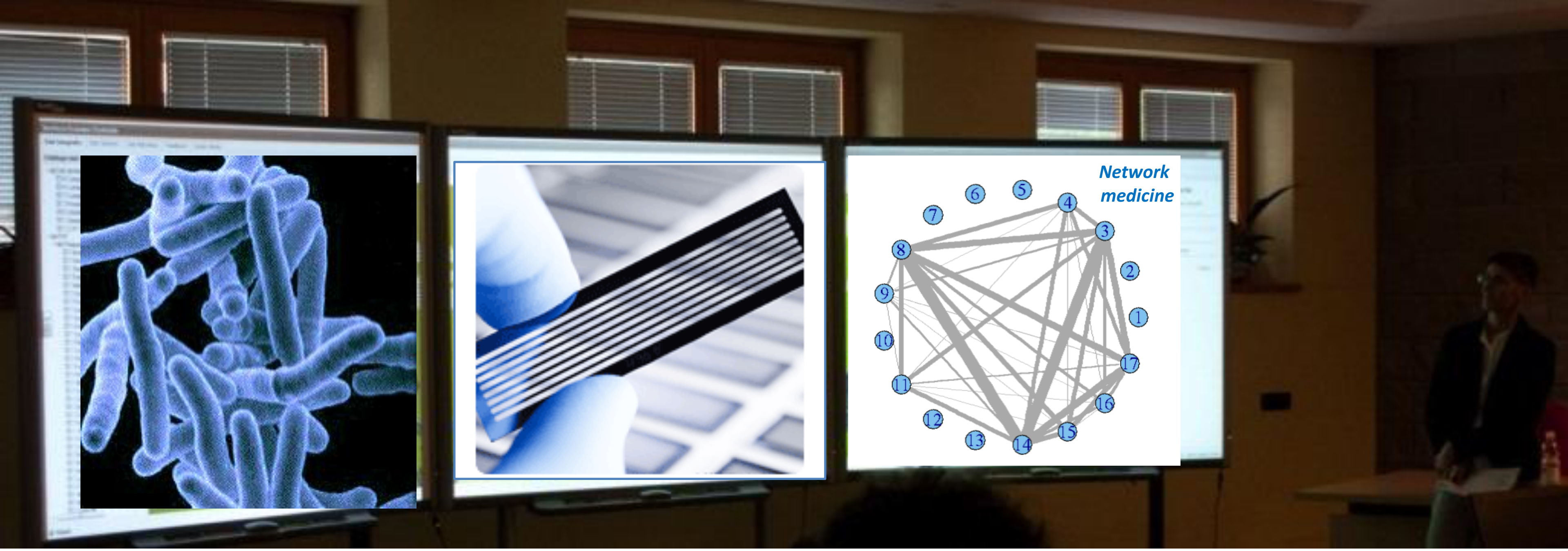


Metagenomics data analytics

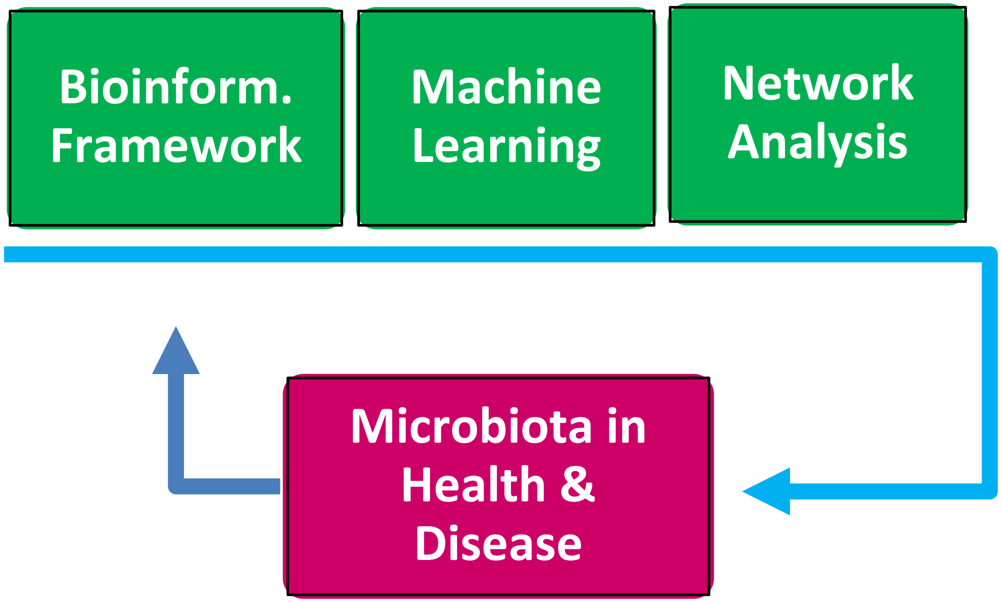
Cesare Furlanello

with A. Zandonà, M. Chierici, G. Jurman

FONDAZIONE BRUNO KESSLER

MAESTRA SUMMER SCHOOL
MINING BIG AND COMPLEX DATA
05 Sept 2016 - Ohrid, Macedonia

Map


Bioinform. Framework
Machine Learning
Network Analysis
Microbiota in Health & Disease

Main concepts

Microbiota

Microorganisms ecosystems inhabiting a particular environment

Microbiome

The community composition, biomolecular repertoire and ecology of microorganisms inhabiting particular environments

(collective genes of the microbiota)

Metagenomics

The application of high-throughput DNA sequencing to profile the genomic composition of a microbial community

- **Taxonomic biomarkers**
- **Functional biomarkers**

Metabolomics

Study of **end products of the metabolism of the host and its microbiota**

Metaproteomics

enabling identification of biomarkers

Metabonomics

comparison with unidentified compounds

Exposomics

cumulative exposures to molecules from the environment

Microbiome impacts on human health

The microbiota affects prenatal and postnatal growth:

Understanding the community structuring could help to prevent and treat disease

Microbiota and diet interact to influence metabolism

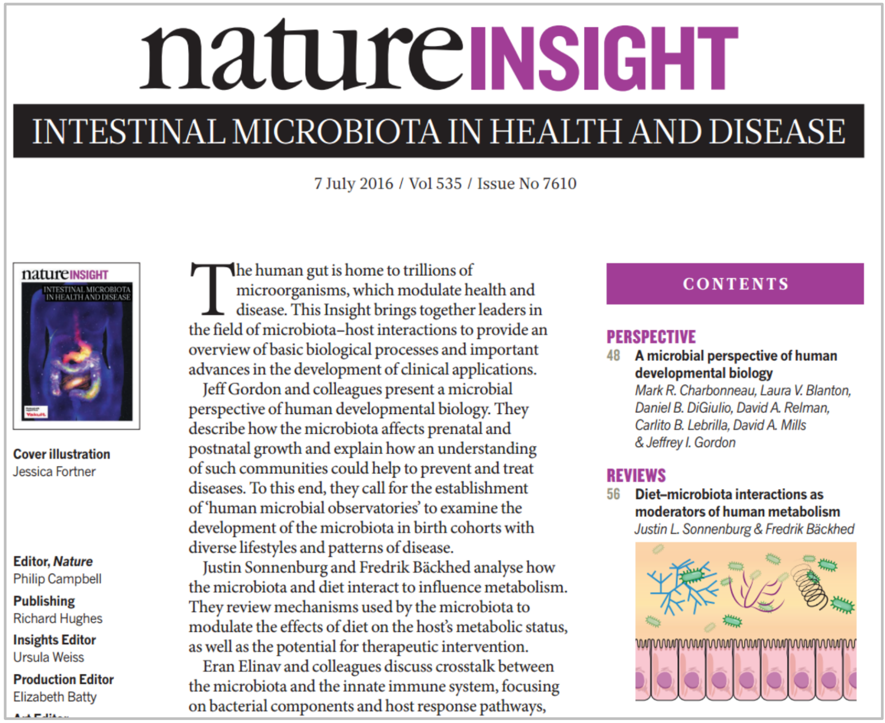
Effects of diet on host metabolic status is modulated → potential for therapeutic interventions

Interaction with pathogenic bacteria

Pathogenic species drive their expansion by exploiting microbiota derived nutrients and triggering inflammation

Specific microbes determine aspects of adaptive immunity

Induction of immune tolerance and conditions (allergy and intestinal inflammation … cancer)



nature INSIGHT

INTESTINAL MICROBIOTA IN HEALTH AND DISEASE

7 July 2016 / Vol 535 / Issue No 7610

Cover illustration
Jessica Fortner

Editor, *Nature*
Philip Campbell

Publishing
Richard Hughes

Insights Editor
Ursula Weiss

Production Editor
Elizabeth Batty

The human gut is home to trillions of microorganisms, which modulate health and disease. This Insight brings together leaders in the field of microbiota–host interactions to provide an overview of basic biological processes and important advances in the development of clinical applications.

Jeff Gordon and colleagues present a microbial perspective of human developmental biology. They describe how the microbiota affects prenatal and postnatal growth and explain how an understanding of such communities could help to prevent and treat diseases. To this end, they call for the establishment of 'human microbial observatories' to examine the development of the microbiota in birth cohorts with diverse lifestyles and patterns of disease.

Justin Sonnenburg and Fredrik Bäckhed analyse how the microbiota and diet interact to influence metabolism. They review mechanisms used by the microbiota to modulate the effects of diet on the host's metabolic status, as well as the potential for therapeutic intervention.

Eran Elinav and colleagues discuss crosstalk between the microbiota and the innate immune system, focusing on bacterial components and host response pathways,

CONTENTS

PERSPECTIVE

REVIEWS

Crosstalk between the microbiota and the innate immune system

Bacterial components and host response pathways can be mutual beneficial, but diseases arise when interaction is disturbed

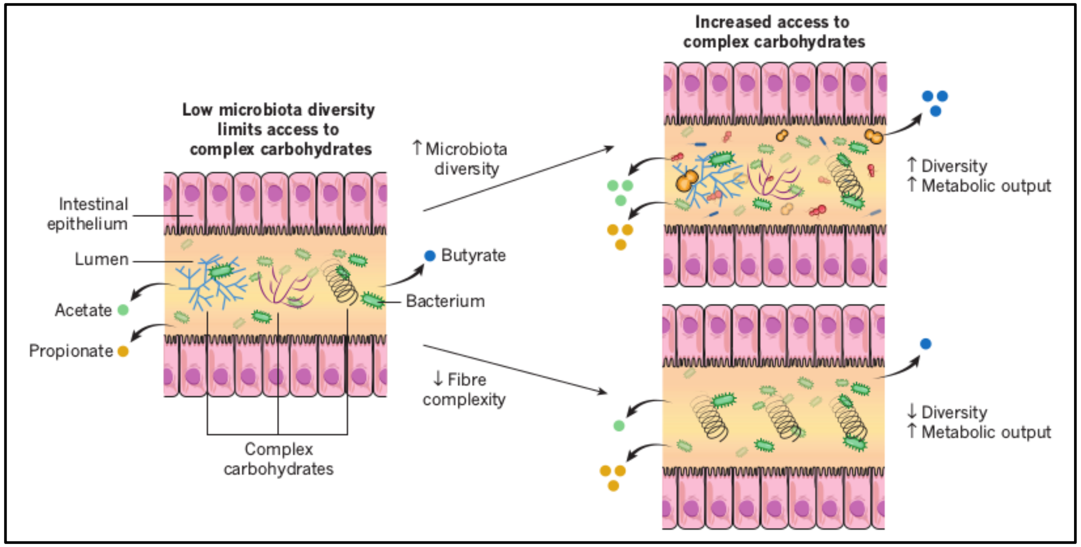
Microbiome-wide association studies DNA Seq, Metabolomics, Computation

Promise of microbiome-based precision diagnostics and therapies

Diet as modulator of gut microbiota

- Microbiota of the human gut responds rapidly to large changes in diet (composition and function of the microbiota shifts over 1–2 days after change in diet)
- Long-term dietary habits are a dominant force in determining the composition of an individual's gut microbiota
- Change in diet can have a highly variable effect on different people owing to the individualized nature of their gut microbiota

[Sonnenburg et al, *Nature*, 2016]

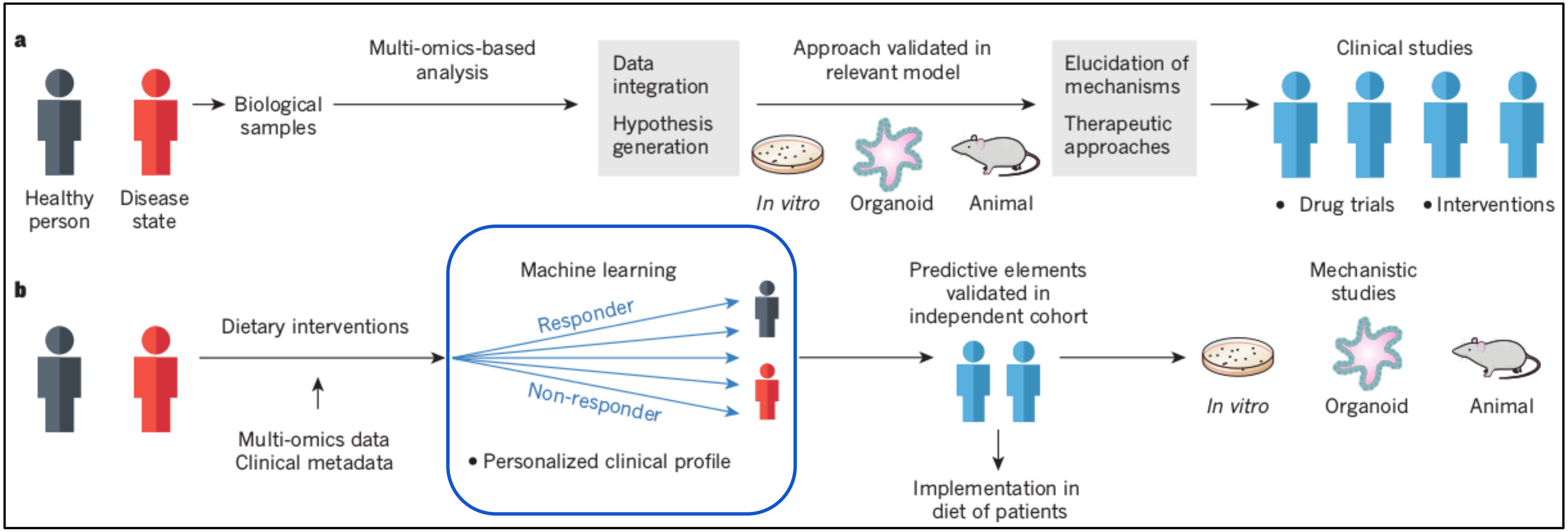


Interactions between the diet and the gut microbiota dictate the production of short-chain fatty acids

Dietary fibre is a source of complex carbohydrates, which are required for the production of **short-chain fatty acids** (i.e.: acetate, butyrate and propionate): anti-inflammatory responses, signalling to the host

Fermentation of fibre in the colon has been shown to **decrease pH levels**, which can help to **increase the diversity of the gut microbiota or results in the reinforcement by certain taxa of a pH that favours their own growth**

ML and diet-based therapeutics



Strategies for modulating the gut microbiota to improve human health

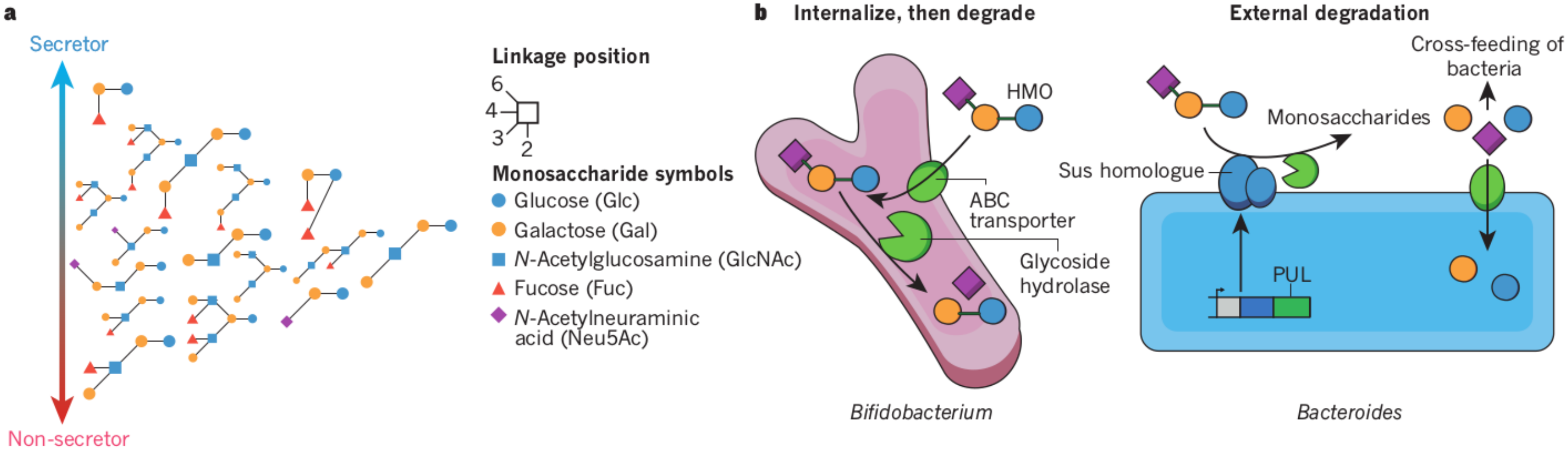
[Sonnenburg et al, *Nature*, 2016]

Machine learning can be used to identify aspects of the clinical profile of individuals (including data on the microbiota) that help to predict the response of others to dietary interventions

Such predictive elements can also be used to guide mechanistic studies in experimental models.

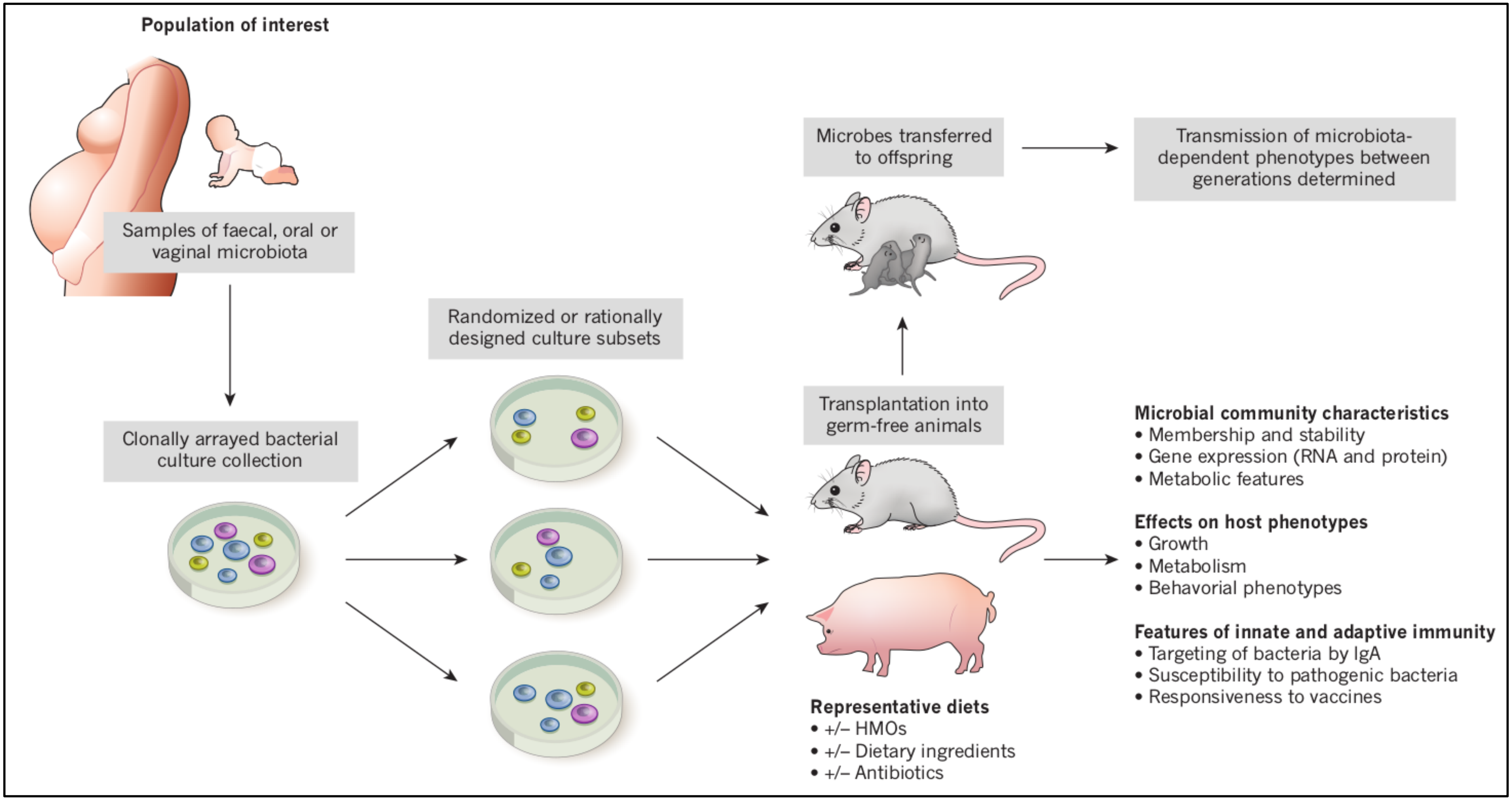
Maternal-fetal microbial landscape

- Vaginal microbiota composition is more stable during pregnancy than at other times during adulthood (*Lactobacillus*-dominated community)
- The initial microbiota of nursing infants is an assemblage of microbes derived from mother's faecal, vaginal and skin microbiota
- Microbes that are transferred to offspring before or during delivery might reflect environmental exposures of the mother during pregnancy (for example, diet)
- Within weeks, development of a **milk-oriented microbiota** occurs: microbiota dominated by ***Bifidobacterium*** species whose primary end fermentation products important sources of energy for colonocytes. Can also result in **'cross-feeding' of secondary consumers,** including potentially pathogenic bacteria in the infant gut.
- Variations in the transfer of microbes from mothers to infants might affect early postnatal development of the child's microbiota, immune system and metabolic processes.



