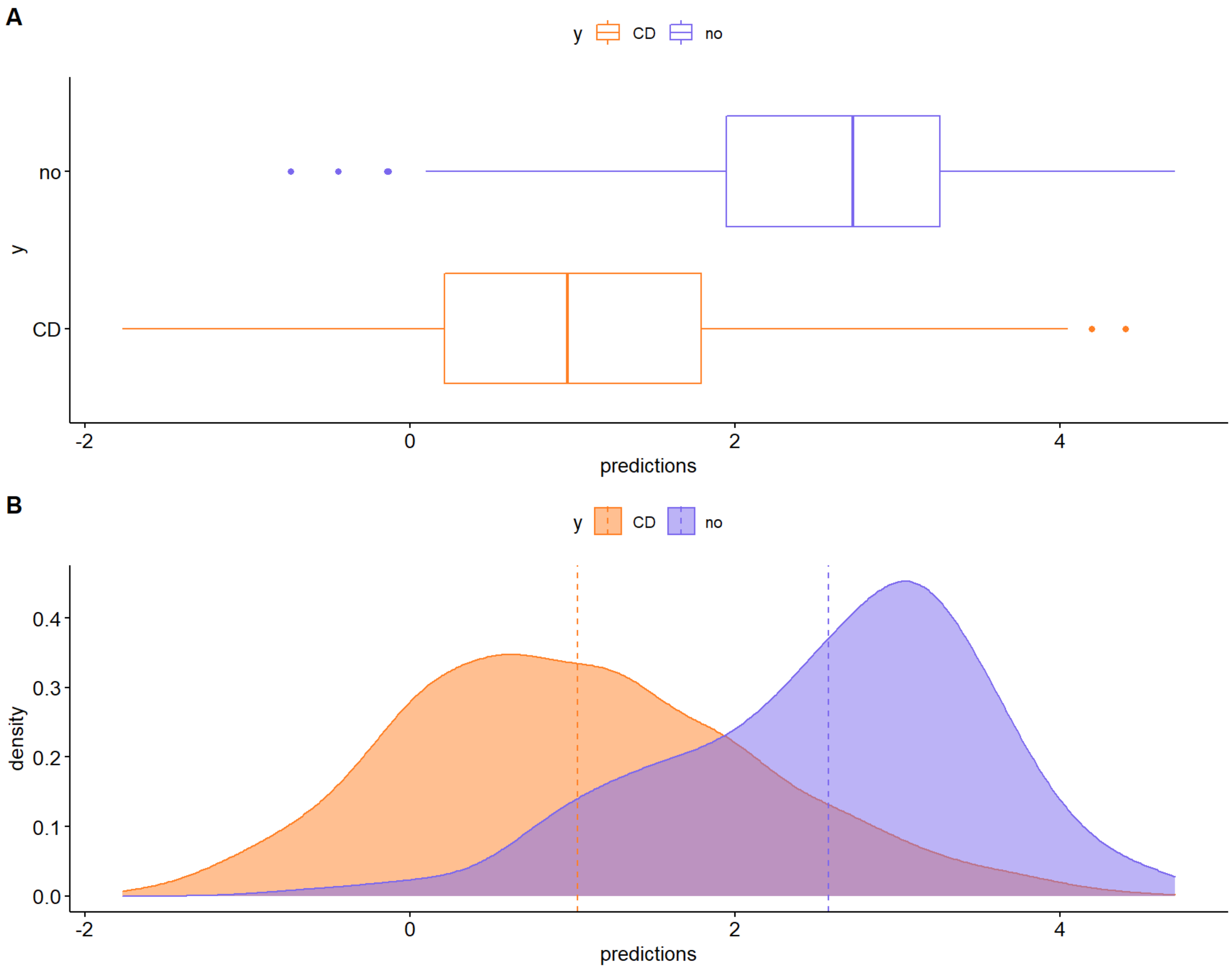
Crohn disease coda4microbiome



Crohn disease coda4microbiome



Crohn disease coda4microbiome



HIV study

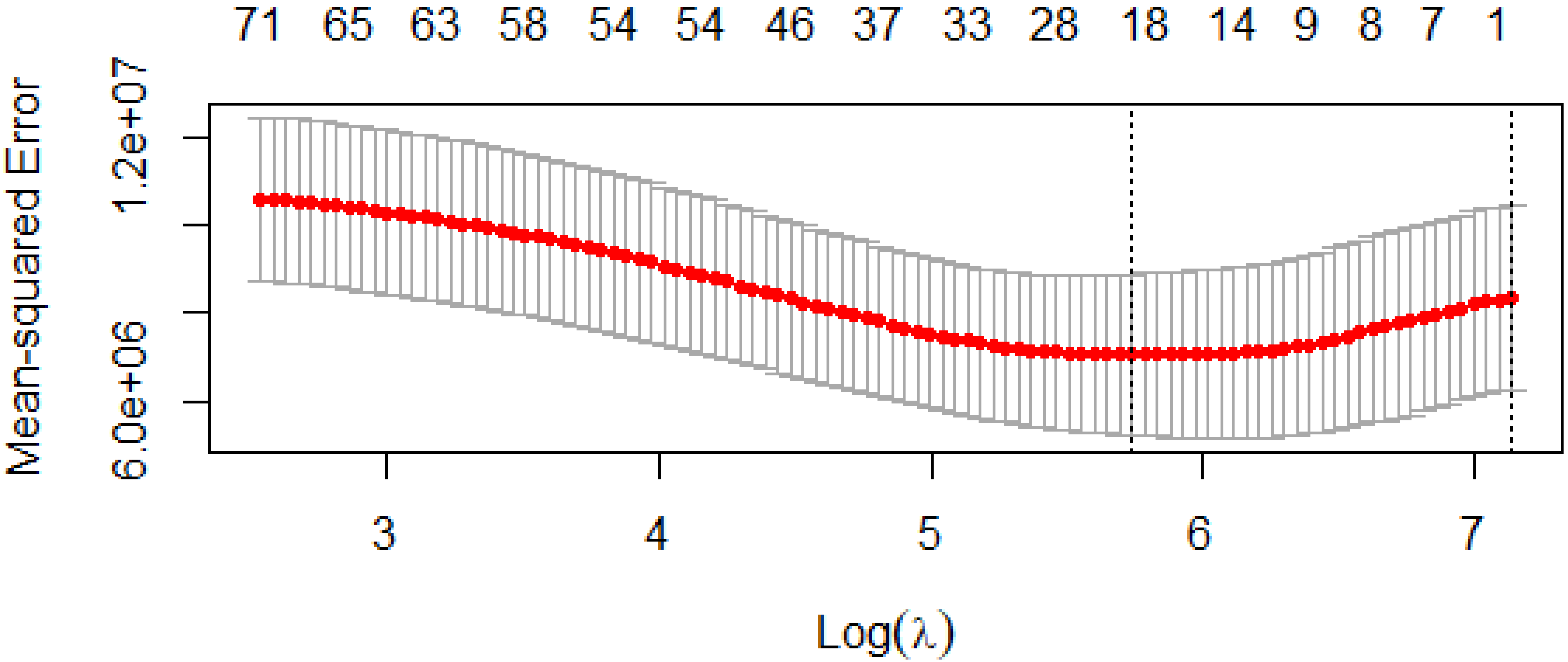
- ***Coda4microbiome with continuous outcome***

