

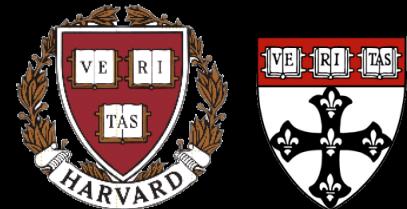


Himel
Mallick



Microbial community statistics and MaAsLin2

Curtis Huttenhower



Harvard T.H. Chan School of Public Health
Department of Biostatistics

07-31-19



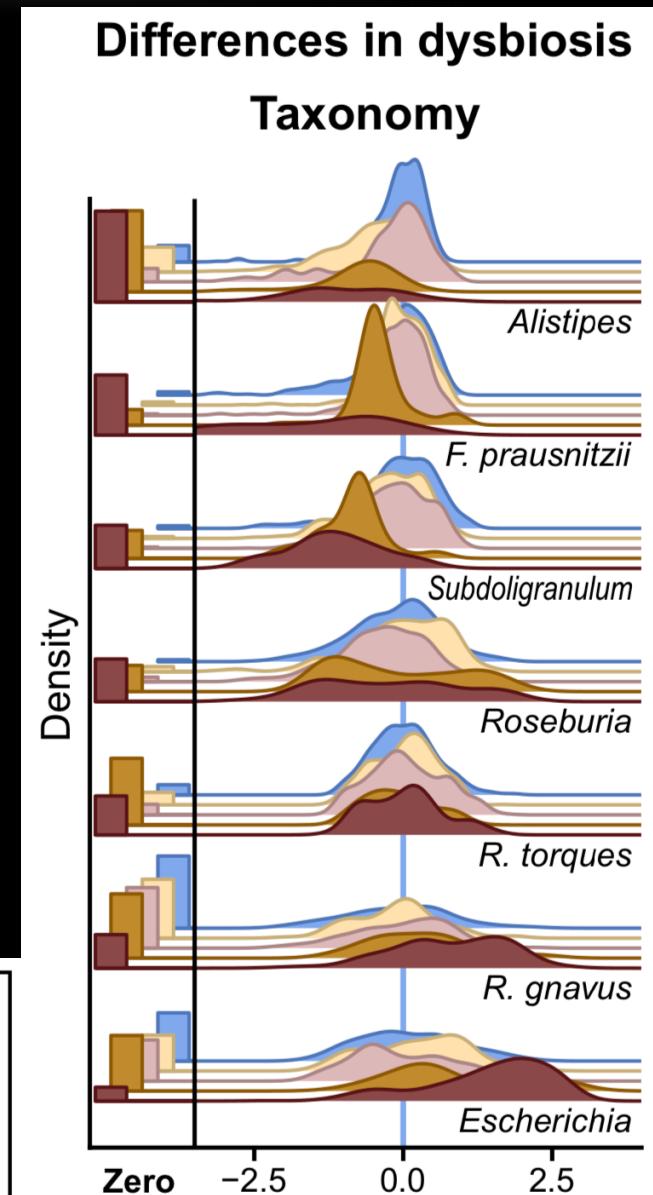
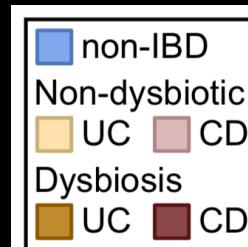
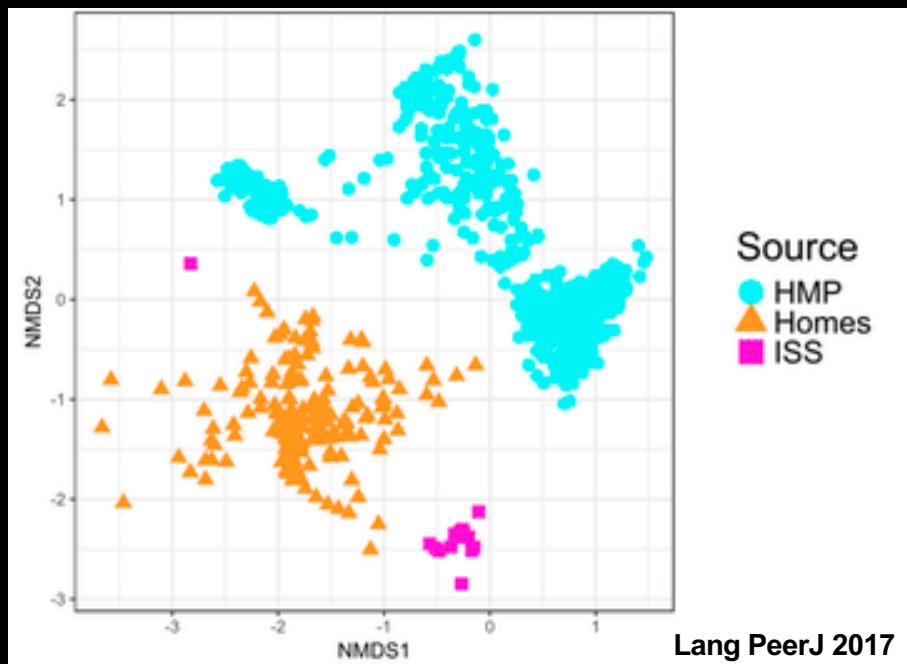


Statistical association discovery in microbial community research

- Omnibus testing
 - **Hypothesis:** Overall community structure (i.e. beta diversity relationships) changes with respect to some exposure / environment / phenotype.
 - Typically fewer tests, but can be difficult to interpret.
- Per-feature (feature-wise) testing
 - **Hypothesis:** There are specific taxa with significantly differential abundance relative to the metadata of interest.
 - Typically many tests, but provides more actionable information.



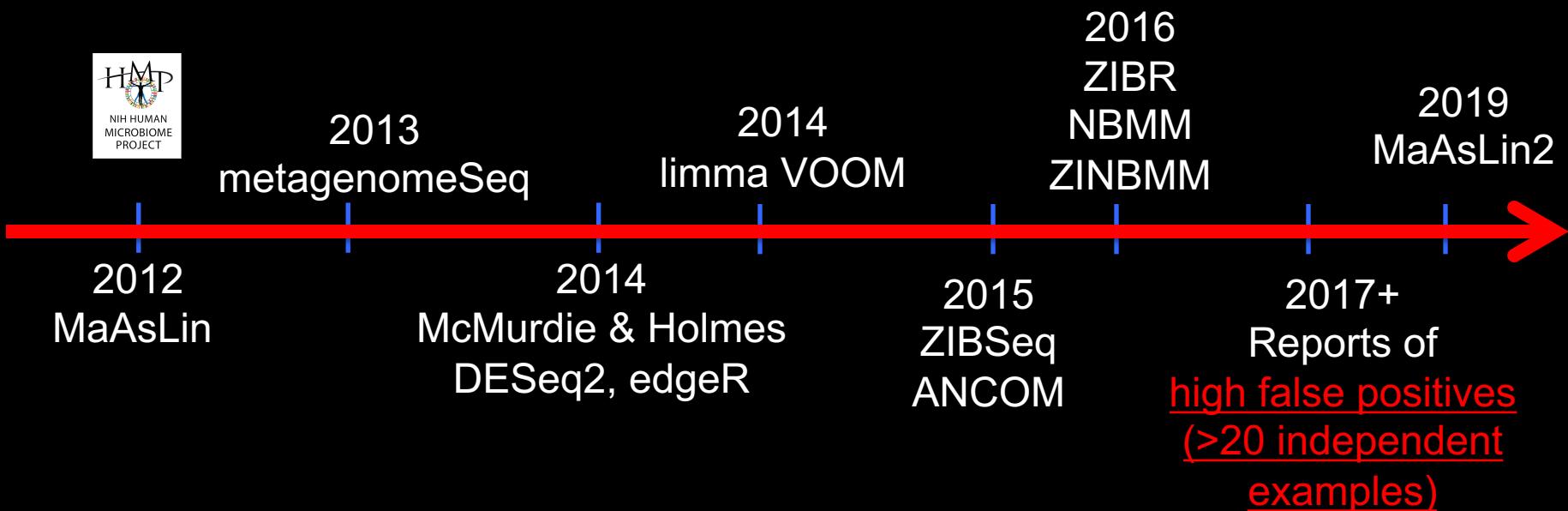
Statistical association discovery in microbial community research





Multivariable association in microbial community studies

Current State-of-the-art



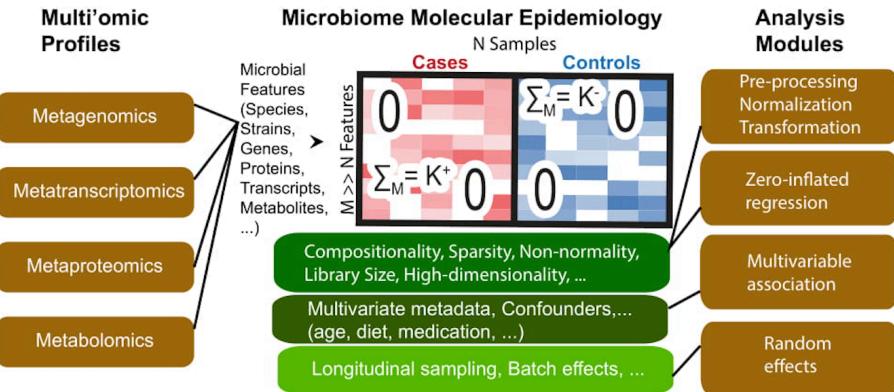
Lack of multivariable, repeated measures methods

