cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

ML4 | MICROBIOME

Working Group 3: Optimisation and standardisation

Cork, June 8th, 2023

Giorgos Papoutsoglou, WG3 Lead

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

JADBio

WG3 Objectives

- Optimise and standardize the use of state-of-the-art ML techniques (WG1) on benchmark data (WG2) to provide SOPs specific to
  - various microbiome data types (16S rRNA amplicons, shotgun metagenomics and metatranscriptomics),
  - human body ecosystems (high/low diversity and variability) and
  - research questions (diagnostics, prognostics, causality)
- Investigate opportunities for automating the established SOPs into pipelines for translational use by clinicians and non-experts.

Deliverables

- D3.1: A decision tree of ML/Stats methods along with optimised parameters suitable for various data types, ecosystems and research questions (disseminated through Web-portal and GitHub).
- D3.2: A publication and white-paper describing the SOPs emanating from D3.1.
- D3.3: A report outlining areas suitable for automation

Administrative information

- Members: 85 (according to the slack channel ☺)
- Slack channel
  - Started on Feb, 2020
  - > messages
- Leadership
  - Magali Berland – Michelangelo Ceci – Magali Berland – Christian Jensen – Giorgos Papoutsoglou (and Sonia always in support!!)
  - Meetings
    - WG: Present in all meetups + 1 dedicated on in Brussels right before COVID hit!
    - Zoom: every month from Sep. 21 onwards

Dissemination

- Workshops and training schools
  - Organizers, trainers and trainees at ML4Microbiome workshops
- STSMs
  - Eliana Ibrahimi, NOVA MATH, FCT NOVA, Lisbon, Portugal
  - Thomas Klammsteiner, University of Ljubljana / Biotechnical Faculty
  - Andrea Mihajlovic, University of Bari, Department of Computer Science
- Papers/Conference presentations
  - Report of the ML4Microbiome Workshop 2021 - Statistical and Machine Learning Techniques for Microbiome Data Analysis. EMBnet Journal 27, e1012http://dx.doi.org/10.14806/ej.
  - Data preprocessing and transformation techniques applied in machine learning modeling of human microbiome data. Preprint.
  - Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Front Microbiol. 2021 Feb 19;12:634511. doi: 10.3389/fmicb.2021.634511. PMID: 33737920; PMCID: PMC7962872.
  - High-performance computing lifts the understanding of insect-based gut microbiomes. Presented at the Austrian-Slovenian HPC Meeting 2021. Online.
  - Searching for consensus in black soldier fly microbiomes. Presented at the 18th International Symposium on Microbial Ecology (ISME18). Lausanne, Switzerland.
  - ...

Recap

Benchmark Datasets

- Ecosystems: gut
- Research question: CRC diagnosis, CRC vs. Adenoma vs. Control
- Shotgun: Saeed (-), Microbiome atlas (-), Public domain curated by Magali’s group (+)
  - ~1600 samples, 755 Controls, 183 Adenoma, 662 CRC
  - AUT, CHN, FRA, GER, IND, ITA, JPN, USA
  - https://doi.org/10.57745/7IVO3E
- 16S data (Laura Marcos)
  - 3 studies, rRNA V4 region
  - 709 samples, 277 Controls, 241 Adenoma, 191 CRC
  - https://hackmd.io/@laurichi13/rJt3ewZut

Analysis results (1)

- Karel Hron
  - Trying different compositionality normalization methods
  - No clear performance increase
- Marta Lopez
  - Checking batch effect correction methods (combat, quantile normalization) for addressing the Country effect
  - Removed the effect, no clear improvement in modeling performance
  - Different preprocessing
- Julia Eckenberger
  - Pipelines: 2 norm. methods, 1 filtering, 3 modeling methods
  - Provided and R script
  - *The type of normalization only had a small effect on the tree-based models while SVMs clearly preferred CLR-transformed data*

Analysis results (2)

- Magali Bertland
  - Testing different preprocessing methods (transformation + filtering) on the WG3 shotgun data
  - 7 different modeling methods
- Sonia Tarazona
  - Testing different preprocessing methods (transformation + filtering) on 6 different 16S datasets
  - 2 different modeling methods
- Alberto Tonda
  - Testing univariate feature selection
  - TPOT autoML

Analysis results (3)