Y = sCD14 inflammation marker

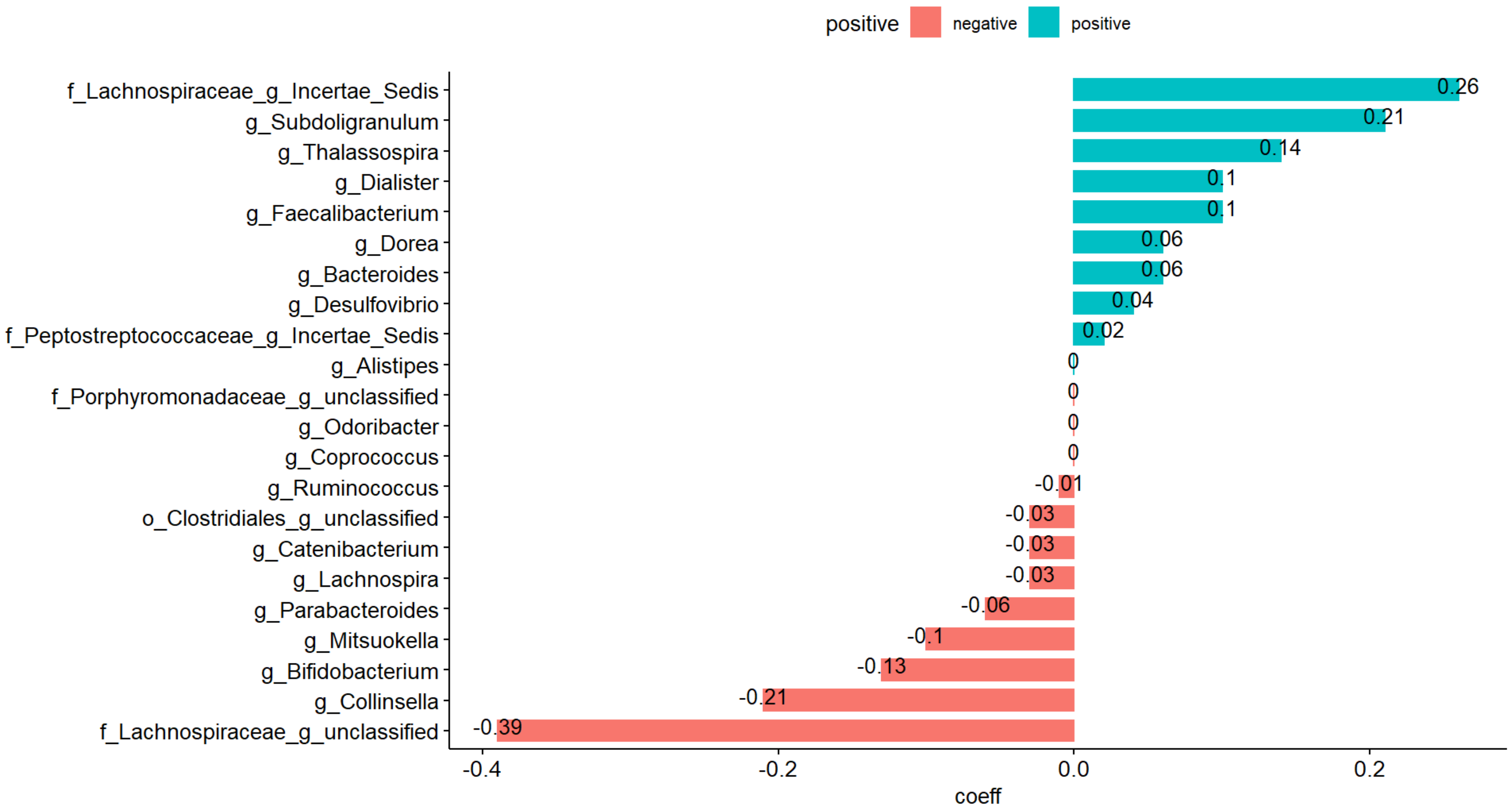
n=151 individuals with HIV

Microbiome data at genus level: k=60 genera

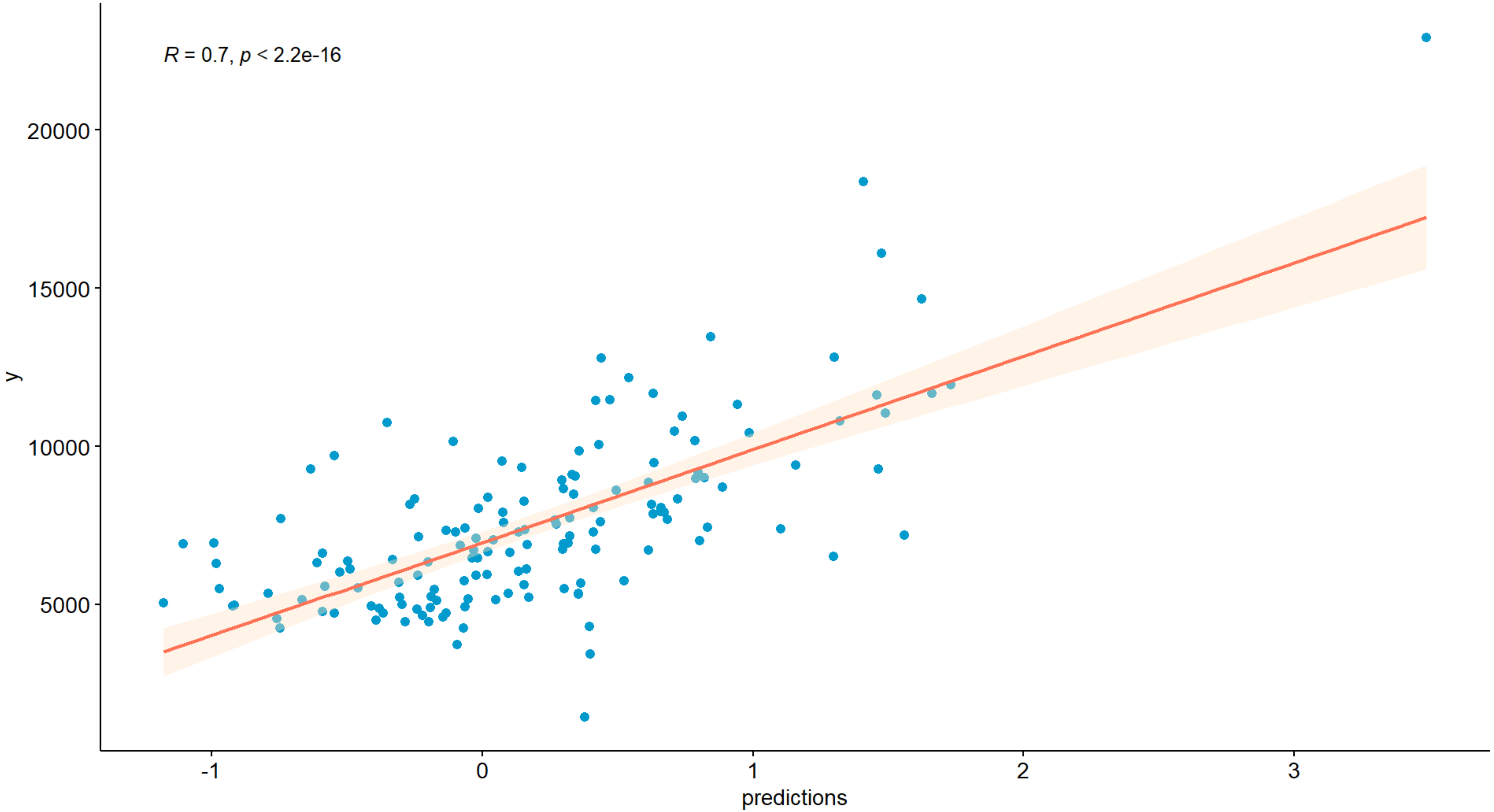
HIV study coda4microbiome



HIV study coda4microbiome



HIV study coda4microbiome



coda4microbiome in longitudinal studies

- ***Low-resolution microbiome longitudinal studies***
  - *Low number of individuals*
  - *Low number of time points*
- ***High-resolution microbiome longitudinal studies***
  - *Mixed models*
  - *Time series*

