European Molecular Biology Laboratory

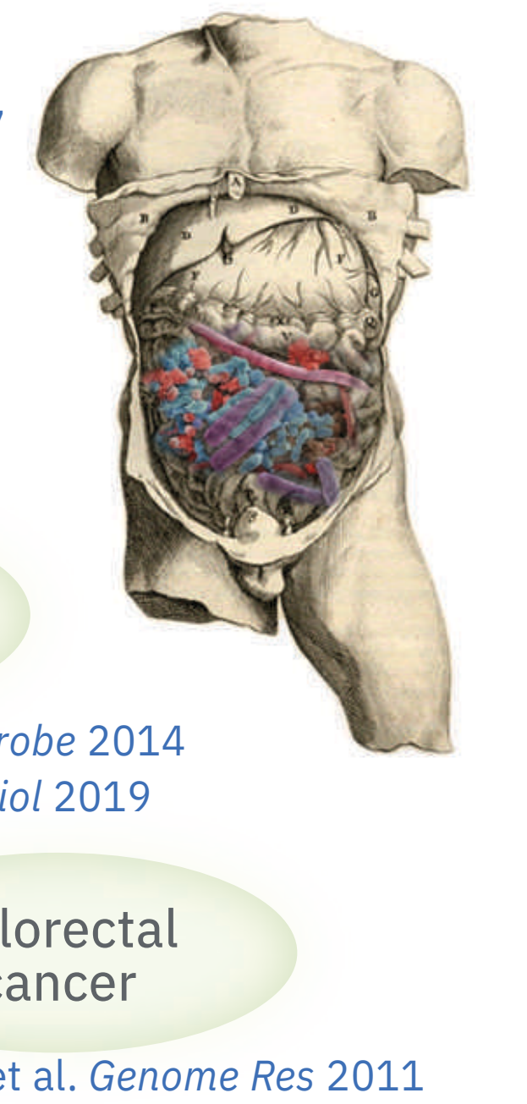
SIAMCAT: user-friendly and versatile machine learning workflows for statistically rigorous microbiome analyses

Jakob Wirbel[1], Konrad Zych[1], Morgan Essex[1], Nicolai Karcher[1], Ece Kartal[1], Guillem Salazar[2], Peer Bork[1], Shinichi Sunagawa[2], Georg Zeller[1]

1) Structural and Computational Biology Unit, EMBL Heidelberg, Germany; 2) Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Switzerland

Introduction

- Changes in microbiome composition are associated with many common conditions
- Microbiome data are extensively mined for biomarkers with diagnostic or therapeutic potential
- Microbiome data analysis presents several challenges, since the data are:
  - not normally distributed,
  - zero-inflated,
  - compositional
- SIAMCAT is an R package using machine learning to infer associations between microbial communities and host phenotypes



Cardiovascular disease

Wang et al. *Nature* 2011
Jie et al. *Nat Commun* 2017

Liver cirrhosis & cancer

Yoshimoto et al. *Nature* 2013
Qin et al. *Nature* 2014

Inflammatory bowel disease

Gevers et al. *Cell Host Microbe* 2014
Franzosa et al. *Nat Microbiol* 2019

Colorectal cancer

Kostic et al. *Genome Res* 2011
Castellarin et al. *Genome Res* 2011
Wirbel et al. *Nat Med* 2019

Comparison to other tools

Tool	Pros	Cons	Assoc. test	ML model
LEfSe Segata et al. *Genome Biol* 2011	• Widely used • Visualizations • Multiclass	• Assumes normal distributions • No predictions or data preprocessing	✓	✗
metagenomeSeq Paulson et al. *Nat Methods* 2013	• On Bioconductor • Multiclass • Attemps to model data distributions (ZIG)	• No predictions • Reported issues with Type I error control Weiss et al. *Microbiome* 2017	✓	✗
MaAsLin Morgan et al. *Genome Biol* 2012	• Visualizations • Multiclass	• Not yet peer-reviewed • No multivariate microbiota models • Not on CRAN/Bioconductor	✓	✗
SIAMCAT	• On Bioconductor • Complete Workflow • Visualizations • Predictions on new data	• Not yet peer-reviewed • Only for case-control designs	✓	✓