[Charbonneau et al, *Nature*, 2016]

Discovery pipeline for developing microbiome characterization



Population of interest

Samples of faecal, oral or vaginal microbiota

Clonally arrayed bacterial culture collection

Randomized or rationally designed culture subsets

Transplantation into germ-free animals

Microbes transferred to offspring

Transmission of microbiota-dependent phenotypes between generations determined

Representative diets
- +/– HMOs
- +/– Dietary ingredients
- +/– Antibiotics

Microbial community characteristics
- Membership and stability
- Gene expression (RNA and protein)
- Metabolic features

Effects on host phenotypes
- Growth
- Metabolism
- Behavioral phenotypes

Features of innate and adaptive immunity
- Targeting of bacteria by IgA
- Susceptibility to pathogenic bacteria
- Responsiveness to vaccines

[Charbonneau et al, *Nature*, 2016]

Test for effects of different community configurations on host biology

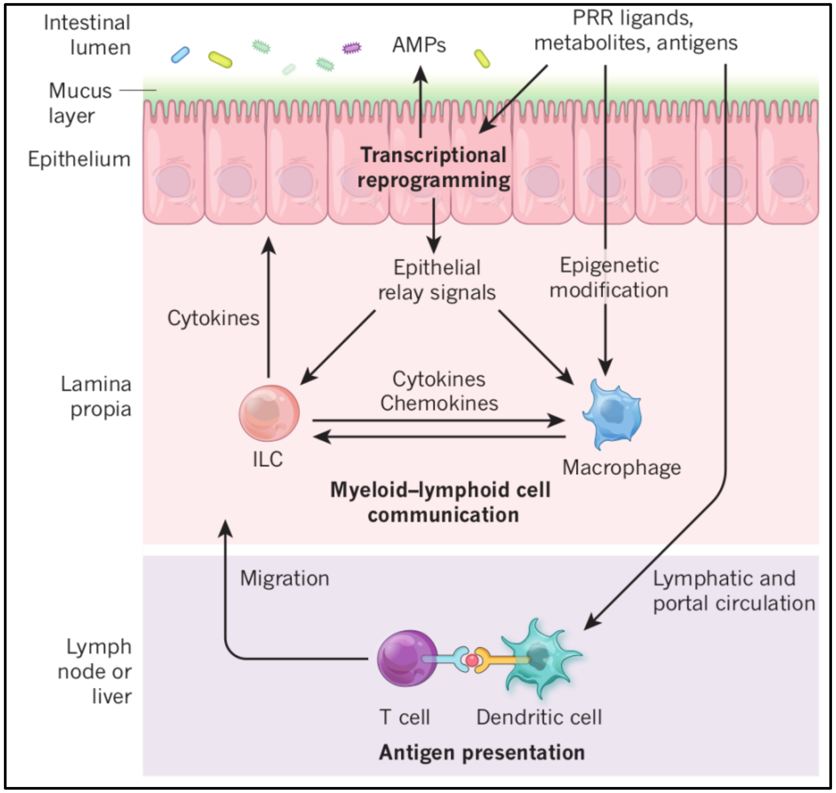
Recipient animals are fed diets representative of those consumed by their microbiota donors, or diets designed to test hypotheses about the role of various components, including HMOs, on microbiota-mediated functions

Gut microbiota and inflammation

Dysbiosis (imbalance in the microbiota) is characterized by

- a reduced diversity of microbes
- a reduced abundance of obligate anaerobic bacteria
- an expansion of facultative anaerobic bacteria in the phylum *Proteobacteria*, mostly members of the family *Enterobacteriaceae*

Intestinal inflammation in people is associated with **Dysbiosis**



[Thaiss et al, *Nature*, 2016]

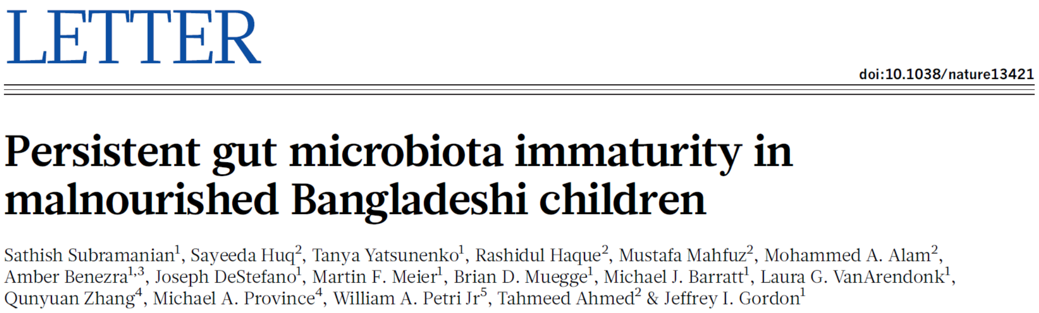
Drivers of changes in the nutritional environment

1. The availability of nutrients in the large intestine is altered during inflammation through changes in the composition of mucous carbohydrates.
2. generation of reactive oxygen species and reactive nitrogen species during inflammation.

Feedback loops between the host and the microbiome

Feedback loops that extend to the underlying lamina propria involve communication between epithelial, myeloid and lymphoid cells using cytokines and chemokines

Microbiome in malnourished children



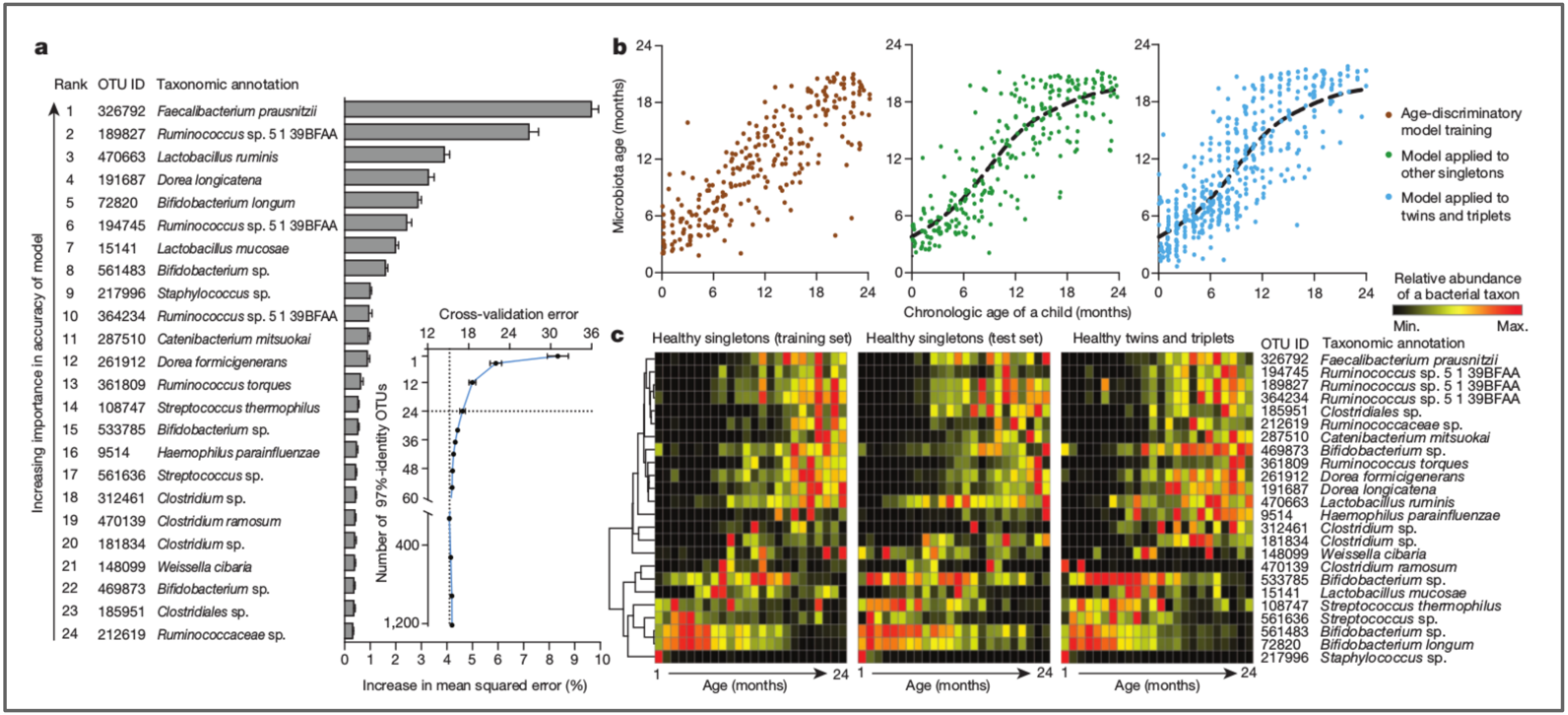
LETTER

doi:10.1038/nature13421

Persistent gut microbiota immaturity in malnourished Bangladeshi children

Sathish Subramanian[1], Sayeeda Huq[2], Tanya Yatsunenko[1], Rashidul Haque[2], Mustafa Mahfuz[2], Mohammed A. Alam[2], Amber Benezra[1,3], Joseph DeStefano[1], Martin F. Meier[1], Brian D. Muegge[1], Michael J. Barratt[1], Laura G. VanArendonk[1], Qunyuan Zhang[4], Michael A. Province[4], William A. Petri Jr[5], Tahmeed Ahmed[2] & Jeffrey I. Gordon[1]

Microbiome in malnourished children



Subramanian et al, *Nature*, 2014

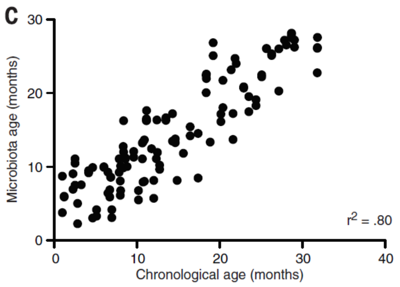
Severe Acute Malnutrition is associated with significant relative microbiota immaturity

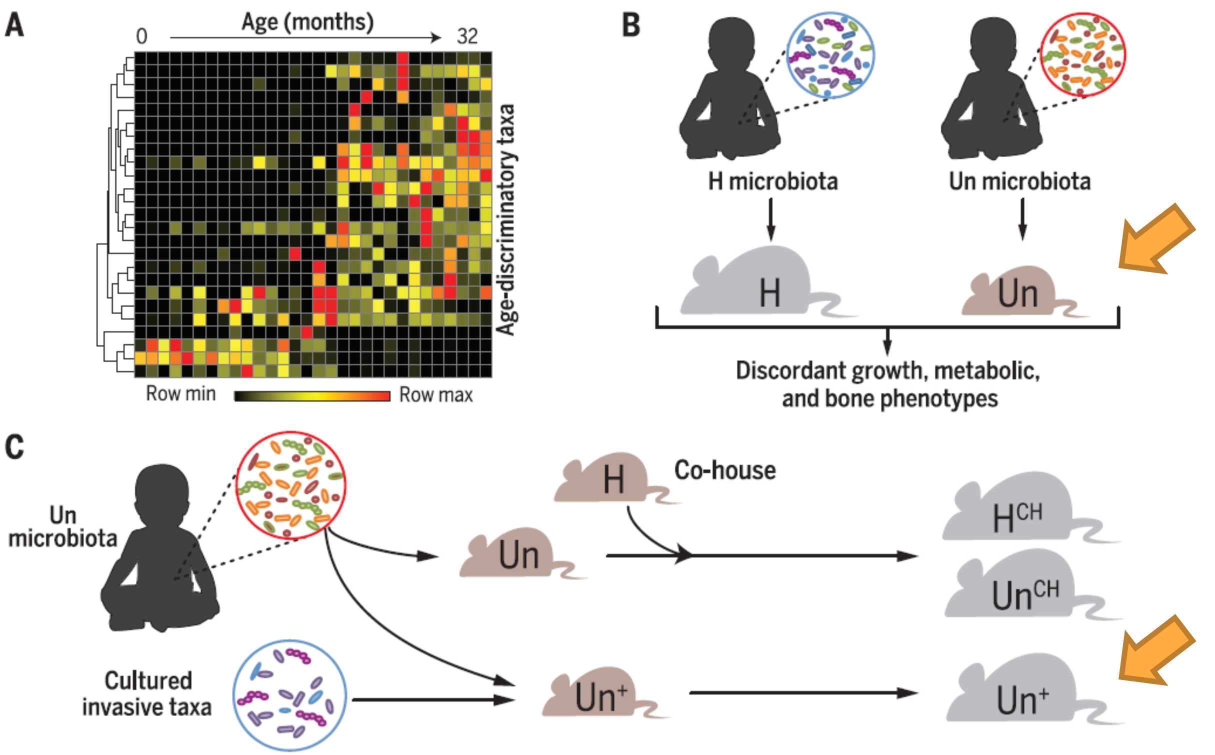
- Machine Learning approach: Random Forest models

APPLICATIONS OF THE “METAGENOMIC CLOCK” IN PRECLINICAL STUDIES

Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children.

Blanton LV et al, Science, Feb 2016

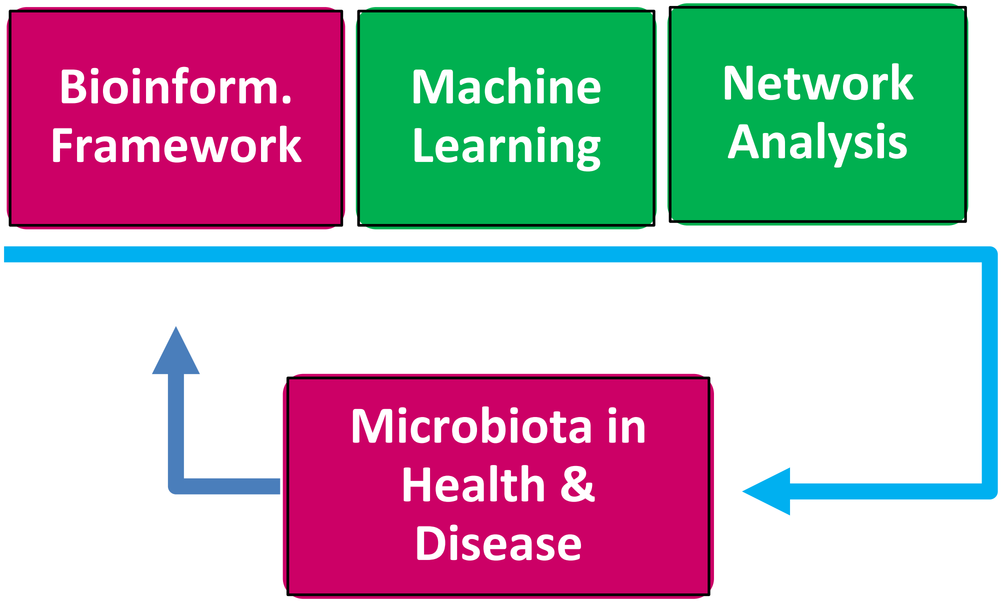




Preclinical evidence that gut microbiota immaturity is causally related to childhood undernutrition. (**A**) A model of normal gut microbial community development in Malawian infants and children, based on the relative abundances of 25 bacterial taxa that provide a microbial signature

- **Model of microbiota:** 36 mo maturation in twin pairs healthy Malawian infants and children by using **RF to regress OTUs against chronological age**, val on 259 h.
- **Undernourished children in a Malawian birth cohort: → immature gut microbiota.**
- Unlike microbiota from healthy children, **immature microbiota transmit impaired growth**, altered bone morphology, and metabolic abnormalities in the muscle, liver, and brain to recipient gnotobiotic mice.

Map/2



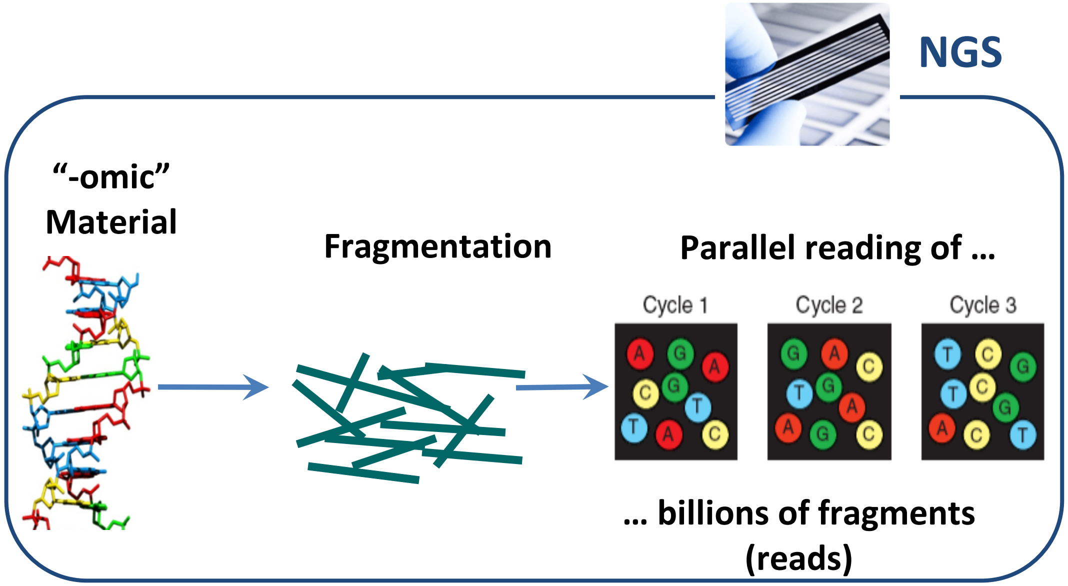
Bioinform. Framework

Machine Learning

Network Analysis

Microbiota in Health & Disease

Next Generation sequencing



NGS

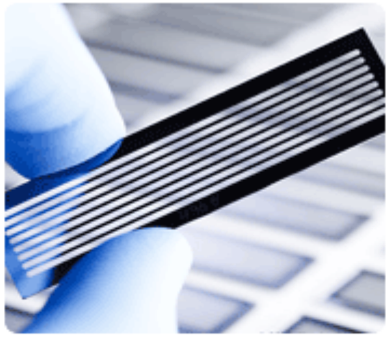
“-omic” Material

Fragmentation

Parallel reading of …

… billions of fragments (reads)

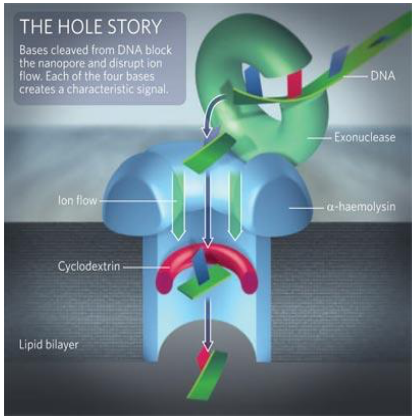
- **Massively parallel** sequencing platforms able to produce **millions of sequences** concurrently, with protocols for DNA, gene expression, methilation, …
- Throughput: up to 25 Gb (~8 human genomes) per day
- More than 85% bases correctly sequenced with accuracy ≥ 99.9% (Illumina HiSeq 2000)

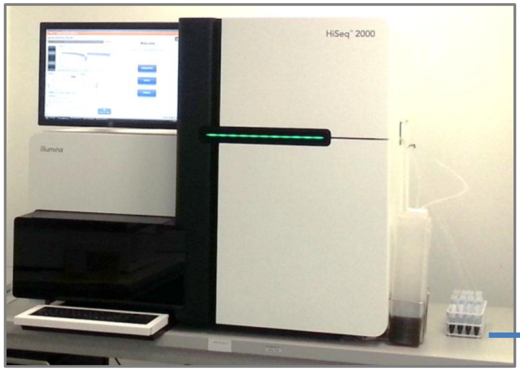


Which platforms for metagenomics markers?

Next Gen Sequencing methods for metagenomics research and clinical applications:

1. Roche 454 Genome Sequencer FLX System
2. Illumina HiSeq / MiSeq
3. Ion Torrent PGM
4. Oxford Nanopore


THE HOLE STORY
Bases cleaved from DNA block the nanopore and disrupt ion flow. Each of the four bases creates a characteristic signal.
DNA
Exonuclease
Ion flow
α-haemolysin
Cyclodextrin
Lipid bilayer



LaBSSAH: Lab.of Biomolecular Sequence and Structure Analysis for Health, a partnership of FBK, UniTN/CIBIO & CNR, with FEM

MinION: electronic single-molecule nanopore sensing (DNA, proteins)

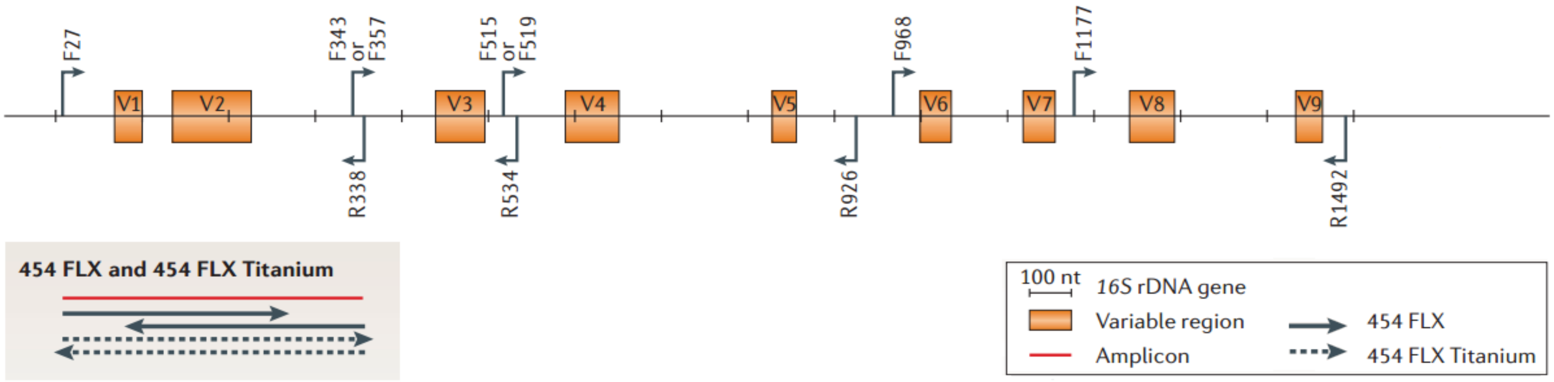
Studying Metagenomics with NGS

Targeted amplicons sequencing

- Only *Gene 'markers'* assumed phylogenetically informative are sequenced
- Most used marker: the gene 16S rRNA, common in all life forms

Whole genome sequencing(WGS)

- Whole (intronic+exonic) genomes from the potential microbiota, incl. fungi and viruses
- Similar 16S are distinguished
- Strains may be identified
- 3 billion 100bp reads (HiSeq), 15 million 36bp (MiSeq)



Experimental and analytical tools for studying the human microbiome. Kuczynski, 2012.

Bioinformatics and the microbiome

Major research areas

1. Sequence Analysis
2. Genome Annotation
3. Computational Biology
4. Meta-transcriptomics
5. Functional Annotations
6. Comparative Genomics
7. Phylogenetics Analysis
8. Networks & Systems Biology

Bioinformatics

A. Sequence Pre-filtering
B. Assembly
C. Gene Prediction
D. Biodiversity
E. Comparative Metagenomics

International Projects (USA,EU)

Meta HIT

HMP
NIH HUMAN
MICROBIOME
PROJECT

iHMP
Integrative Human Microbiome Project

The Inflammatory Bowel Disease Multi'omics Database

Multi-Omic Microbiome Study-Pregnancy Initiative

Onset of Type 2 Diabetes

Major reference databases

16S rRNA

green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

Release gg_13_5_99 (2013/05):
- 202,421 bacterial and archaeal sequences

silva
high quality ribosomal RNA databases

Release 115 (SSURef NR):
- 418,497 bacterial sequences
- 17,530 archaeal sequences
- 43,698 eukaryotic sequences

WGS

HMP
NIH HUMAN MICROBIOME PROJECT

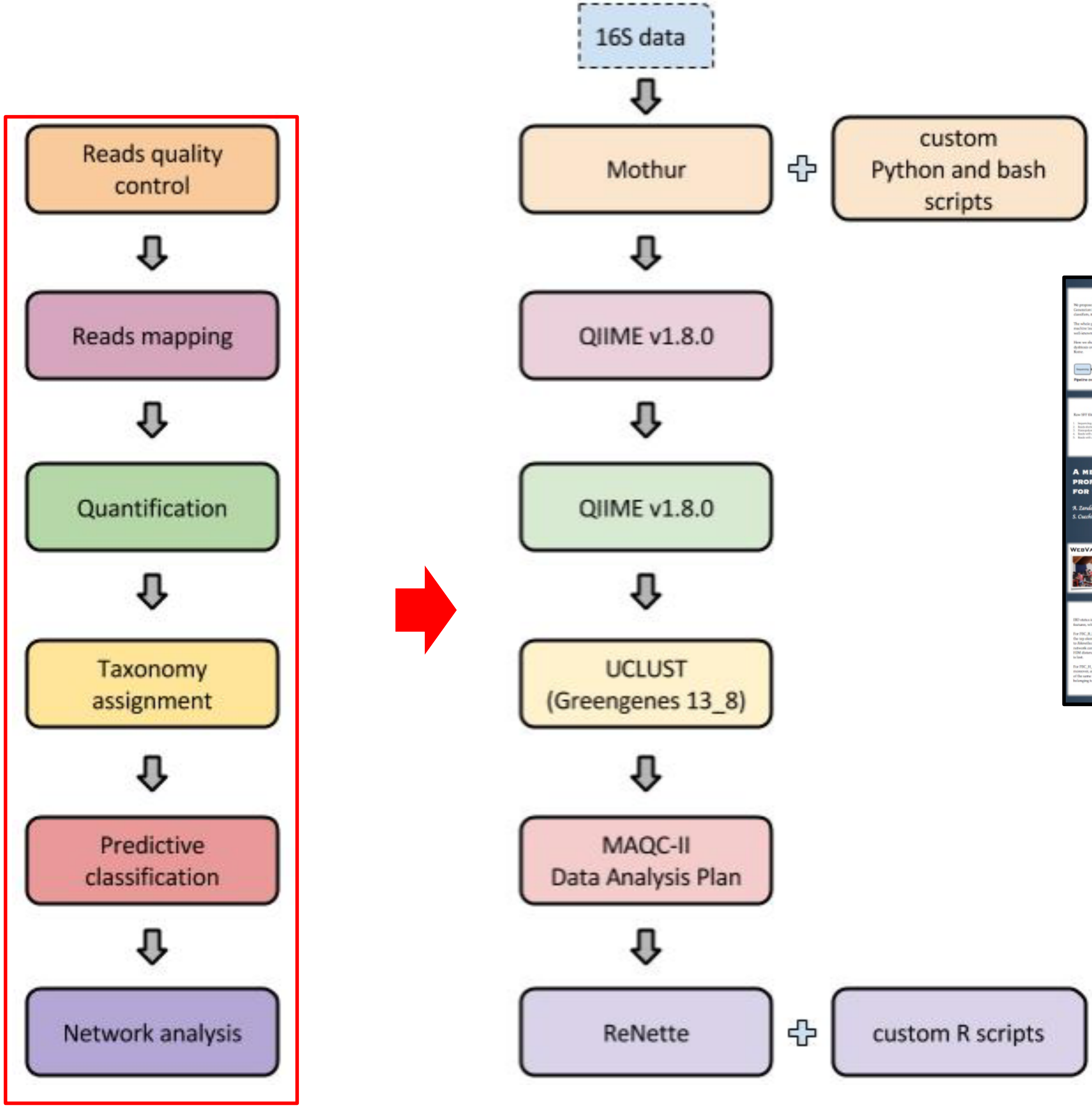
Ref. metagenome (*http://www.hmpdacc.org/HMREFG/*):
- 1,253 Bacteria
- 97 Archaea
- 326 Eukaryotes
- 1,420 Viruses