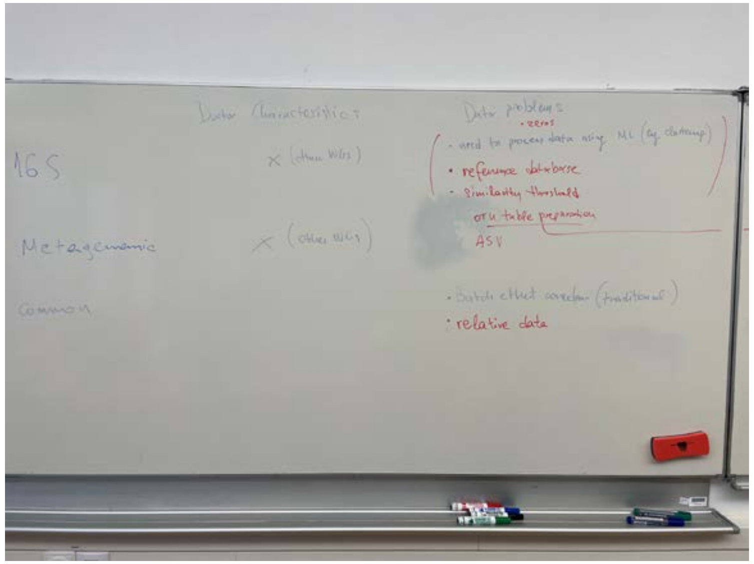
- Christian Jensen
  - Working on some other 16S datasets
  - Doing robust PCA, no feature selection
  - Random forest, SVMs
  - Best model: Compositional transformation + SVMs
- Michelangelo Ceci – Gianvito Pio
  - Analysing the 16S datasets
  - Feature selection + Modeling (Random forest + boosting trees)
- Giorgos Papoutsoglou
  - Using JADBio on WG3 shotgun data

Paper drafting

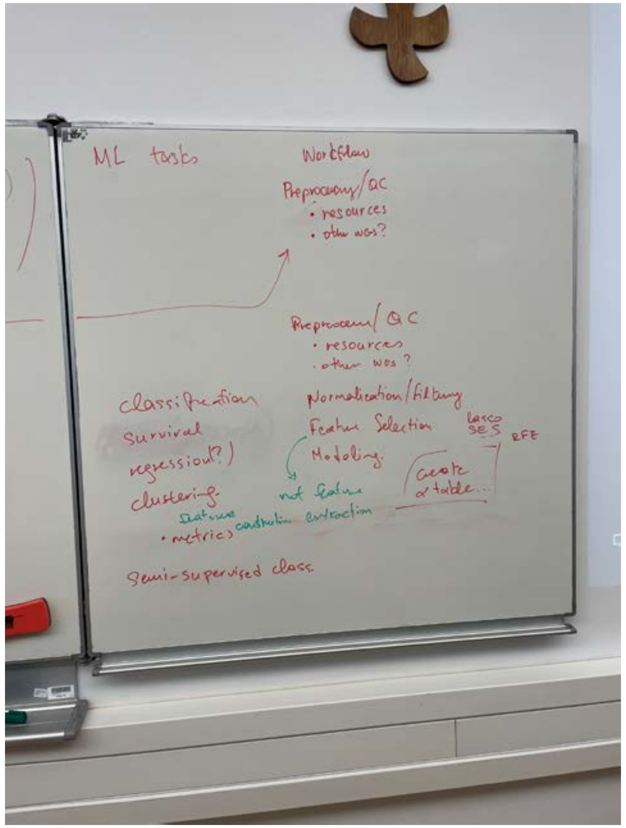
- Plan started after the Tirana meeting
- Turku – Budapest established the final concept to describe
  - How a data analyst currently executes a microbiome data analysis? [Normalization/Filtering –> Feature Selection –> Modeling]
  - For each step in the workflow describe the
    - biological, methodological, and technical problems/constraints pertaining (or not) to microbiome data
    - algorithms and their hyperparameters designed for each ML task (linear vs nonlinear ones, if applicable)
  - Put everything together
    - Estimation protocols
    - Explainability of results

