

> Integrating Human Gut Microbiome Data at Scale: From Individuals to Multi-Layered Features in Clinical and Nutritional Studies.

Julien Tap
INRAE MICALIS
AI SCHOOL 2025



INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr



Wolfgang Huber @wolfganghuber · 8 h

I think we should stop using the term 'biologists' for people who cannot use computer code.



8



14



39



TECH

Accenture plans on ‘exiting’ staff who can’t be reskilled on AI amid restructuring strategy

PUBLISHED FRI, SEP 26 2025•7:18 AM EDT | UPDATED SAT, SEP 27 2025•8:33 AM EDT

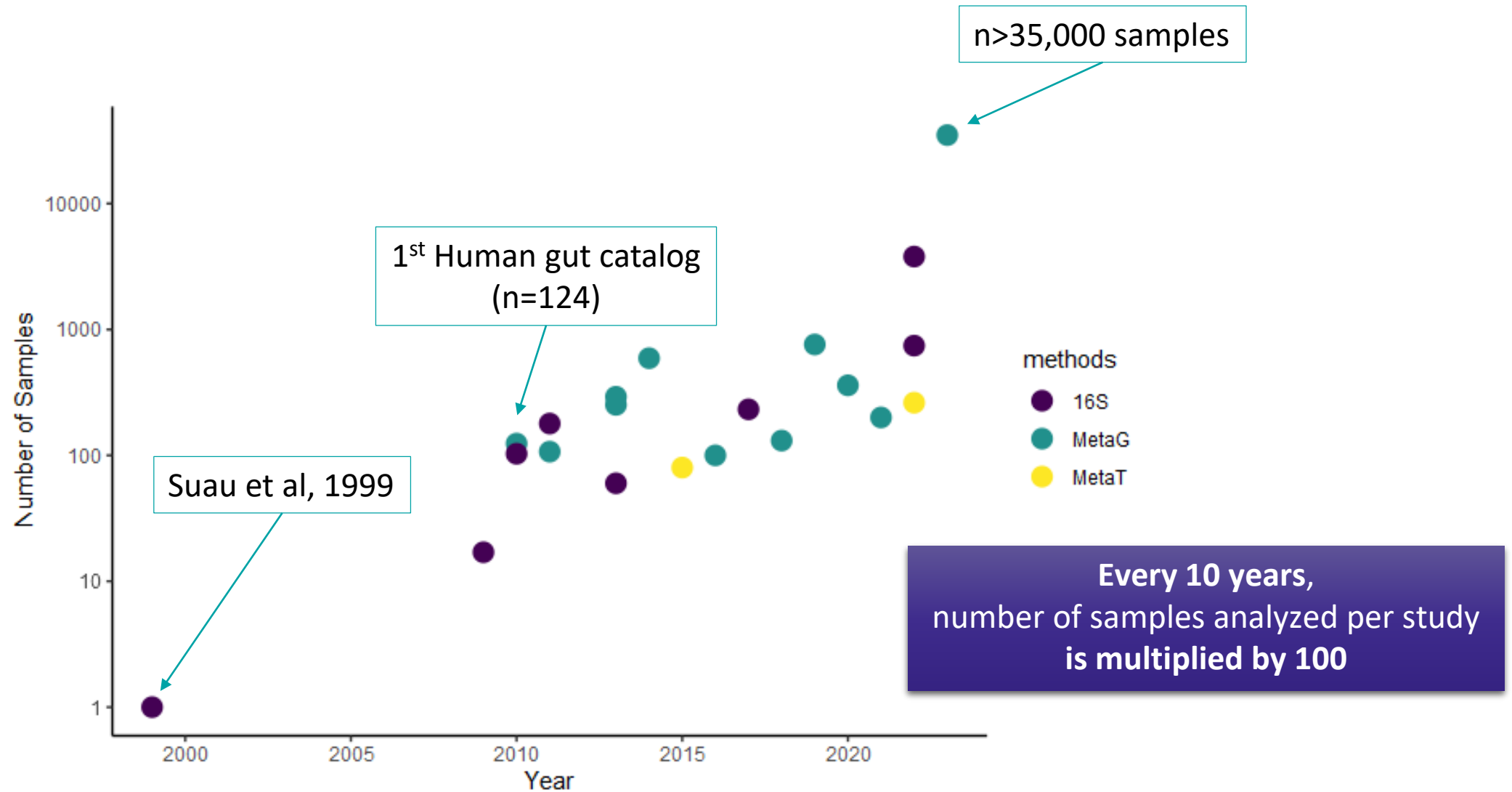


Sawdah Bhaimiya

SHARE



> Gut microbiome studies: exponential growth of sample numbers



> Human gut microbiota analysis in a nutshell

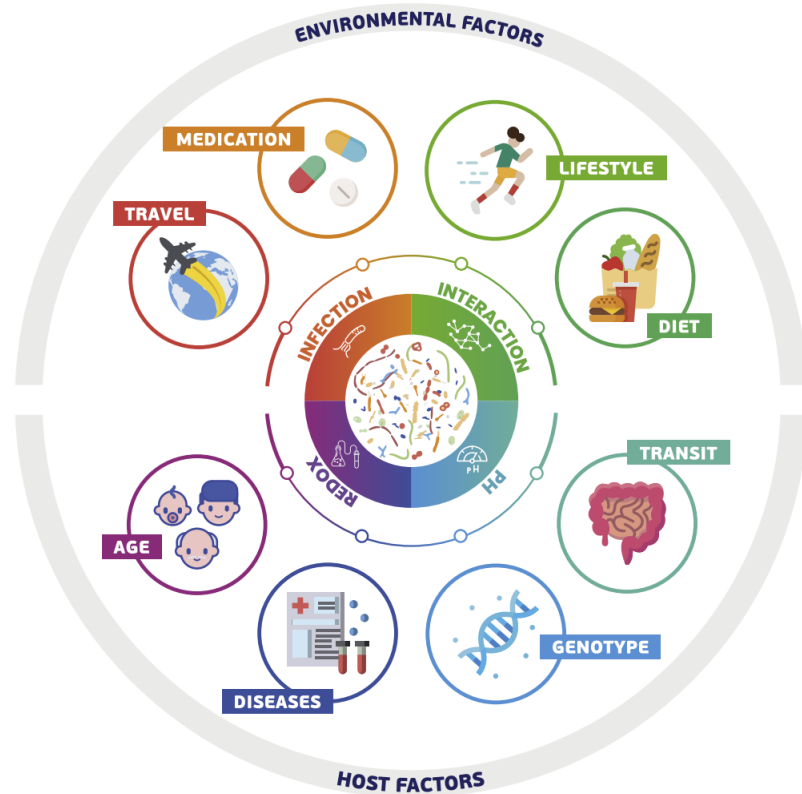


Fig. 1 Overview of host and environmental factors that are associated with gut microbiota variation, in addition to intrinsic factors.

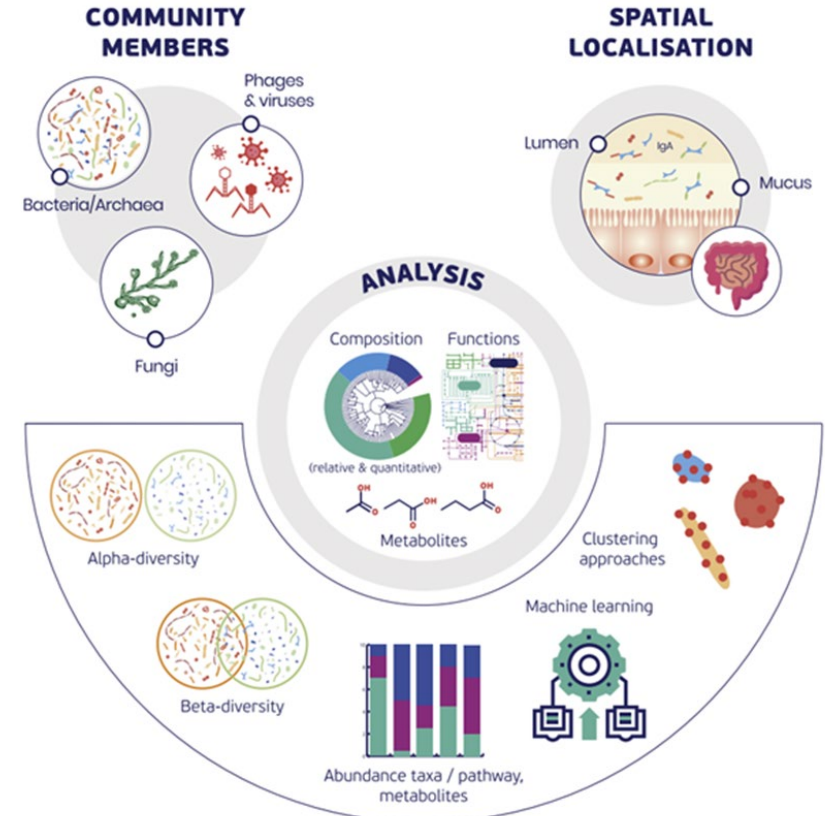


Fig. 2 Overview of the study of gut microbiota, including community members, spatial localization, and analysis.



INRAE

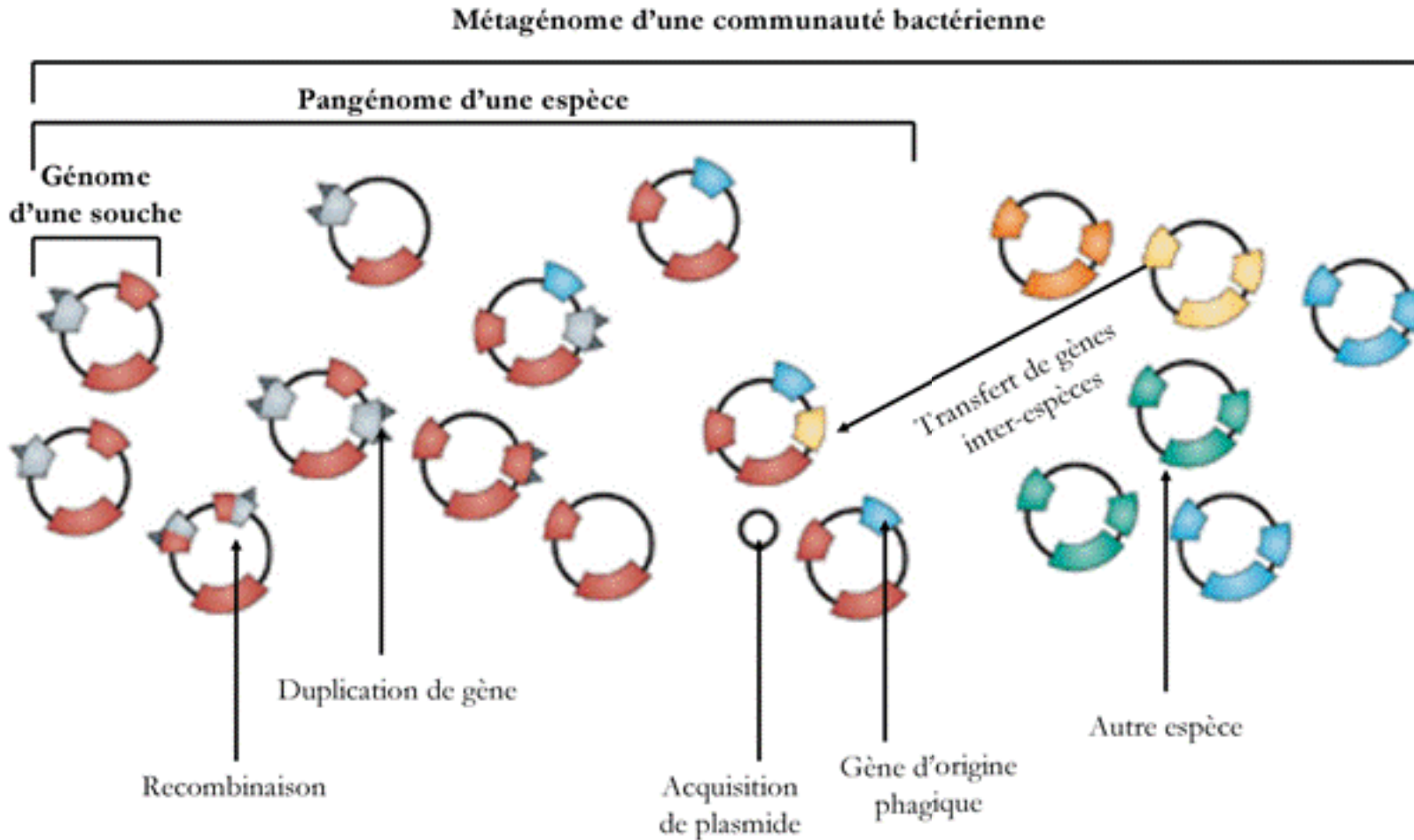
Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

Tap et al., 2022

p. 5

> Microbiome omics data challenges



Number of features to analyze exponentially increase



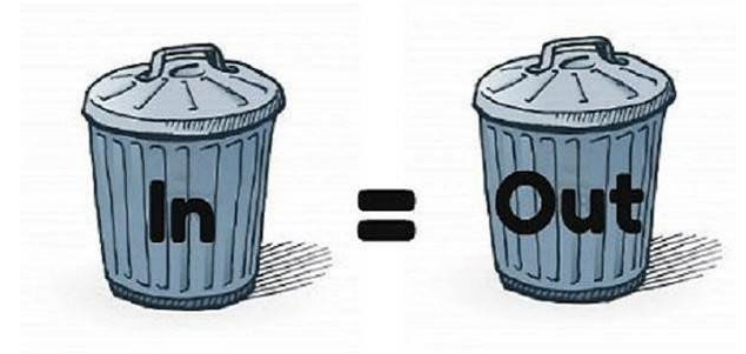
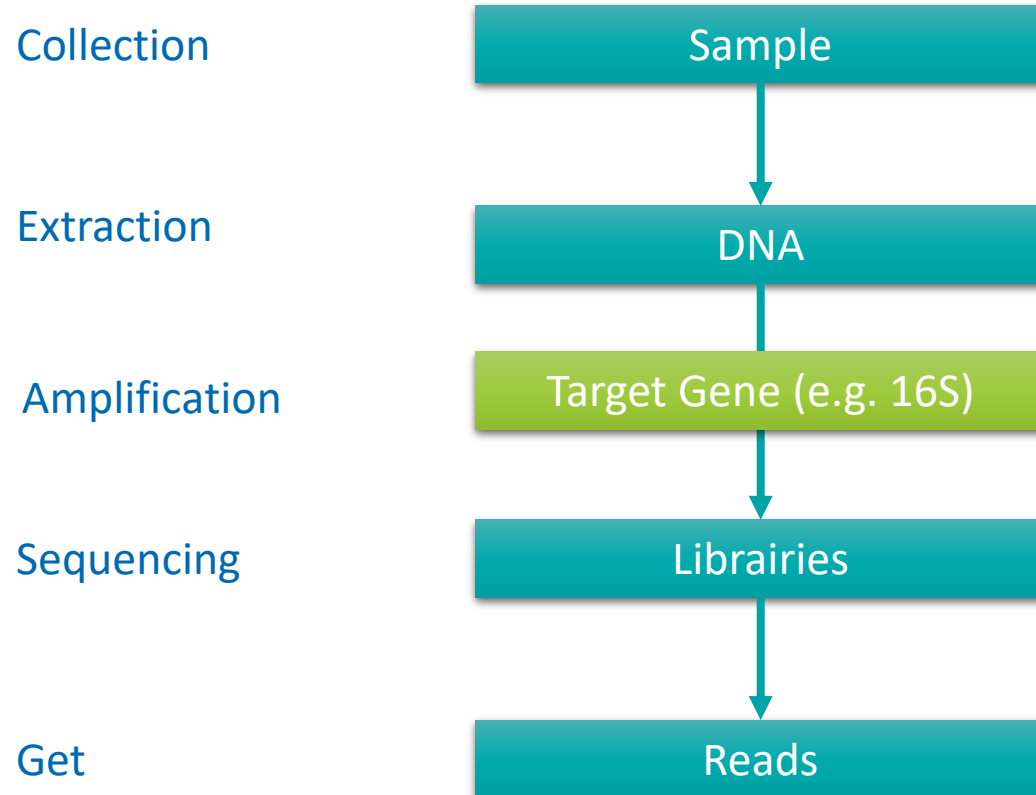
Organizing and structuring these metagenomic objects



Combine those features with internal and external metadata

> The metagenomic approach

IT'S EASY!



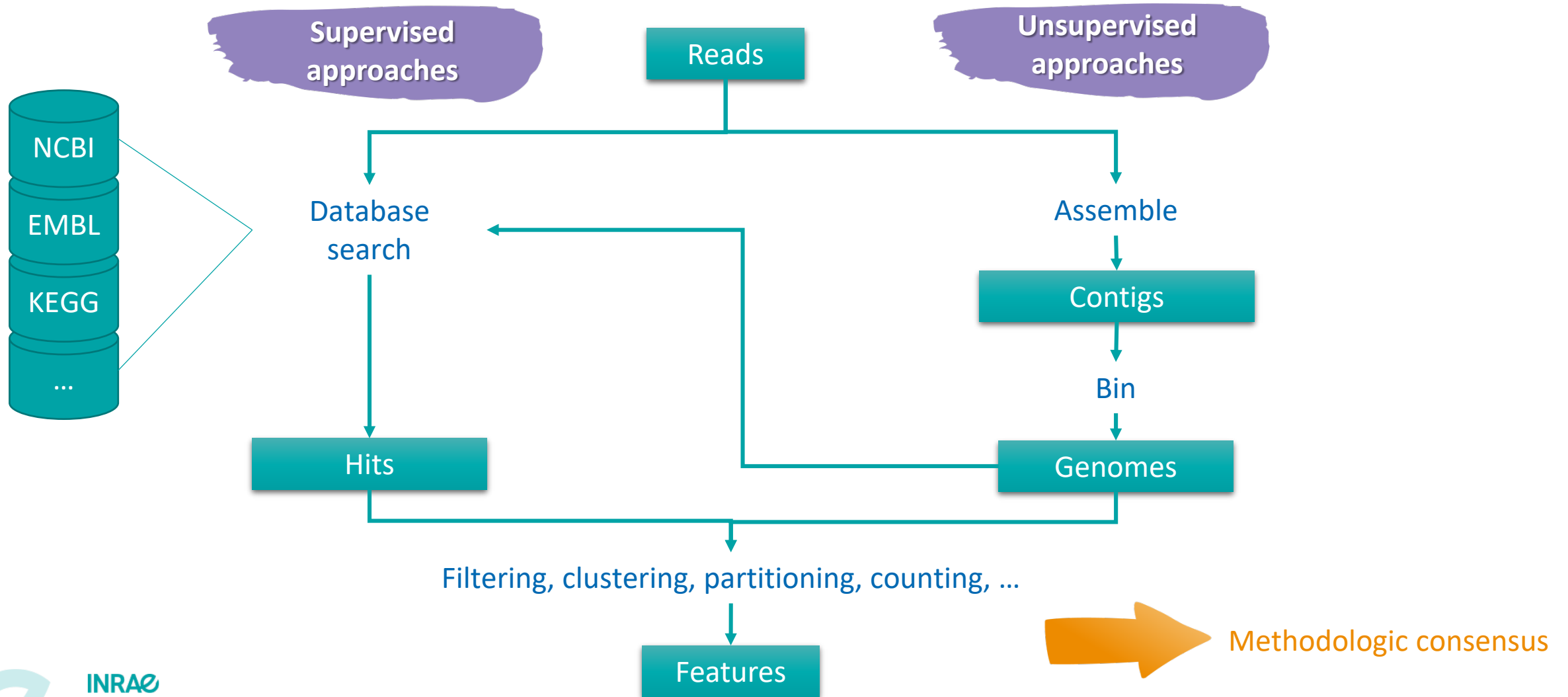
Garbage in = Garbage out

Each stage of the process
can be affected by the
protocol used

 Close to a standard

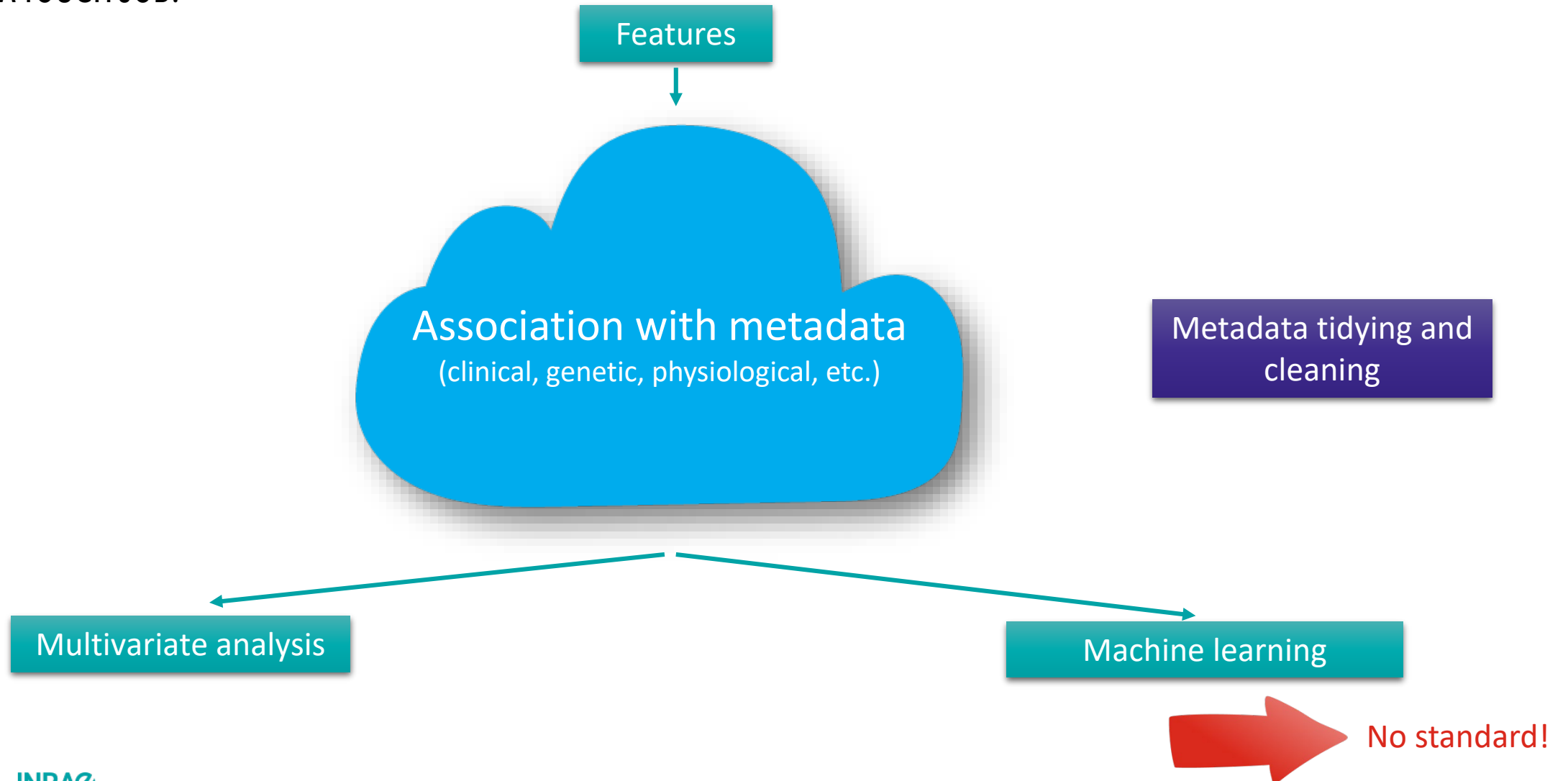
> Metagenomic analysis

IT'S COMPLEX!

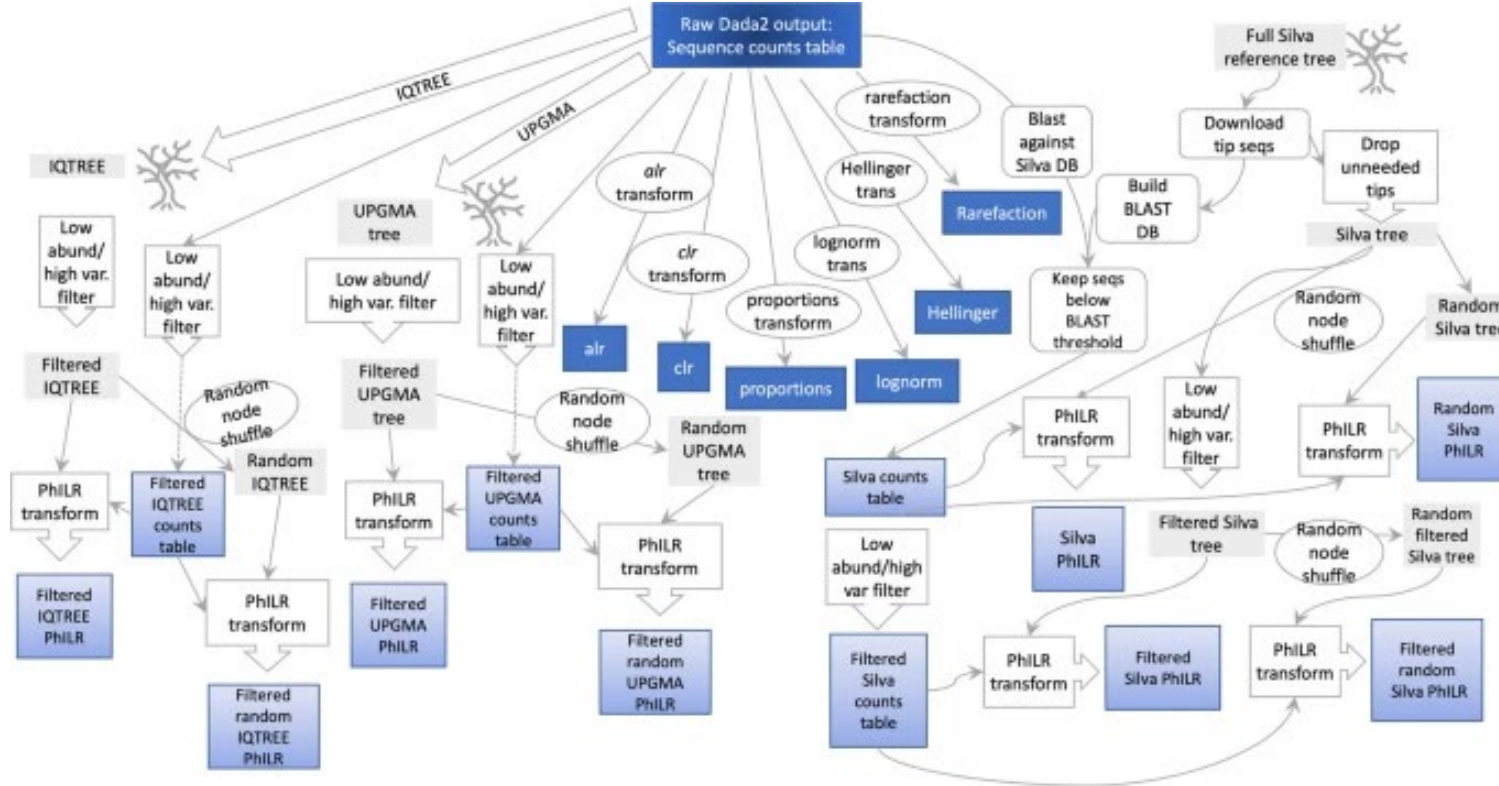


> Statistical analysis in metagenomics

IT'S A TOUGH JOB!