NCBI

Ref. metagenome (2013/06):
- 2,367 Bacteria, Archaea
- 35 Fungi
- 2,397 Viruses

Also: KEGG, COG, GO, EggNOG

Bioinformatics + ML Framework



Reads quality control

Reads mapping

Quantification

Taxonomy assignment

Predictive classification

Network analysis

16S data

Mothur + custom Python and bash scripts

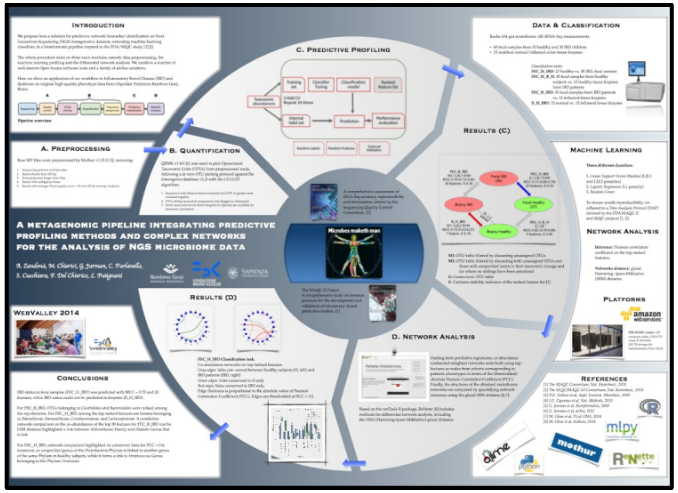
QIIME v1.8.0

QIIME v1.8.0

UCLUST (Greengenes 13_8)

MAQC-II Data Analysis Plan

ReNette + custom R scripts

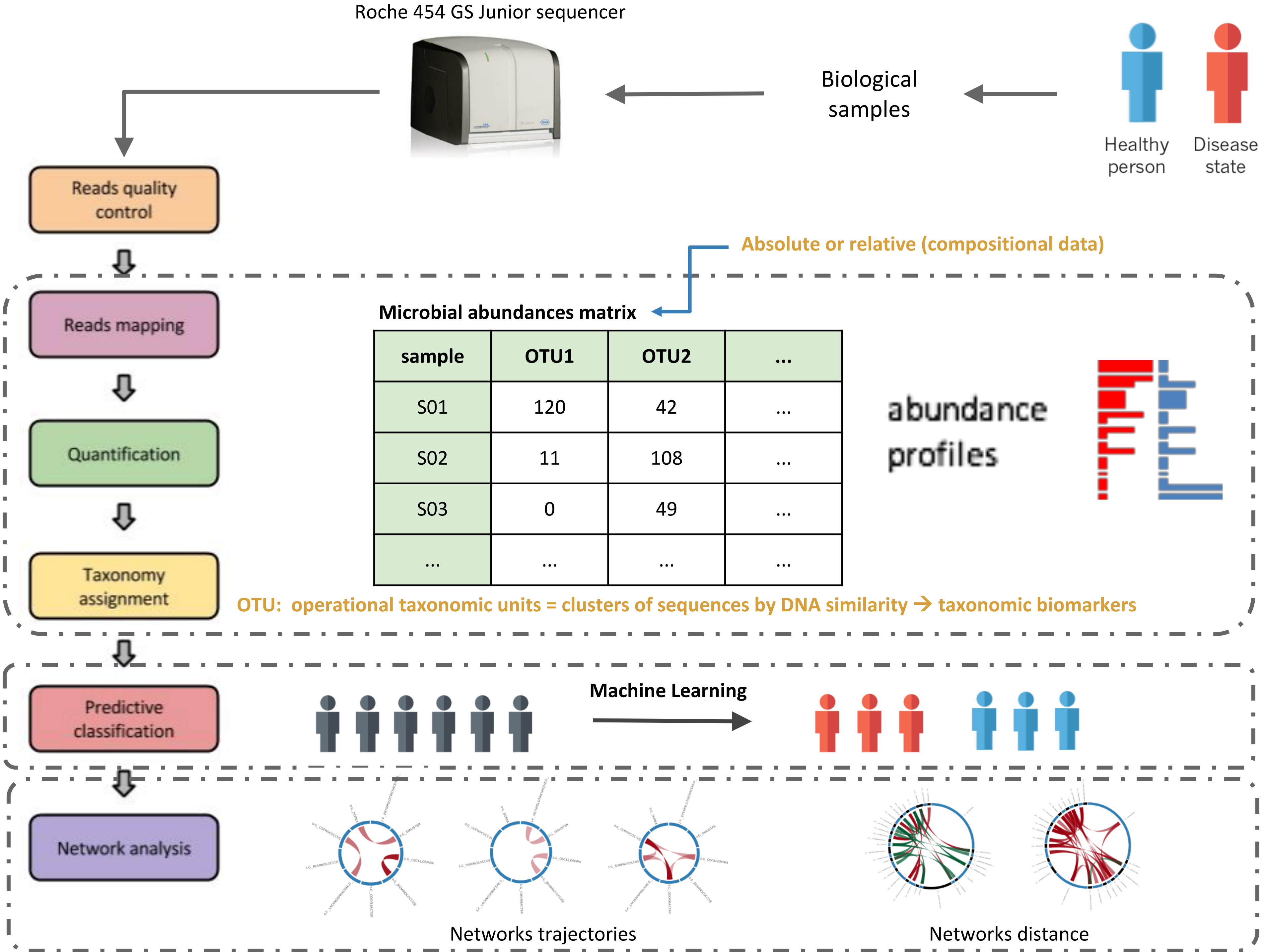


Zandonà, Chierici, Jurman, Del Chierico, Cucchiara, Putignani, Furlanello

NIPS-MLCB Workshop 2014

Machine Learning in Computational Biology: Montreal- **Dec 13, 2014**

The FBK Kore HPC cluster - *May 2013:* ***~1000 cores in 196*** *multi-processor sockets ("blades"); about 5 TB RAM, 25 TB scratch area, 200 TB for genomics, 100 utenti (SON of Grid Engine queue system) –* ***now connected to FEM campus***



Roche 454 GS Junior sequencer

Biological samples

Healthy person

Disease state

Reads quality control

Reads mapping

Quantification

Taxonomy assignment

Predictive classification

Network analysis

Absolute or relative (compositional data)

Microbial abundances matrix

sample	OTU1	OTU2	...
S01	120	42	...
S02	11	108	...
S03	0	49	...
...	...	...	...

abundance profiles

OTU: operational taxonomic units = clusters of sequences by DNA similarity → taxonomic biomarkers

Machine Learning

Networks trajectories

Networks distance

A warning about compositional data

Two types of metagenomic data: absolute vs relative abundance (compositional data)

For each sample, sum of microbial abundance is equal to 1
(growth or decay is connected to decay or growth of all others)

- Traditional **Pearson correlation** analysis treating the observed data as absolute abundances of the microbes may lead to spurious results with relative abundances.

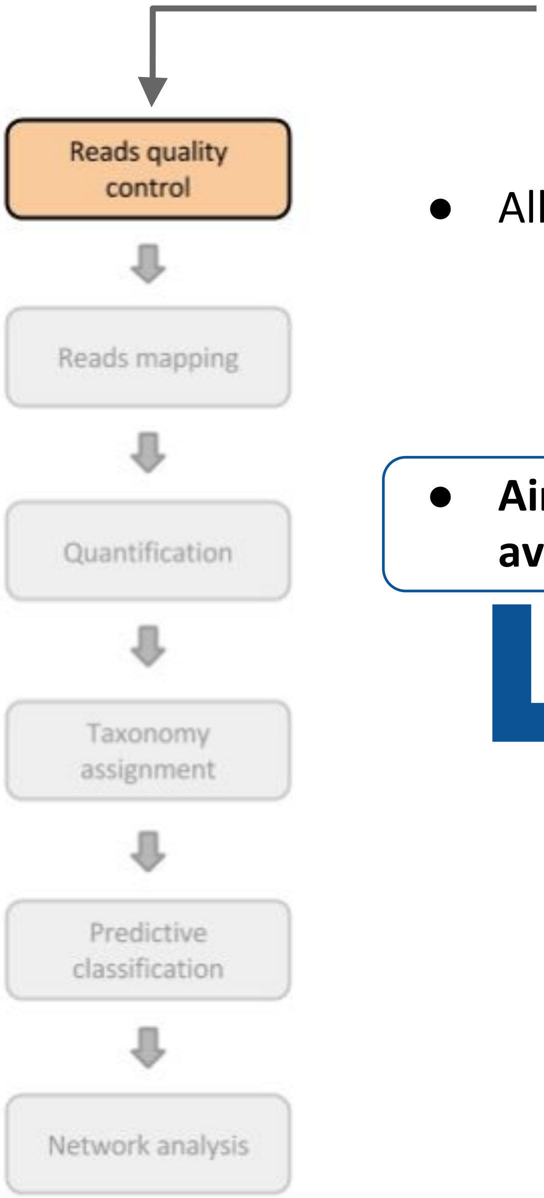
- Special care and appropriate methods are required prior to correlation analysis for these compositional data.

CCLasso: novel method based on least squares with ℓ1 penalty to infer the correlation network for latent variables of compositional data from metagenomic data.
An effective alternating direction algorithm from augmented Lagrangian method is used to solve the optimization problem.

[Fang et al, *Bioinformatics*, 2015]

Roche 454 GS Junior sequencer

Biological samples



- All sequencing platforms have artefacts

main artefact: **long homopolymers**

```
CTTCGGGTGCGTTTTTTTTGCCCC
CTTCGGGTGCG-TTTTTTTGCCCC
CTTCGGGTGCGTTTTTTTTGCCCC
CTTCGGGTGCG-TTTTTTTGCCCC
CTTCGGGTGCG-TTTTTTTGCCCC
```

- **Aim: getting qualitative and quantitative information about data available for further analysis**

http://www.mothur.org

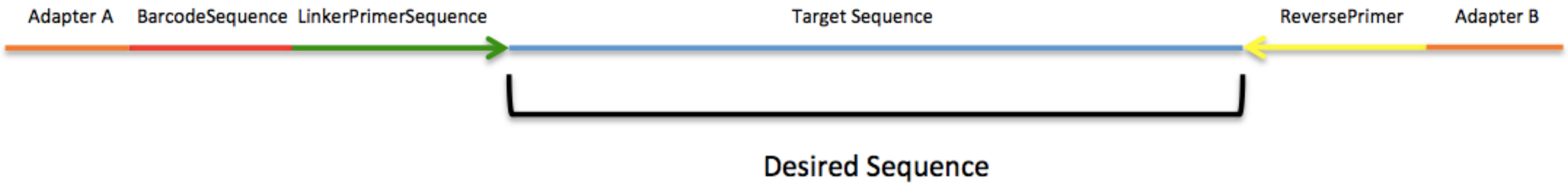
mothur

trim.seqs()

remove:

- ➔ redundancies
- ➔ low quality reads
- ➔ long homopolymers
- ➔ primers and adapters

Trimming primers and adaptors

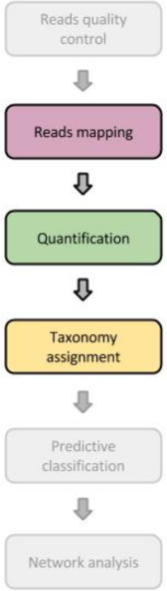


The adapter and primer sequences do not correspond to the bases at the 3' end of the reference genome sequence

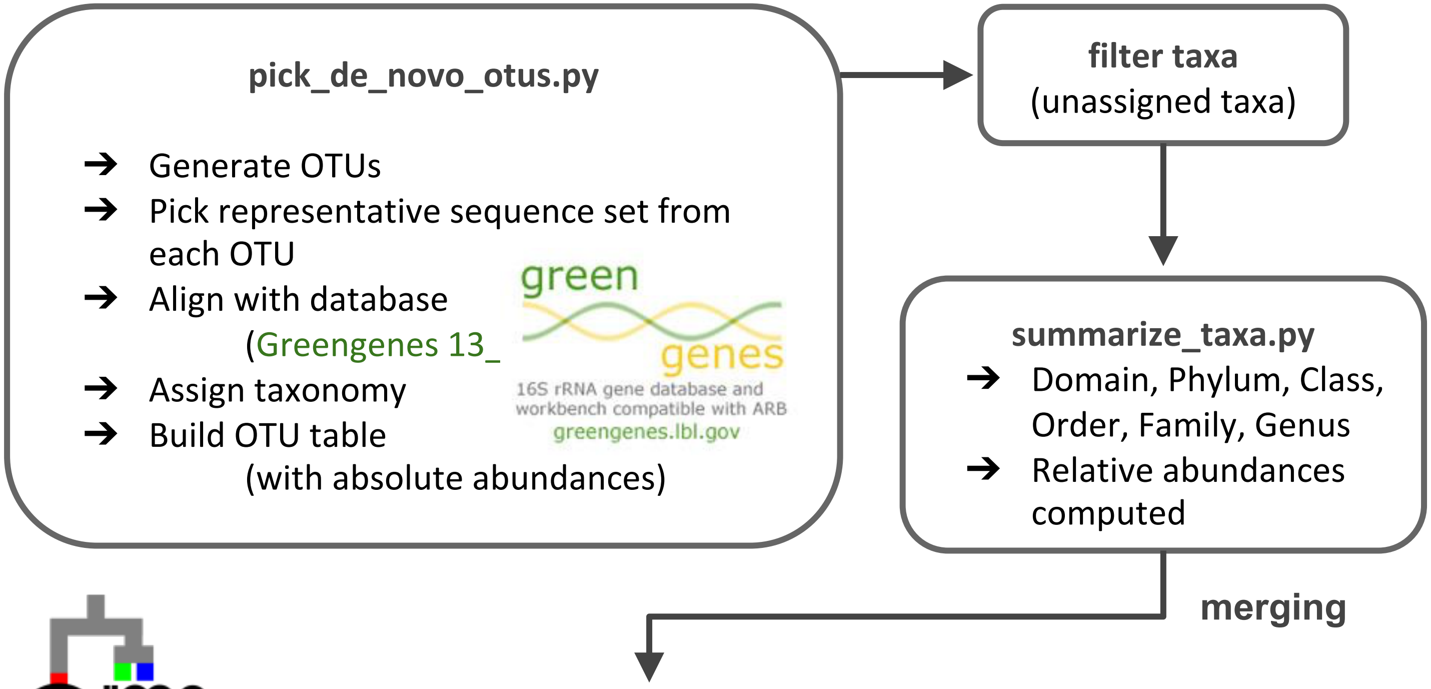
→

This can cause an otherwise mappable sequence not to align

Introns and primer sequence frequently flank the sequence of amplified exons. Unless removed by trimming, any of these artifacts will distort your sequence assembly and downstream sequence analysis.



Assigning Taxa



pick_de_novo_otus.py

- ➔ Generate OTUs
- ➔ Pick representative sequence set from each OTU
- ➔ Align with database (Greengenes 13_
- ➔ Assign taxonomy
- ➔ Build OTU table (with absolute abundances)

filter taxa
(unassigned taxa)

summarize_taxa.py

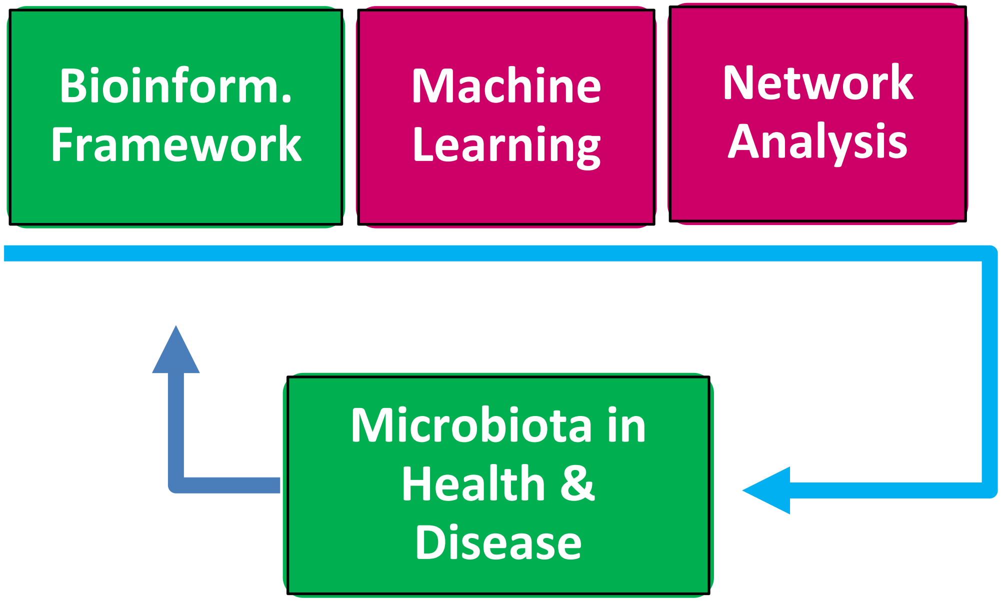
- ➔ Domain, Phylum, Class, Order, Family, Genus
- ➔ Relative abundances computed

merging

OTU table

sample	OTU1	OTU2	...
S01	120	42	...
S02	11	108	...
S03	0	49	...
...	...	...	...

Map/3



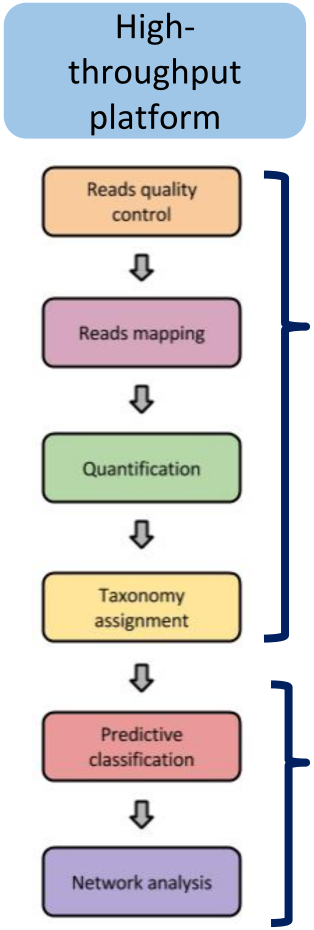
Bioinform. Framework

Machine Learning

Network Analysis

Microbiota in Health & Disease

Conceptual pipelines: meta-blocks



High-throughput platform

Reads quality control

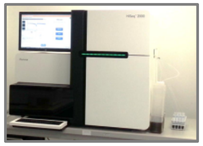
Reads mapping

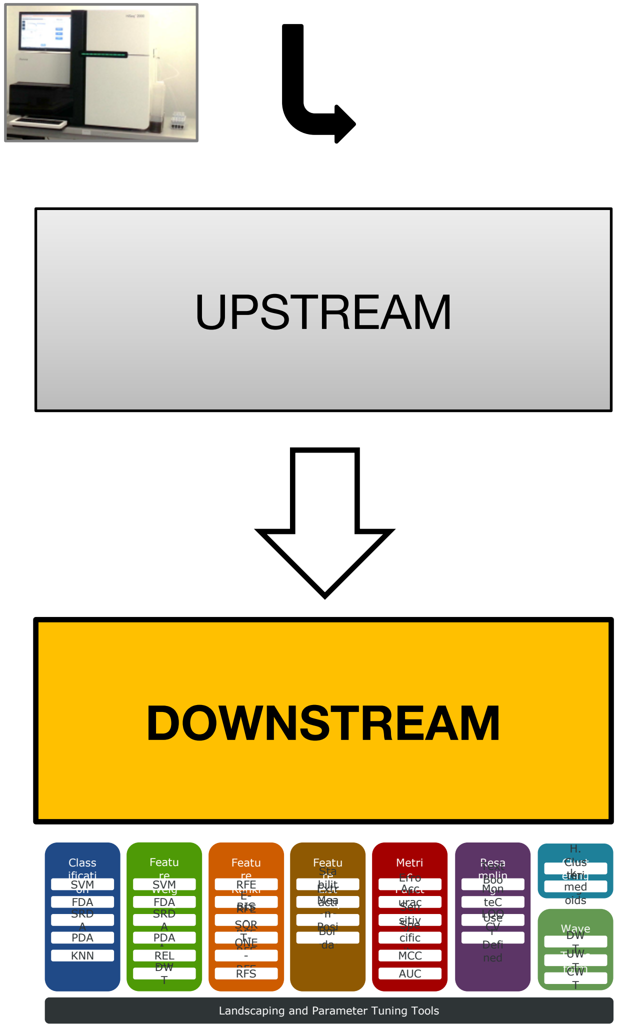
Quantification

Taxonomy assignment

Predictive classification

Network analysis





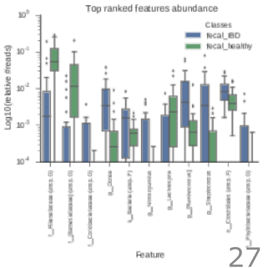
UPSTREAM

DOWNSTREAM

Landscaping and Parameter Tuning Tools

Data preparation
— QC
— preprocessing

"Sense-making"
— Machine learning
— Networks



The MAQC/SEQC initiatives

A set of guidelines for predictive profiling

***(*2014: for high-throughput sequencing with NGS)**

1. Predictive models can be derived from high-throughput data,
2. But they need to be carefully developed and independently tested
3. Reproducibility requires substantial effort.

nature biotechnology

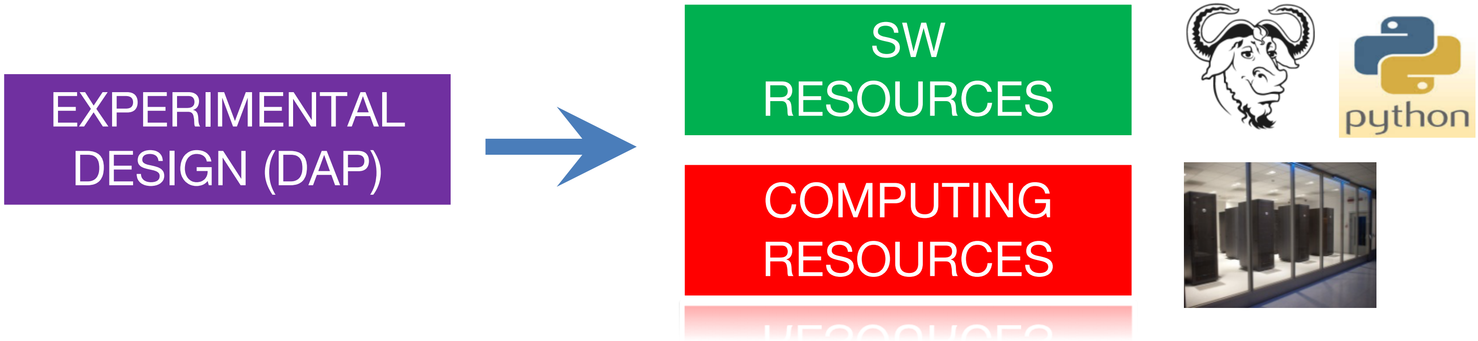
VOLUME 32 NUMBER 9 SEPTEMBER 2014
www.nature.com/naturebiotechnology

THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY

Focus on RNA sequencing quality control (SEQC)
ABRF evaluation of RNA-seq
Genome editing in hexaploid wheat

Need for Data Analysis Protocols

A **Data Analysis Protocol (DAP)** must be defined that details all the procedures used to develop the predictive classifiers, **including the data preprocessing**



EXPERIMENTAL DESIGN (DAP)