Budapest brainstorming discussion

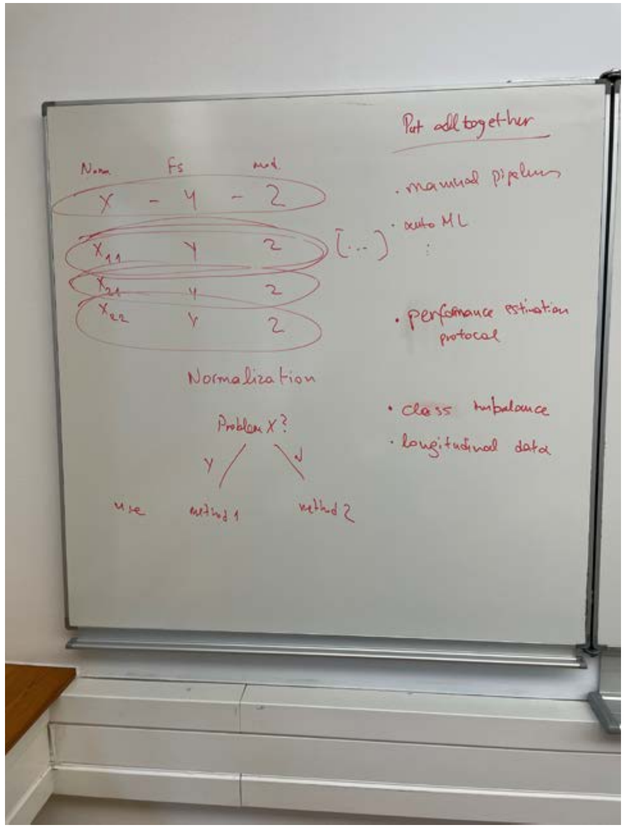
Data description



Workflow



Put all together



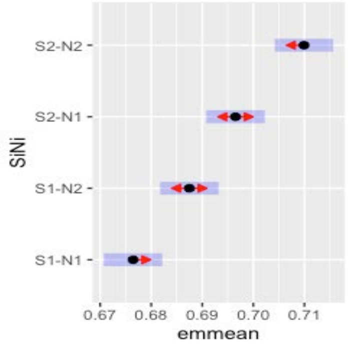
Current state - Results

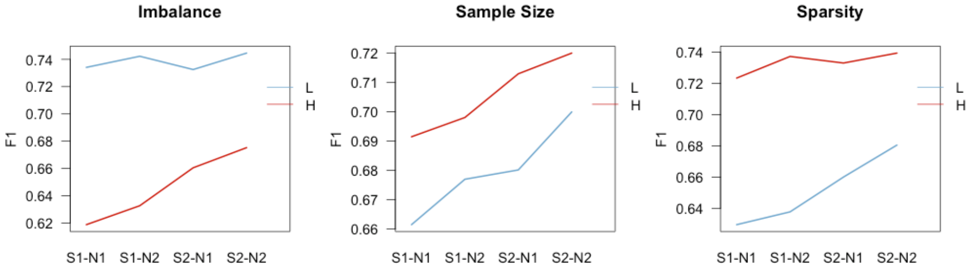
Decision tree

- Started drafting, Sep. ‘21
  - Decided not to include any bioinformatic analysis (data preparation)
- Could not reach to a consensus
  - Too many methods available
  - It would mainly have been based on known trees for ML analysis
- Decided to write Practical Advices for each of the analysis steps

Conference: SEIO 2022 (Granada, Spain)
Oral presentation: Disease prediction from microbiome data
Authors: Sonia Tarazona & Camila Nieto

- Data: Shotgun Illumina Sequencing (stool samples) → 6 datasets (healthy/diseased) from Pasolli et al., *PLoS Computational Biology* (2016): Cirrhosis, Colorectal cancer, IBD, Obesity and T2 Diabetes. n ∈ [100,350]
- Preprocessing: Comparison of 4 strategies combining two prevalence filters (removing zeroes and 20%) and two normalizations (TSS and CLR). In all cases, outlier detection with PCA.
- ML methods: PLS-DA and RF (also SVM but discarded because of low performance). Hyperparameters optimization through repeated k-fold CV (k = 10, r = 5) and F1-score as error metric. No variable selection.
- Results
  - In general, 20% prevalence filtering (S2) combined with CLR (N2) rendered better F1-score.
  - Sparsity benefits classification.
  - For balanced classes, pre-processing effect is not important.
  - For unbalanced classes and/or lower sample sizes, S2-N2 works significantly better.





