BioData School Alb 2025

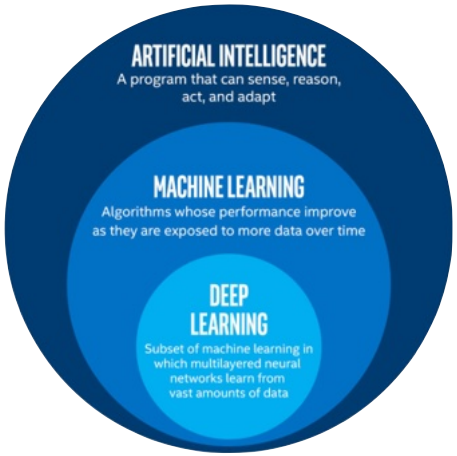
December 11th 2025

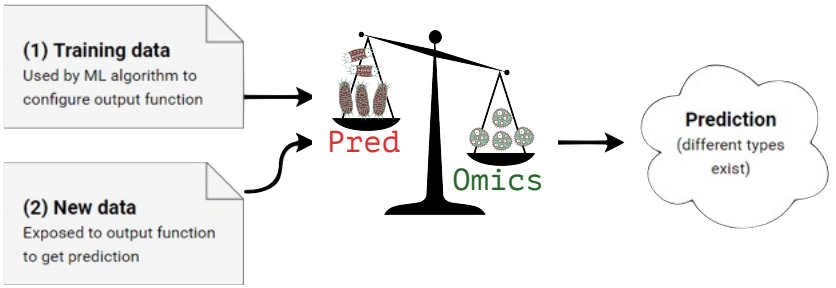
Harnessing AI in Precision Health:
Interpretable AI for Microbiome Applications

Edi Prifti. PhD, DR

Institut de Recherche pour le Développement

UMI 209 UMMISCO

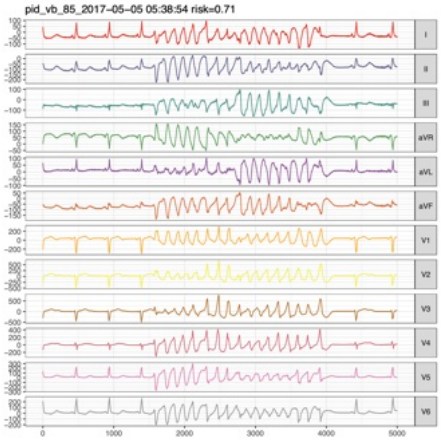




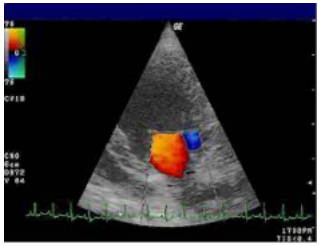
Application fields and the data

Health

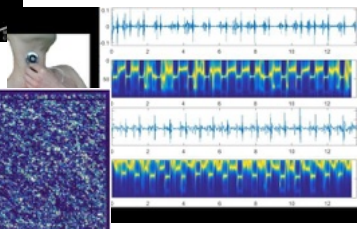
biosignals (ECG, EEG), high throughput sequencing (genomics, metagenomics, barcoding, RNASeq), CT/IRM 3D, text

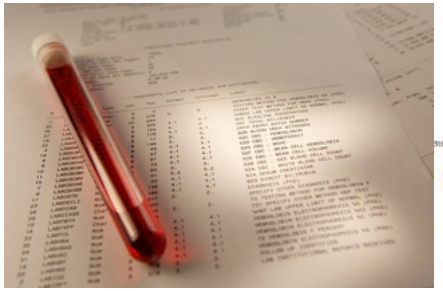


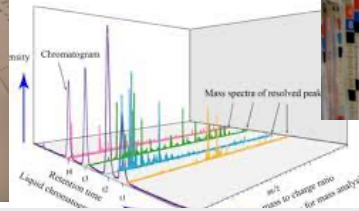

```
>gi|528476558|ref|NC_018928.2| Homo sapiens chromosome 17
TCCACTCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTC
AGCACATCAGCTCTGAGGCCTCCACTCACTGAGACAATAGACCCAAG
TCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTCCACTC
AACAGCTCTGAGGGCTCCACTCACTGAGACAACAGACCCAAGAACAA
>gi|528476670|ref|NC_018912.2| Homo sapiens chromosome 1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTA
TAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCAA
```



A 12-year old girl with known hyperagglutinability, presented to the emergency department with a 2-week history of headeaches and facial weakness. Neurologic examination indicated sensorineural hearing loss on the right side with Weber's test lateralizing to the left, and the Rinne's test demonstrating bone conduction greater than air conduction on the right. Magnetic resonance imaging of the head revealed severe structural defects of the right petrous temporal bone. No indication of cerebral infarction.

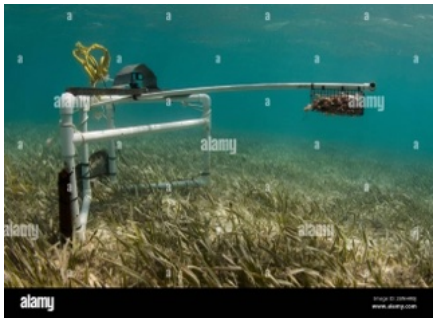


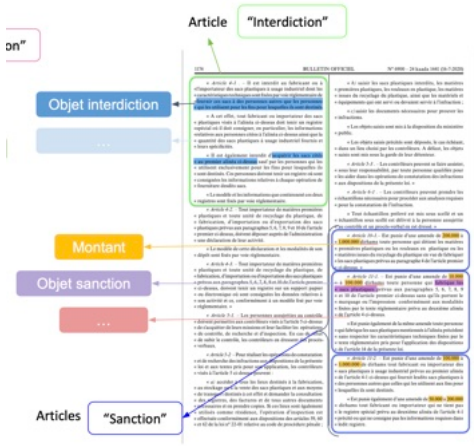




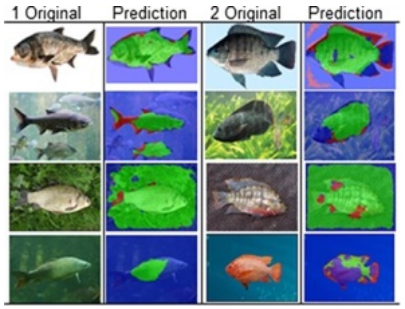
Environnement

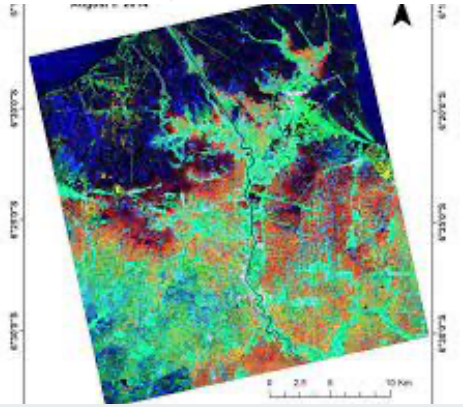
Images, scans 2D, 3D, metabarcoding, satellite plant distribution, texte, video, populations, ecosystems






```
>gi|528476558|ref|NC_018928.2| Homo sapiens chromosome 17
TCCACTCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTC
AGCACATCAGCTCTGAGGCCTCCACTCACTGAGACAATAGACCCAAG
TCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTCCACTC
AACAGCTCTGAGGGCTCCACTCACTGAGACAACAGACCCAAGAACAA
>gi|528476670|ref|NC_018912.2| Homo sapiens chromosome 1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTA
TAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCAA
```

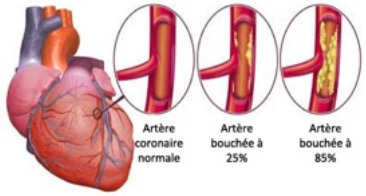




Non-Communicable Diseases (NCDs) **are exponentially more prevalent.** Modelling the **role of the environnement.**

Obesity

Diabetes



CVD

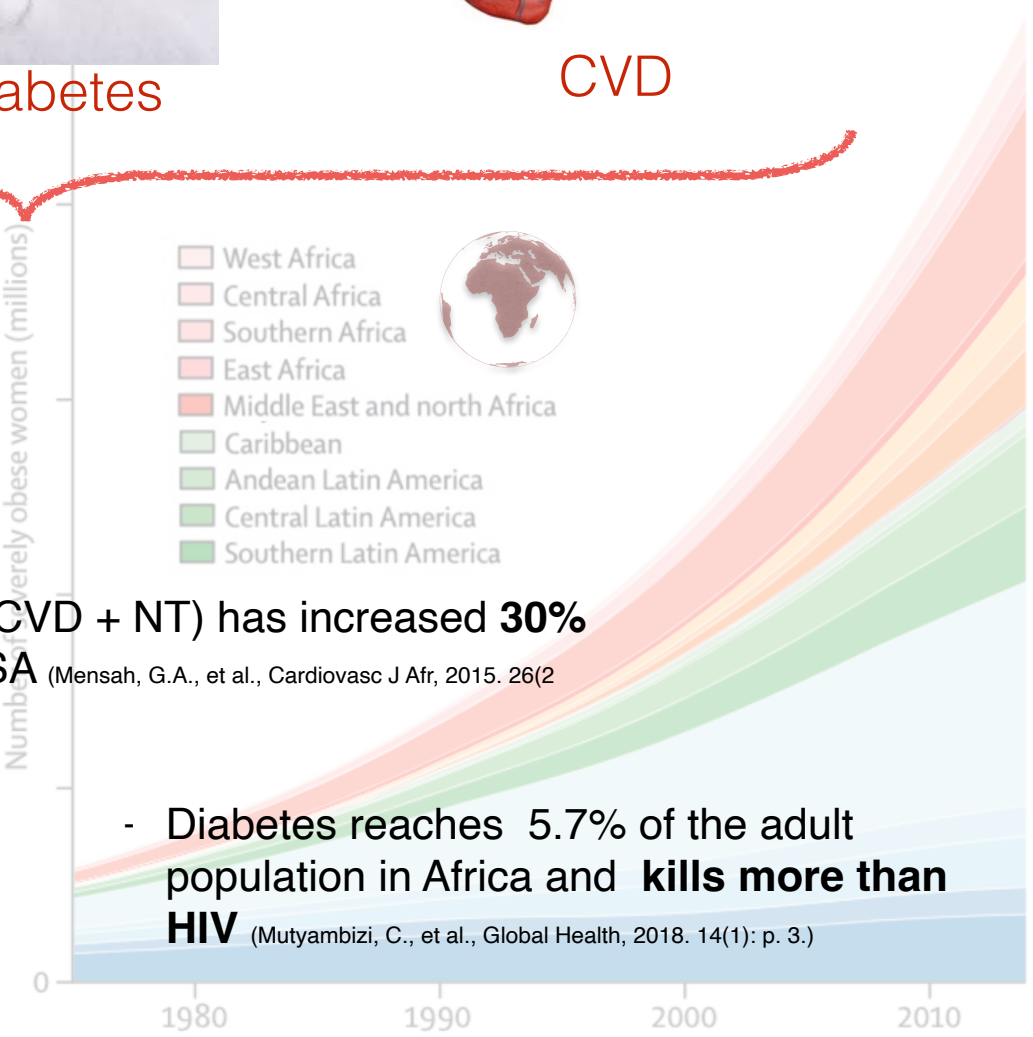
- Morbid obesity has increased worldwide **11-fold in 40 years**

(Mutyambizi, C., et al., Global Health, 2018. 14(1): p. 3.)

GBD Risk factors collaborators., Lancet, 2017.

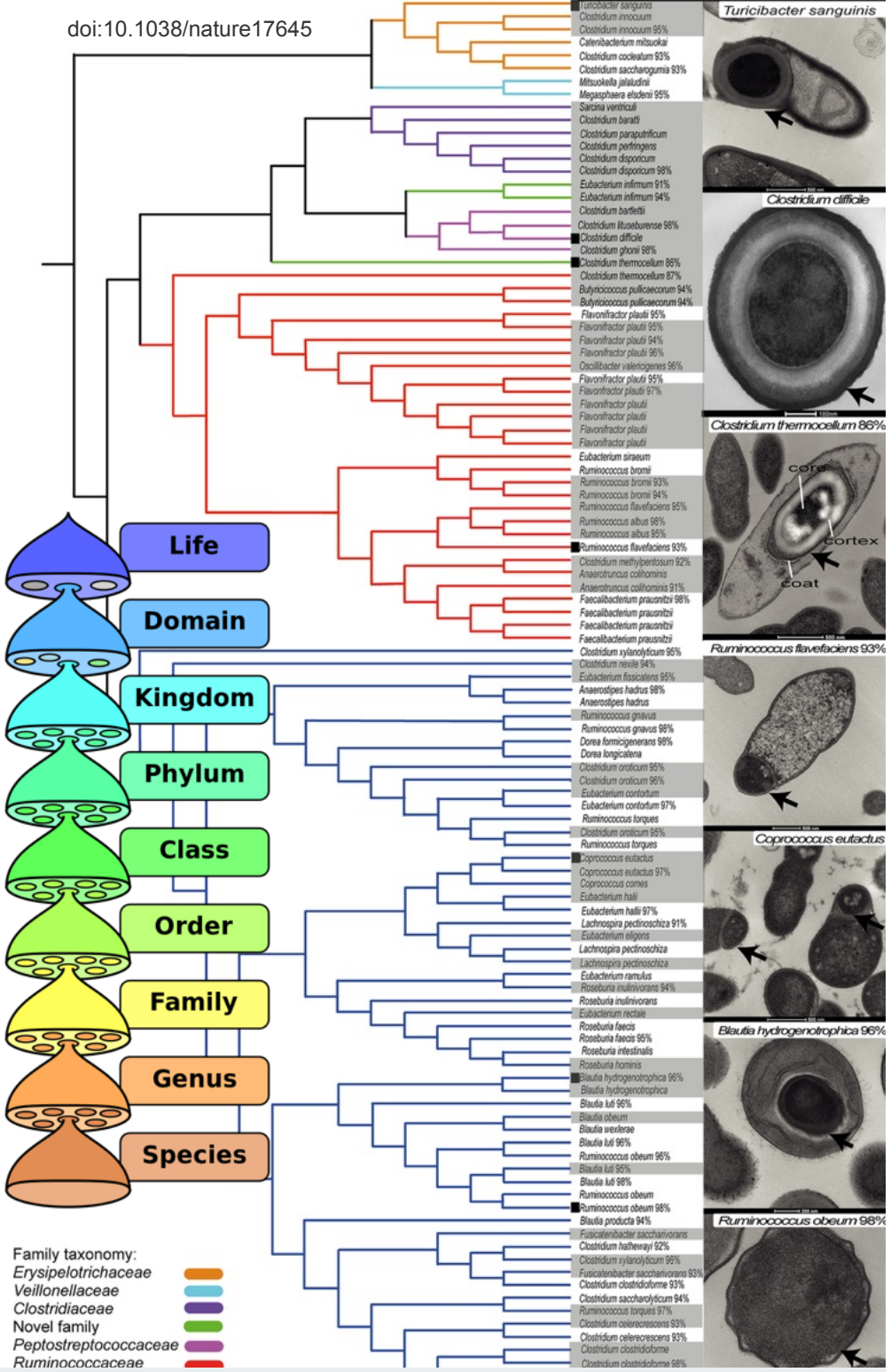
- Mortality due to(CVD + NT) has increased **30% en 30 ans** en SSA (Mensah, G.A., et al., Cardiovasc J Afr, 2015. 26(2 Suppl 1): p. S6-10.)

Collaboration, N.C.D.R.F., Lancet, 2016. 387(10026): p.1377-1396.