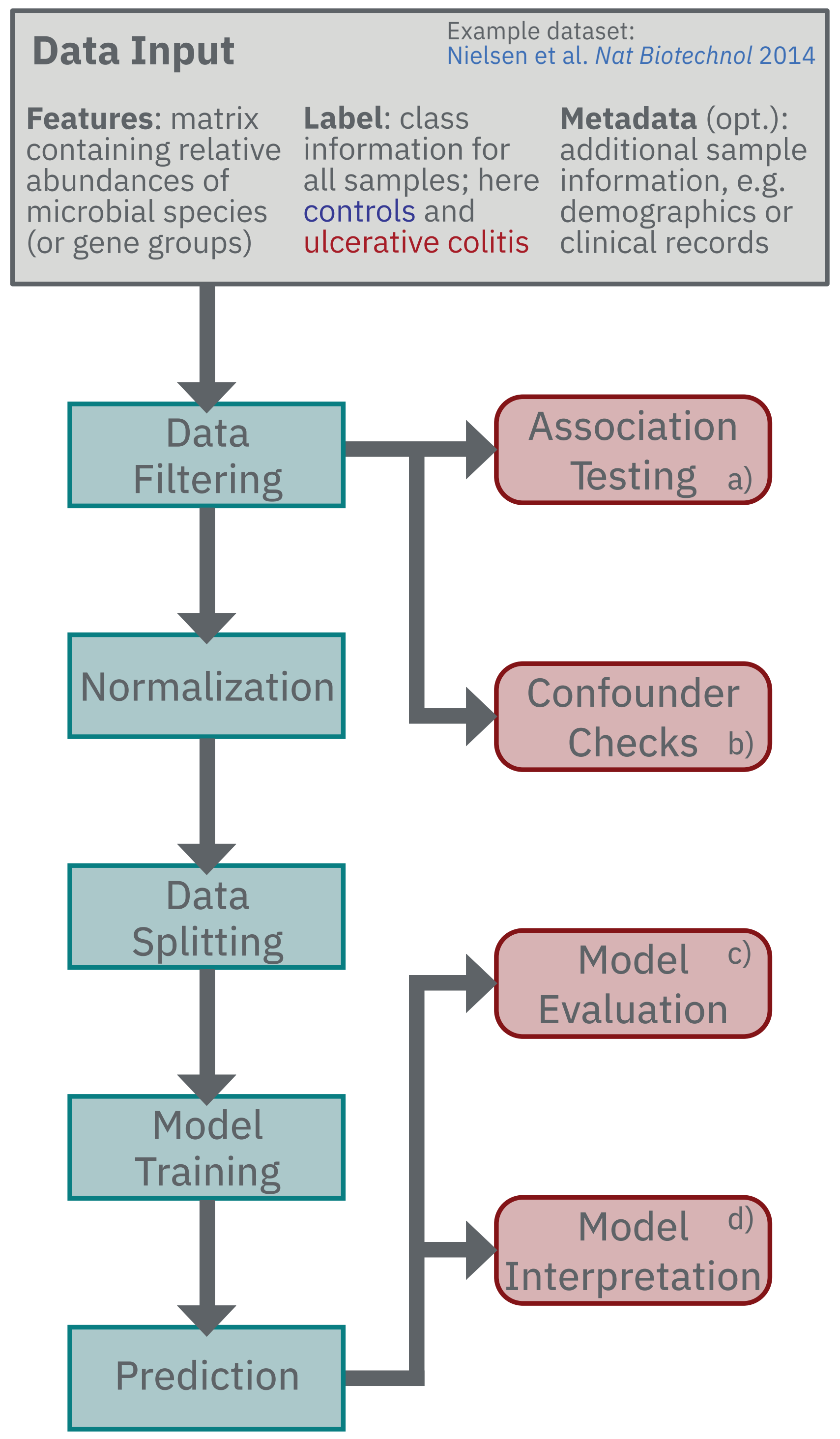
Further reading

siamcat.embl.de

microbiome-tools.embl.de

SIAMCAT: user-friendly and versatile machine learning workflows for statistically rigorous microbiome analyses
Preprint on bioRxiv.org

