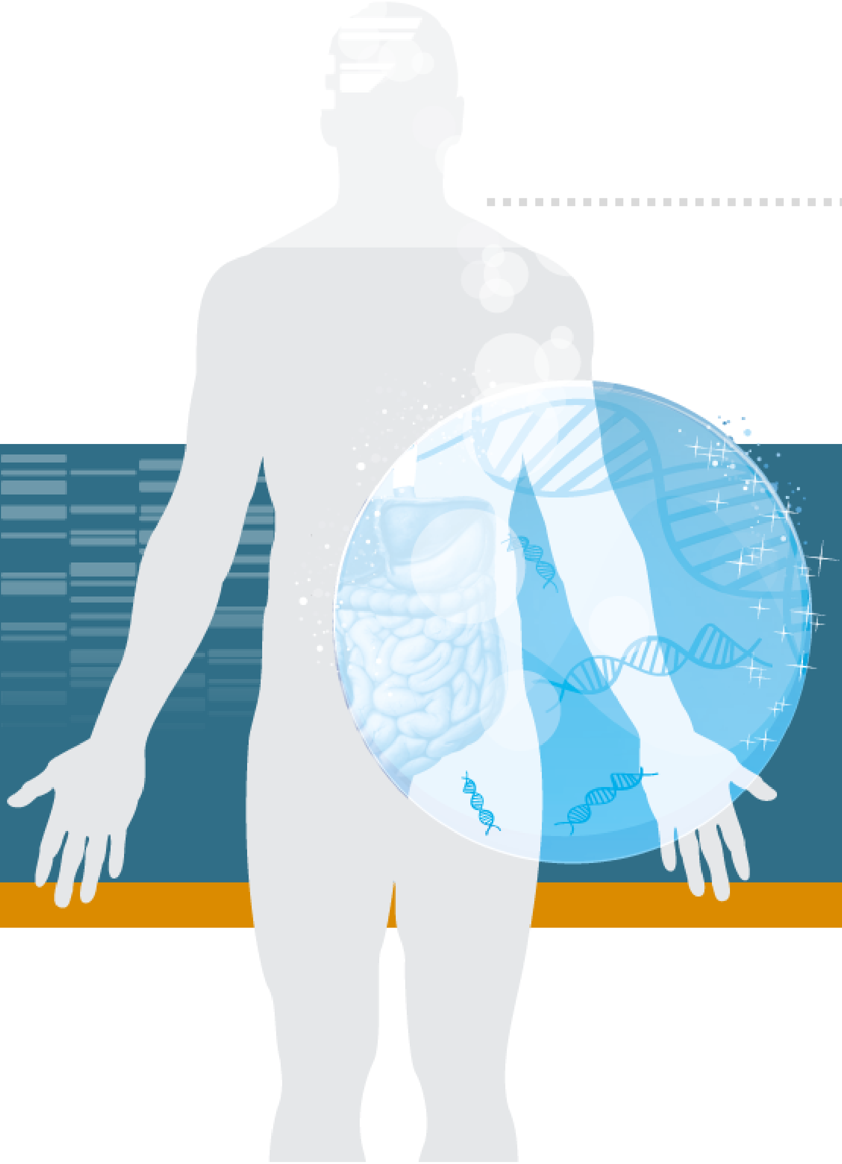


Network inference and analysis for gut microbiota exploration

Hanna Julienne / MetaGenoPolis / Team InfoBioStat

People differ by gut bacterial gene counts

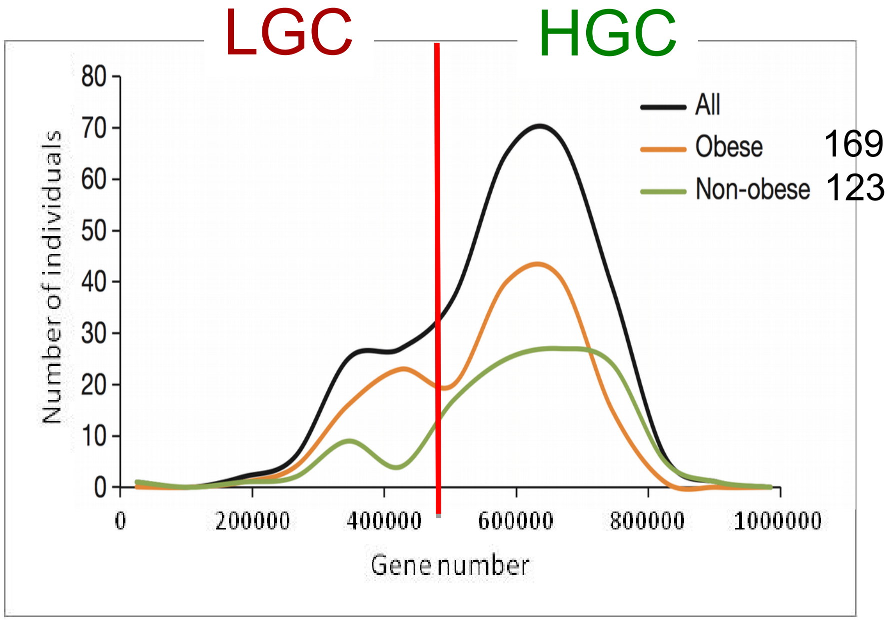
metaHIT obesity cohort (292 Danes)

Slide kindly provided by
Emmanuelle Le Chatelier

genes count

Reference catalogue

	Individuals						
item	**Ind 1**	**Ind 2**	**Ind 3**	**Ind 4**	**Ind 5**	**Ind 6**	**Ind 7**
1	0	36	2	0	43	106	1250
2	0	27	193	0	44	103	8
3	0	31	0	0	0	0	0
4	152	59	282	1	0	0	0
5	115	0	0	1	0	29	2
6	90	783	26	0	2	0	0
7	104	1616	0	0	0	0	5
8	0	82	0	0	0	0	0
9	2	0	0	0	0	0	0
10	23	239	1302	10	0	190	0
11	30	183	900	13	0	172	0
12	27	228	1120	6	0	324	0
13	103	0	0	0	0	0	0
14	0	30	269	0	0	0	0
15	0	0	0	0	0	95	0
16	1250	6002	468	607	492	141	8023
17	0	0	0	0	0	0	0
18	0	9	108	0	0	55	0
19	0	0	0	3	0	0	0
9900000 **3300000**	0	36	2	0	43	106	1250



(3.3M gene catalogue)

¼ LGC, enriched in obese

ARTICLE

doi:10.1038/nature12506

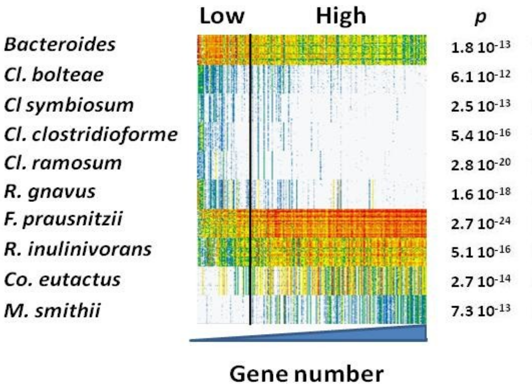
Richness of human gut microbiome correlates with metabolic markers

Le Chatelier et al, Nature 2013

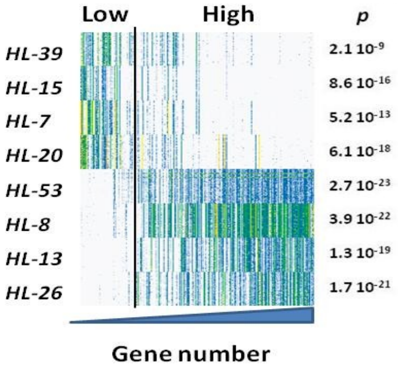
People differ by gut bacterial gene counts and bacterial communities

Slide kindly provided by
Emmanuelle Le Chatelier

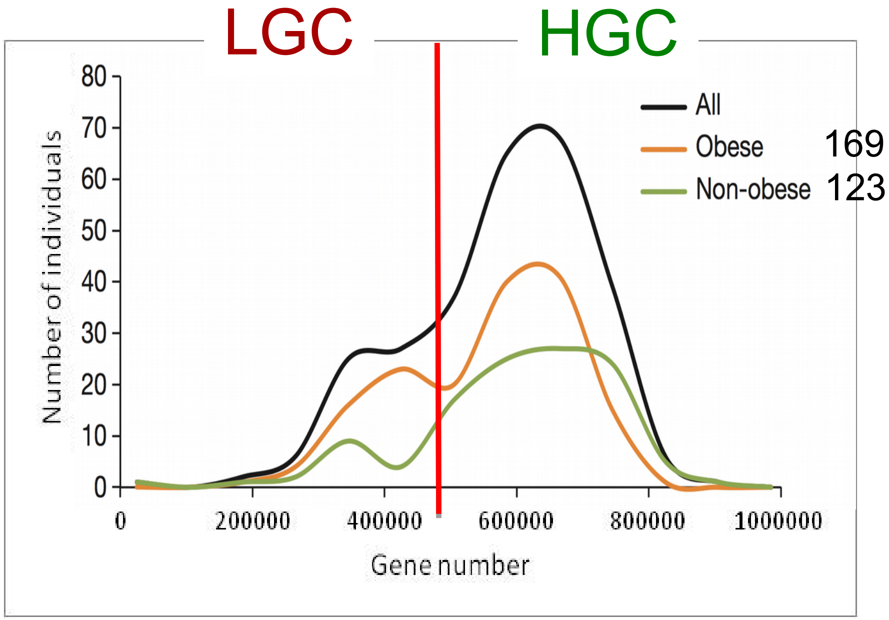
58 richness MGS



Known species n=10



Unknown MGS n=48



(3.3M gene catalogue)

¼ LGC, enriched in obese

ARTICLE

doi:10.1038/nature12506

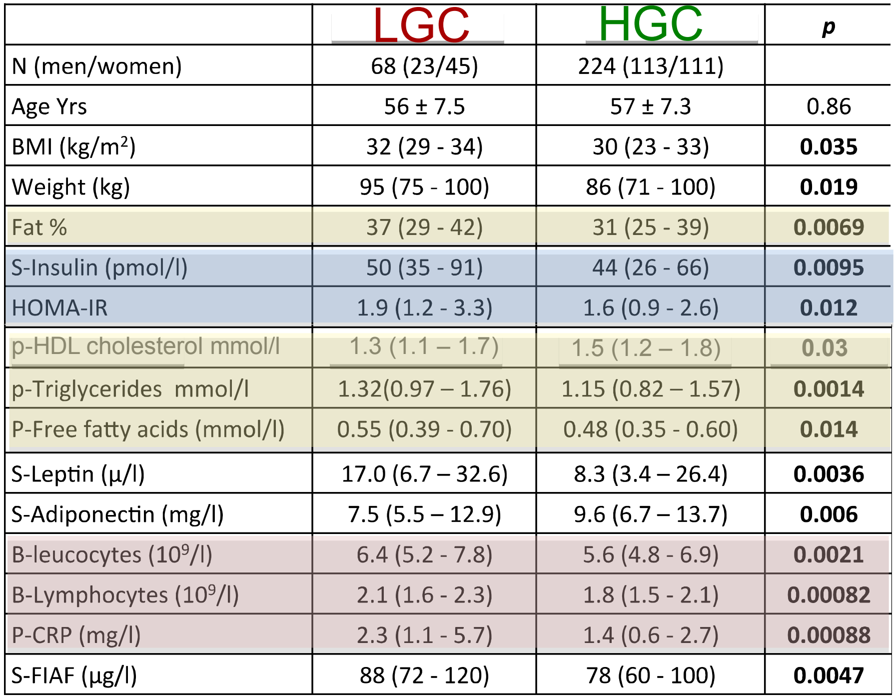
Richness of human gut microbiome correlates with metabolic markers

Le Chatelier et al, Nature 2013

low richness: a risk phenotype ?

low diversity (LGC) individuals have less healthy metabolic & inflammatory traits

increased adiposity, dyslipidemia, insulin resistance, inflammation,
predispose for type 2 diabetes, cardio-vascular disease, cancer ...



	LGC	HGC	*p*
N (men/women)	68 (23/45)	224 (113/111)	
Age Yrs	56 ± 7.5	57 ± 7.3	0.86
BMI (kg/m^2)	32 (29 - 34)	30 (23 - 33)	**0.035**
Weight (kg)	95 (75 - 100)	86 (71 - 100)	**0.019**
Fat %	37 (29 - 42)	31 (25 - 39)	**0.0069**
S-Insulin (pmol/l)	50 (35 - 91)	44 (26 - 66)	**0.0095**
HOMA-IR	1.9 (1.2 - 3.3)	1.6 (0.9 - 2.6)	**0.012**
p-HDL cholesterol mmol/l	1.3 (1.1 – 1.7)	1.5 (1.2 – 1.8)	**0.03**
p-Triglycerides mmol/l	1.32(0.97 – 1.76)	1.15 (0.82 – 1.57)	**0.0014**
P-Free fatty acids (mmol/l)	0.55 (0.39 - 0.70)	0.48 (0.35 - 0.60)	**0.014**
S-Leptin (µ/l)	17.0 (6.7 – 32.6)	8.3 (3.4 – 26.4)	**0.0036**
S-Adiponectin (mg/l)	7.5 (5.5 – 12.9)	9.6 (6.7 – 13.7)	**0.006**
B-leucocytes (10^9/l)	6.4 (5.2 - 7.8)	5.6 (4.8 - 6.9)	**0.0021**
B-Lymphocytes (10^9/l)	2.1 (1.6 - 2.3)	1.8 (1.5 - 2.1)	**0.00082**
P-CRP (mg/l)	2.3 (1.1 - 5.7)	1.4 (0.6 - 2.7)	**0.00088**
S-FIAF (µg/l)	88 (72 - 120)	78 (60 - 100)	**0.0047**