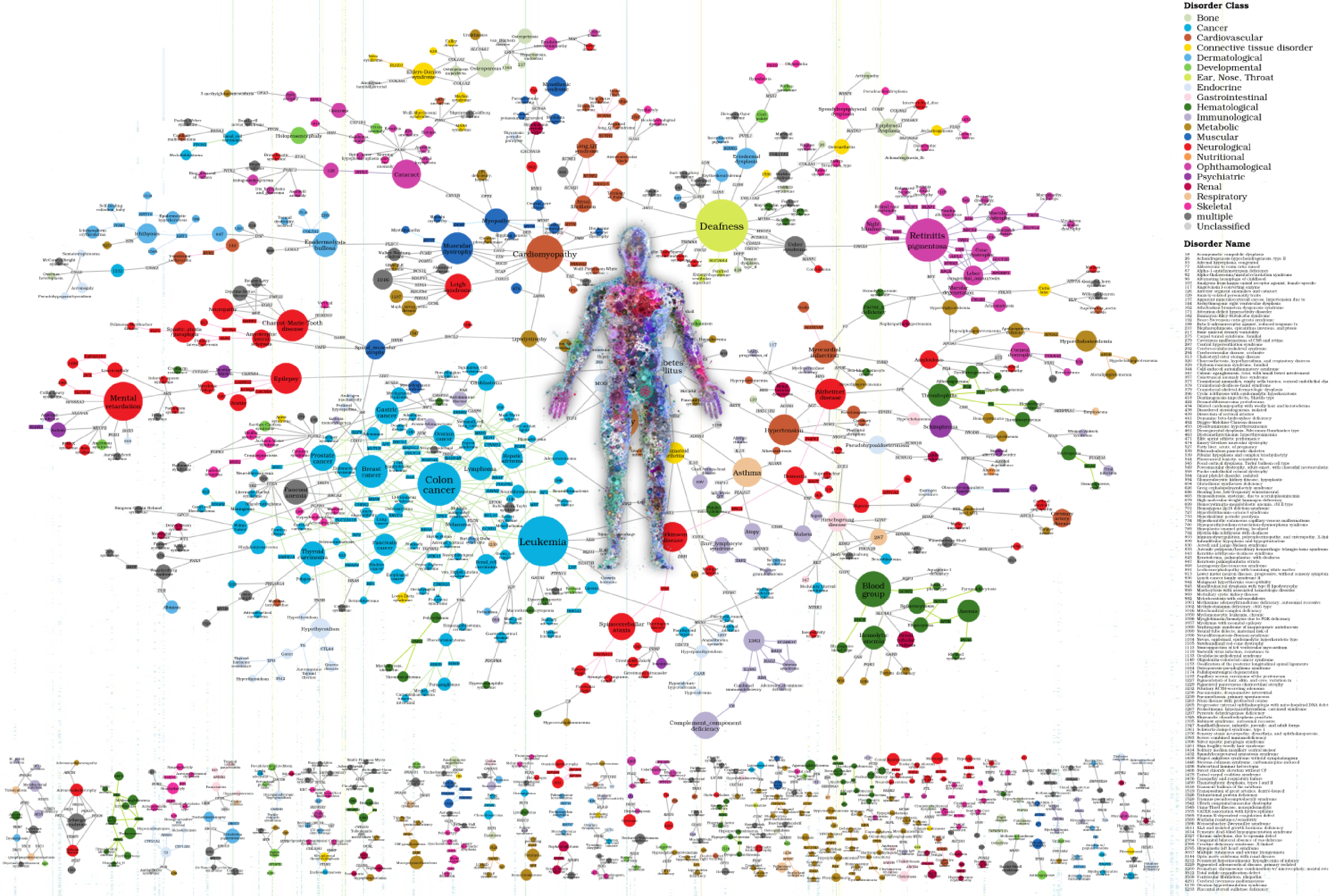
- Diabetes reaches 5.7% of the adult population in Africa and **kills more than HIV** (Mutyambizi, C., et al., Global Health, 2018. 14(1): p. 3.)

- 2Kg of microbes; **1/2 of the cells is bacterial**; **400 times more bacterial DNA**.
- **~5000** species, **~300** in average in a given sample
- Important in **metabolism**, **immune system education**, **barrier** with the outside world, etc.



The human disease network

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) *Proc Natl Acad Sci USA* 104:8685-8690



Supporting Information Figure 13 | Bipartite-graph representation of the diseasome. A disorder (circle) and a gene (rectangle) are connected if the gene is implicated in the disorder. The size of the circle represents the number of distinct genes associated with the disorder. Isolated disorders (disorders having no links to other disorders) are not shown. Also, only genes connecting diseases are shown.

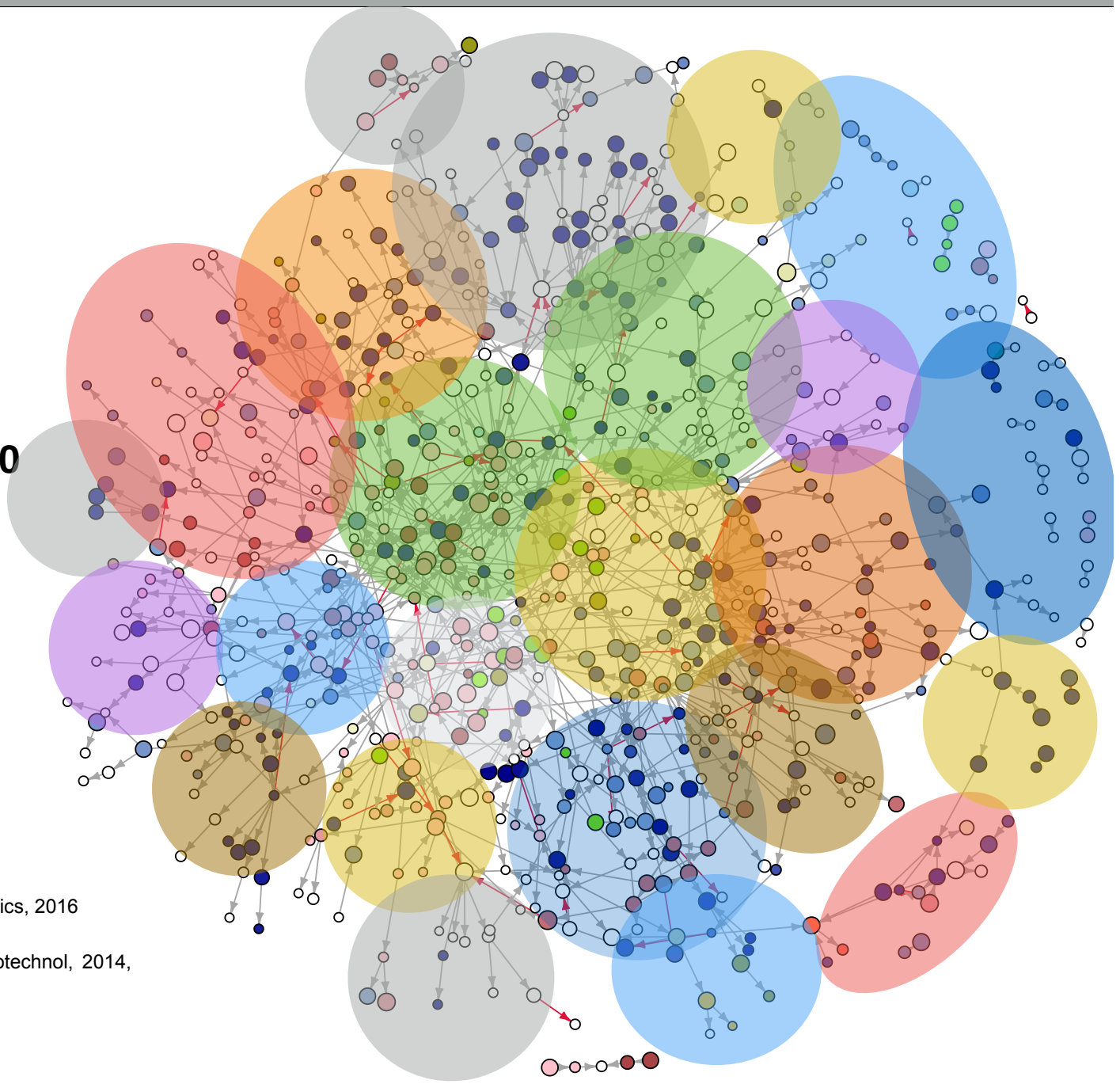
The gut microbiome is very **complex to model**


Severine Affeldt


Nataliya Sokolovska


Jean-Daniel Zucker

- **p ~ 5000** (variables) **> n ~ 100** (observations)
- exponential complexity
- highly sparse data
- limited annotations
- **scalenet**: local reconstruction (eigenvectors / laplacian)



S Affeldt, N Sokolovska, E Prifti, JD Zucker. BMC bioinformatics, 2016

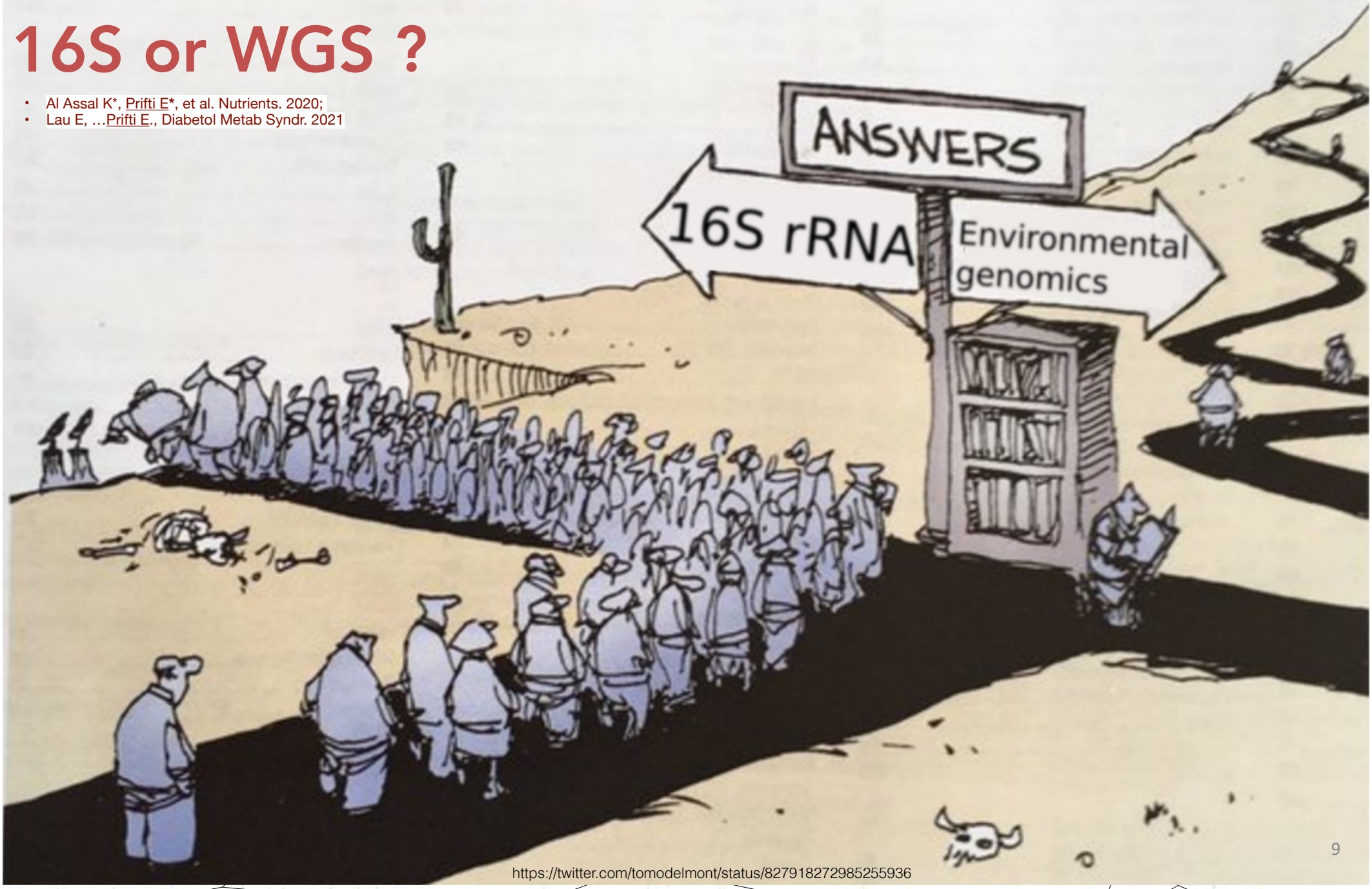
S Affeldt, N Sokolovska, E Prifti, JD Zucker. IJCNN, 2017

Nielsen, H. B.*, Almeida, M.*, …Prifti, E., et al. Nature Biotechnol, 2014, 32(8), 822-8

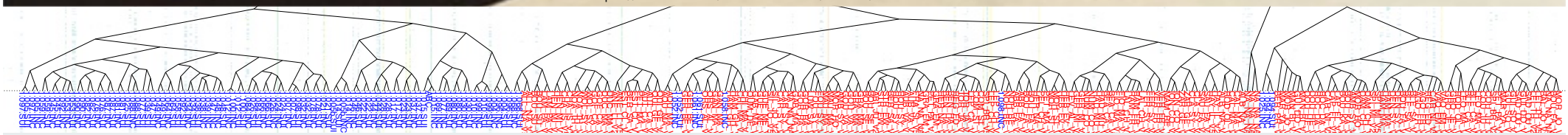
Cougoul et al 2019

16S or WGS ?

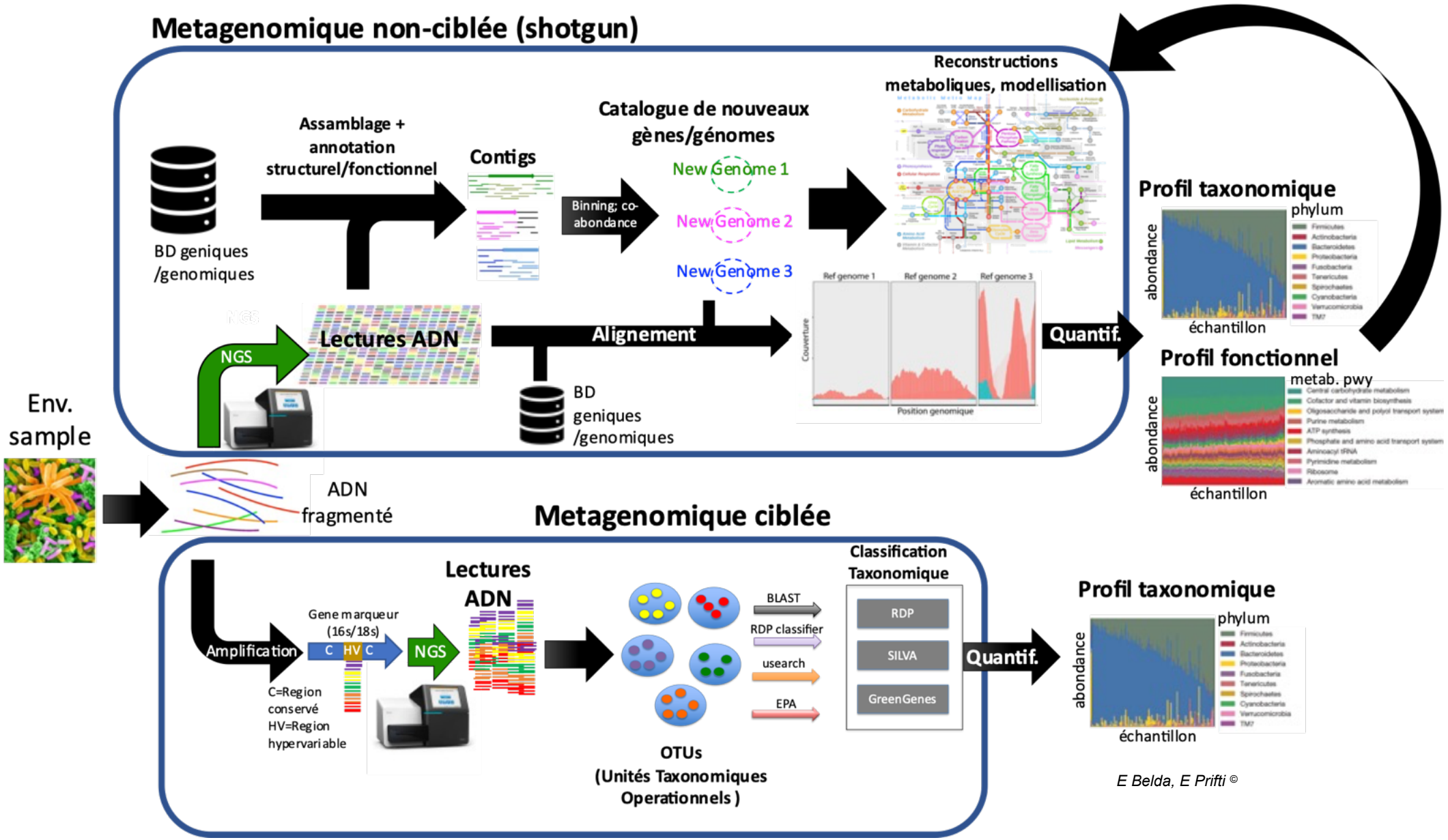
- Al Assal K*, Prifti E*, et al. Nutrients. 2020;
- Lau E, …Prifti E., Diabetol Metab Syndr. 2021



https://twitter.com/tomodelmont/status/827918272985255936

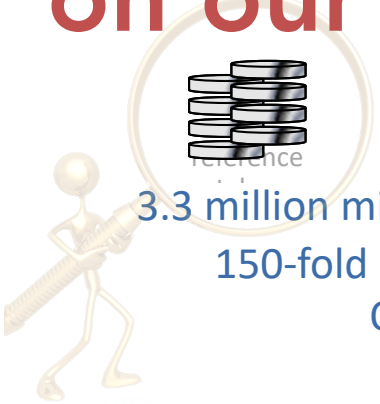


There are two main approaches of microbiome "quantitative" profiling



E Belda, E Prifti ©

The way how we see the world depends on our reference



3.3 million microbial gene catalogue

150-fold human genome !!!