SW RESOURCES

COMPUTING RESOURCES

Reads quality control

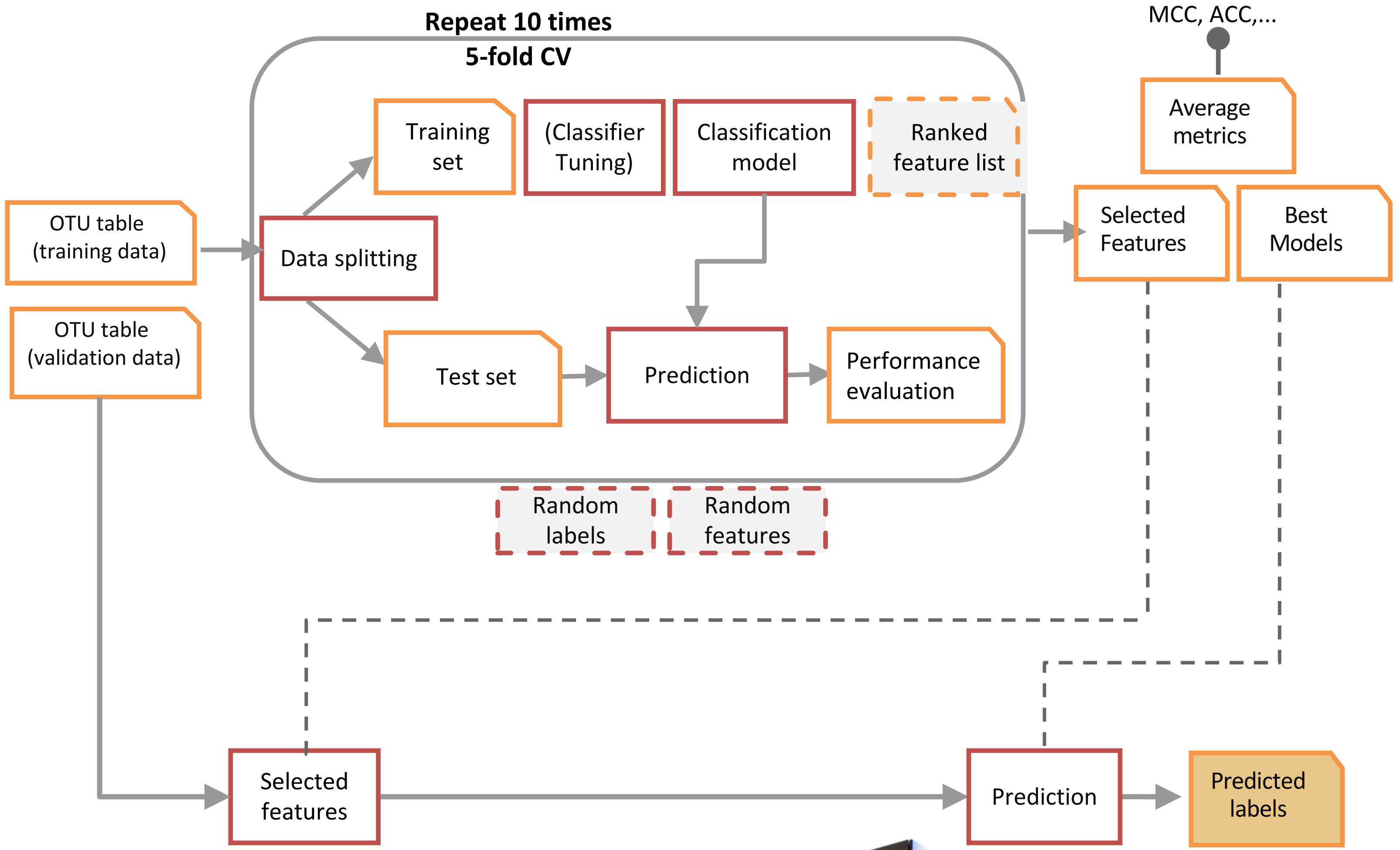
Reads mapping

Quantification

Taxonomy assignment

Predictive classification

Network analysis



Repeat 10 times

5-fold CV

MCC, ACC,...

OTU table (training data)

OTU table (validation data)

Data splitting

Training set

(Classifier Tuning)

Classification model

Ranked feature list

Test set

Prediction

Performance evaluation

Average metrics

Selected Features

Best Models

Random labels

Random features

Selected features

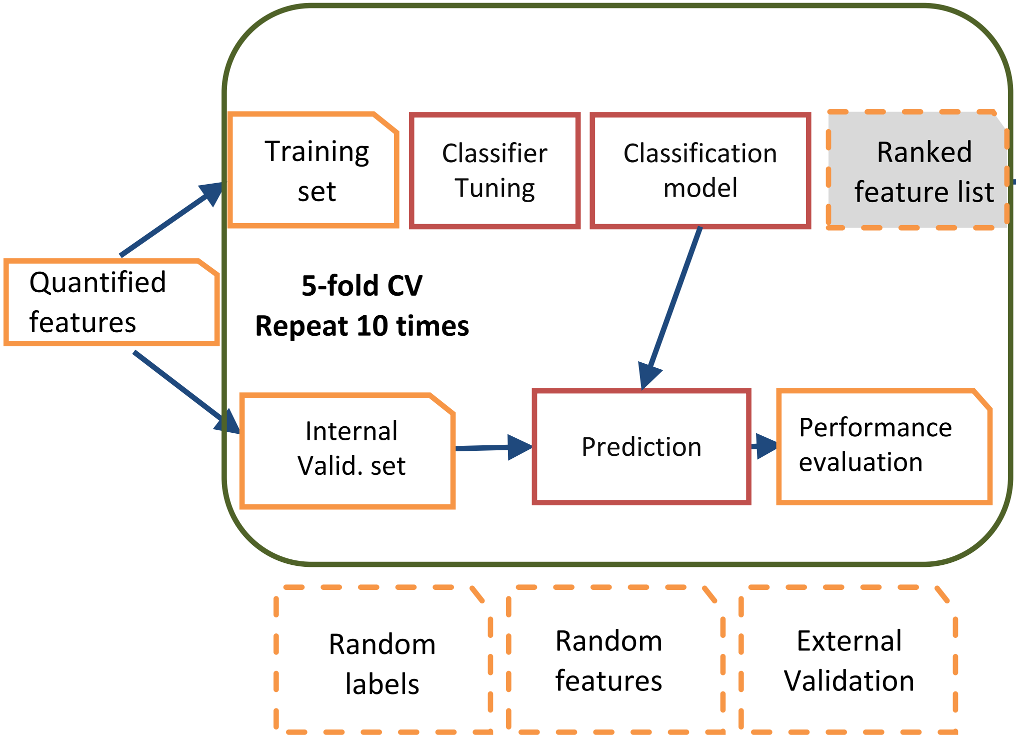
Prediction

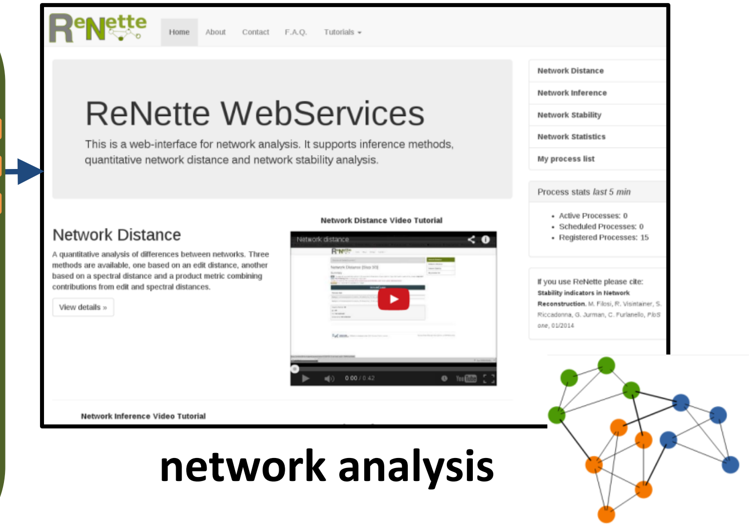
Predicted labels

DOWNSTREAM

Sense-making

A MAQC-II/SEQC Data Analysis Plan





ReNette

Home | About | Contact | F.A.Q. | Tutorials

ReNette WebServices

This is a web-interface for network analysis. It supports inference methods, quantitative network distance and network stability analysis.

Network Distance

A quantitative analysis of differences between networks. Three methods are available, one based on an edit distance, another based on a spectral distance and a product metric combining contributions from edit and spectral distances.

View details »

Network Distance Video Tutorial

Network Inference Video Tutorial

Network Distance
Network Inference
Network Stability
Network Statistics
My process list

Process stats *last 5 min*

- Active Processes: 0
- Scheduled Processes: 0
- Registered Processes: 15

If you use ReNette please cite: **Stability indicators in Network Reconstruction.** M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, C. Furlanello, *PloS one*, 01/2014

network analysis

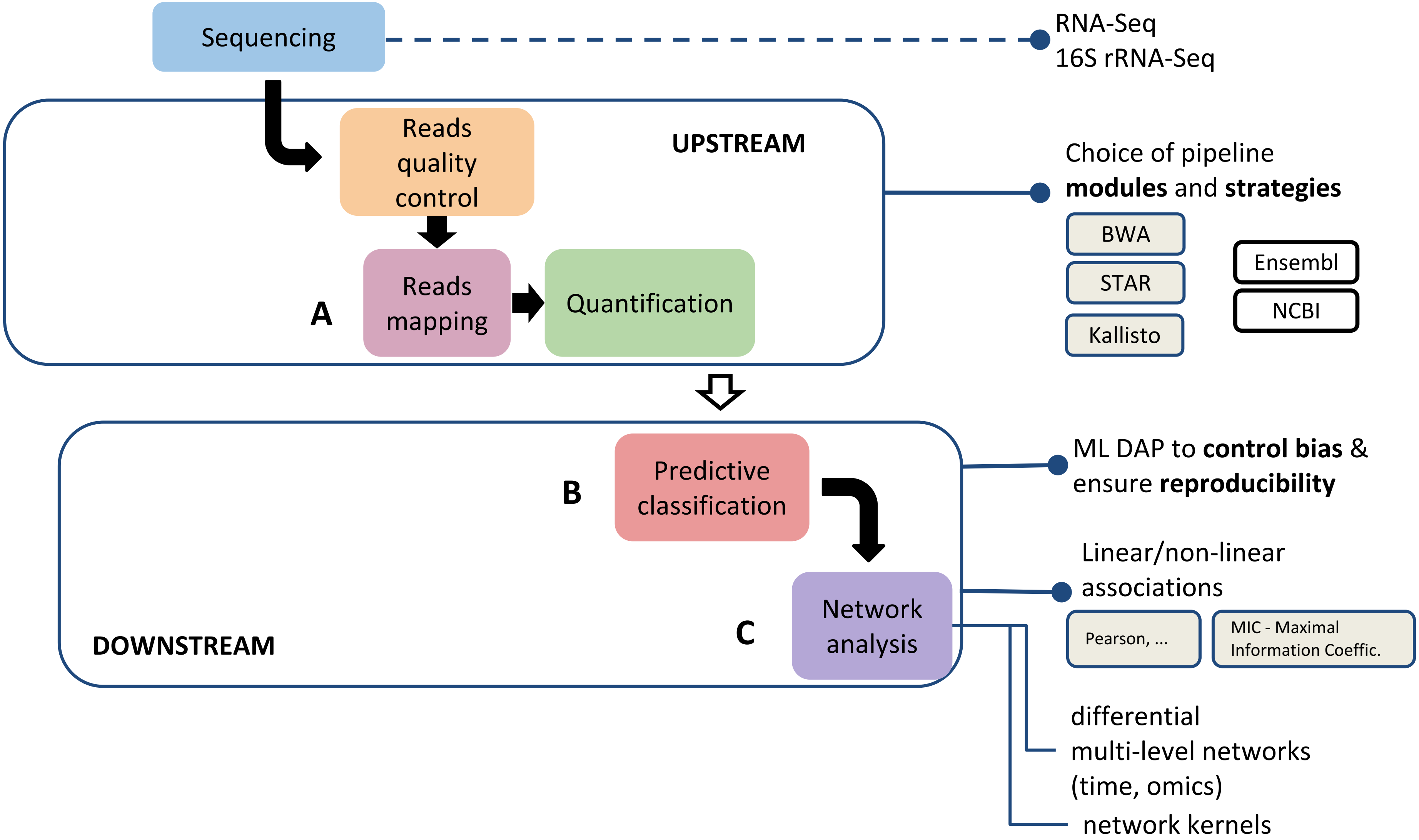
For network analysis of metagenomics data we apply **ReNette (**based on the **netTools R package)**

- Filosi M *et al.* **ReNette: a web-infrastructure for reproducible network analysis**. *bioRxiv*, Aug 2014
- Zandonà, et al **A metagenomic pipeline integrating predictive profiling methods and complex networks for the analysis of NGS microbiome data.** NIPS-MLCB Machine Learning in Computational Biology: Montreal, Dec 13, 2014

Used in

- Su Z *et al*. A comprehensive assessment of RNA-Seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotech*, 2014
- Wang C *et al*. The concordance between RNA-Seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotech*, 2014
- Zhang W *et al*. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, 2015

Summary of decisions/Challenges



Sequencing

RNA-Seq
16S rRNA-Seq

UPSTREAM

A

Reads quality control

Reads mapping

Quantification

Choice of pipeline **modules** and **strategies**

BWA

STAR

Kallisto

Ensembl

NCBI

DOWNSTREAM

B

Predictive classification

C

Network analysis

ML DAP to **control bias** & ensure **reproducibility**

Linear/non-linear associations

Pearson, …

MIC - Maximal Information Coeffic.

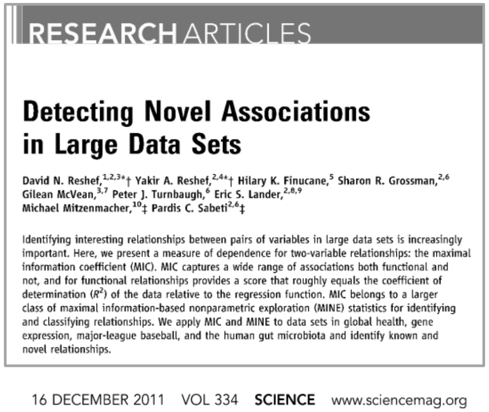
differential
multi-level networks
(time, omics)

network kernels

A. Characterization of the features of interest (e.g. transcriptome);
B: Identification of predictive biomarkers; **C**: Co-abundance networks inference and analysis

MINEPY

in metagenomics networks: a novel tool to quantify NON LINEAR ASSOCIATIONS between abundance of microbial taxa



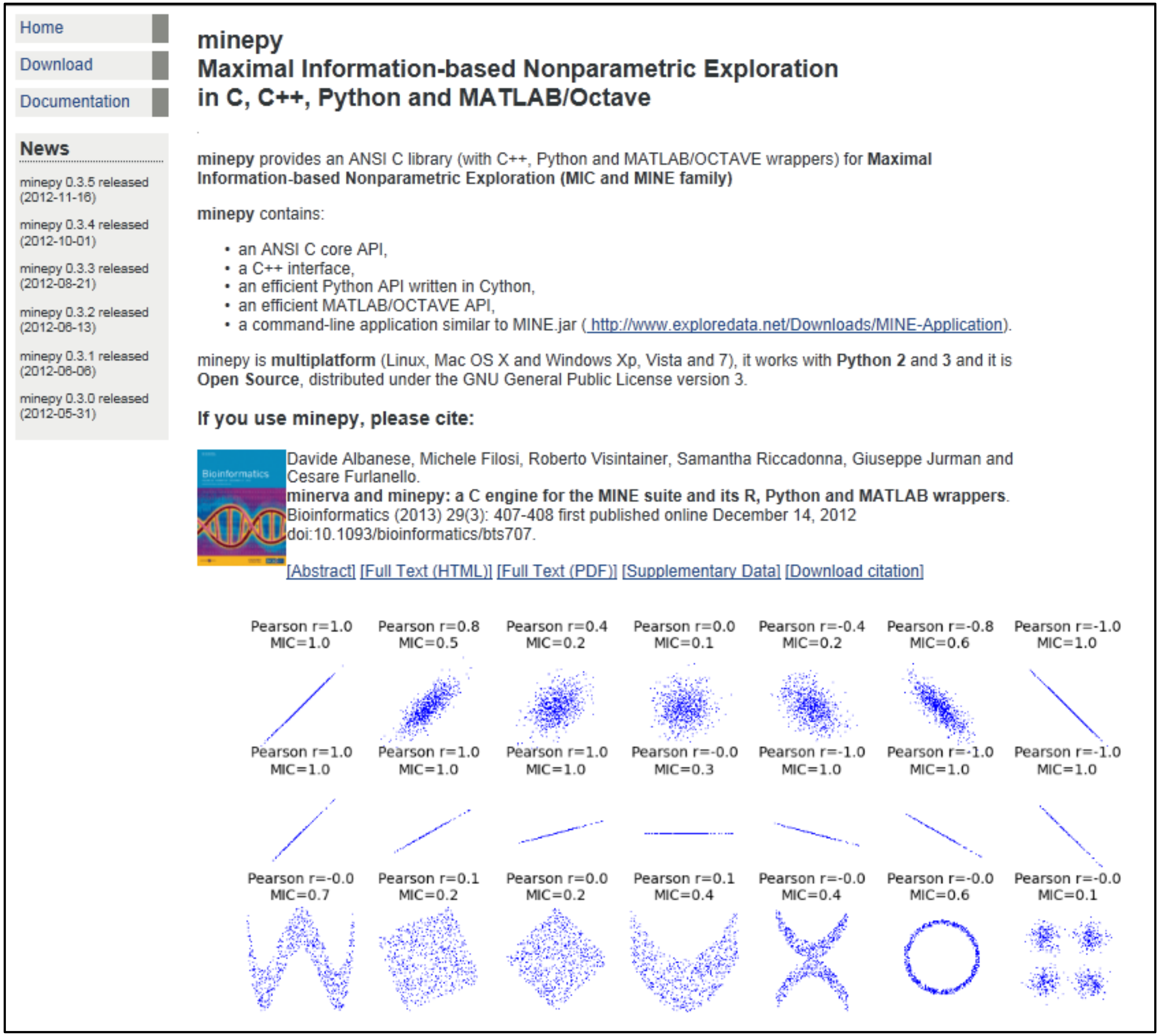
RESEARCH ARTICLES

Detecting Novel Associations in Large Data Sets

David N. Reshef,[1,2,3*†] Yakir A. Reshef,[2,4*†] Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[6] Eric S. Lander,[2,8,9] Michael Mitzenmacher,[10‡] Pardis C. Sabeti[2,6‡]

Identifying interesting relationships between pairs of variables in large data sets is increasingly important. Here, we present a measure of dependence for two-variable relationships: the maximal information coefficient (MIC). MIC captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships. We apply MIC and MINE to data sets in global health, gene expression, major-league baseball, and the human gut microbiota and identify known and novel relationships.

16 DECEMBER 2011 VOL 334 **SCIENCE** www.sciencemag.org



News

minepy 0.3.5 released (2012-11-16)

minepy 0.3.4 released (2012-10-01)

minepy 0.3.3 released (2012-08-21)

minepy 0.3.2 released (2012-06-13)

minepy 0.3.1 released (2012-06-06)

minepy 0.3.0 released (2012-05-31)

minepy
Maximal Information-based Nonparametric Exploration in C, C++, Python and MATLAB/Octave

minepy provides an ANSI C library (with C++, Python and MATLAB/OCTAVE wrappers) for **Maximal Information-based Nonparametric Exploration (MIC and MINE family)**

minepy contains:

- an ANSI C core API,
- a C++ interface,
- an efficient Python API written in Cython,
- an efficient MATLAB/OCTAVE API,
- a command-line application similar to MINE.jar (http://www.exploredata.net/Downloads/MINE-Application).

minepy is **multiplatform** (Linux, Mac OS X and Windows Xp, Vista and 7), it works with **Python 2** and **3** and it is **Open Source**, distributed under the GNU General Public License version 3.

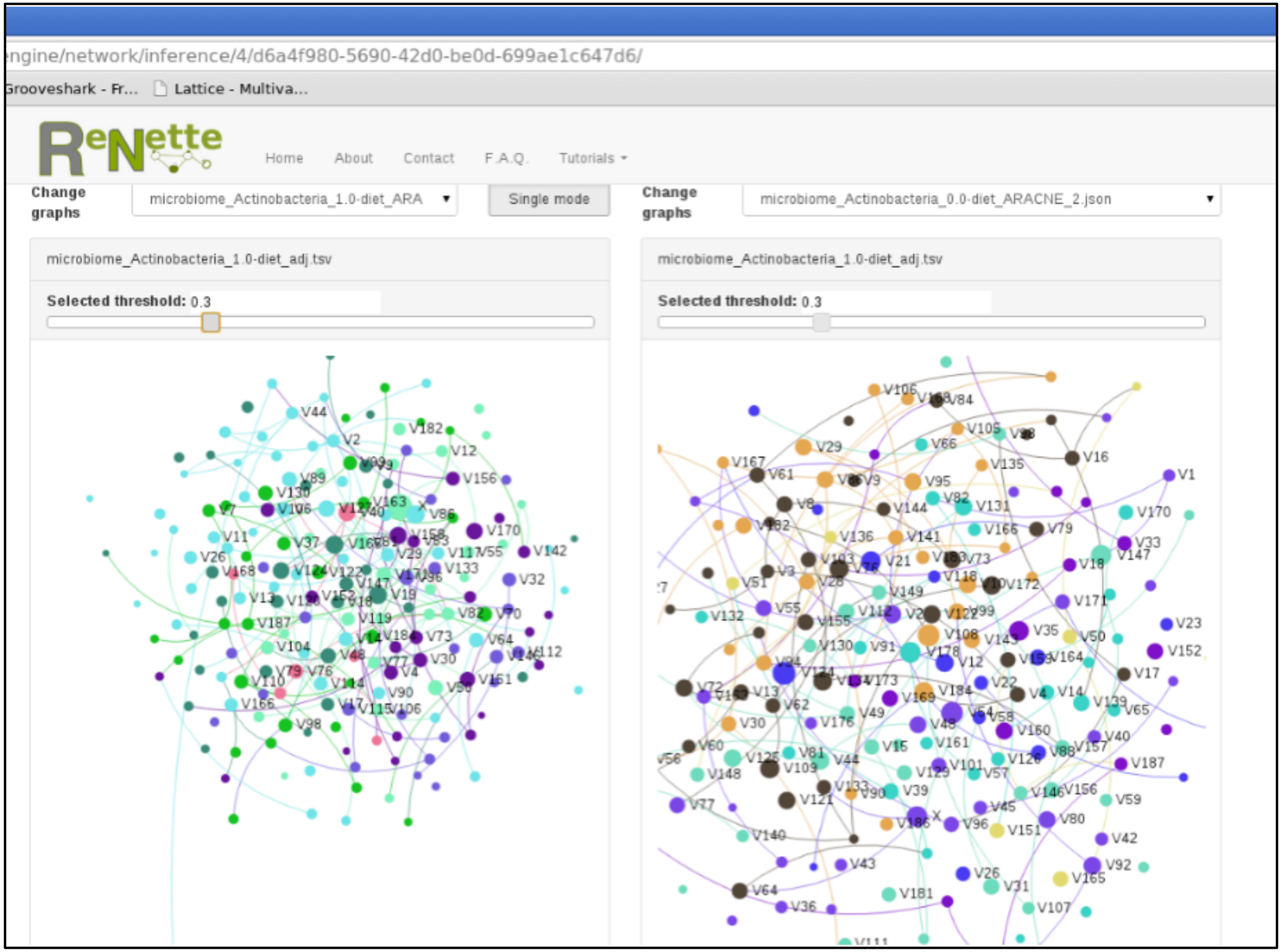
If you use minepy, please cite:

Davide Albanese, Michele Filosi, Roberto Visintainer, Samantha Riccadonna, Giuseppe Jurman and Cesare Furlanello.
minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers.
Bioinformatics (2013) 29(3): 407-408 first published online December 14, 2012
doi:10.1093/bioinformatics/bts707.

[Abstract] [Full Text (HTML)] [Full Text (PDF)] [Supplementary Data] [Download citation]

Albanese et al (Bioinformatics 2013): an open source implementation of MINE MINEPY (Python) , MINERVA (in R), also in MATLAB, Octave C++.