➤ Tools Benchmarks



“Overall, we found that the compositionally aware data transformations such as alr, clr, and ilr (PhILR) performed generally slightly worse or only as well as compositionally naïve transformations.”

> Tools Benchmarks

Summary of the compared DAA methods.

Method	Normalization strategy	Performs log (or CLR) transformation	Model	Can incorporate covariates	Provides CIs	Typical run-time
ALDEx2	CLR	x	Linear	x	x	30 s
ANCOM-BC2	Bias correction	x	Linear	x	x	20 s
corncob	TSS		Beta-binomial	x	x	10 s
DESeq2	RLE		Negative binomial	x	x	7 s
edgeR	TMM		Negative binomial	x		0.5 s
fastANCOM	Reference taxa	x	Linear	x	x	0.01 s
LDM	TSS		Linear	x		30 s
limma voom	TMM	x	Linear	x	x	0.07 s
LinDA	Bias correction	x	Linear	x	x	0.2 s
LogR	–		Binary logistic	x	x	0.6 s
MaAsLin2/t-test	TSS	x	Linear	x	x	2 s
metagenomeSeq	CSS	x	Linear		x	0.8 s
ORM/Wilcoxon	TSS		Ordinal (proportional odds)	x	x	3 s
ZicoSeq	Reference taxa		Linear	x		20 s

Overall, the best performance was obtained by analyzing TSS normalized counts, i.e. relative abundances, with a nonparametric method (ORM/Wilcoxon), log-transformed TSS normalized counts with the t-test/linear regression (MaAsLin2/t-test), or the presence/absence of taxa with logistic regression (LogR).



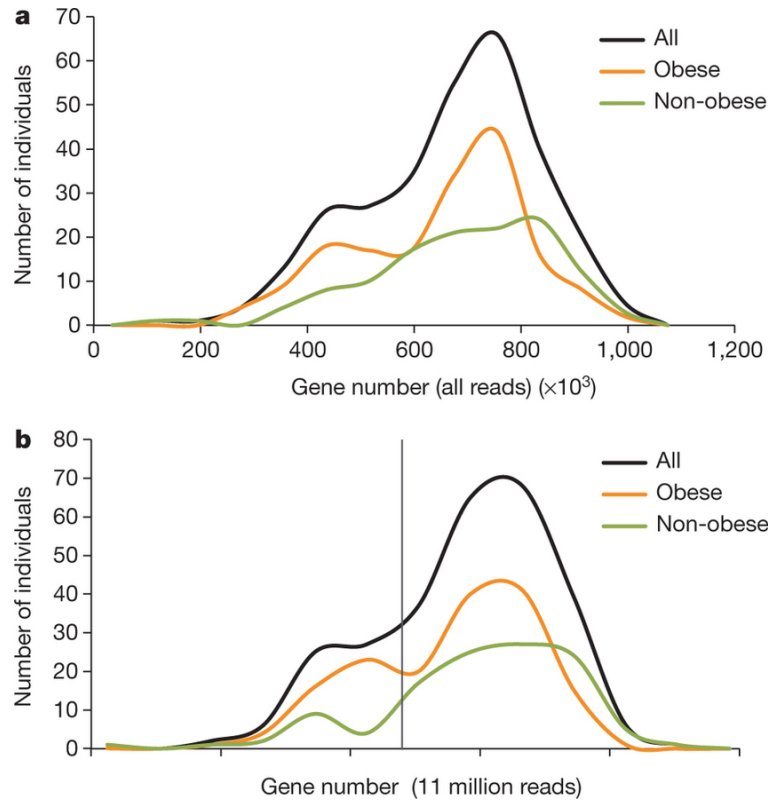
INRAE

Human Gut Microbiome Data at Scale – IA School

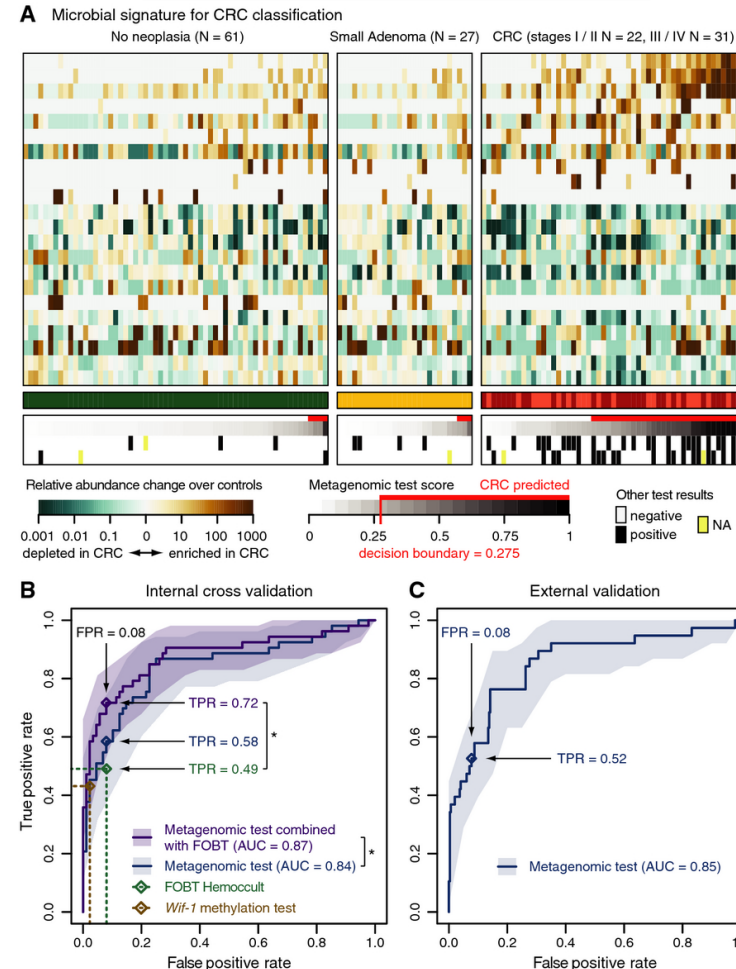
2025.10.01 / Julien Tap / fme.micalis.fr

➤ The association of metagenomic data with clinical studies

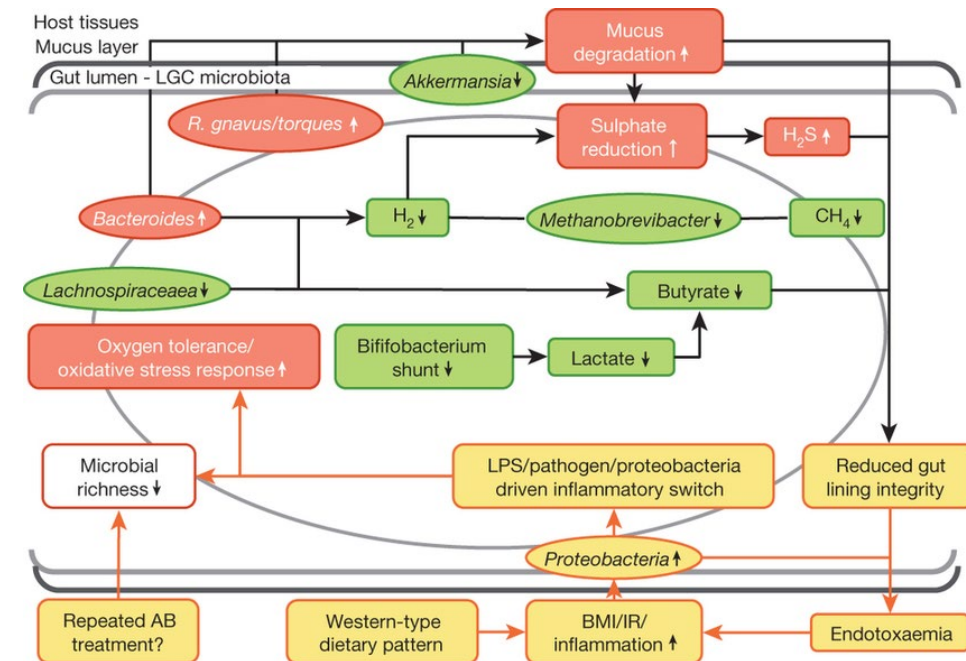
Gene richness



Biomarkers from machine learning

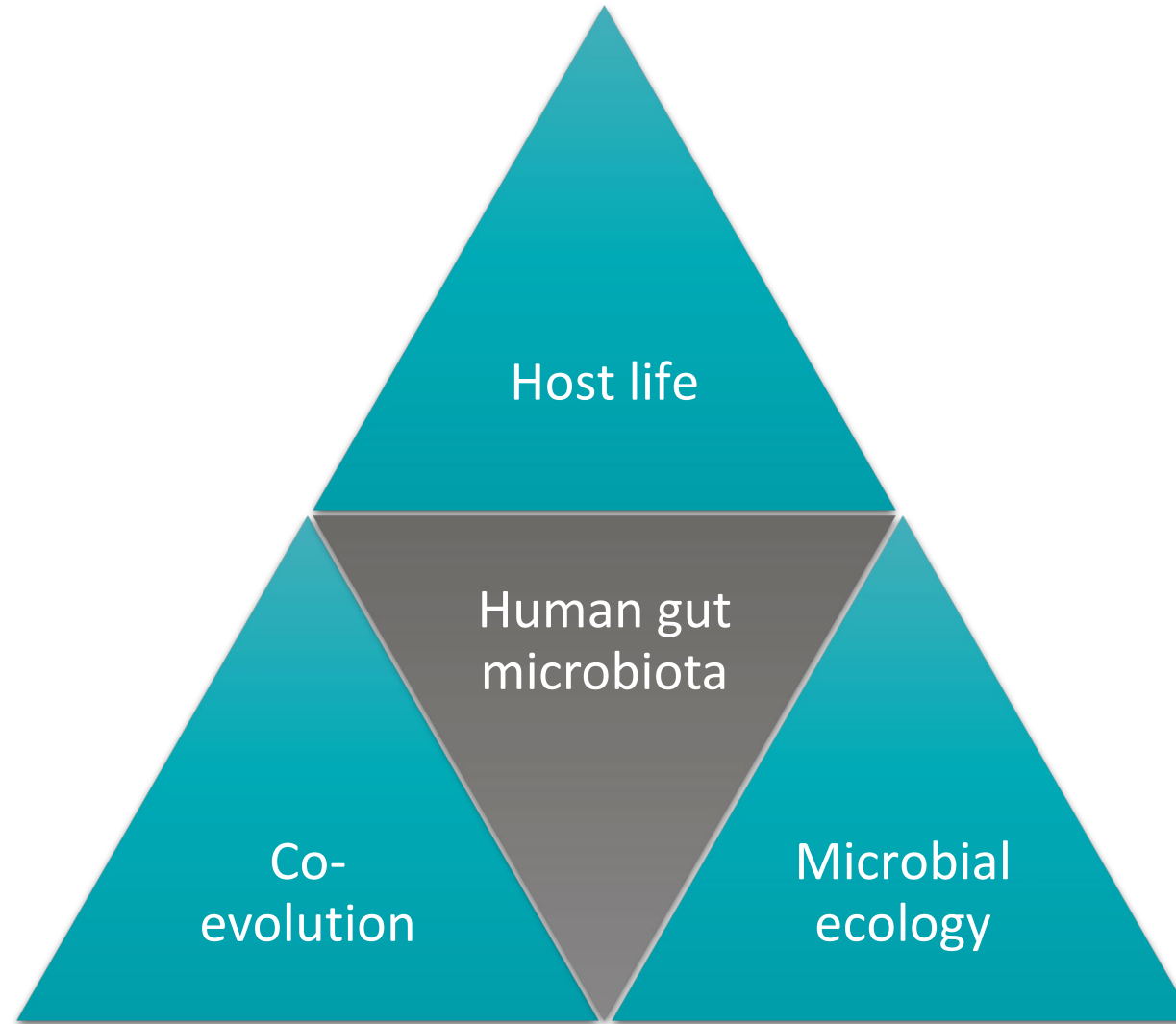


Descriptive analysis of potential functions



<https://github.com/raeslab/GMMs>
<https://siamcat.embl.de/>

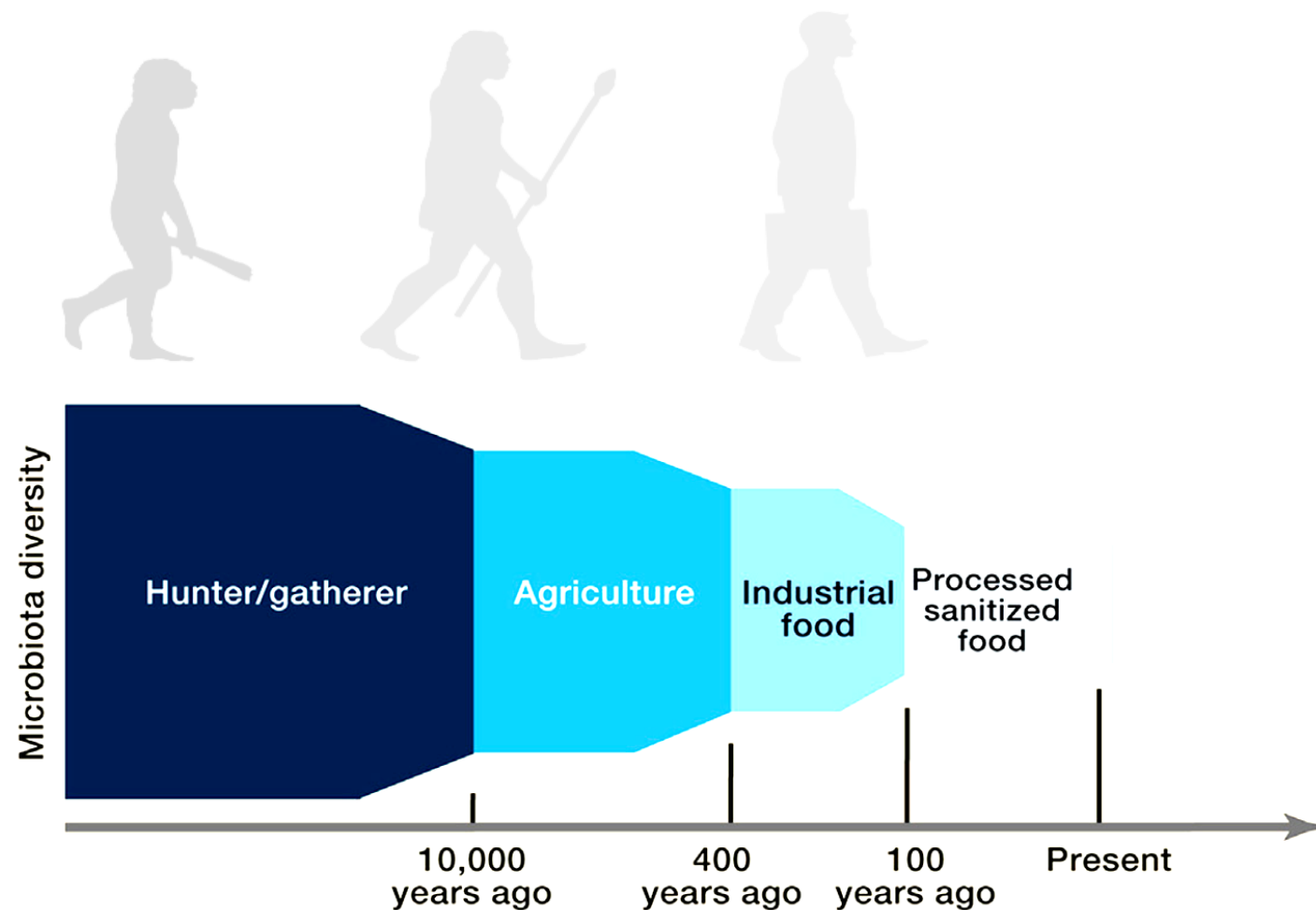
> Different factors shapes the human gut microbiome



> Diet and lifestyle effect host microbiome symbiosis

14

Are we uncoupling our metagenome and our genome?



Hunters - gatherers

Sedentarization

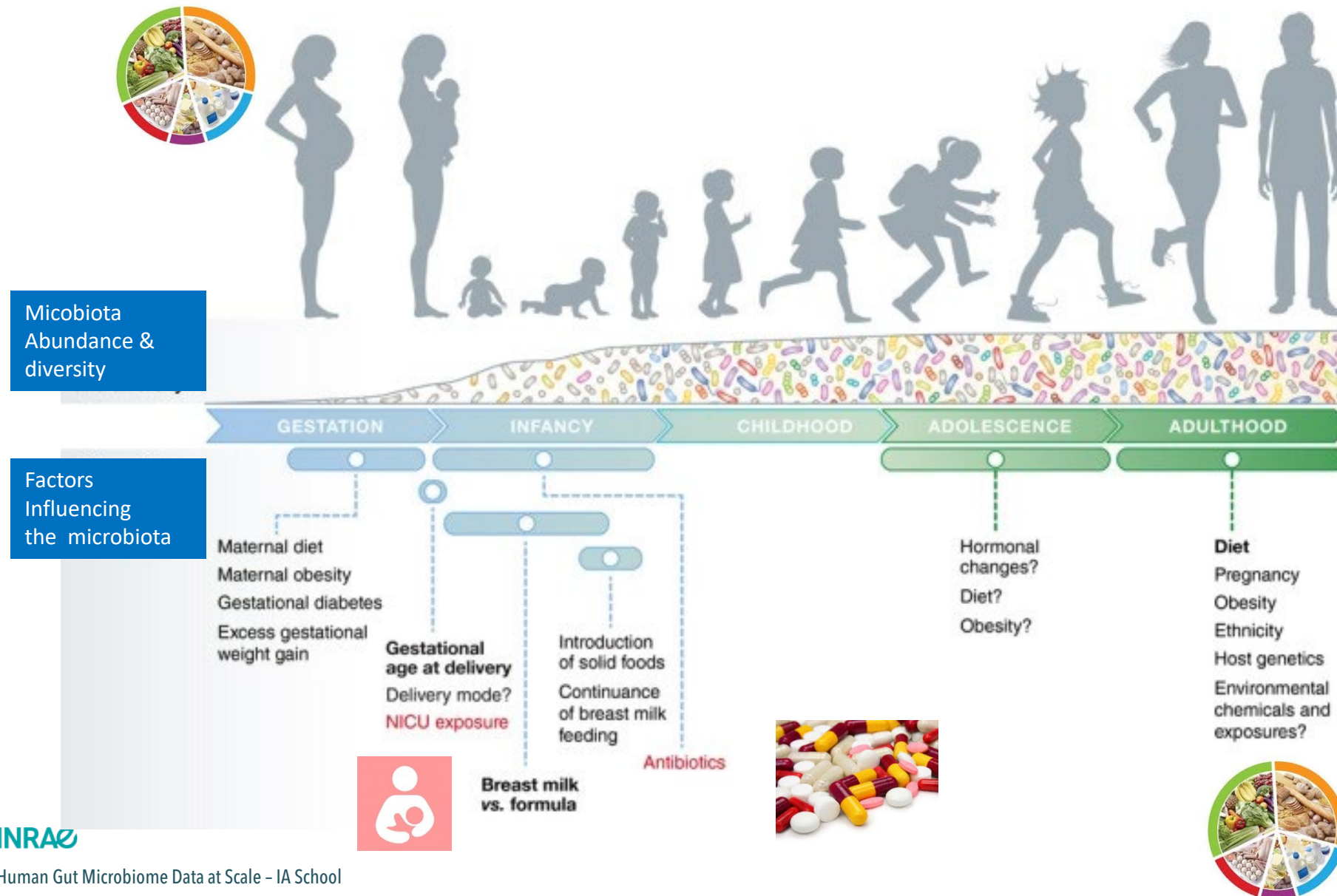
Livestock and crops

Hygiene/medication

Refined products

Gut microbiome throughout life span

15



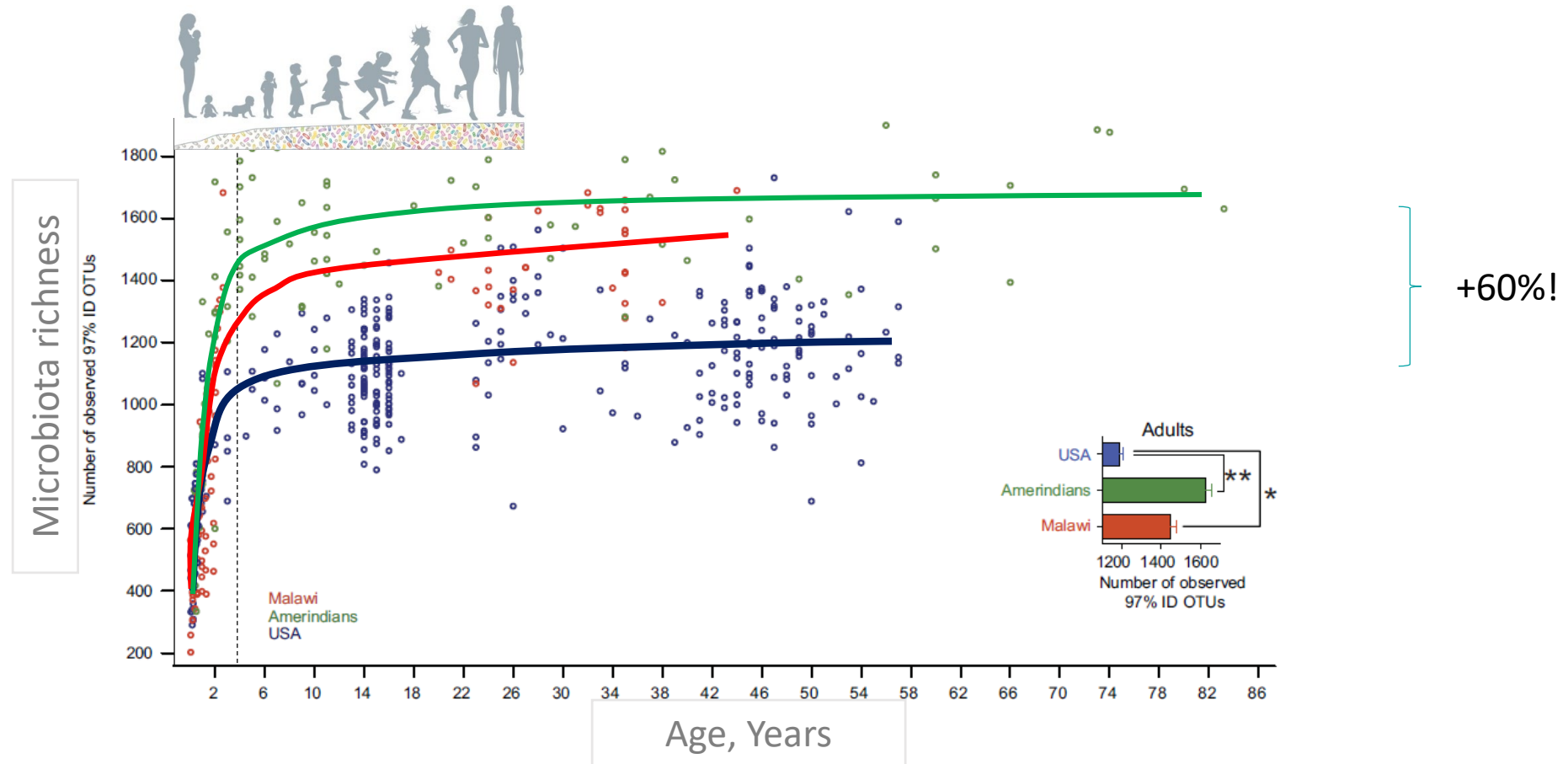
INRAE

Human Gut Microbiome Data at Scale - IA School

2025.10.01 / Julien Tap / fme.micalis.fr

Gut Microbiome diversity associated with the region of birth during the lifespan

16



INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

Yatsuneneko et al, 2012

p. 16



Microbiome states resilience as key factor



Healthy state



Resistance



Recovery



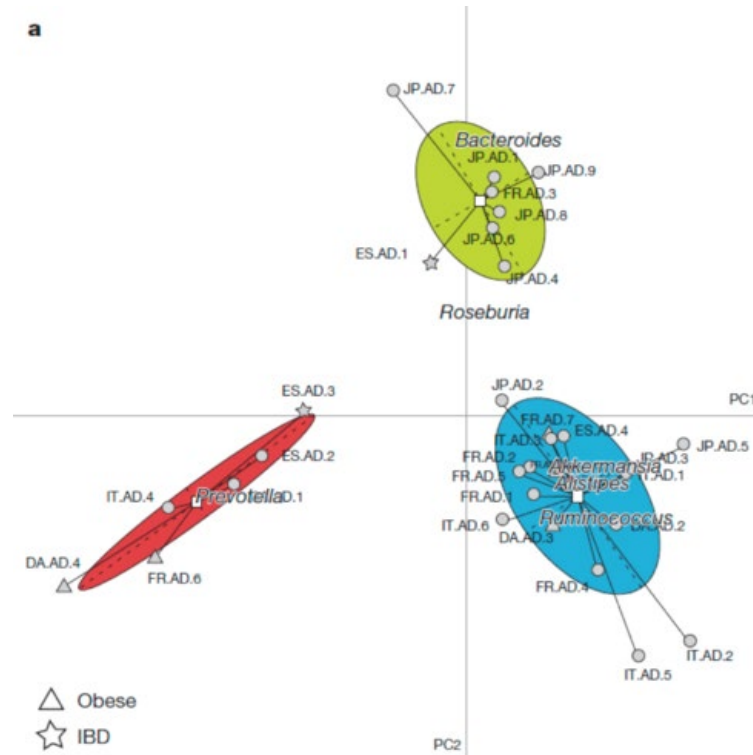
Healthy state



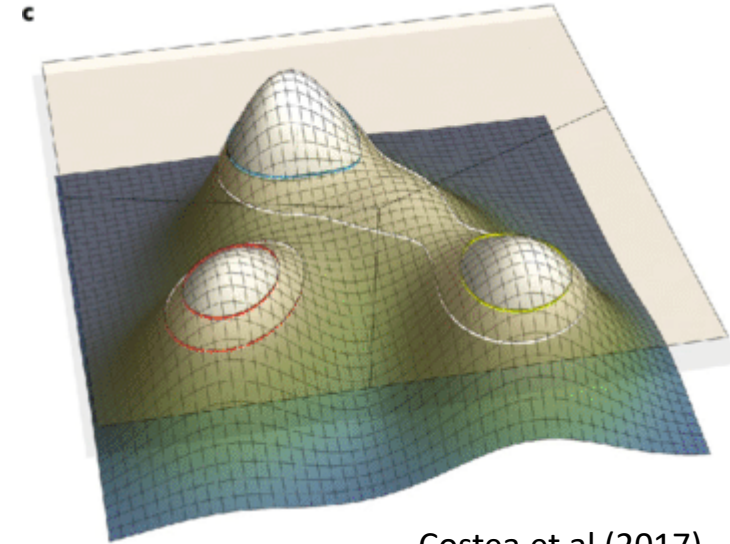
Degraded state



➤ Several attempts to decipher the microbiome structure



Arumugam, Raes et al (2010)



Costea et al (2017)