coda4microbiome in longitudinal studies

IDENTIFICATION OF DYNAMIC MICROBIAL SIGNATURES IN LONGITUDINAL STUDIES

https://www.biorxiv.org/content/10.1101/2022.04.25.489415v1

M.Luz Calle, Antonio Susin

doi: https://doi.org/10.1101/2022.04.25.489415

- ***Longitudinal compositions***

Given a microbiome composition $X = (X_1, X_2, \dots, X_k)$

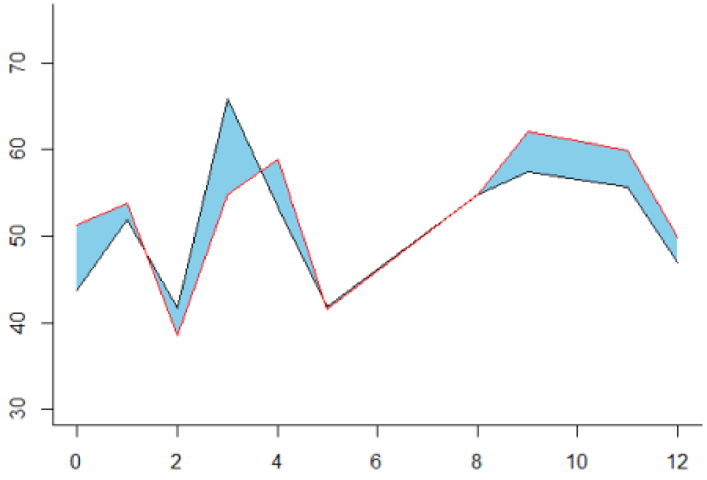
Subject *i* has been observed in L_i time points, $(t_{i1}, t_{i2}, \dots, t_{iL_i})$.

The **log-ratio trajectory** between components A and B for individual *i* is:

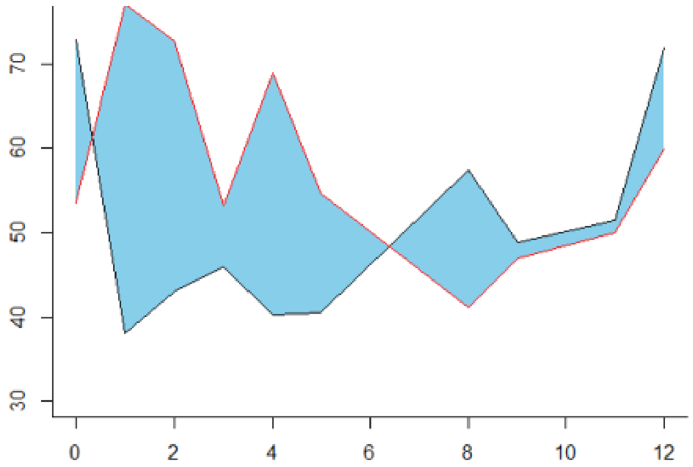
$$\log(X_{iA}/X_{iB}) =$$

$$(\log(X_{iA}/X_{iB})(t_{i1}), \log(X_{iA}/X_{iB})(t_{i2}), \ldots, \log(X_{iA}/X_{iB})(t_{iL_i}))$$

Log-ratio trajectories



Similar profiles = small area between the curves



Different profiles = large area between the curves

- ***Summary of log-ratio trajectories***

We summarize the log-ratio trajectories within two time points l_1 *and* l_2 *as:*

$$s_i(A,B) = \int_{l_1}^{l_2} log(X_{iA}/X_{iB})(t)dt$$

where the values of the log-ratio for t $\notin (t_{i1}, t_{i2}, \dots, t_{iL_i})$ *are linearly interpolated.*

Since the integral is linear:

$$s_i(A,B) = \int_{l_1}^{l_2} logX_{iA}(t)\, dt - \int_{l_1}^{l_2} logX_{iB}(t)\, dt$$

Thus, the number of integrals to be calculated is of the order of K, the number of taxa, instead of $K(K-1)/2$*, the number of pairwise log-ratios.*

- ***Pairwise log-ratio analysis in longitudinal studies***

The pairwise log-ratio summary for components A and B , $s(A, B)$, can be tested for association with the phenotype Y with a generalized linear model (glm) adjusted for some covariates Z:

$$g\left(E(Y)\right) = \ \beta_0 + \beta_1 s(A, B) + \gamma' \cdot Z$$

where Z = $(Z_1, Z_2, \dots, Z_r)$ *are non-compositional*

- ***Microbiome signature based on log-ratio analysis***

We consider glm penalized regression on the log-ratio summaries for all pairs of taxa:

$$g\big(E(Y)\big) = \beta_0 + \sum_{j \in J, (j_1, j_2) = J_{12}} \beta_j \cdot s(j_1, j_2)$$

The regression coefficients are estimated to minimize the loss function $L(\beta)$ *subject to a penalization on the regression coefficients,* $P(\beta)$

$$\hat{\beta} = \underset{\beta}{argmin}\{L(\beta) + P(\beta)\}$$

$P(\beta) = \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\beta\|_1$ *(elastic-net)*

- ***Microbiome signature based on log-ratio analysis***

For the linear regression model the loss function is given by the residual sum of squares

$$\hat{\beta} = \underset{\beta}{argmin}\{\|Y - S\beta\|_2^2 + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1\},$$

where S is the matrix of all log-ratio summaries and has dimension n by $K(K-1)/2$.