SIAMCAT Workflow



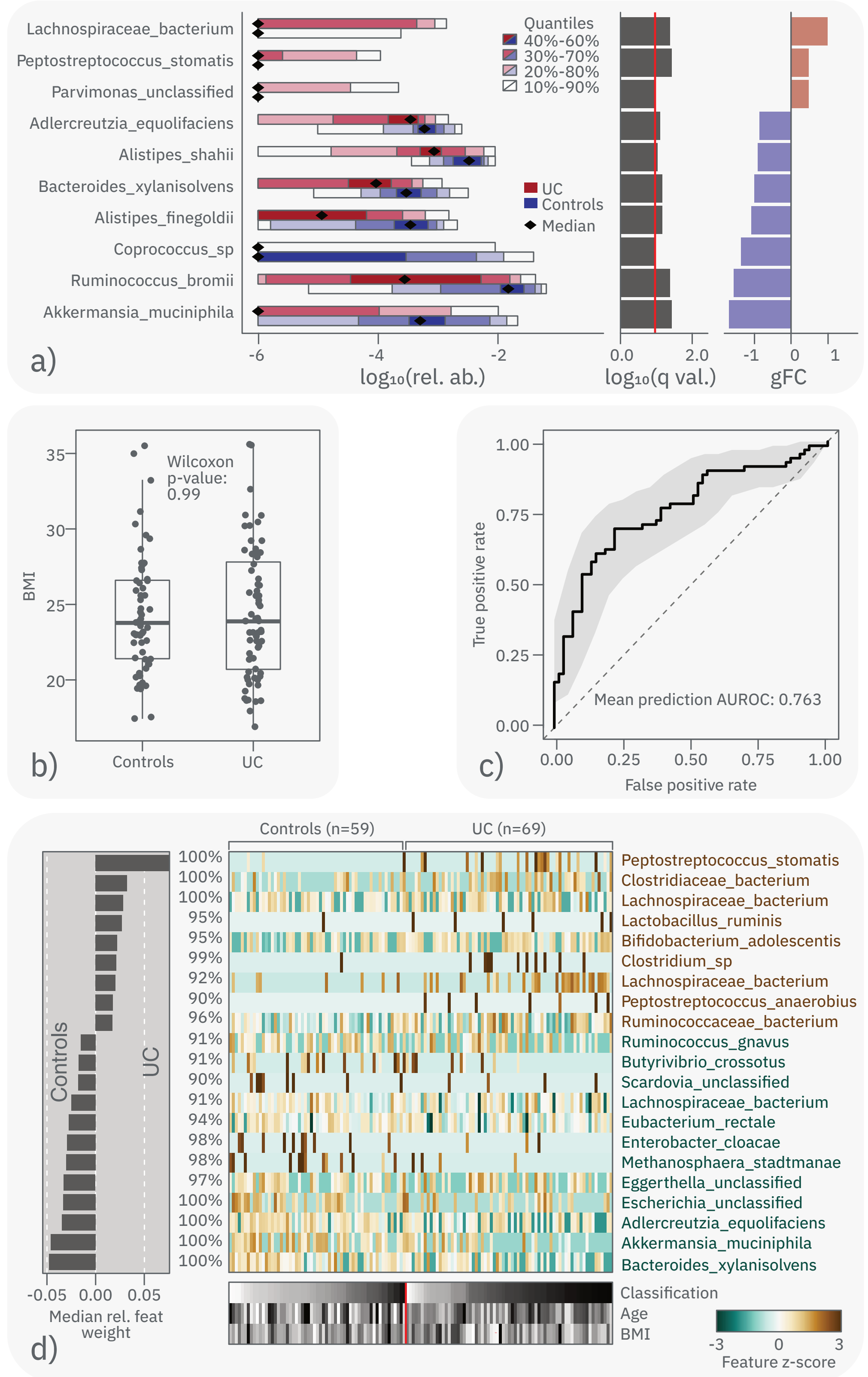
Data Input

Example dataset: Nielsen et al. *Nat Biotechnol* 2014

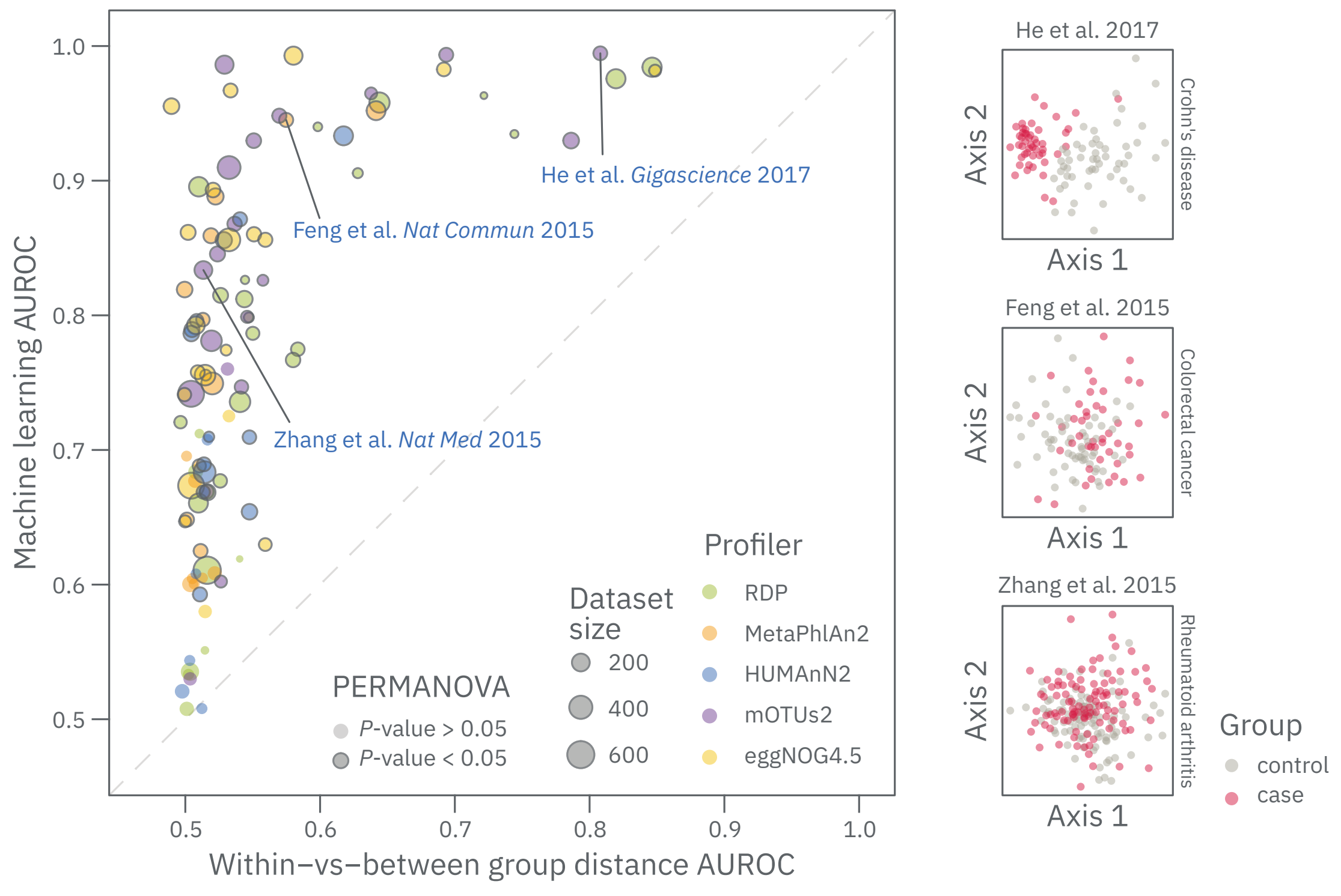
Features: matrix containing relative abundances of microbial species (or gene groups)

Label: class information for all samples; here controls and ulcerative colitis

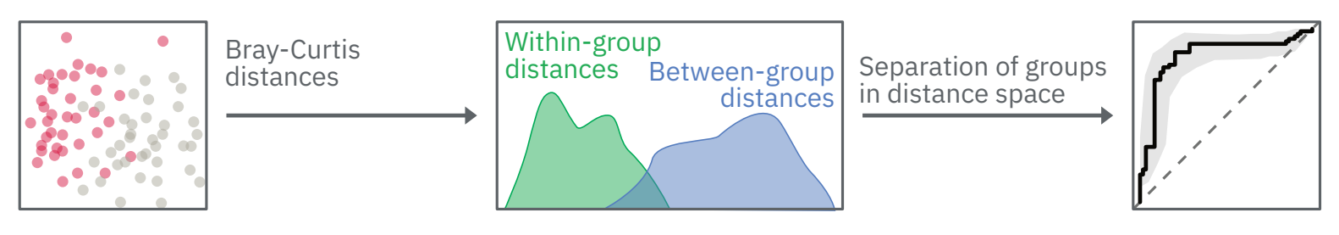
Metadata (opt.): additional sample information, e.g. demographics or clinical records