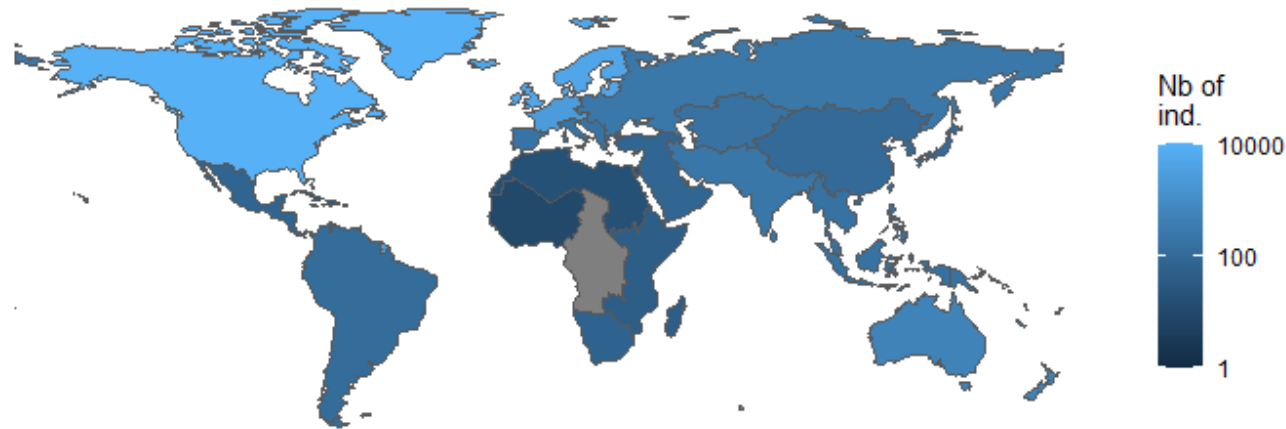
Dirichlet Multinomial Mixture (DMM) modeling

Limits:

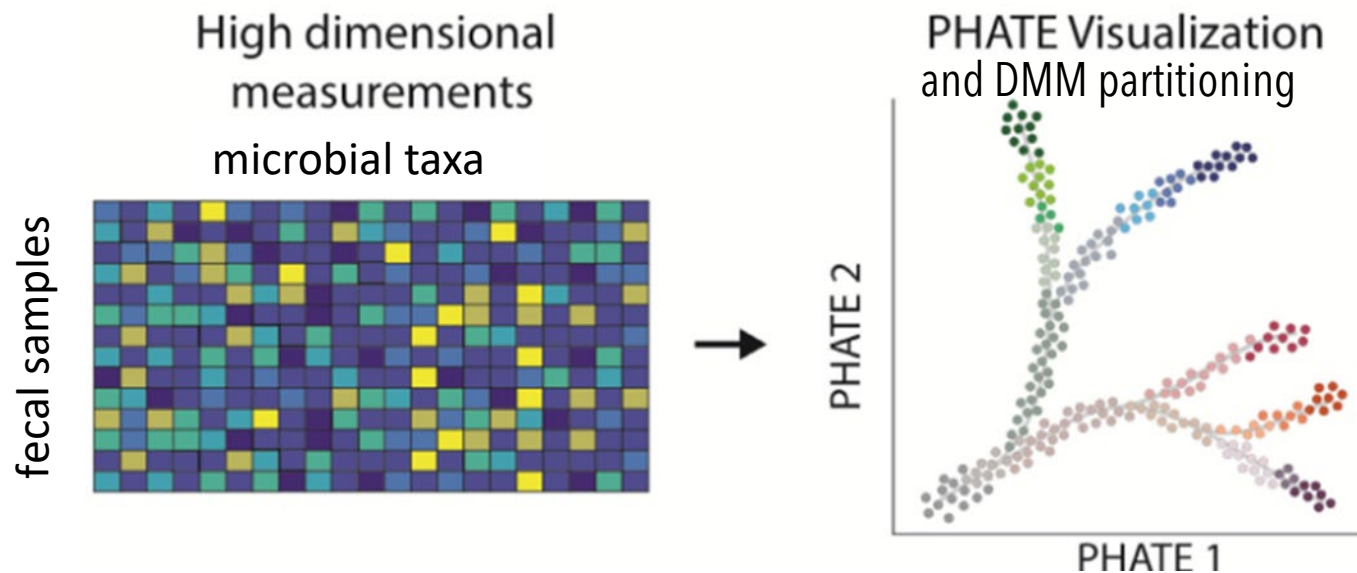
Few association with diet

Few stability assessment (e.g. with time-series)

> 35,000 human gut microbiome samples analyzed



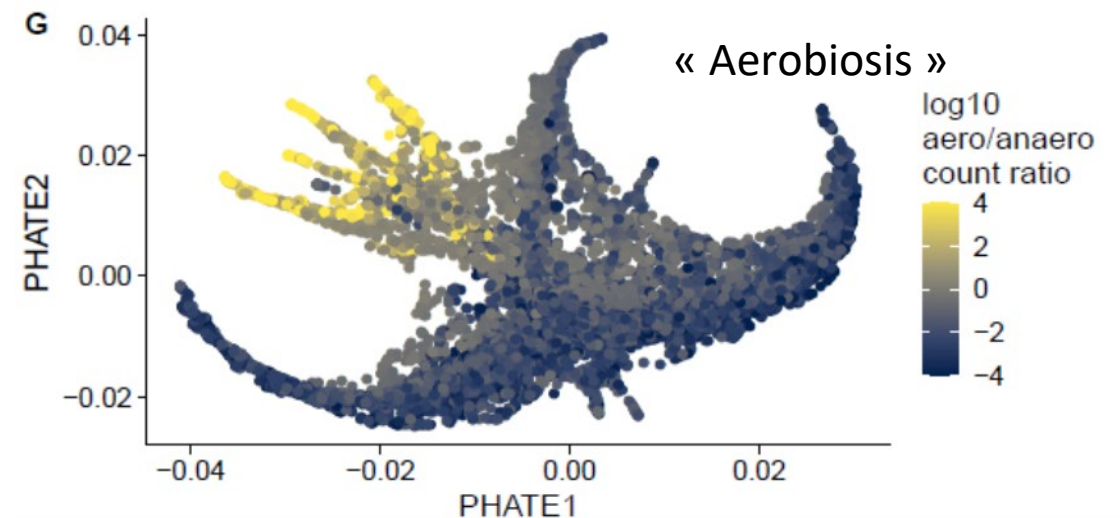
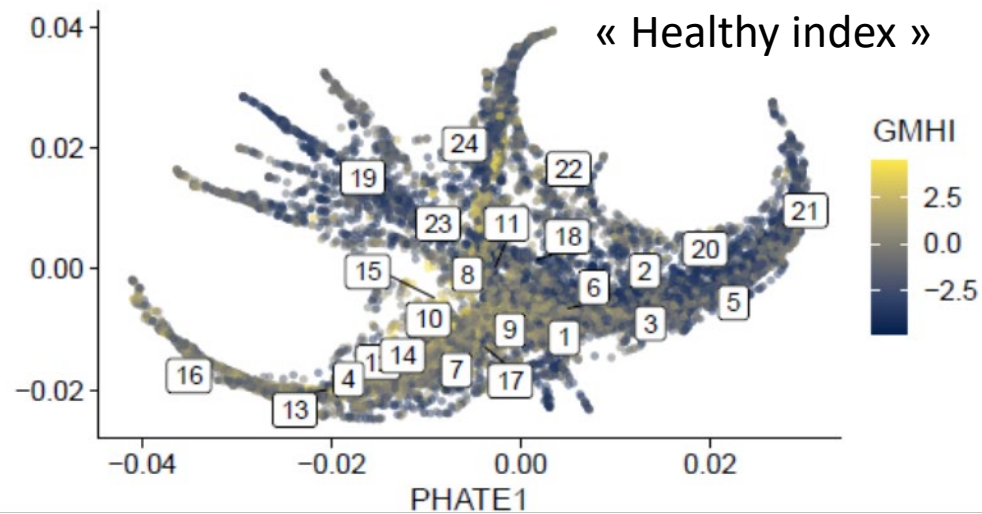
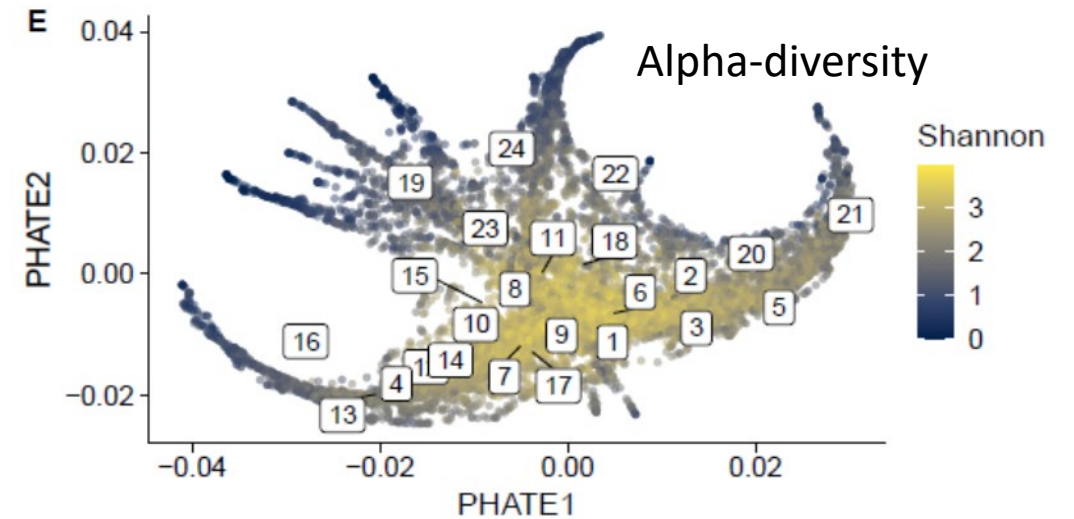
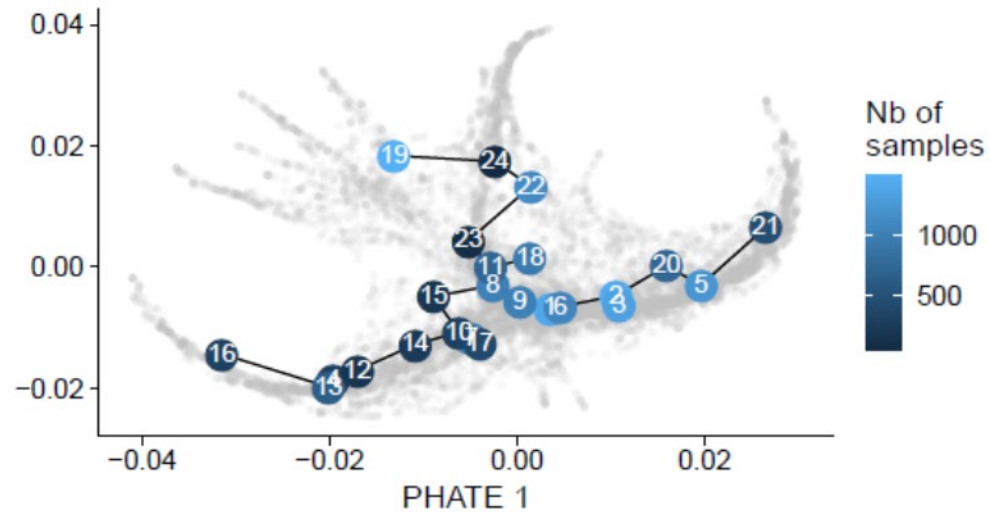
New computational method that help observe **global and local structure** at the same time



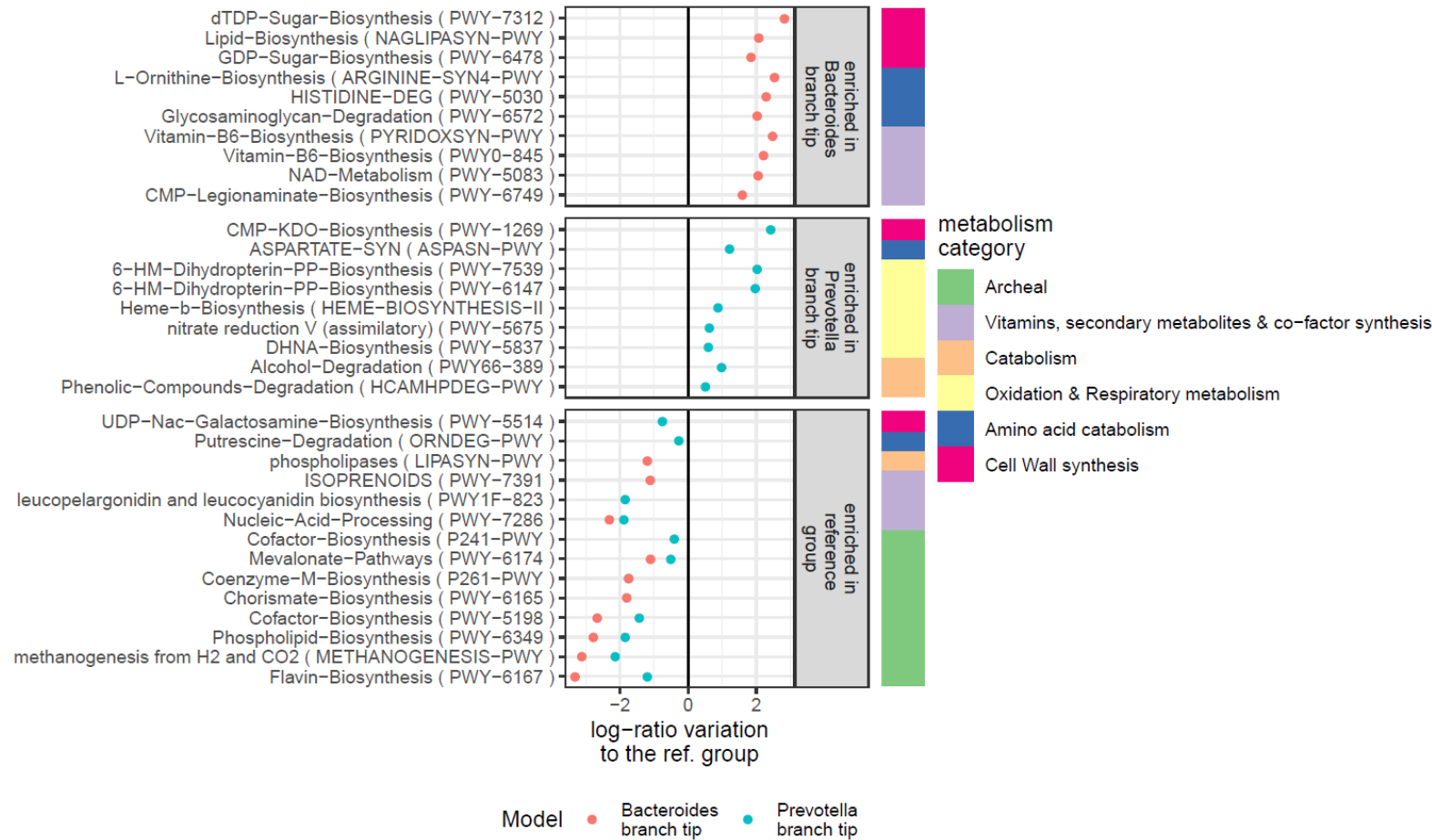
Time-series analysis to detect microbiome states

Region of birth, Lifestyle, diet, and health modelization

➤ Human Gut microbiome structured as branches



> Low-diversity tips of branches display functional shifts



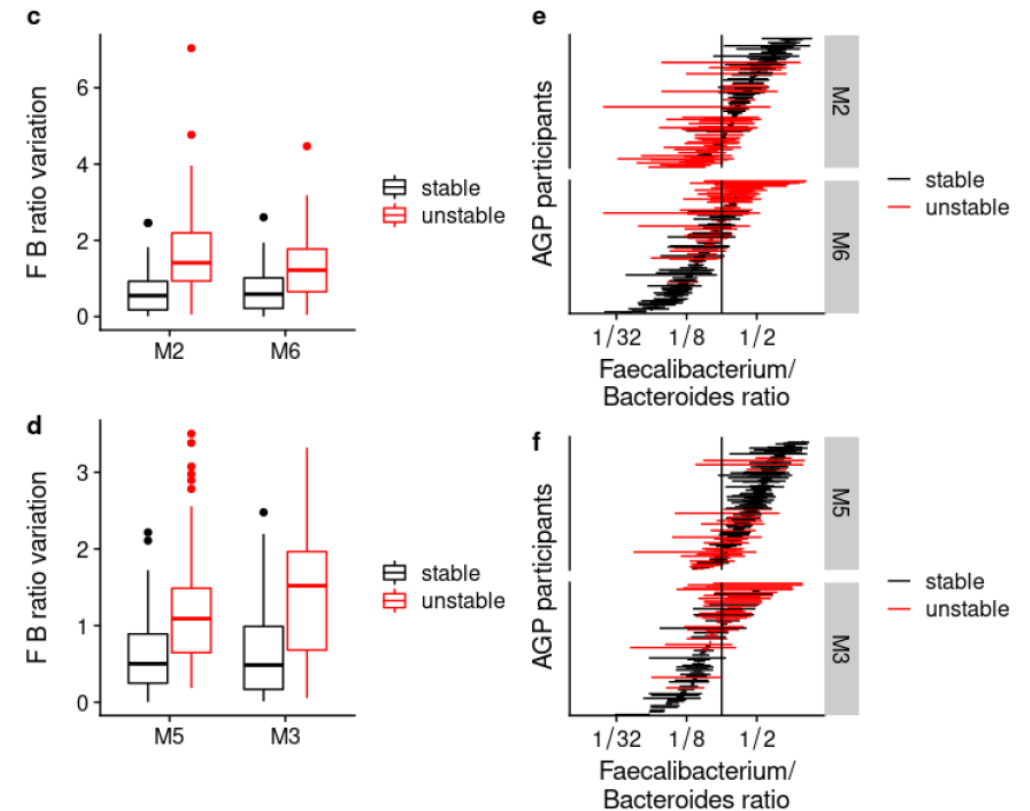
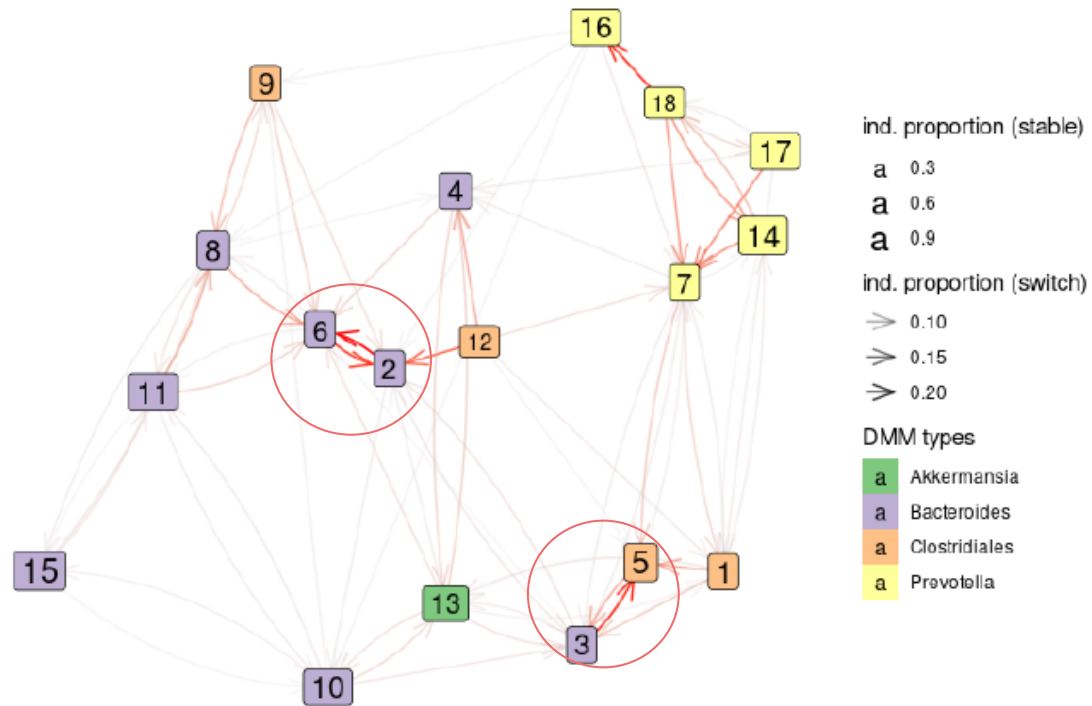
amino acid catabolism increase
in *Bacteroides* branch tip

oxidative stress response increase
in *Prevotella* branch tip

multinomial regressions models (songbird)
<https://doi.org/10.1038/s41467-019-10656-5>

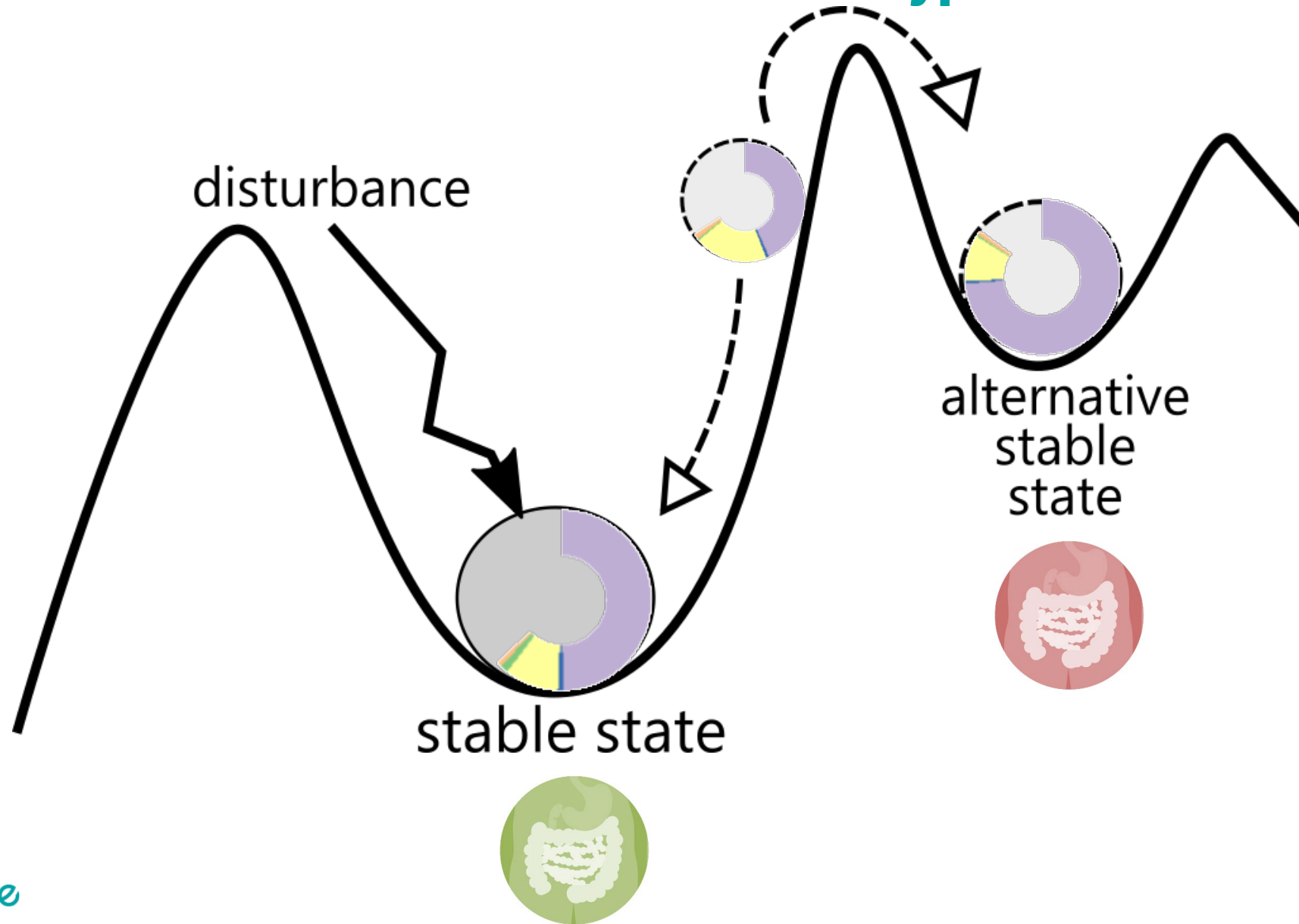
Dynamics between partitions are associated to branches

827 participants 2,998 samples 2,171 time-pairs

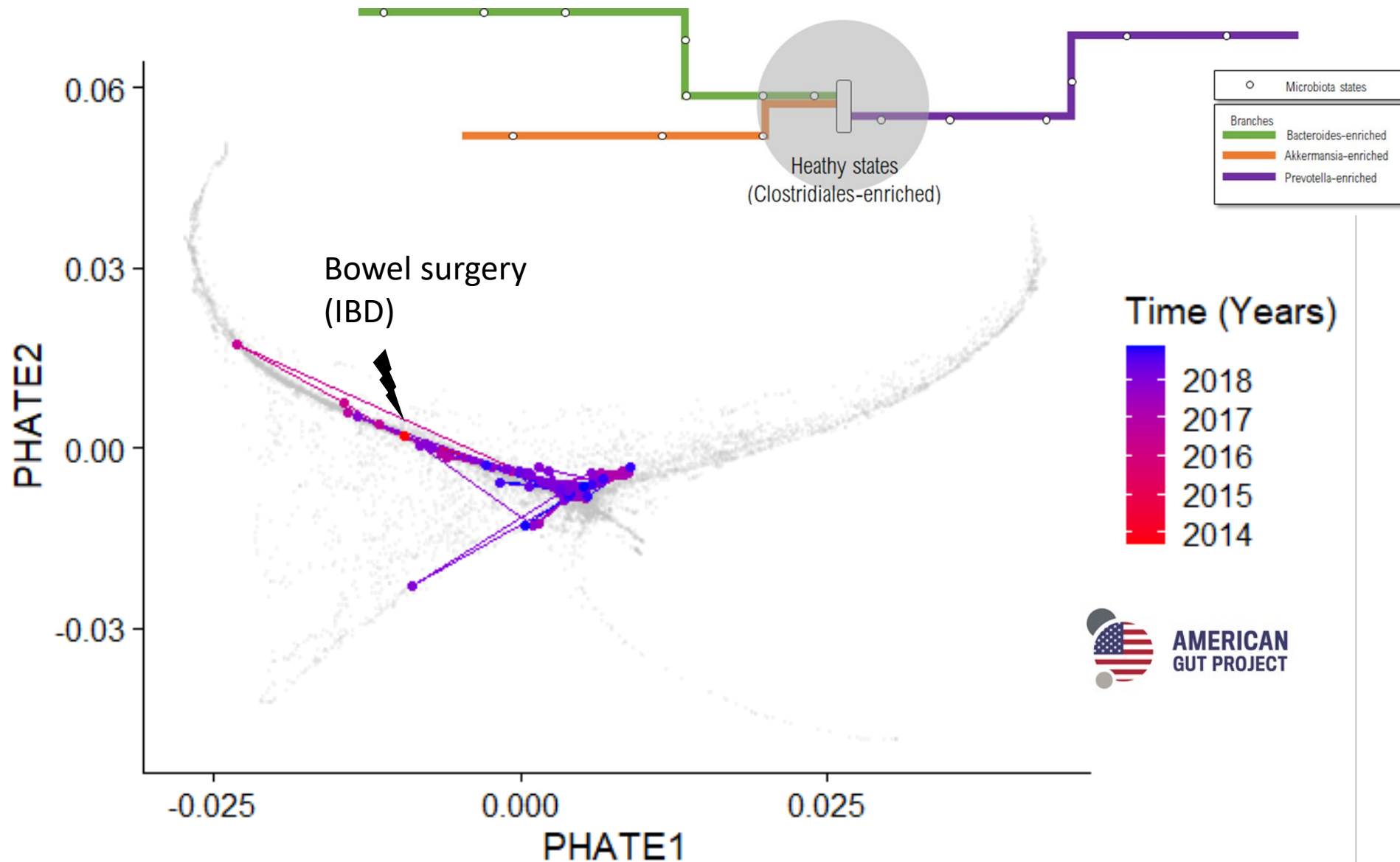


Faecalibacterium/Bacteroides as a tipping element explaining the switch between some states