Qin et al Nature, 2010

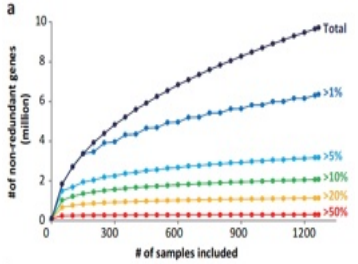
3.3M genes

3.9M genes

5.1M genes

10M genes

MetaHit HMP and others



Adapted from Gevers, et al, 2012, Plos One

The wheel is turning as we go and our view of the world becomes more complex


Nicolas Pons


E. Le Chatelier


Mathieu Almeida

UHGP
>170M
proteins
205k genomes

nature biotechnology

RESOURCE

https://doi.org/10.1038/s41587-020-0603-3

Check for updates

OPEN

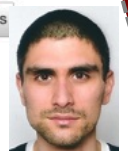
A unified catalog of 204,938 reference genomes from the human gut microbiome

Alexandre Almeida[1,2]✉, Stephen Nayfach[3,4], Miguel Boland[1], Francesco Strozzi[5], Martin Beracochea[1], Zhou Jason Shi[6,7], Katherine S. Pollard[6,7,8,9,10,11], Ekaterina Sakharova[1], Donovan H. Parks[12], Philip Hugenholtz[12], Nicola Segata[13], Nikos C. Kyrpides[3,4] and Robert D. Finn[1]✉

Comprehensive, high-quality reference genomes are required for functional characterization and taxonomic assignment of the human gut microbiota. We present the Unified Human Gastrointestinal Genome (UHGG) collection, comprising 204,938 non-redundant genomes from 4,644 gut prokaryotes. These genomes encode >170 million protein sequences, which we collated in the Unified Human Gastrointestinal Protein (UHGP) catalog. The UHGP more than doubles the number of gut proteins in comparison to those present in the Integrated Gene Catalog. More than 70% of the UHGG species lack cultured representatives, and 40% of the UHGP lack functional annotations. Intraspecies genomic variation analyses revealed a large reservoir of accessory genes and single-nucleotide variants, many of which are specific to individual human populations. The UHGG and UHGP collections will enable studies linking genotypes to phenotypes in the human gut microbiome.

https://www.nature.com/articles/s41587-020-0603-3#article-info

Published: 20 July 2020


Florian Plaza Onate


Dusko Ehrlich

+

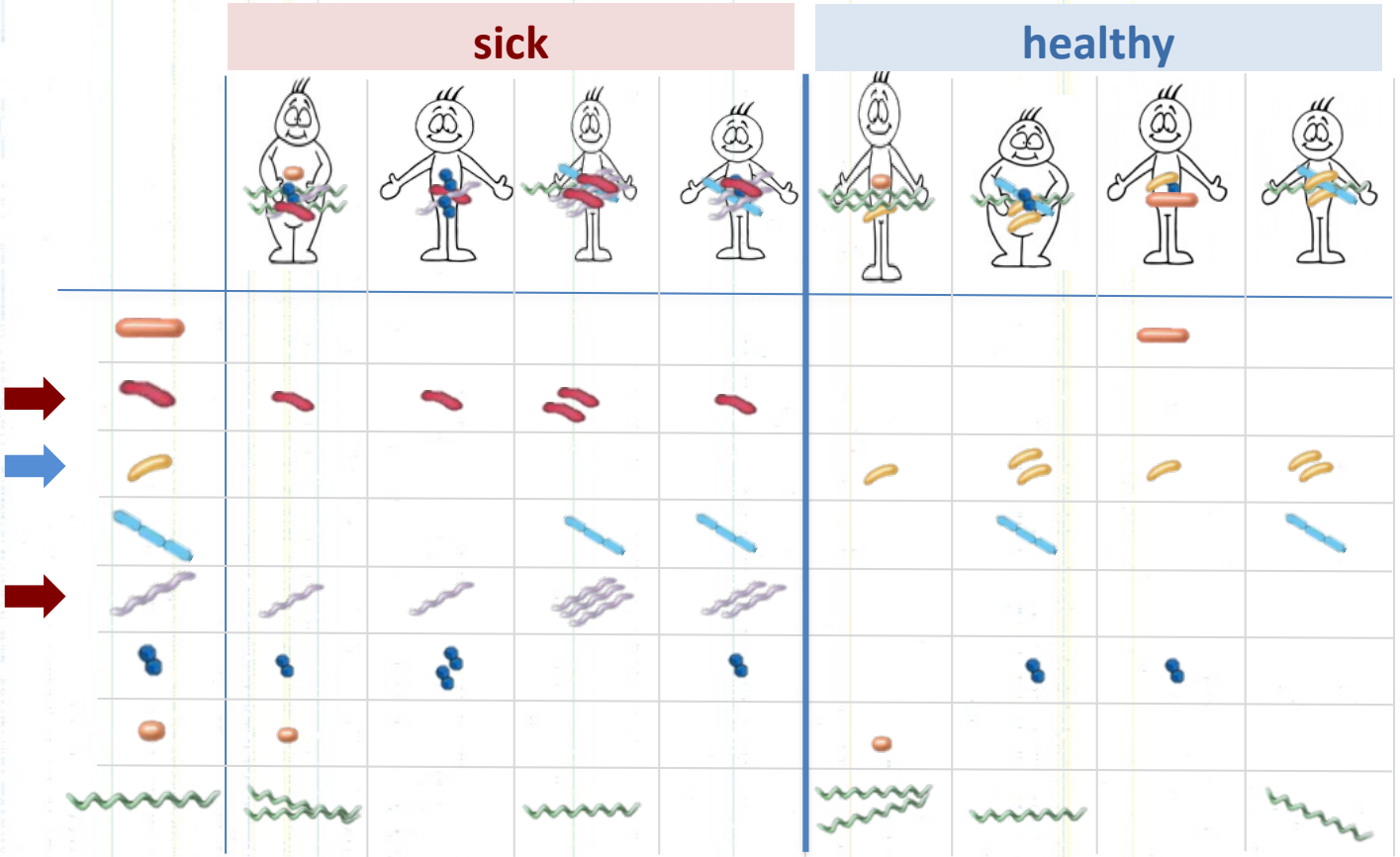
teams

metagenopolis mgps.eu

INTEGROMICS

Le Chatelier et al Nature, 2013
Cotillard et al Nature, 2013
Prifti et al. IHMC, 2013
Nielsen, et al. Nat Biotech. 2014
Li et al. Nat Biotech 2014
Almeida et al The ISME Journal, 2016

The gut **microbiome offers** important **disease classification**



Stratification of the individuals & personalised medicine.

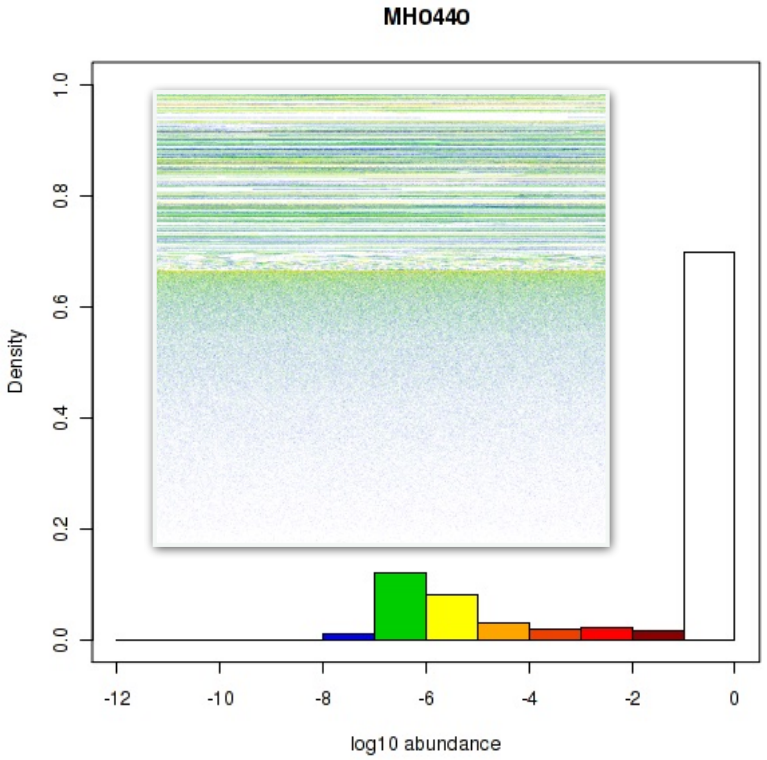
Health *vs.* disease biomarkers.

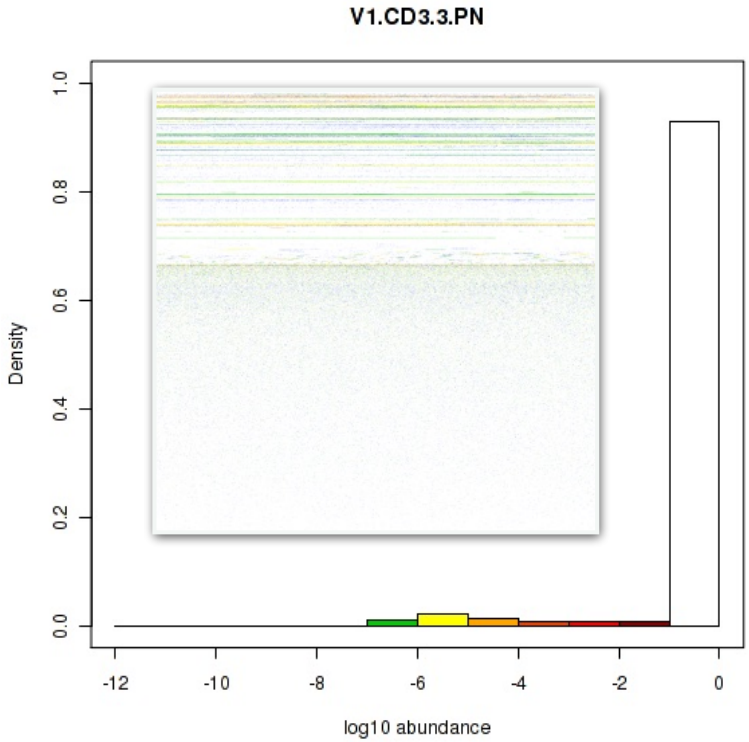
identify microbiota signature used as a diagnostic or prognostic tools.

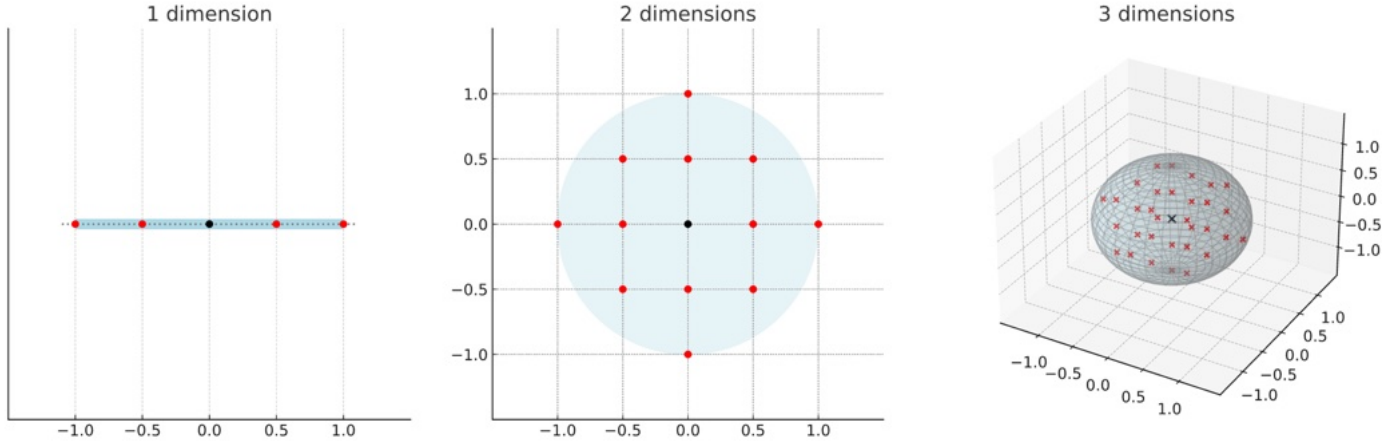
Interventional medicine.

modify the gut ecosystem using prebiotics, probiotics, microbiota transplantation or microbiome targeted drugs.

These are **not straightforward** data to work with. They are **compositional** and full of **empty space**.





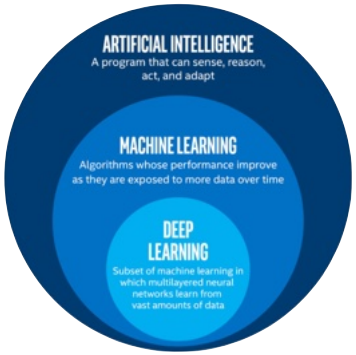


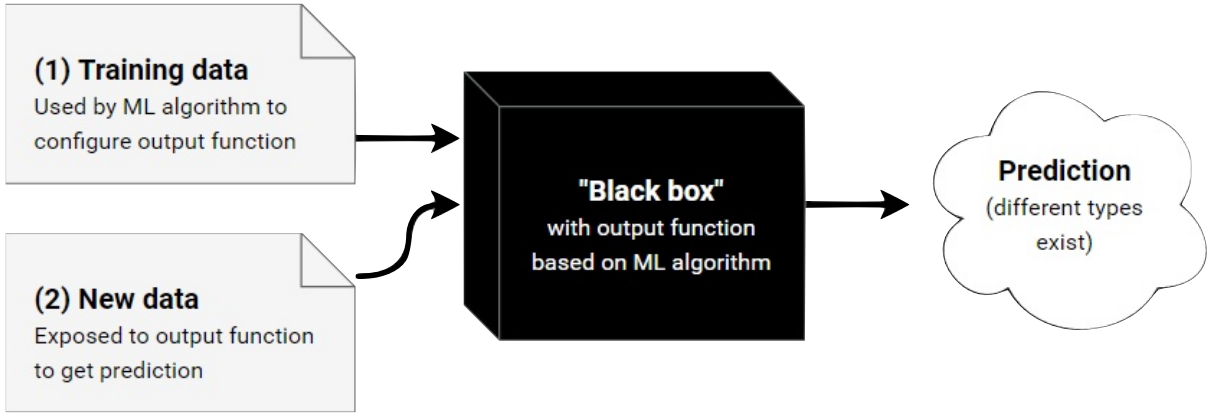
10 000 000 dimensions ?

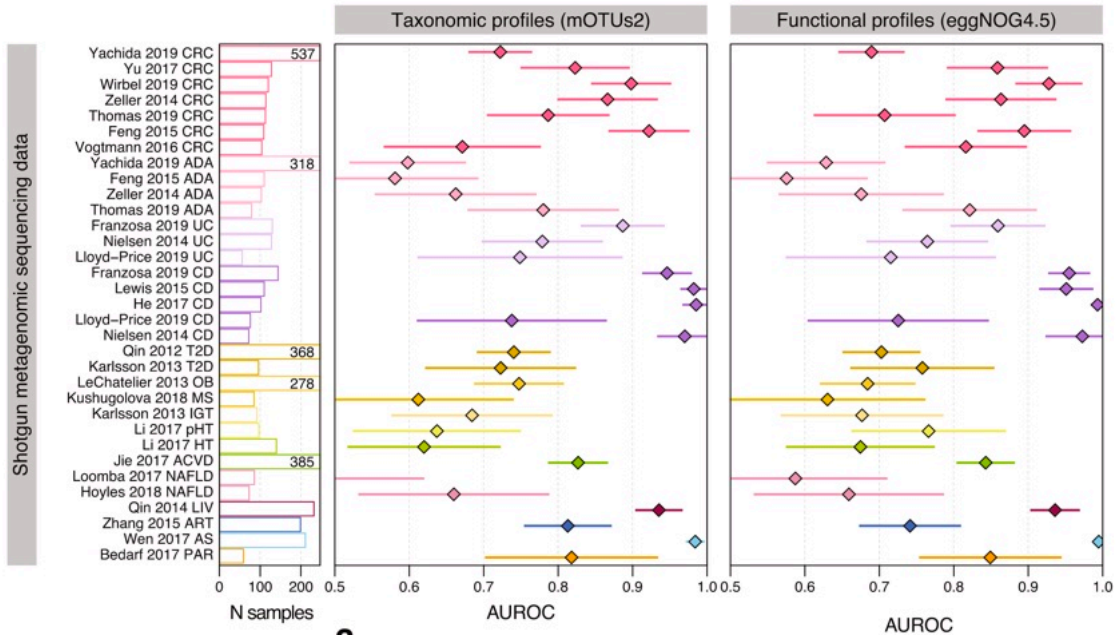
5000 dimensions ?