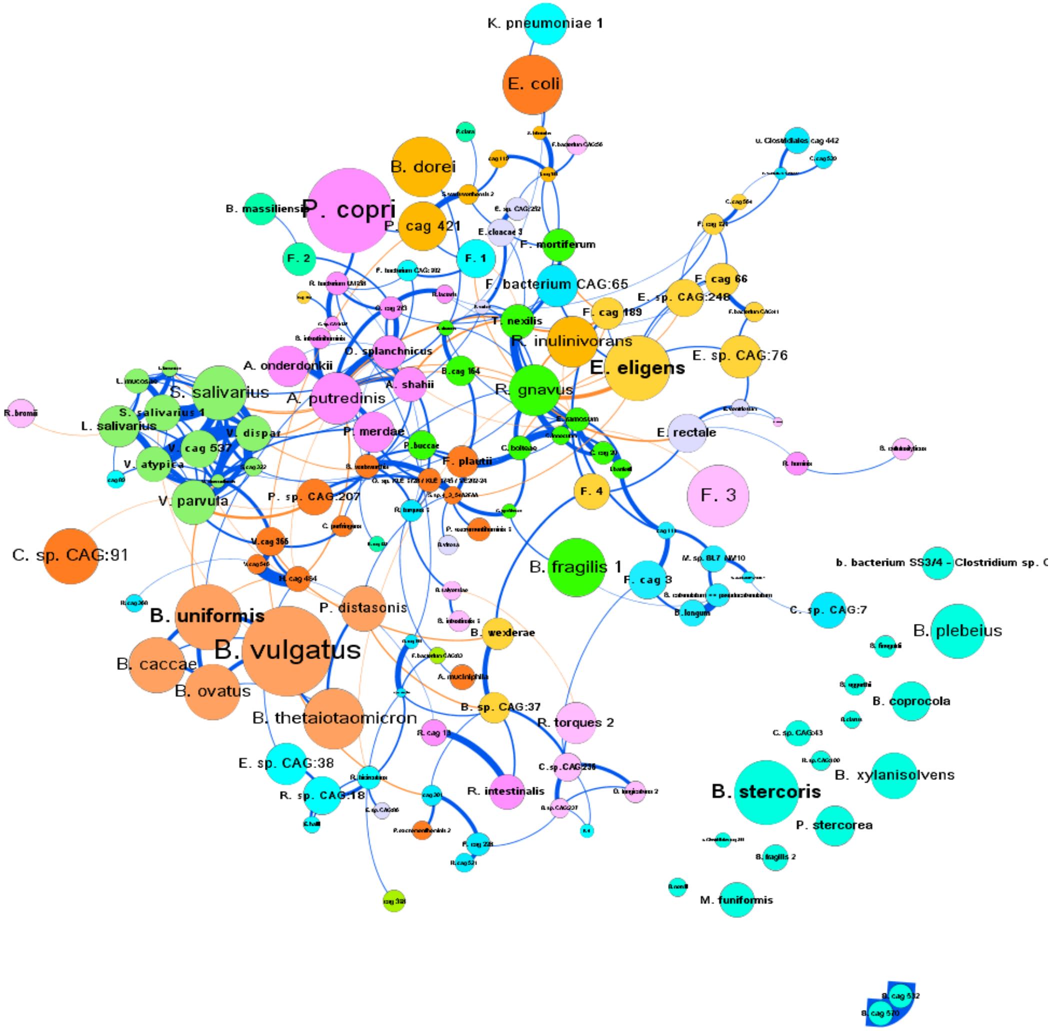
Le Chatelier et al, Nature 2013

Slide kindly provided by
Emmanuelle Le Chatelier

Complex network analysis for microbiota exploration

- Global indicator of microbiota status (health-dysbiosis): **richness √**
- Retrieve **differentially abundant species** between health and dysbiosis **√**
- **We would like to represent the modulation of the whole ecosystem between health and dysbiosis**
- A network is a natural object to represent interactions between MGS. Our goals could be:
  - **Detect antagonism**: resource competition or interference competition (direct harm by secretion of toxin)
  - **Detect co-occurrence**: True symbiosis (cross-feeding) or profiting from the same conditions
  - **Detect keystone species**: if the keystone species abundance varies the whole ecosystem changes

MGS Network representation



- Node = MGS
- **Undirected** graph
- **Weighted** graph
- Co-occurence edges in blue
- Antagonism edges in red

Chosen method

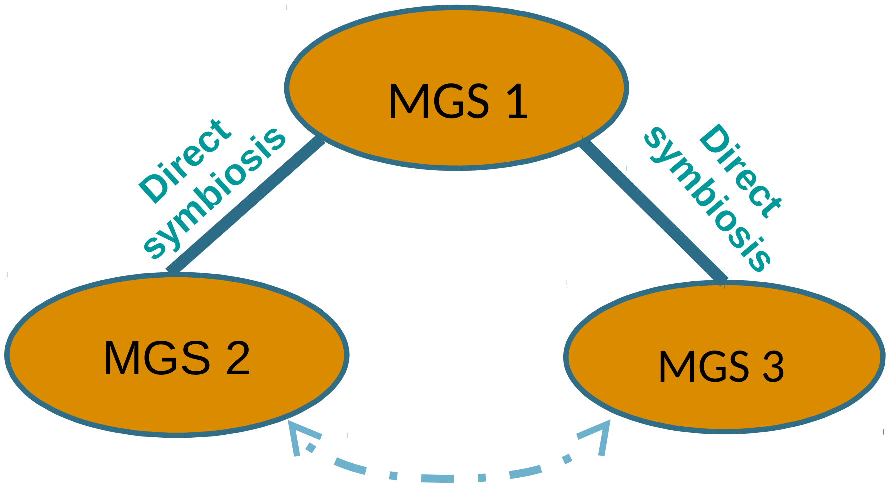
- We use a Glasso (Graphical Lasso) method to infer the network

Sparse and Compositionally Robust Inference of Microbial Ecological Networks

Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, Richard A. Bonneau

Published: May 7, 2015 • http://dx.doi.org/10.1371/journal.pcbi.1004226

- MGS abundances are represented by a gaussian markov random field
- The main advantage over correlation based methods is that indirect correlations are not inferred



Strong indirect correlation

Spiec Easi pipeline

- MGS abundances are log transformed
- Find the inverse covariance matrix that satisfy this criterion :

Likelihood Regularisation

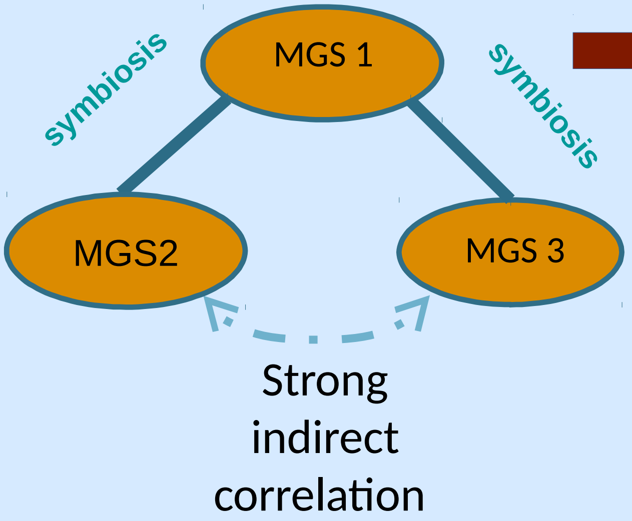
$$\widehat{\Omega}(\lambda) = \arg\min_{\Omega \succ 0} \left\{ -\ell(\Omega) + \lambda \|\Omega\|_1 \right\}$$

Fit the data at the closest

Sparsify
Avoid overfitting

- $\widehat{\Omega}(\lambda)$ Is also the **adjacency matrix** of the graph :

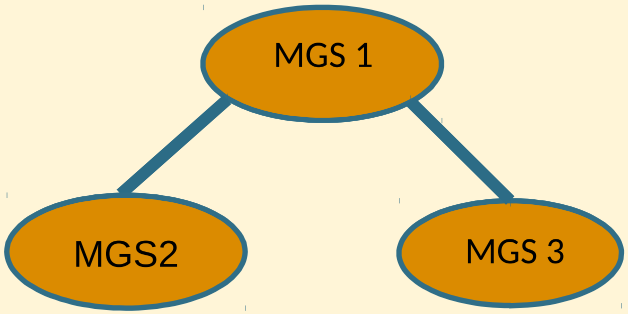
Reality :



Inferred Ω :

	MGS1	MGS2	MGS3
MGS1	.	0,5	0,5
MGS2	0,5	.	.
MGS3	0,5	.	.

Inferred graph :



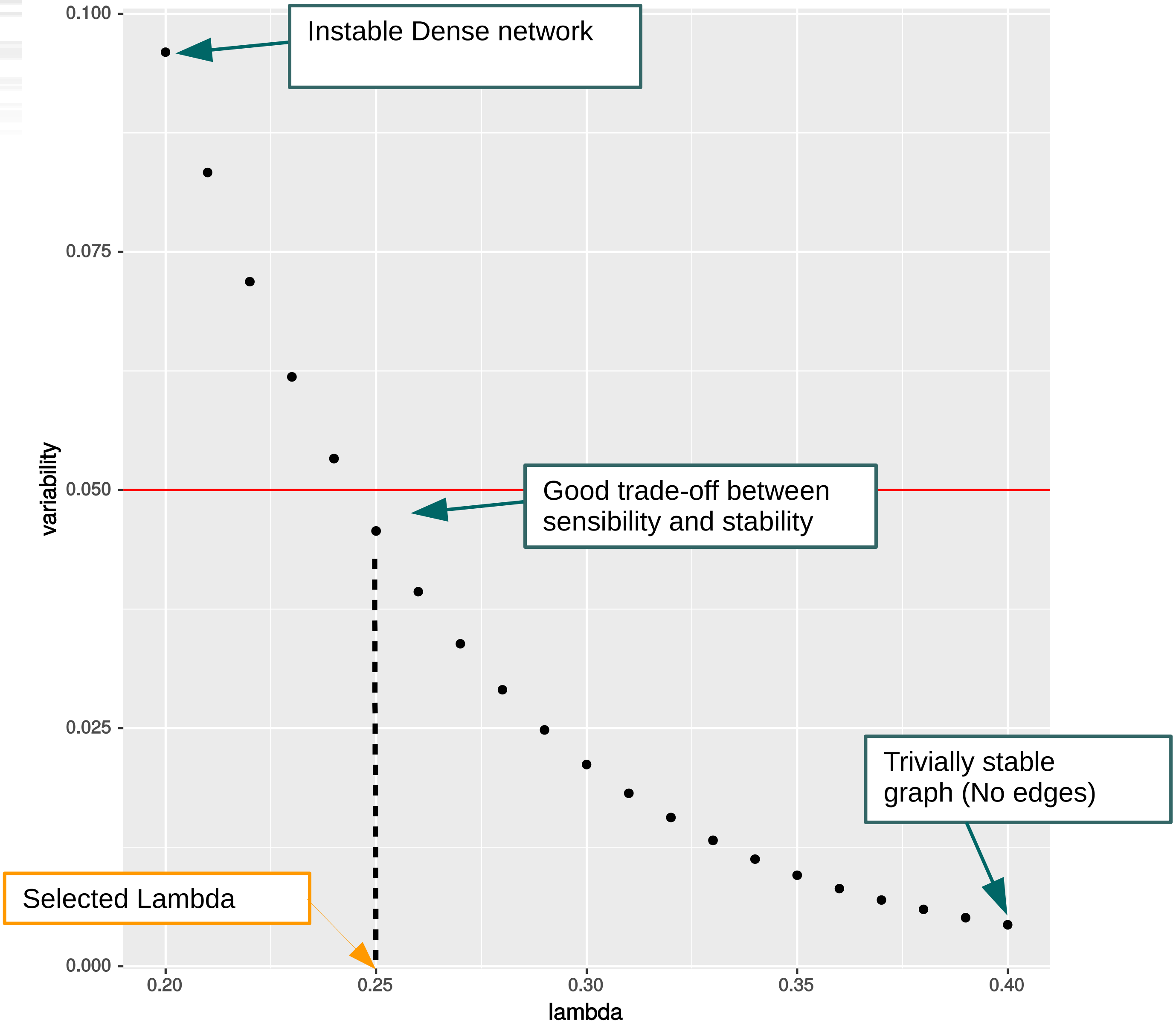
Select λ to guarantee stability of the edges

$$\widehat{\Omega}(\lambda) = \underset{\Omega \succ 0}{\arg\min} \left\{ \overset{\text{Likelihood}}{\boxed{-\ell(\Omega)}} + \overset{\text{Regularisation}}{\boxed{\lambda ||\Omega||_1}} \right\}$$

- StARS Stability Approach to Regularization Selection (Liu, Roeder, & Wasserman, 2010)
  - To select the best λ:
    - For a serie of λ, we draw N sub-samples representing 80 % of the cohort and we infer a network
    - We compute a stability score of the graph over sub-sampling for each λ
    - We select the lowest lambda (which gives the denser graph) that satify an arbitrary stability treshold

Variability of graph edges versus λ

The variability is the fraction of subsamples that disagree averaged over all edges



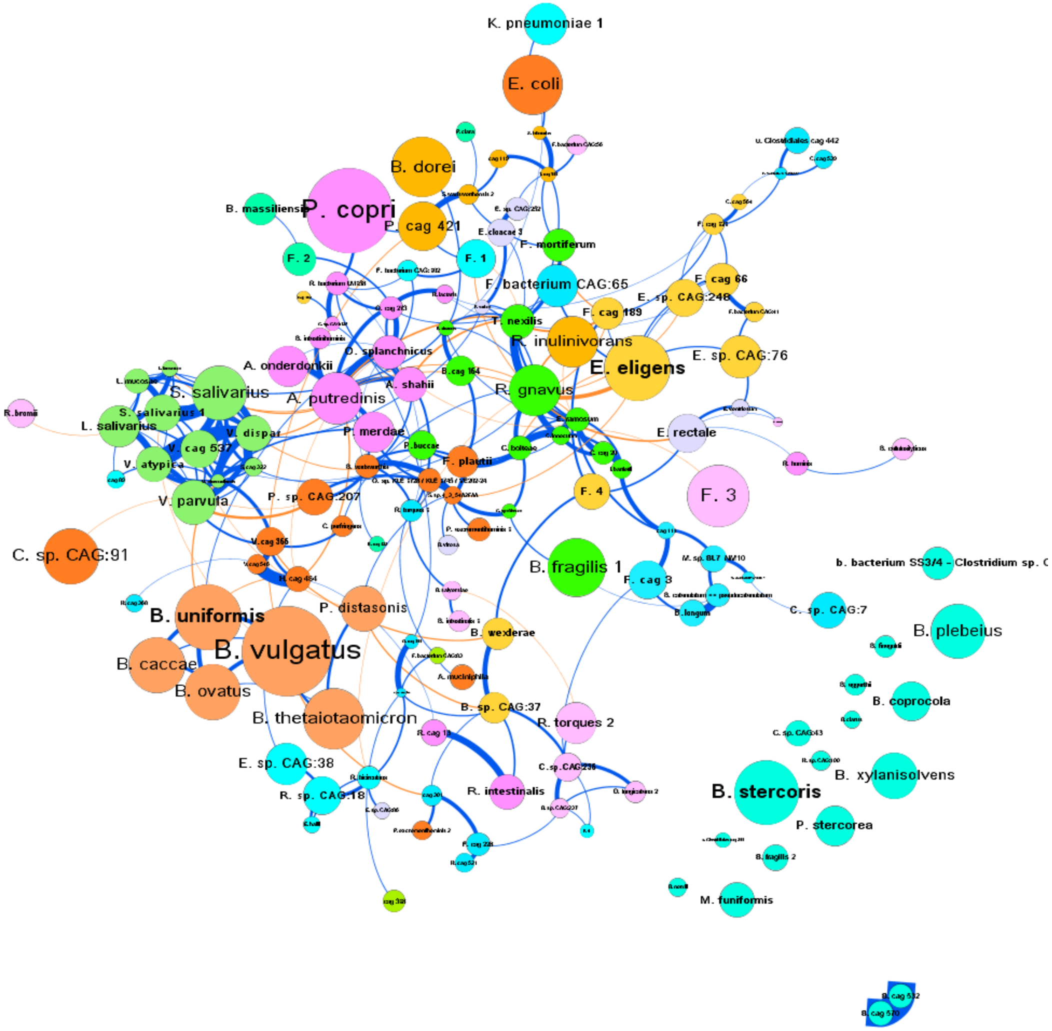
Instable Dense network

Good trade-off between sensibility and stability

Trivially stable graph (No edges)