Microbiome: network differences



ngine/network/inference/4/d6a4f980-5690-42d0-be0d-699ae1c647d6/

Change graphs: microbiome_Actinobacteria_1.0-diet_ARA

Change graphs: microbiome_Actinobacteria_0.0-diet_ARACNE_2.json

microbiome_Actinobacteria_1.0-diet_adj.tsv

Selected threshold: 0.3

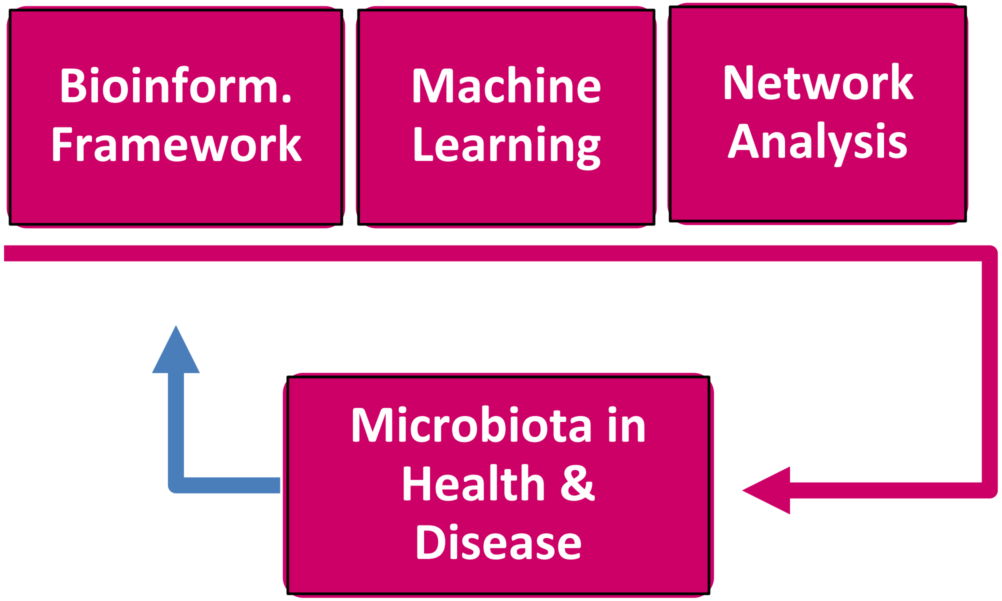
microbiome_Actinobacteria_1.0-diet_adj.tsv

Selected threshold: 0.3

The open source R package nettools and the dedicated web interface ReNette: a complete implementation of the stability indicators and HIM with different network inference methods (e.g. MIC)

FONDAZIONE BRUNO KESSLER

Map/4


Bioinform. Framework

Machine Learning

Network Analysis

Microbiota in Health & Disease

Example 1: Diet Induced Diversity

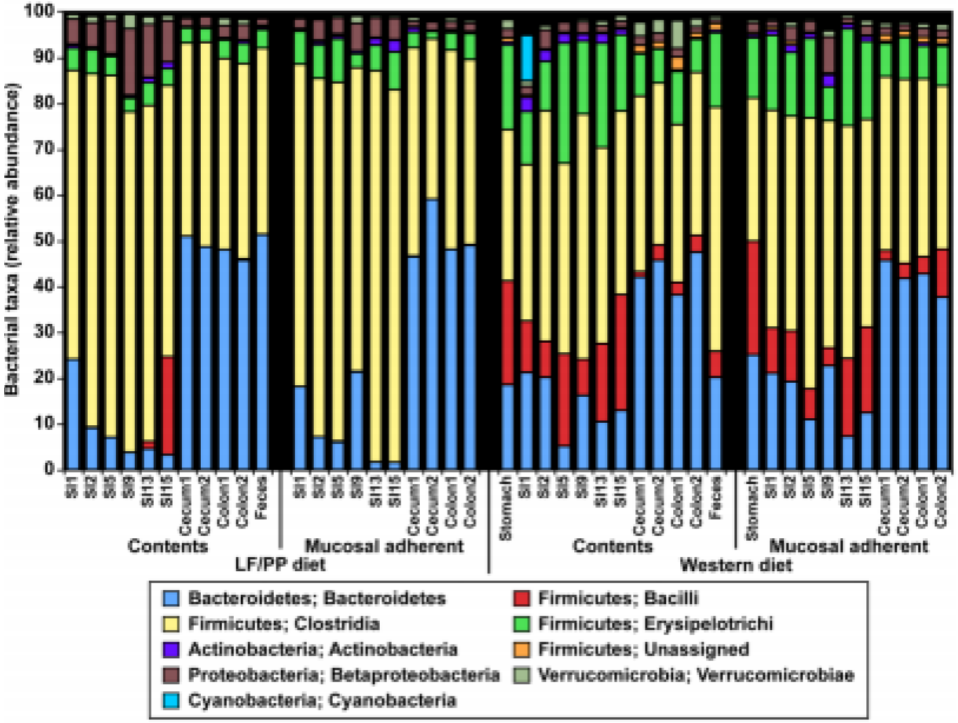
Diet induced diversity

“The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice.”

[Turnbaugh P et al, 2009]

- Illumina GA II gut microbiome 16S rRNA-seq
- 389 low-fat, plant polysaccharide-rich (LF) diet 269 high-fat, high-sugar (Western) diet

- TASK. Compare the network co-occurrence structure



Difference induced by diet: NETWORKS

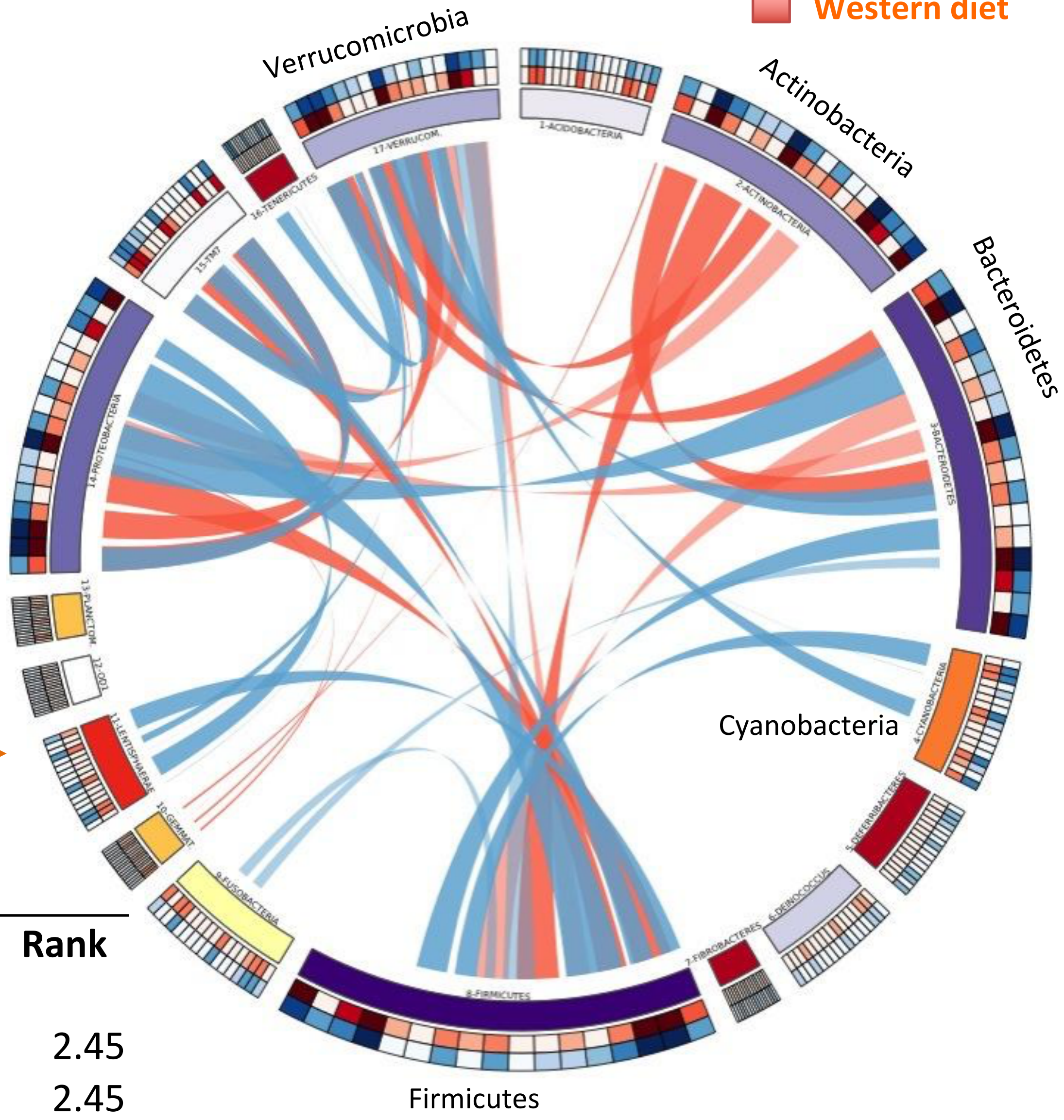
🟦 Low-fat diet

🟥 Western diet

- ONLY IN WESTERN DIET MICE
  Co-occurrence of Actinobacteria with Bacteroidetes, Firmicutes e Verrucomicrobia

- ONLY IN LOW-FAT DIET MICE
  Co-occurrence of Cyanobacteria with Firmicutes and Verrucomicrobia

- **Western vs LF wrt taxonomy**



Nodes = Phyla (~ OTU abbondance)
Weighted edges: non linear MIC assoc.

Top 5 discriminant nodes

Phylum	Western	LF	Total	Rank
Deferribacteres	1.60E-06	0	6.53E-07	2.45
Fibrobacteres	1.55E-06	0	6.32E-07	2.45
Tenericutes	1.25E-05	0	5.12E-06	2.45
Lentisphaerae	1.34E-05	8.76E-07	5.98E-06	2.09
Cyanobacteria	1.67E-05	1.66E-06	7.81E-06	1.93

FONDAZIONE
BRUNO KESSLER

Example 2

Gut microbiota and GI in children with Autism Spectrum Disorder

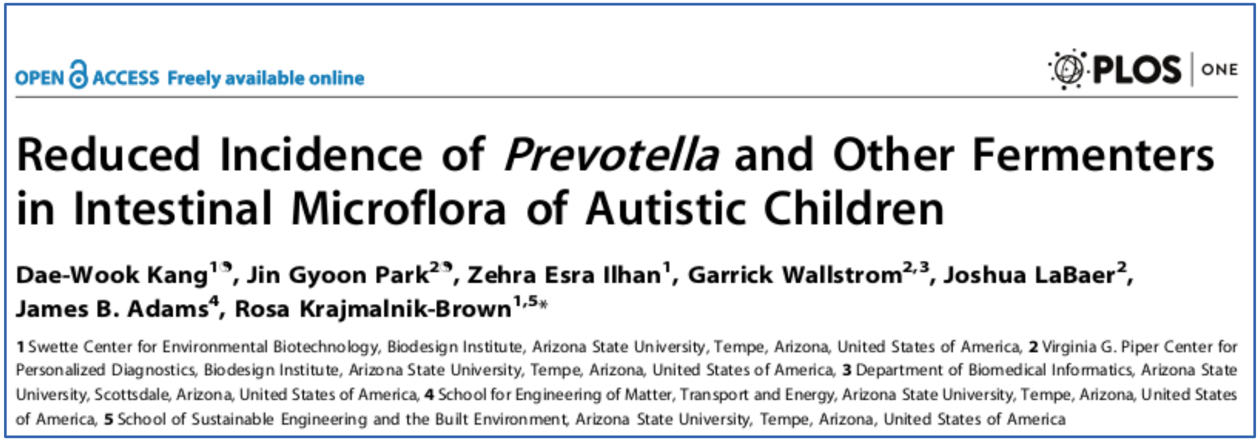
[Kang *et al*, 2013]

- Platform: Pyrosequencing 16S rDNA, Roche 454 FLX-Titanium
- Mean: 24 695 reads per sample per campione
- Bioinformatics Pipeline: FBK (taxa level: 712 species)
- **39 children (3-16 y) in 2 classes: 20 neurotypically developed, 19 ASD**

  ASD Phenotype: ADI-revised, ADOS, ATEC, PDD-BI

  GI: Gastro-Intestinal Severity Index, diet patterns survey*

  TASK. Marker characterizing autism and GI condition



OPEN ACCESS Freely available online

PLOS | ONE

Reduced Incidence of *Prevotella* and Other Fermenters in Intestinal Microflora of Autistic Children

Dae-Wook Kang[1¶], Jin Gyoon Park[2¶], Zehra Esra Ilhan[1], Garrick Wallstrom[2,3], Joshua LaBaer[2], James B. Adams[4], Rosa Krajmalnik-Brown[1,5*]

1 Swette Center for Environmental Biotechnology, Biodesign Institute, Arizona State University, Tempe, Arizona, United States of America, **2** Virginia G. Piper Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, Arizona, United States of America, **3** Department of Biomedical Informatics, Arizona State University, Scottsdale, Arizona, United States of America, **4** School for Engineering of Matter, Transport and Energy, Arizona State University, Tempe, Arizona, United States of America, **5** School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, Arizona, United States of America

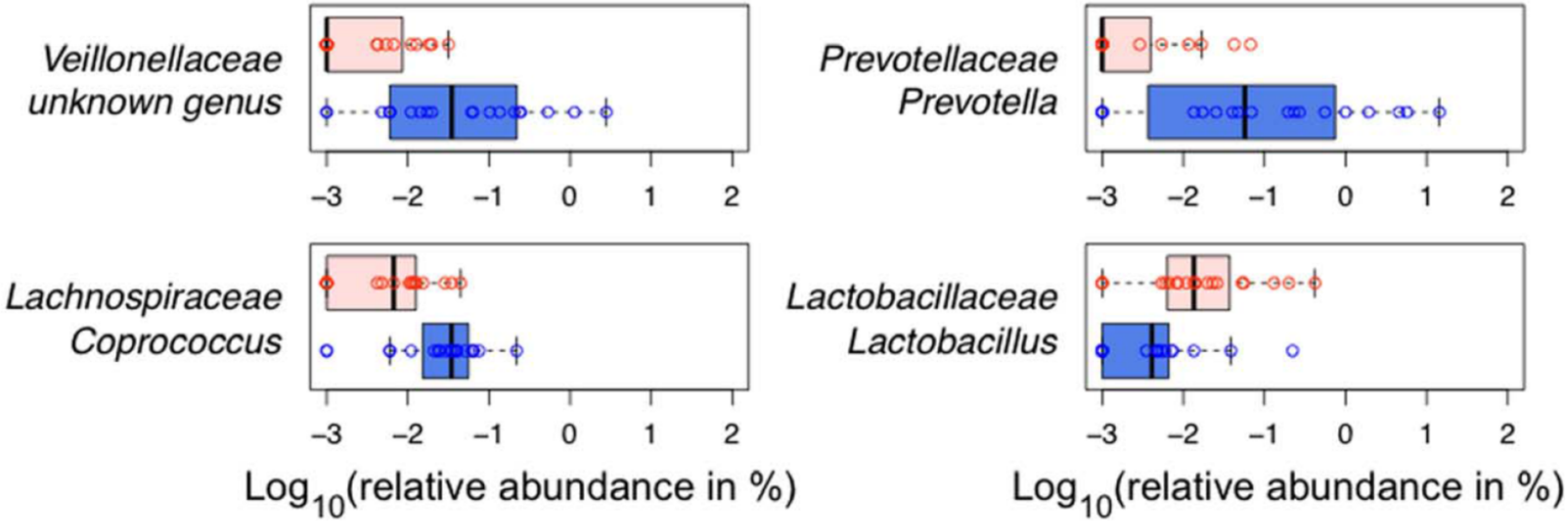
Results (Kang 2013)

Kang 2013

a. Limited association between 6-GSI score and severity of ASD

b. Difference in microbiome composition (richness, diversity)

c. Genus level: significant difference for 4 OTUs, specifically for *Prevotella*, confirmed with qPCR, also for subgenus



○ Autism

● Neurotipical

Results (FBK 2014)

a. Complete replication, from reads to biomarker extraction, based on the FDA/SEQC Data Analysis Plan: classifier Support Vector Machine*

Taxonomic level (NCBI, 340 genera-712 species), which after filtering 105 genus, 195 species
RISULTATI:
70 species: Acc 72% (CI 0.69-0.76), OR: 7.11, **with 3 sp in *Prevotellae***

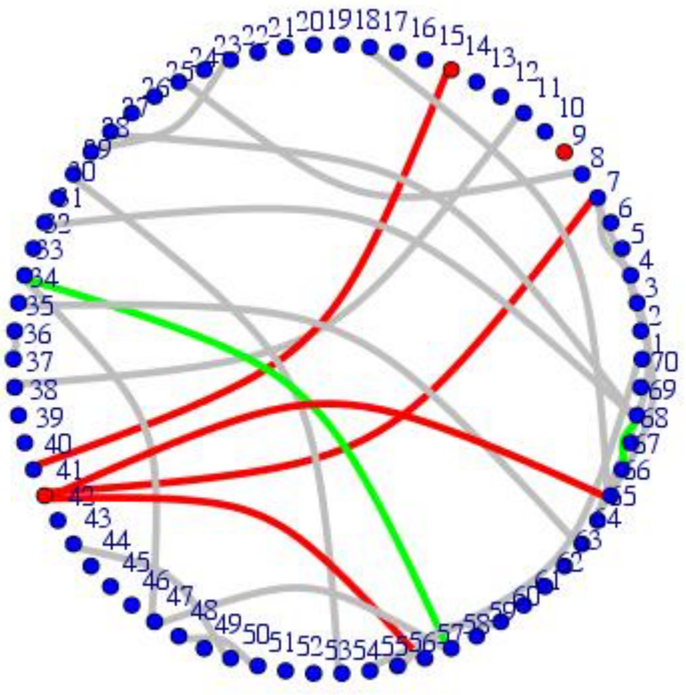
b. Top 70 OTUs then used to develop co-abundance networks
   - **For all OTU pairs : Pearson correlation on normalized number of reads (method: TMM-edgeR)**
   - **Consider separately neurotypical development and ASD cases**

GOAL: identify network difference

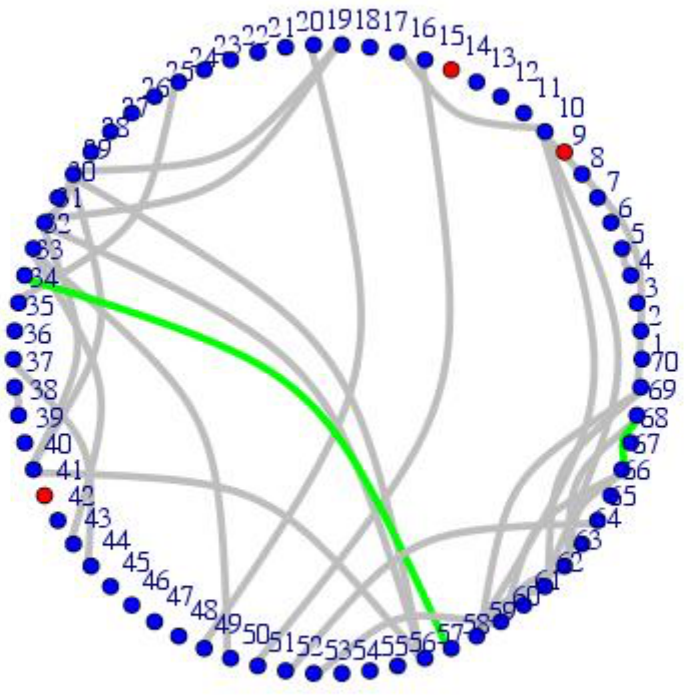
***** (SVM-L2R/L2loss dual),

FONDAZIONE BRUNO KESSLER

Network dysbiosis



Neurotipical



ASD

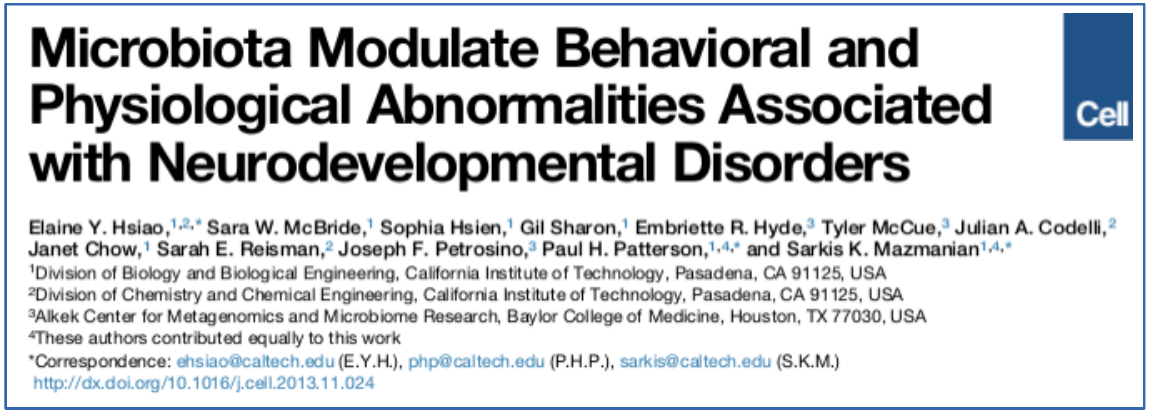
● OTU ● *Prevotellae*

🟥 Link: *Prevotellae* – altra OTU

🟩 conserved

⬜ non conserved

Microbiota & Behaviour

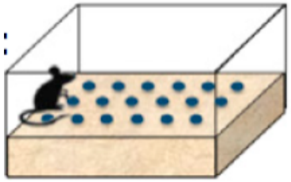


Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders

Elaine Y. Hsiao,[1,2,*] Sara W. McBride,[1] Sophia Hsien,[1] Gil Sharon,[1] Embriette R. Hyde,[3] Tyler McCue,[3] Julian A. Codelli,[2] Janet Chow,[1] Sarah E. Reisman,[2] Joseph F. Petrosino,[3] Paul H. Patterson,[1,4,*] and Sarkis K. Mazmanian[1,4,*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[3]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX 77030, USA
[4]These authors contributed equally to this work
*Correspondence: ehsiao@caltech.edu (E.Y.H.), php@caltech.edu (P.H.P.), sarkis@caltech.edu (S.K.M.)

http://dx.doi.org/10.1016/j.cell.2013.11.024



Hsiao *et al.*, 19 Dec, 2013

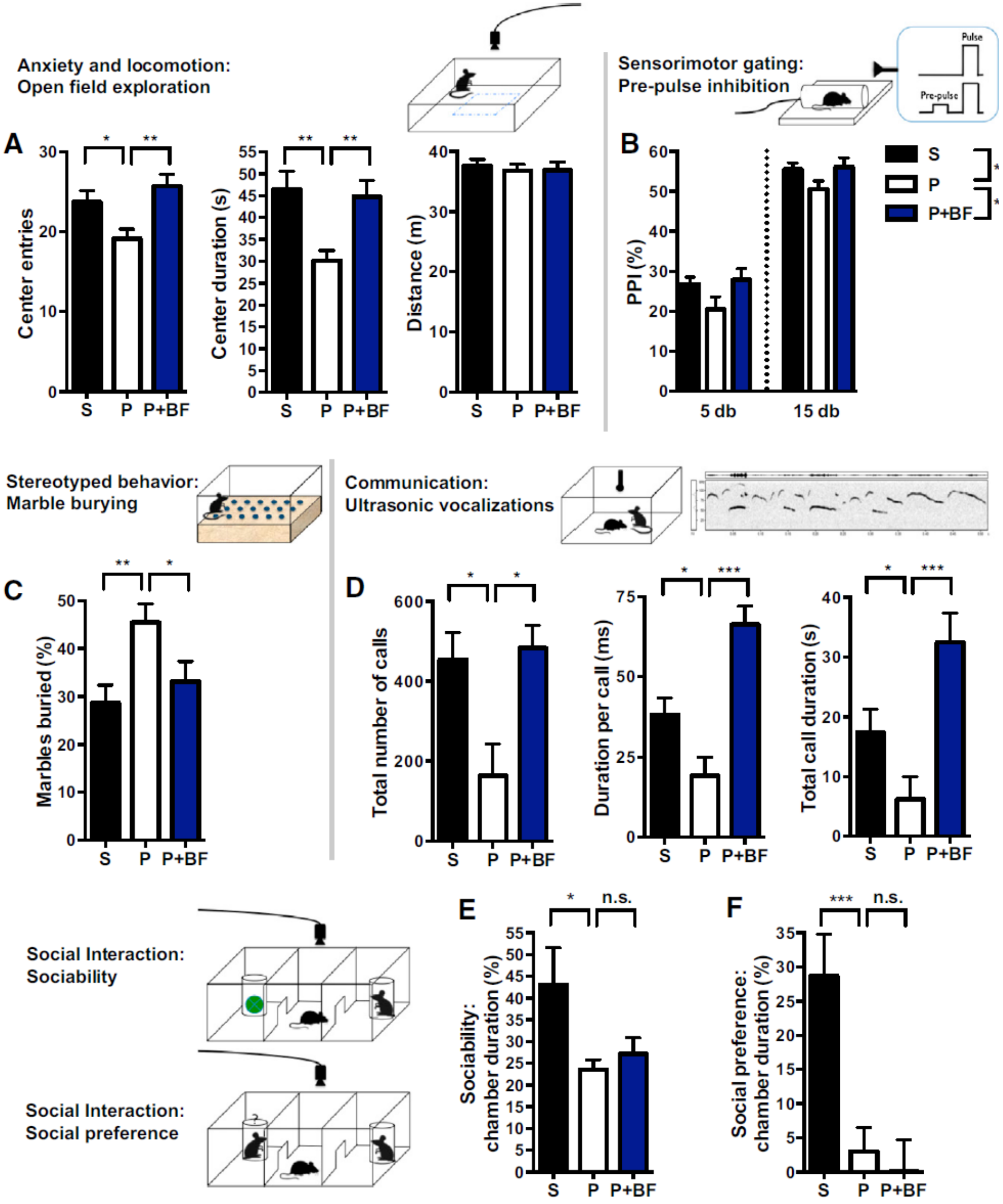
Mice: 30 sequenced on 16S rRNA - Roche 454-Titanium