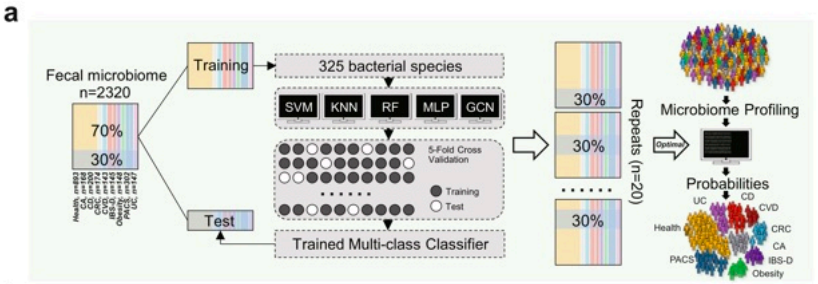
Using the gut microbiome and AI to find **novel biomarkers**

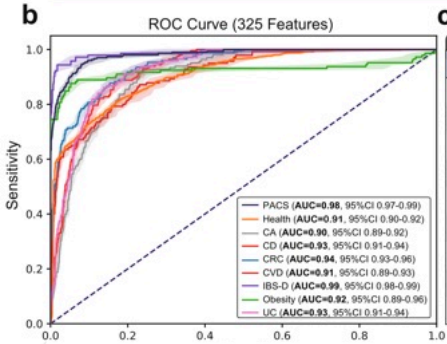
...











c

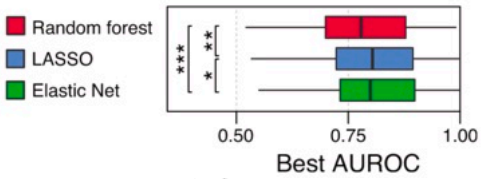
Thresholds at highest Youden index

Phenotype	Threshold	Sensitivity	Specificity	Accuracy	Youden index
Health	0.36586	0.81	0.83	0.82	0.64
CA	0.08028	0.93	0.76	0.77	0.69
CD	0.09361	0.88	0.83	0.83	0.71
CRC	0.09666	0.88	0.85	0.85	0.72
CVD	0.06786	0.87	0.77	0.78	0.64
IBS-D	0.12423	0.94	0.98	0.98	0.93
Obesity	0.10805	0.88	0.95	0.94	0.82
PACS	0.15348	0.95	0.92	0.92	0.87
UC	0.09891	0.86	0.86	0.86	0.72

SOFTWARE

Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox

Jakob Wirbel[1], Konrad Zych[1,2], Morgan Essex[1,3], Nicolai Karcher[1,4], Ece Kartal[1], Guillem Salazar[5], Peer Bork[1,6,7,8], Shinichi Sunagawa[5] and Georg Zeller[1*]



nature communications

Article

https://doi.org/10.1038/s41467-022-34405-3

Faecal microbiome-based machine learning for multi-class disease diagnosis

Received: 2 September 2022

Accepted: 21 October 2022

Published online: 10 November 2022

Qi Su[1,2,3,4,6], Qin Liu[1,2,3,4,6], Raphaela Iris Lau[1,2,3], Jingwan Zhang[1,2,3,4], Zhilu Xu[1,2,3,4], Yun Kit Yeoh[1], Thomas W. H. Leung[2], Whitney Tang[1,2,3], Lin Zhang[1,2,3,4], Jessie Q. Y. Liang[2,3,4], Yuk Kam Yau[1,2,3], Jiaying Zheng[1,2,3], Chengyu Liu[1,2,3], Mengjing Zhang[1,2,3], Chun Pan Cheung[1,2,4], Jessica Y. L. Ching[1,2,3], Hein M. Tun[1,3,5], Jun Yu[2,3,4], Francis K. L. Chan[1,2,3,4] &

Back to the drawing board !

Develop a method that is Interpretable by nature, full of constraints and thought for the microbiome data

OXFORD (GIGA)n SCIENCE

GigaScience, 9, 2020, 1–11

doi: 10.1093/gigascience/giaa010
Research

RESEARCH

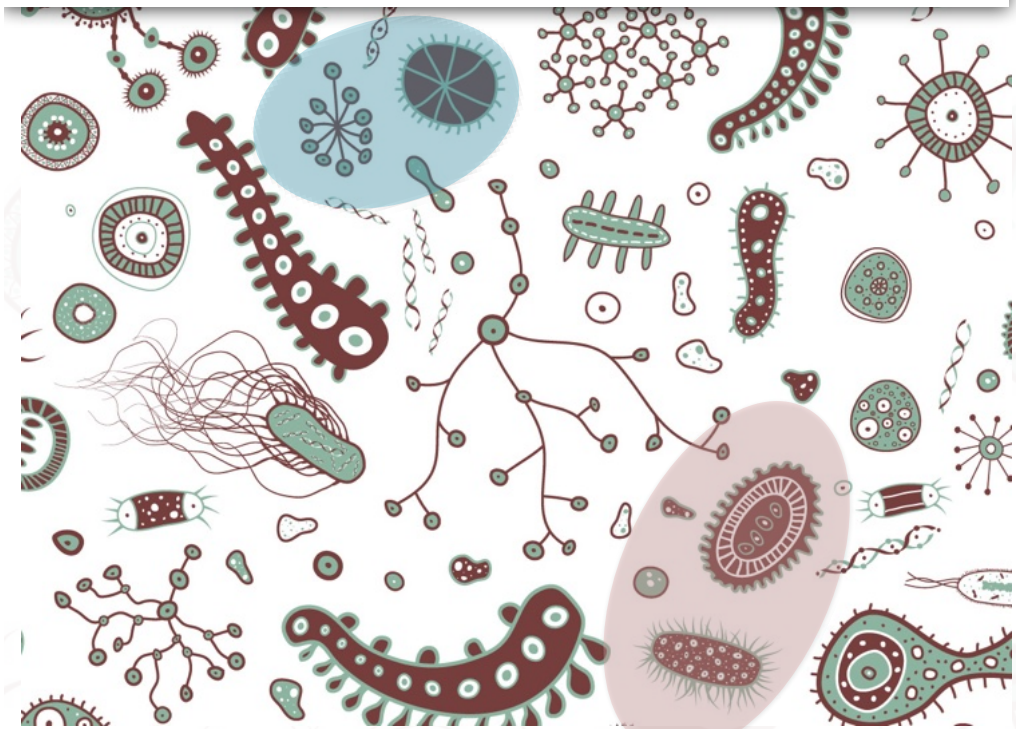
Interpretable and accurate prediction models for metagenomics data

Edi Prifti [1,2,*], Yann Chevaleyre[3], Blaise Hanczar[4], Eugeni Belda[2], Antoine Danchin[5], Karine Clément[6,7] and Jean-Daniel Zucker [1,2,6,*]

[1]IRD, Sorbonne University, UMMISCO, 32 Avenue Henri Varagnat, F-93143 Bondy, France; [2]Institute of Cardiometabolism and Nutrition, ICAN, Integromics, 91 Boulevard de l'Hopital, F-75013, Paris, France;

https://academic.oup.com/gigascience/article/9/3/giaa010/5801229

#Veillonella unclassified +#Clostridium perfringens
< 18.0 % then class = healthy



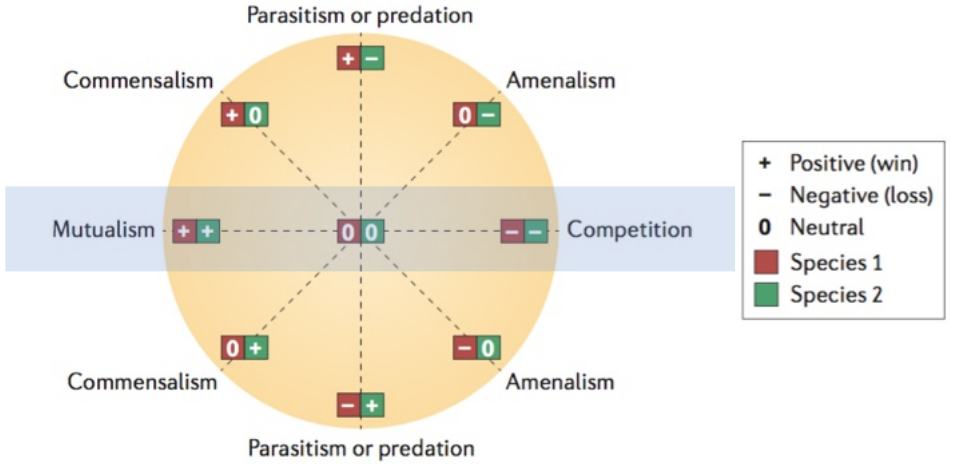


Figure 1 | **Summary of ecological interactions between members of different species.** The wheel display introduced by Lidicker[1] has been adapted to summarize all possible pairwise interactions. For each interaction partner, there are three possible outcomes: positive (+), negative (–) and neutral (0). For instance, in parasitism, the parasite benefits from the relationship (+), whereas the host is harmed (–); this relationship is thus represented by the symbol pair +–.

Faust, Raes. Nature Reviews Microbiology 10, no. 8 (August 1, 2012): 538–550.

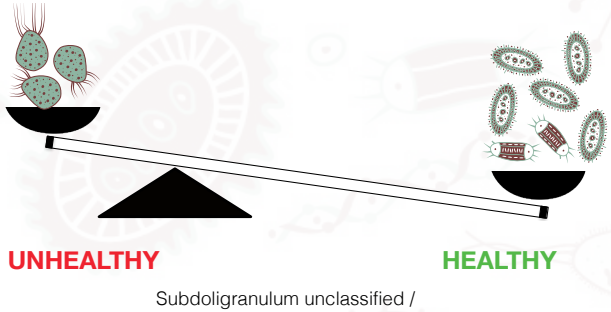
Inspired by ecological relations we proposed the **Bin/Ter/Ratio** (BTR) models, which quantify interactions by comparing abundance.

Veillonella unclassified + Clostridium perfringens
< 18.0 % then class = healthy

BINARY

Streptococcus_anginosus + Veillonella_unclassified - Alistipes_indistinctus
< 8.3 % then class = healthy

TERNARY



Subdoligranulum unclassified /
(Megasphera micronuciformis + Streptococcus anginosus)
> 81 then class = healthy

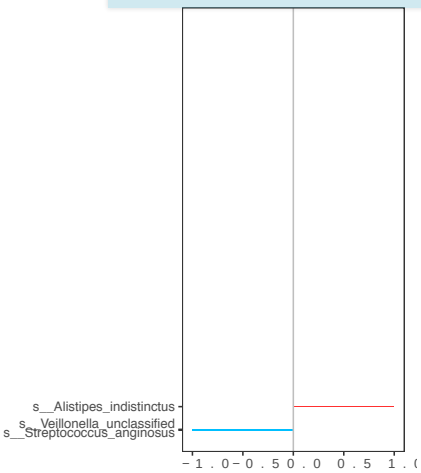
RATIO

- Prifti et al. GigaScience, 2020
- Nguyen, Prifti et al. Sci Rep, 2020.

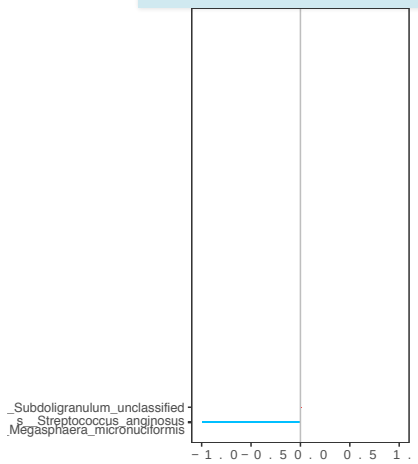
BTR models are **extremely simple** …

A + B > *s*

A + B - C > *s*

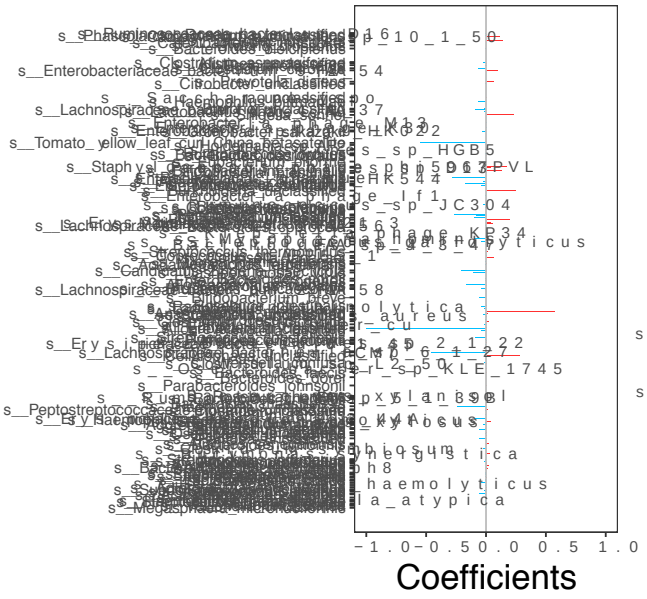


A / (B + C) > *s*

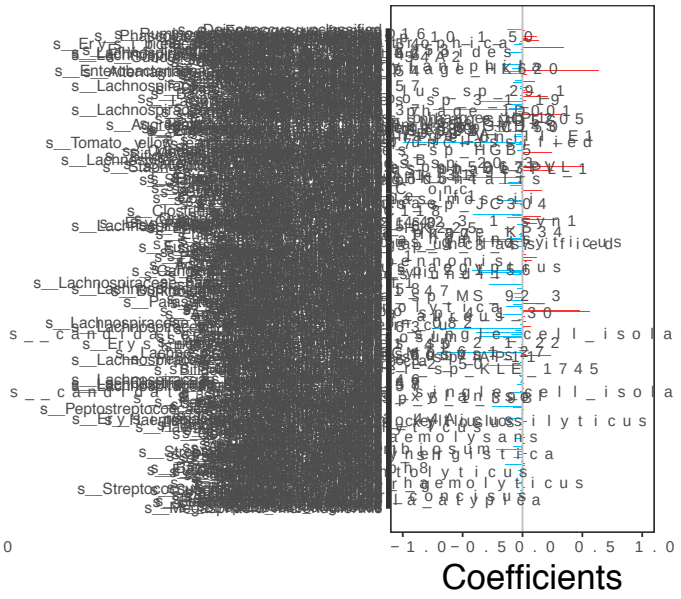


… compared with **SOTA**

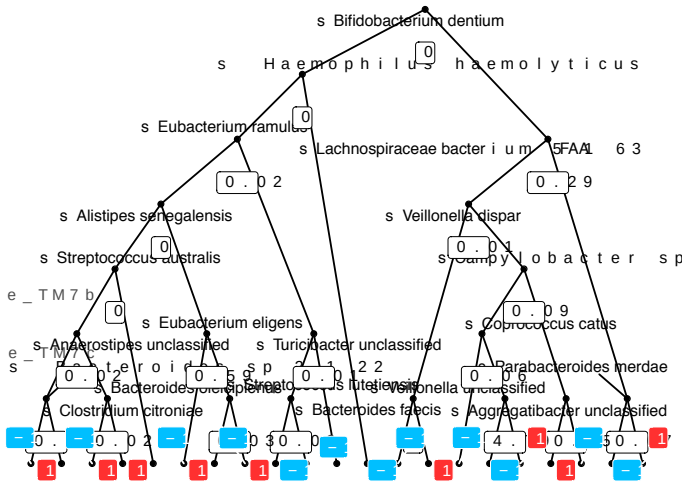
D ENET



E SVM



F RF (1 of 500



ARTICLE

doi:10.1038/nature13568

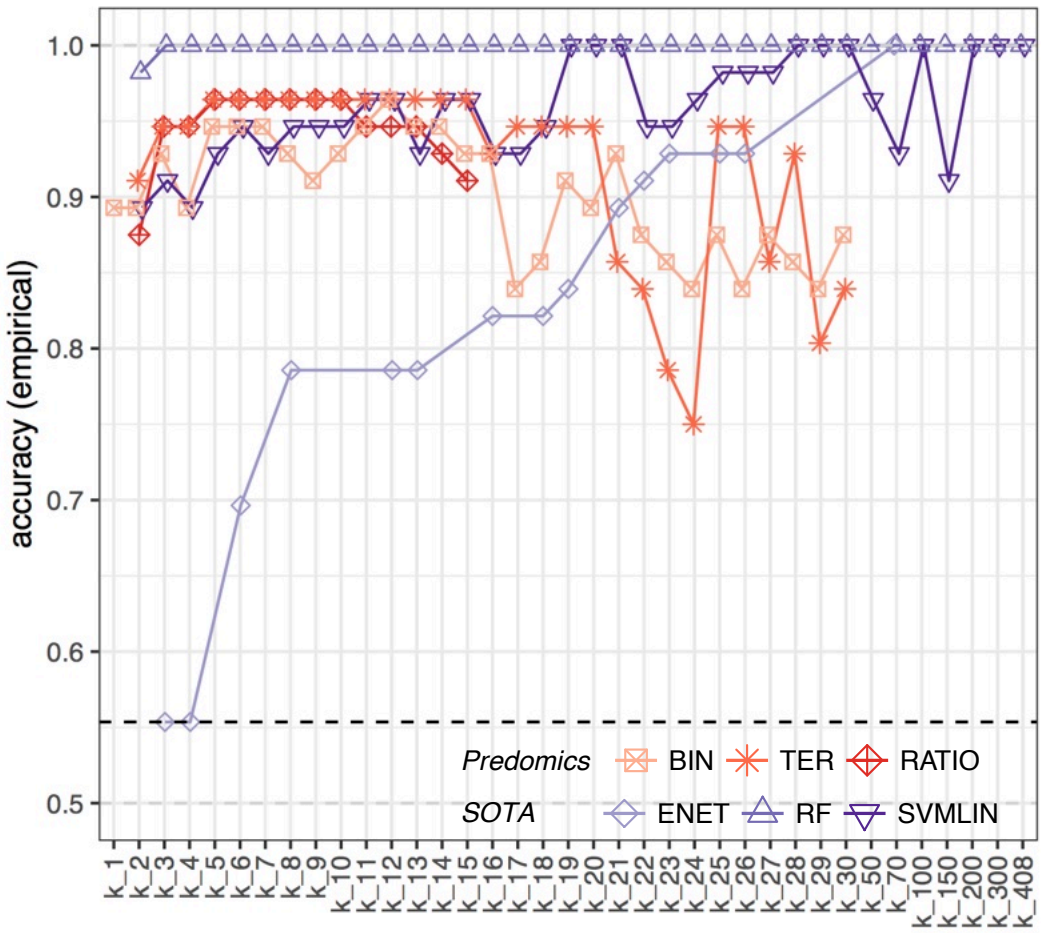
Alterations of the human gut microbiome in liver cirrhosis