metagenopolis
mgps.eu

Evaluation of preprocessings for machine learning applications

Giacomo Vitali –Stéphane Béreux
Magali Berland

MetaGenoPolis

Centre de recherche INRAE de Jouy-en-Josas
Domaine de Vilvert, Bât.325
78 350 Jouy-en-Josas France

by INRAE

Machine Learning Model tested

- Random Forest
- PLS – Partial least square
- Earth – spline regression (can be appied to classification also)
- Pam – Partition around medioids (normally a clustering algorithm)
- Glmboost - Gradient Boosting with Component-wise Linear Models
- Glmnet – Generalized linear model with elastic net penality
- GBM – Gradient boosting machine

Dataset: colorectal cancer (CRC) use-case

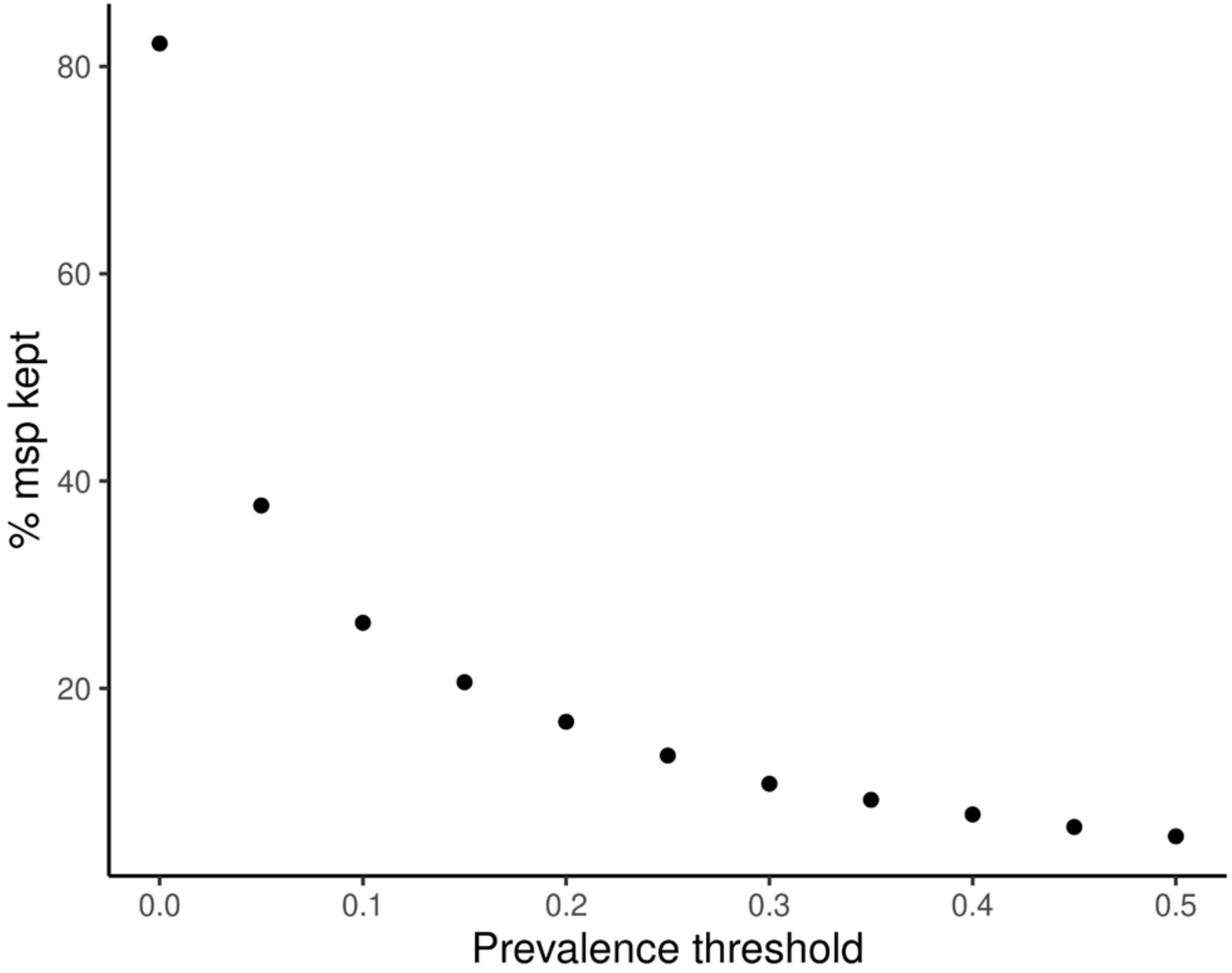
Shotgun data

- **1600 samples**, from 10 publicly available studies
- **8 countries**, over 3 continents (Europe, America, Asia)