Lack of comprehensive benchmarking

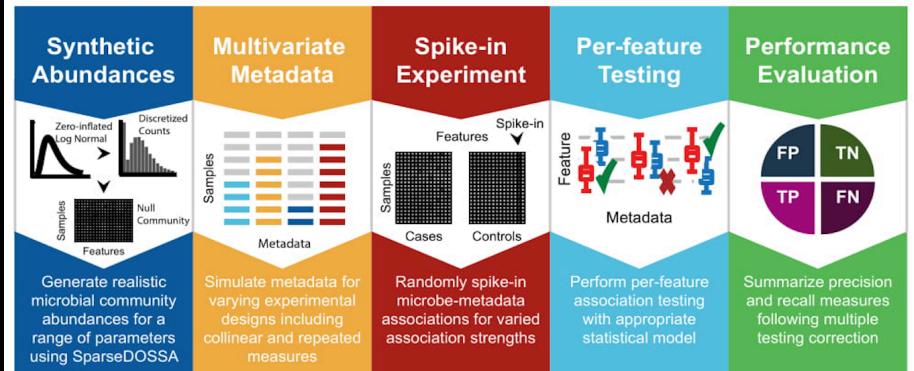


MaAsLin2 methodology and validation

A. MaAsLin2 - A framework for microbiome epidemiology



B. Comprehensive benchmarking using synthetic abundances





MaAsLin2 methodology and validation

A. MaAsLin2

Multi'omic
Profiles

Metagenom

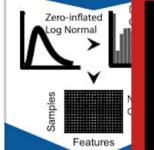
Metatranscri

Metaproteo

Metabolom

B. Comparison

Synthetic
Abundances



Generate reali
microbial comm
abundances fo
range of param
using SparseDC

Association
strength governed
by **linear effect size**
parameter, 100
replications per
dataset and method

10% microbial
features and 20%
metadata features
**(for multivariable
metadata design)**
are randomly
spiked-in

Methodology

Analytical Mo

Pre-pro

Norm

Trans

Zero

regre

Mult

asso

Ran

ef

abundanc

Perfo

Evalu

ture

re

Testing

Summar

and rec

followin

testin

g

ate

riate

model

Summary

and rec

followin

testin

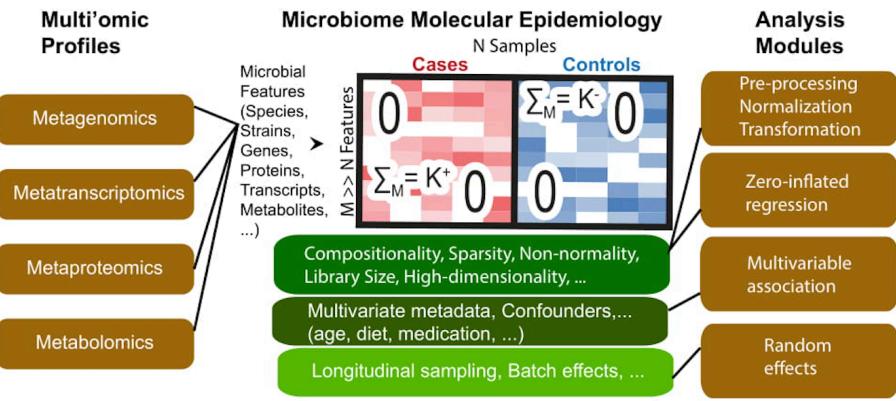
g

Parameter	Values
Zero-inflation	TRUE, FALSE
Random effect	TRUE, FALSE
Metadata design	Univariate, Continuous Univariate, Binary Multivariable, Independent Multivariable, Collinear
Number of subjects	10, 20, 50, 100, 200
Number of repeated measures	2, 5, 10, 20
Number of microbes	100, 200, 500
Number of metadata (multivariable design)	5, 10, 20
Effect size	1, 2, 5, 10
Read depth	10K, 50K, 100K



MaAsLin2 methodology and validation

A. MaAsLin2 - A framework for microbiome epidemiology



B. Comprehensive benchmarking using synthetic abundances



- Linear Model (LM)
- Linear Model with Depth Covariate (LM2)
- Compound Poisson Linear Model (CPLM)
- Negative Binomial GLM
- Zero-inflated Negative Binomial
- Zero-inflated Beta
- Zero-inflated CPLM (ZICP)

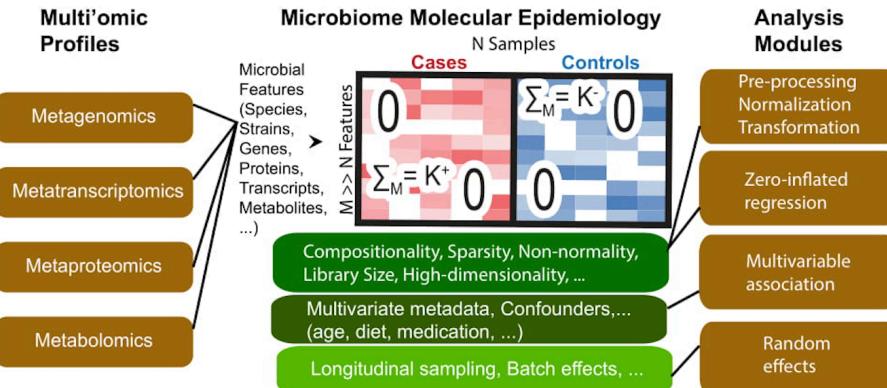
More than 75 methods from all possible legitimate combinations, between 10-10,000 tests per dataset and method

- DESeq2
- edgeR
- ANCOM
- limma
- limma2
- limma VOOM
- metagenomeSeq
- metagenomeSeq2
- Spearman
- Wilcoxon

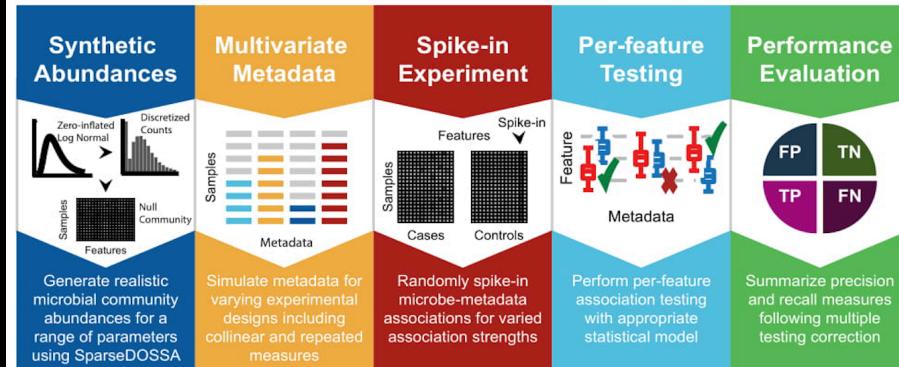


MaAsLin2 methodology and validation

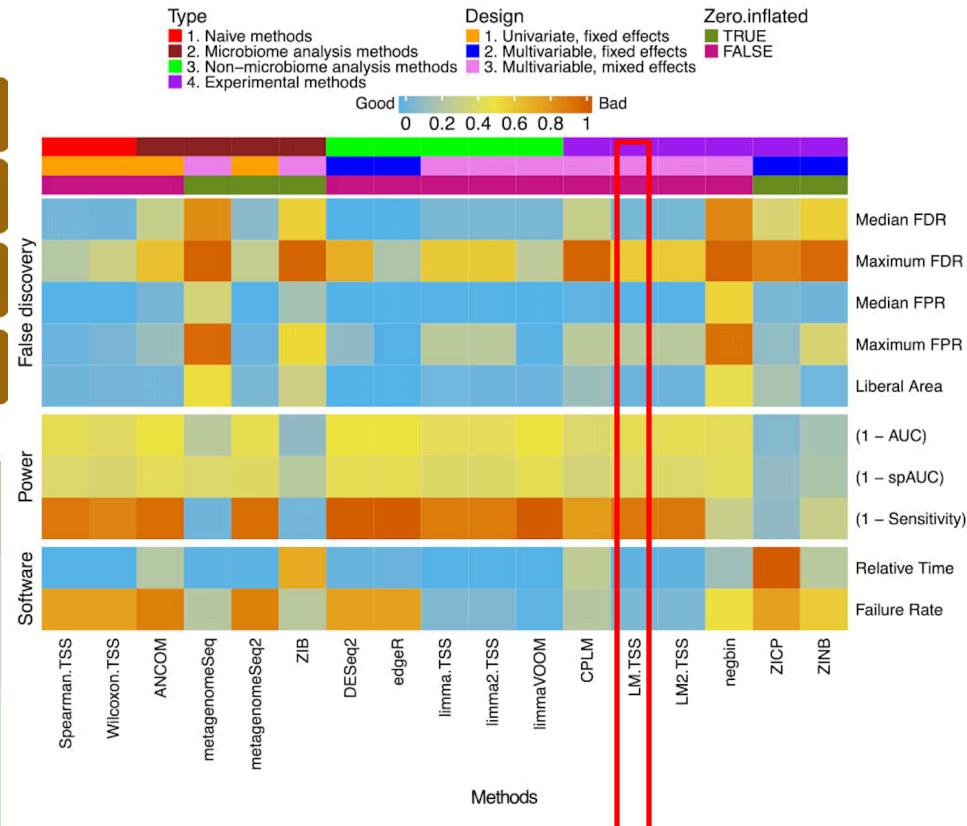
A. MaAsLin2 - A framework for microbiome epidemiology