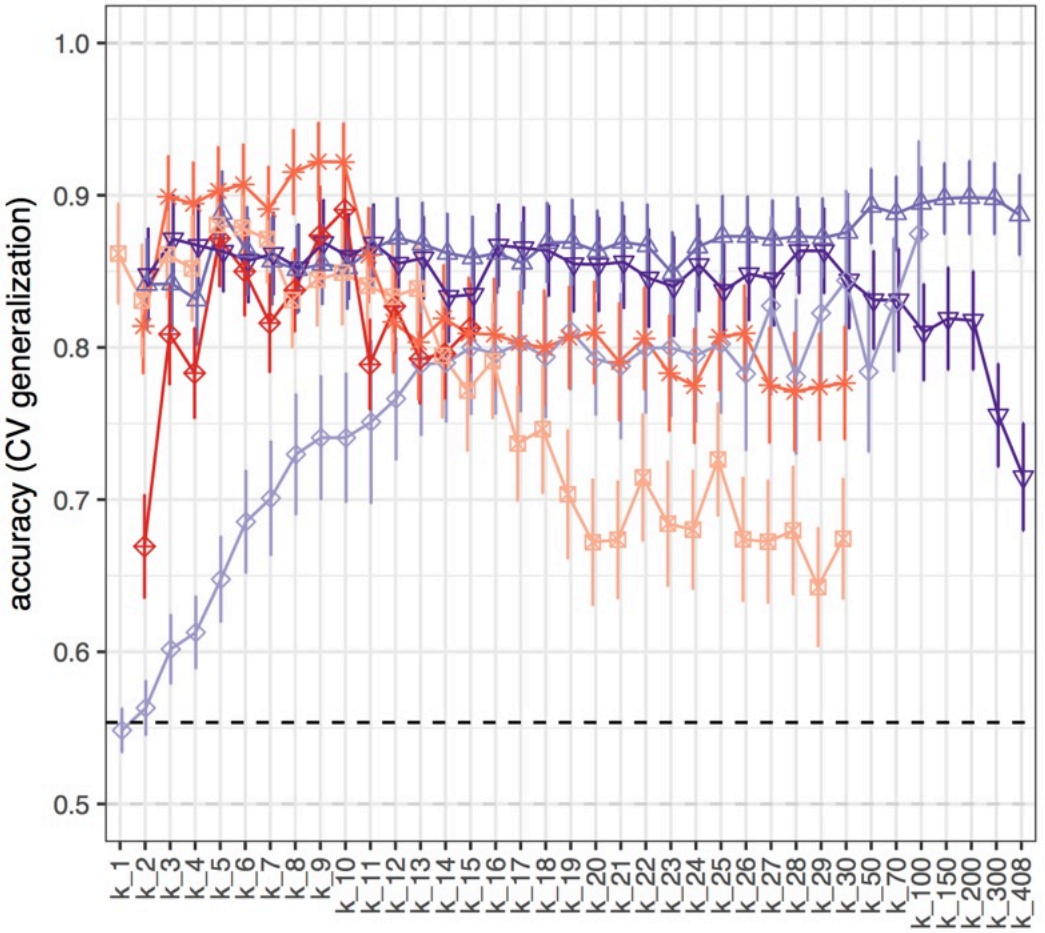
Nan Qin[1,2*], Fengling Yang[1*], Ang Li[1*], Edi Prifti[3*], Yanfei Chen[1*], Li Shao[1,2*], Jing Guo[1], Emmanuelle Le Chatelier[3], Jian Yao[1,2], Lingjiao Wu[1], Jiawei Zhou[1], Shujun Ni[1], Lin Liu[1], Nicolas Pons[3], Jean Michel Batto[3], Sean P. Kennedy[3], Pierre Leonard[3], Chunhui Yuan[1], Wenchao Ding[1], Yuanting Chen[1], Xinjun Hu[1], Beiwen Zheng[1,2], Guirong Qian[1], Wei Xu[1], S. Dusko Ehrlich[3,4], Shusen Zheng[2,5] & Lanjuan Li[1,2]

Accessible, curated metagenomic data through ExperimentHub

Edoardo Pasolli[*,1], Lucas Schiffer[*,2], Audrey Renson[2], Valerie Obenchain[3], Paolo Manghi[1], Duy Tin Truong[1], Francesco Beghini[1], Faizan Malik[2], Marcel Ramos[2], Jennifer B. Dowd[2,4], Curtis Huttenhower[5,6], Martin Morgan[3], Nicola Segata[^,1], Levi Waldron[^,2]

Cirrhosis stage 2 - species dataset

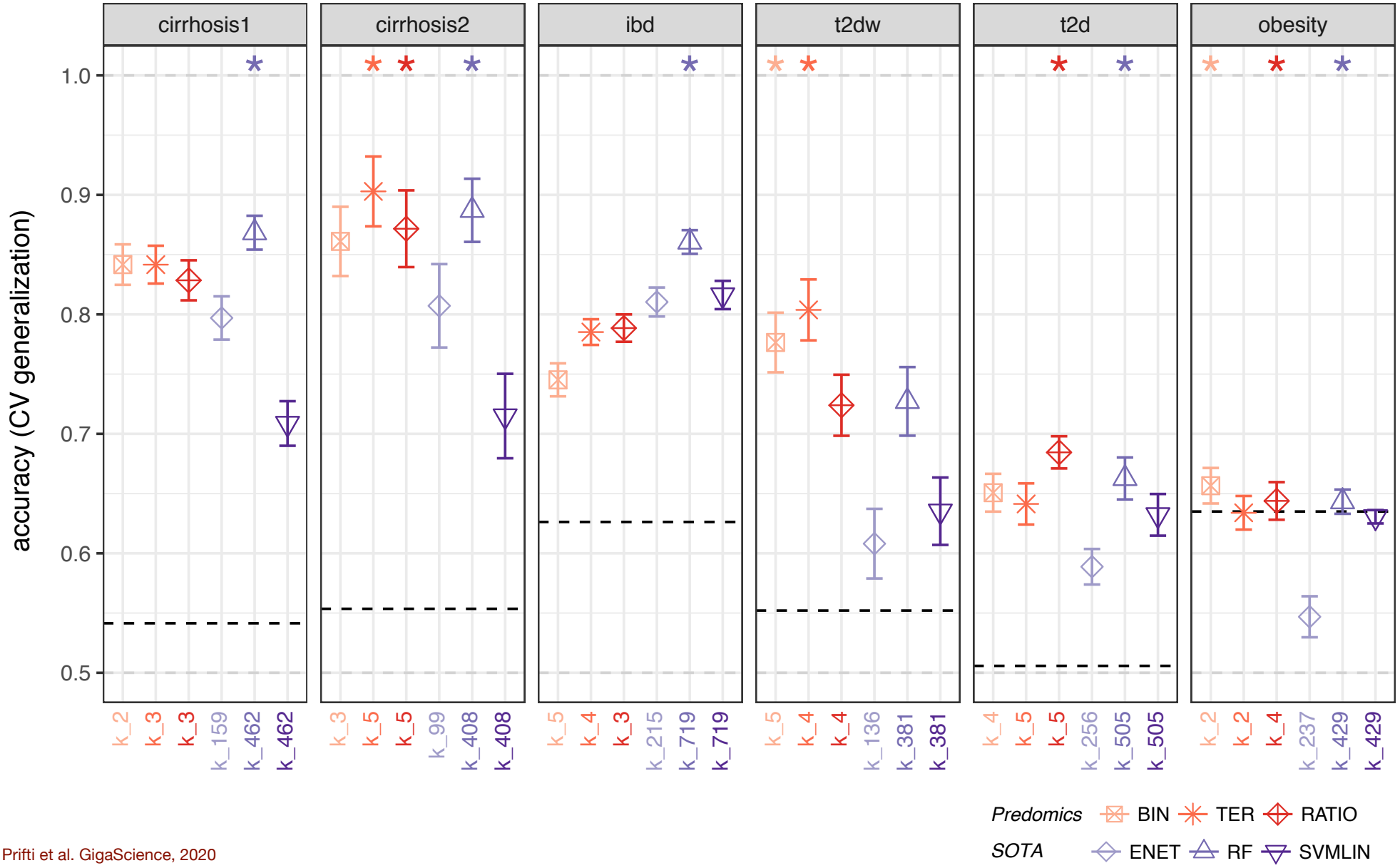




- large exploration of models
- k-selection based on training db using penalty (acc-k*penalty); penalty = 0.1/100
- cross-validation of the algorithms = 10*10 fold; statistical tests possible

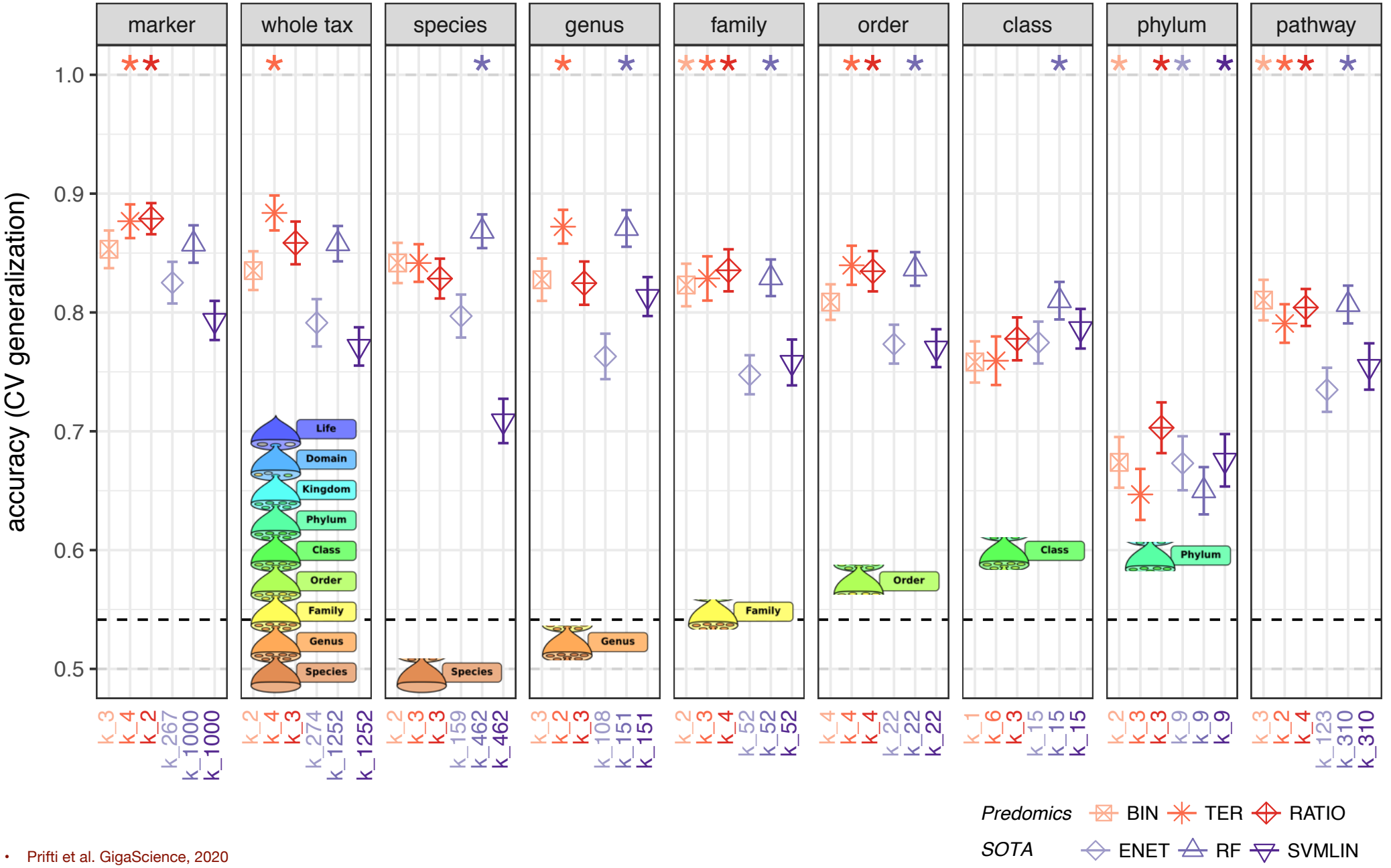
- Prifti et al. GigaScience, 2020

BTR models display **similar performance** with SOTA across datasets



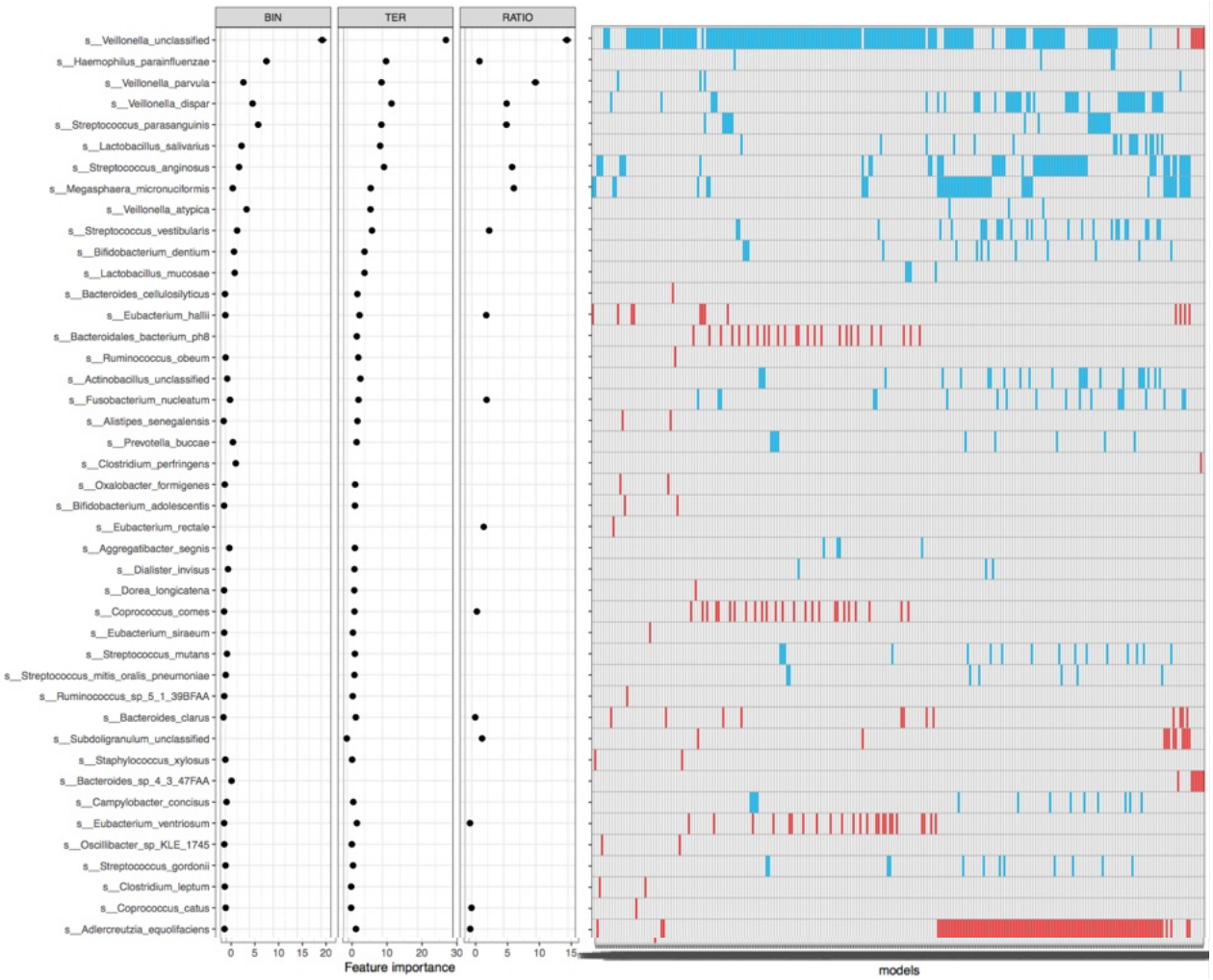
- Prifti et al. GigaScience, 2020

Predictive performance **decreases with taxonomic specificity** (Cirrhosis db)