The result of the penalized optimization provides a set of ***selected pairs of taxa****, those with a non-null estimated coefficient.*

- ***Microbiome signature based on log-ratio analysis***

The linear predictor of the generalized linear model is the ***microbiome signature*** *associated with phenotype* Y:

$$M = \sum_{j \in J, (j_1, j_2) = J_{12}} \widehat{\beta}_j \cdot s(j_1, j_2) = \sum_{k=1}^{K} \hat{\alpha}_k \cdot \int_{l_1}^{l_2} logX_k(t)\, dt$$

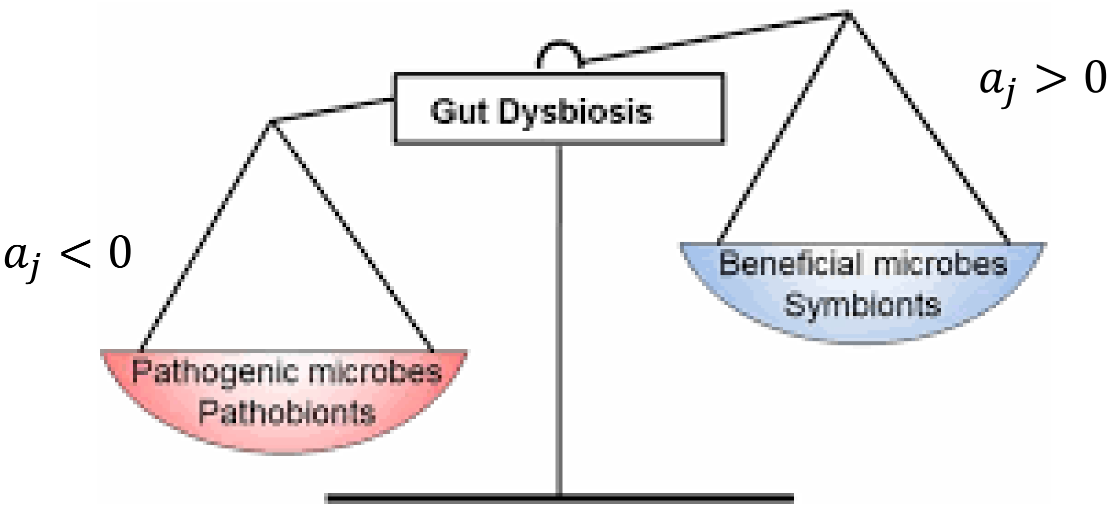
where $\hat{\alpha}_k = \sum_{j:k \in J_{12}(j)} \widehat{\beta}_j$

$$= \int_{l_1}^{l_2} \left(\sum_{k=1}^{K} \hat{\alpha}_k \cdot logX_k(t) \right) dt$$

- ***Microbiome signature based on log-ratio analysis***

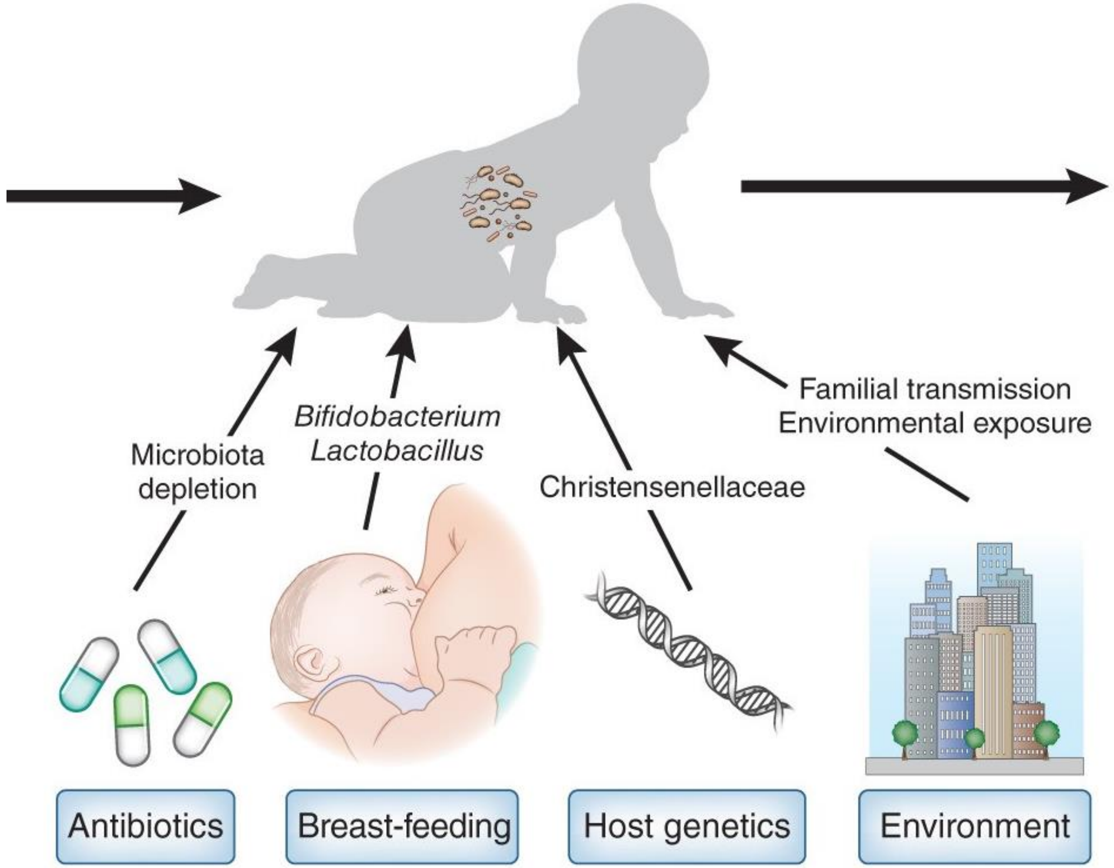
*Thus, the microbiome signature M is **the integral of the trajectory of a log-contrast function** involving the selected taxa:*

$$M = \int_{l_1}^{l_2} \left(\sum_{k=1}^{K} \hat{\alpha}_k \cdot log X_k(t) \right) dt \qquad with \ \ \textstyle\sum_{k=1}^{K} \hat{\alpha}_k = 0.$$