B. Comprehensive benchmarking using synthetic abundances

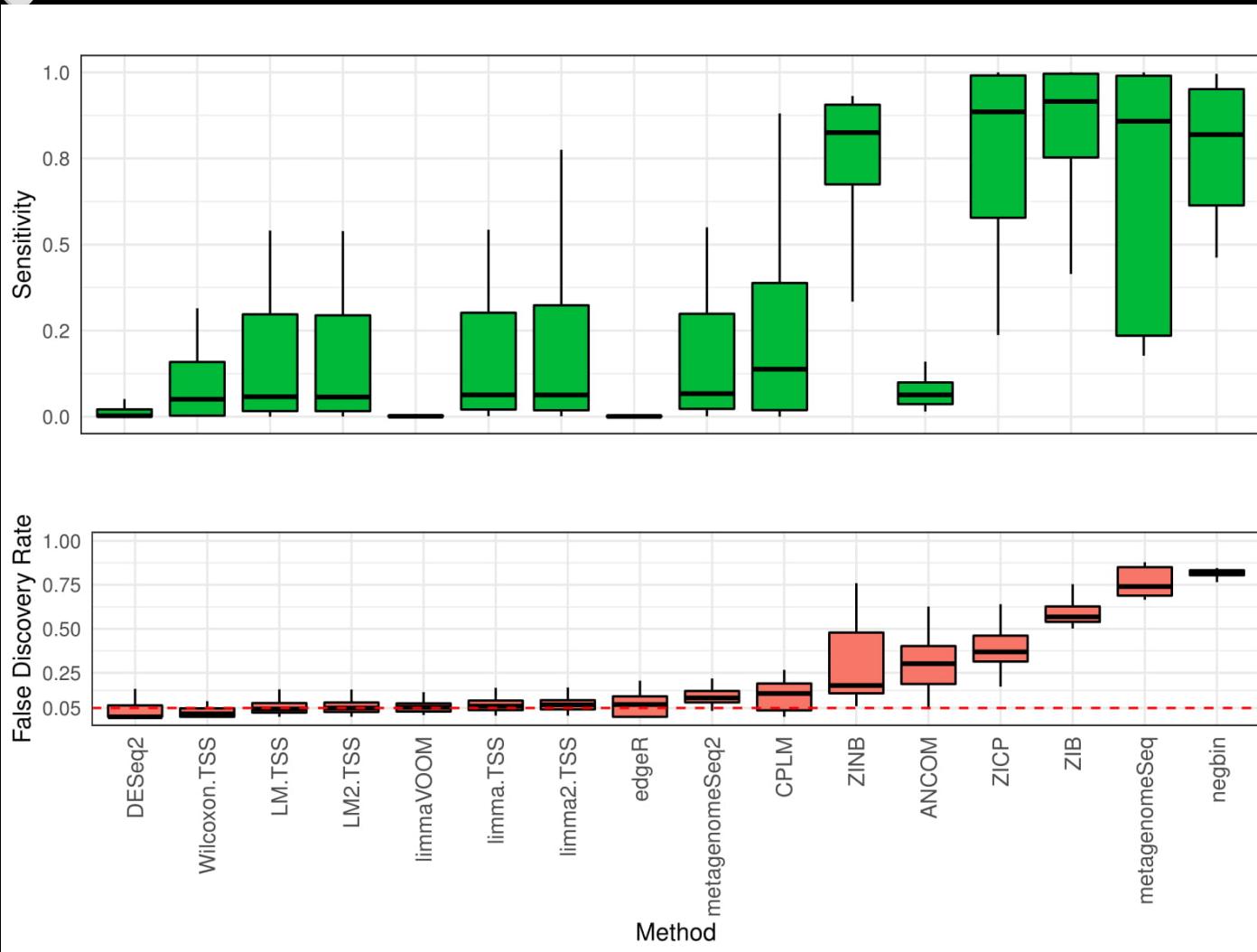


C. Large-scale evaluation of multivariable association methods





MaAsLin2 controls false discovery rate while maintaining power in differential abundance analysis

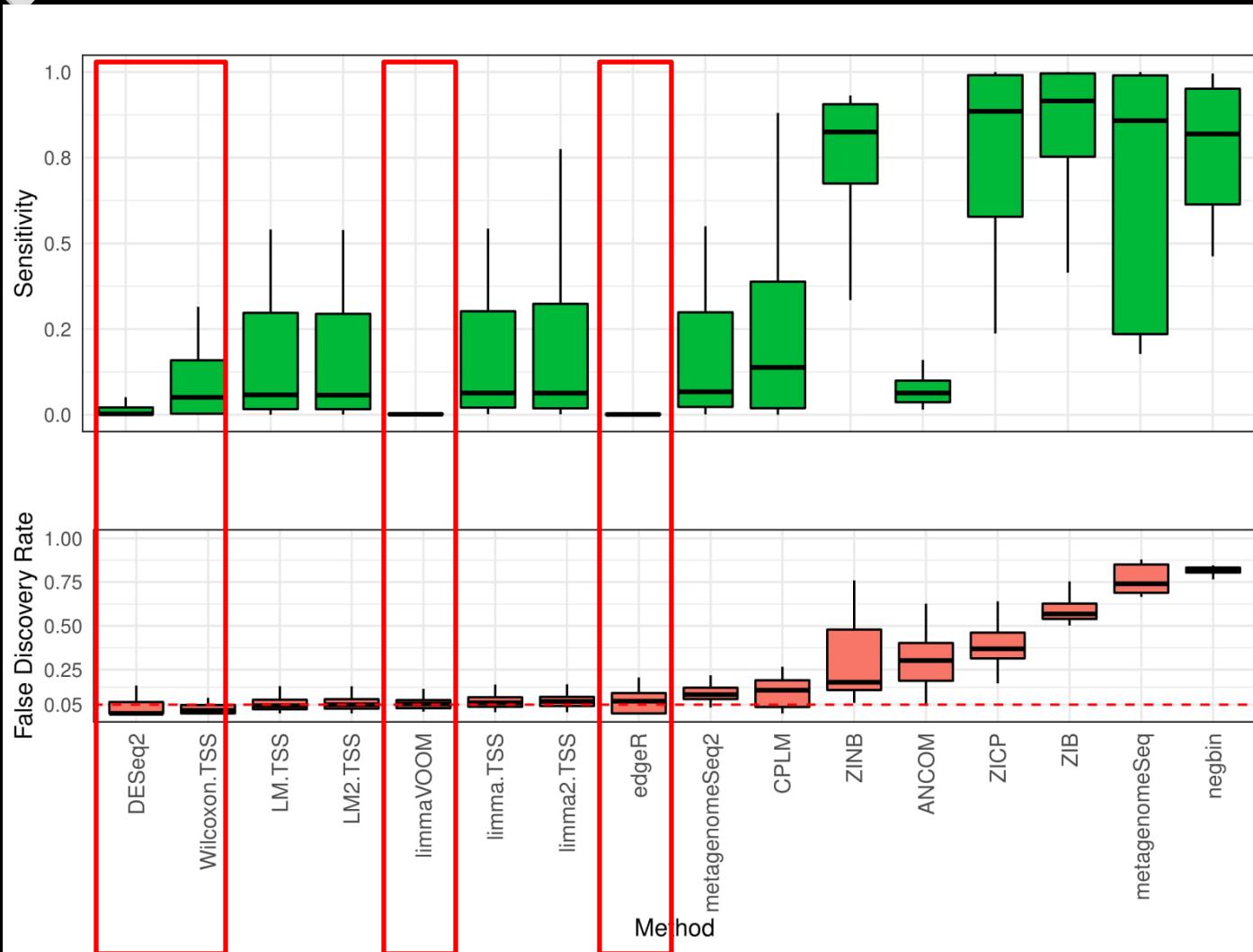


Four major patterns:

- (i) low-sensitivity, low-FDR,
- (ii) low-sensitivity, high-FDR,
- (iii) high-sensitivity, high-FDR,
- and
- (iv) high-sensitivity, low-FDR