Selected Lambda

MGS Network representation



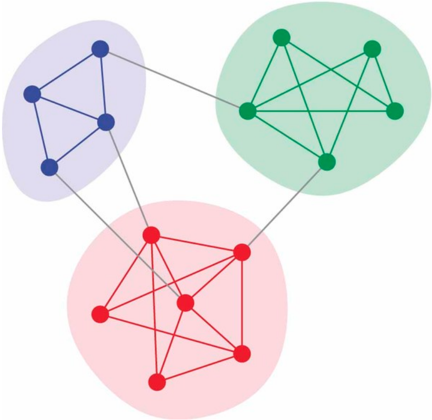
- Node = MGS
- **Undirected** graph
- **Weighted** graph
- Co-occurence edges in blue
- Antagonism edges in red

Interpretation of MGS Network

- A Network over of ~ 200 MGS species is a complex object and a challenge for interpretation, **possible strategies to decipher it are:**
  - Use the network to compute community of co-abundant MGS
  - Use centrality measures (PageRank, degree, betweenness centrality) to find keystone species
  - Focus on MGS of interest (differentially abundant, fonctional specificity)

Graph Community

- Recently, complex networks have become a reccurent object in many scientific fields :
  - Collaboration network
  - Ecology species network
  - Gene regulation network
- Globally those networks are relatively sparse but dense locally. Few nodes are more connected together than with the rest of the graph : These nodes form a **community**
- Given a partition of nodes in k groups, modularity assesses how well this partition correspond to k communities



Definition : fraction of intra community edges minus the expected fraction of intra community edges for a graph with the same community partition but with random edges between them

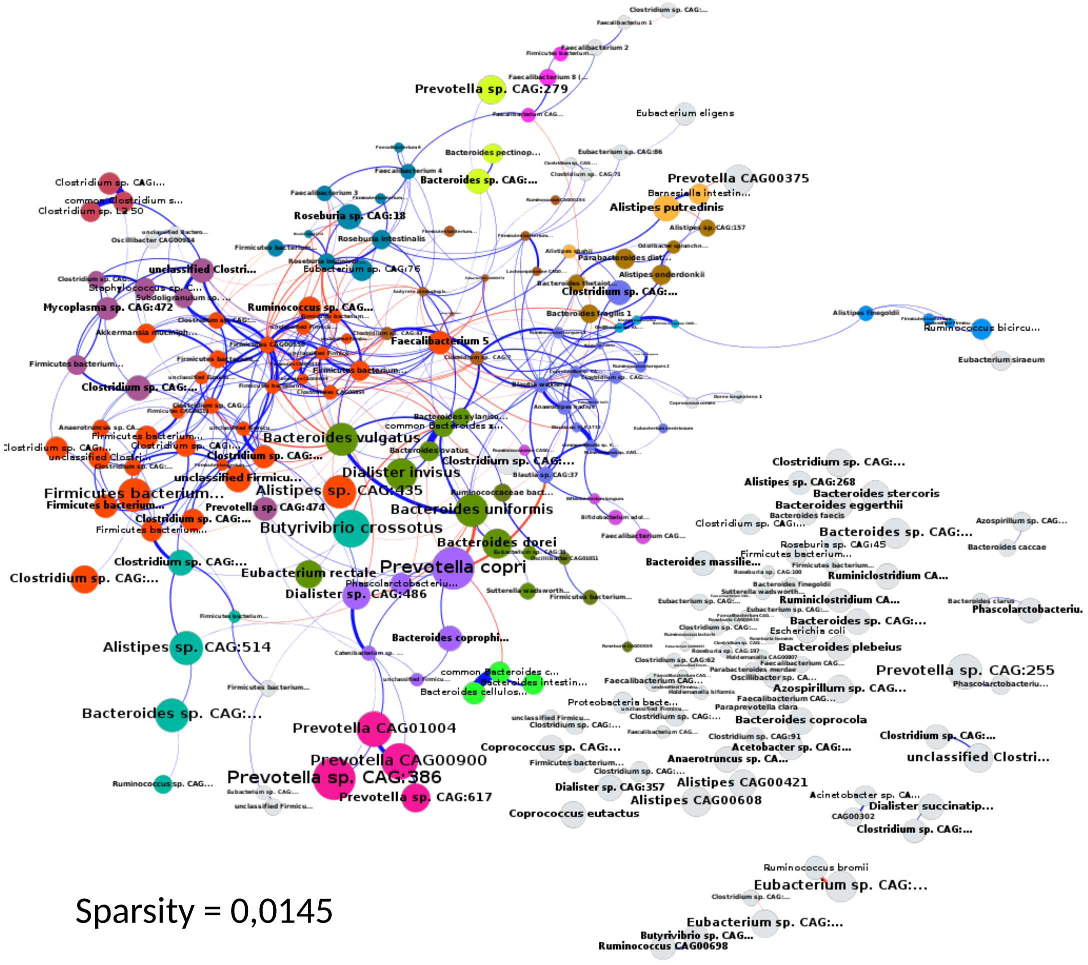
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i * k_j}{2m} \right] \delta(c_i, c_j)$$

MGS graph and community on the MetaHit cohort

- 233 species
- 155 species in the main connected component
- 133 in communities

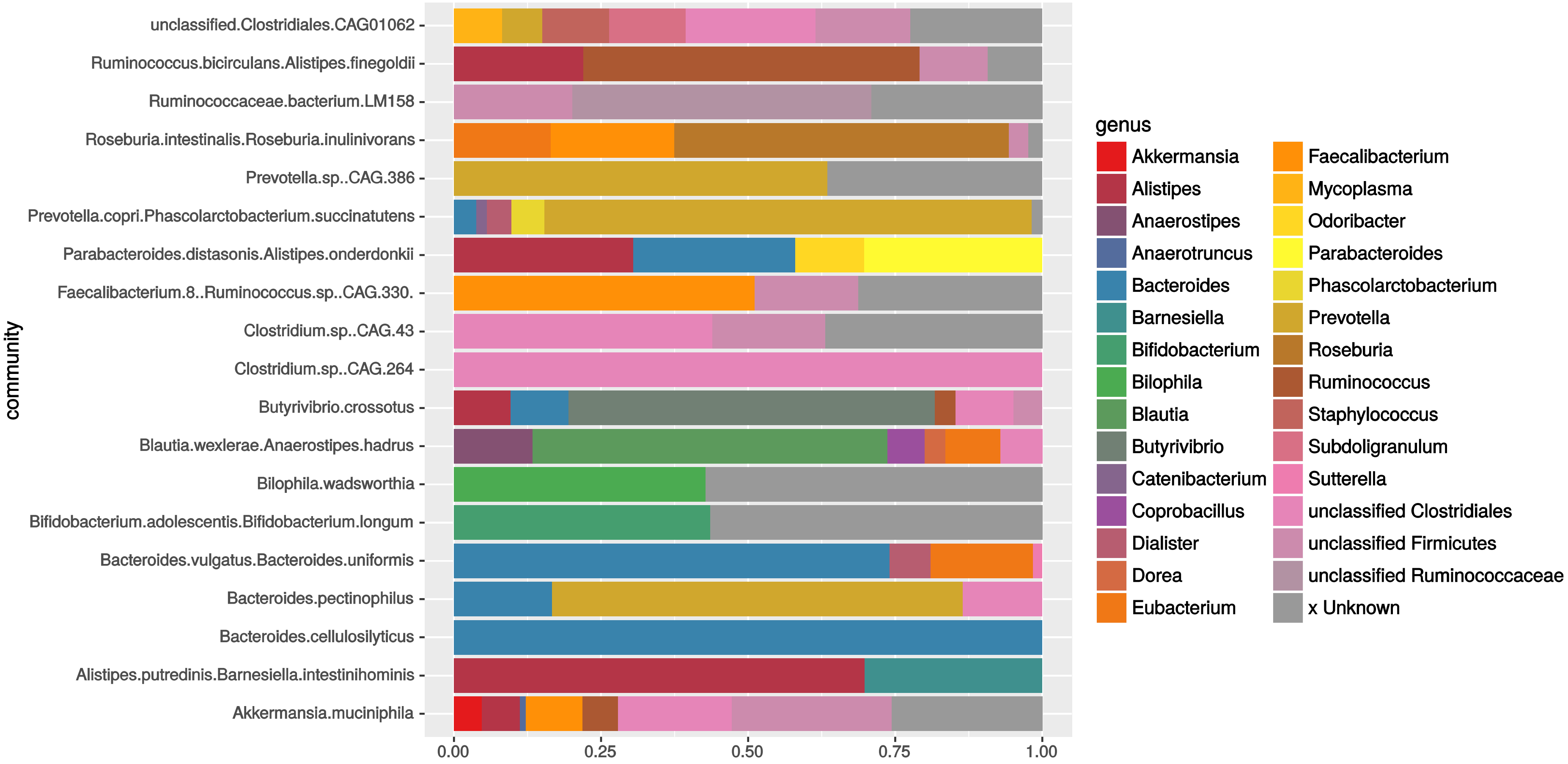
We achieve a fairly good modularity score with 19 communities (0.71, maximum 1).

Each color represents a community

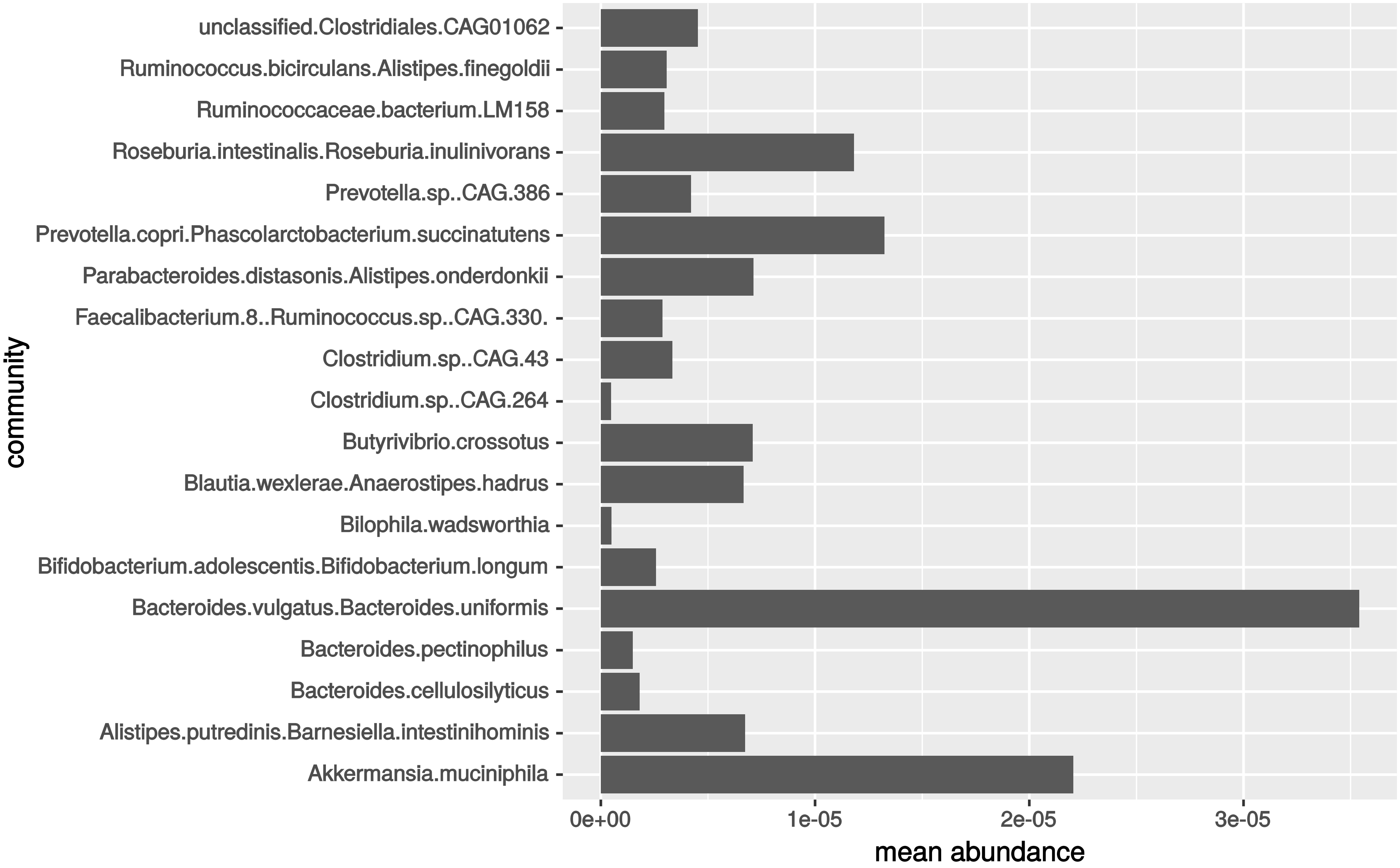


Sparsity = 0,0145

Relative abundance of genus in communities



Mean abundance (in all samples) of each community

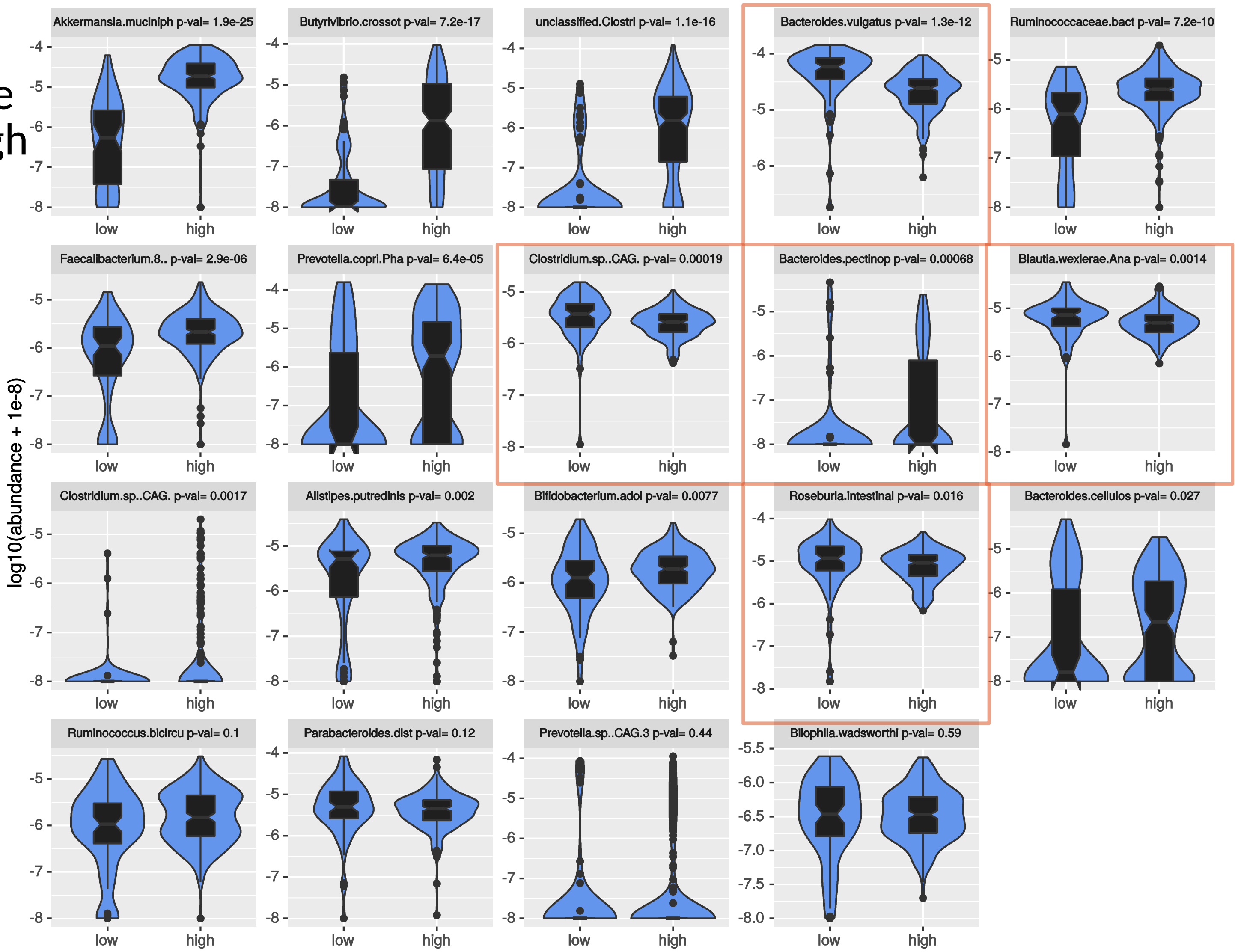


Communities are very relevant to understand the microbiota difference between low and high gene counts

