SIAMCAT - Statistical Inference of Associations between Microbial Communities And host phenoTypes

Jakob Wirbel[1], Konrad Zych[1], Morgan Essex[1], Kersten Breuer[1], Nicolai Karcher[1], Paul Igor Costea[1], Shinichi Sunagawa[2], Peer Bork[1], Georg Zeller [1]

(1) Structural & Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

(2) Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland

Motivation

- Changes in microbiome composition are associated with many common conditions
- Microbiome data are extensively mined for biomarkers with diagnostic and therapeutic potential
- Microbiome data anaysis is challenging because microbiome data is
  - not normally distributed
  - zero-inflated
  - compositional
- SIAMCAT is an R package using machine learning to infer associations between microbial communities and host phentoypes

SIAMCAT compared to other tools

Tool	Pros	Cons	Assoc. testing	ML model
LEfSe Segata *et al.* Genome Biol 2011	• Widely used • Visualizations • Multiclass	• Assumes normal distribution • 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MaAsLin Morgan *et al.* Genome Biol 2012	• Visualizations • Multiclass	• Tool not yet peer-reviewed • No multivariate microbiota models • Not on CRAN/Bioconductor	✓	✗
SIAMCAT	• On Bioconductor • Complete workflow • Visualizations • Predictions on new data	• Not yet peer-reviewed • Only for case-control designs	✓	✓

Getting started with SIAMCAT

SIAMCAT

http://siamcat.embl.de/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

@SIAMCAT_dev
@JakobWirbel

EMBL MICROBIOME TOOLS

microbiome-tools.embl.de

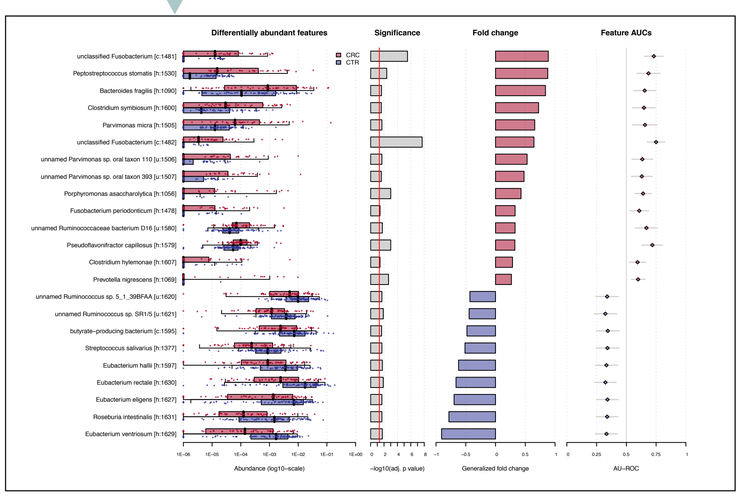
Workflow

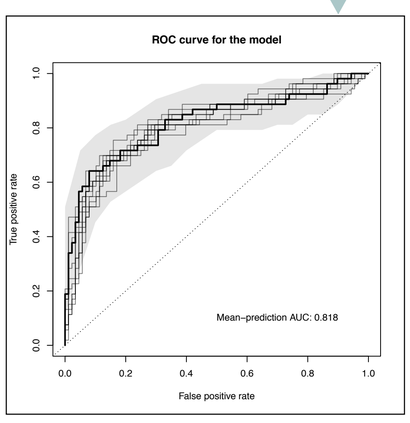
Data input

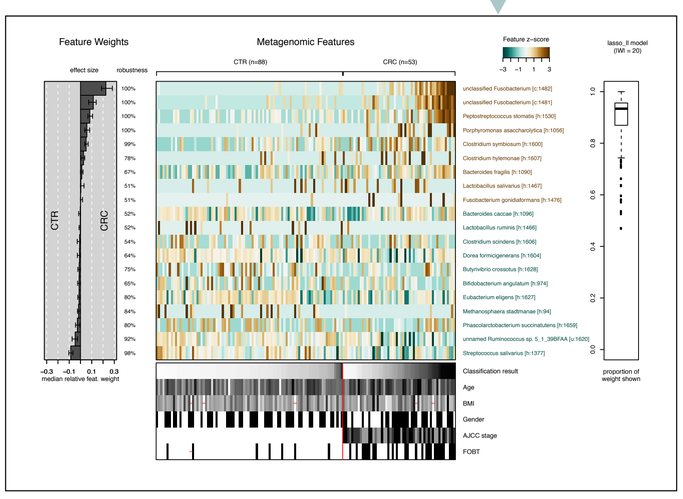
Example dataset:
Zeller *et al.* MSB 2014