MaAsLin2 controls false discovery rate while maintaining power in differential abundance analysis

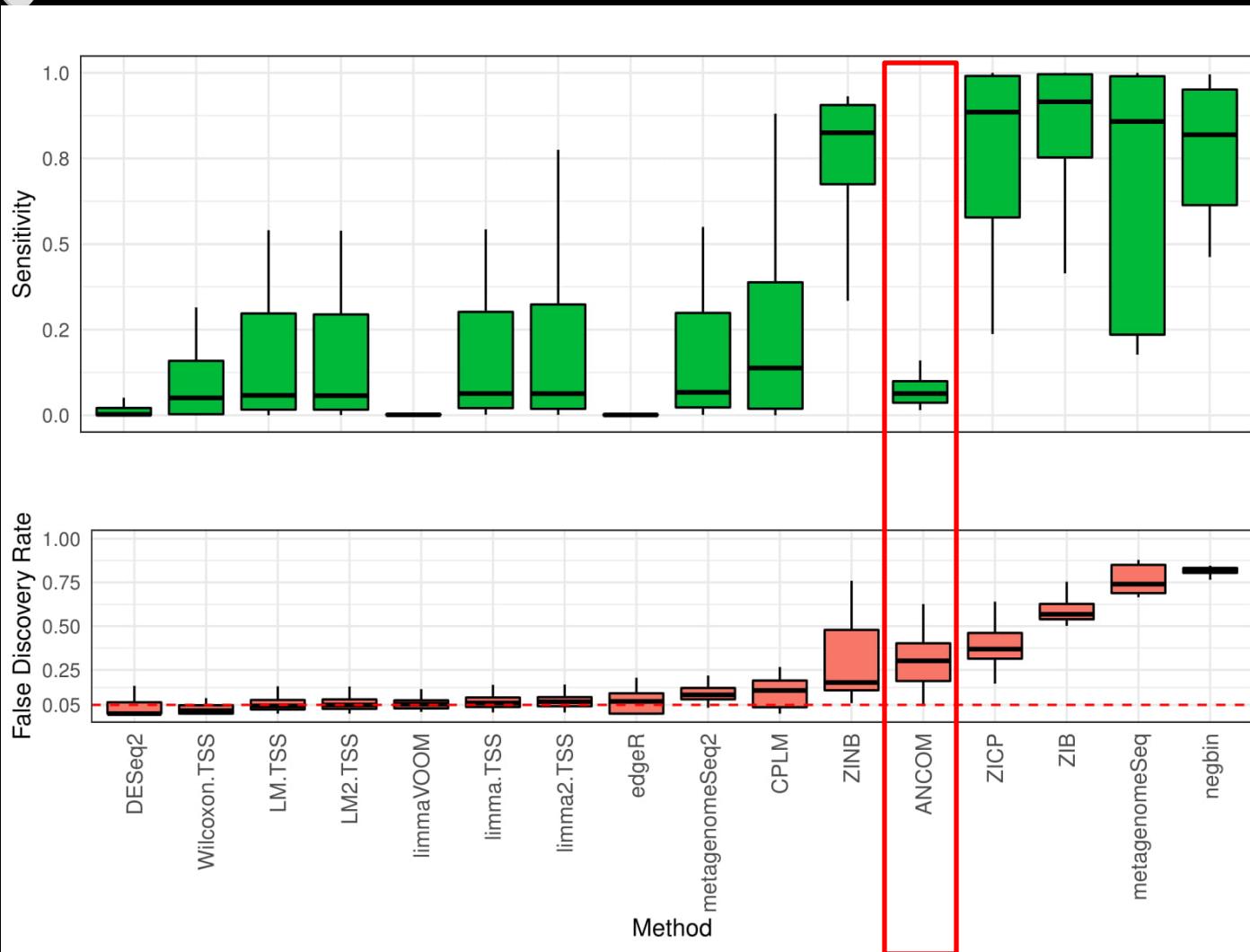


Four major patterns:

- (i) low-sensitivity, low-FDR,
- (ii) low-sensitivity, high-FDR,
- (iii) high-sensitivity, high-FDR,
- and
- (iv) high-sensitivity, low-FDR



MaAsLin2 controls false discovery rate while maintaining power in differential abundance analysis

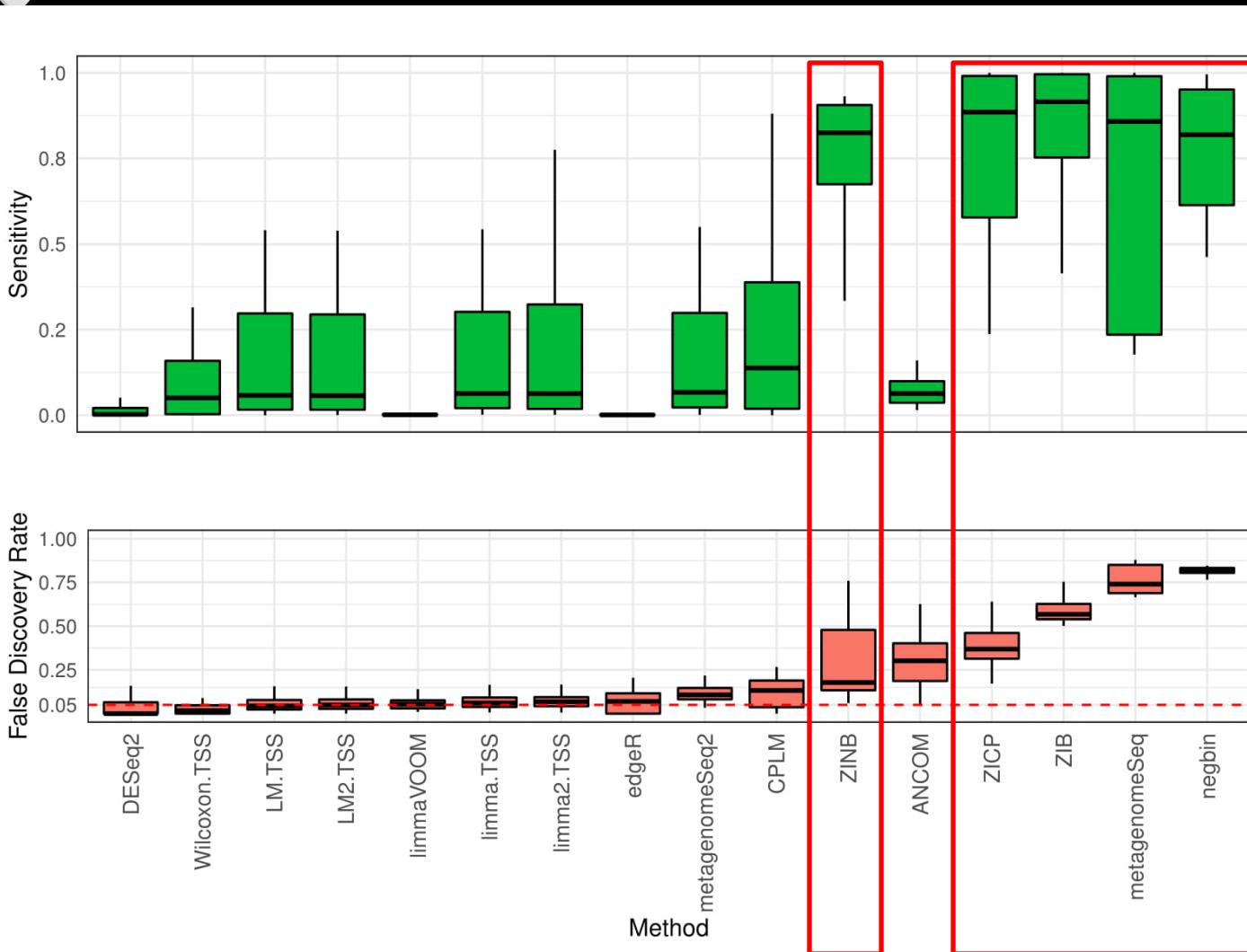


Four major patterns:

- (i) low-sensitivity, low-FDR,
- (ii) low-sensitivity, high-FDR,
- (iii) high-sensitivity, high-FDR,
- and
- (iv) high-sensitivity, low-FDR



MaAsLin2 controls false discovery rate while maintaining power in differential abundance analysis