Machine learning vs. PERMANOVA

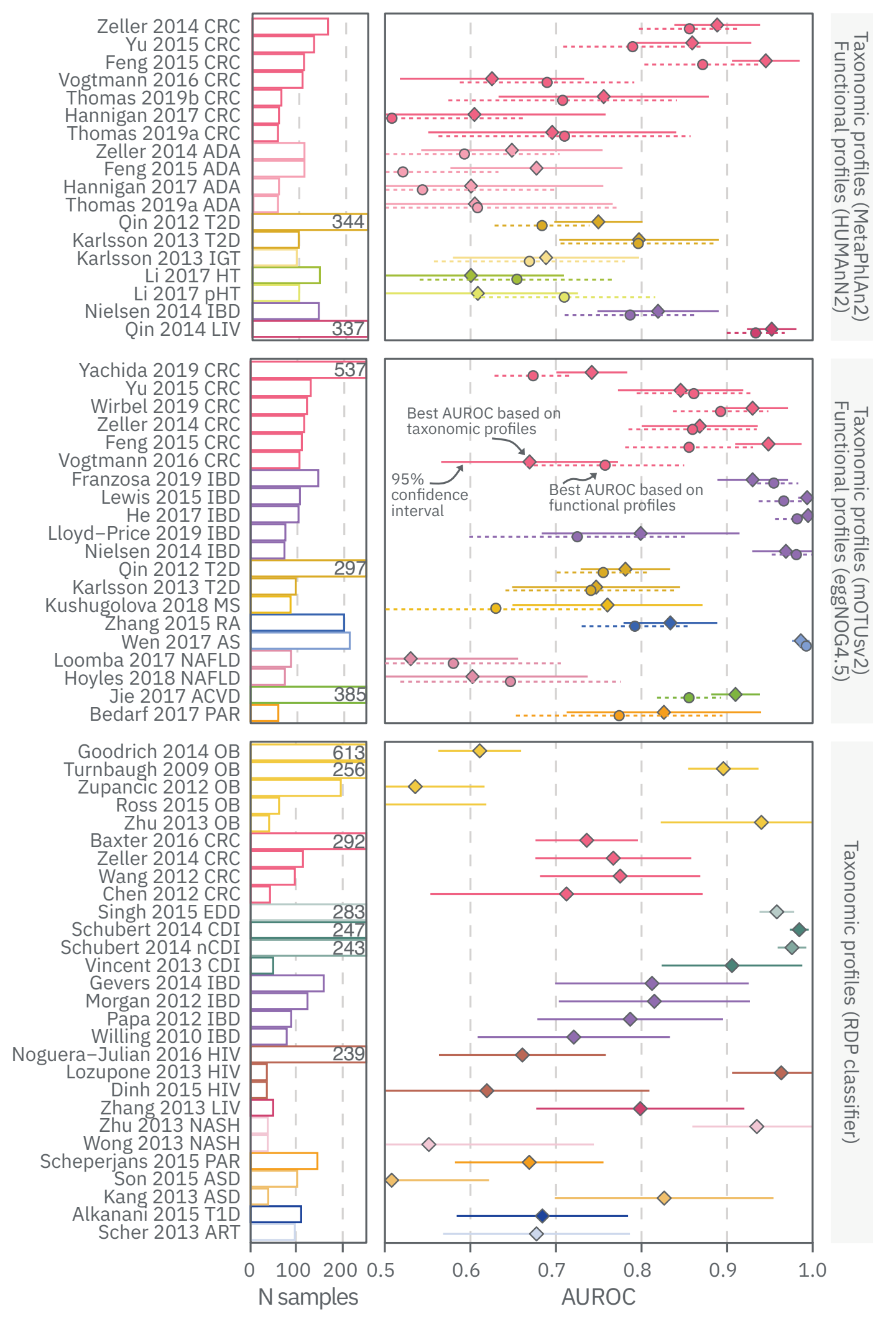


To **compare** the results of **machine learning workflows** with commonly used analyses based on **ecological distances**, we included many datasets in a **machine learning meta-analysis**.
For each dataset, we computed an AUROC based on a **machine learning model** and an AUROC based on **within- and between-group Bray-Curtis dissimilarity** as a measure of separation.



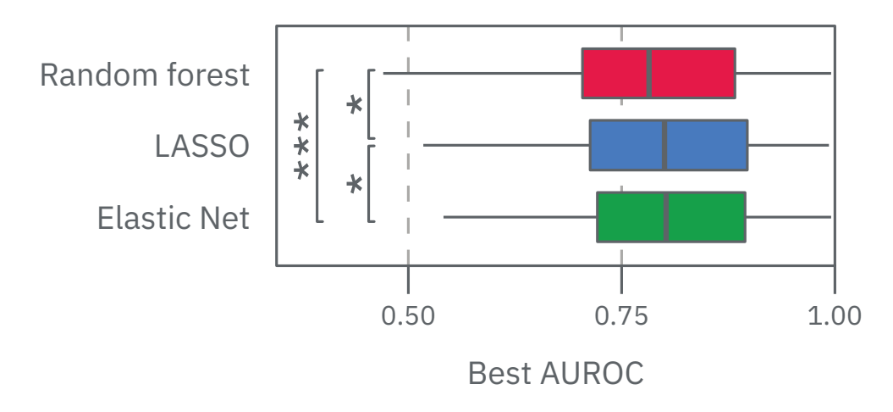
For many datasets, **machine learning** classifiers show very **good accuracy**, while **ecological distances** exhibit **no or only very little separation** between groups. This suggests that the differences between controls and diseased samples are not global, but more nuanced and may be easier to detect with machine learning approaches.