- **Features**: matrix containing species (or gene group) relative abundances for all samples
- **Label:** class information for each sample (e.g. healthy control versus colorectal cancer)
- **Metadata** (optional): other sample information, can be checked for confounders

Data filtering → **Data normalization** → **Data splitting** → **Model training** → **Prediction** → **Model interpretation** → **Model evaluation** → **Confounder check** → **Association testing**

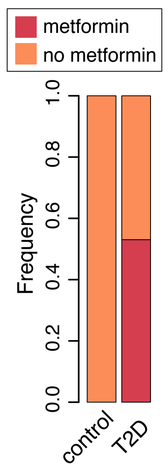


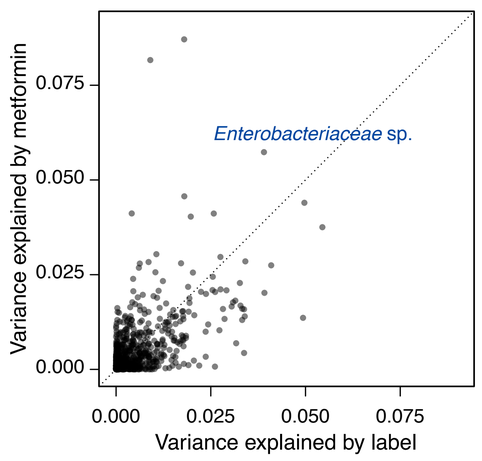


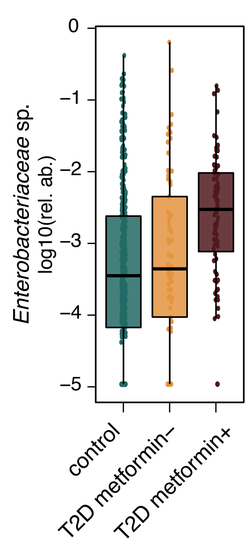


SIAMCAT can detect confounders

Medication with metformin can have dramatic influence on the microbiome composition, potentially confounding studies investigating a microbiome signature for type 2 diabetes (T2D) (see **Forslund *et al.* Nature 2015**).

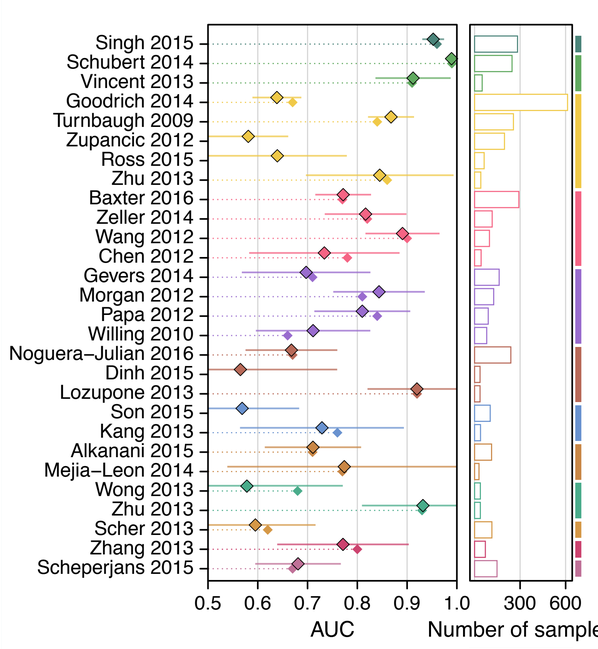






SIAMCAT can detect such confounders and quantify their influence on single species, giving the user a visual aid for confounder identification.

SIAMCAT is a versatile tool



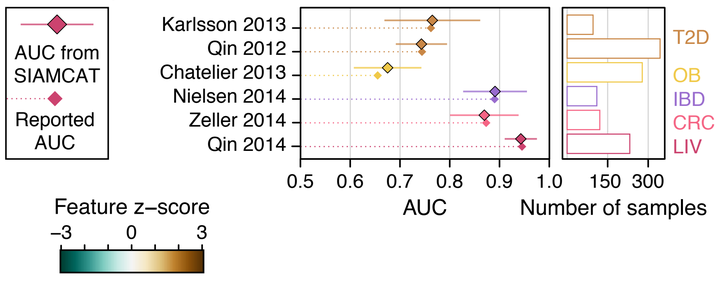
Very different forms of input data can be used as basis for SIAMCAT.