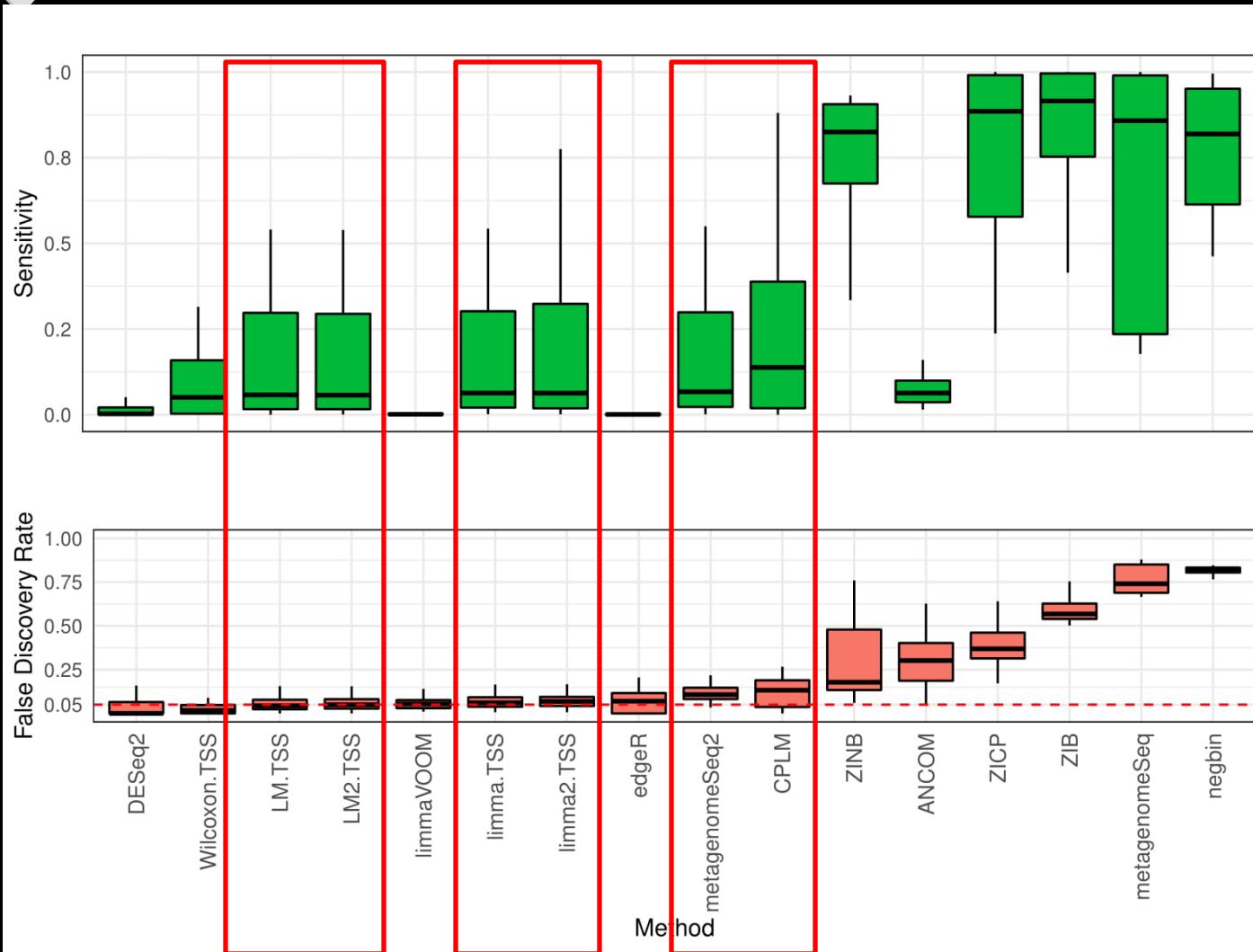


Four major patterns:

- (i) low-sensitivity, low-FDR,
- (ii) low-sensitivity, high-FDR,
- (iii) high-sensitivity, high-FDR,
- and
- (iv) high-sensitivity, low-FDR



MaAsLin2 controls false discovery rate while maintaining power in differential abundance analysis



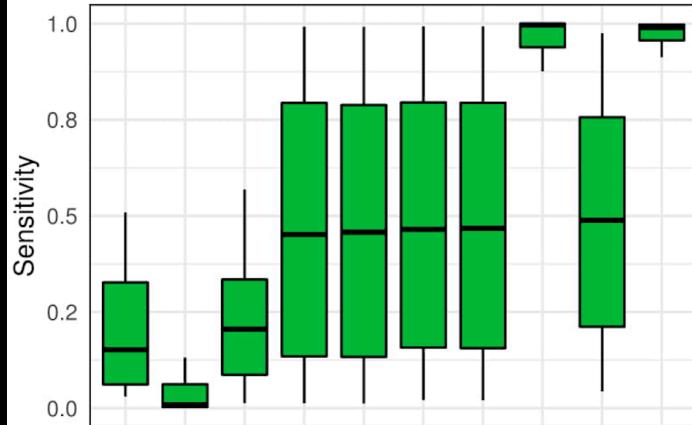
Four major patterns:

- (i) low-sensitivity, low-FDR,
- (ii) low-sensitivity, high-FDR,
- (iii) high-sensitivity, high-FDR,
- and
- (iv) high-sensitivity, low-FDR

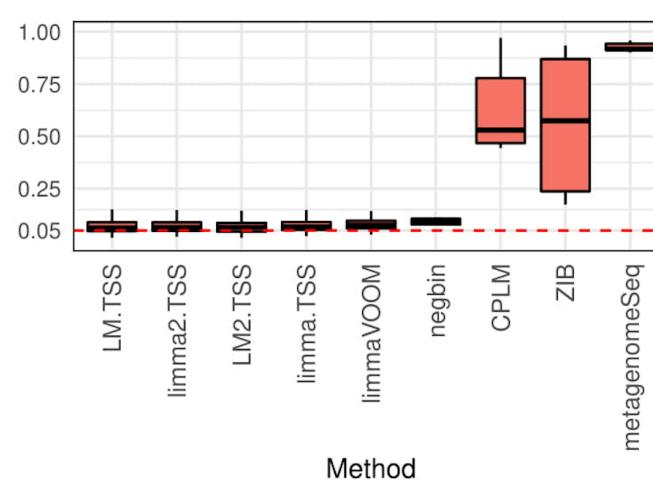
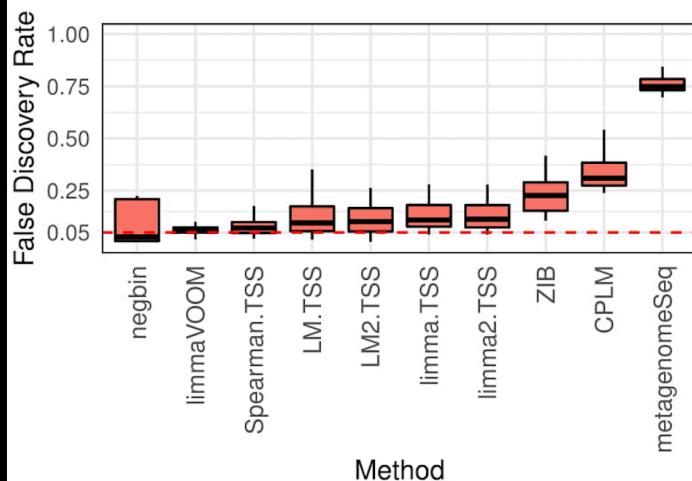
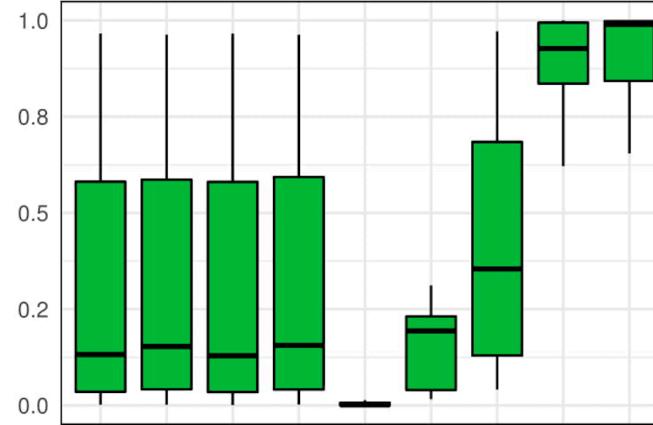


Still true for multiple covariates, repeated measures (random effects)