10 subjects with maternal immune activation (MIA) exhibit atypical behaviours ASD-like (e.g. stereotypic, anxiety, reduced communication and socialization …) + GSI

1. **Microbiota is diverse from 10 mice fed with placebo**
2. ***Bacteroides fragilis* corrects the behavioural trait** (10 MIA treated)

B. *fragilis*

1. Improves gut barrier integrity
2. Corrects species level abnormalities
3. Ameliorates autism-related behavioral abnormalities in MIA offspring



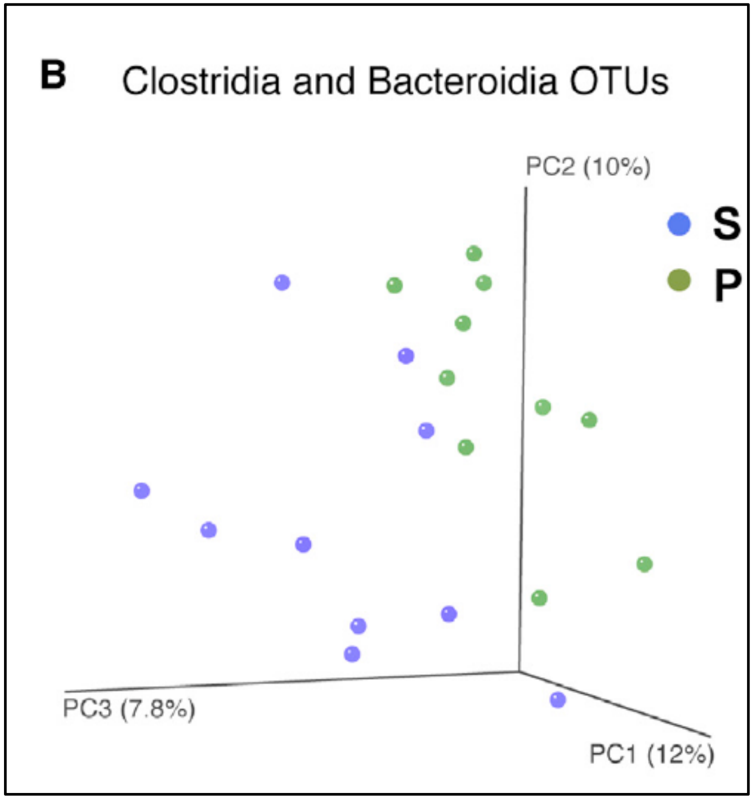
Hsiao *et al.*, 19 Dec, 2013

Results (Hsiao et al 2013)

a. Limited diversity differences between MIA or control adults

b. Significant philogenetic distance between microbic communities: OUT structure change is the main drivers of difference

c. 1474 OTUs identified, of which 67 discriminants (19+ controls, 48 MIA+), with alteration in OTU mixtures for Bacteroidia and Clostridia classes



Results (FBK)

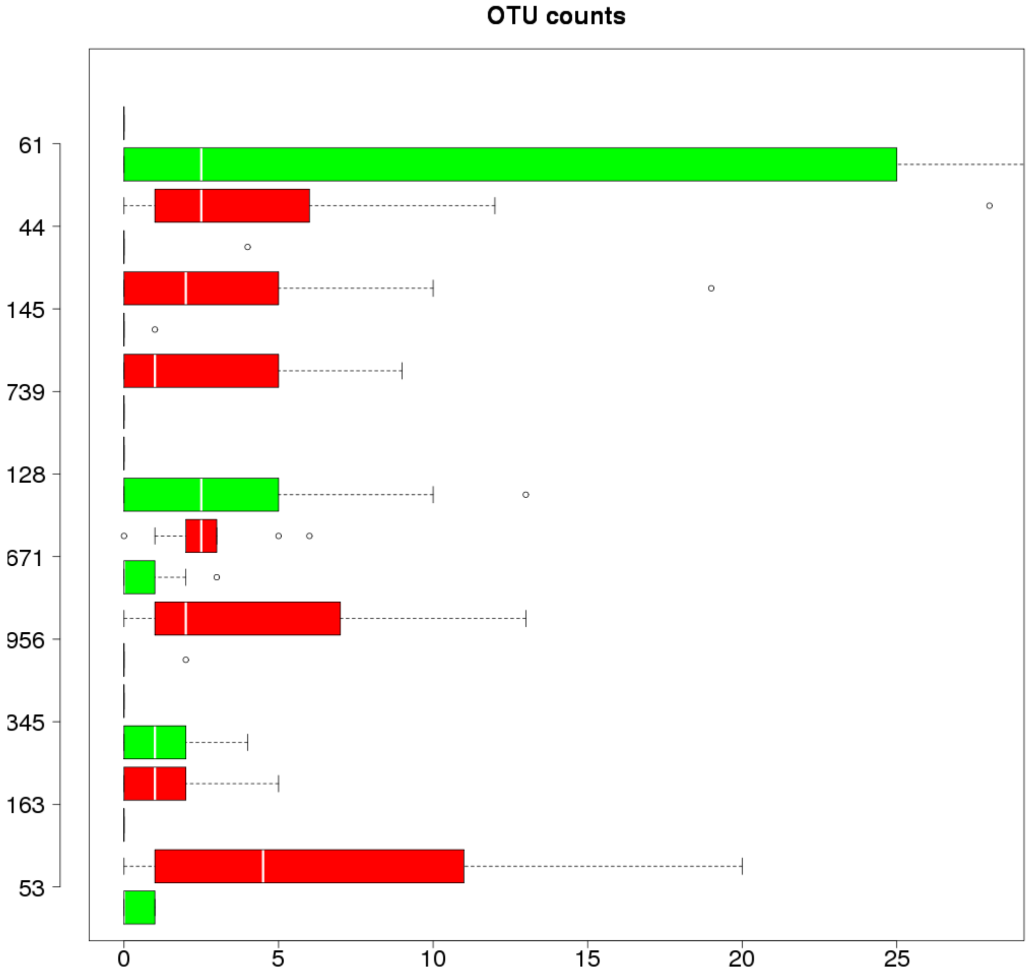
a. Analysis on 1474 OTUs (Hsiao 2013): after filtering: 351 OTUs

b. Data Analysis plan from FDA/SEQC, with SVM-L2R/L2loss dual

c. **RESULTS:**
10 OTUs: Acc 93% (CI 0.89-0.97),
OR > 100

NB: our top 10 markers are discriminants in Hsiao 2013

OTU classes:
Erysipelotrichi: 61
Bacteroidia: 44, 739, 671
Clostridiae: 145, 128, 956, 345, 53



IBD OPBG clinical dataset*

TRACKING GUT MICROBIOTA DYSBIOSIS AND HOST RESPONSE TO PREVENT IBD AND IBS THROUGHOUT LIFE

Objectives of the bioinformatics analysis:

1. Identification of **omics markers** as IBD/IBS predictors
2. Development of a dysbiosis scale useful to stratify the risk for IBS/IBD.

Outcomes (for clinical tests):

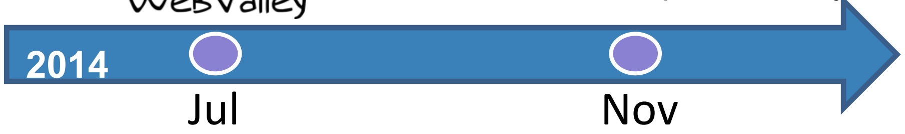
1. New laboratory tests for IBD and IBS (biomarkers)
2. Evaluation of the different staging of the dysbiosis status (risk factor)
3. Support to intervention protocols

Pyrosequencing:
barcoded pyrosequencing **V1-V3** regions of the 16S rRNA gene (amplicon size 520 bp) on GS Junior platform (**Roche 454**)

DATASET 1:
- Fecal IBD/healthy
- Paired biopsies IBD/ctrl

WebValley

DATASET 2:
- Biopsies healthy

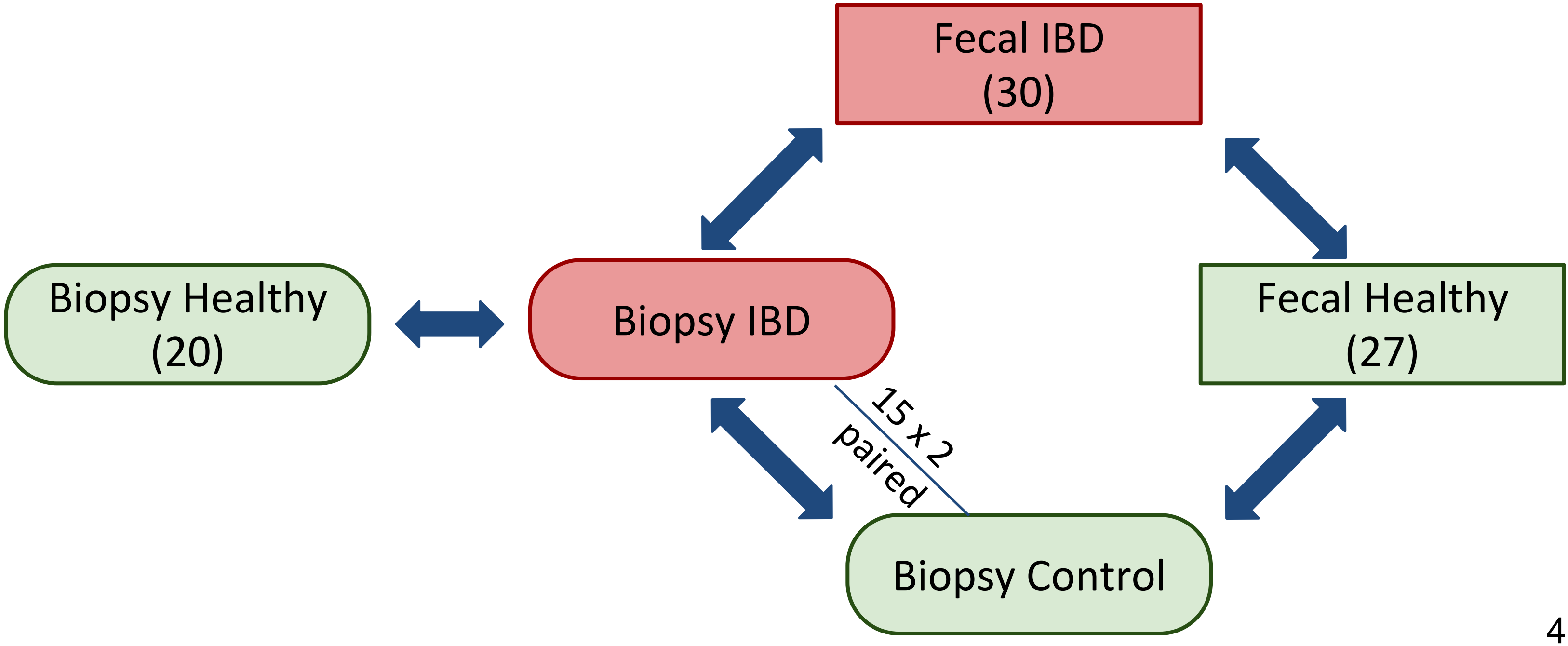


*CREDITS:
- OPBG (**Lorenza Putignani**)
- Dip. Univ. Pediatria e Neuropsichiatria Infantile, Sapienza Università di Roma (**S. Cucchiara**)

IBD OPBG clinical dataset

Roche 454 GS Junior gut microbiome 16S rRNA-Seq

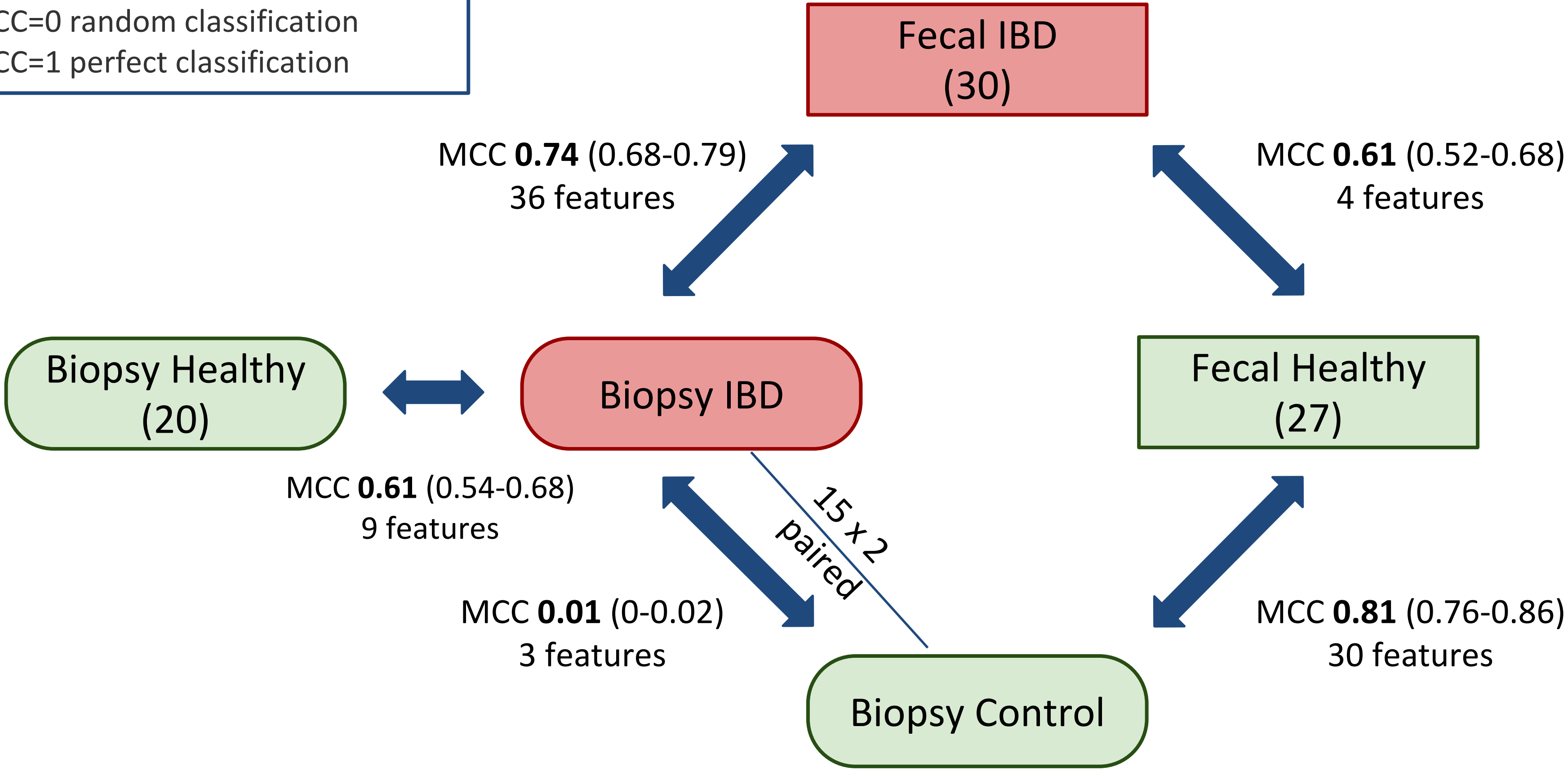
- **30 IBD** vs **27 healthy** children (fecal samples)
- **15 paired** (inflamed/control) **biopsies** from colon
- **20** colon **biopsies** from **healthy** individuals
- Age: 4 -19 years old



IBD Classification models

Matthews correlation coefficient (**MCC**): Indicator of predictive performance

MCC=0 random classification
MCC=1 perfect classification



Fecal IBD (30)

MCC **0.74** (0.68-0.79)
36 features

MCC **0.61** (0.52-0.68)
4 features

Biopsy Healthy (20)

Biopsy IBD

Fecal Healthy (27)

MCC **0.61** (0.54-0.68)
9 features

15 x 2 paired

MCC **0.01** (0-0.02)
3 features

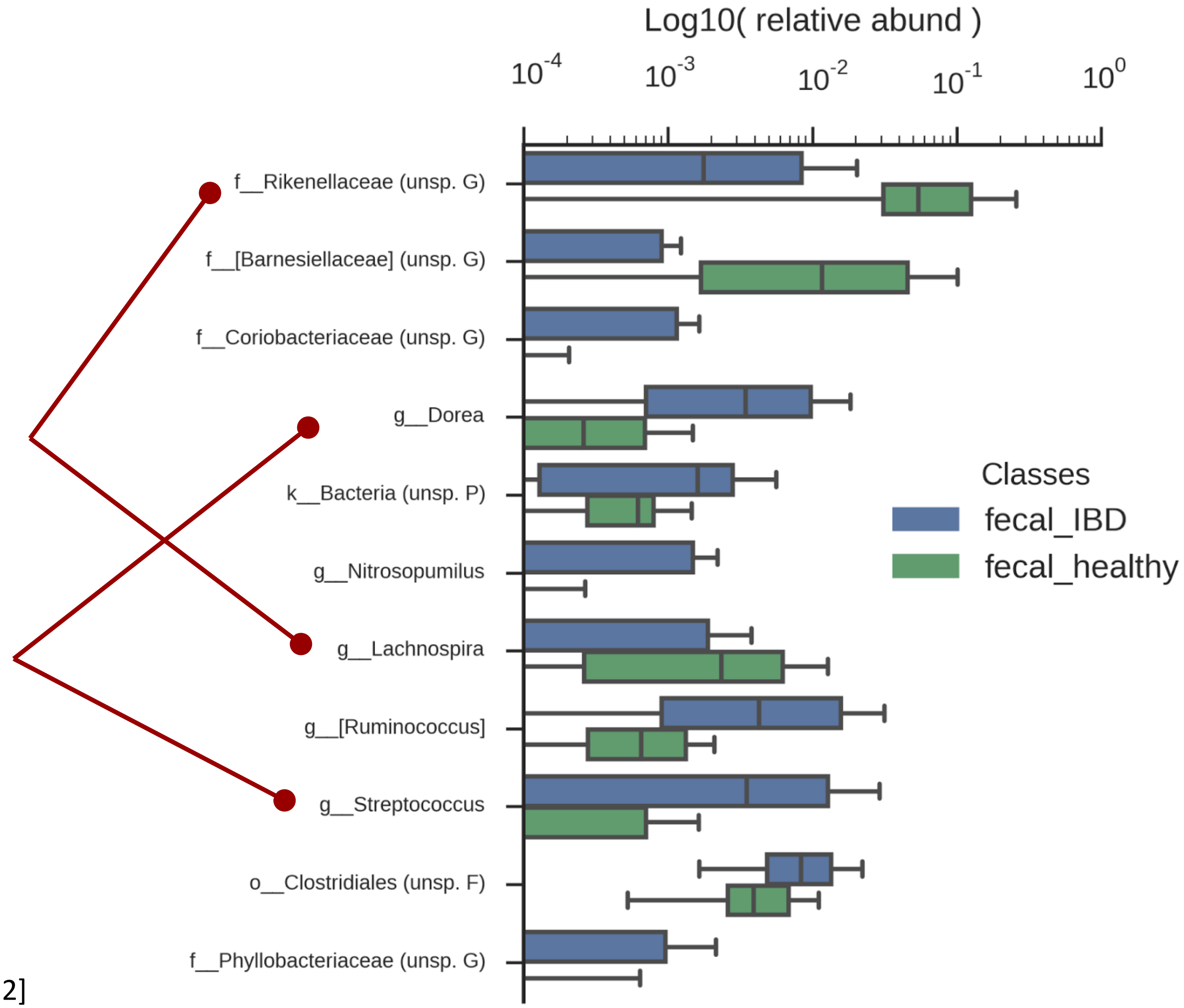
MCC **0.81** (0.76-0.86)
30 features

Biopsy Control

Top discriminant features

Papa et al, 2012

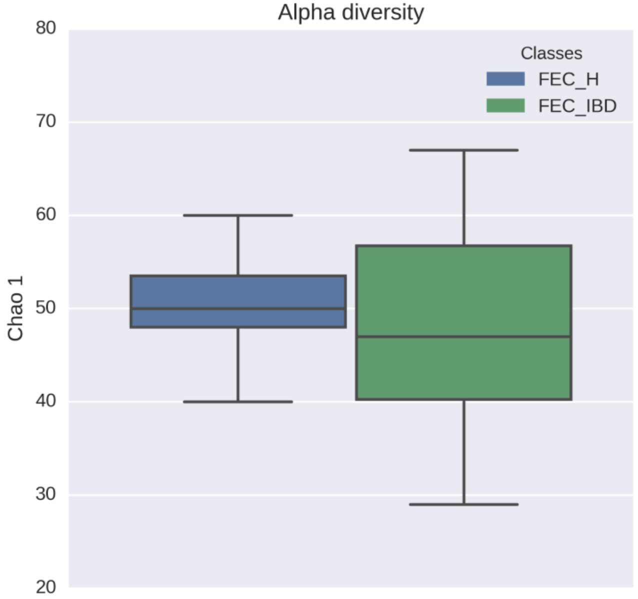
Gevers et al, 2014



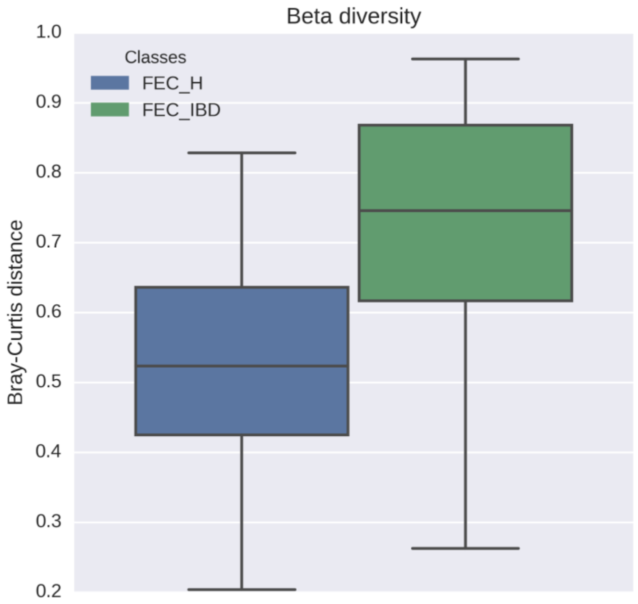
[Gevers et al, *Cell Host & Microbe*, 2014]

[Papa et al, *PLoS ONE*, 2012]