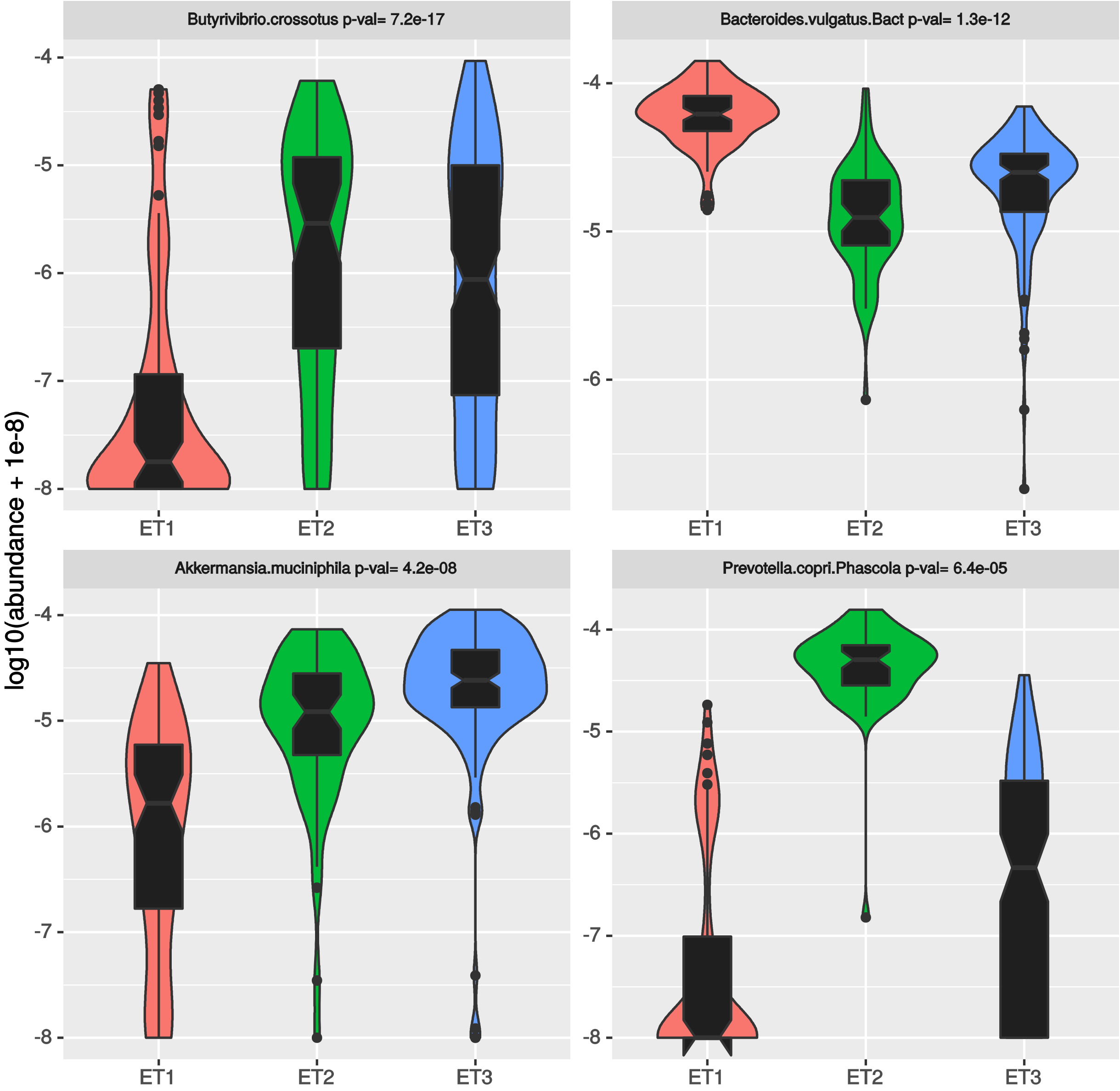
Community abundance is the sum of the abundance of species in the community

Low-gene count microbiota is depleted in most communities except :

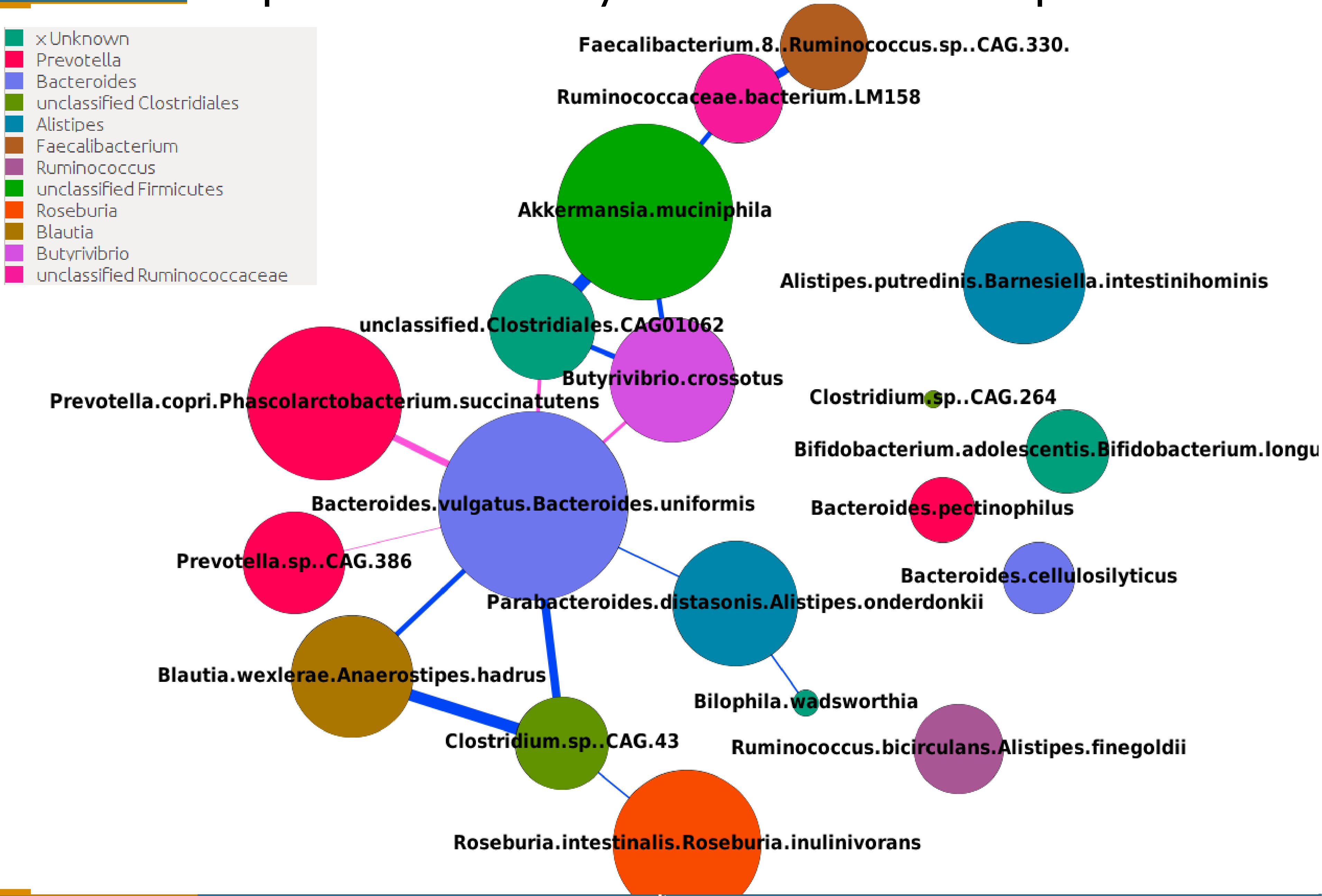
« Bacteroides vulgatus » community, « Blautia wexlerea » , « Roseburia intestinal », « Clostridium sp » community