A Univariate, Continuous



B Multivariable, Mixed

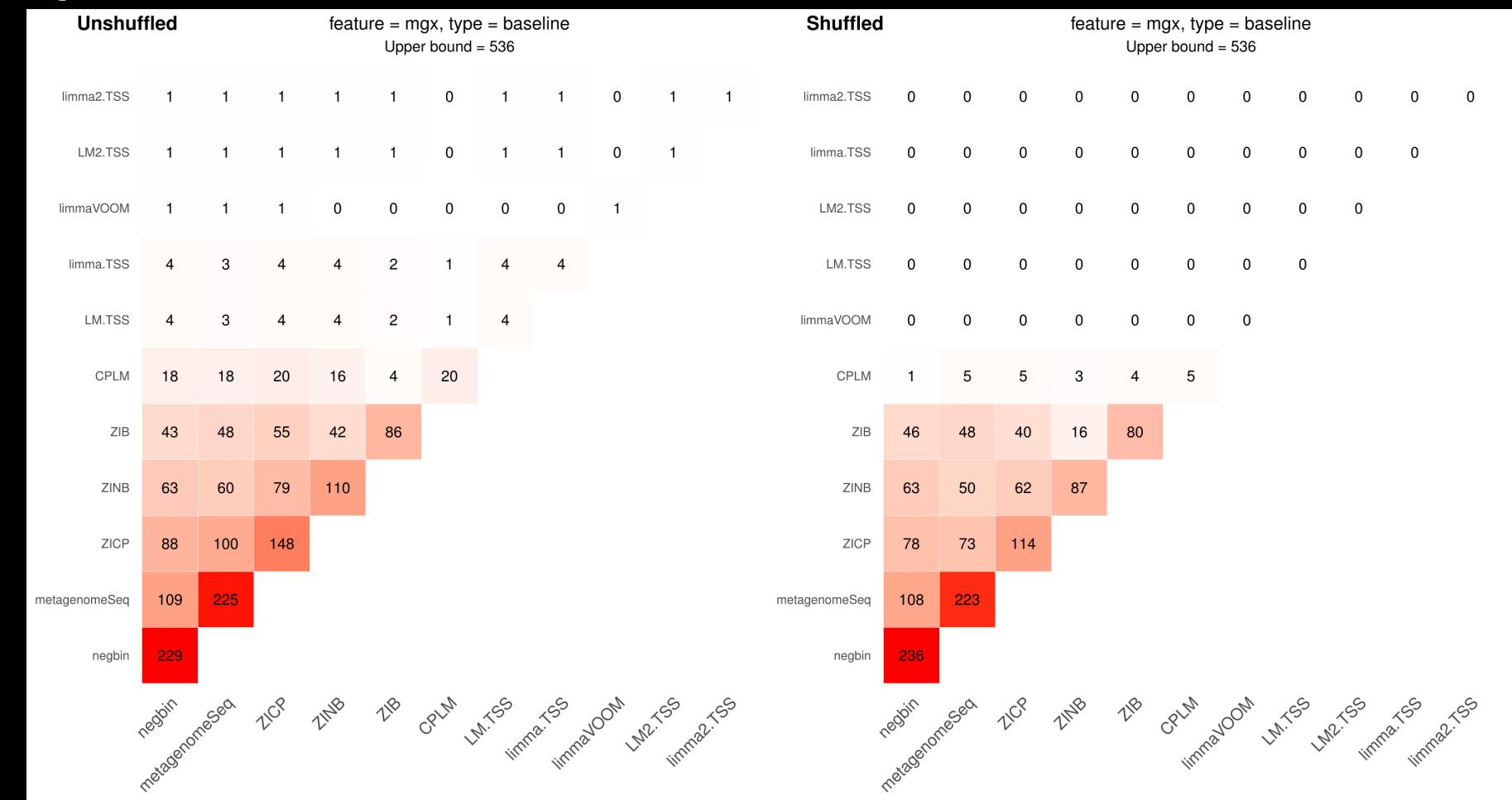


Again,
similar trend
for repeated
measures,
both for
univariate
and
multivariable
cases.

Results hold
for **varying
number of
repeated
measures.**

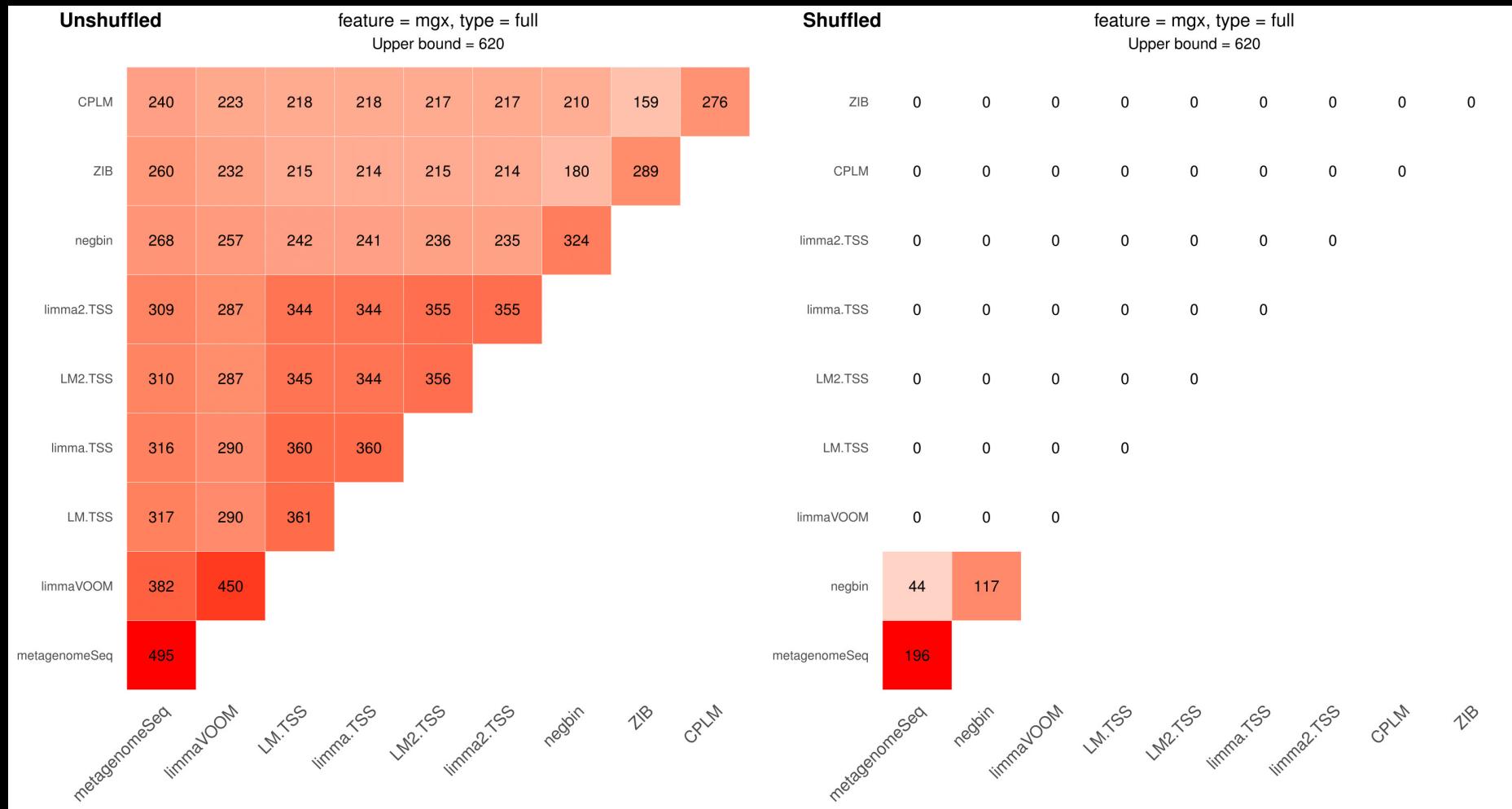


Association discovery in the HMP2 (baseline time point, IBD diagnosis)