- All sequences have been downloaded and processed the same way
  - Mapping of the reads onto the 10.4 million genes IGC2 reference catalog
  - Generation of the gene abundance profiling table (rarefaction and FPKM normalization)
  - Generation of the Metagenomic Species (MGS) abundance table from 100 marker genes
- **Metadata available**: health status and phenotype (healthy, patient, adenoma, CRC stage), country, BMI, gender, age, gene and MGS richness
- Accessible here: https://doi.org/10.57745/7IVO3E

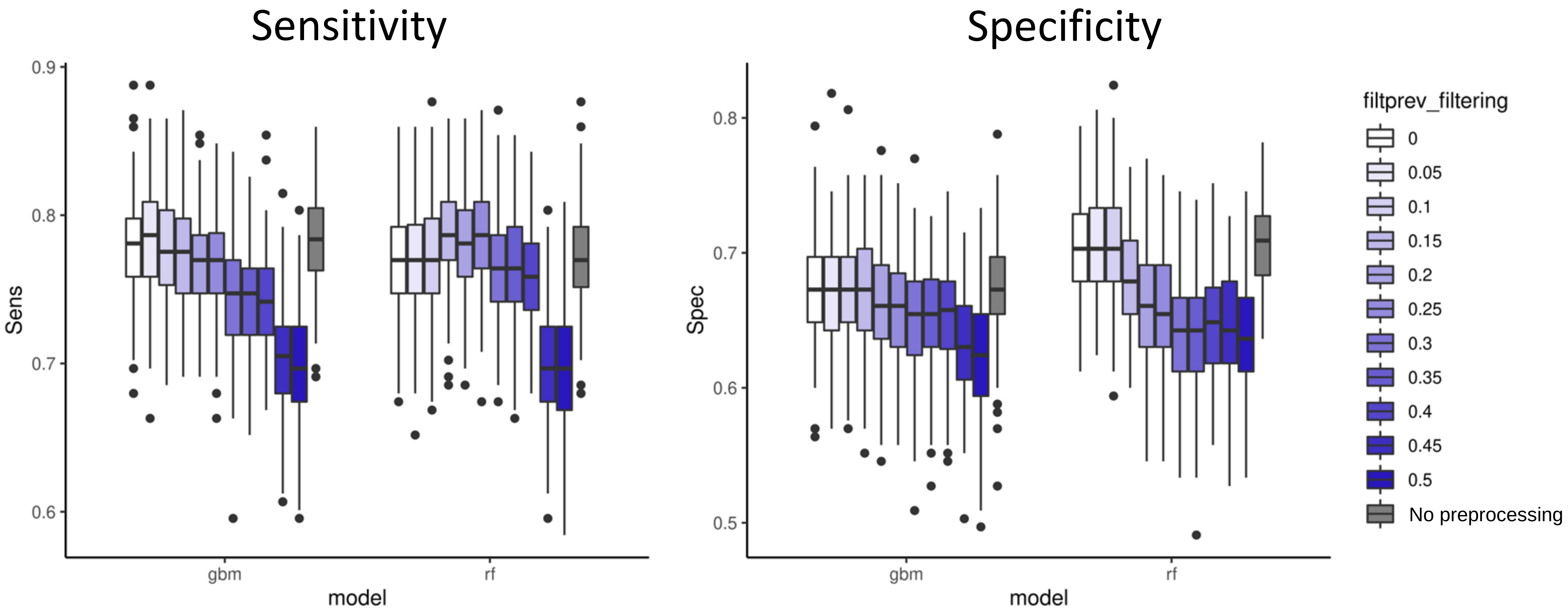
Introduction to the filtering process

- A ***fixed*** threshold for fpkm values: retain features with a total abundance across samples > 5e-06 – Always applied

- A ***variable*** threshold of prevalence (0-0.5): retained features with X prevalence across samples