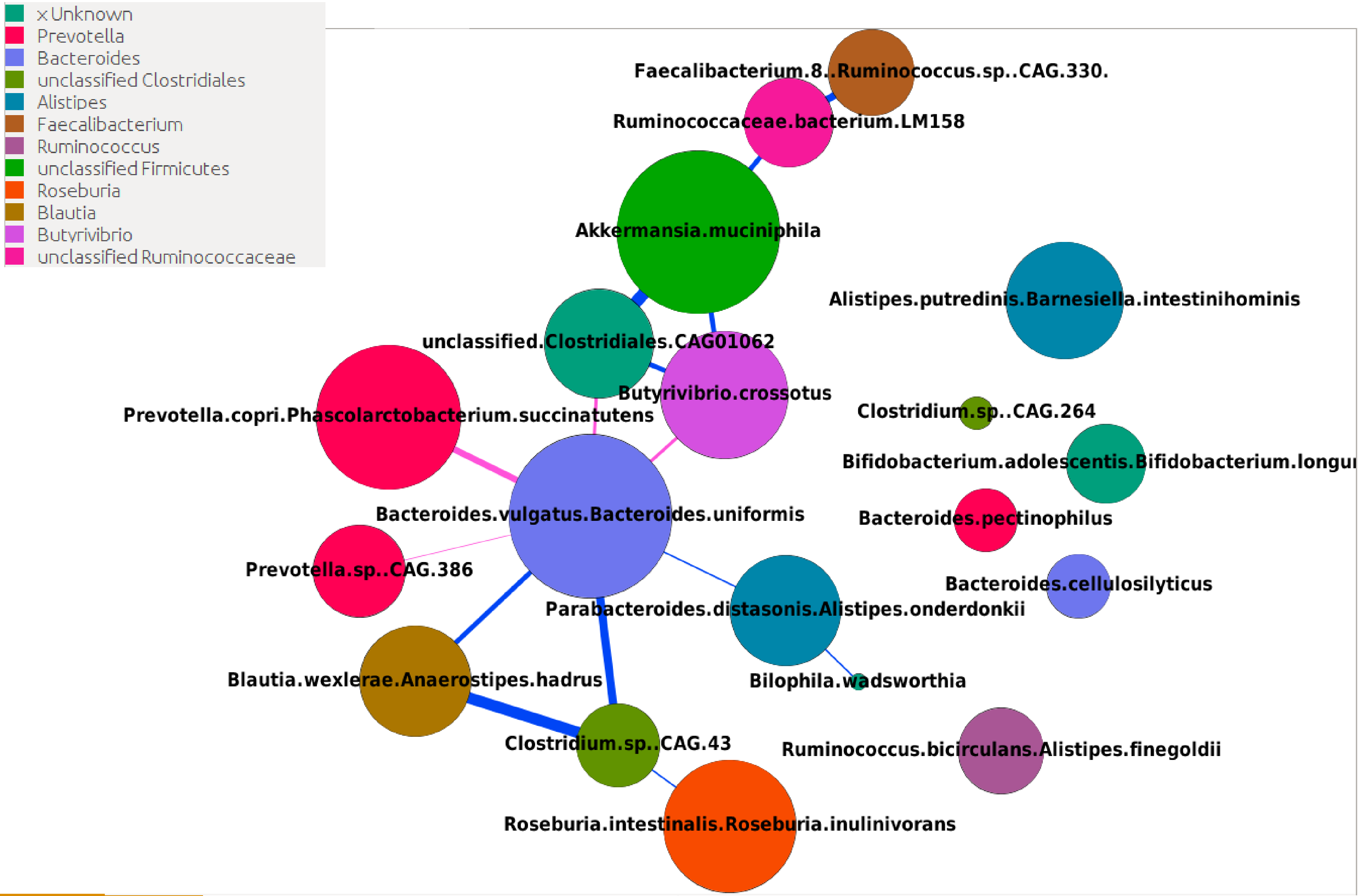
Enterotypes and Communities



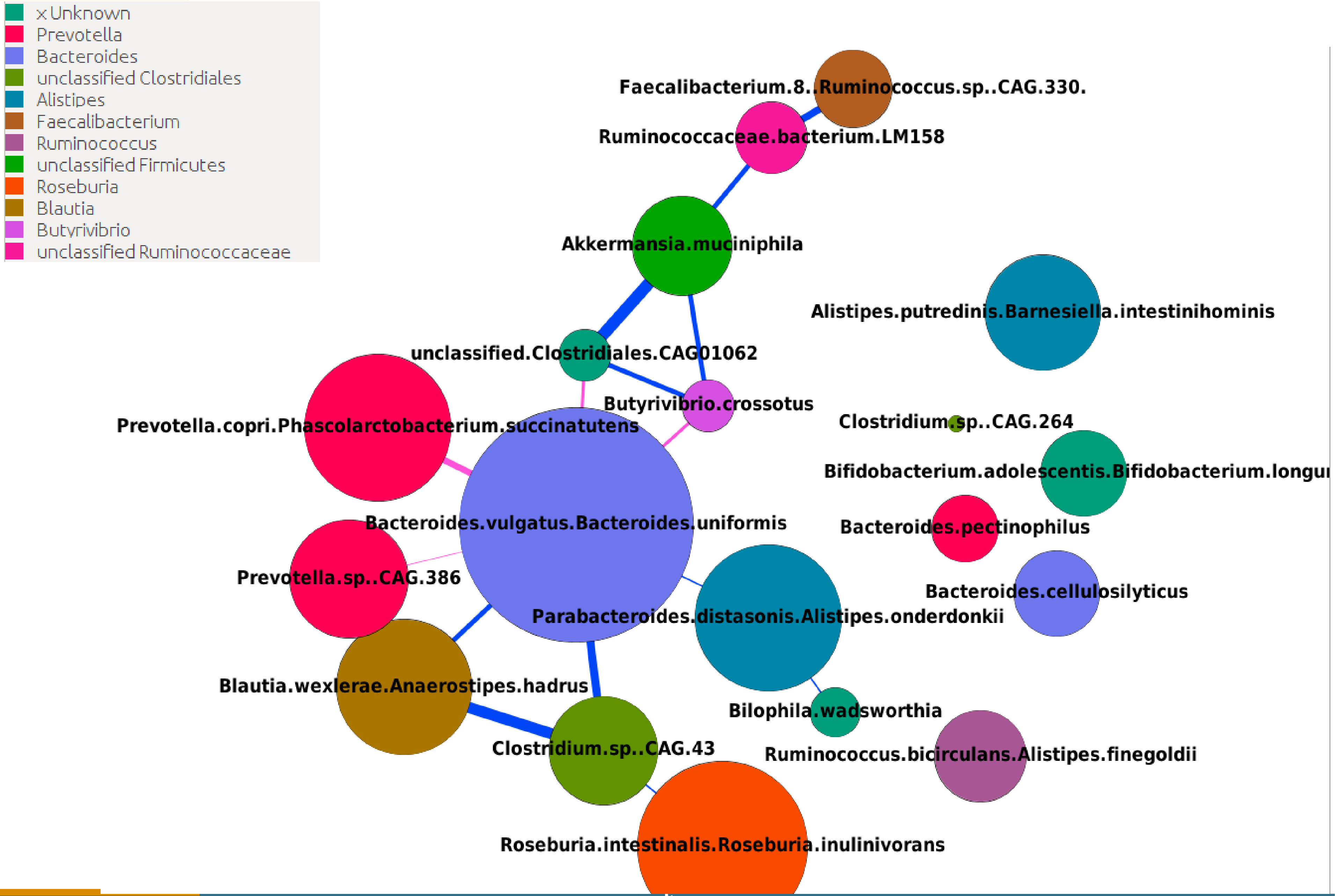
Graph of community abundance: all samples



Graph community: high gc

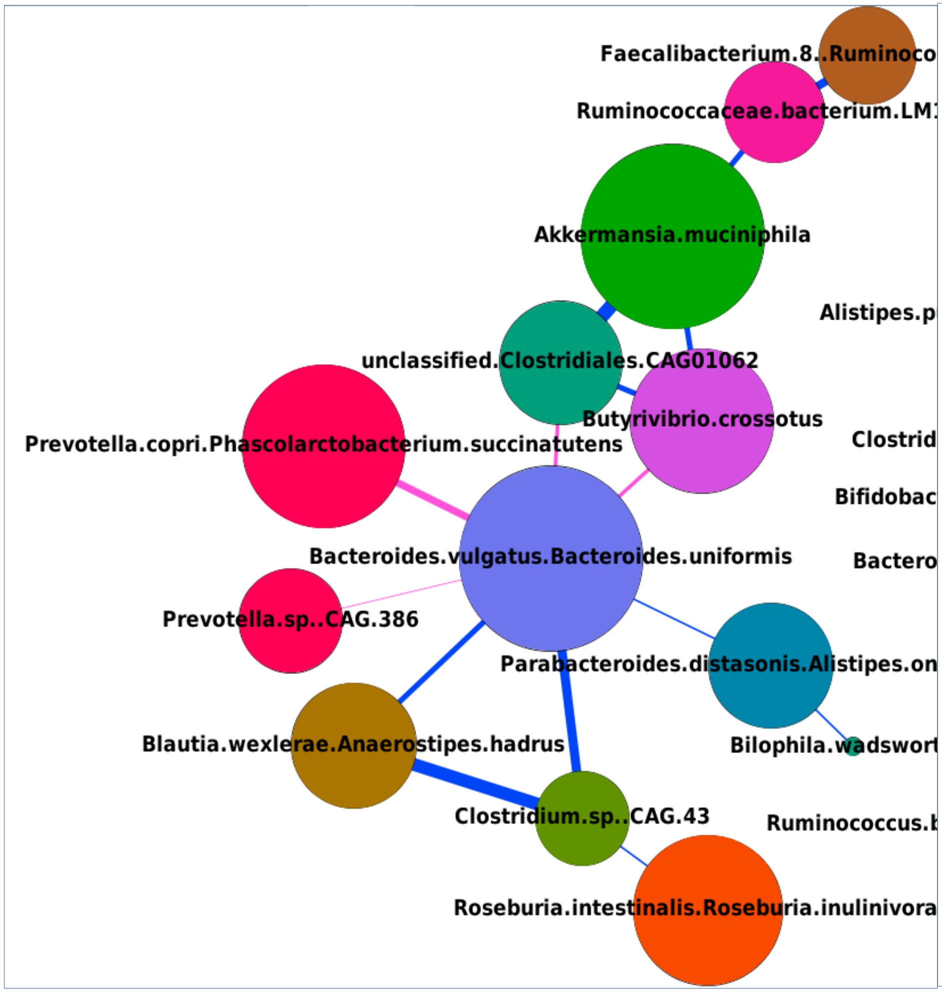


Graph of Community dynamics: low gc individuals

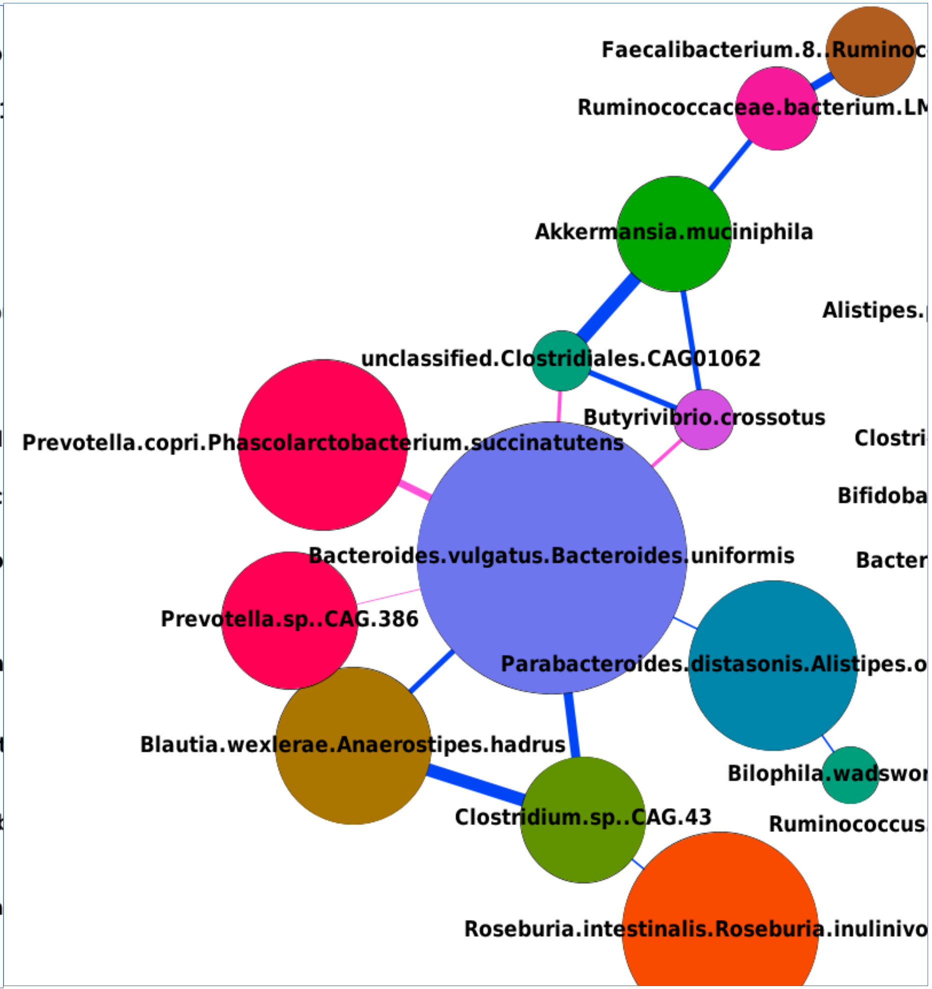


Dynamic of dysbiosis

High-GC graph :

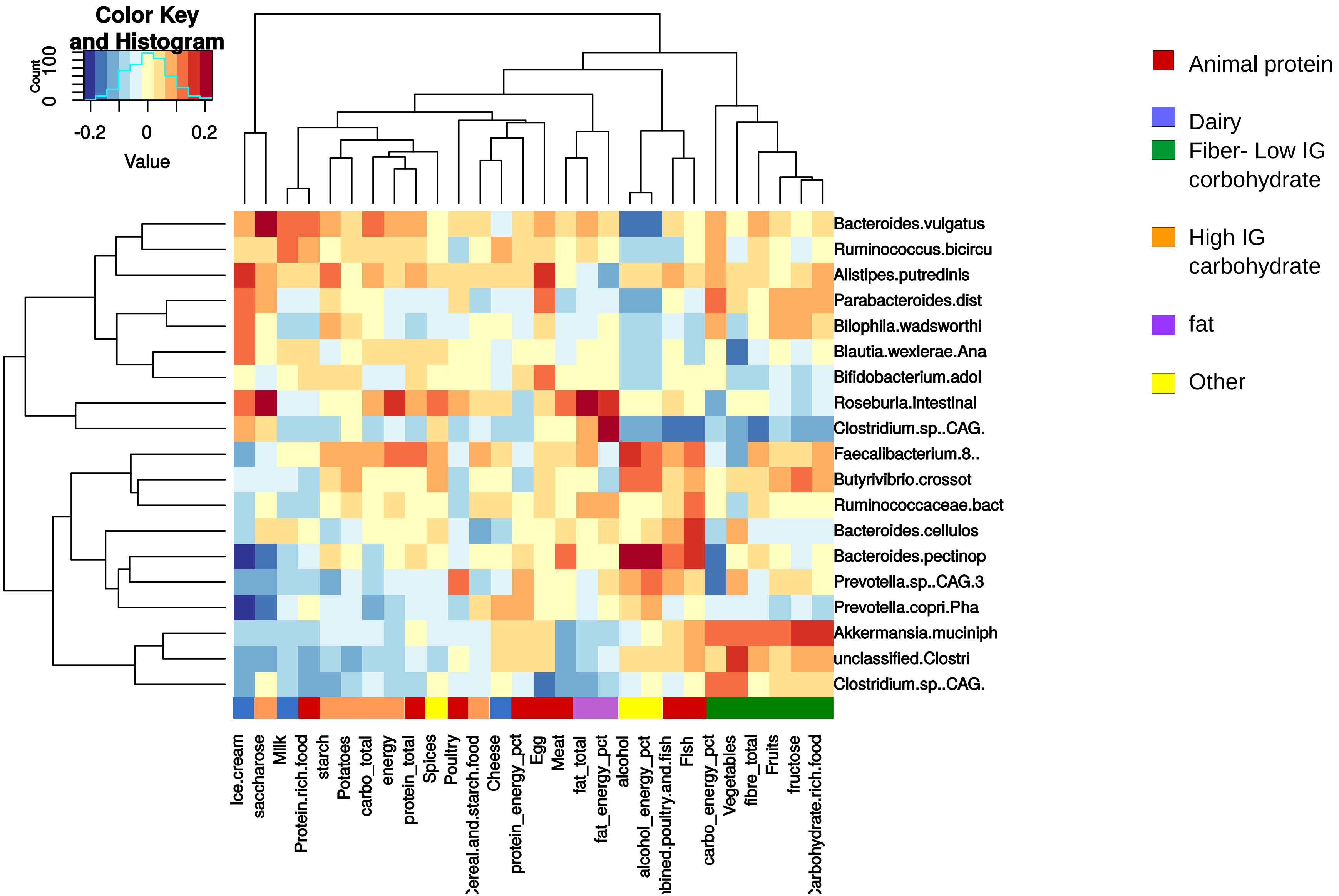


Low-GC graph :