Early childhood and the microbiome (ECAM) study

n=42 individuals (30 BD, 12 FD)
Y = diet (BD vs FD)
X = microbiome 36 genera



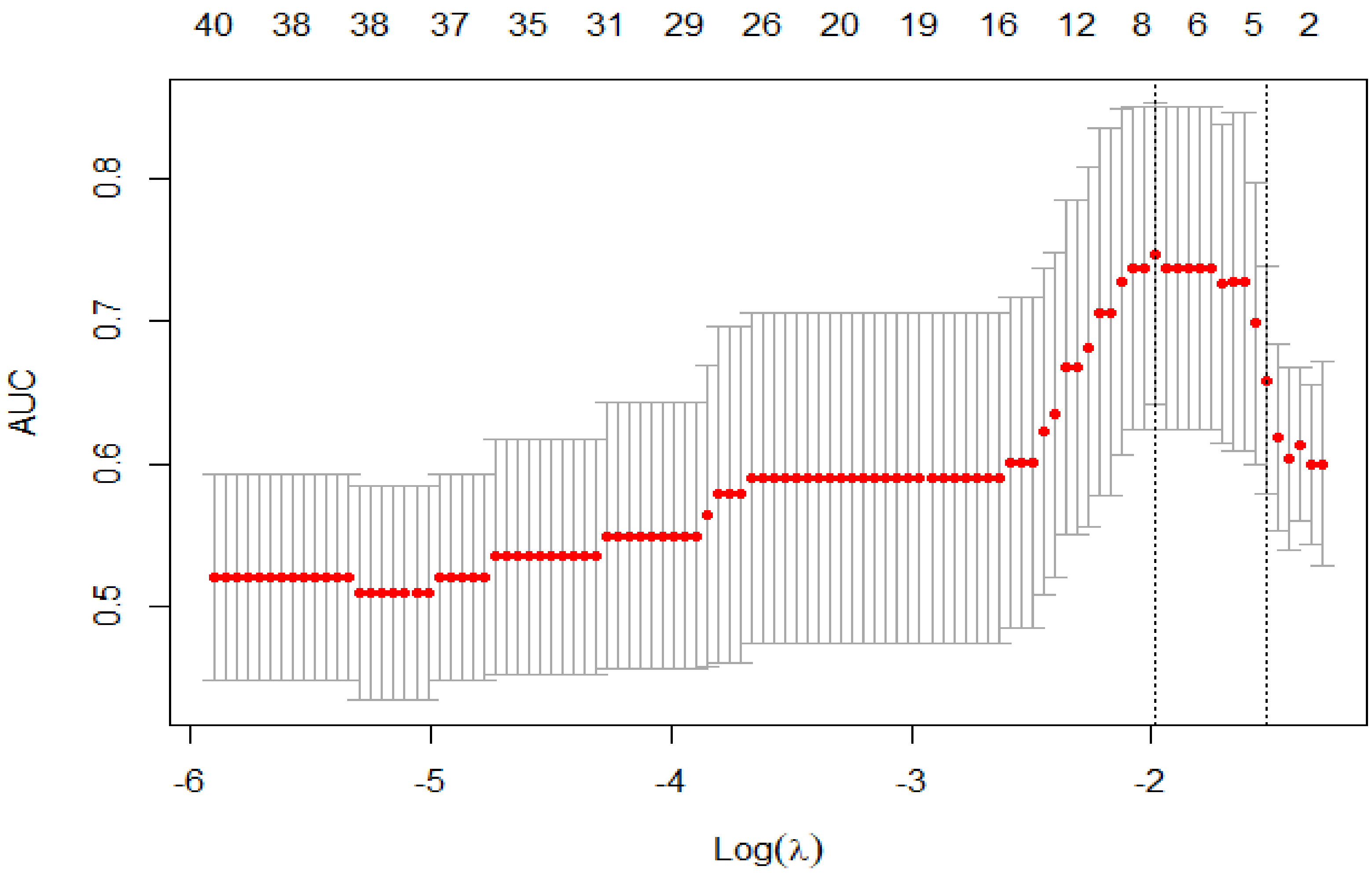
From Tamburini et al. 2016

Early childhood and the microbiome (ECAM) study

n=42 individuals (30 BD, 12 FD)
Y = diet (BD vs FD)
X = microbiome 36 genera