Machine learning meta-analysis

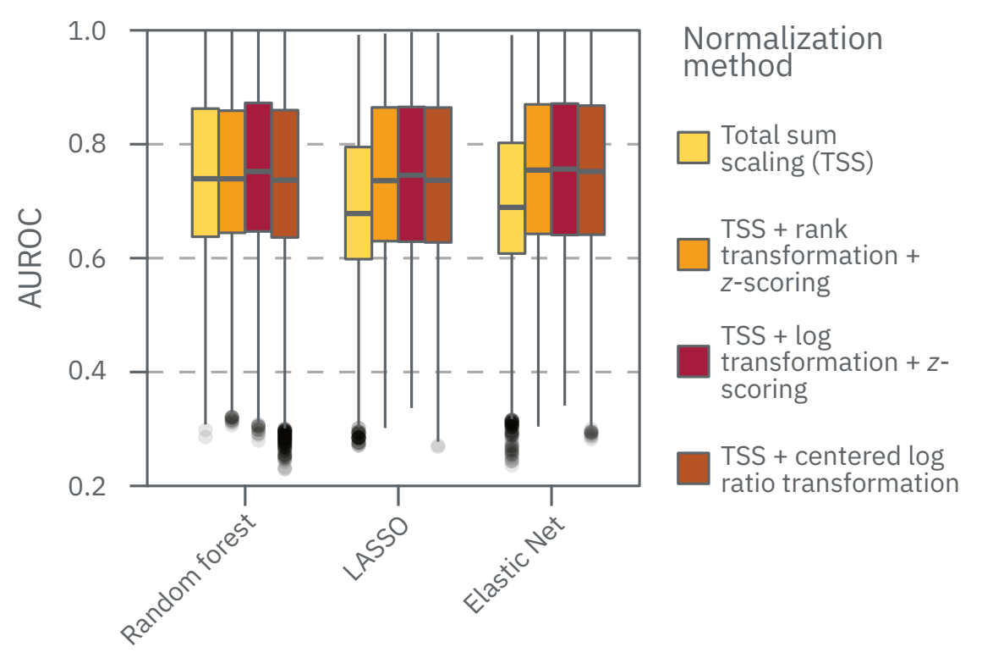


To demonstrate how SIAMCAT can be applied to a **wide range of different types of input data**, we performed a large-scale machine learning meta-analysis of case-control gut metagenomic datasets, including taxonomic profiles based on **16S rRNA** gene sequencing (from Duvallet et al. *Nat Commun* 2017) and functional or taxonomic profiles based on **shotgun metagenomic** sequencing, profiled with many different methods (Pasolli et al. *Nat Methods* 2017, Milanese et al. *Nat Commun* 2018, Huerta-Cepas et al. *Nucleic Acids Res* 2016).

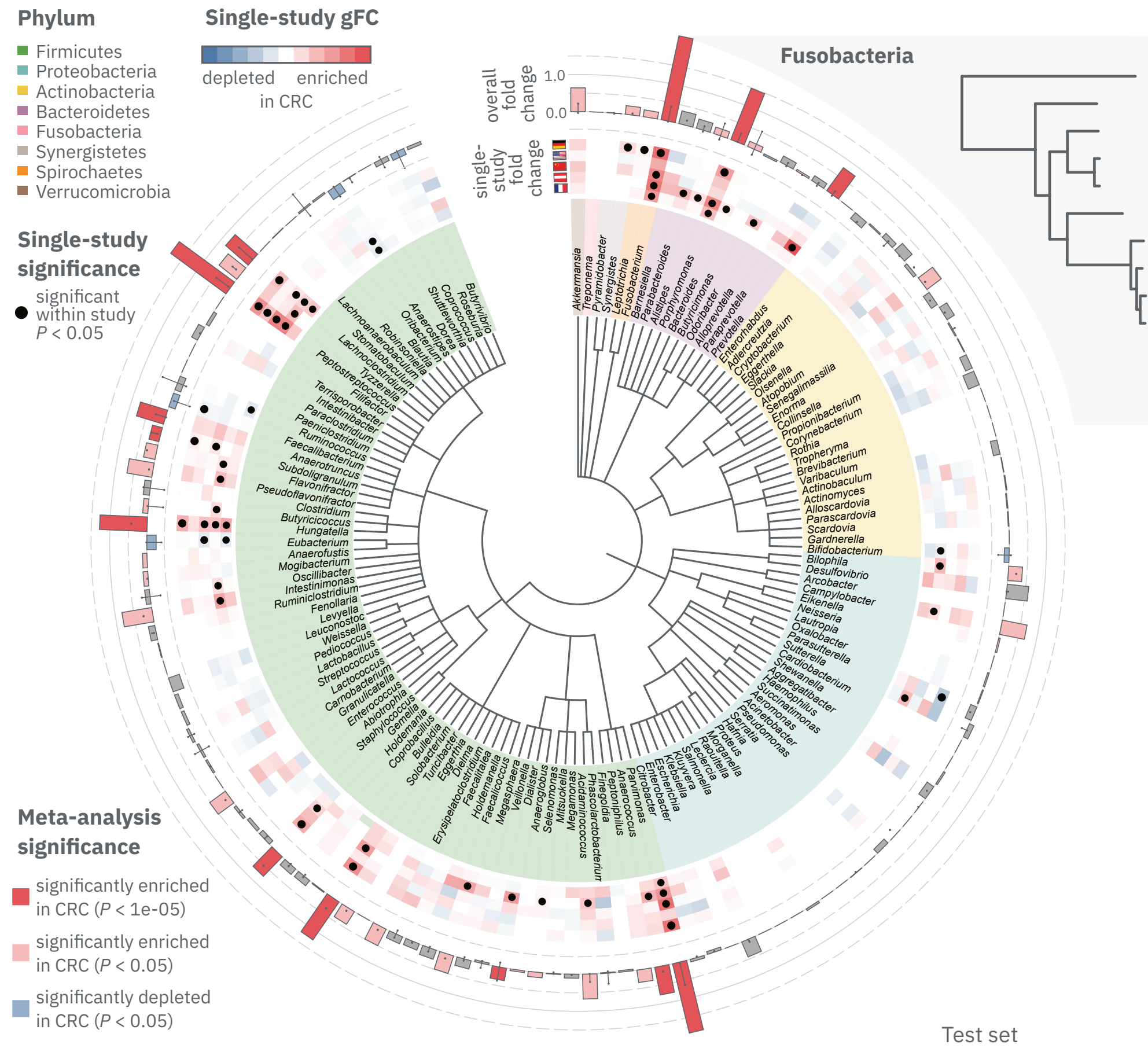
With this, we aimed to **explore** the space of possible machine learning **workflow configurations** and **hyper-parameter combinations**.