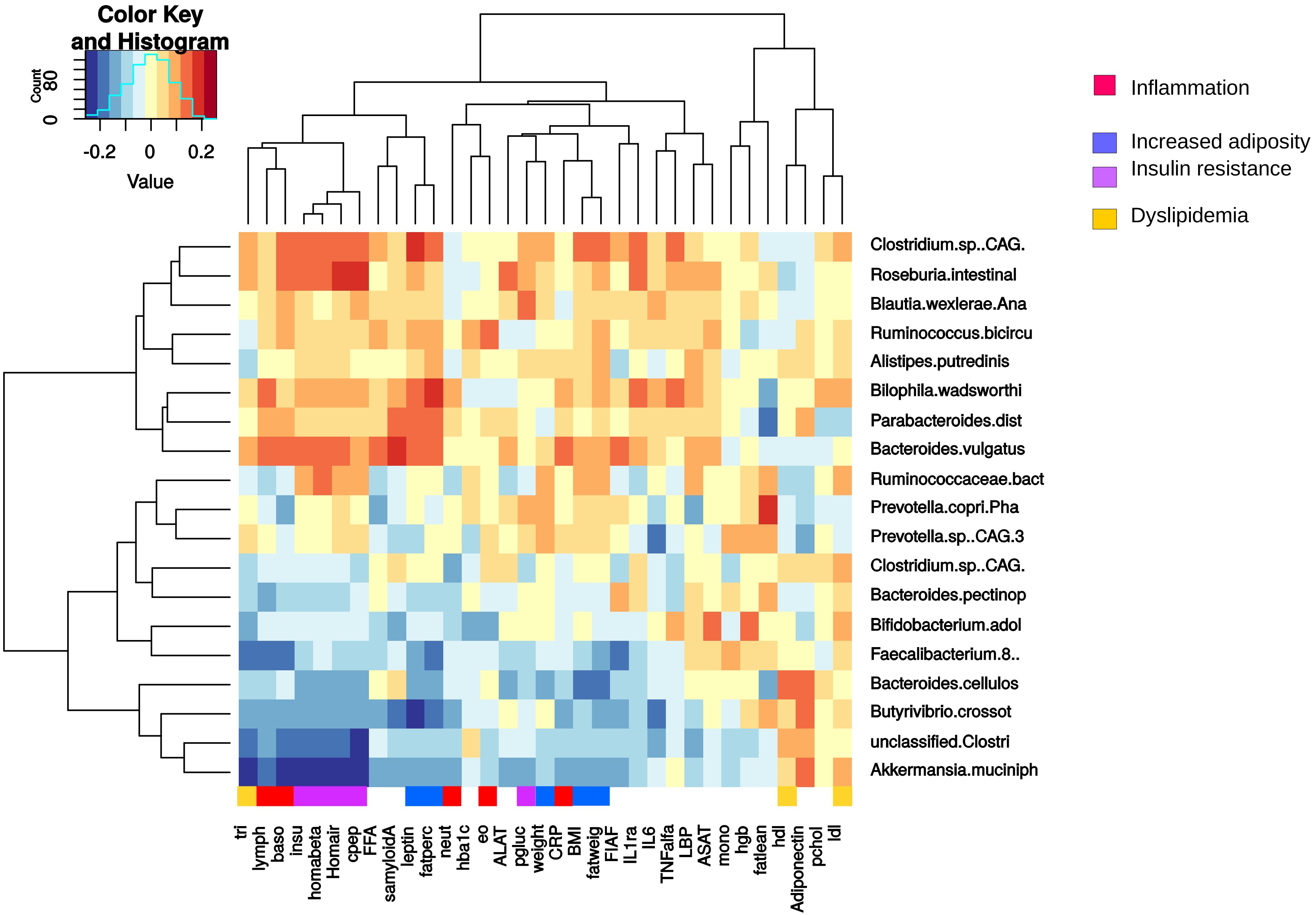


Dynamic of dysbiosis

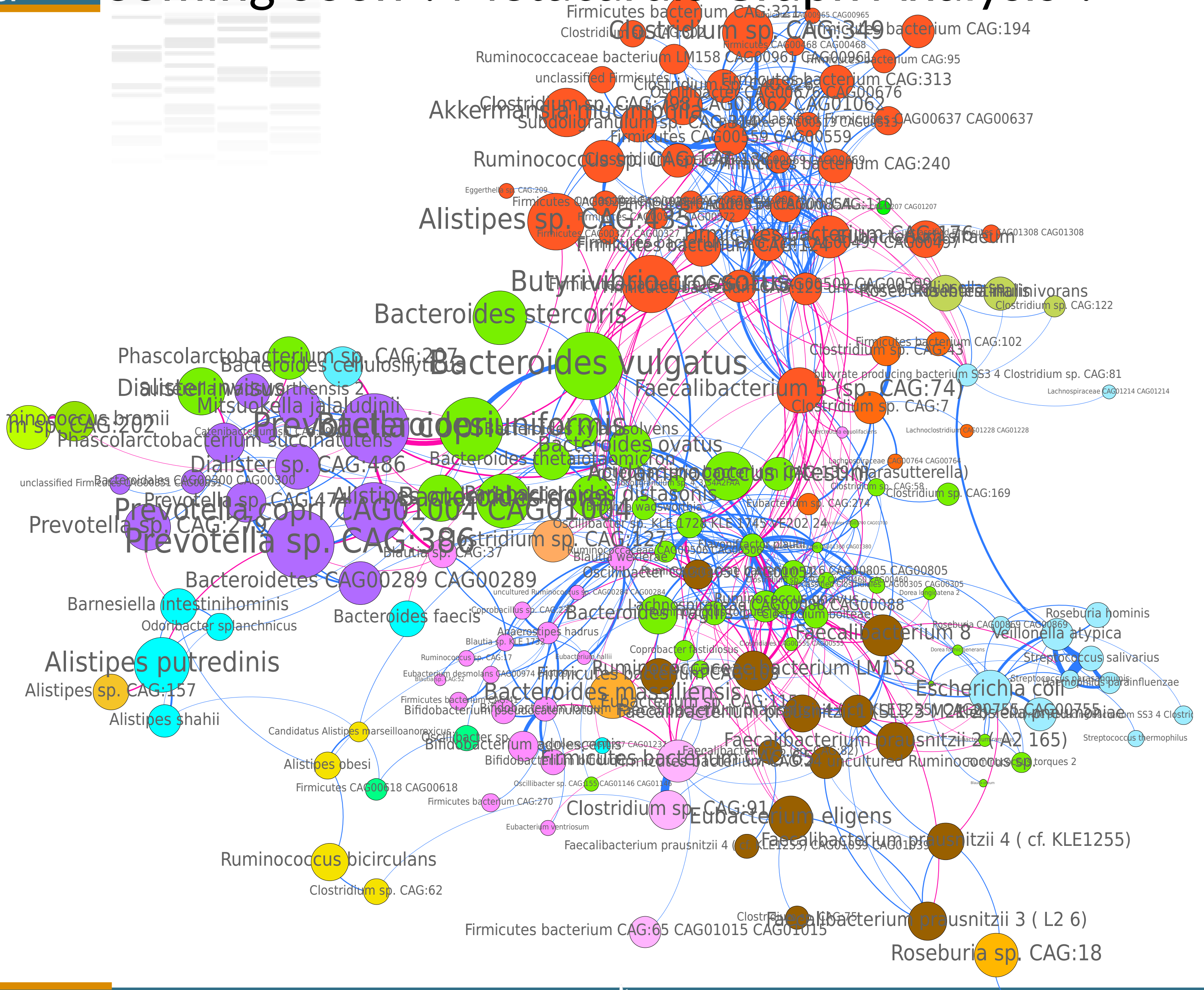
Nutrition – Community - Spearman correlations



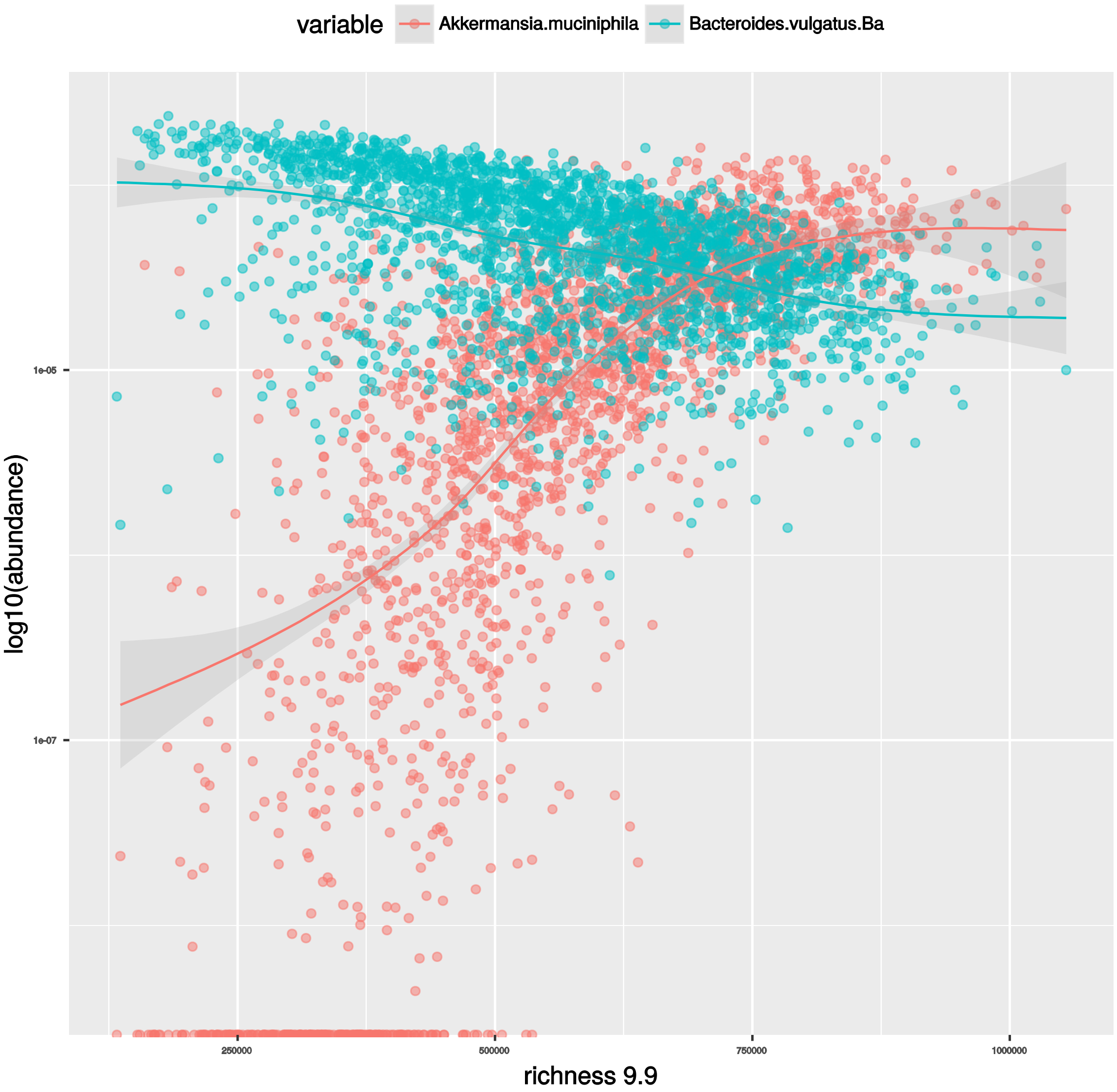
Clinical phenotype – Community - Spearman correlations



Coming Soon : Metacardis Graph Analysis !



Community abundance versus richness in Metacardis



Species level analysis versus community level analysis

Grouping species in community increase contrast between low richness samples and high richness samples.

This “denoising effect” increase statistical power which is really interesting for the high-dimensional noisy data we are studying in metagenomics.

Conclusions

- MGS **graph communities** :
  - **synthesize** gut microbiota on few variables
  - show the **modulation** of gut microbiota between high gene count and low gene count status
  - Are linked to **nutrition** and **clinical traits** such as insulin resistance, adiposity and inflammation

Perspectives

- Characterize the functional content of communities
  - 1 Community = 1 function ?
  - Does the loss of one community lead to the loss of specific function?
  - Can functional annotation can help us to understand the nature of edges (cross-feeding, toxin secretion, competition for nutrients…) ?

- Use the graphical model under the graph to simulate the increase of one MGS and see the effect on the ecosystem
- Assess the robustness of communities on several cohorts

Acknowledgements

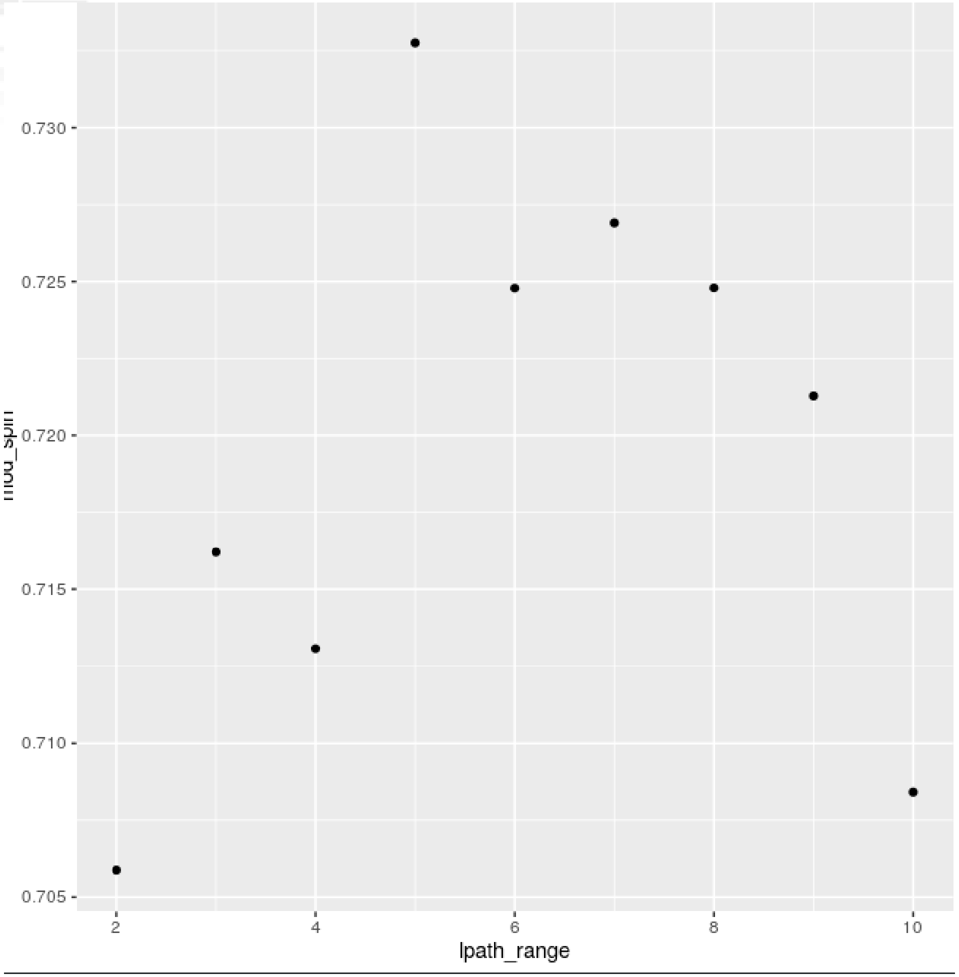
Ariane Basigliani
Marie Jeammet
Anne-Sophie Alvarez
S. Dusko Ehrlich
Franck Gauthier
Emmanuelle Le Chatelier
Nicolas Pons
Pierre Léonard
Nicolas Maziers
Florian Plaza-Onate
Magali Berland
Florence Thirion
Kevin Weiszer