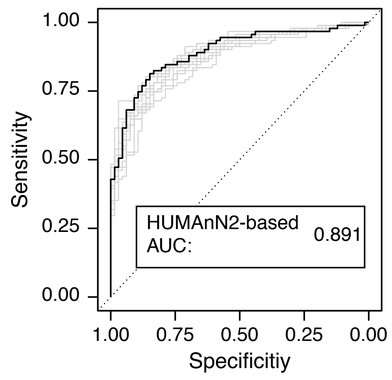
Here, we re-analyzed the data from **Duvallet *et al.* Nat Commun 2017** (**16S-based** genus absolute abundance, left) and **Pasolli *et al.* PLOS Comput Biol 2016** (**MetaPhlAn2** species-level relative abundance, below).

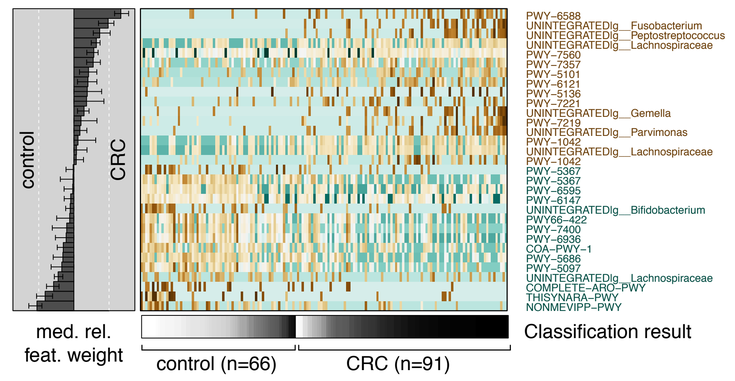
SIAMCAT shows very similar performances compared to the values reported in the paper.

SIAMCAT models can also be trained on gene group abundances, e.g. **HUMAnN2** data

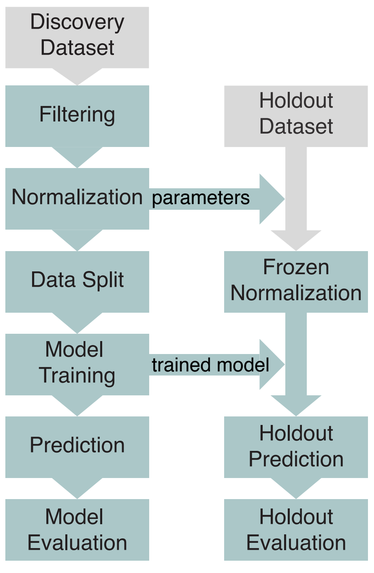
Below: LASSO model trained on HUMAnN2 profiles from **Zeller *et al.* MSB 2014** with feature selection folded into the cross-validation

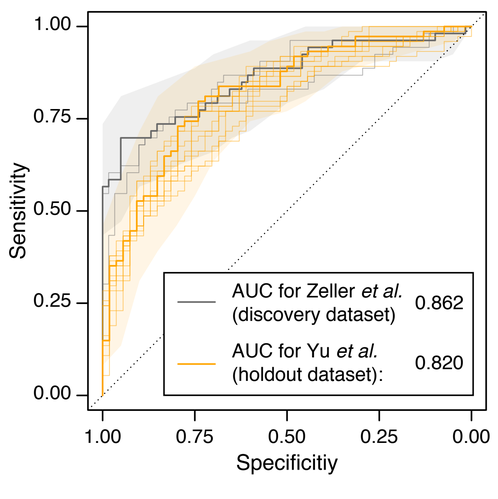






SIAMCAT enables holdout testing



SIAMCAT models can easily be applied to holdout datasets, for example in the context of a **meta-analysis** (see **Wirbel *et al.* Nat Med 2019**).

As an example, we can train a model on the colorectal cancer dataset from **Zeller *et al.* MSB 2014** and apply it to the data from **Yu *et al.* Gut 2017**.