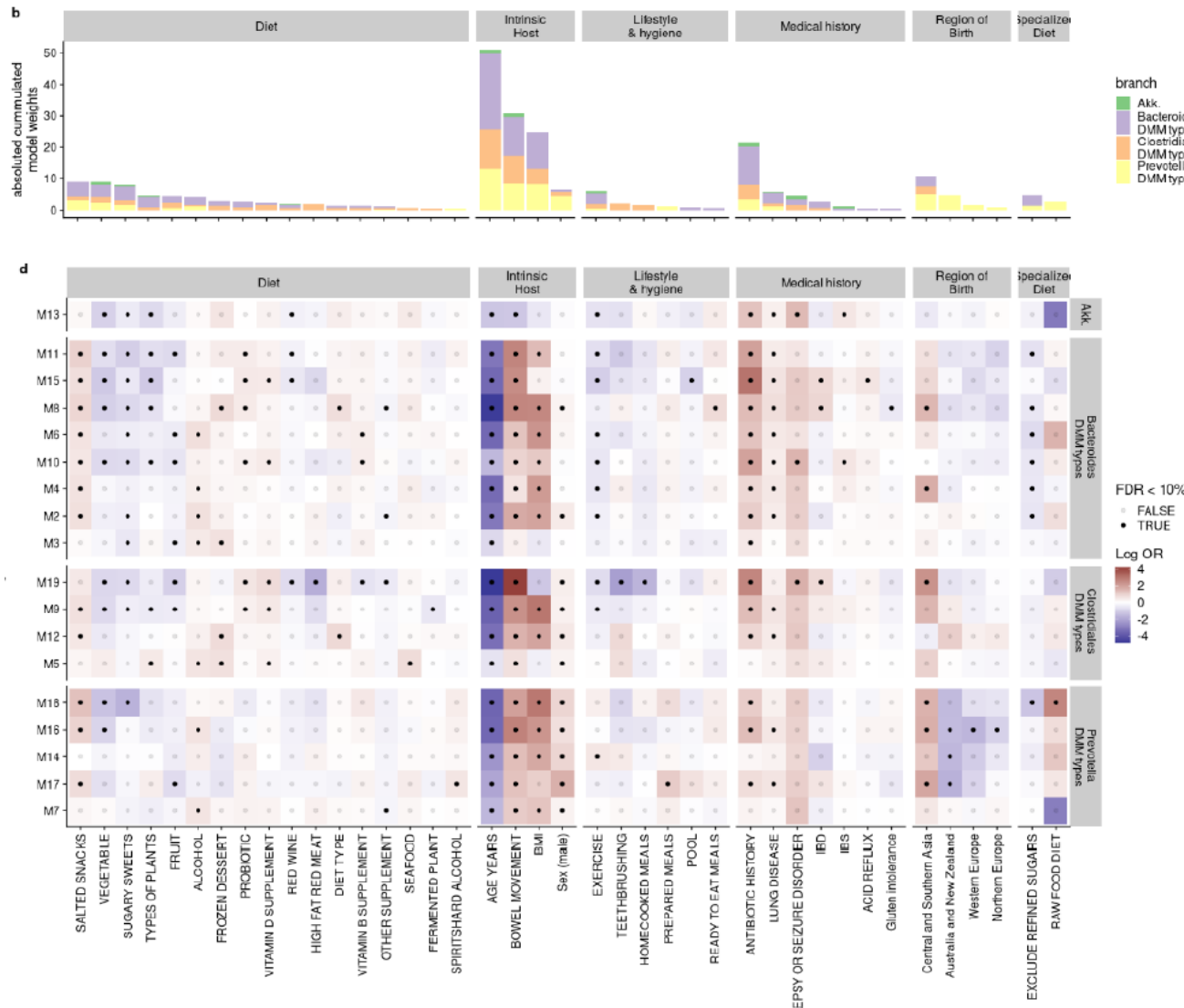
Gut microbiome alternative stable states hypothesis



➤ Branches help to monitor gut microbiome recovery



> Gut microbiome states are differentially associated with host and env factors



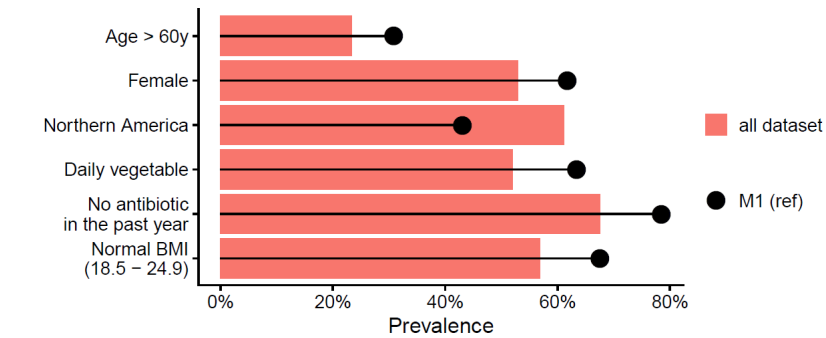
Sex, exercises frequency and **region of birth** associated differentially between branches

Lifestyle and hygiene associated with altered state within *Bacteroides* branch

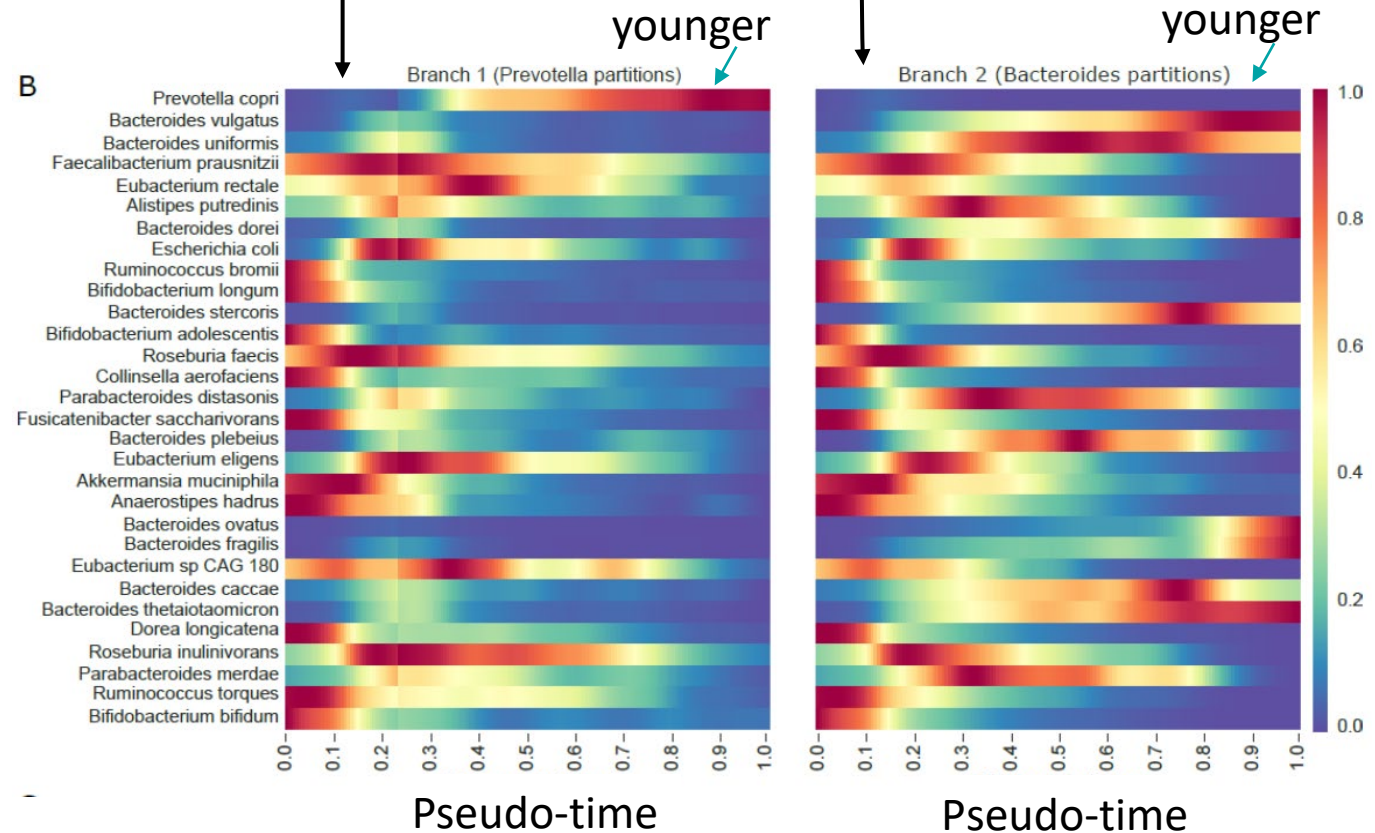
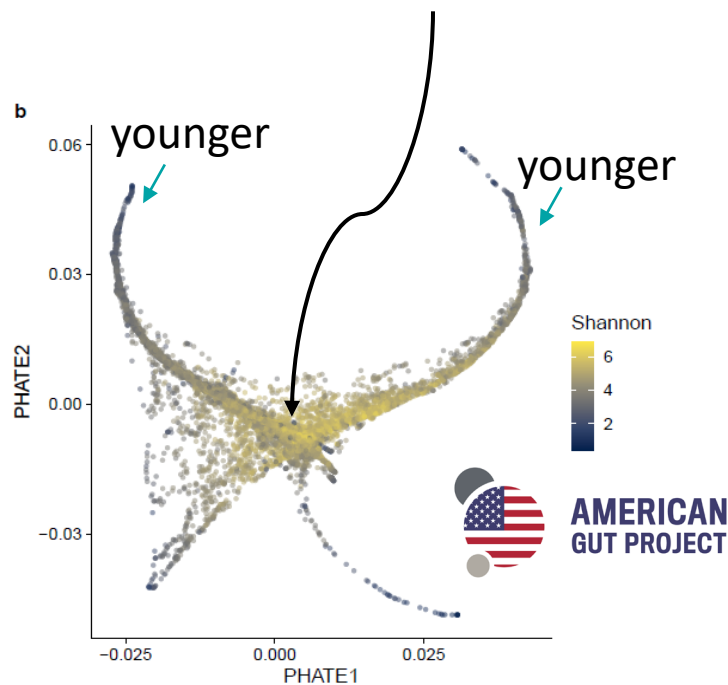
Specialized diet associated with altered state within *Prevotella* branch

multinomial logistic regression (nnet R packages)

> Do we already observe a intergenerationally mass extinction ?



Older and healthy diet and lifestyle

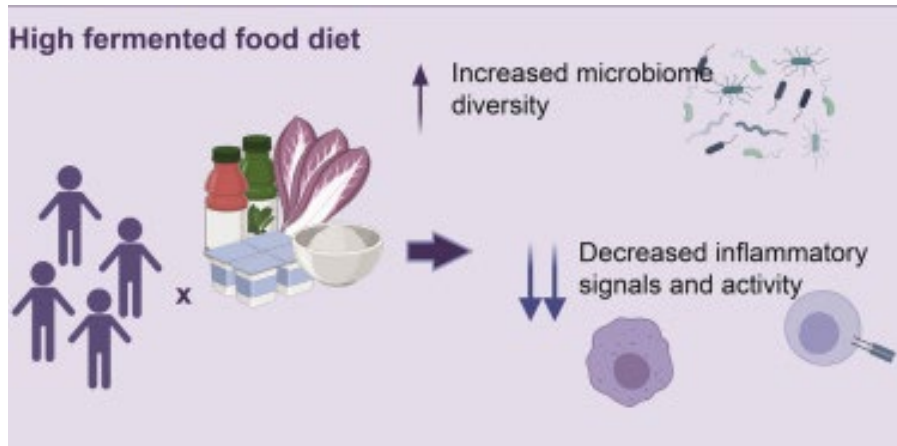


Wishbone

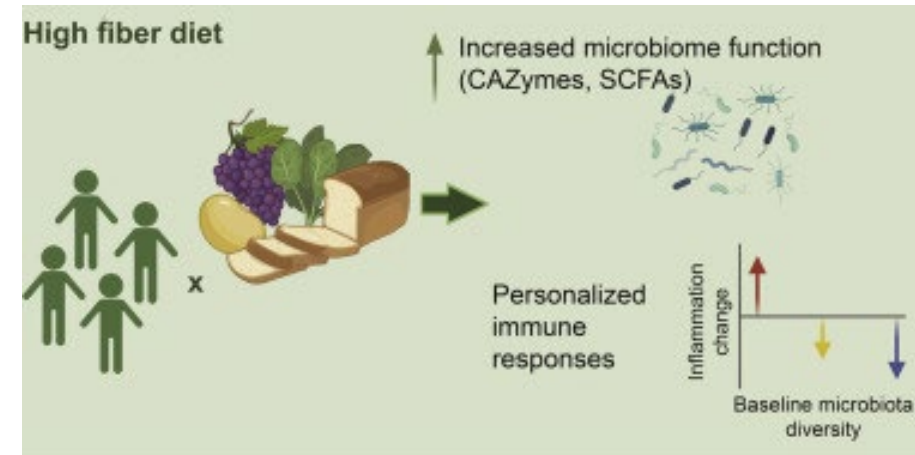
➤ Harnessing the microbial potential of fermented foods to avoid the point of no return



> How fermented foods and dietary fiber can fit precision nutrition for health?



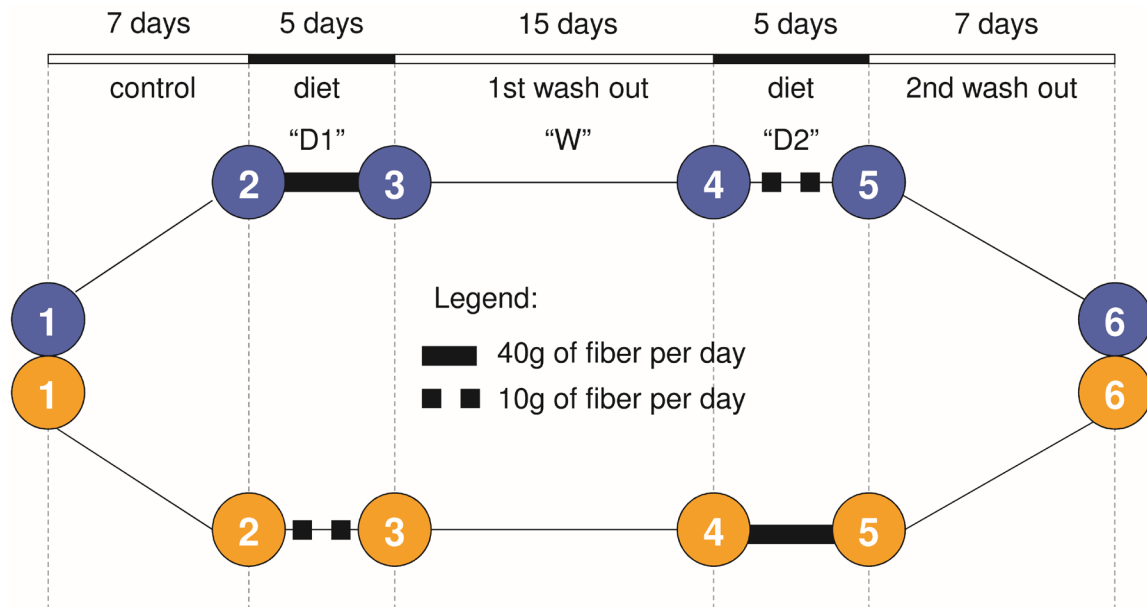
Fermented foods valuable solution in countering the decreased microbiome diversity



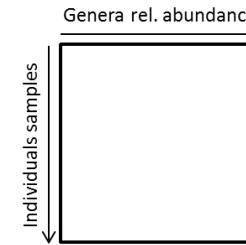
Personalized immune as function of gut microbiota baseline

Wastyk et al., 2021

> Combining time-series and meta-omics from dietary fibers challenge study



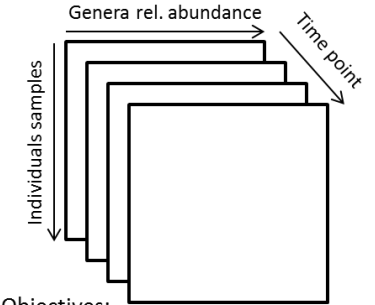
1-Table approach (e.g. PCA)



Objectives:

- data structure description
- Samples ordination
- Correlation between genera

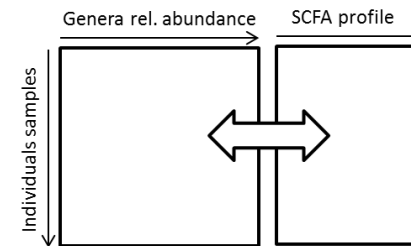
K-Tables approach (e.g. PTA)



Objectives:

- Microbiota structure stability
- Samples time point ordination
- Time point contribution

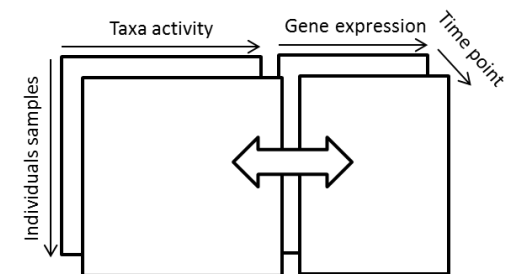
Paired-Tables approach (e.g. COIA)



Objectives:

- Relation between genera and SCFA
- co-structure (co-inertia) description
- co-structure statistical test by permutation

K-Paired-Tables approach (e.g. COIA+PTA)

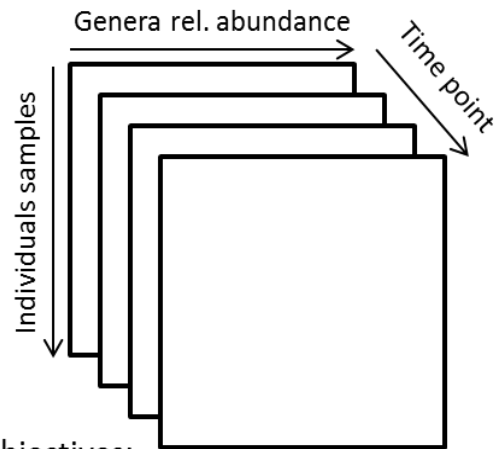


Objectives:

- co-structure stability between gene exp. and taxa
- Time point contribution to this stability

Composition and times series associations using K-tables approach

K-Tables approach (e.g. PTA)

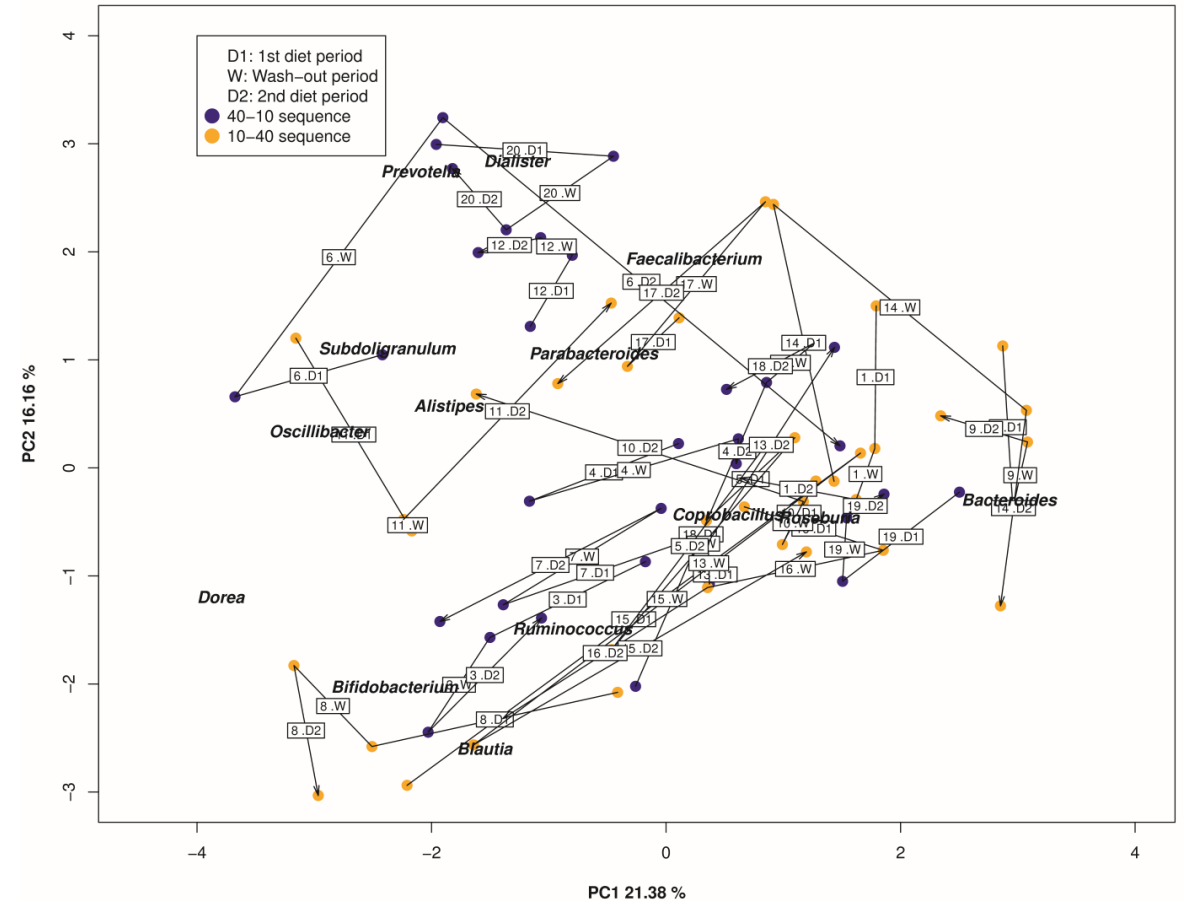


Objectives:

- Microbiota structure stability
- Samples time point ordination
- Time point contribution



Times series Partial triadic analysis



Lesser contribution after diet challenges



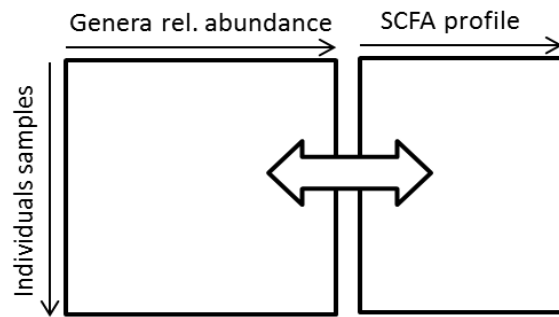
INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

➤ Metabolites and composition associations using Paired-Tables approach

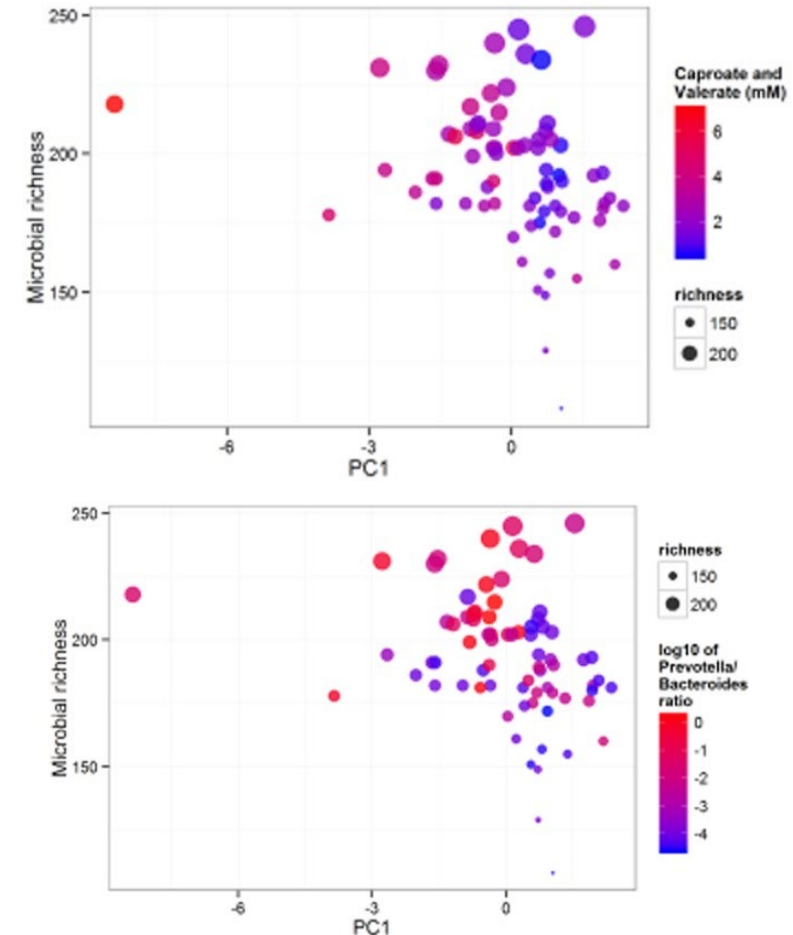
Paired-Tables approach (e.g. COIA)



Objectives:

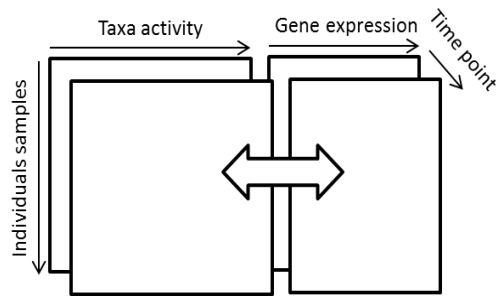
- Relation between genera and SCFA
- co-structure (co-inertia) description
- co-structure statistical test by permutation