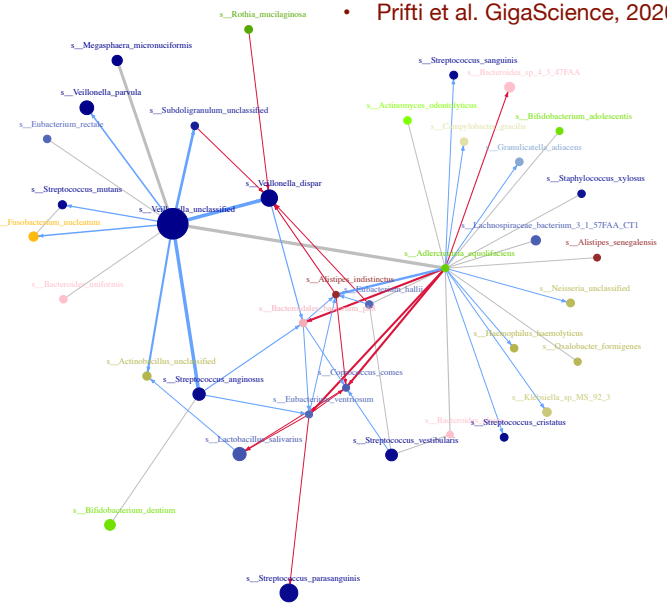


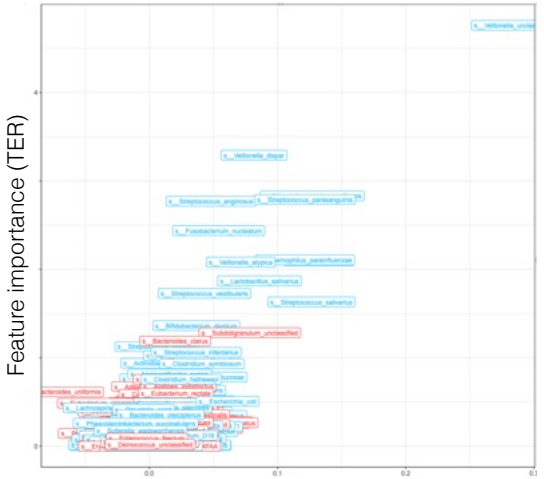
- Prifti et al. GigaScience, 2020

The analysis of the Family of best models (FBM) **provides the most predictive features**



- Prifti et al. GigaScience, 2020





main collaborators


Jean-Daniel Zucker


Yann Chevaleyre


Blaise Hanczar

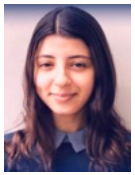
undergrads (algorithmic)


Lucas Robin

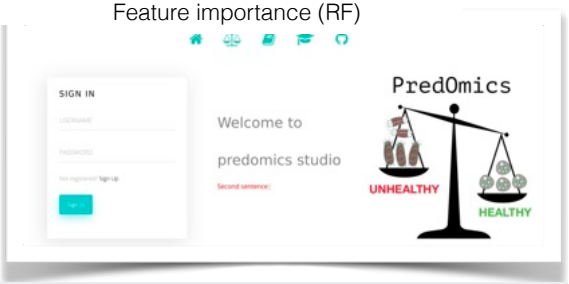

Shasha Cui


Magali C. Thorez

Predomics App


Fatima Souafi

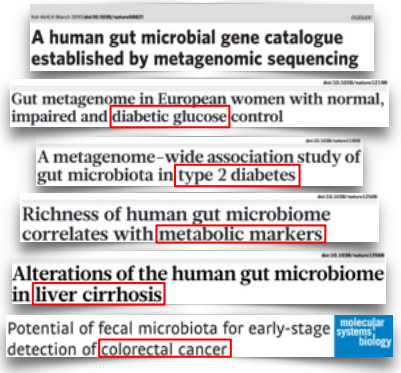

Youcef Sklab

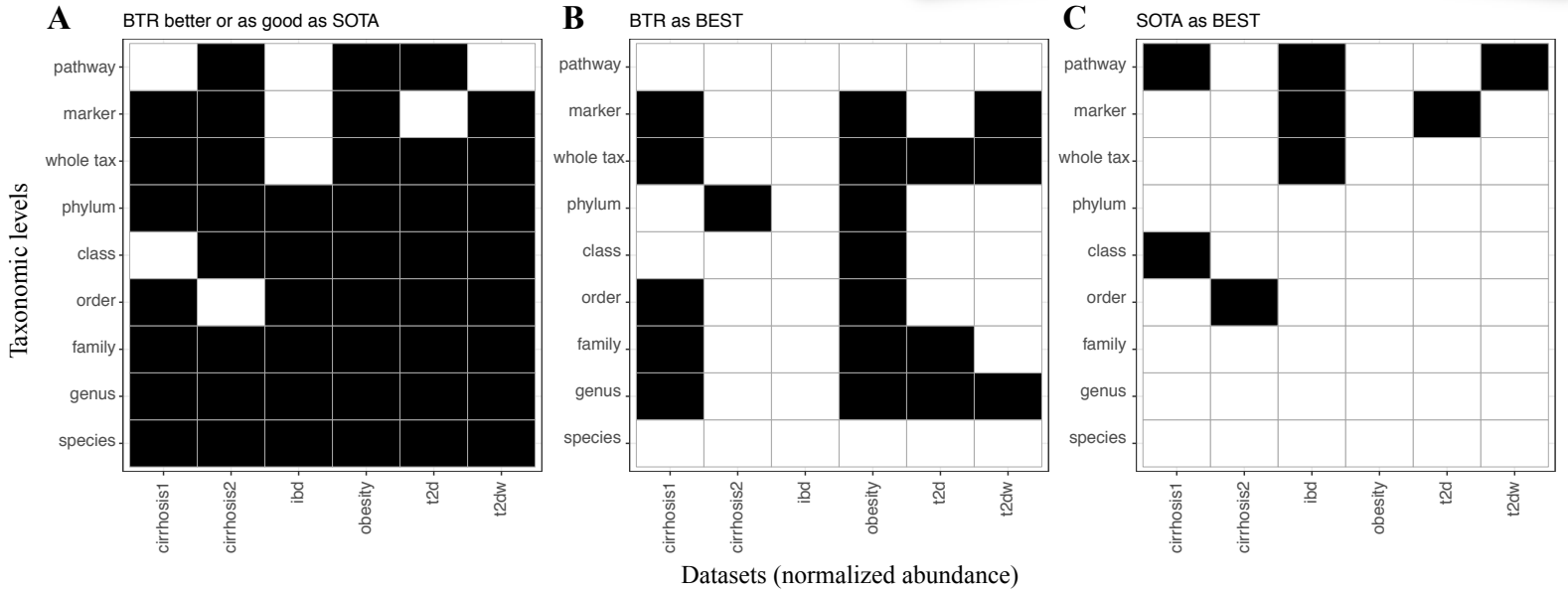


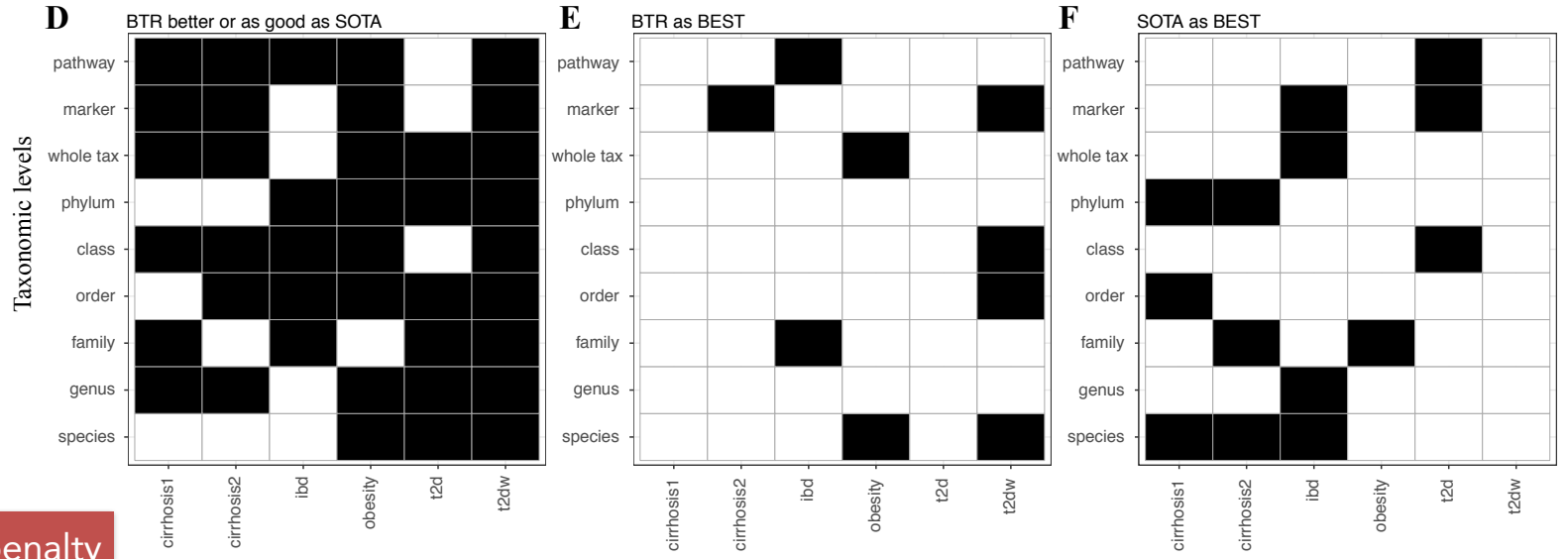
Predictive performance **decreases with taxonomic specificity** (Cirrhosis db)

Accessible, curated metagenomic data through ExperimentHub

Edoardo Pasolli[*,1], Lucas Schiffer[*,2], Audrey Renson[2], Valerie Obenchain[3], Paolo Manghi[1], Duy Tin Truong[1], Francesco Beghini[1], Faizan Malik[2], Marcel Ramos[2], Jennifer B. Dowd[2,4], Curtis Huttenhower[5,6], Martin Morgan[3], Nicola Segata[^,1], Levi Waldron[^,2]





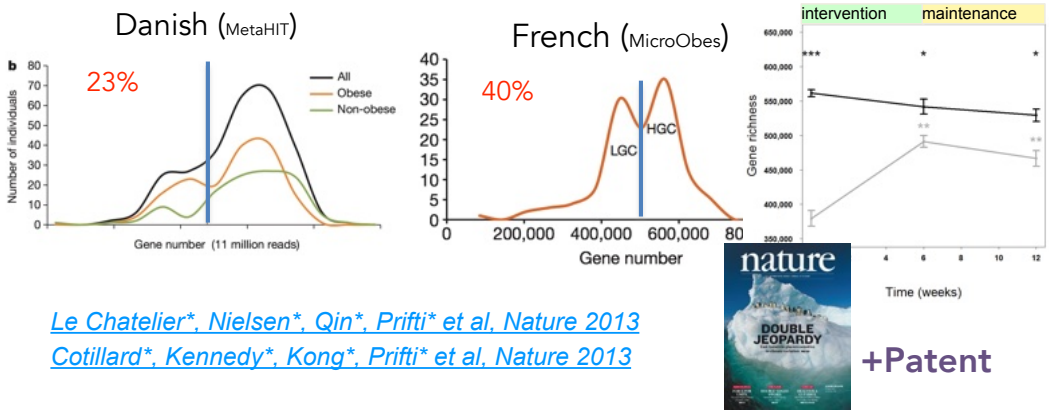


w. penalty

- Prifti et al. GigaScience, 2020

We have applied these models with **success in numerous human cardiometabolic diseases**

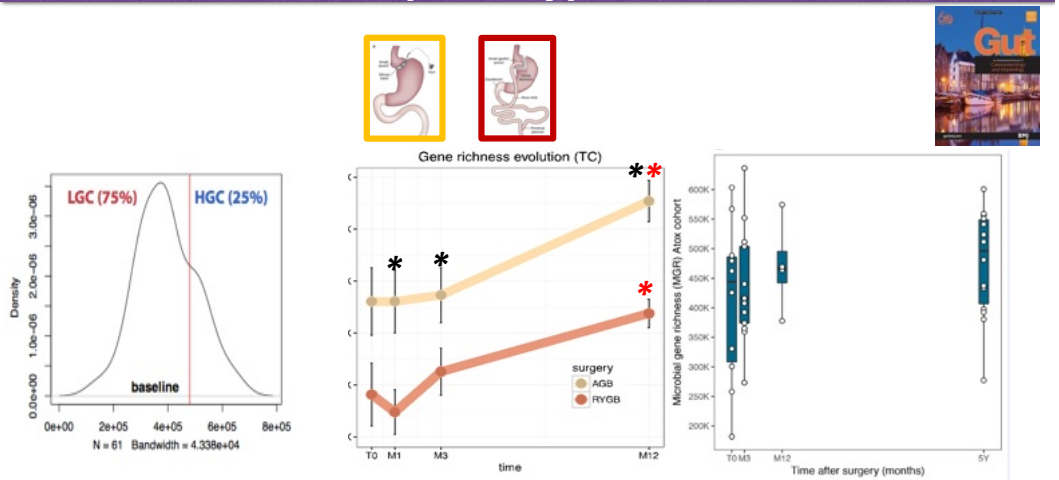
Discovery of the bimodal distribution of gut microbiome richness and its association with the success of dietary interventions



+Patent

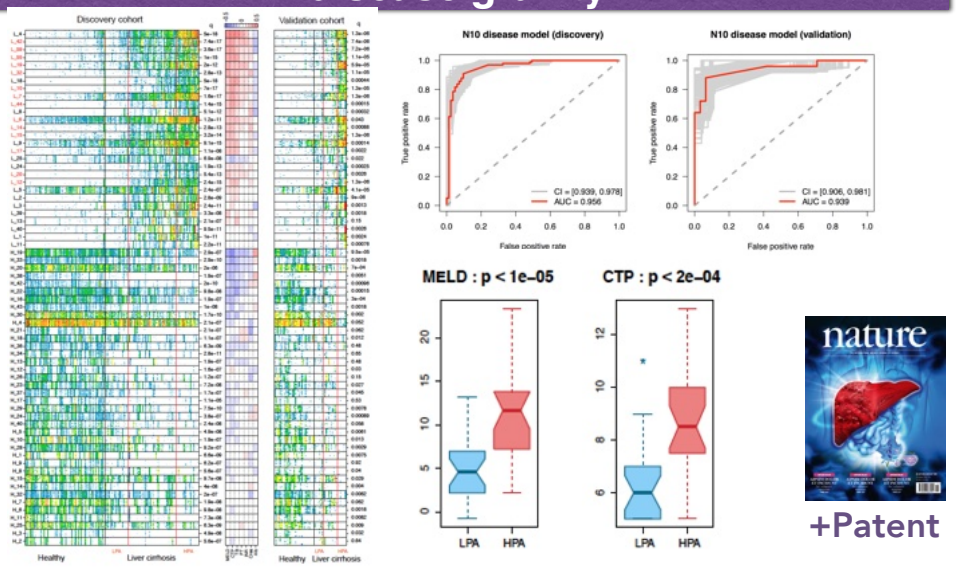
[Le Chatelier*, Nielsen*, Qin*, Prifti* et al, Nature 2013]
[Cotillard*, Kennedy*, Kong*, Prifti* et al, Nature 2013]

Bariatric surgery improves gut microbial richness, although it can not entirely restore it as other phenotypes.