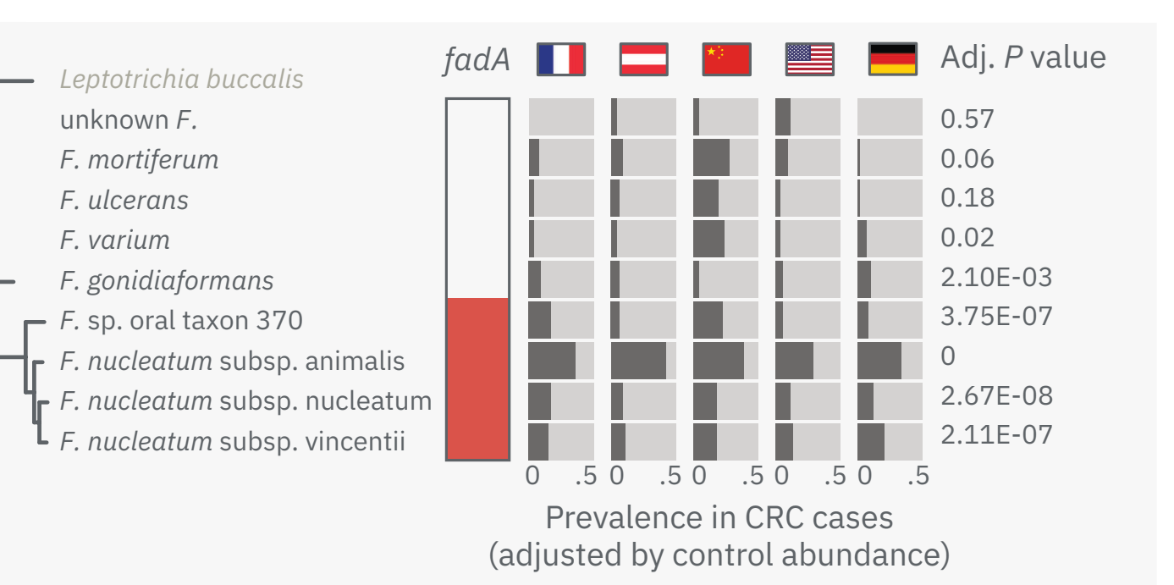


Interestingly, the **Elastic Net** algorithm on average **outperforms** other algorithms, but requires **appropriately normalized** data.