Co-inertia between composition and metabolites correlate with microbial richness



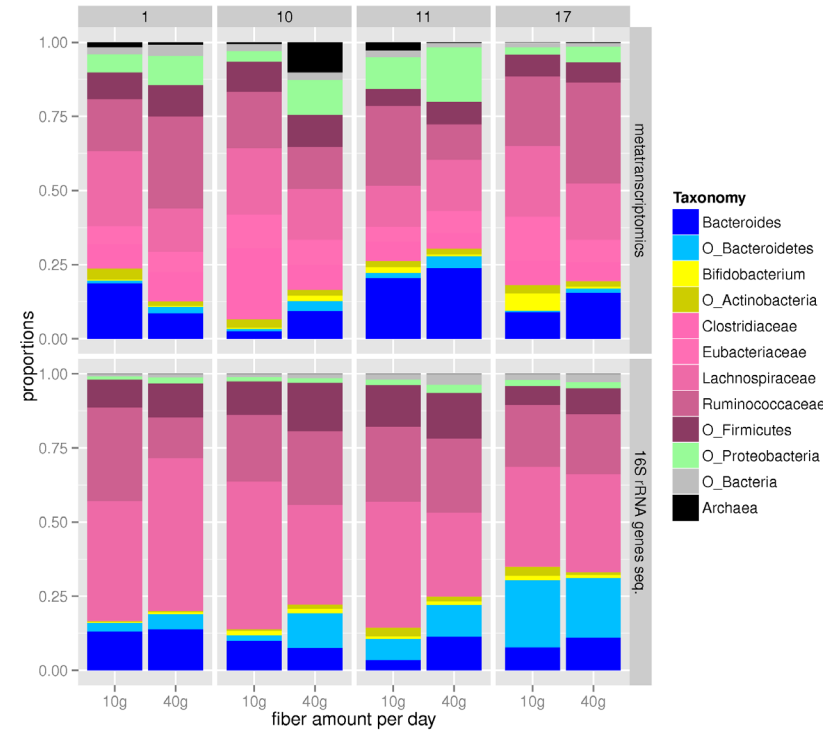
Integrate composition and metatranscriptomics after challenges using K-Paired-tables approaches

K-Paired-Tables approach (e.g. COIA+PTA)



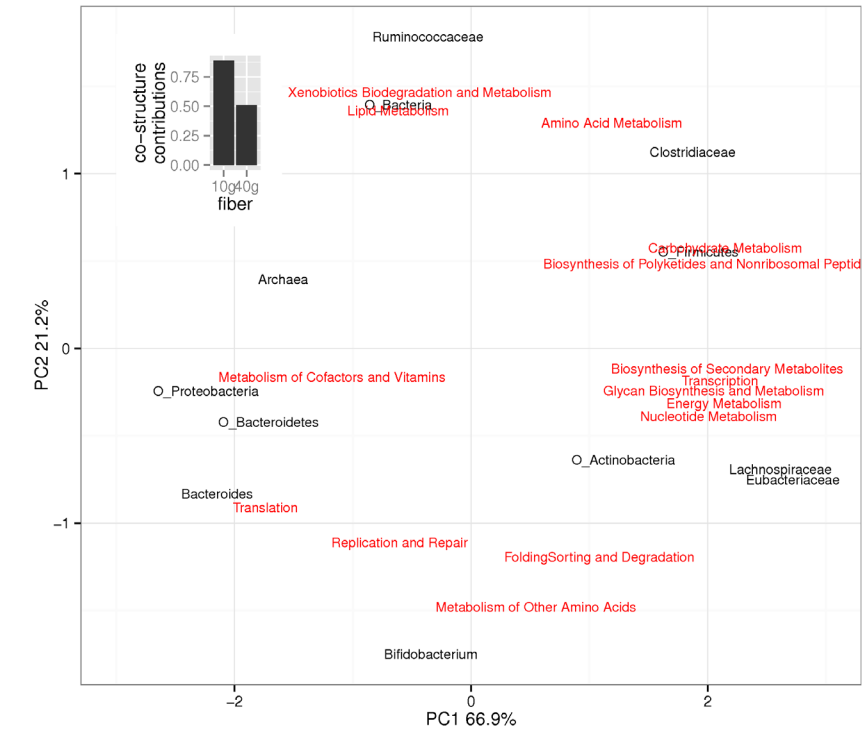
Objectives:

- co-structure stability between gene exp. and taxa
- Time point contribution to this stability



Good overlap between metatranscriptomics composition and 16S-based composition

Co-Inertia Partial Triadic Analysis



The link between MetaT and composition is lower after 40g of fiber than 10g

> Fermented food strains are active in gut microbiome

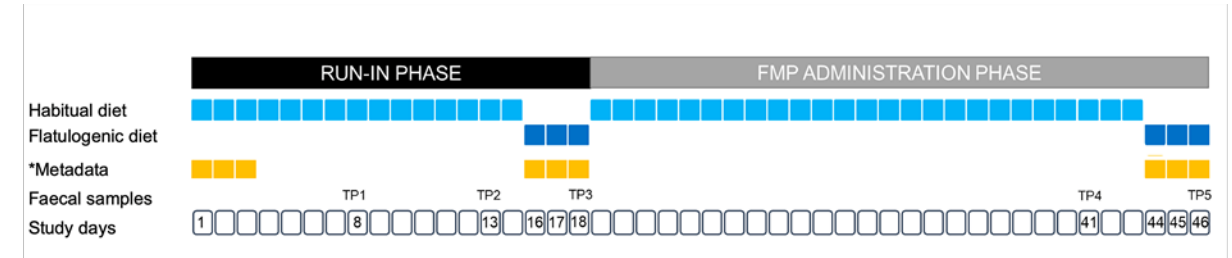
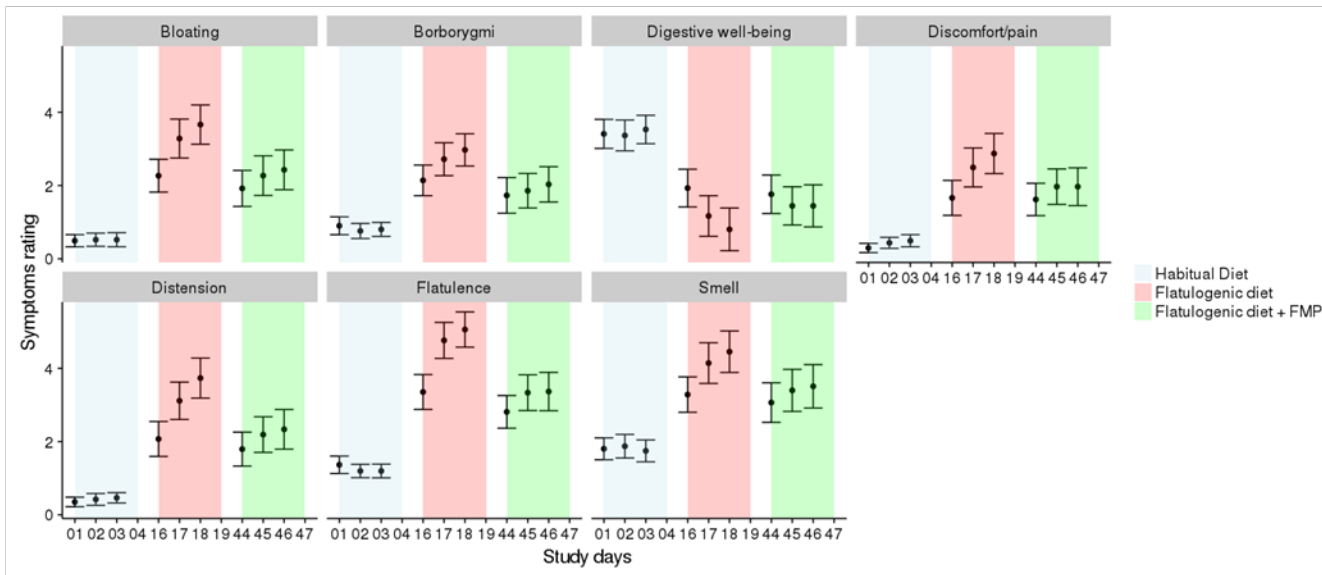
FMP product

Streptococcus salivarius subsp. *thermophilus* strains (CNCM I-2773, CNCM I-2130, CNCM I-2272),

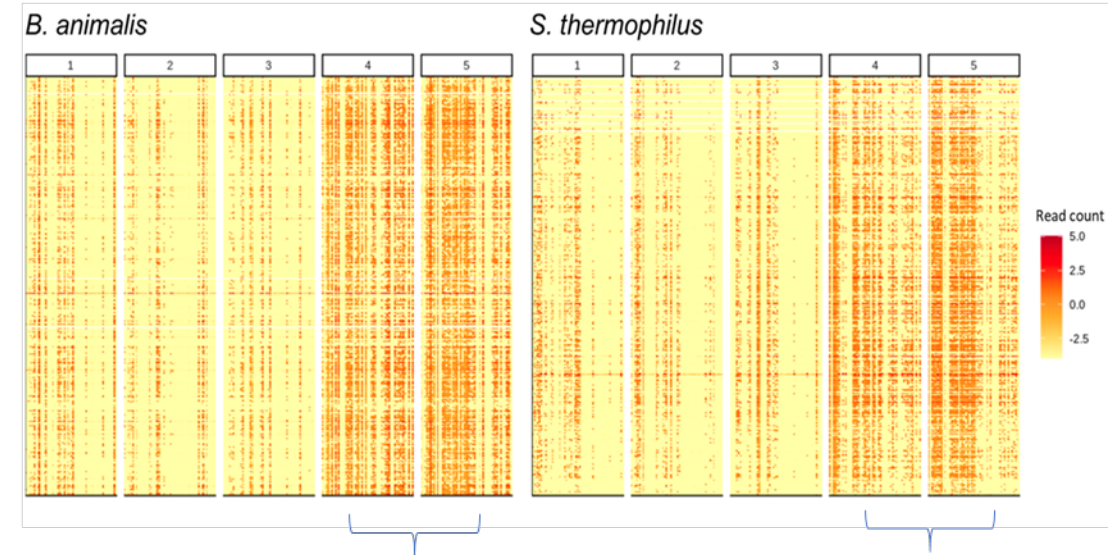
Lactobacillus delbrueckii subsp. *bulgaricus* (CNCM I-1519),

Bifidobacterium animalis subsp. *lactis* (CNCM I-2494)

Lactococcus lactis subsp. *lactis* (CNCM I-1631).

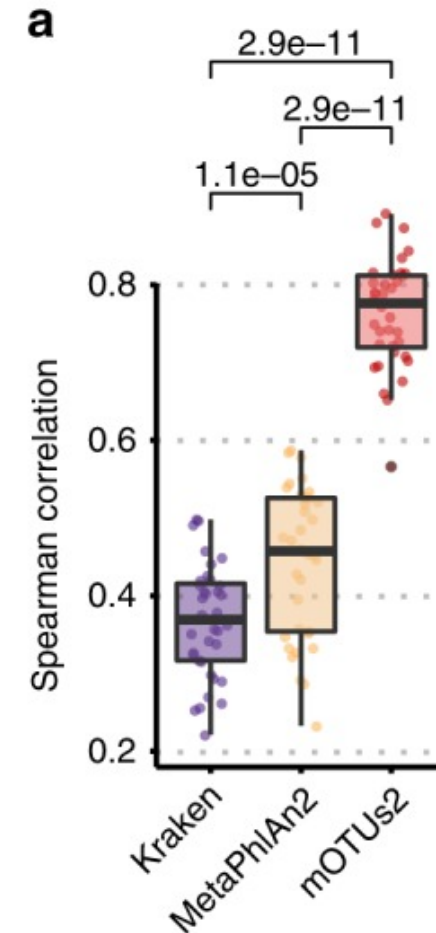
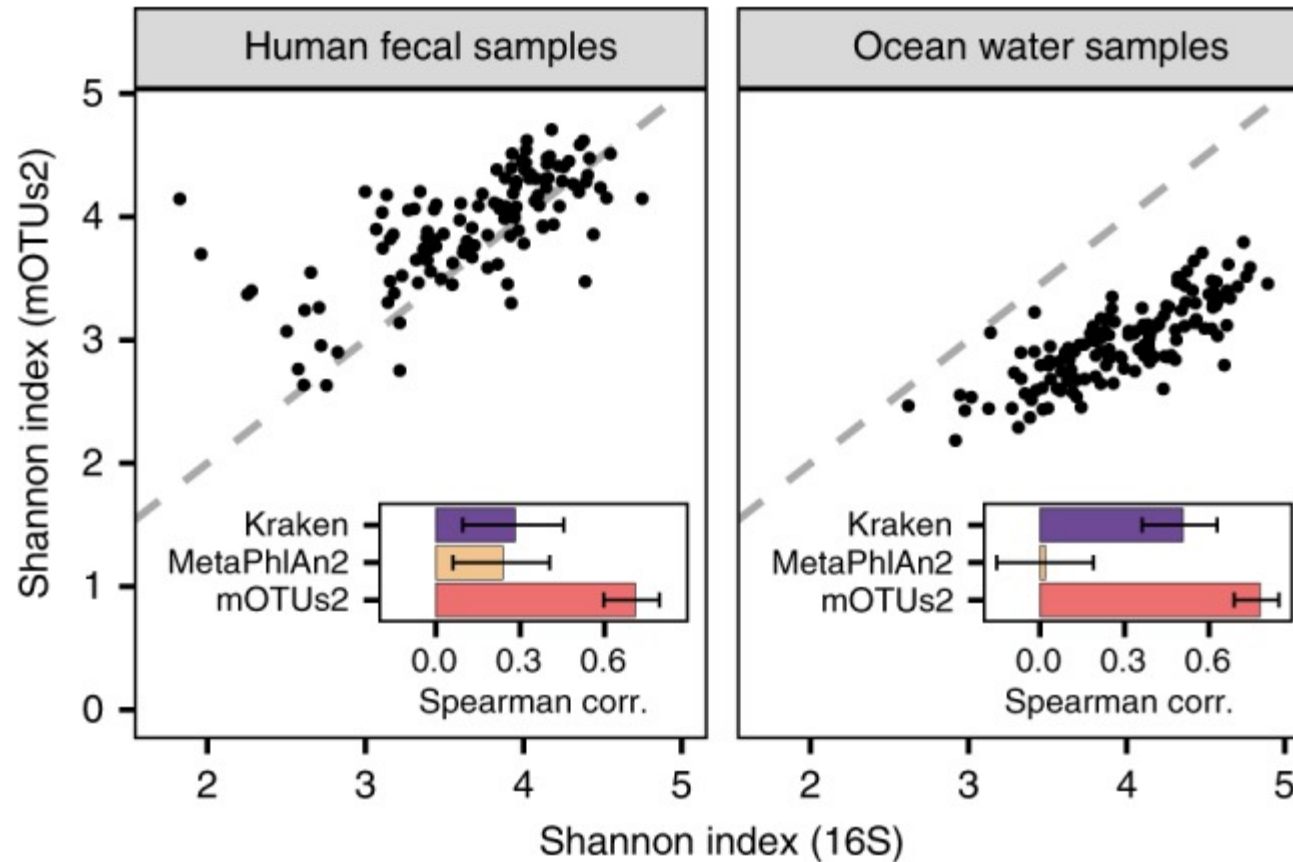


flatulogenic diet (61% carbohydrates, 25% proteins and 14% fat, 27 g of fiber per day)

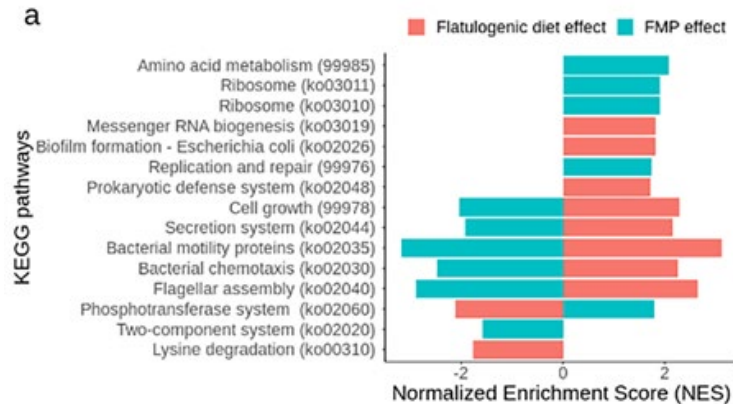


After FMP consumption, FMP active strains detected by metatranscriptomics

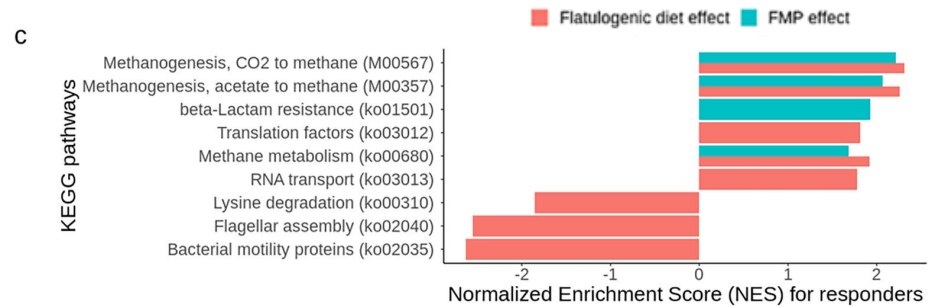
Metatranscriptomics as proxy for composition analysis



➤ Fermented food consumption was associated with gut microbiota metabolic changes



Flatulogenic diet might increase cell motility and cell growth whereas FMP consumption exhibited the opposite effect

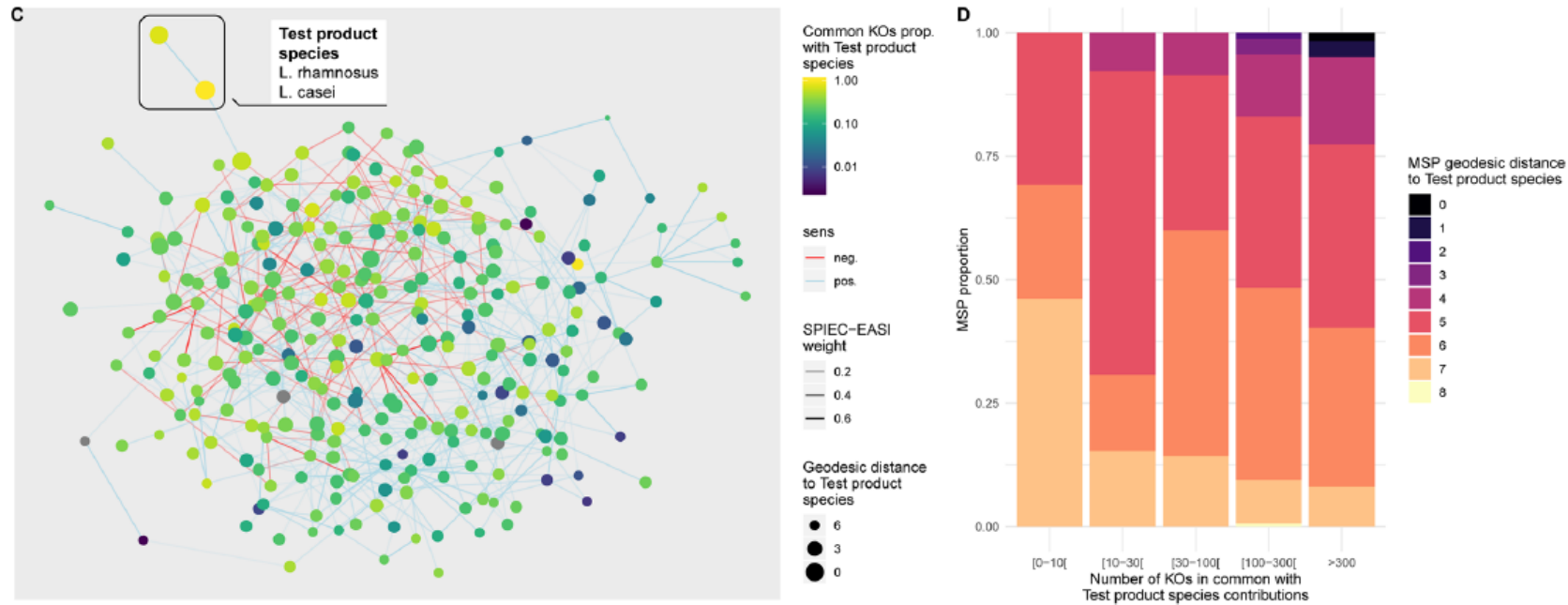


After FMP intake (TP5), higher responder group showed higher methanogenesis activity than low-responders

limma-voom + FSEA

> Gut microbiome functional enrichment by fermented food is personalized

Modular coalescence between gut and food species



“Engraftment” as function of existing niche and metabolic pathway

Alvarez et al., 2020

> What about Bifidobacterium Ecology in Adult Gut Microbiome?

Gut metagenomes



• Metadata •



Antibiotics
intake



Age



Western /
non-Western



Health /
disease

• Gut microbiome •



Gut microbiome profile
(~ 9,000 subjects)

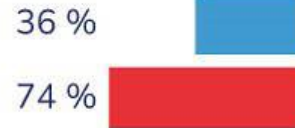


Bifidobacterium community
Dirichlet Multinomial Mixtures
(~ 5,000 subjects)

Partitions depleted
of *B. longum*
or *B. adolescentis*



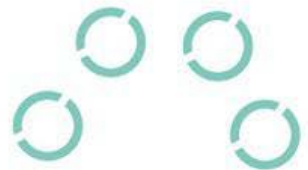
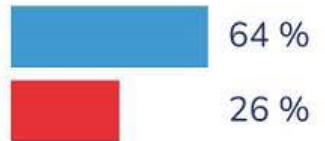
↑ *E. coli*



Partitions enriched
in both *B. longum*
and *B. adolescentis*



↑ Butyrate
producers



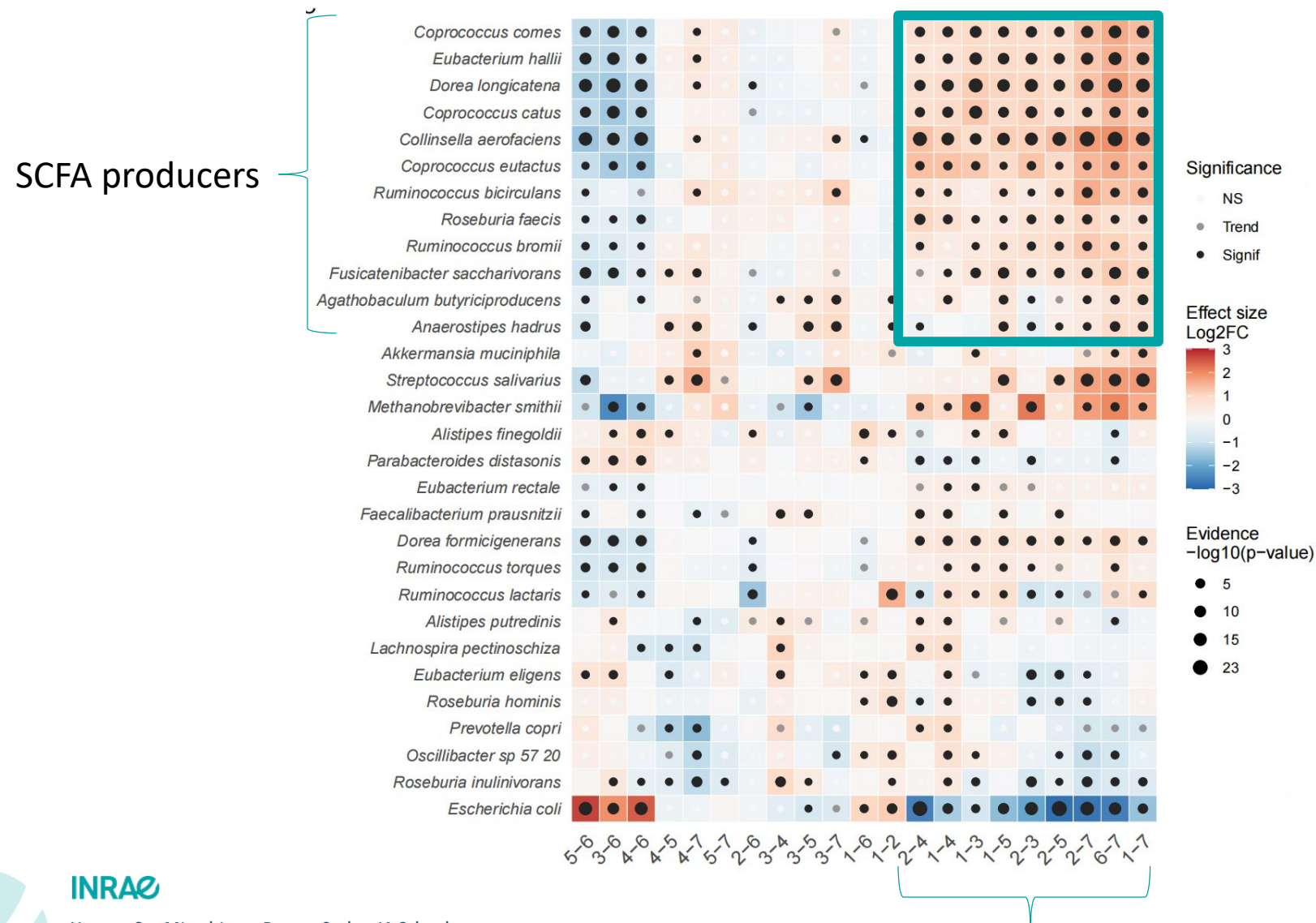
Gut microbiome
alpha-diversity

Gut microbiome
composition

Health / Disease
(Proportion of partitions
in healthy subjects (blue)
or subjects with diseases (red))

***Bifidobacterium* MAGs**
Pangenomics comparisons

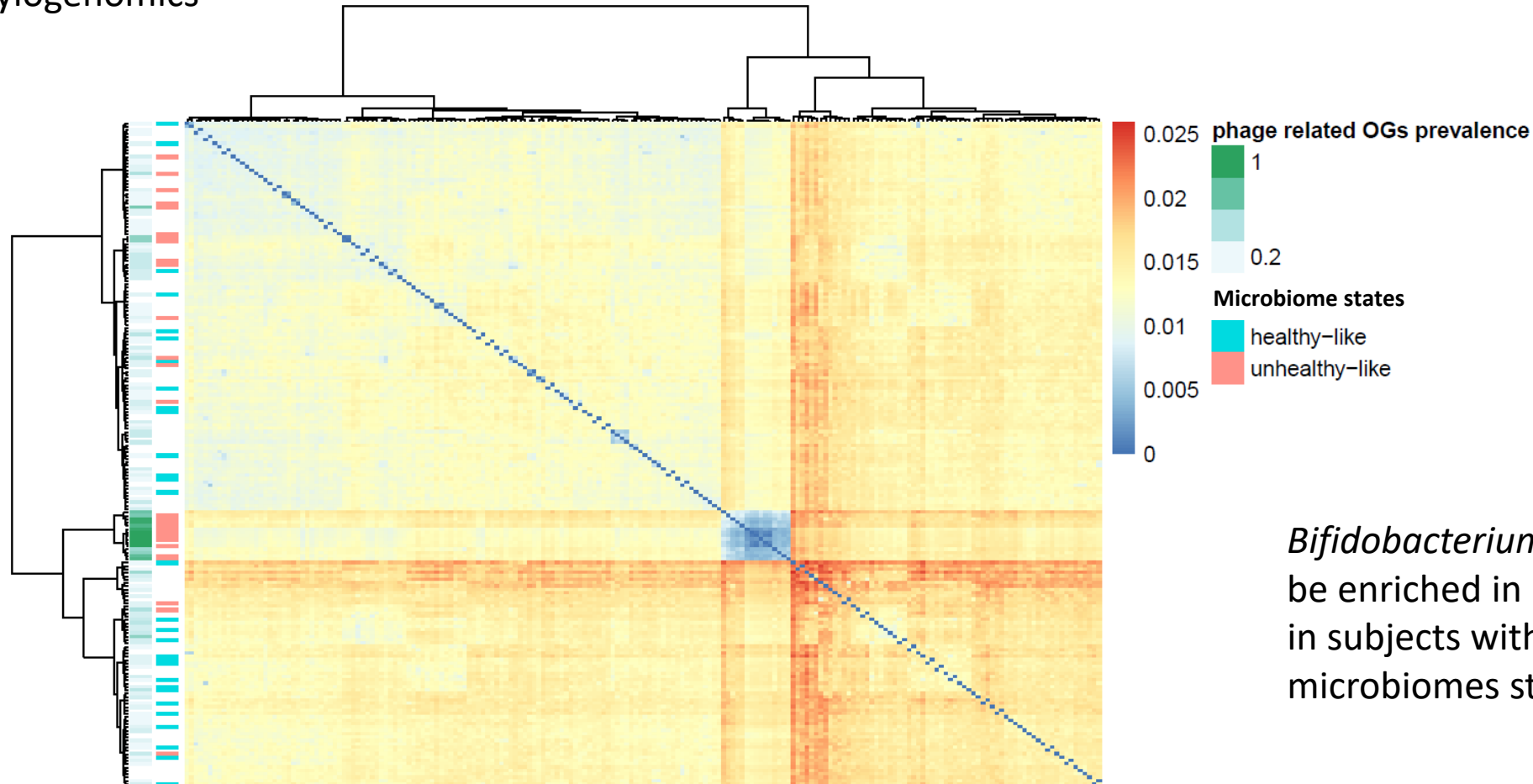
➤ High SCFA producers in gut microbiome with High abundant B.adolescentis