[Aron-Wisnewsky J*, Prifti E*, Belda E*…Clement K Gut 2018;]
[Al Assal K*, Prifti E*, et al. Nutrients. 2020;]
[Lau E, …Prifti E., Diabetol Metab Syndr. 2021]

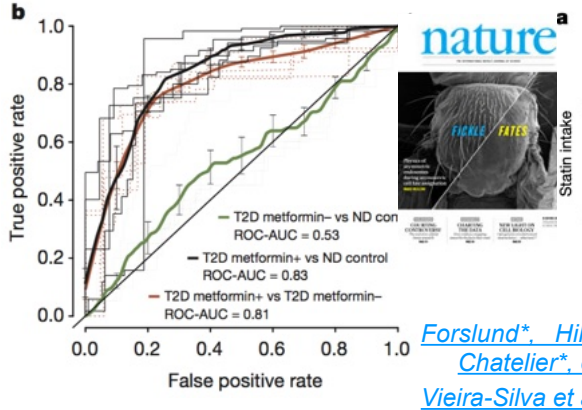
Liver cirrhosis patients have an oral originated microbiome, which is altered and can predict disease gravity.

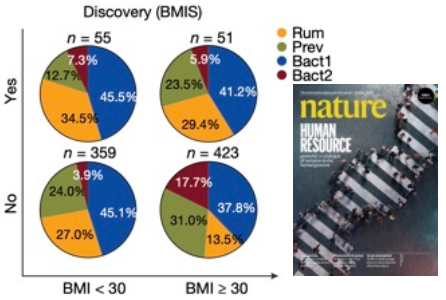


+Patent

[Qin*, Yang*, Li*, Prifti* et al, Nature, 2014]

Medications confounding effects should be taken into account compte (metformin in T2D).

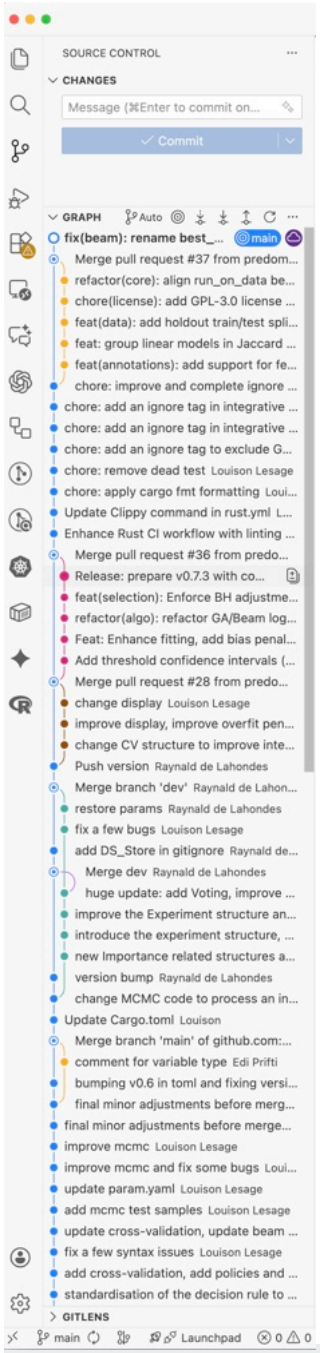


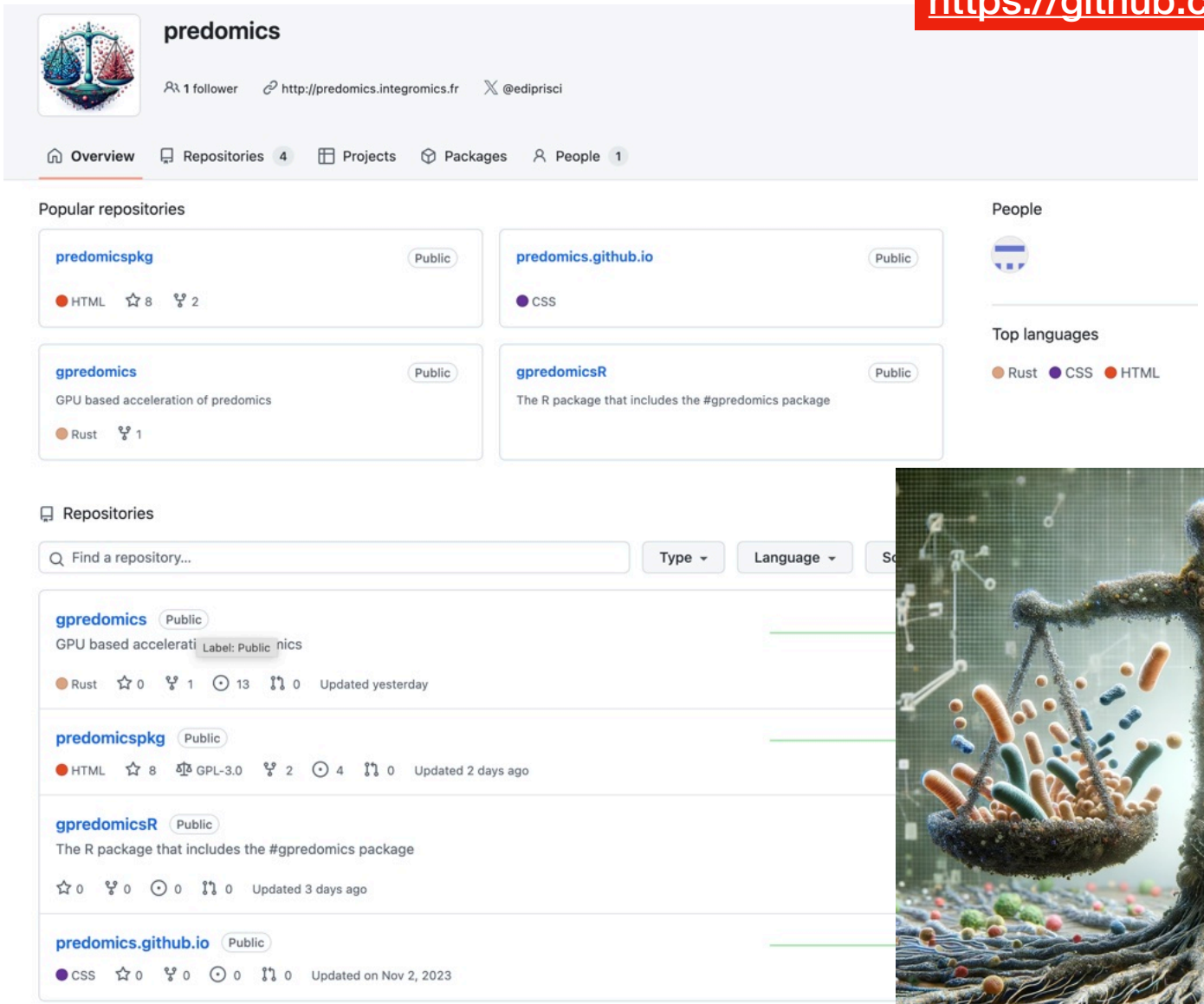


[Forslund*, Hildebrand*, Nielsen*, Falony*, Le Chatelier*, et al, Nature, 2014]
[Vieira-Silva et al Nature, 2020]

An **open source project** which is actively improving (multiclass, GPU, R and python integration, Shiny app and more) …

https://github.com/predomics/

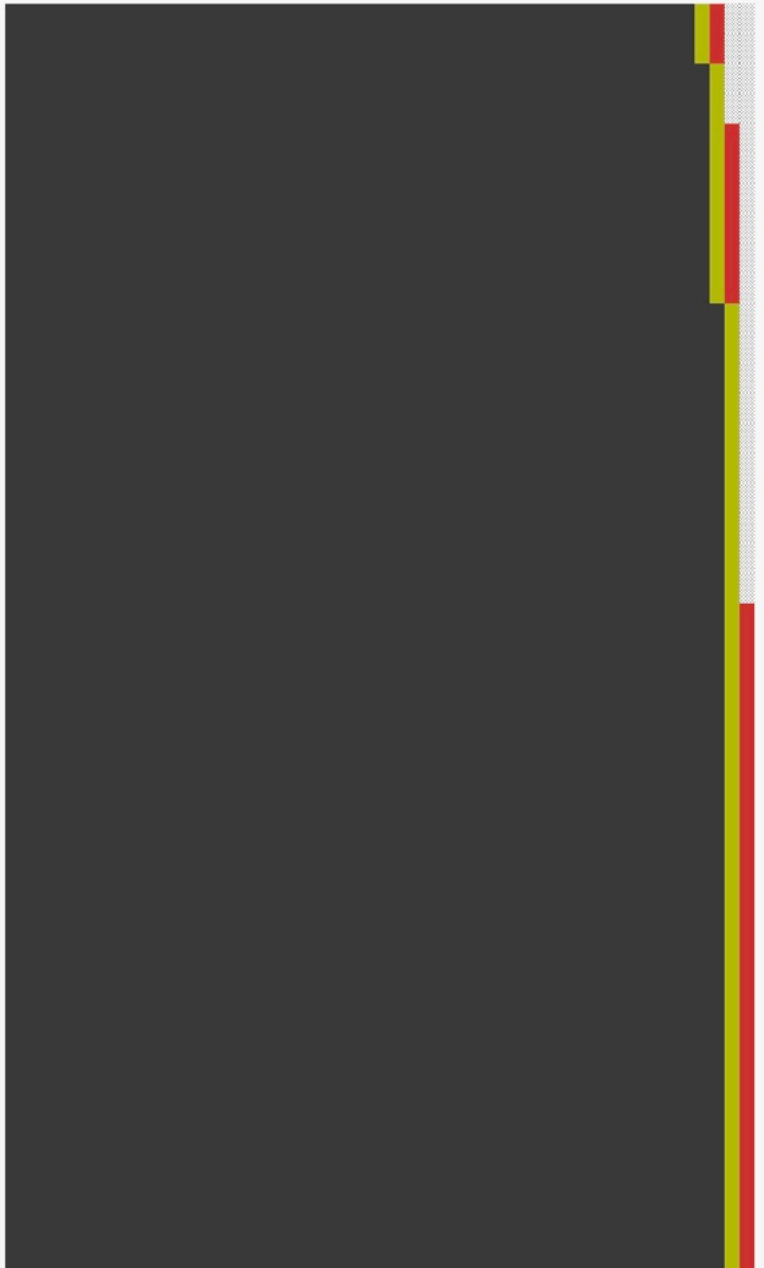




The gpredomics parameter space

```yaml
general:
  seed: 42                                  # used in parent selection, child conception (cross over) and mutation, all of which is single thread
  algo: ga                                  # ga for genetic algorithm, beam for beam algorithm and mcmc for MCMC-based algorithm
  cv: false                                 # should cross-validation be enabled?
  thread_number: 8                          # the number of thread used in feature selection and fit computation
  gpu: true                                 # should Gpredomics use GPU ? (ga and beam only)
  language: ter,ratio                 # possible values are ter,bin,ratio,pow2, see README.md for detail. A comma separated list (no spaces) is accepted, which means the initial population
  data_type: raw,prev                     # possible values are raw,prev,log, see README.md for detail. Same as above, comma separated list is fine.
  epsilon: 1e-5                             # this is only usefull for data_type prevalence (where it is a threshold) or log (where it replaces values below)
  fit: auc                                  # possible values are auc,specificity,sensitivity (classification), see README.md for details
  k_penalty: 0.0001                         # this penalty is deduced from fit function multiplied by k, the number of variables used in the model
  fr_penalty: 0.0                           # used only when fit is specificity or sensitivity, deduce (1 - symetrical metrics) x fr_penalty to fit
  nb_best_model_to_test: 100                # nb of models to test in the last generation (default to 10, 0 means all models)
  #log_base: ""                             # uncomment to print log and results in log_file_name
  log_level: info                          # possible values are trace, debug, info, warning or error
  display_level: 2                          # precision in variable display (0=anonymized features, 1=feature line index, 2=feature names (default))
  display_colorful: true                    # should the terminal results be coloured to make them easier to read?
  keep_trace: false                         # keep this setting to false when using gpredomics as a binary
  save_exp: exp.mp                         # uncomment to save experiment in timestamp-save_exp, which can be reloaded with --load timestamp-save_exp. Extention should be .json, .mp/.msg

data:
  X: "../../../data/X/nsatDisc.tsv"                          # the features of the train data set
  y: "../../../data/y/yDisc_OS12p.tsv"                          # the class description of the train data set (0=class 0, 1=class 1 (the class to be predicted), 2=unknown status)
  Xtest: "../../../data/X/nsatValid.tsv"                       # the features of the test data set
  ytest: "../../../data/y/yValid_OS12p.tsv"                       # the class description of the test data set
  features_maximal_number_per_class: 0      # 0: all significant features ; else first X significant features (per class!) sorted according to their pvalue/log_abs_bayes_factor
  feature_minimal_prevalence_pct: 10        # per class, e.g. features are retained if any of the class reaches this level
  feature_minimal_feature_value: 1e-4       # features which mean is below that value are discarded
  feature_selection_method: wilcoxon        # possible values are wilcoxon, studentt and bayesian_fisher. wilcoxon is recommanded in most cases.
  feature_maximal_pvalue: 0.05              # features with differences less significant (p value above that threshold) than this will be removed
  feature_minimal_log_abs_bayes_factor: 2   # features with a fewer log absolute bayes factor will be removed (bayesian method only)
  classes:
    - "OS12-"
    - "OS12+"
    - "unknown"

cv:
  inner_folds: 5                            # number of folds used to penalize overfit if overfit_penalty > 0
  overfit_penalty: 5                        # this penalty is deduced from fit function (fit = fit on k-1 - abs(delta with last fold) * overfit_penalty)
  outer_folds: 5                            # number of folds used for cross-validation (launch algorithm on each k-1 folds then merge Families of Best Models).
  fit_on_valid: true                        # if true, FBM is based on validation fold fit favouring generalisation, else on k-1 folds ; DO NOT fit on valid without external validation
  cv_best_models_ci_alpha: 0.05             # alpha for the Family of Best Models confidence interval based on the best fit on validation fold. Smaller alpha, larger best_model range.

importance:
  compute_importance: false                 # should importance be computed ?
  n_permutations_oob: 100                   # number of permutations per feature for OOB importance.
  scaled_importance: true                   # scale importance by feature prevalence inside folds.
  importance_aggregation: mean              # aggregation method for importances: "mean" or "median".

voting:
  vote: true                                # should voting be activated ?
  use_fbm: true                             # should only "Family of Best Models" be kept (alpha = 0.05) ?
  min_perf: 0.5                             # required sensitivity AND specificity to be judge ; >=0.5 avoid "single-choice oriented" judges
  min_diversity: 10                         # required diversity between judges
  method: Majority                          # Majority: class 1 if > threshold, else class 0 (if = no classification) VS Consensus: non-classification (rejection) if the threshold is not
```

```
INFO [gpredomics] Training using Genetic Algorithm
------------------------------------------------------
INFO [gpredomics::data] Selecting features...
INFO [gpredomics::data] 447 features selected
INFO [gpredomics::ga] Population size: 4991, kmin 1, kmax 199
INFO [gpredomics::ga] Legend:    [≠ diversity filter]    [↺ resampling]    [▮: AUC]    [▮: penalized fit]
```



```
INFO [gpredomics::ga] #1       | best: Ternary:Log            0 1 [k=99, age=0]
INFO [gpredomics::ga] #2       | best: Ternary:Log            0 1 [k=99, age=1]
INFO [gpredomics::ga] #3       | best: Ternary:Log            0 1 [k=80, age=0]
INFO [gpredomics::ga] #4       | best: Ternary:Log            0 1 [k=92, age=0]
INFO [gpredomics::ga] #5       | best: Ternary:Log            0 1 [k=98, age=0]
INFO [gpredomics::ga] #6       | best: Ternary:Log            0 1 [k=98, age=1]
INFO [gpredomics::ga] #7       | best: Ternary:Log            0 1 [k=98, age=2]
INFO [gpredomics::ga] #8       | best: Ternary:Log            0 1 [k=98, age=3]
INFO [gpredomics::ga] #9       | best: Ternary:Log            0 1 [k=98, age=4]
INFO [gpredomics::ga] #10      | best: Pow2:Log               0 1 [k=112, age=0]
INFO [gpredomics::ga] #11      | best: Ternary:Log            0 1 [k=110, age=0]
INFO [gpredomics::ga] #12      | best: Ternary:Log            0 1 [k=110, age=1]
INFO [gpredomics::ga] #13      | best: Ternary:Log            0 1 [k=113, age=0]
INFO [gpredomics::ga] #14      | best: Ternary:Log            0 1 [k=113, age=1]
INFO [gpredomics::ga] #15      | best: Ternary:Log            0 1 [k=113, age=2]
INFO [gpredomics::ga] #16      | best: Pow2:Log               0 1 [k=108, age=0]
INFO [gpredomics::ga] #17      | best: Pow2:Log               0 1 [k=108, age=1]
INFO [gpredomics::ga] #18      | best: Pow2:Log               0 1 [k=108, age=2]
INFO [gpredomics::ga] #19      | best: Pow2:Log               0 1 [k=108, age=3]
INFO [gpredomics::ga] #20      | best: Pow2:Log               0 1 [k=108, age=4]
INFO [gpredomics::ga] #21      | best: Ternary:Log            0 1 [k=106, age=0]
INFO [gpredomics::ga] #22      | best: Ternary:Log            0 1 [k=106, age=1]
INFO [gpredomics::ga] #23      | best: Ternary:Log            0 1 [k=106, age=2]
INFO [gpredomics::ga] #24      | best: Ternary:Log            0 1 [k=106, age=3]
INFO [gpredomics::ga] #25      | best: Ternary:Log            0 1 [k=106, age=4]
INFO [gpredomics::ga] #26      | best: Ternary:Log            0 1 [k=106, age=5]
INFO [gpredomics::ga] #27      | best: Ternary:Log            0 1 [k=106, age=6]
INFO [gpredomics::ga] #28      | best: Ternary:Log            0 1 [k=106, age=7]
INFO [gpredomics::ga] #29      | best: Ternary:Log            0 1 [k=106, age=8]
INFO [gpredomics::ga] #30      | best: Ternary:Log            0 1 [k=106, age=9]
INFO [gpredomics::ga] #31      | best: Ternary:Log            0 1 [k=122, age=0]
INFO [gpredomics::ga] #32      | best: Ratio:Log              0 1 [k=109, age=0]
INFO [gpredomics::ga] #33      | best: Ratio:Log              0 1 [k=109, age=1]
INFO [gpredomics::ga] #34      | best: Ratio:Log              0 1 [k=109, age=2]
INFO [gpredomics::ga] #35      | best: Ratio:Log              0 1 [k=109, age=3]
INFO [gpredomics::ga] #36      | best: Ratio:Log              0 1 [k=109, age=4]
INFO [gpredomics::ga] #37      | best: Ratio:Log              0 1 [k=109, age=5]
INFO [gpredomics::ga] #38      | best: Pow2:Prevalence        0 1 [k=135, age=0]
INFO [gpredomics::ga] #39      | best: Pow2:Prevalence        0 1 [k=135, age=1]
INFO [gpredomics::ga] #40      | best: Ratio:Log              0 1 [k=122, age=0]
INFO [gpredomics::ga] #41      | best: Ratio:Log              0 1 [k=122, age=1]
INFO [gpredomics::ga] #42      | best: Ratio:Log              0 1 [k=122, age=2]
```