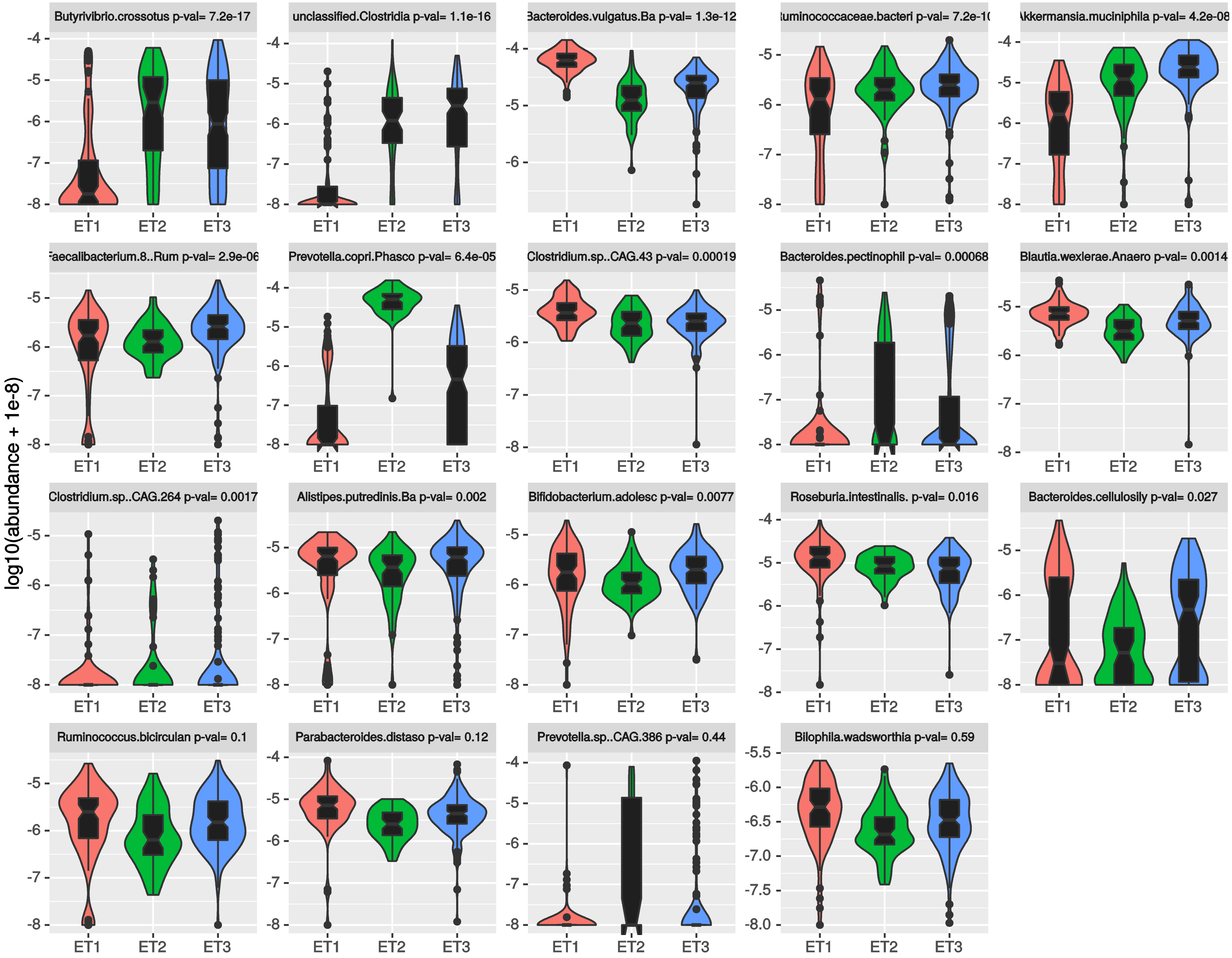
Meteor

Nicolas Pons

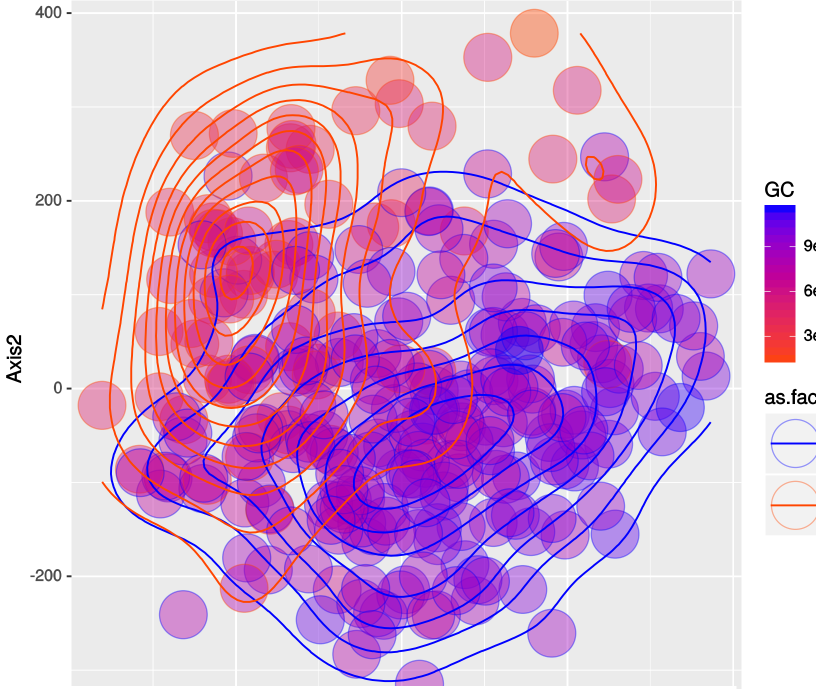
*MetaO*MiNER

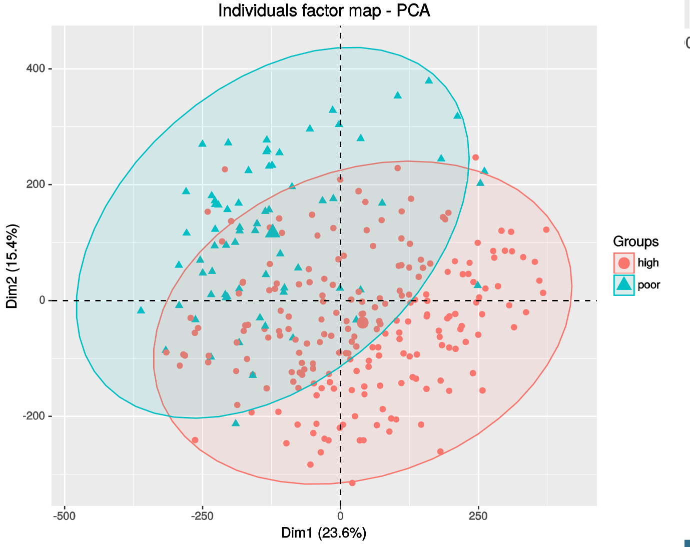
Emmanuelle Le Chatelier
Edi Prifti

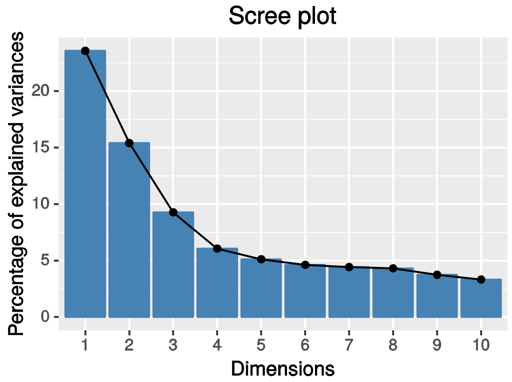

0.730
0.725
0.720
0.715
0.710
0.705
2
4
6
8
10
lpath_range

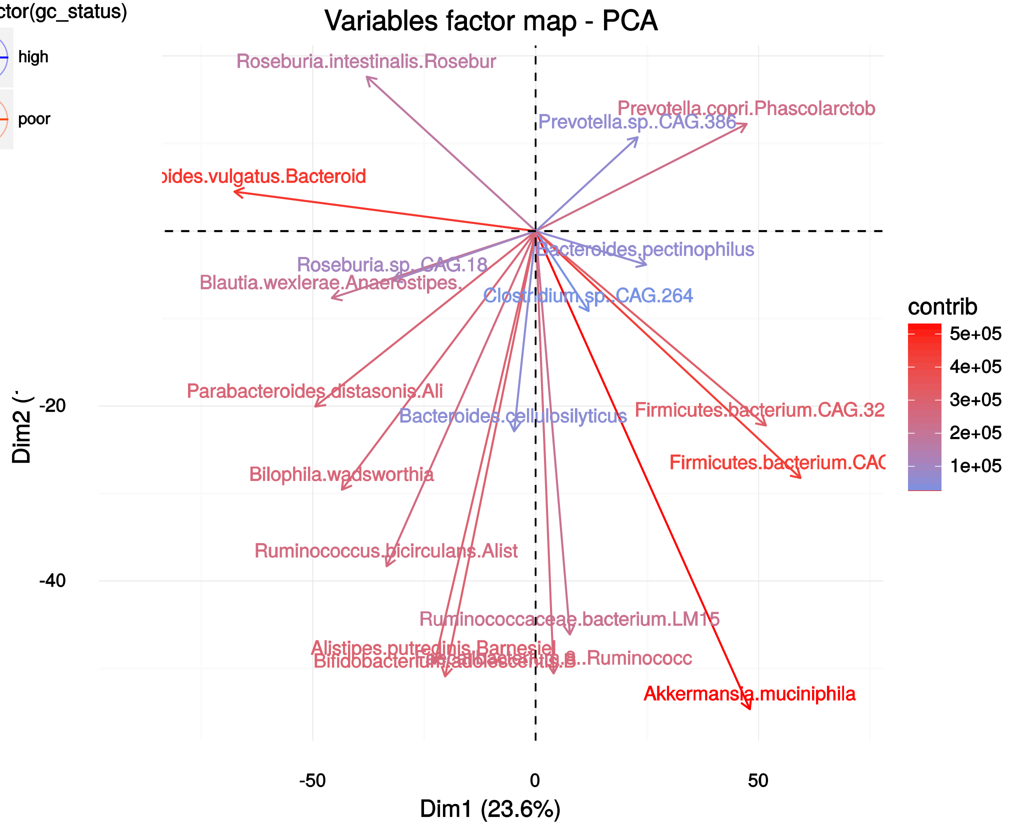

Butyrivibrio.crossotus p-val= 7.2e-17
unclassified.Clostridia p-val= 1.1e-16
Bacteroides.vulgatus.Ba p-val= 1.3e-12
uminococcaceae.bacteri p-val= 7.2e-1
Akkermansia.muciniphila p-val= 4.2e-08
aecalibacterium.8..Rum p-val= 2.9e-06
Prevotella.copri.Phasco p-val= 6.4e-05
Clostridium.sp..CAG.43 p-val= 0.00019
Bacteroides.pectinophil p-val= 0.00068
Blautia.wexlerae.Anaero p-val= 0.0014
Clostridium.sp..CAG.264 p-val= 0.0017
Alistipes.putredinis.Ba p-val= 0.002
Bifidobacterium.adolesc p-val= 0.0077
Roseburia.intestinalis. p-val= 0.016
Bacteroides.cellulosily p-val= 0.027
Ruminococcus.bicirculan p-val= 0.1
Parabacteroides.distaso p-val= 0.12
Prevotella.sp..CAG.386 p-val= 0.44
Bilophila.wadsworthia p-val= 0.59
log10(abundance + 1e-8)
ET1
ET2
ET3
-4
-5
-6
-7
-8
-5.5
-6.0
-6.5
-7.0
-7.5
-8.0