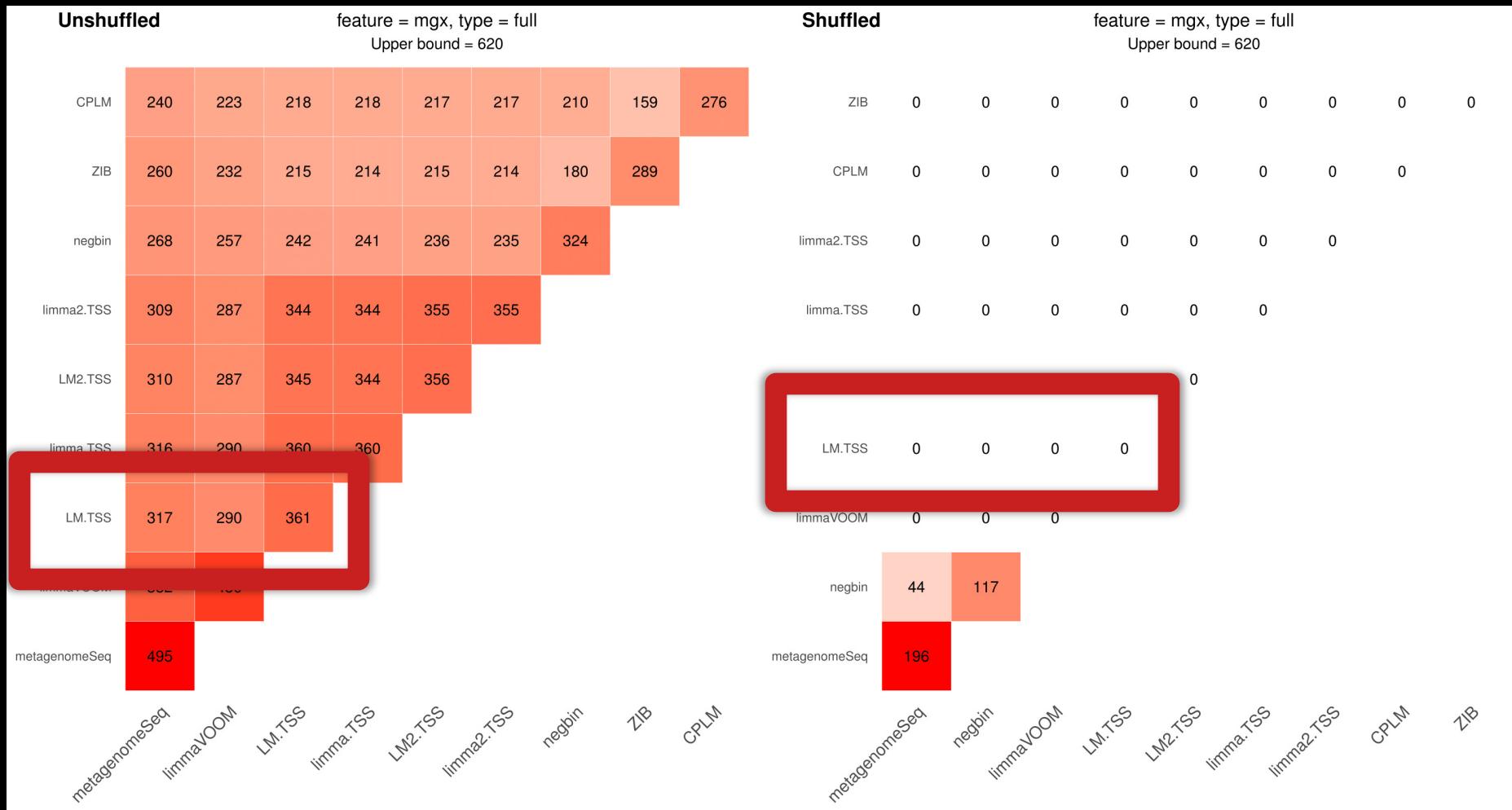


Association discovery in the HMP2 (all time points, IBD dysbiosis)



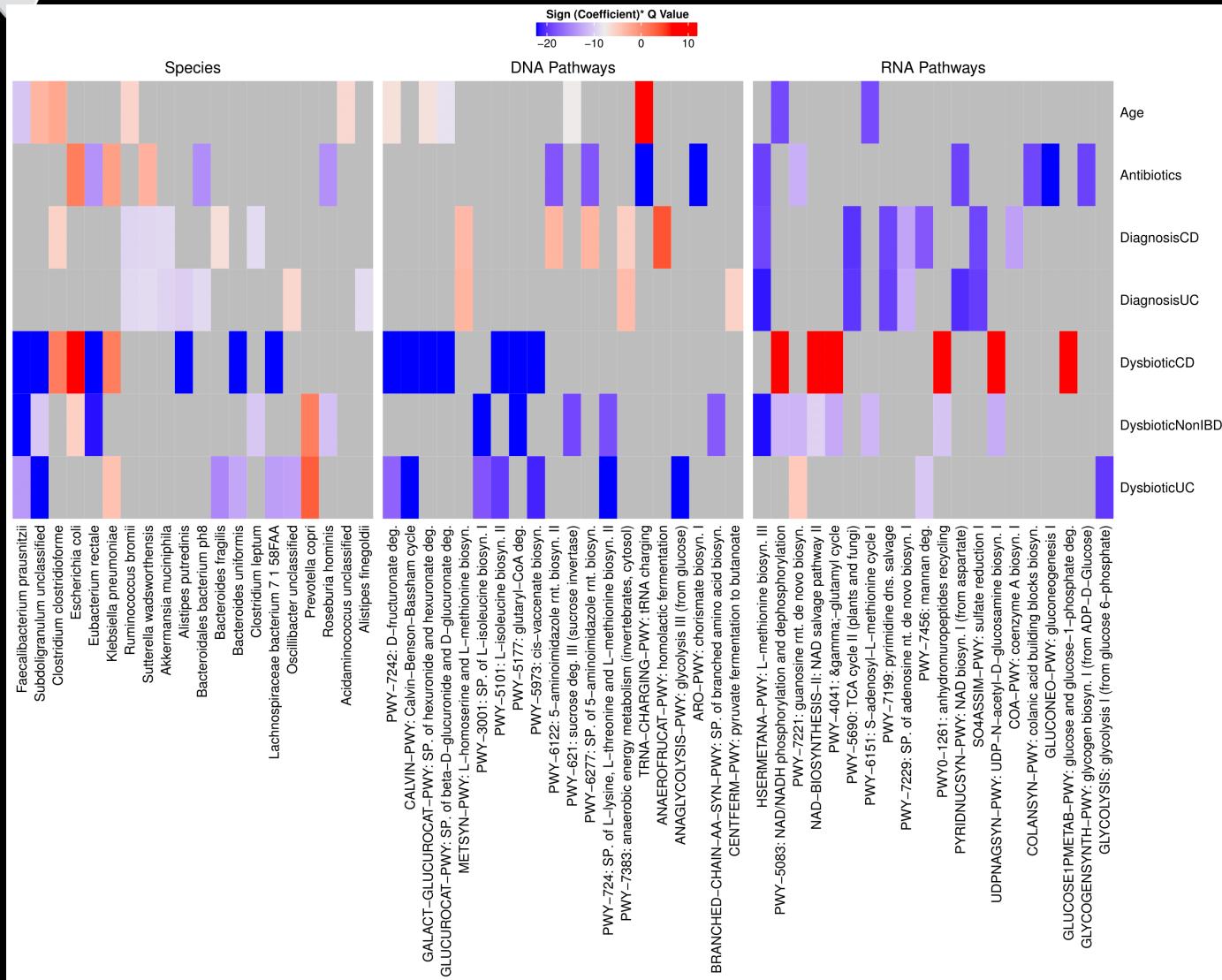


Association discovery in the HMP2 (all time points, IBD dysbiosis)





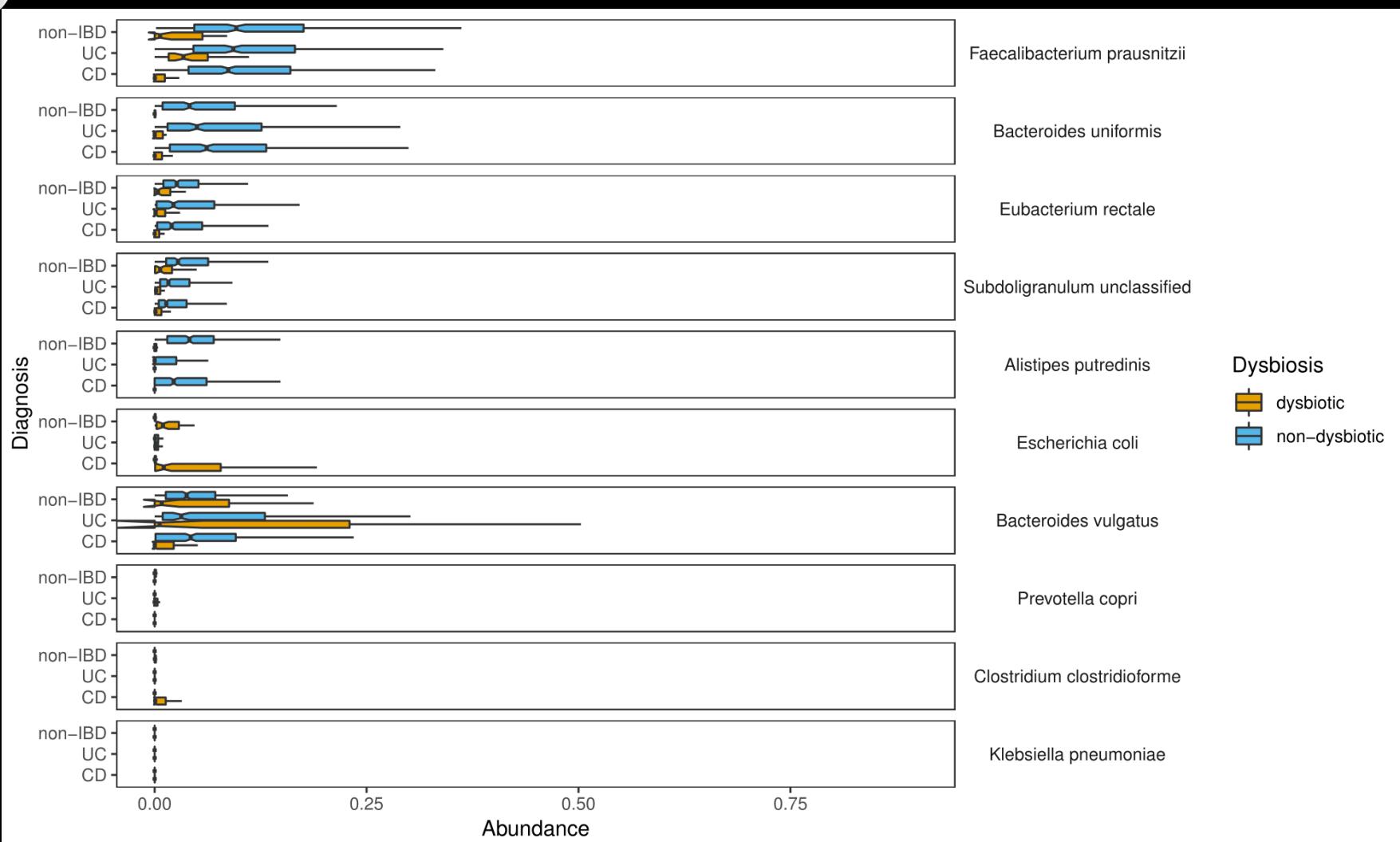
MaAsLin2 multi'omic associations from the Integrated Human Microbiome Project



Increase in facultative anaerobes including *E. coli* during inflammation at the expense of anaerobic SCFA producers including *F. prausnitzii* and *R. hominis*.