Model #1 Ternary:Prevalence [k=18] [gen:138] [fit:0.868] AUC 0.914/0.533 | accuracy 0.857/0.512 | sensitivity 0.800/0.312 | specificity 0.886/0.708
Class R: (msp_0134⁰ + msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1399⁰ + msp_1461⁰ + msp_1698⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0212⁰ + msp_0850⁰ + msp_0874⁰ + msp_1060c⁰ + omsp_2357⁰) ≥ 2

Model #2 Ternary:Prevalence [k=20] [gen:189] [fit:0.865] AUC 0.889/0.522 | accuracy 0.797/0.465 | sensitivity 0.889/0.469 | specificity 0.750/0.462
Class R: (msp_0208⁰ + msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_0592⁰ + msp_0657⁰ + msp_0785⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1399⁰ + msp_1698⁰ + omsp_1410⁰) - (msp_0850⁰ + msp_1060c⁰ + msp_1556⁰ + omsp_0212⁰ + omsp_2357⁰) ≥ 2

Model #3 Ternary:Prevalence [k=19] [gen:109] [fit:0.864] AUC 0.910/0.497 | accuracy 0.789/0.504 | sensitivity 0.911/0.547 | specificity 0.727/0.462
Class R: (msp_0134⁰ + msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1698⁰ + msp_1729⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0212⁰ + msp_0850⁰ + msp_1060c⁰ + msp_1556⁰ + omsp_0212⁰ + omsp_2357⁰) ≥ 0

Model #4 Ternary:Prevalence [k=17] [gen:190] [fit:0.861] AUC 0.913/0.540 | accuracy 0.850/0.543 | sensitivity 0.778/0.312 | specificity 0.886/0.769
Class R: (msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1399⁰ + msp_1461⁰ + msp_1698⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0212⁰ + msp_0850⁰ + msp_0874⁰ + msp_1060c⁰ + omsp_2357⁰) ≥ 2

Model #5 Ternary:Prevalence [k=20] [gen:190] [fit:0.861] AUC 0.912/0.493 | accuracy 0.842/0.527 | sensitivity 0.778/0.391 | specificity 0.875/0.662
Class R: (msp_0134⁰ + msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1399⁰ + msp_1461⁰ + msp_1698⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0850⁰ + msp_0874⁰ + msp_1060c⁰ + msp_1381⁰ + msp_1556⁰ + omsp_2357⁰) ≥ 1

Model #6 Ternary:Prevalence [k=18] [gen:197] [fit:0.859] AUC 0.905/0.509 | accuracy 0.827/0.512 | sensitivity 0.822/0.422 | specificity 0.830/0.600
Class R: (msp_0134⁰ + msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1399⁰ + msp_1461⁰ + msp_1729⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0212⁰ + msp_0850⁰ + msp_0874⁰ + msp_1060c⁰) ≥ 2

Model #7 Ternary:Prevalence [k=21] [gen:197] [fit:0.858] AUC 0.898/0.536 | accuracy 0.805/0.504 | sensitivity 0.867/0.484 | specificity 0.773/0.523
Class R: (msp_0208⁰ + msp_0254⁰ + msp_0294⁰ + msp_0740⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1296c⁰ + msp_1356⁰ + msp_1461⁰ + msp_1698⁰ + msp_1729⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0212⁰ + msp_0793⁰ + msp_0850⁰ + msp_1060c⁰ + msp_1556⁰ + omsp_0212⁰) ≥ 2

Model #8 Ternary:Prevalence [k=19] [gen:177] [fit:0.858] AUC 0.912/0.514 | accuracy 0.835/0.512 | sensitivity 0.867/0.422 | specificity 0.818/0.600
Class R: (msp_0134⁰ + msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1399⁰ + msp_1461⁰ + msp_1698⁰ + msp_1729⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0212⁰ + msp_0850⁰ + msp_0874⁰ + msp_1060c⁰) ≥ 2

Model #9 Ternary:Prevalence [k=21] [gen:101] [fit:0.856] AUC 0.901/0.521 | accuracy 0.812/0.504 | sensitivity 0.911/0.500 | specificity 0.761/0.508
Class R: (msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_0740⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1296c⁰ + msp_1356⁰ + msp_1461⁰ + msp_1698⁰ + msp_1729⁰ + omsp_1410⁰) - (msp_0173⁰ + msp_0212⁰ + msp_0793⁰ + msp_0850⁰ + msp_1060c⁰ + msp_1556⁰ + omsp_0212⁰) ≥ 1

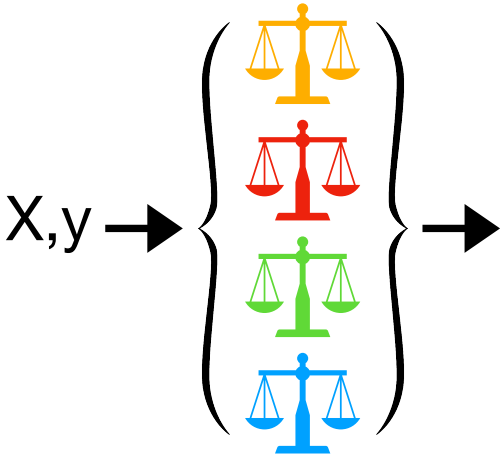
Model #10 Ternary:Prevalence [k=15] [gen:138] [fit:0.854] AUC 0.883/0.485 | accuracy 0.782/0.481 | sensitivity 0.933/0.531 | specificity 0.705/0.431
Class R: (msp_0254⁰ + msp_0294⁰ + msp_0414⁰ + msp_0785⁰ + msp_0942⁰ + msp_1048⁰ + msp_1146⁰ + msp_1244⁰ + msp_1356⁰ + msp_1399⁰ + omsp_1410⁰) - (msp_0850⁰ + msp_1060c⁰ + omsp_0212⁰ + omsp_2357⁰) ≥ 1

```
═══════════════════════════════════════════════════════════════════════════════
                                VOTING ANALYSIS
═══════════════════════════════════════════════════════════════════════════════

Majority jury [14 experts] | AUC 0.923/0.503 | accuracy 0.859/0.504 | sensitivity 0.932/0.484 | specificity 0.821/0.524 | rejection rate 0.038/0.031

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ DETAILED METRICS ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

──────────────── CONFUSION MATRIX (TRAIN) ────────────────

         | Pred 1 | Pred 0 | Abstain
Real 1   |     41 |      3 |       1
Real 0   |     15 |     69 |       4

──────────────── CONFUSION MATRIX (TEST) ─────────────────

         | Pred 1 | Pred 0 | Abstain
Real 1   |     30 |     32 |       2
Real 0   |     30 |     33 |       2


~~~~~~~~~~~~~~~~~~~~~~~~~~~ PREDICTIONS BY SAMPLE (TEST) ~~~~~~~~~~~~~~~~~~~~~~~~~~~

──────────────────────────────────────────────────────────────────────────────────
Sample          | Real | Predictions     | Result | Consistency
──────────────────────────────────────────────────────────────────────────────────
──────── ERRORS (12 shown of 62, sorted by inconsistency) ────────
SAMN37671286    |    0 | 01110110011100 → 1     | ✗      |   57.1%
SAMN37671581    |    1 | 01100110101000 → 0     | ✗      |   57.1%
SAMN37671687    |    0 | 01101010101101 → 1     | ✗      |   57.1%
SAMN37671700    |    0 | 00100100111111 → 1     | ✗      |   57.1%
SAMN37671479    |    0 | 11011111000011 → 1     | ✗      |   64.3%
SAMN37671506    |    0 | 01100110011111 → 1     | ✗      |   64.3%
SAMN37671669    |    1 | 00000000101111 → 0     | ✗      |   64.3%
SAMN37671683    |    1 | 00100010100110 → 0     | ✗      |   64.3%
SAMN37671688    |    0 | 01100110101111 → 1     | ✗      |   64.3%
SAMN37671705    |    0 | 01111110101100 → 1     | ✗      |   64.3%
SAMN37671715    |    1 | 01110000101000 → 0     | ✗      |   64.3%
SAMN37671732    |    0 | 01100110011111 → 1     | ✗      |   64.3%
──────── ABSTENTIONS (4 shown of 4, sorted by inconsistency) ────────
SAMN37671287    |    1 | 00100100011111 → 2 | ~      |   50.0%
SAMN37671511    |    0 | 00101001010111 → 2 | ~      |   50.0%
SAMN37671682    |    1 | 01010010101011 → 2 | ~      |   50.0%
SAMN37671694    |    0 | 01010010101011 → 2 | ~      |   50.0%
──────── CORRECT (4 shown of 63, sorted by inconsistency) ────────
SAMN37671280    |    1 | 00111111100100 → 1 | ✓      |   57.1%
SAMN37671512    |    0 | 01100000101101 → 0 | ✓      |   57.1%
SAMN37671519    |    1 | 10101111001001 → 1 | ✓      |   57.1%
SAMN37671522    |    0 | 00100000101111 → 0 | ✓      |   57.1%
 ... 109 additional samples not shown
```

What else ?

Fabien Kambu

Multi-class classification OVO/OVA

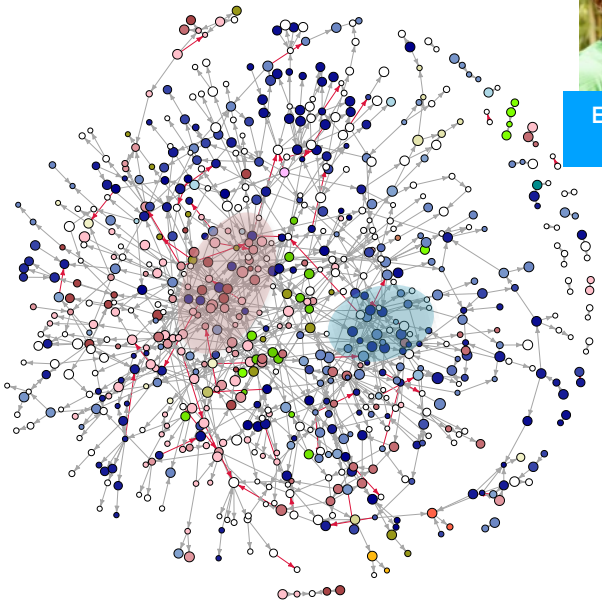


GPU/rust acceleration
Model robustness
Industrial applications

Louison Lesage

Raynald de Lahondès

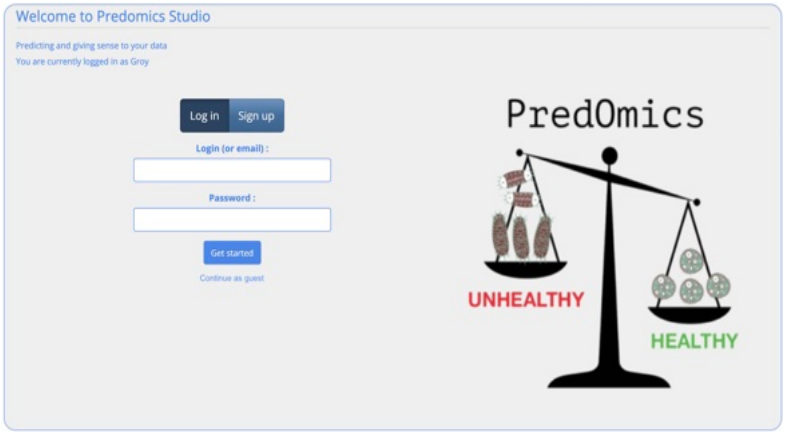
Ecosystem awareness of predictive signatures

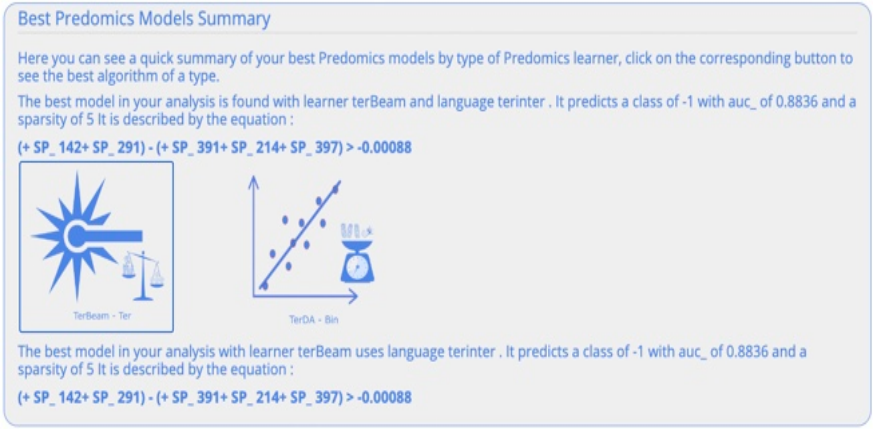


Estephe Kana

- F. Kambu …, Prifti E. In prep
- Prifti E et al In prep

A shiny web application





Gaspar Roy

Take-away messages …

- **See the Data, don't trust blindly the models and p-values**
- The gut microbiome's role in our health is still under-appreciated mostly because of challenges with standards, quantification, reference and catalogs…
- AI is powerful in numerous fields, but can be **challenging in biomedical data**, which are highly resolutive but **with a limited number of observations.**
- **Interpretability and clinical validation** are key for downstream applications.
- **Ad hoc specialised tools** and **methods** such as predomics… need to be developed and **adapted to the fields of study. Highly constrained models such as BTR @predomics** are both **extremely simple** and **accurate and Interpretable**
- **There is always room for improvement. Don't let a good idea die, follow up… it usually pays off**

Institut de Recherche pour le Développement FRANCE
French National Research Institute for Sustainable D

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Inserm
La science pour la santé
From science to health

We are hiring

3 postdocs +

1 PhD in AI

PARIS

Do you want to **put your skills, time and effort into exciting and useful projects that can save lives**? This is what we offer you by joining a **dynamic, agile and passionate team**.

PARIS

Position 1: (2 year postdoc; DL + ECG)

- **Deep learning engineer**: transformer models **embedding** research, **robustness** and **interpretability** for ECG (electrocardiograms) clustering in the context of sudden death risk.
- Apply **(CV + letter + 3 references)** to **talents@ummisco.fr**

Position 2: (3 year postdoc; DL + ECG)

- **Data scientist in AI and ECG.** Analysing one of the world's largest and diverse ECG datasets.
- Apply **(CV + letter + 3 references)** to **talents@ummisco.fr**

Position 3: (3 year postdoc; DL + ECG)

- **Data manager and biostatistics**. Analysing one of the world's largest and diverse ECG datasets.
- Apply **(CV + letter + 3 references)** to **talents@ummisco.fr**

Position 4: (3 year; PhD; DL + ECG)

- **PhD position**: multimodal transformer based models **for ECG** research.
- Apply **(CV + letter + 3 references)** to **talents@ummisco.fr**

+ **Multiple** internships


Thanks
Merci
Faleminderit