`coda4microbiome::coda_glmnet_longitudinal`

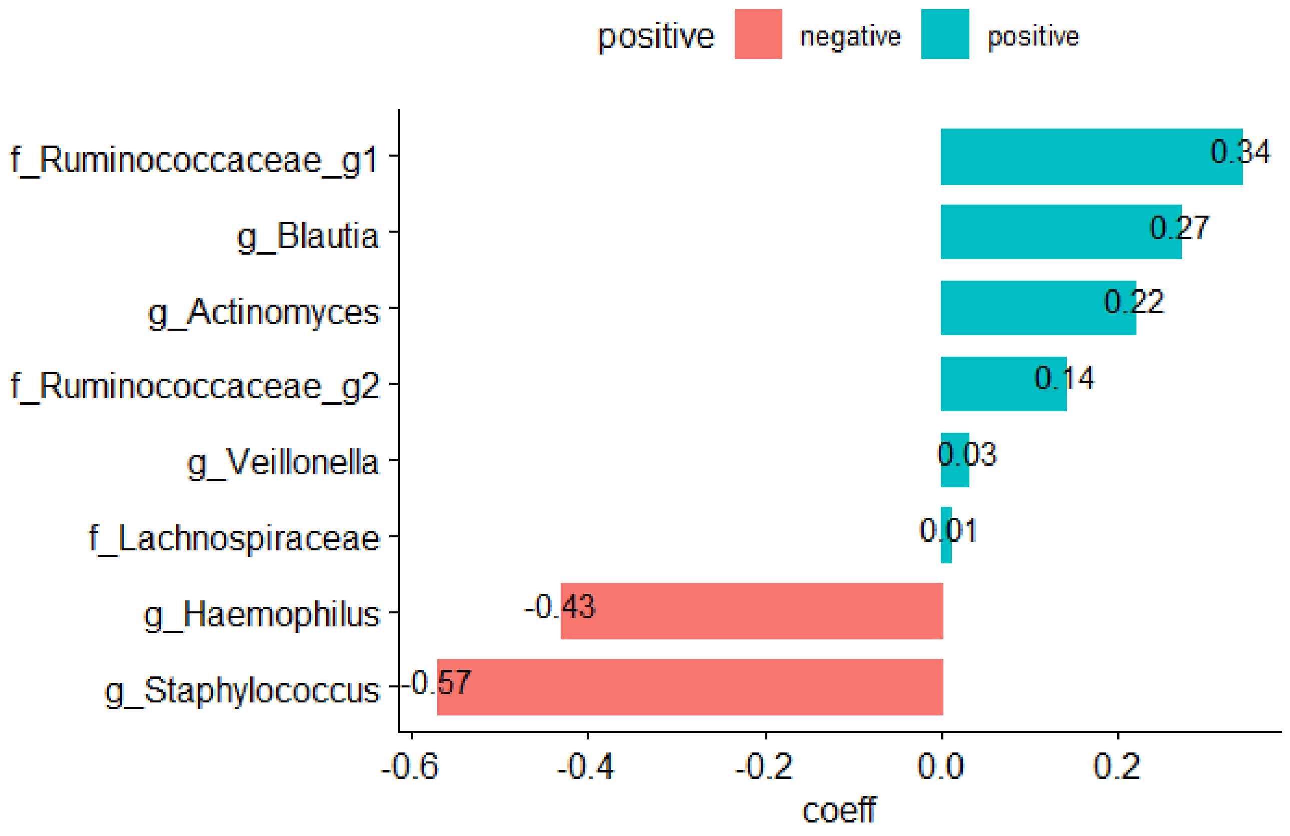
Penalization:
Optimal number of log-ratios to retain?



Early childhood and the microbiome (ECAM) study

n=42 individuals (30 BD, 12 FD)
Y = diet (BD vs FD)
X = microbiome 36 genera

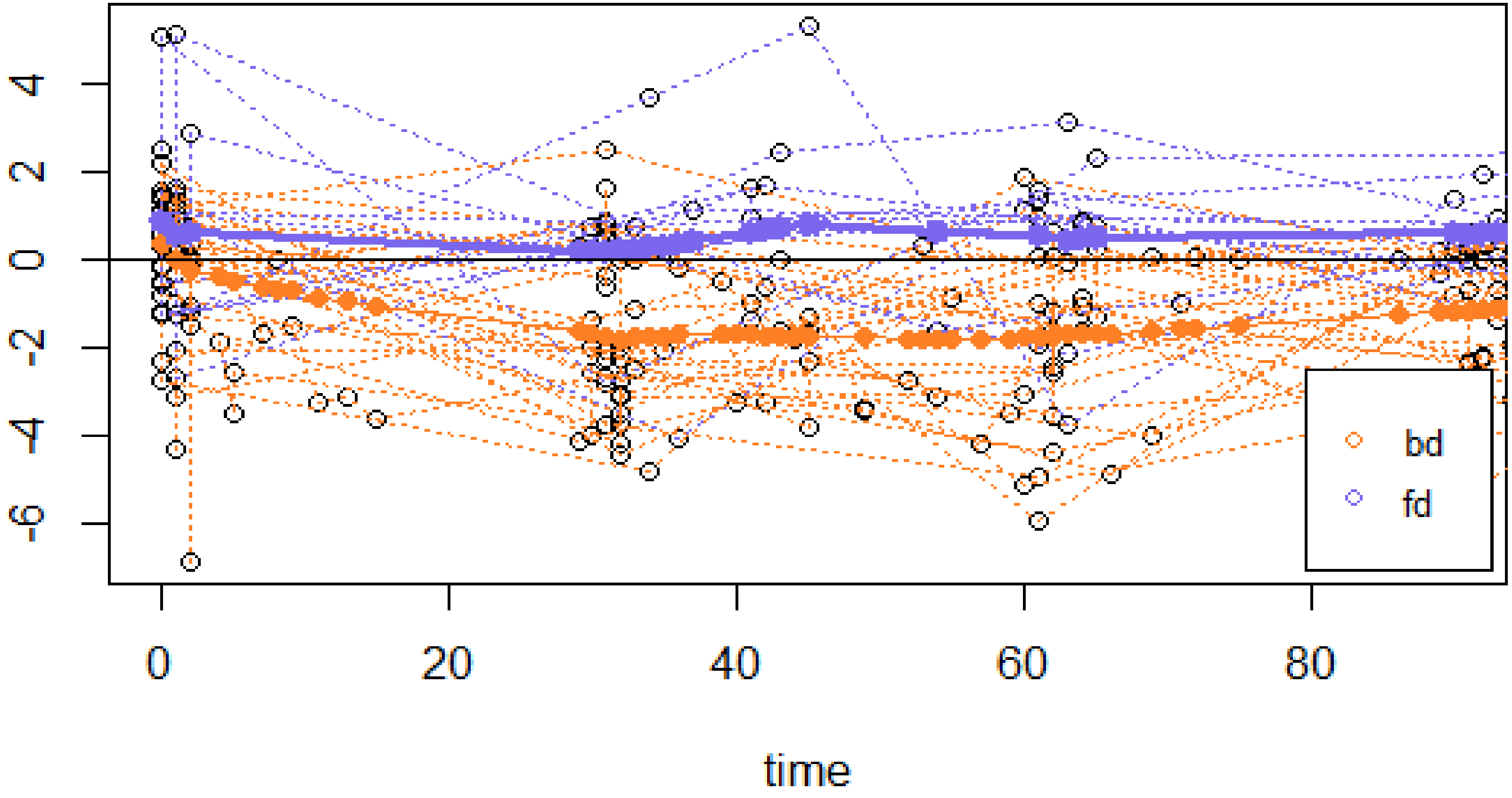
Selected taxa:



Early childhood and the microbiome (ECAM) study

n=42 individuals (30 BD, 12 FD)
Y = diet (BD vs FD)
X = microbiome 36 genera

Signature trajectories:



Early childhood and the microbiome (ECAM) study

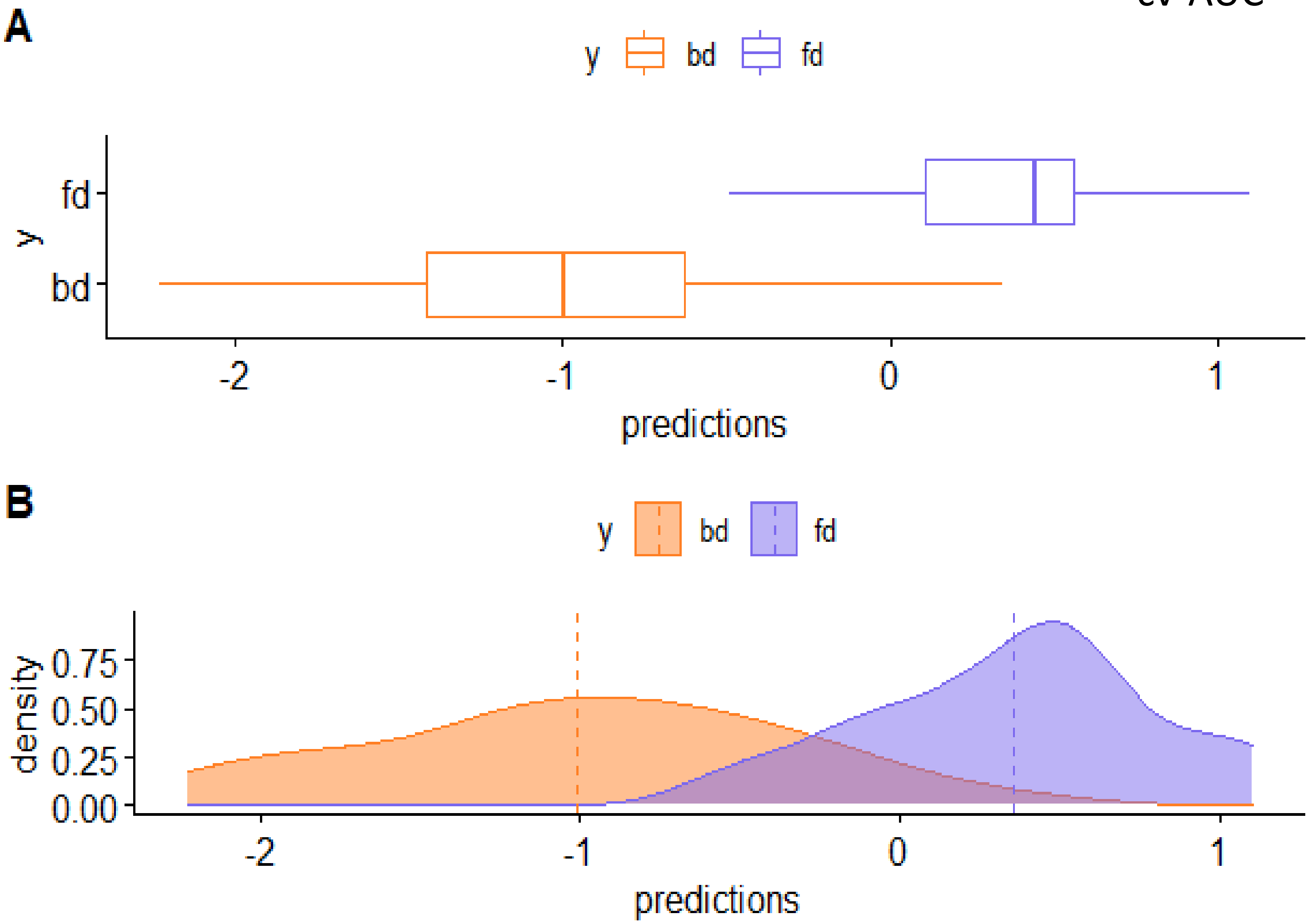
n=42 individuals (30 BD, 12 FD)

Y = diet (BD vs FD)

X = microbiome 36 genera

Discrimination accuracy:

cv-AUC = 0.74 (sd=0.10)


A
y
bd
fd
fd
bd
y
predictions
B
y
bd
fd
density
predictions