High B.ado gut microbiome versus others ➡ « Healthy-like states » ^{p. 40}

> Altered states showed functional differences within species

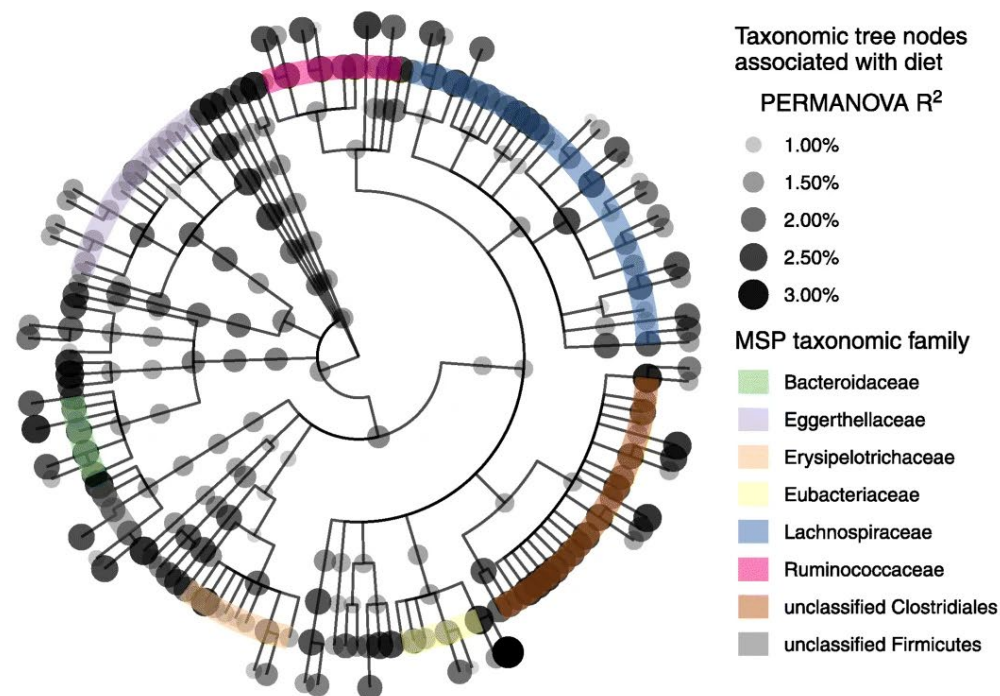
400 *B. bifidum* MAGS
phylogenomics



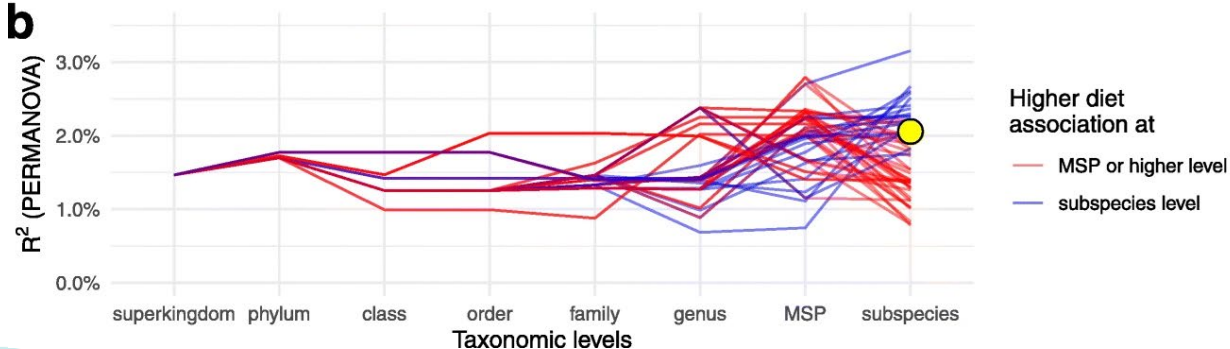
Bifidobacterium bifidum strains may be enriched in phage-related genes in subjects with altered gut microbiomes states

Some species are more associated with diet at subspecies levels

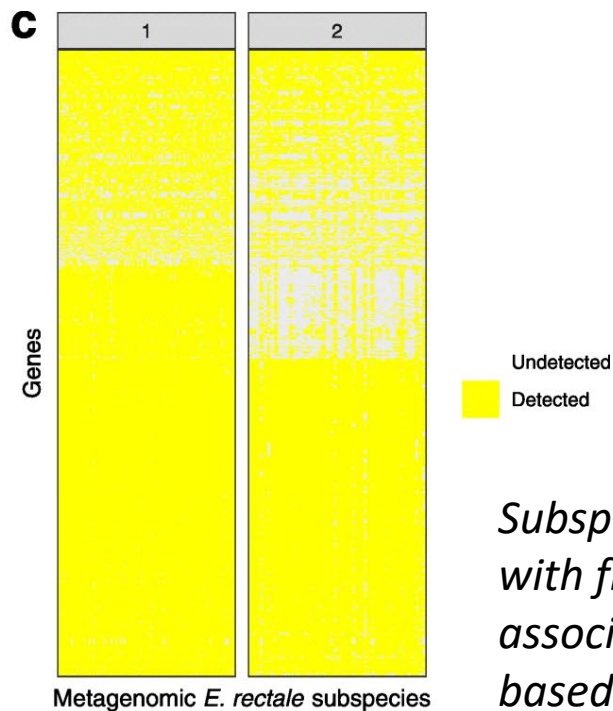
a



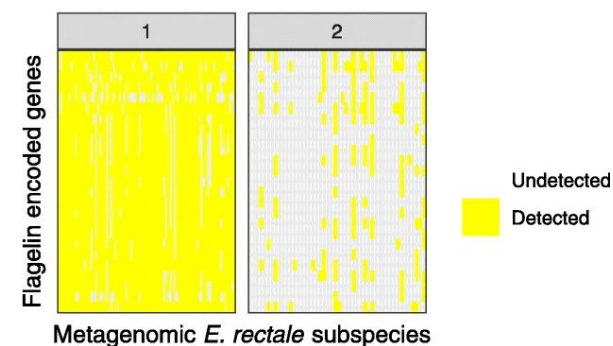
b



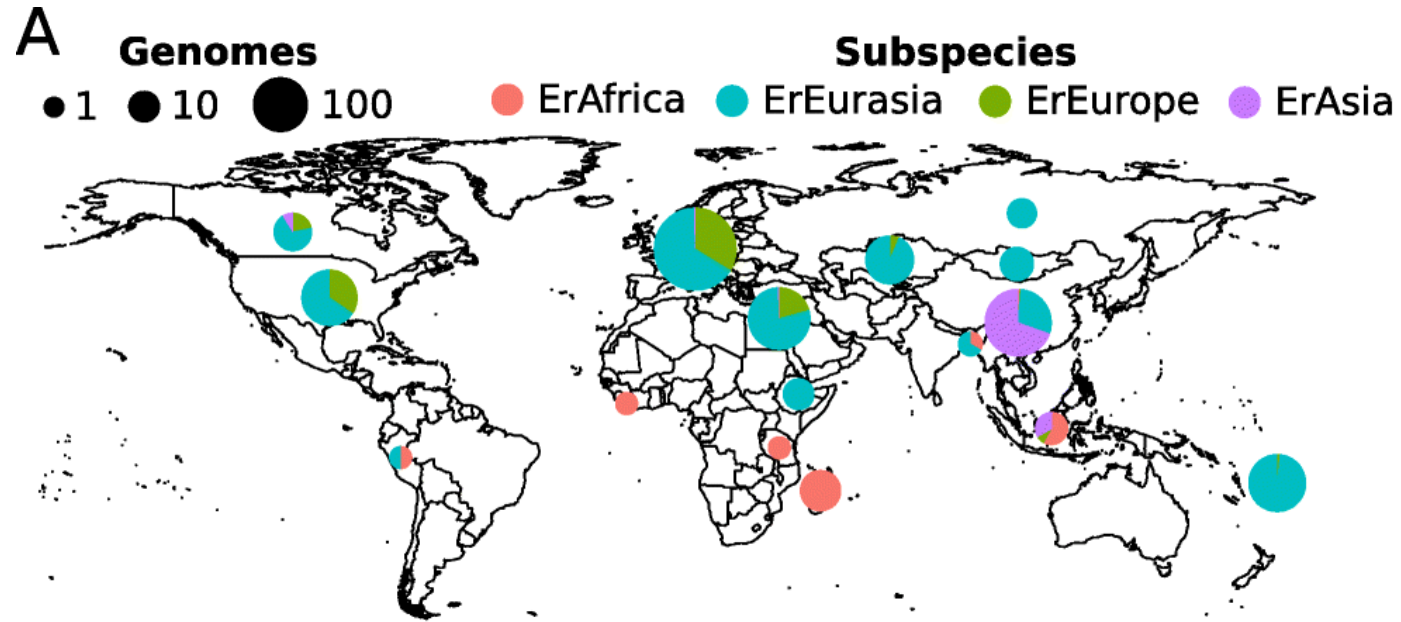
c



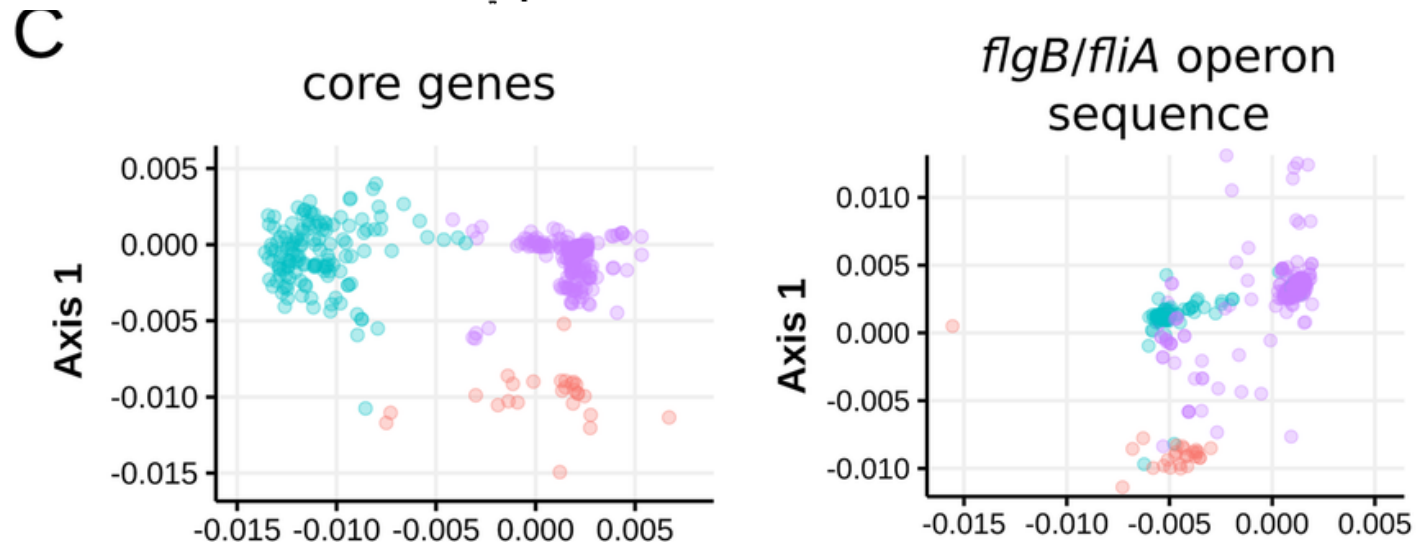
Subspecies E. rectale with flagellin associated with animal based product

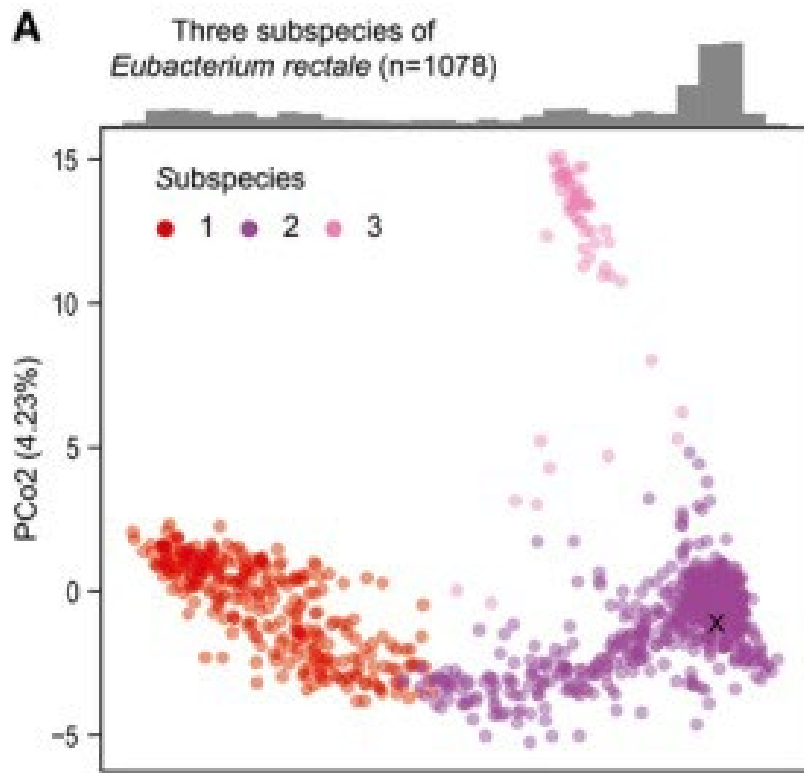


Tap et al. Microbiome (2021)

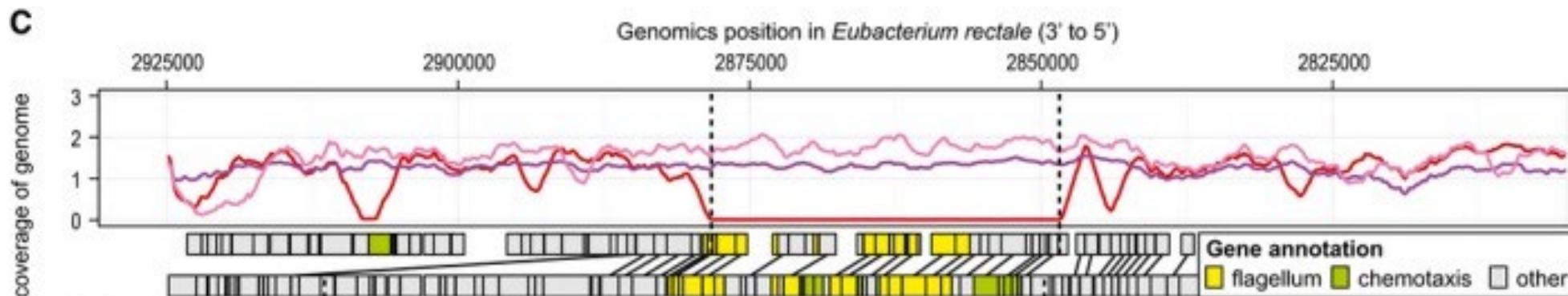
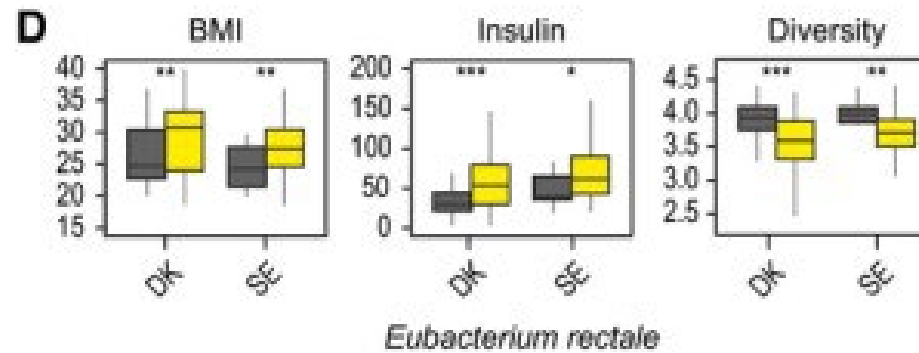


ErEurope strains
are immotile due
to loss of motility
operons

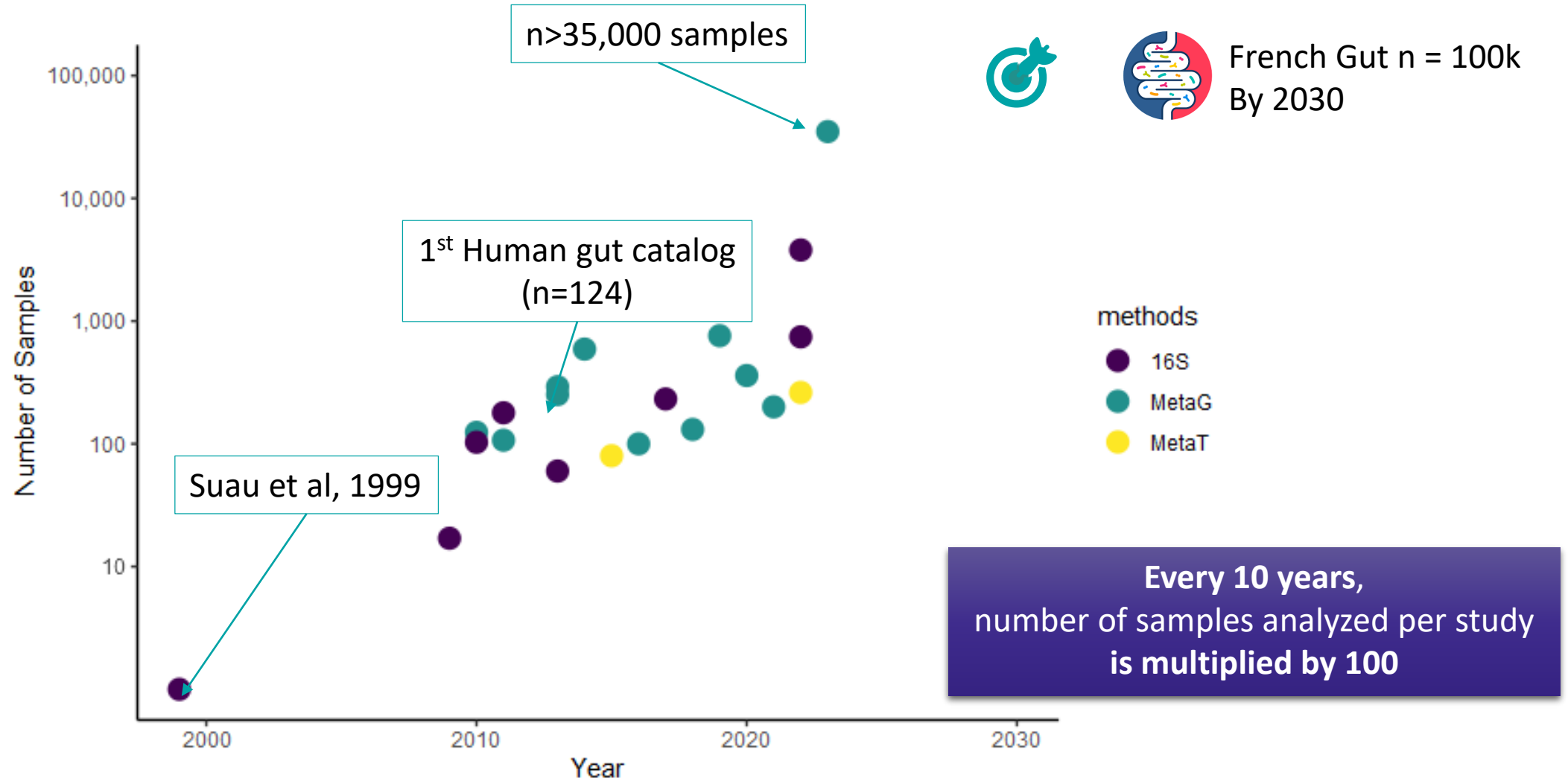




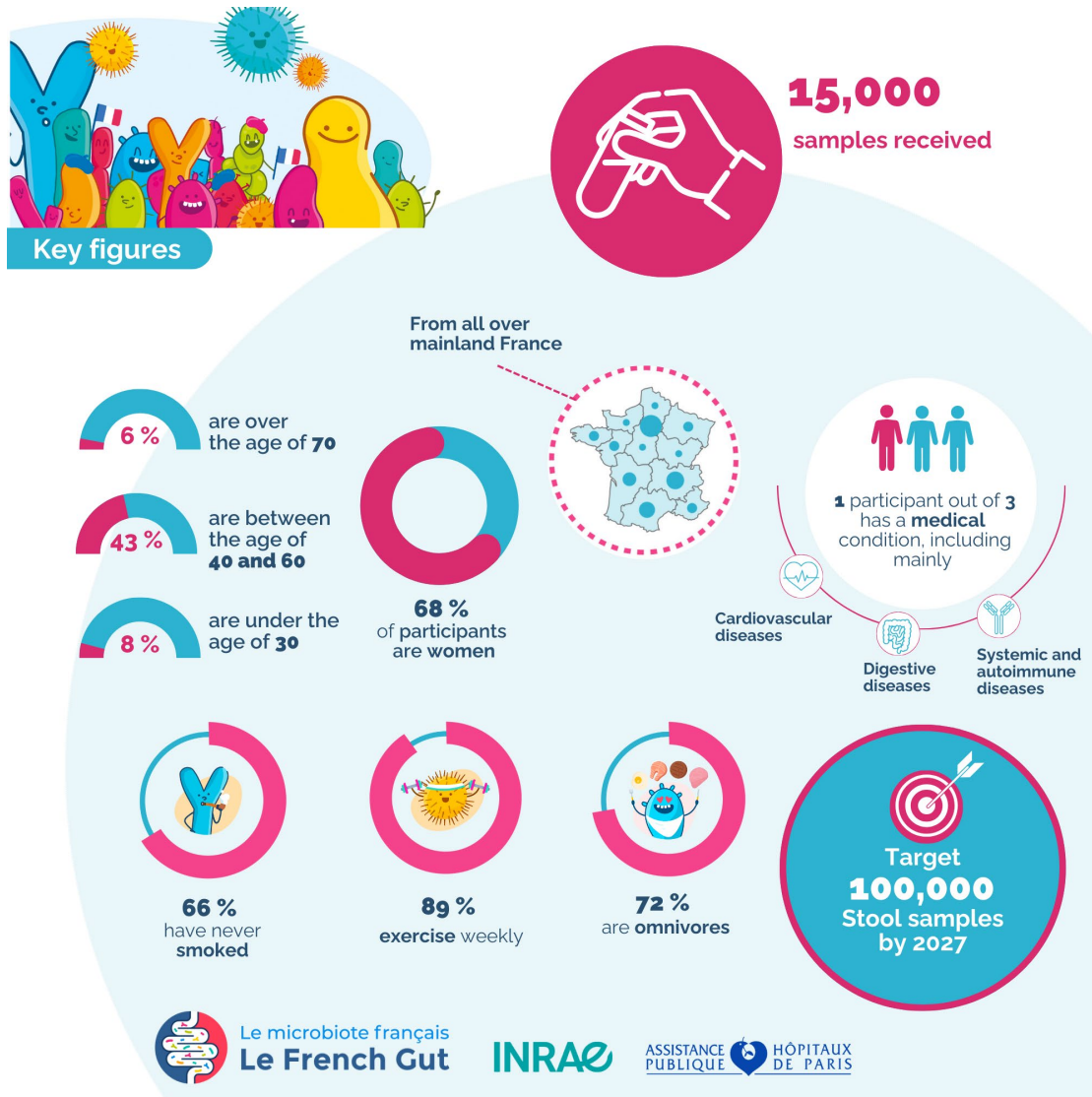
BMI and insulin resistance are significantly higher in individuals who predominantly harbor the flagellum-carrying *E. rectale* subspecies



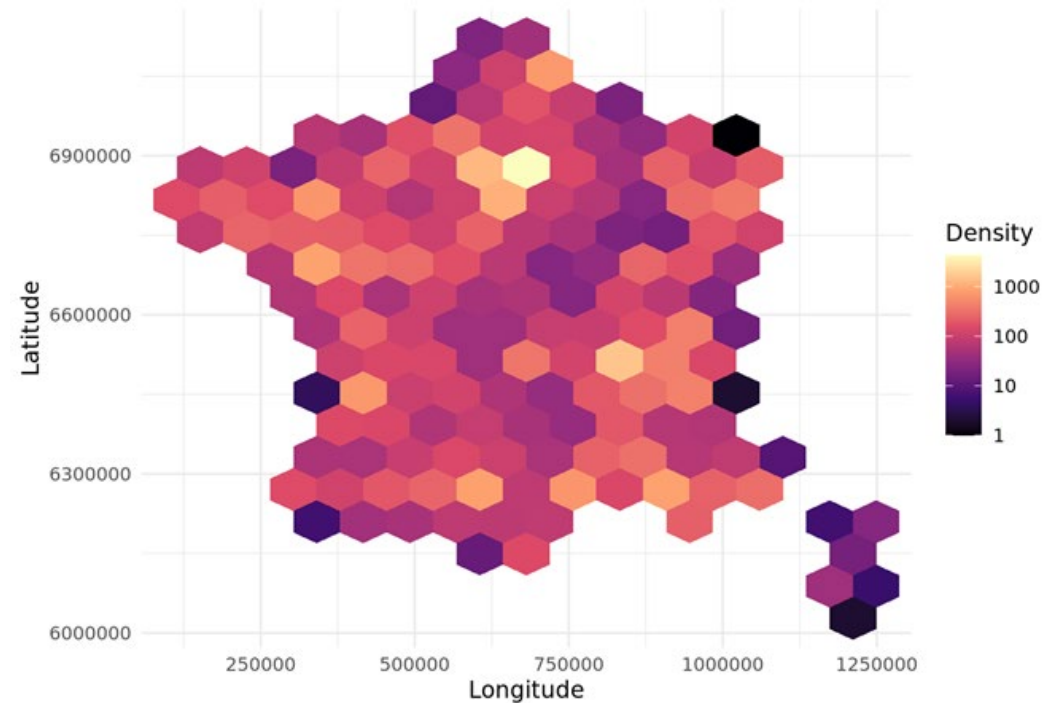
Gut microbiome studies: exponential growth of sample numbers