Colorectal cancer meta-analysis using SIAMCAT



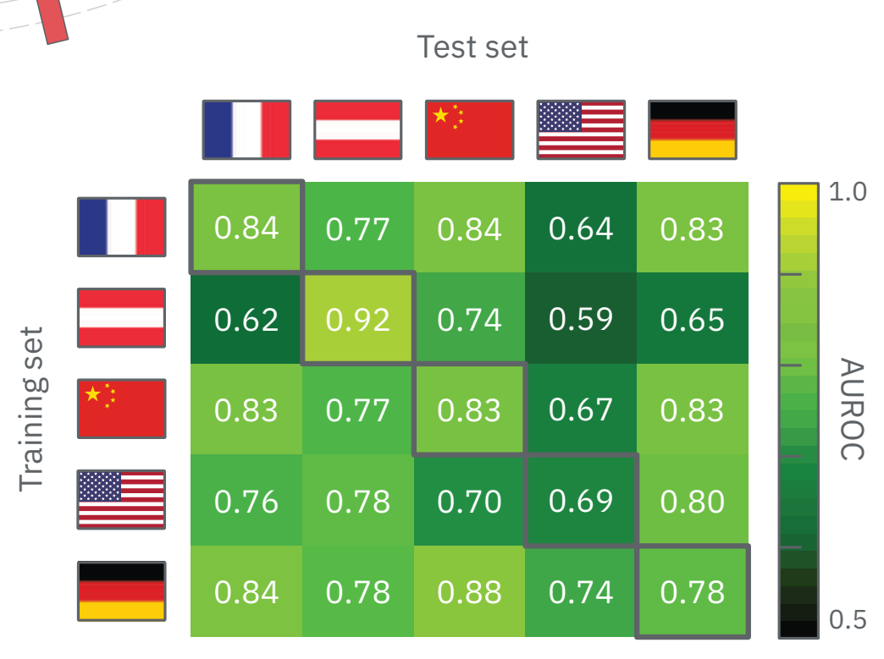


In a recent colorectal cancer (CRC) meta-analysis, we analyzed five metagenomic datasets with broad geographical and technical variability:

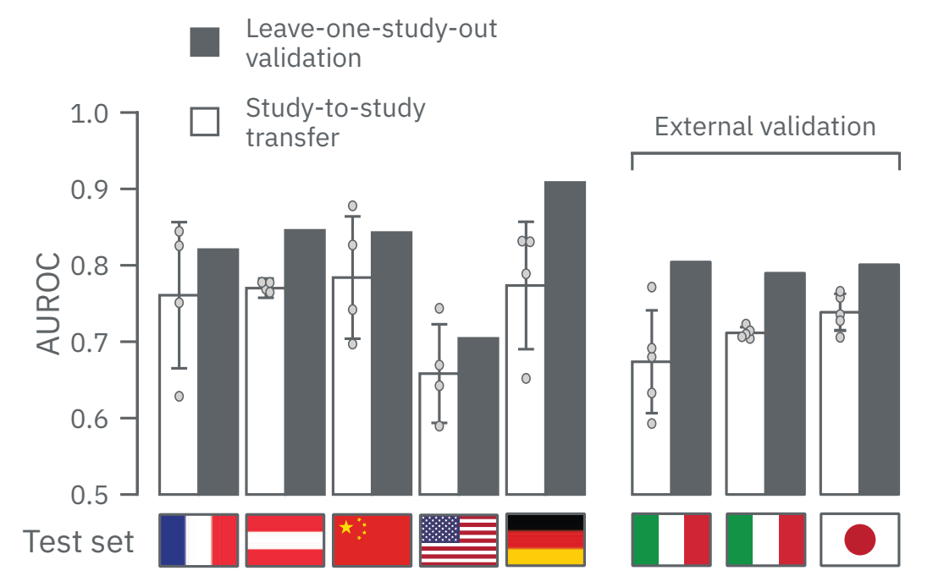
Zeller et al. *Mol Syst Biol* 2014
Feng et al. *Nat Commun* 2015
Yu et al. *Gut* 2016
Vogtmann et al. *PLoS ONE* 2016
Wirbel et al. *Nat Med* 2019 **(Meta-analysis publication)**

We first used SIAMCAT to test for **differential abundance** between control and CRC samples within each dataset, highlighting **consistent enrichment** of several bacterial taxa (e.g. *Fusobacteria*) in CRC patients.

Using the SIAMCAT **machine learning workflow**, we could establish gut microbial signatures that are **highly predictive for CRC**.
Especially models trained on data from multiple studies maintained high performance with an **AUROC of 0.8** or higher on the test set, even for completely external validation from Thomas et al. *Nat Med* 2019 or Yachida et al. *Nat Med* 2019.



Training set \ Test set	🇫🇷	🇦🇹	🇨🇳	🇺🇸	🇩🇪
🇫🇷	0.84	0.77	0.84	0.64	0.83
🇦🇹	0.62	0.92	0.74	0.59	0.65
🇨🇳	0.83	0.77	0.83	0.67	0.83
🇺🇸	0.76	0.78	0.70	0.69	0.80
🇩🇪	0.84	0.78	0.88	0.74	0.78



jakob.wirbel@embl.de
@JakobWirbel
@SIAMCAT_dev
@ZellerGroup

de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

EMBL Heidelberg
Meyerhofstraße 1 · 69177 Heidelberg · Germany
T +49 6221 3870 · www.embl.org