Best models comparison for a range of filtering values



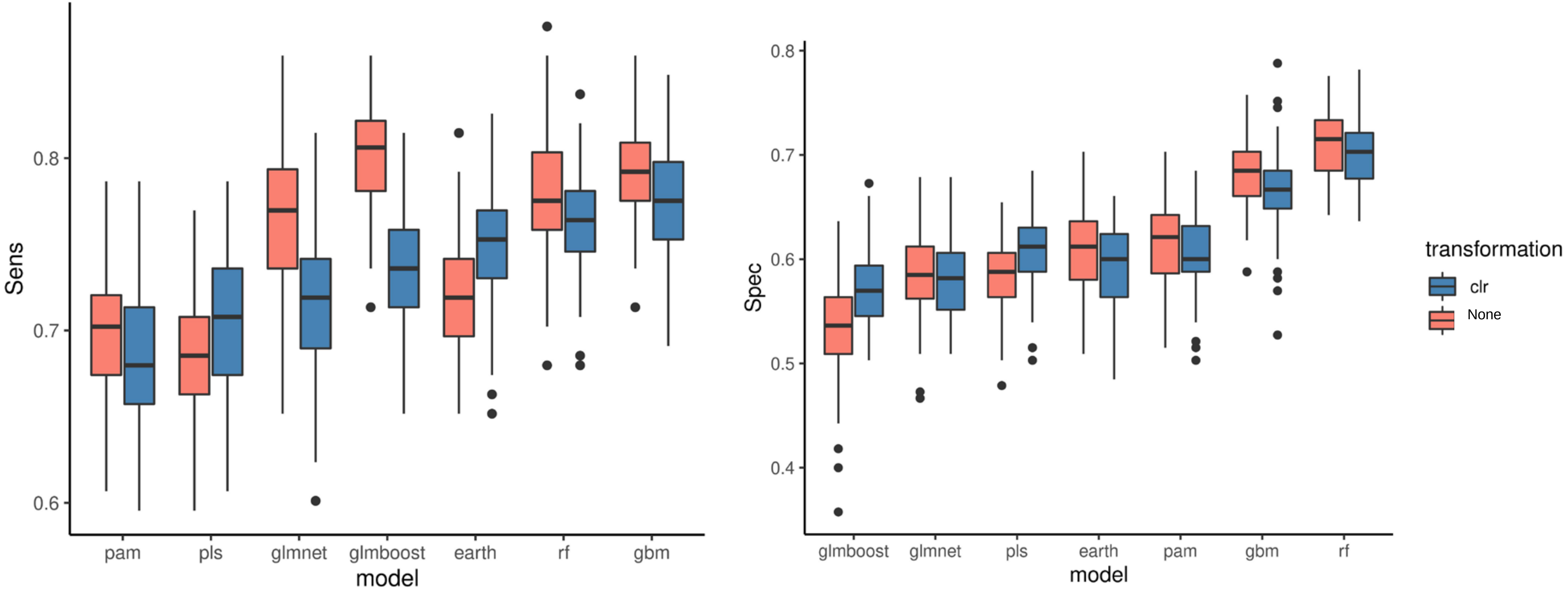
Sensitivity

Specificity

Main messages:

- A small filtering slightly improved the performances
- A strong filtering decreased the performances
- No preprocessing at all is also a valid option

Compositional data – CLR transformation



Main messages:

- CLR transformation slightly improved the performances for PLS model
- For the majority of the models, the CLR transformation decreased the performances