Le French Gut aka "le microbiote français"



Participant Living Locations Heatmap

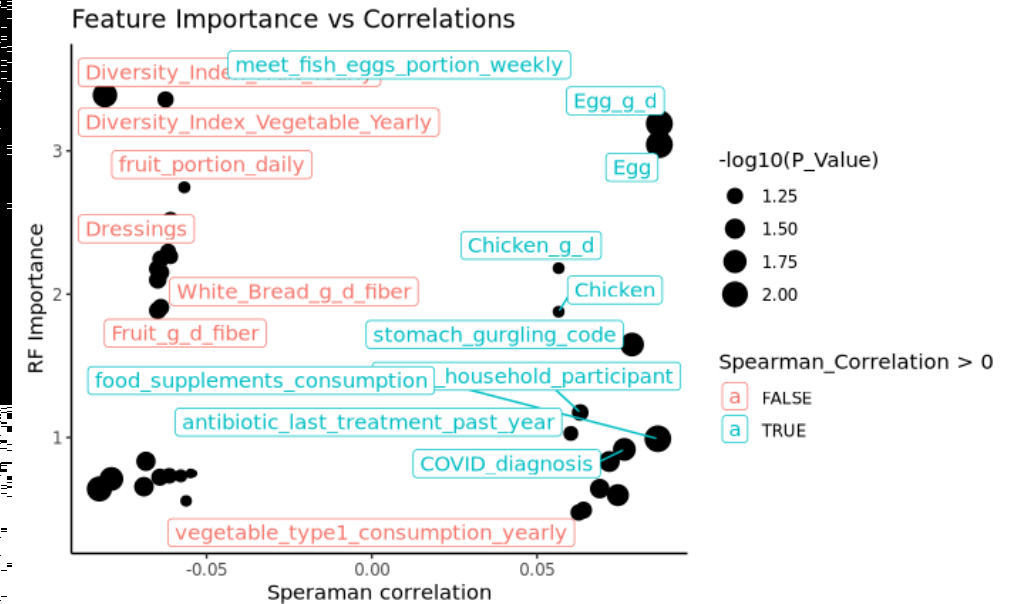
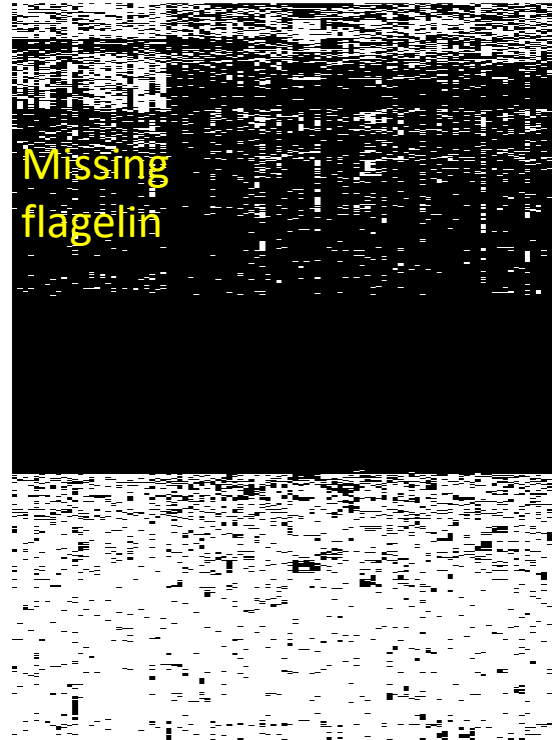
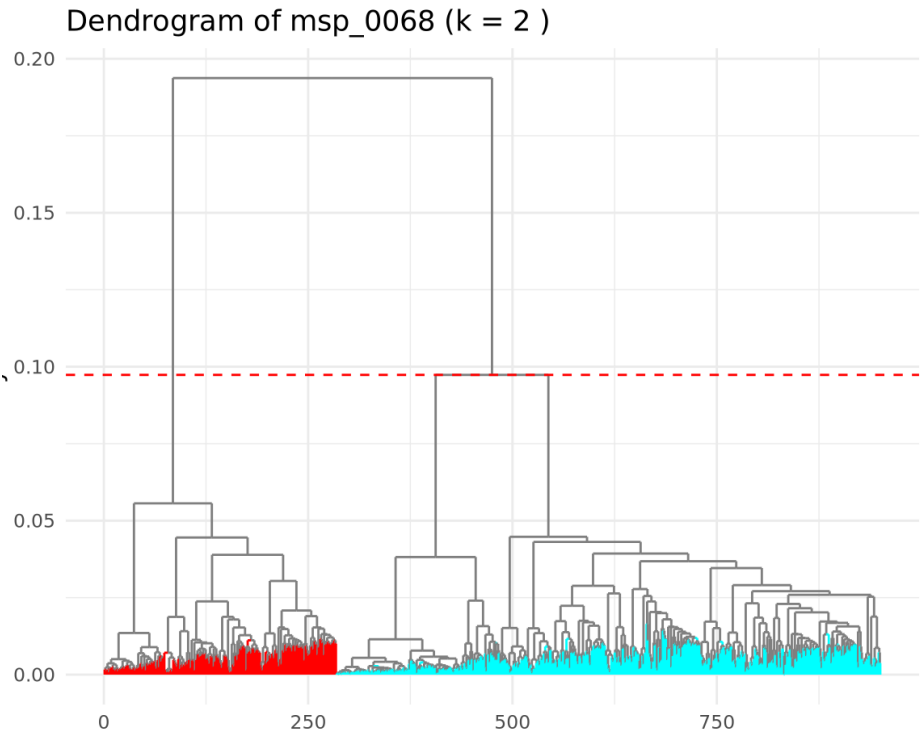


INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

➤ Metagenopolis Meteor pipeline can recapitulate findings in French Gut data



Diet profile and habits associated with *E. rectale* strains with flagelin



<https://github.com/metagenopolis/meteor>

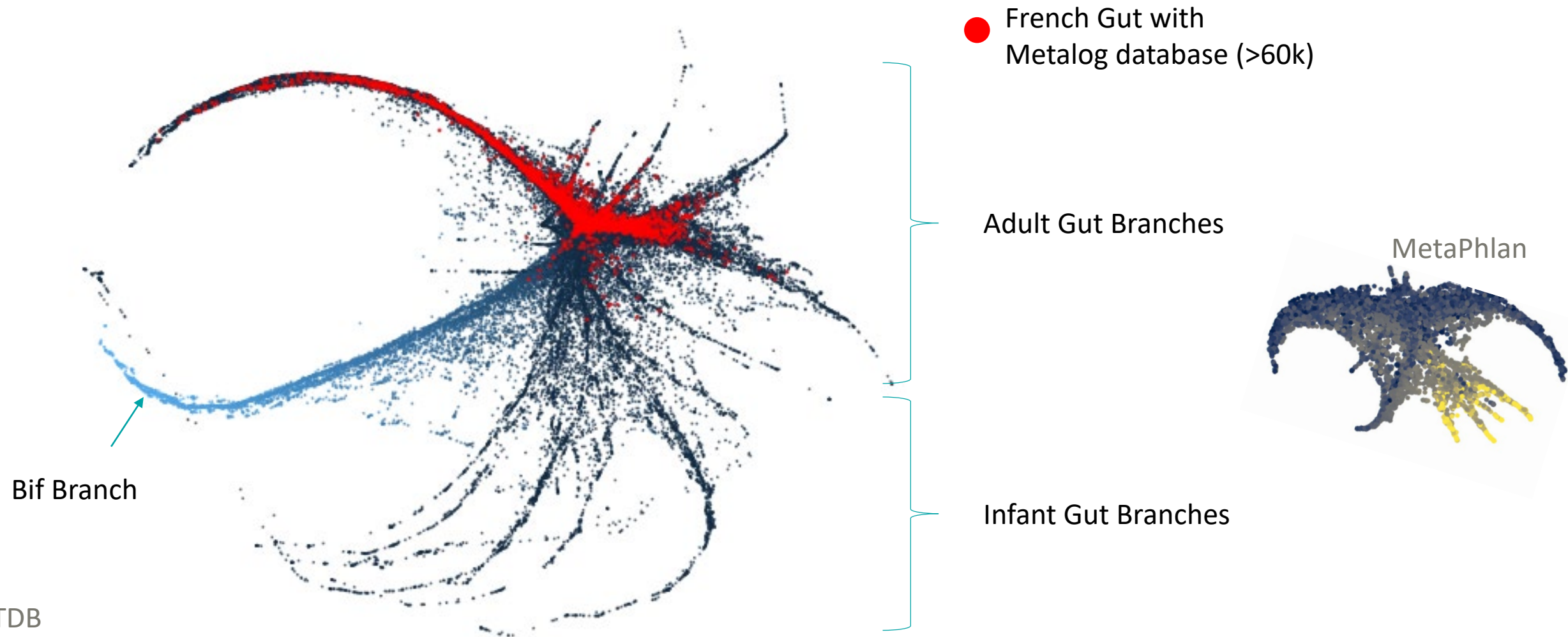


INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

French Gut microbiome participants were spread into branches



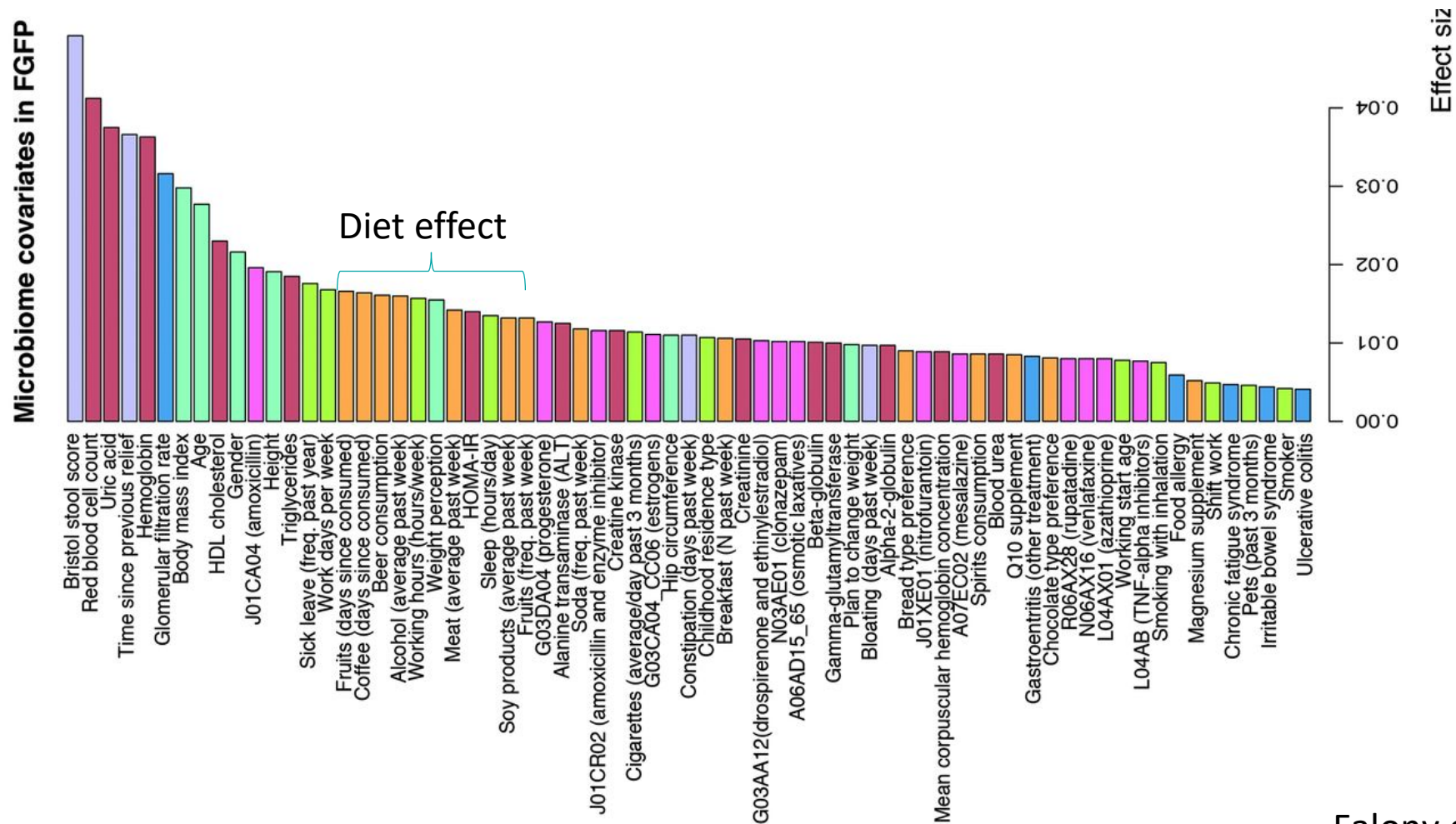
INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

<https://github.com/KrishnaswamyLab/phateR>
<https://metalog.embl.de/>

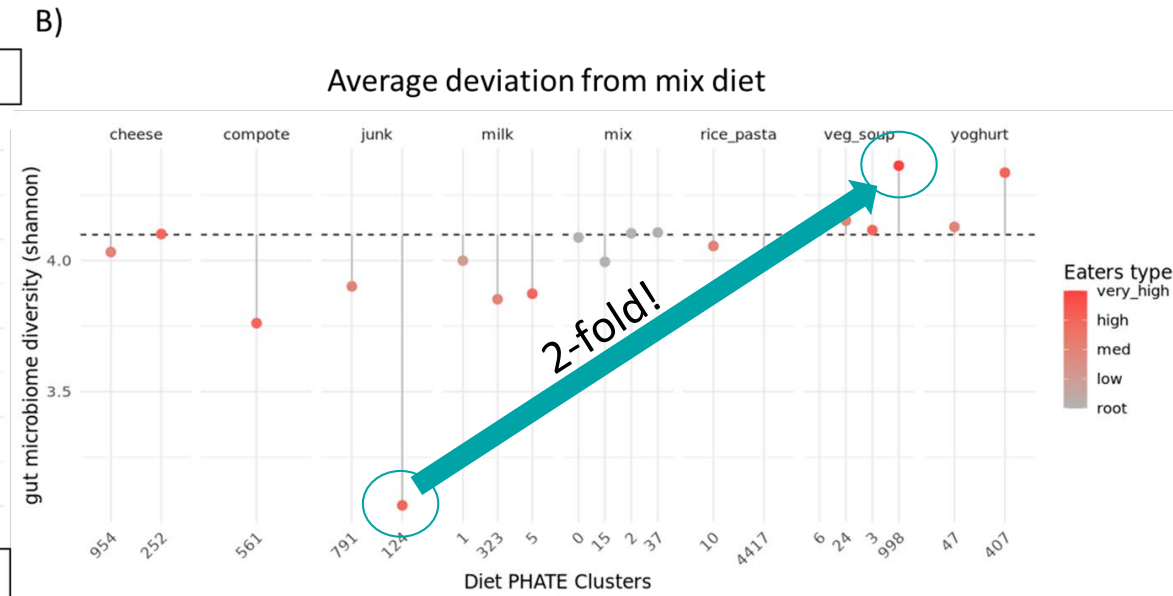
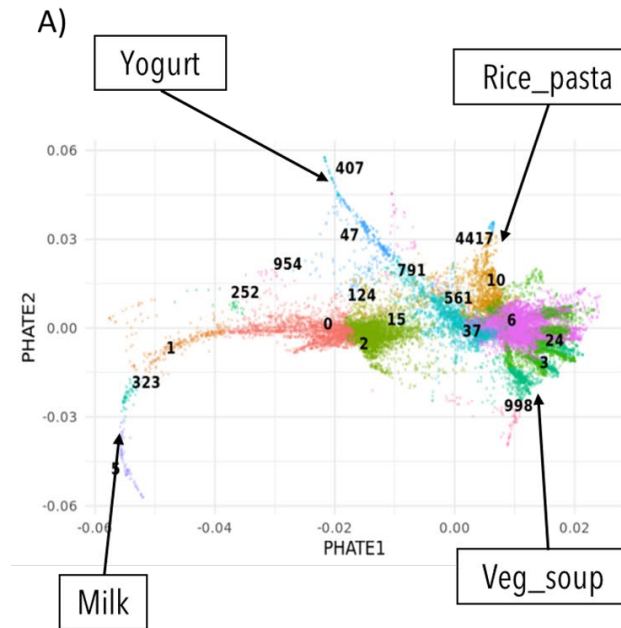
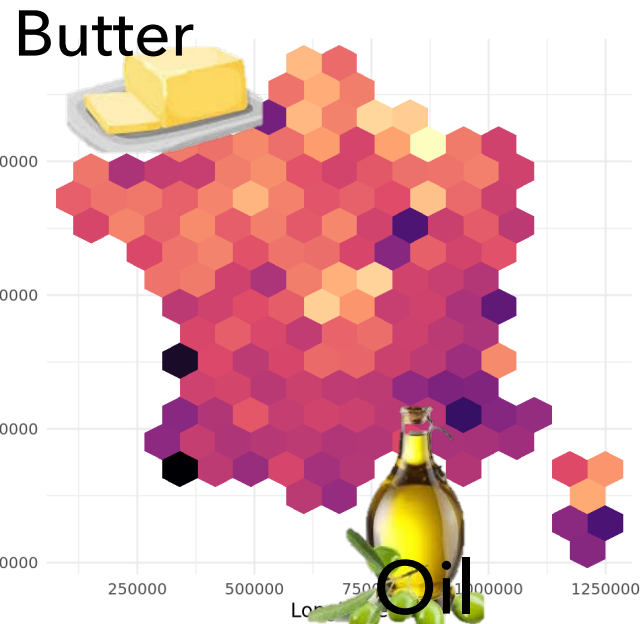
➤ distance-based redundancy analysis showed minor associations with diet compared to physiological covariates



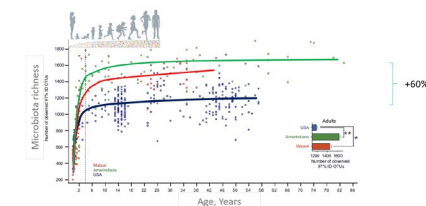
Falony et al. Science. 2016

dbRDA - Vegan p. 49

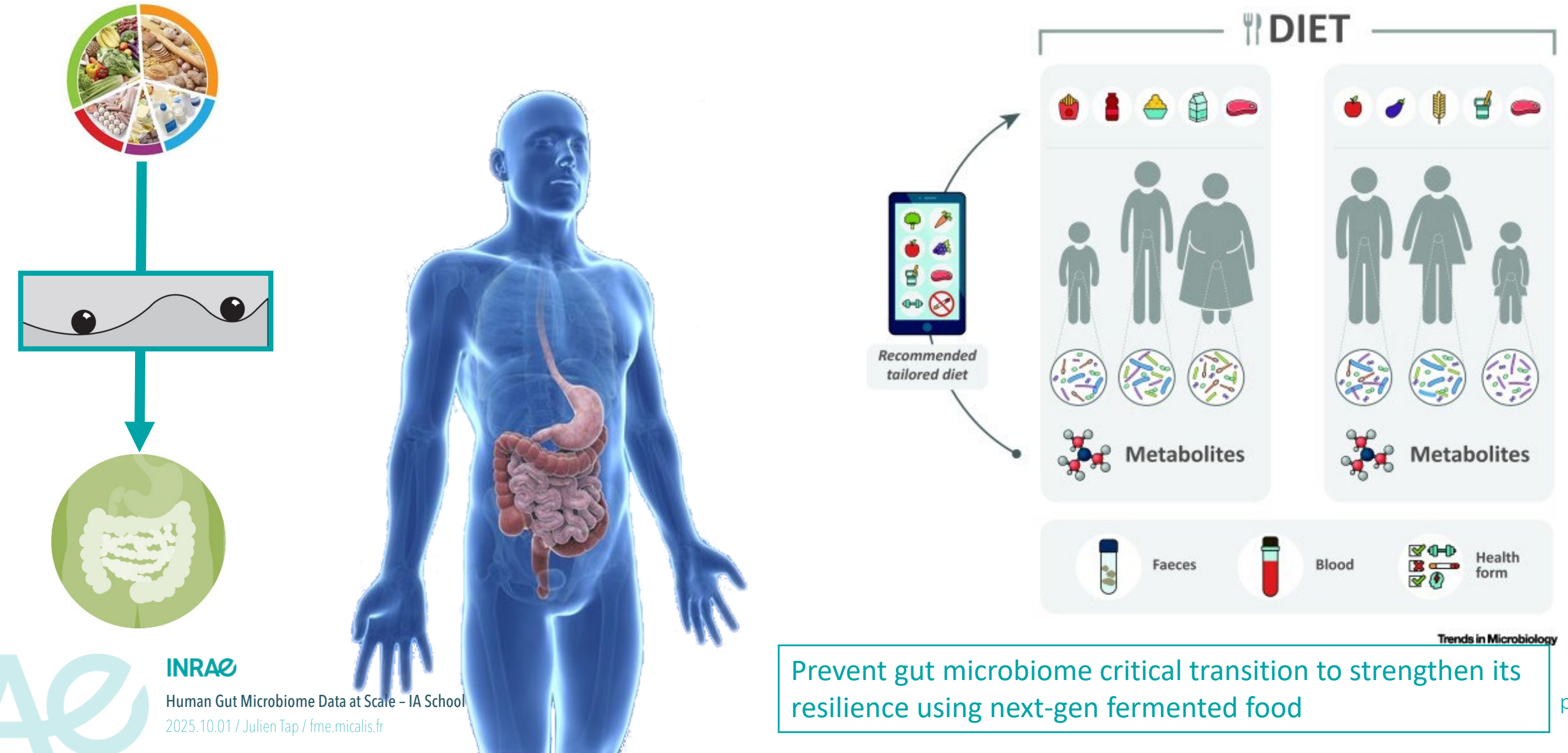
➤ Dietary branches analysis can explain up to 50% gut microbiome richness



With large data, any subtle difference is significant, so the effect size is the way to interpret data



> Endgame : From gut microbiome states to next-gen fermented food



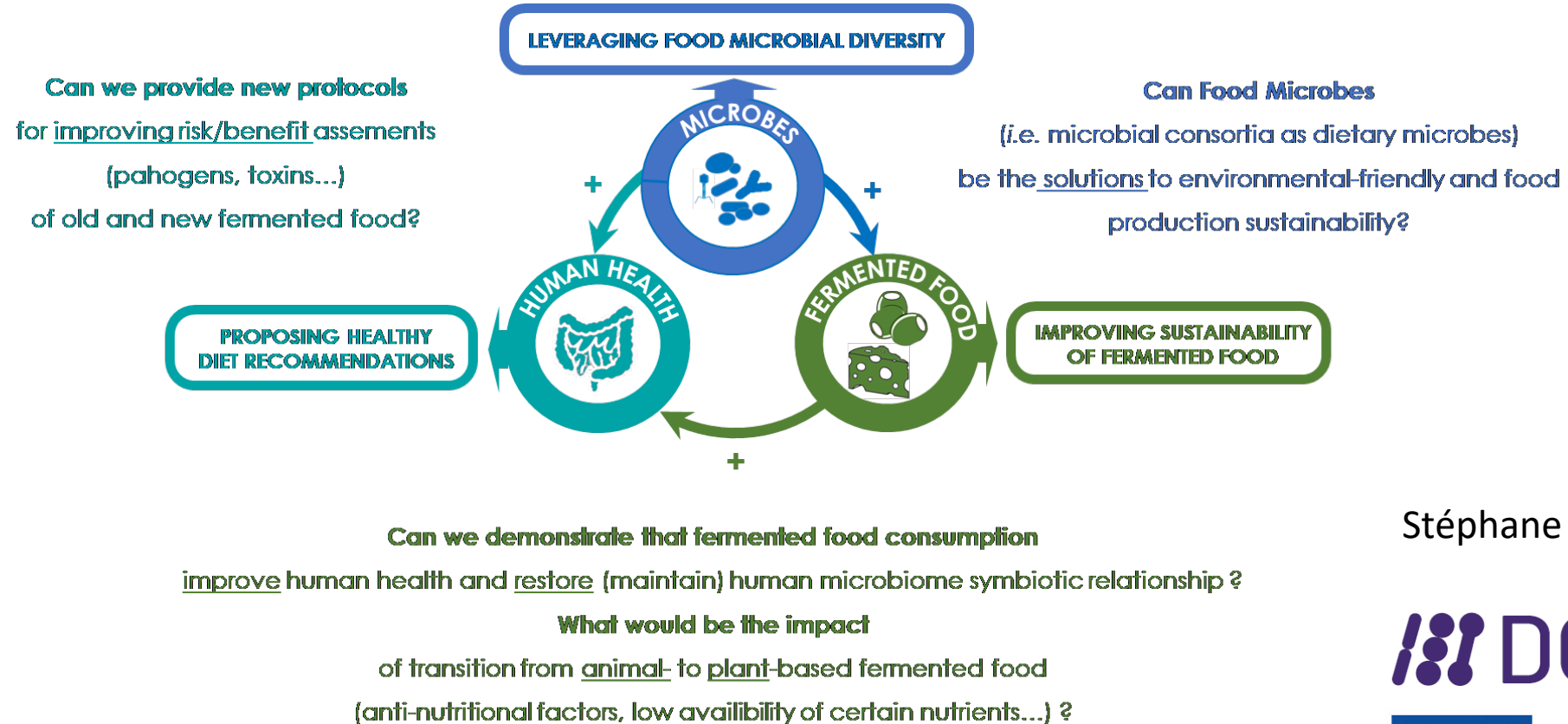
INRAE

Human Gut Microbiome Data at Scale – IA School
2025.10.01 / Julien Tap / fme.micalis.fr

Prevent gut microbiome critical transition to strengthen its resilience using next-gen fermented food

Trends in Microbiology

> How to design next-gen and sustainable fermented food?



Stéphane Chaillou (Coord.)