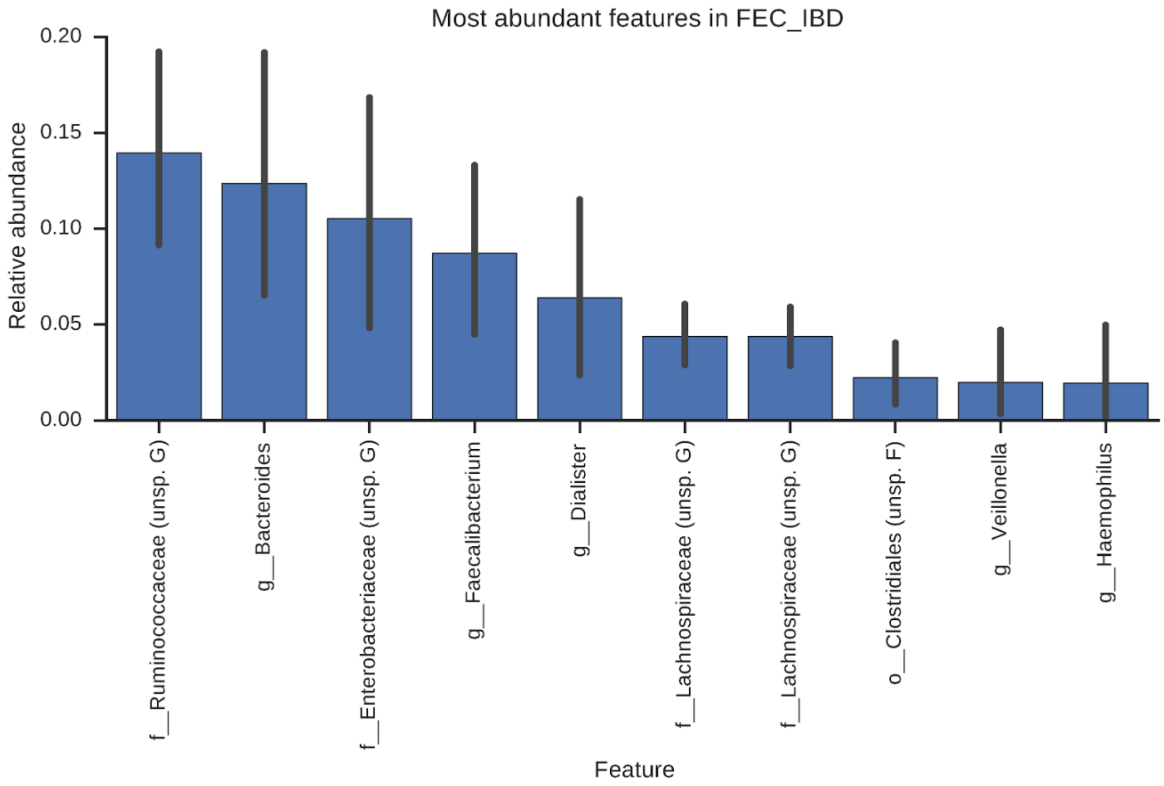
Microbiome characterization

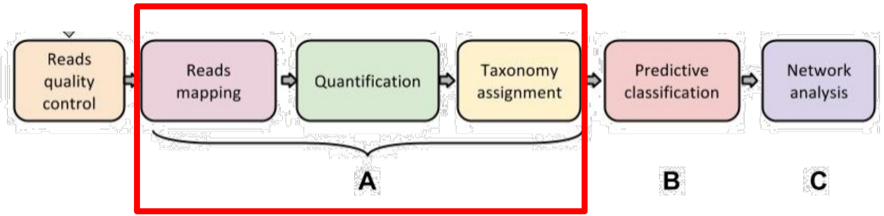

Alpha diversity

P-value = 0.3534

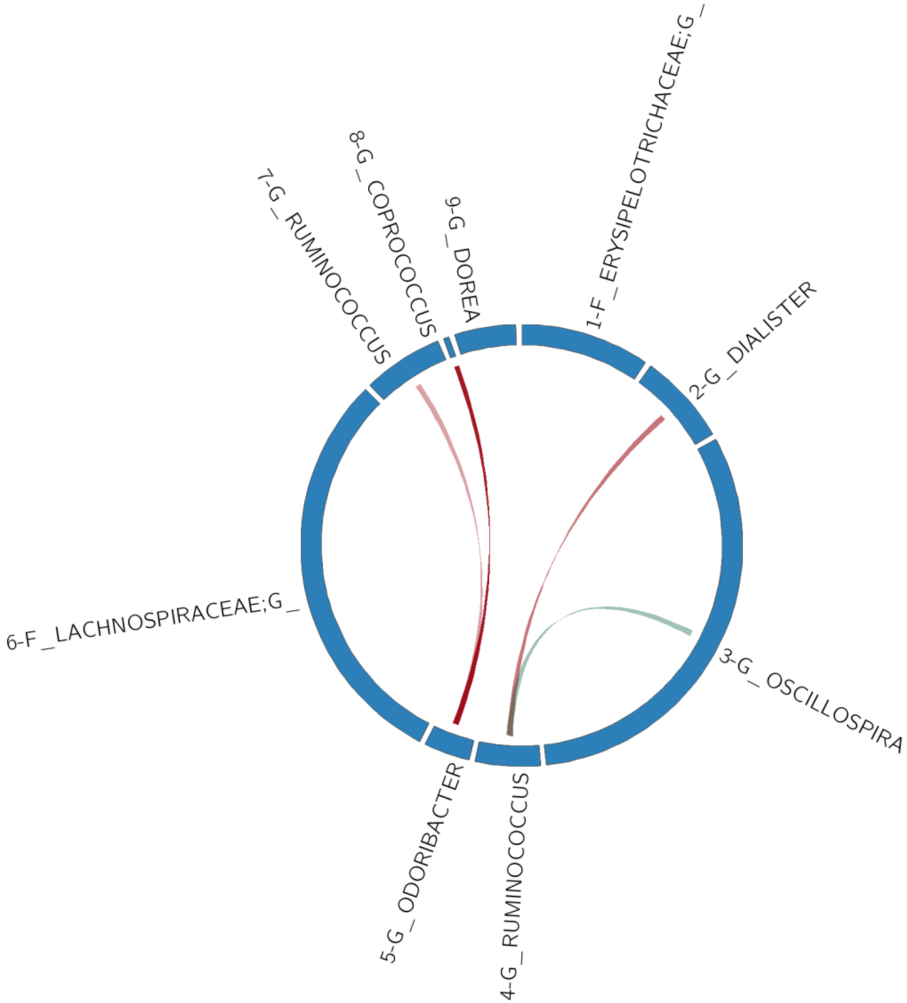

Beta diversity

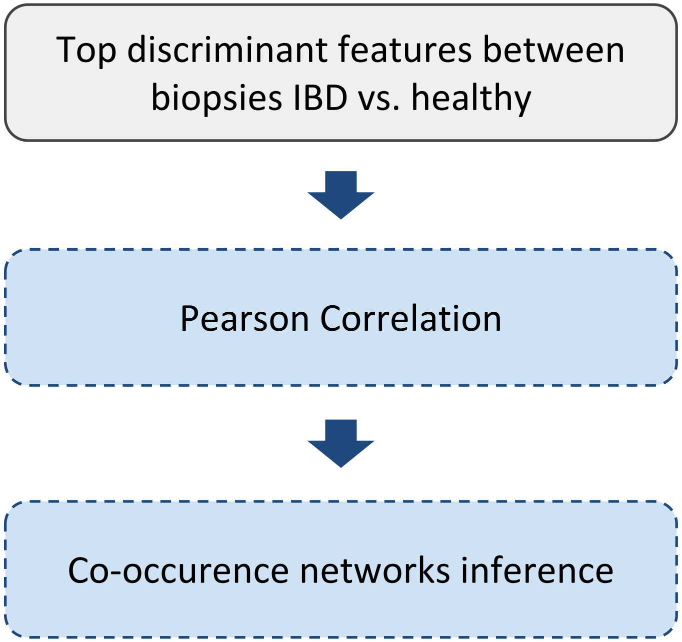
P-value = 1.053e-50


Most abundant features in FEC_IBD



Networks: IBD vs. healthy




Top discriminant features between biopsies IBD vs. healthy

Pearson Correlation

Co-occurence networks inference

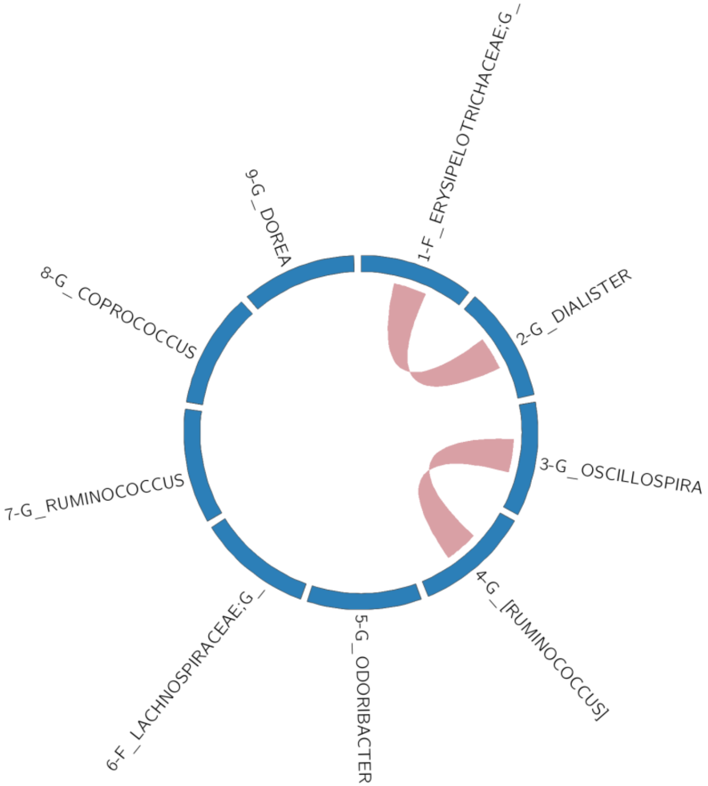
🟩 Links conserved in healthy only

🟥 Links conserved in IBD only

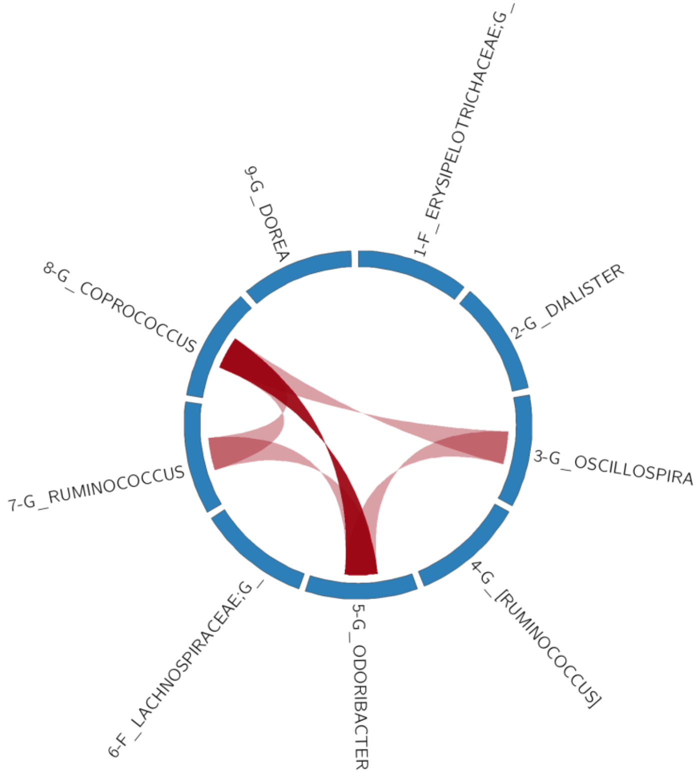
Edge's color intensity ∝ absolute value of Pearson correlation coefficient (PCC)

Shown links with PCC >0.65

Calprotectin level is associated to increasing dysbiosis in Biopsy Networks



10-24 [mg/kg]

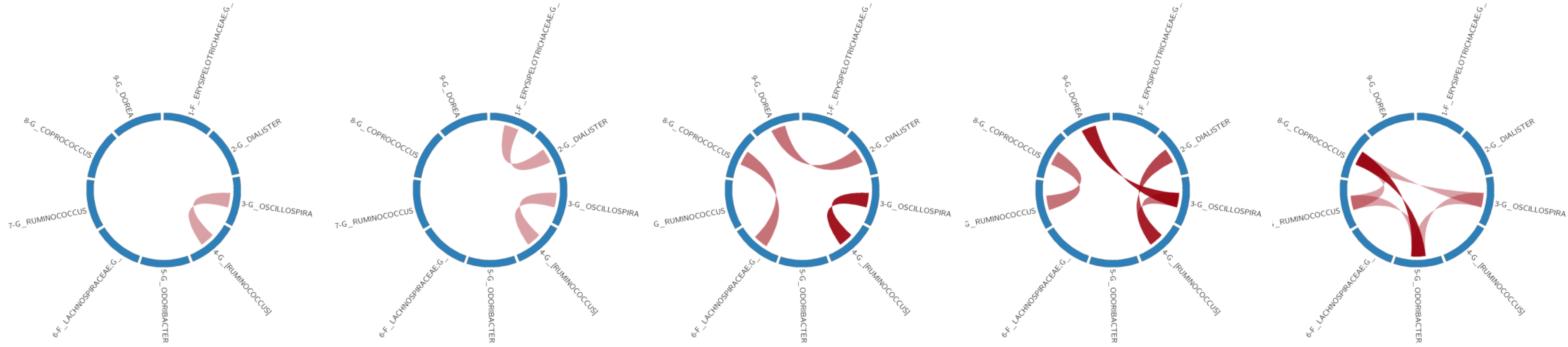


124-370 [mg/kg]

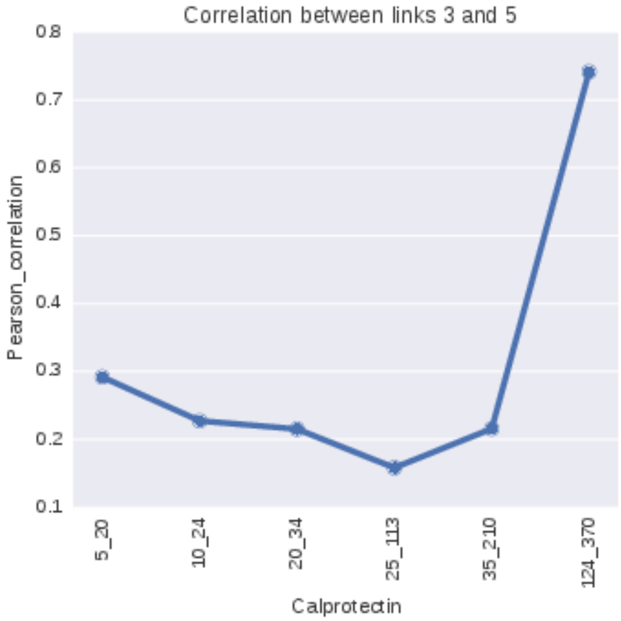
Healthy : Calprotectin < 50 mg/kg

Phenotype: Cucchiara Lab - Rome

Calprotectin level is associated to increasing dysbiosis in Biopsy Networks

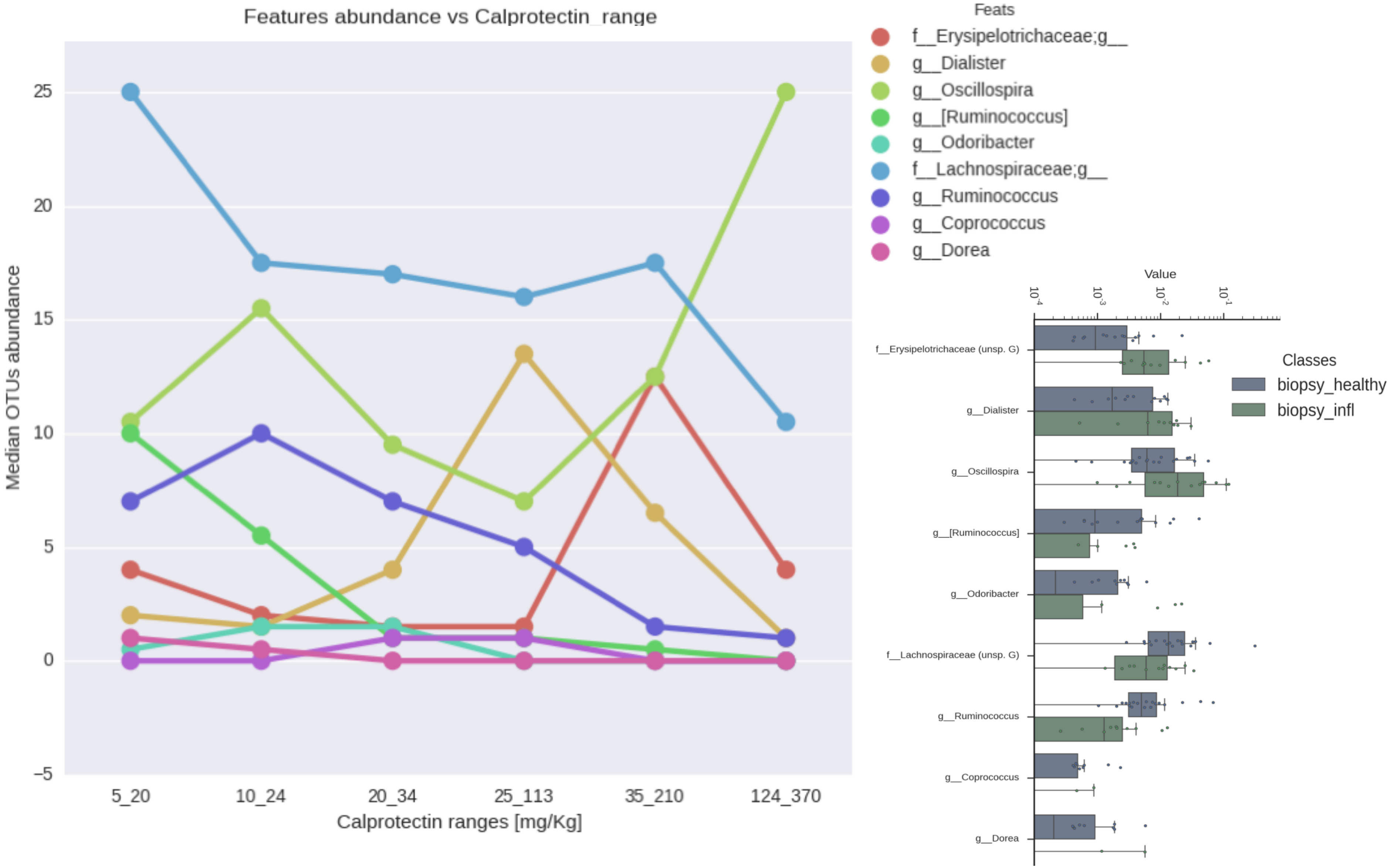


5-20 10-24 20-34 25-113 124-370



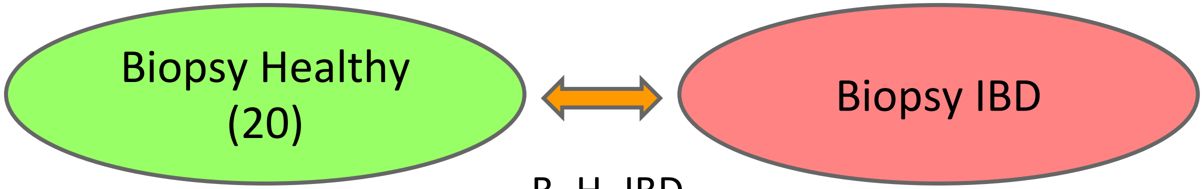
1. Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae
2. Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister
3. Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira
4. Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus
5. Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Odoribacteraceae;g_Odoribacter
6. Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_
7. Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus
8. Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus
9. Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea

Markers patterns vs Calprotectin



Biopsy IBD Networks

Co-occurrence nets for Pearson Correlation, for stronger links only (PCC > 0.5), taxonomic assignment 6 levels deep: 20% presence filter > 3510 OTUs table led to 168 OTUs,

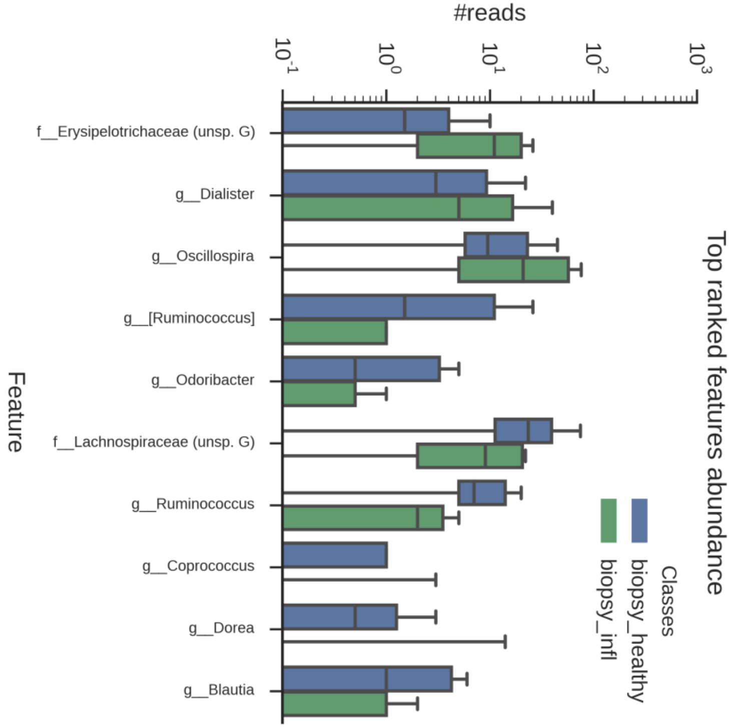

Biopsy Healthy (20)

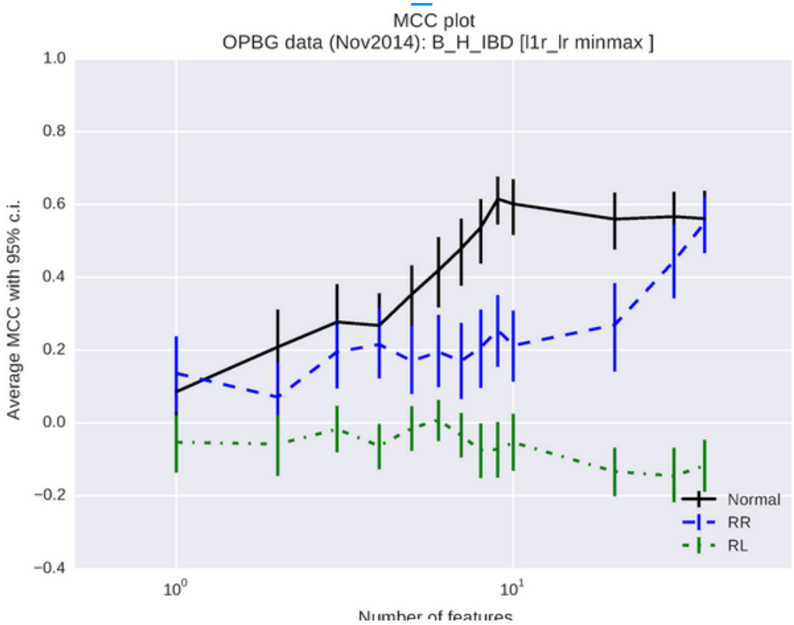
Biopsy IBD

B_H_IBD

MCC **0.61 (0.54-0.68**)

9 features





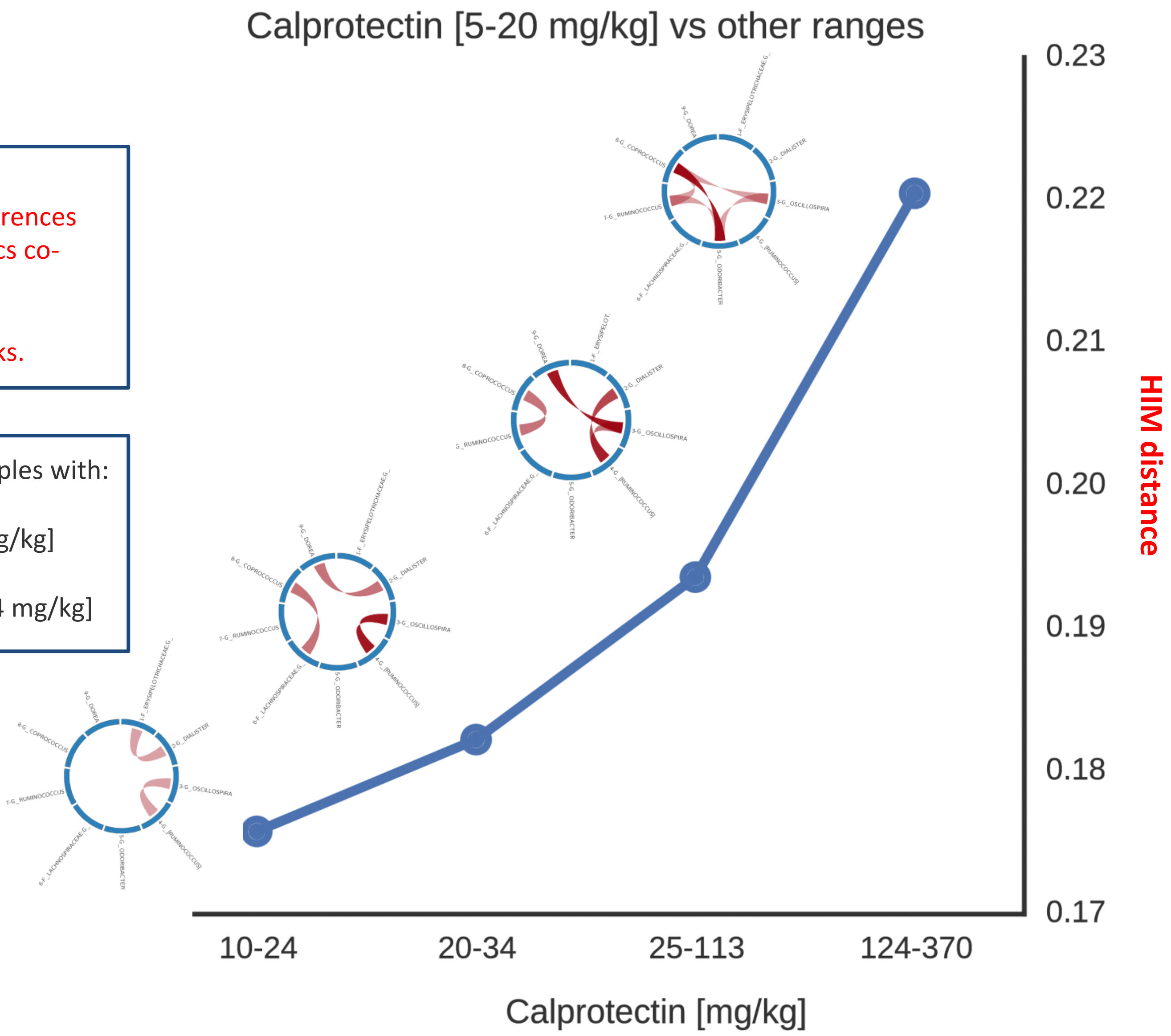
Biopsy networks trajectories

HIM distance:
a quantitative method for evaluating differences between networks, here for metagenomics co-occurence.