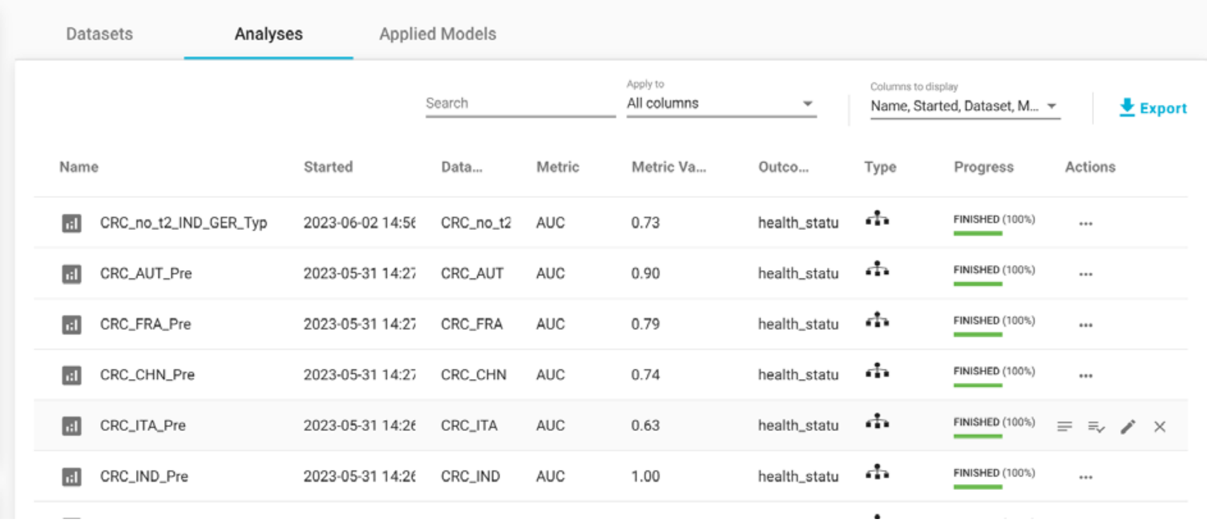
Thanks

fibers
Science
Prebiotics
Explore
Postbiotics
Innovation
Health
Functional
Diversity
Research
Drugs
Cohort
Knowledge
Probiotics
Prevent
Discover
Antibiotic resistance
Metagenomic
Nutrition
gut
metagenopolis mgps.eu
Polyphenols
Enrich
Microbiome
Citizen
Collection
Protect

ML MICROBIOME

Working with COST data (WG3)

- Research question: CRC diagnosis
  - ~1600 samples, 755 Controls, 183 Adenoma, 662 CRC
  - AUT, CHN, FRA, GER, IND, ITA, JPN, USA
- Results
  - Variable AUCs (0.63-0.9)
  - Lot’s of technical artifacts
    - Country: split samples
    - DE: instrument model
    - IND: control vs case
    - JPN: timepoint
  - Signatures: 8 species up to 25
  - SES + Random Forests seem to work very nicely
  - https://docs.google.com/spreadsheets/d/1mREhuCoAj5SmcJ1bUT_El7dGlQ2UWQiO/edit?usp=drive_link&ouid=115018265883606062272&rtpof=true&sd=true



Name	Started	Data...	Metric	Metric Va...	Outco...	Type	Progress	Actions
CRC_no_t2_IND_GER_Typ	2023-06-02 14:56	CRC_no_t2	AUC	0.73	health_statu		FINISHED (100%)	...
CRC_AUT_Pre	2023-05-31 14:27	CRC_AUT	AUC	0.90	health_statu		FINISHED (100%)	...
CRC_FRA_Pre	2023-05-31 14:27	CRC_FRA	AUC	0.79	health_statu		FINISHED (100%)	...
CRC_CHN_Pre	2023-05-31 14:27	CRC_CHN	AUC	0.74	health_statu		FINISHED (100%)	...
CRC_ITA_Pre	2023-05-31 14:26	CRC_ITA	AUC	0.63	health_statu		FINISHED (100%)	≡ ≡✓ ✎ ✕
CRC_IND_Pre	2023-05-31 14:26	CRC_IND	AUC	1.00	health_statu		FINISHED (100%)	...

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

JADBio

Responses from WG3 members

- Excluded those who did not do a comparative analysis
- WG3 Data Analysis.xlsx - Google Sheets

Take home messages

- Sample size and feature size define the methods to try
- Preprocessing
  - Compositional preprocessing/filtering does not affect the predictive performance
  - check the selected features?
- Feature selection
  - important for identifying technical artefacts
  - SES is a good starting point
- Modeling
  - Random Forests are a good starting point

Useful links

- Slack channel: https://ca18131.slack.com/
- Shotgun dataset: https://doi.org/10.57745/7IVO3E
- 16S dataset: https://hackmd.io/@laurichi13/rJt3ewZut
- Bioinformatic processing for shotgun data: https://ca18131.slack.com/files/UUNS11R38/F02NBMW5KSM/2021-11-17-bioinformatic-processing.pdf
- Bioinformatic processing for 16S data: https://ca18131.slack.com/files/U015ZFHBXEW/F02RDAMGKTJ/16sdataset_processing.pdf.pdf
- WG3 (white) Paper: https://docs.google.com/document/d/1tfL58ckp43XDrSglykYejOSqC2_tU4KCC2ugs9TduPk/edit?usp=sharing

THANKS !!!