 **DOMINO**



Funded by
the European Union

INRAE

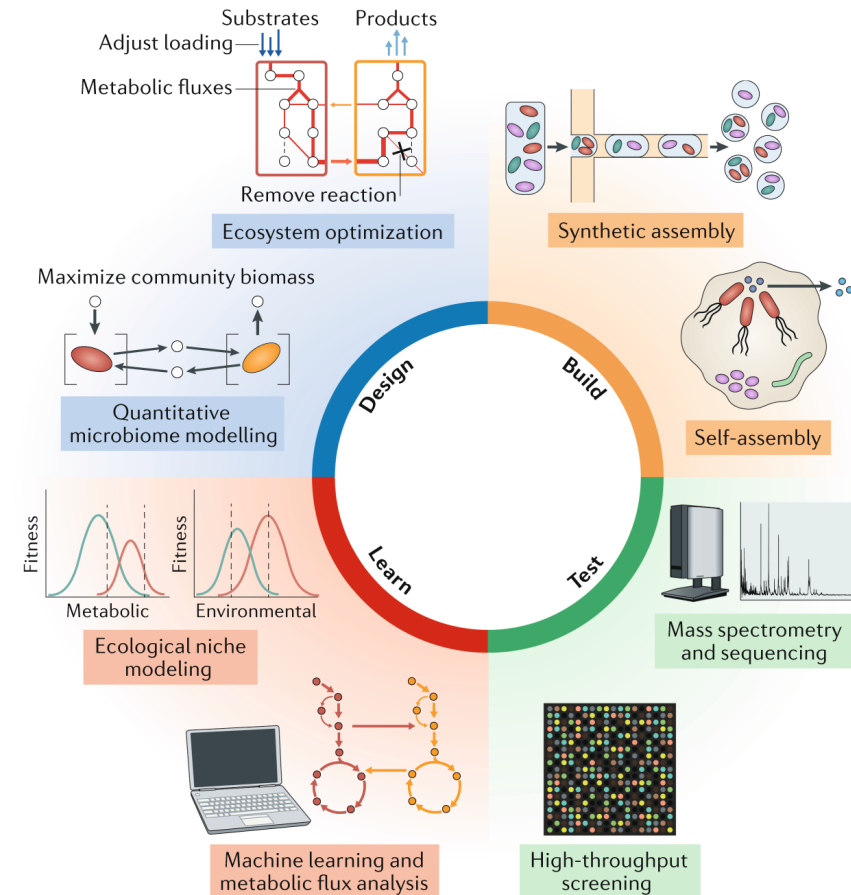
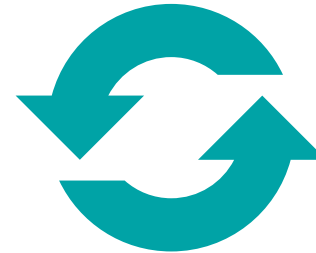
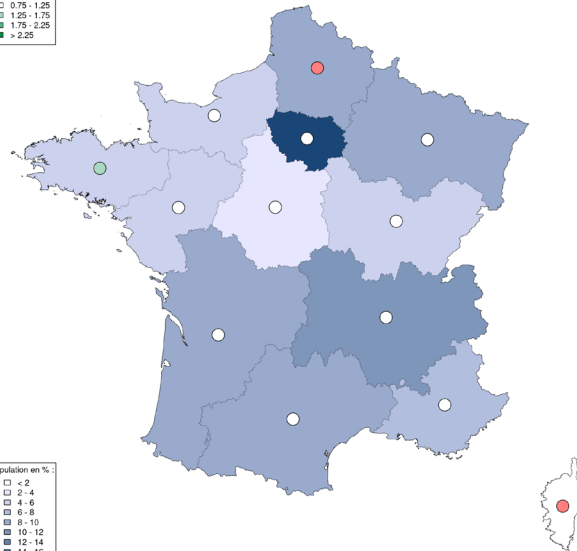
Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

Big data population scale study connect to microbiome engineering



Le microbiote français :
Le French Gut

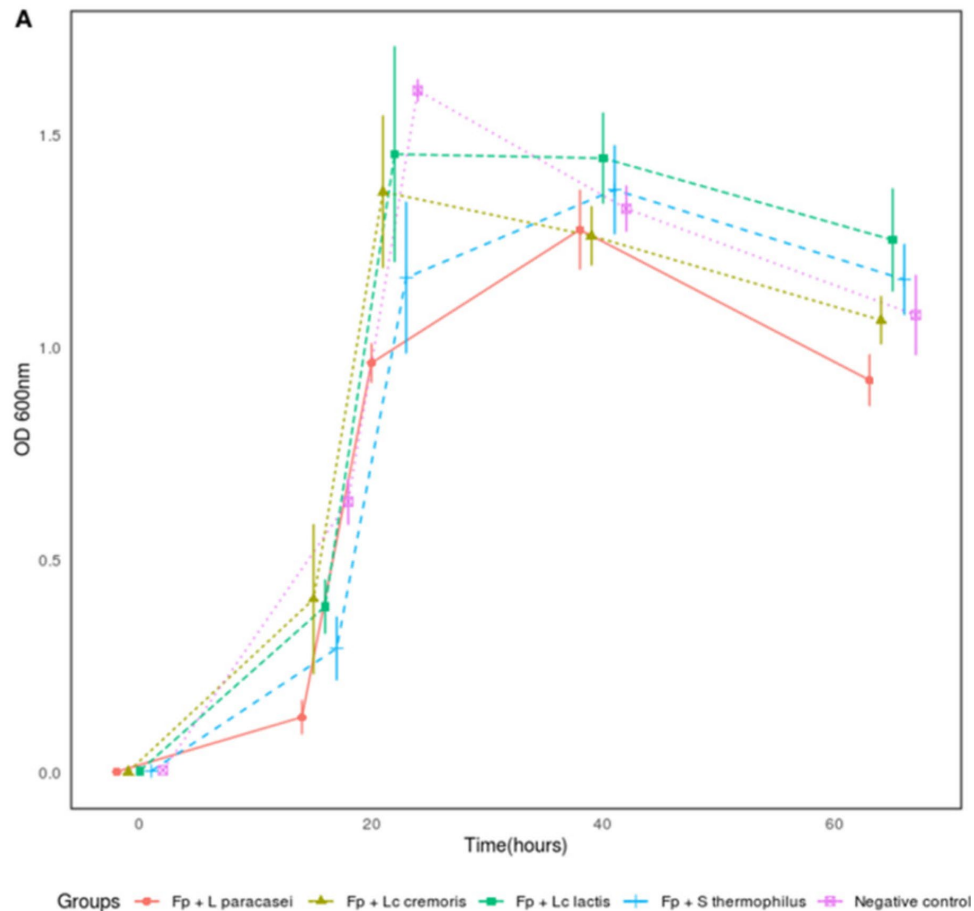


INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

➤ Towards selecting strains with abilities to prevent tipping point in gut microbiota



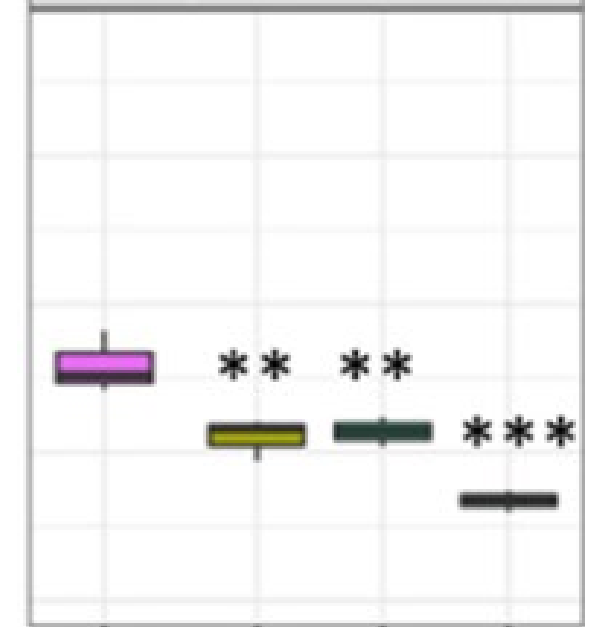
umax



F.prau

F.Prau
+ LABs lysate

Decay slope

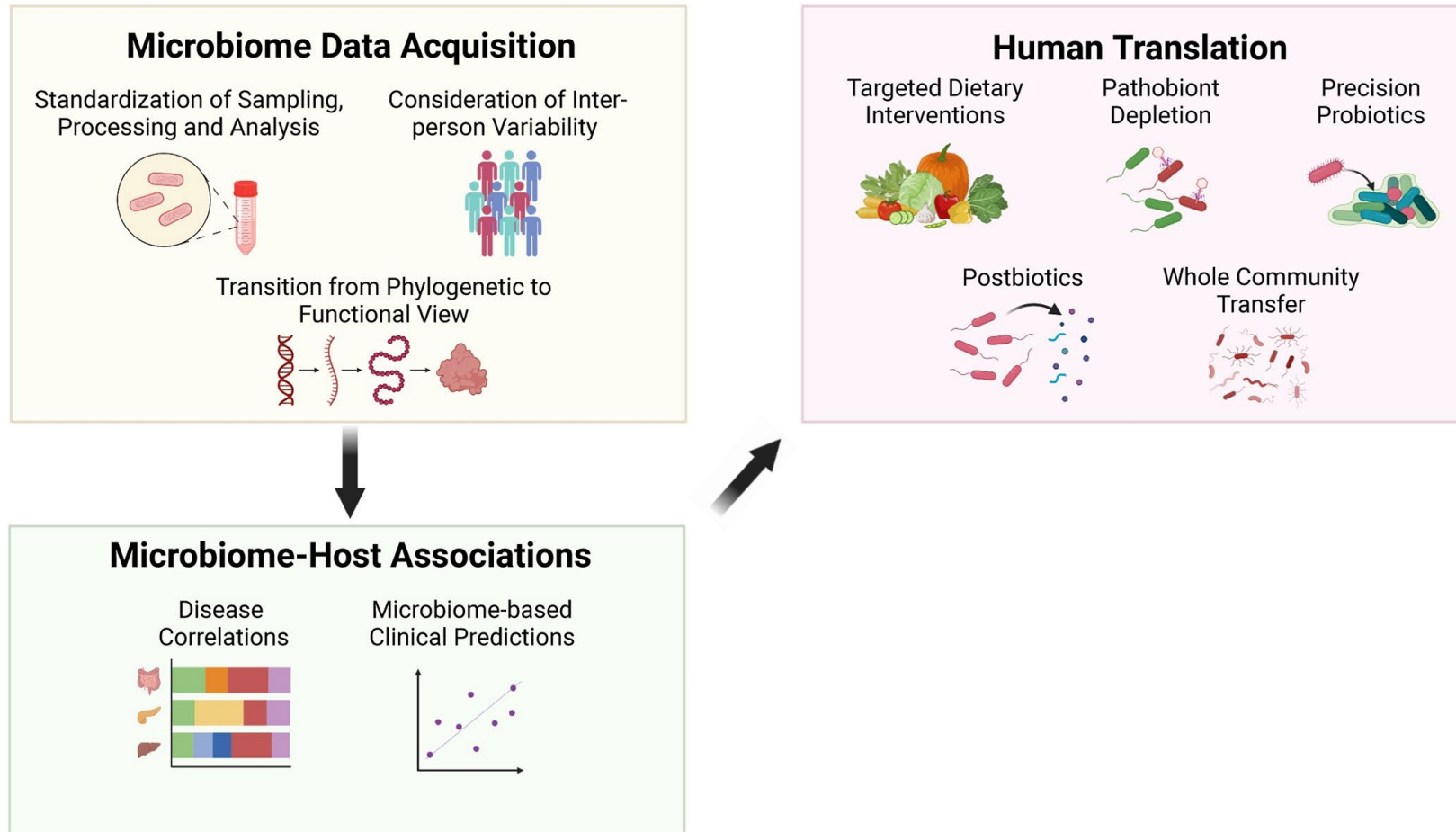


F.prau

F.Prau
+ LABs lysate

Increase of peptidoglycan levels for cell wall formation *F.prau*

From big data study to the determination of causality at scale



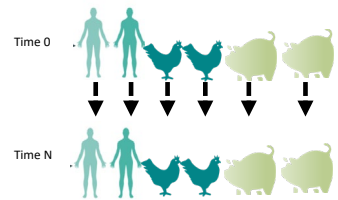
➤ Microbiome Workflows at INRAE Micalis institute

Contact:
Lionel Rigotier-Gois,
Emmanuelle Maguin
Moez Rhimi



Specificities:
Anaerobic microorganisms
Single cell dispensing
Oriented culturomics by genomics

Cohorts of individuals



Microbiota
samples
(eubiosis)



Sample
preparation
(serial dilution)

B-SIGHT 2nd



Single cell dispensing
(microfluidic combined
with microscopy &
fluorescence)



Genomic-guided media
(solid & liquid)

Incubator



Incubation &
growth monitoring

Characterization of isolates



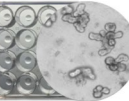
MALDI Biotyper®



Dereplication,
Species identification,
Strain typing
(MALDI, 16S rRNA & gyrB
& genome sequencing)



Gut on chip
Core
Facility



Phenotyping
(Host-microbe
interactions
In vitro, consortia,
Gut on chip,
ex vivo & *in vivo*)

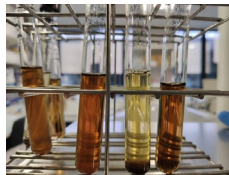
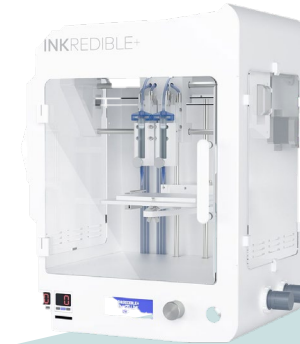


INRAE

Human Gut Microbiome Data at Scale - IA School

2025.10.01 / Julien Tap / fme.micalis.fr

➤ Towards Phenotyping & Consortia design



Hungate tubes

Droplet screening

Microfermenters

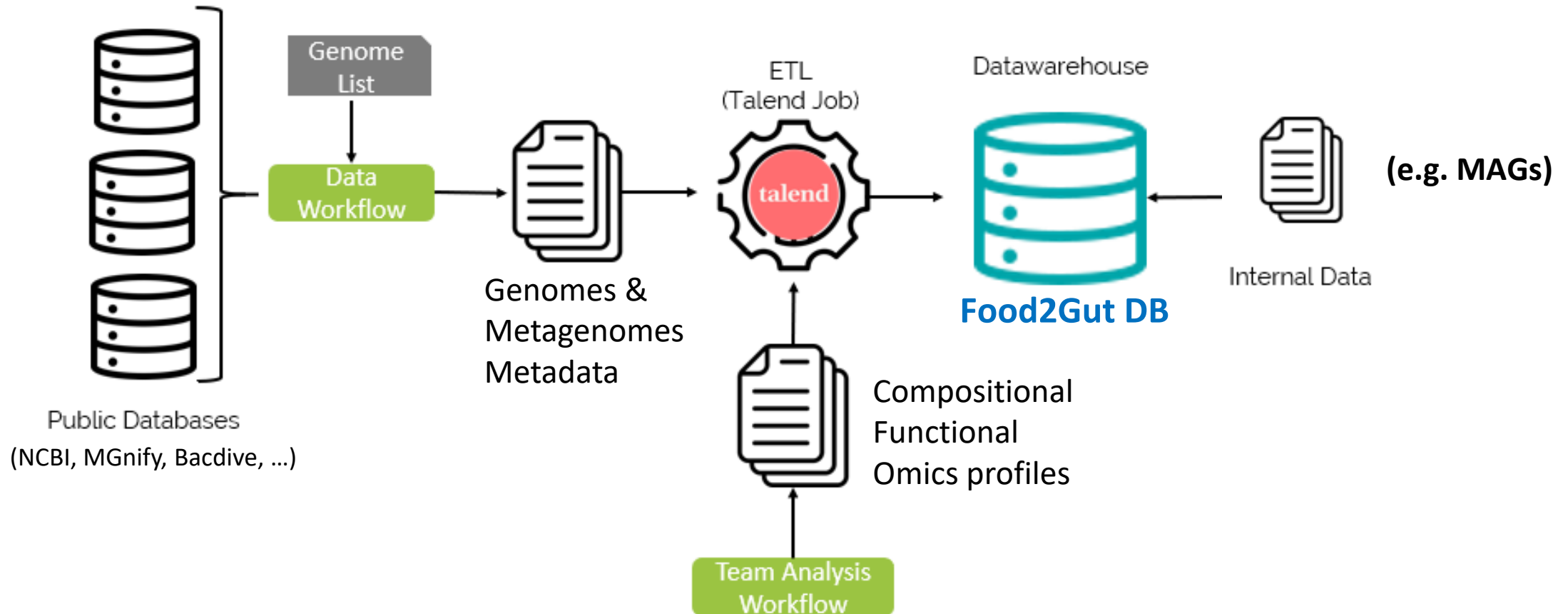
Fermenters

3D bioprinting

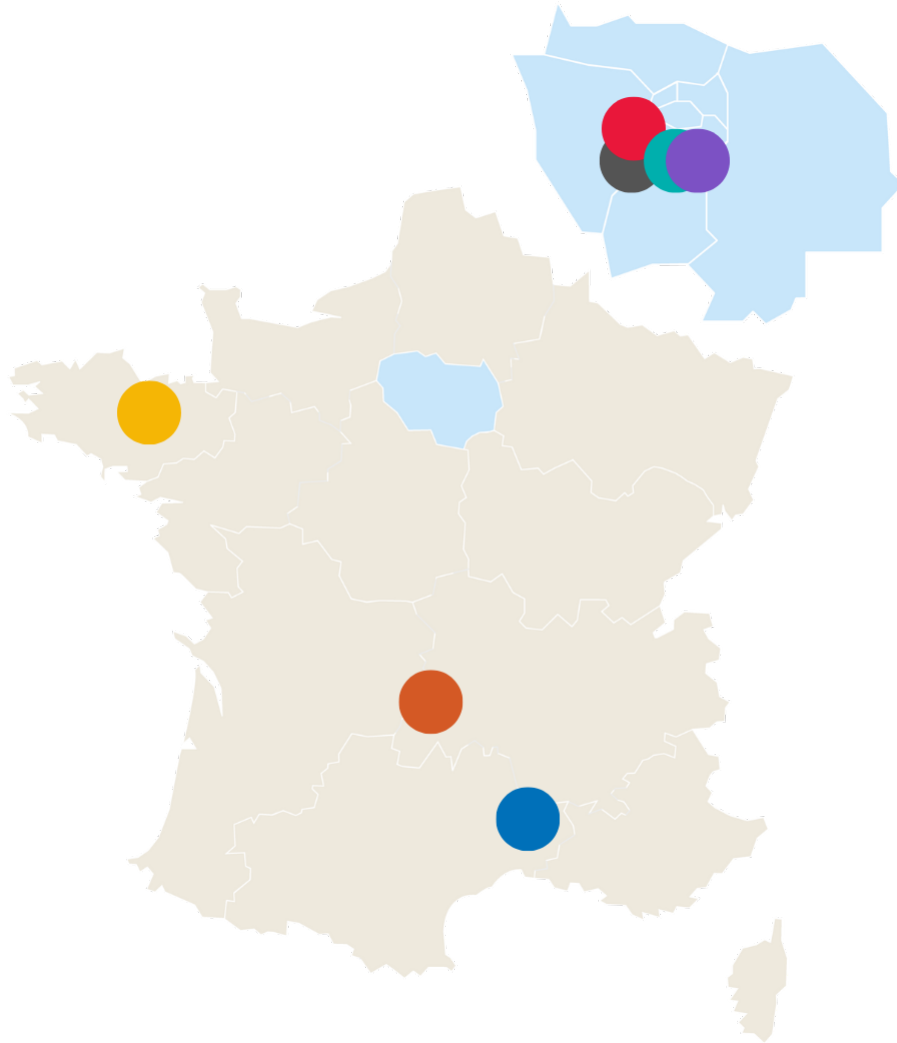
Contact:
Romain Briandet
Stéphane Chaillou



Integration of multi-omics data with metadata



National distributed platform set-up



Innovative processes development

Screening platform

- › Selection of ferments & growth conditions
- › Metabolic properties
- › Directed evolution

Fermentation platform

- › Bioreactors (liquid, solid); 1 to 10 L
- › Production of ferments & fermented foods
- › DSP; recovery of co-products



Analytical platform

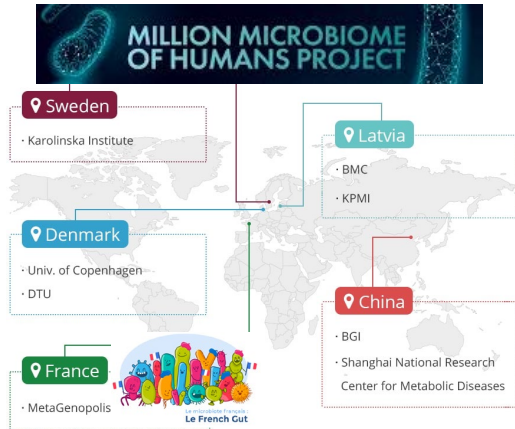
- › Fermentation monitoring; raw material compositions
- › Nutritional analysis & contaminants
- › Physico-chemical properties

Food workshop

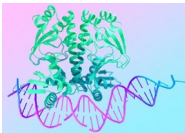
Pre-treatment of raw materials
Recipes and prototyping
Conservation

Microbiome science trends: From large data to precision-based models

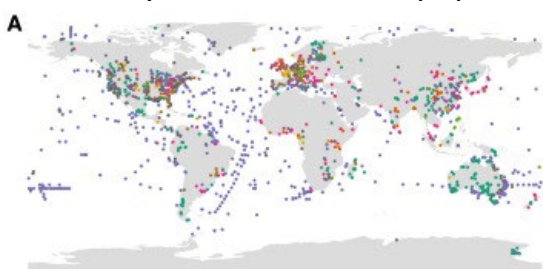
Worldwide Population scale studies



Machine learning
Deep learning
Generative AI



Discovery of antimicrobial peptides



INRAE

Human Gut Microbiome Data at Scale – IA School

2025.10.01 / Julien Tap / fme.micalis.fr

High throughput microbiology



Droplet screening

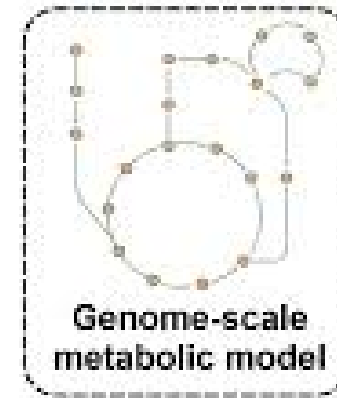


culturomics

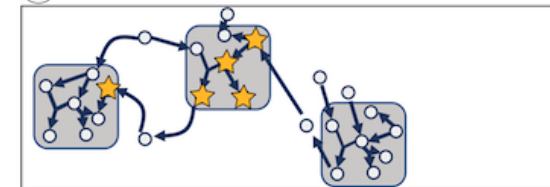


Multi-agent
Federating AI
(Mathieu Almeida)

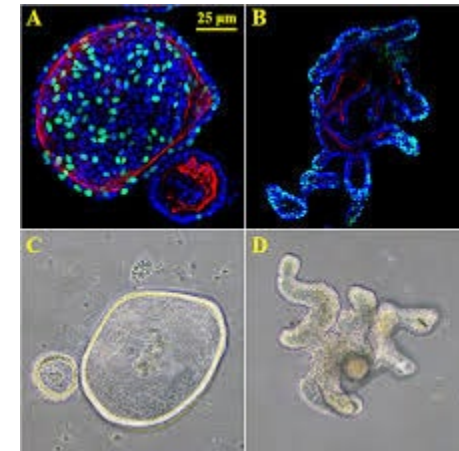
Metabolic models



③ Metabolic cooperation potential added value



Ex-vivo models



> Advanced “real” AI for omics

Techniques and applications

- LLLM for genome assembly binning, taxonomic identification and functional annotation
- ML/DL to predict the microbiote composition
- ML to capture the compositionality of microbiomes

-
- Variant discovery and strain-level identification

-
- High throughput 3D structure prediction of proteins and complexes

-
- Predict metabolic profiles from microbe abundance

Examples

- BERTax, VAMB
- cNode
- DeepCoda, Coda4Microbiome

-
- DeepVariant

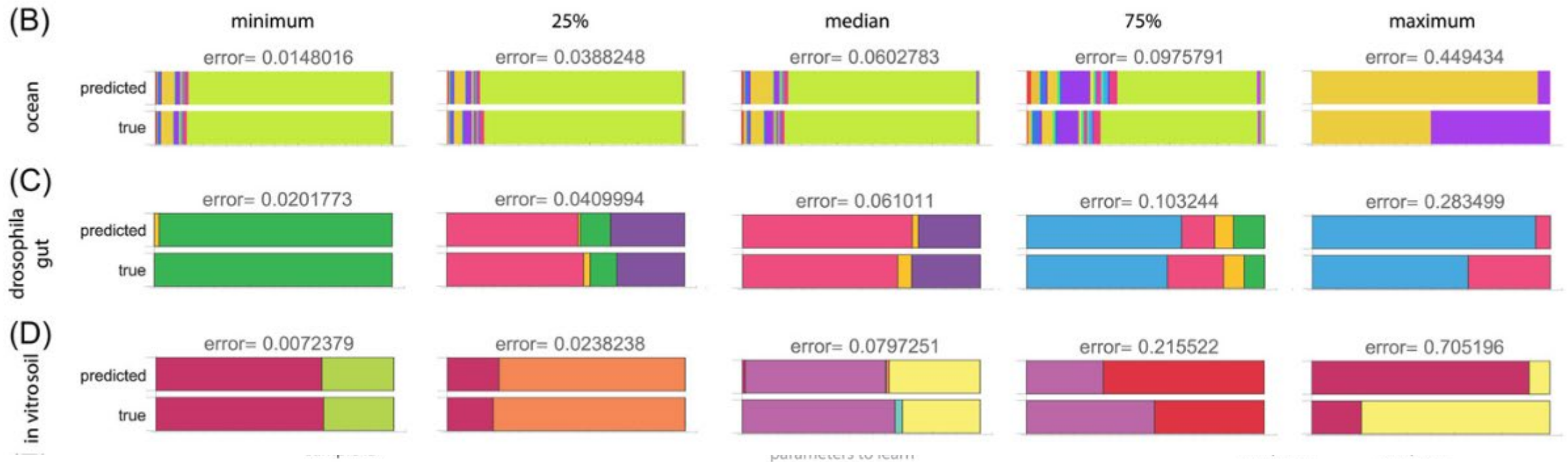
-
- [Alpha|Rosetta|ESM]Fold
 - AlphaMissense, ProGen

-
- mmvec / McMLP



cNode (Michel-Mata et al. 2022)

A Deep Learning framework to predict microbiome compositions from species assemblages



Take home messages- large data microbiome versus “AI”

Anchor analyses in ecology:

Always connect your analysis to a clear ecologically relevant research question, ensuring that findings are interpretable beyond purely statistical associations.

Integrate ecological modeling:

Use ecological theories and models (e.g., succession, fluxes, network dynamics) to guide interpretation and avoid black-box conclusions.

Diversify analytical pipelines:

When possible, cross-validate results using multiple approaches (statistics, bioinformatics pipelines, AI/machine learning, mechanistic models) to reduce false discoveries and enhance robustness.

Interact deeply with your data:

Explore, visualize, and critically question both raw data and method outputs: do not let the algorithm dictate your interpretation.

Prioritize effect size over significance:

While p-values provide a basic filter, effect sizes and confidence intervals better reflect the biological relevance of findings.

Learn across disciplines:

Use and adapt advanced statistical and modeling methods from other fields (climate science, economics, physics) that already tackle complexity, fluxes, and high-dimensionality.



Acknowledgments

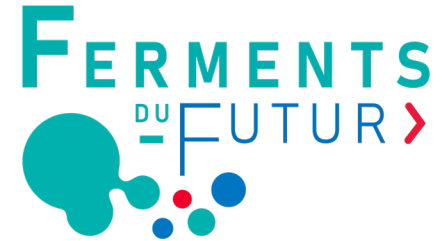


UC San Diego

Knight lab



Joel Doré et al.



Damien Paineau et al.



Manichanh lab



Boris Le Neve
Muriel Derrien



INRAE

Human Gut Microbiome Data at Scale – IA School
2025.10.01 / Julien Tap / fme.micalis.fr