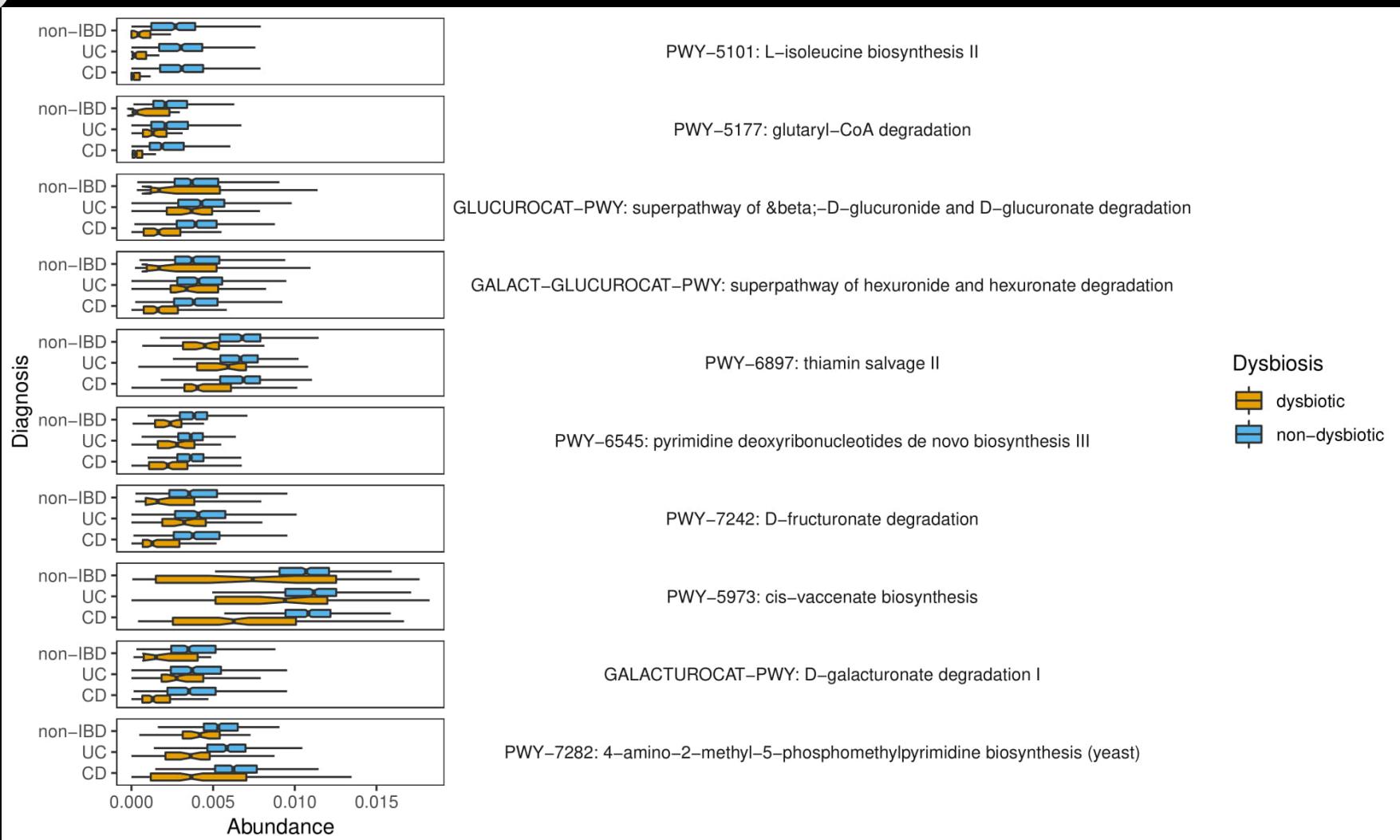


Top Associations (Species)



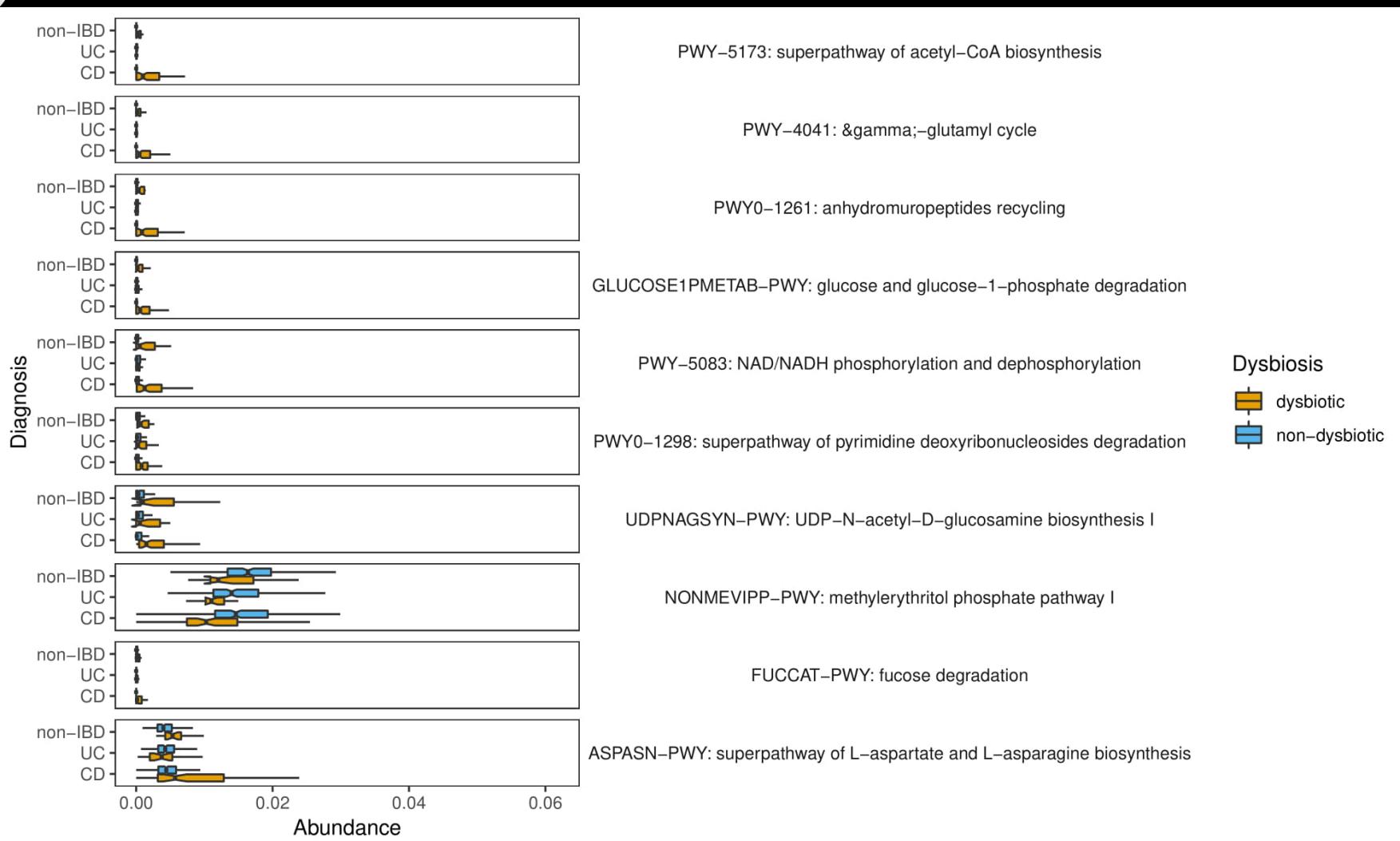


Top Associations (DNA Pathways)





Top Associations (RNA Pathways)





Summary of Findings

- **MaAsLin2** for microbial community association testing based on simple and highly flexible linear models and validated the method using realistic synthetic data.
- **Alarming rate of false positives** for several state-of-the-art methods including zero-inflated and count models, expanding indications from the literature.
- Application to HMP2 dataset reproduced **established and novel ecological results** for multiple feature types.
- **Software and preprint available:**
 - <http://huttenhower.sph.harvard.edu/maaslin2>
(first hit in Google with keyword ‘maaslin2’)



MaAsLin2 tutorial

<https://bitbucket.org/biobakery/biobakery/wiki/maaslin2>

MaAsLin

Associate arbitrarily
complex metadata with
microbiome features