PCA on community vector

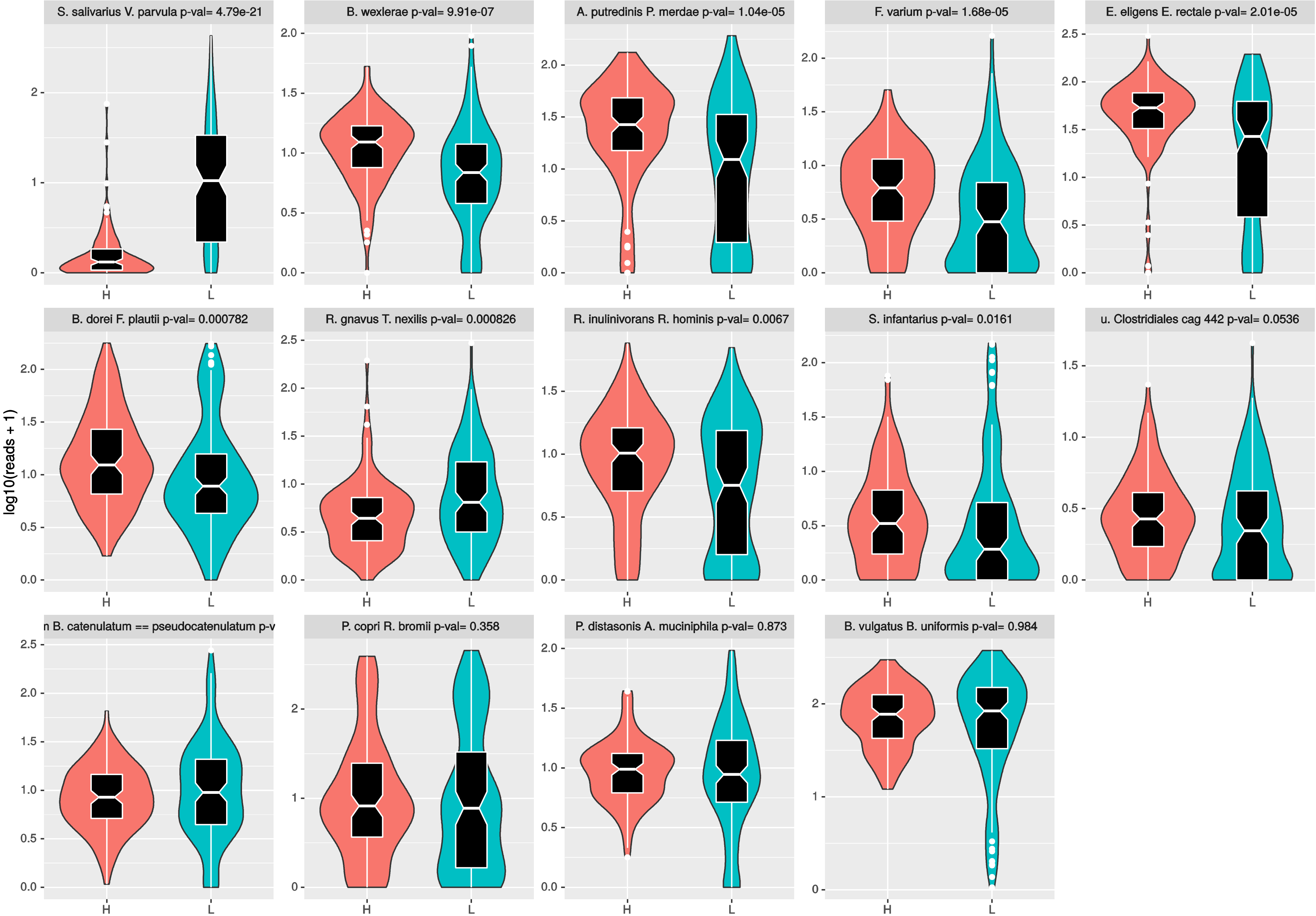






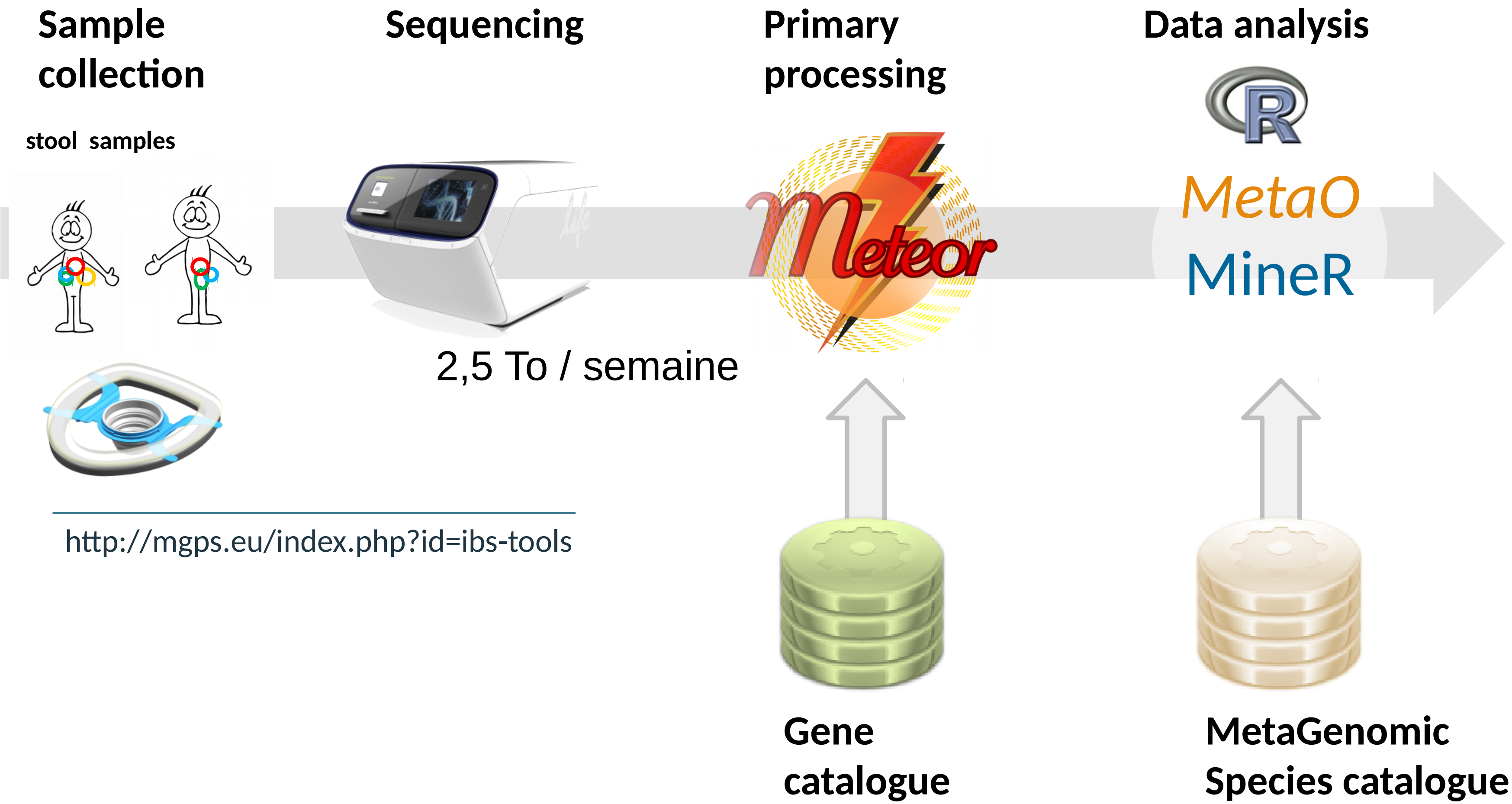


Regroupement en communauté LC



- Augmente le contraste entre Healthy et liverC dans pas mal de cas (cadre rouge). Non systématique
- Synthèse globale de l'état du microbiote plus pertinente que le regroupement par phylum

Quantitative metagenomic pipeline



Sample collection

stool samples

http://mgps.eu/index.php?id=ibs-tools

Sequencing

2,5 To / semaine

Primary processing

Data analysis

MetaO

MineR

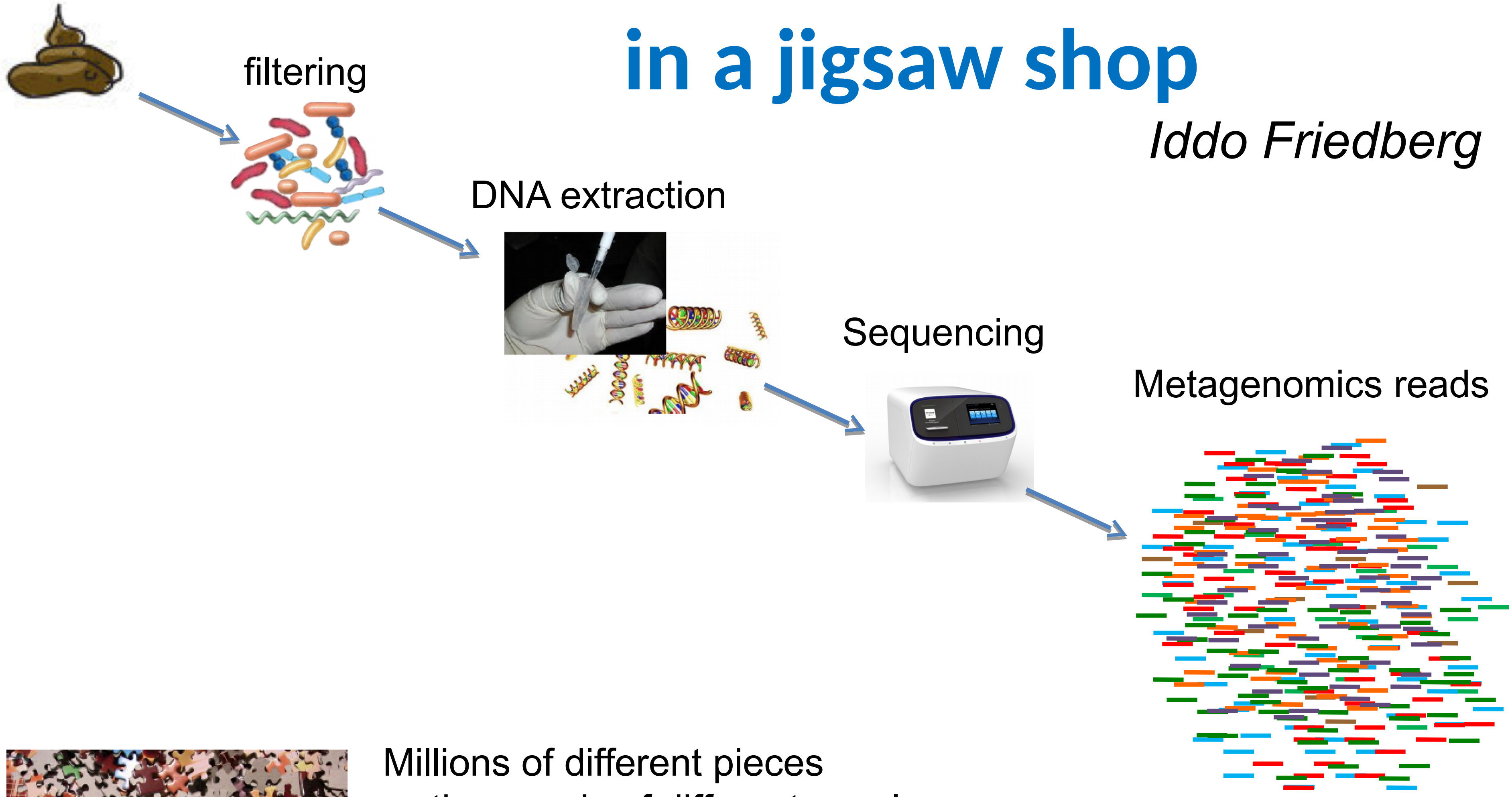
Gene catalogue

MetaGenomic Species catalogue

Slide kindly provided by Magali Berland

Metagenomics is like a disaster in a jigsaw shop

Iddo Friedberg



filtering

DNA extraction

Sequencing

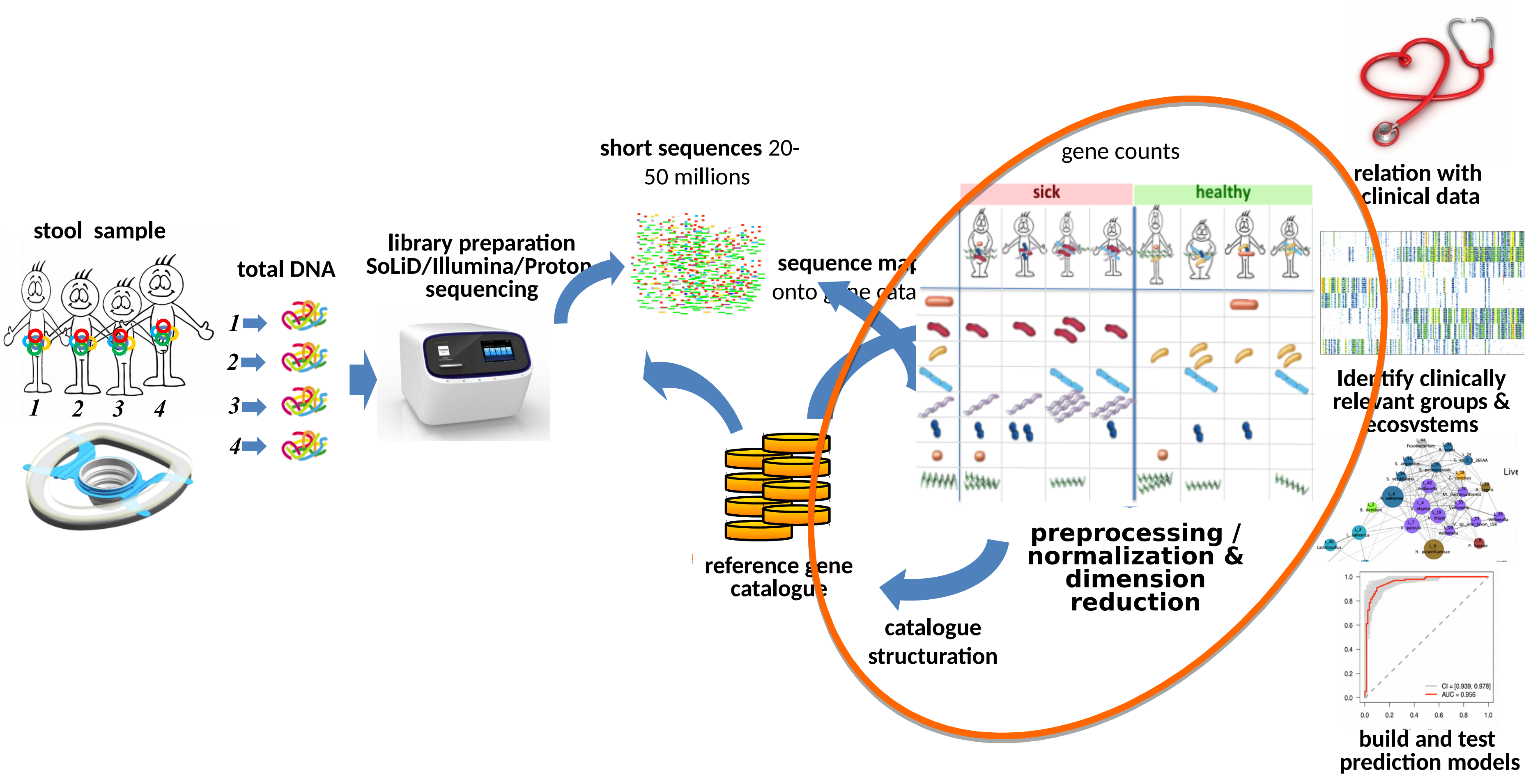
Metagenomics reads

Millions of different pieces

- thousands of different puzzles
- all mixed together
- most of the pieces are missing
- no boxes to refer to
- some pieces do not match

Quantitative metagenomics pipeline

sample collection → **sequencing** → **reference construction** → **gene profiling** → **bioinformatics / statistics analyses**



stool sample

total DNA

library preparation SoLiD/Illumina/Proton sequencing

short sequences 20-50 millions

sequence ma
onto g ne cata

reference gene catalogue

catalogue structuration

preprocessing / normalization & dimension reduction

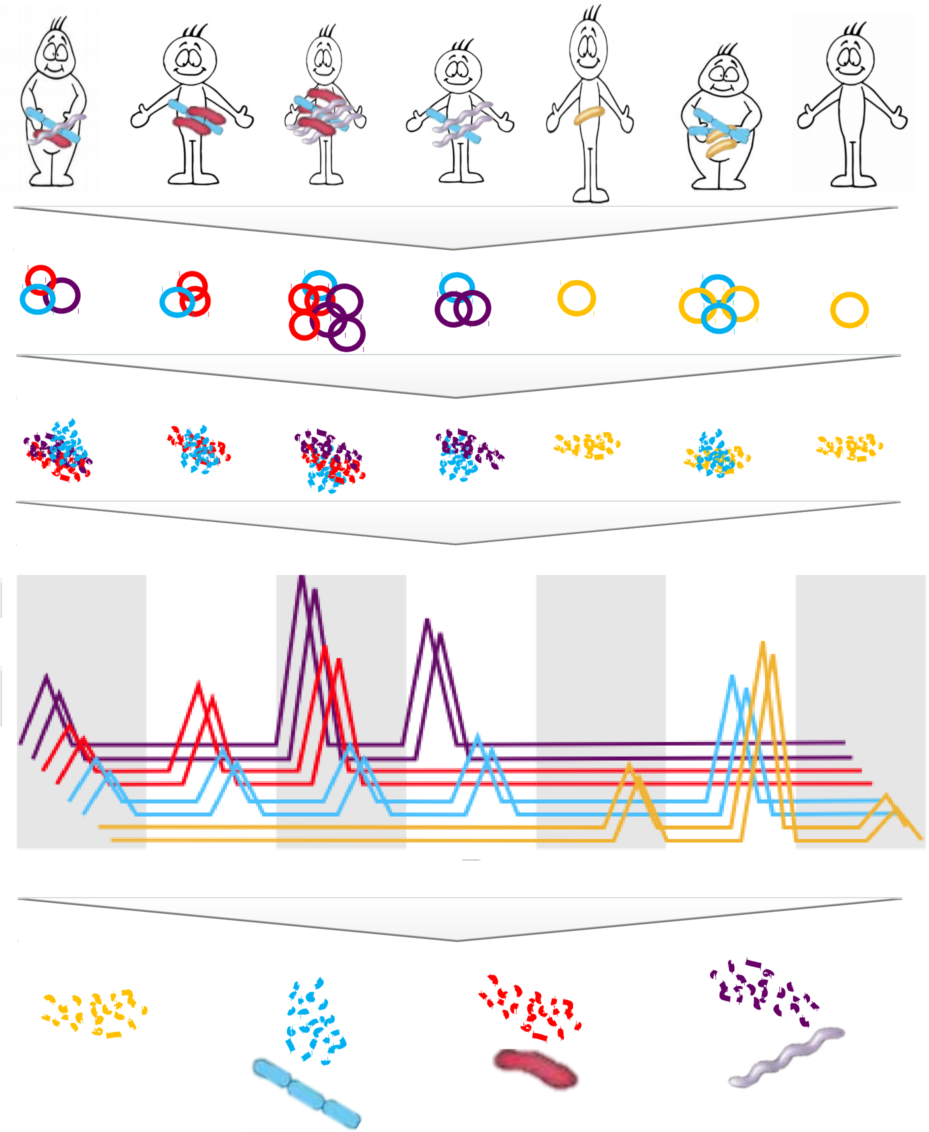
relation with clinical data

Identify clinically relevant groups & ecosystems

build and test prediction models

Clustering genes into MetaGenomic Species (MGS)

Slide kindly provided by
Emmanuelle Le Chatelier



intestinal
gut bacteria

DNA extraction

high throuput
sequencing

abundance profiles of
gene catalogue

organize the catalogue
by co-variance

Dynamic of Dysbiosis