Low distance means more similar networks.

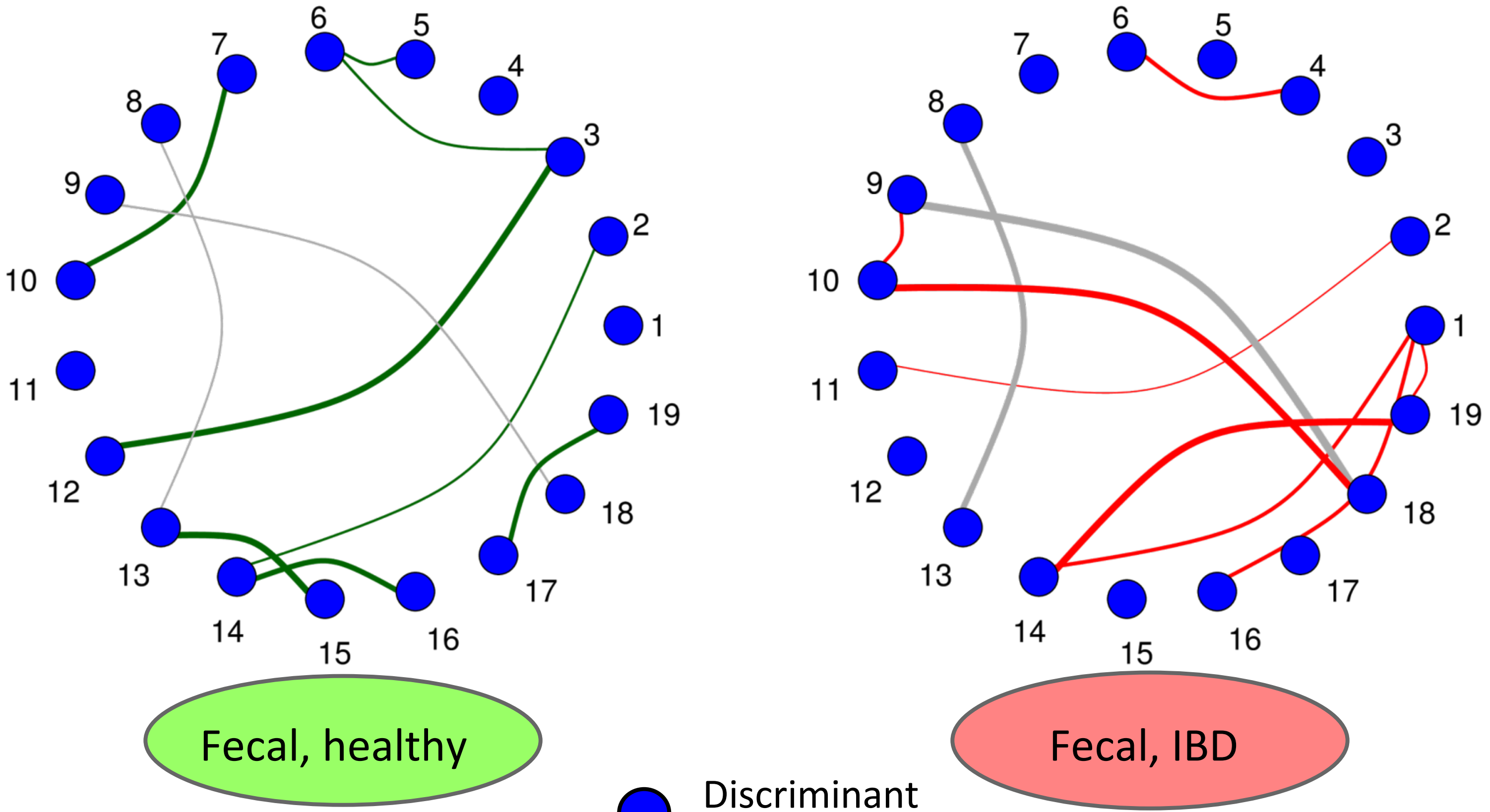
HIM distances between networks on samples with:

lowest levels of calprotectin [5-20 mg/kg]
vs.
increasing levels of calprotectin [10 -374 mg/kg]



Networks: healthy vs IBD

Co-occurrence nets for Pearson Correlation, for stronger links only (PCC > 0.5)



● Discriminant taxa

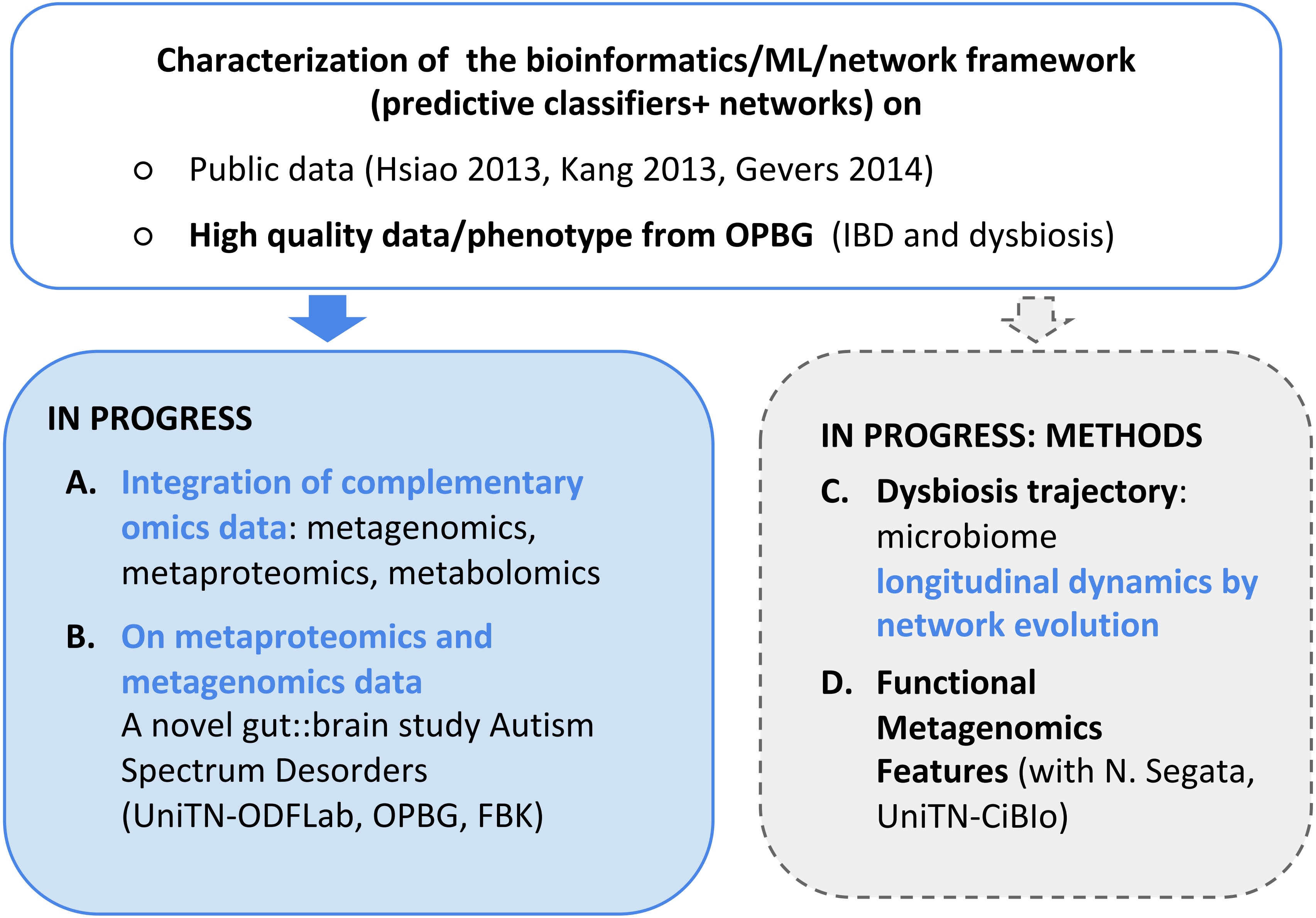
1: f_Barnesiellaceae
3: g_Dorea
8: g_Streptococcus
9: o_Clostridiales
12: g_Collinsella
13: (p_Proteobacteria);c_Gammaproteobacteria
15: p_Proteobacteria
18: f_Lachnospiraceae

▨ Conserved links

🟩 Links conserved in healthy only

🟥 Links conserved in IBD only

Summary 1



Characterization of the bioinformatics/ML/network framework (predictive classifiers+ networks) on

- Public data (Hsiao 2013, Kang 2013, Gevers 2014)
- **High quality data/phenotype from OPBG** (IBD and dysbiosis)

IN PROGRESS

A. **Integration of complementary omics data**: metagenomics, metaproteomics, metabolomics

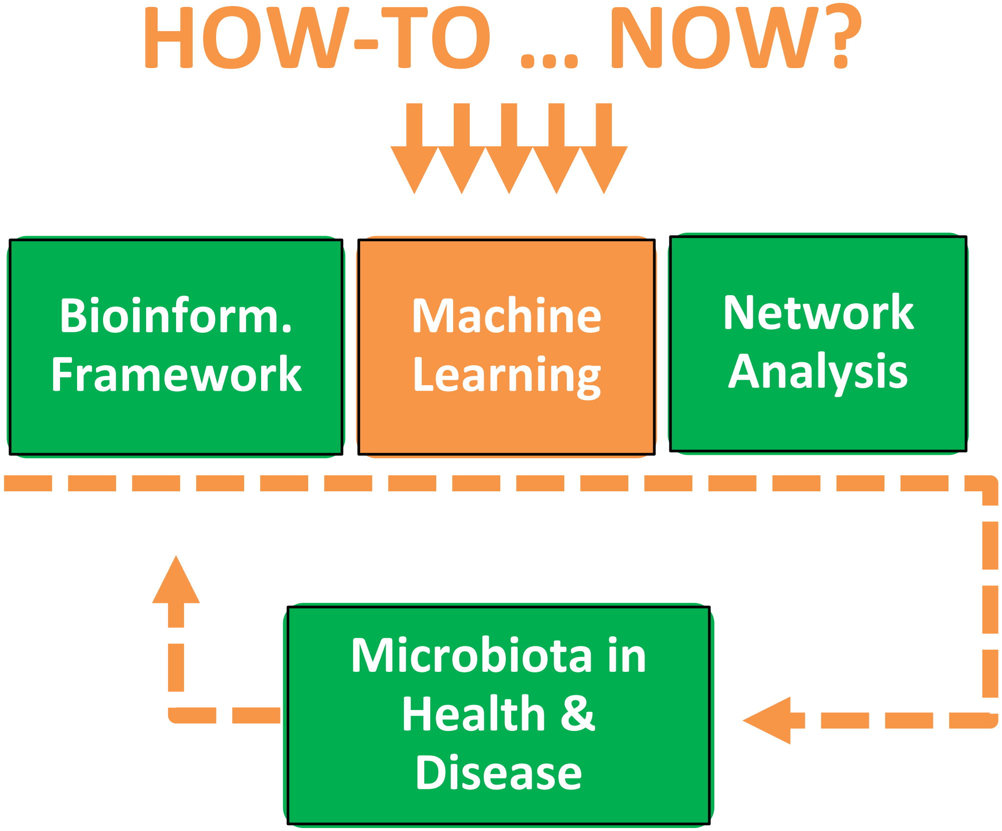
B. **On metaproteomics and metagenomics data**
A novel gut::brain study Autism Spectrum Desorders (UniTN-ODFLab, OPBG, FBK)

IN PROGRESS: METHODS

C. **Dysbiosis trajectory**: microbiome **longitudinal dynamics by network evolution**

D. **Functional Metagenomics Features** (with N. Segata, UniTN-CiBIo)

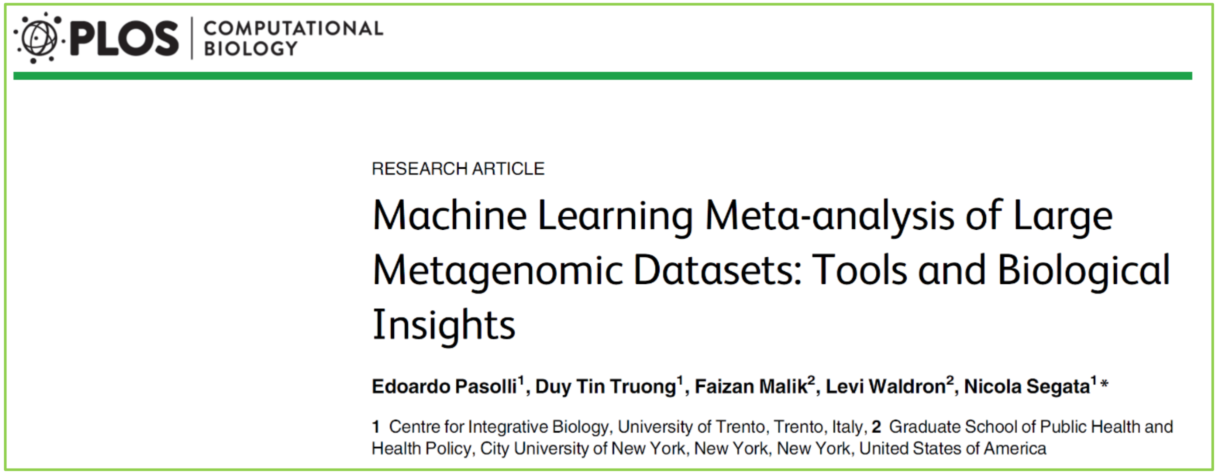
Map/5


HOW-TO … NOW?
Bioinform. Framework
Machine Learning
Network Analysis
Microbiota in Health & Disease

ML Framework for Metagenomics

MetAML

(July 2016)



PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights

Edoardo Pasolli[1], Duy Tin Truong[1], Faizan Malik[2], Levi Waldron[2], Nicola Segata[1]*

1 Centre for Integrative Biology, University of Trento, Trento, Italy, **2** Graduate School of Public Health and Health Policy, City University of New York, New York, New York, United States of America

A framework for validating computational tools for ML tasks in metagenomics

- 8 large-scale studies («shotgun» aka whole-genome, 2424 samples):
  Liver Cirrhosis, Colorectal Cancer, Inflammatory Bowel Disease, Obesity, Type2 Diabetes, HMP Controls (~1K, no disease)
- **Quantitative species/subspecies-level taxonomic profiling** with MetaPhlAn2
  Species **(~ 500 features)** vs strain **(~100 000 features)**
  from 30-70 ML reads
- **Support the systematic assessment of Models transferred between studies, possibly on full archives on clinical outcomes.**

MetAML
RESULTS

A Data Analysis Plan oriented to meta-analysis (Leave-One-Dataset-Out)

- SVM and Random Forests classifiers
- Lasso, Elastic net, regularized multiple log regr, ANN, Bayes. Logistic Regression

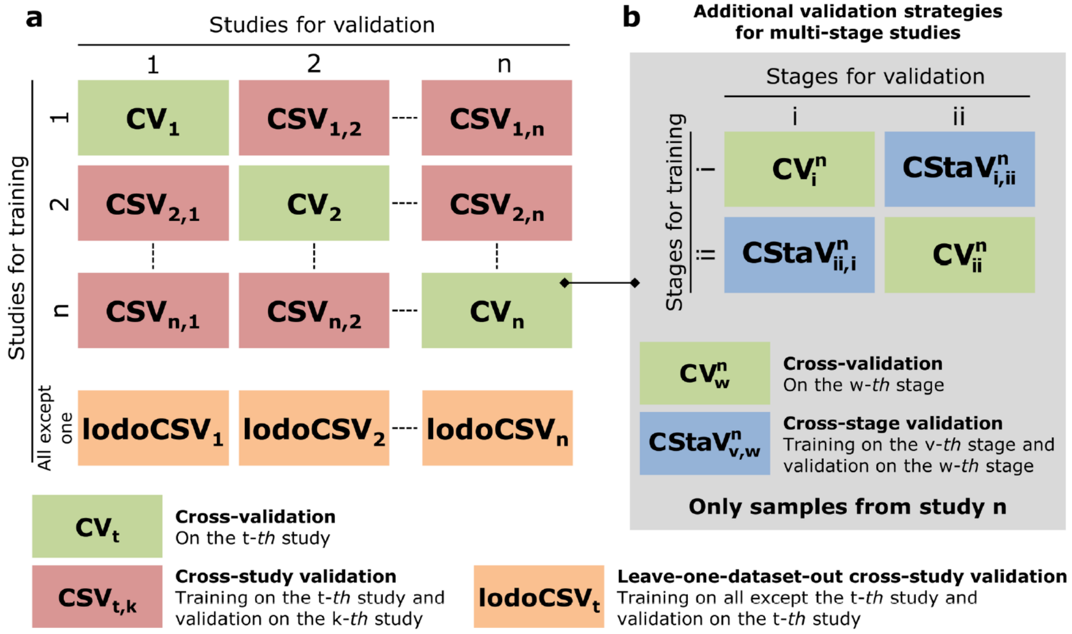
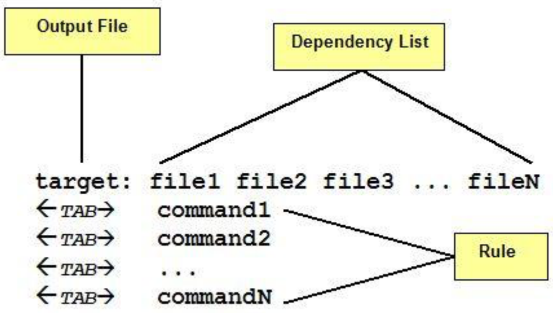


Fig 1. Validation strategies implemented in the developed framework. (**a**) Main strategies include cross-validation on single studies and cross-validation across multiple studies. (**b**) Additional strategies when multiple stages are available from the same study.

doi:10.1371/journal.pcbi.1004977.g001

1. **Good disease prediction from metagenomic data in cv studies**
2. **RF advantage at species level**
3. **Best: strain-level markers and feature selection (with linear SVM > RF)**
4. **Extension to non-disease classification (gender, body site)**
5. **Cross-stage (labs …) generalization is OK**
6. **Generalization improved by including healthy samples from other cohorts**
7. **Good Cross-disease prediction ("general non-healthy status" = dysbiosis)**

For reproducibility and upscaling



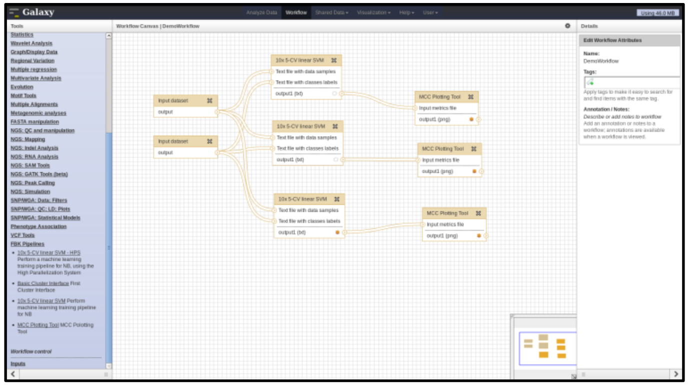
```
target: file1 file2 file3 ... fileN
←TAB→    command1
←TAB→    command2
←TAB→    ...
←TAB→    commandN
```

Pipelines as Makefiles

- Better automation
- Built-in control of parallelization
- Improved reproducibility

Galaxy Workflow Modeler

- Automatic recording of analysis steps & parameters
- Allows non-computational investigators to run complex pipelines



- 10x 5-CV linear SVM - HPS Perform a machine learning training pipeline for NB, using the High Parallelization System
- Basic Cluster Interface First Cluster Interface
- 10x 5-CV linear SVM Perform machine learning training pipeline for NB
- MCC Plotting Tool MCC Polotting Tool

Name:
DemoWorkflow

Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:
Describe or add notes to workflow
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Pushing pipelines on the Cloud

- Completely scalable infrastructure
- Use of computing resources as a service
- Pay-as-you-go

amazon web services™

Google Cloud Platform

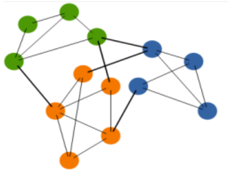
Summary 2

Hunting patterns in metagenomes with ML

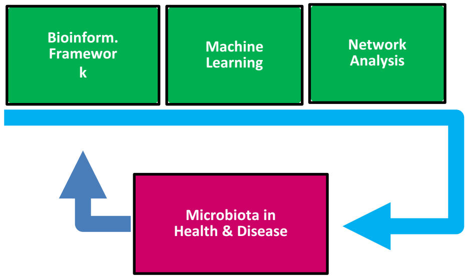
1. Questions start from high throughput metagenomics (aim to whole-genome, 100K features)

- ML framework: now available for a quick start

2. Bioinformatics pipelines



- The FDA/SEQC protocols for predictive markers
- Differential Network Analysis



Example 1: Markers and Diet (gut microbiome)
Example 2: Gut:brain axis (autism)
Example 3: Pediatric Dysbiosis

Visit us at FBK: Povo/Trento Campus

Acknowledgments

Foto Carlo Baroni - Archivio Fotografico : FBK

MPBA / FBK

Giuseppe Jurman, Marco Mina, Roberto Visintainer, Michele Filosi, **Marco Chierici**, Calogero Zarbo, **Alessandro Zandonà**
Silvano Paoli, Roberto Flor

mlpy machine learning py

WebValley

Collaborations

Weida Tong (FDA), Leming Shi (Fudan Univ & FDA), D. Cavalieri, C. De Filippo, K. Tuohy (FEM), A. Barla (UniGE), B. Di Camillo, G. Toffolo (UniPD), A. Quattrone, O. Jousson, N. Segata (CiBIO), GP Tonini (CdS), Louise & Mike Showe (Wistar Inst.), Victor Moreno (ICO Barcelona), **A. Tozzi, L. Putignani, A. Alisi, F. Del Chierico, P. Vernocchi, D. Fruci (OPBG)**, **S. Cucchiara, S. Isoldi** (Uni Sapienza), P. Venuti (UniTN), P. Zanini - Unifarm