Metagenomics data analytics

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

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

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)

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
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

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

[Charbonneau et al, Nature, 2016]
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**

**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

Persistent gut microbiota immaturity in malnourished Bangladeshi children

Satish Subramanian1, Sayeeda Huq2, Tanya Yatsunenko1, Rashidul Haque2, Mustafa Mahfuz2, Mohammed A. Alam2, Amber Benezra1, Joseph DeStefano1, Martin F. Meier1, Brian D. Muegge1, Michael J. Barratt1, Laura G. VanArendonk1, Qunyuan Zhang1, Michael A. Province1, William A. Petri Jr1, Tahmeed Ahmed2 & Jeffrey I. Gordon1

doi:10.1038/nature13421
Microbiome in malnourished children

Severe Acute Malnutrition is associated with significant relative microbiota immaturity

- Machine Learning approach: Random Forest models

APPLICATIONS OF THE “METAGENOMIC CLOCK” IN PRECLINICAL STUDIES


- 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.
Bioinform. Framework

Machine Learning

Network Analysis

Microbiota in Health & Disease
Next Generation sequencing

- Massively parallel sequencing platforms able to produce millions of sequences concurrently, with protocols for DNA, gene expression, methylation, ...
- 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

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)

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)

The Inflammatory Bowel Disease Multi'omics Database
Multi-Omic Microbiome Study-Pregnancy Initiative
Onset of Type 2 Diabetes

The Integrative Human Microbiome Project

The Integrative Human Microbiome Project

The Integrative Human Microbiome Project
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

**WGS**
- **silva**: high quality ribosomal RNA databases
  - Release 115 (SSURef NR):
    - 418,497 bacterial sequences
    - 17,530 archaeal sequences
    - 43,698 eukaryotic sequences

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

- **NCBI**: National Center for Biotechnology Information
  - Ref. metagenome (2013/06):
    - 2,367 Bacteria, Archaea
    - 35 Fungi
    - 2,397 Viruses

Also: KEGG, COG, GO, EggNOG
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

- Reads quality control
- Reads mapping
- Quantification
- Taxonomy assignment

Microbial abundances matrix

<table>
<thead>
<tr>
<th>sample</th>
<th>OTU1</th>
<th>OTU2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>120</td>
<td>42</td>
<td>...</td>
</tr>
<tr>
<td>S02</td>
<td>11</td>
<td>108</td>
<td>...</td>
</tr>
<tr>
<td>S03</td>
<td>0</td>
<td>49</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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

Machine Learning

- Predictive classification
- Network analysis

Networks trajectories

Networks distance

Absolute or relative (compositional data)

Biological samples

Healthy person

Disease state
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 $\ell_1$ penalty to infer the correlation network for latent variables of compositional data. An effective alternating direction algorithm from augmented Lagrangian method is used to solve the optimization problem.

[Fang et al, *Bioinformatics*, 2015]
- All sequencing platforms have artefacts

main artefact: long homopolymers

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

http://www.mothur.org

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

OTU table

<table>
<thead>
<tr>
<th>sample</th>
<th>OTU1</th>
<th>OTU2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>120</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>S02</td>
<td>11</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>S03</td>
<td>0</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

merging
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

Data preparation
- QC
- Preprocessing

“Sense-making”
- Machine learning
- Networks

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

Ranked feature list

Performance evaluation

Selected features

Best Models

MCC, ACC,...

Average metrics

OTU table (training data)

OTU table (validation data)

Data splitting

Test set

Prediction

5-fold CV

Repeat 10 times

(Classifier Tuning)

Classification model

Selected Features

Predicted labels
A MAQC-II/SEQC Data Analysis Plan

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


Used in

- Wang C et al. The concordance between RNA-Seq and microarray data depends on chemical treatment and transcript abundance. Nature Biotech, 2014
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

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.exploredatav.net/Downloads/MINE-Application).

MINEPY is multipurpose (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:

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

Machine Learning

Network Analysis

Microbiota in Health & Disease
Example 1: Diet Induced Diversity

Diet induced diversity
[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

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

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

- **Western vs LF wrt taxonomy**

**Top 5 discriminant nodes**

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Western</th>
<th>LF</th>
<th>Total</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deferribactere</td>
<td>1.60E-06</td>
<td>0</td>
<td>6.53E-07</td>
<td>2.45</td>
</tr>
<tr>
<td>Fibrobacteres</td>
<td>1.55E-06</td>
<td>0</td>
<td>6.32E-07</td>
<td>2.45</td>
</tr>
<tr>
<td>Tenericutes</td>
<td>1.25E-05</td>
<td>0</td>
<td>5.12E-06</td>
<td>2.45</td>
</tr>
<tr>
<td>Lentisphaerae</td>
<td>1.34E-05</td>
<td>8.76E-07</td>
<td>5.98E-06</td>
<td>2.09</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>1.67E-05</td>
<td>1.66E-06</td>
<td>7.81E-06</td>
<td>1.93</td>
</tr>
</tbody>
</table>

Nodes = Phyla (~ OTU abundance)
Weighted edges: non linear MIC assoc.
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**
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
- Neurotypical
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),
Network dysbiosis

- 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 McGee,3 Julian A. Codelli,2 Janet Chow,1 Sarah E. Reisman,2 Joseph F. Petrosino,2 Paul H. Patterson,1,4,* and Sarkis K. Mazmanian1,4,*

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 Akeel Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX 77030, USA
4 These authors contributed equally to this work

*Correspondence: erhsiao@caltech.edu (E.Y.H.), ph@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
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
Clostridia: 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

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

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

DATASET 2:
- Biopsies healthy

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

Biopsy Healthy (20)

Biopsy Control

Biopsy IBD

MCC 0.61 (0.54-0.68)
9 features

MCC 0.01 (0-0.02)
3 features

MCC 0.74 (0.68-0.79)
36 features

MCC 0.61 (0.52-0.68)
4 features

Fecal IBD (30)

MCC 0.81 (0.76-0.86)
30 features

Fecal Healthy (27)

15 x 2 paired
Top discriminant features

- f__Rikenellaceae (unsp. G)
- f__[Barnesiellaceae] (unsp. G)
- f__Coriobacteriaceae (unsp. G)
- g__Dorea
- k__Bacteria (unsp. P)
- g__Nitrosopumilus
- g__Lachnospira
- g__[Ruminococcus]
- g__Streptococcus
- o__Clostridiales (unsp. F)
- f__Phyllobacteriaceae (unsp. G)

Classes
- fecal_IBD
- fecal_healthy

Papa et al, 2012

Gevers et al, 2014

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

Networks: IBD vs. healthy

Top discriminant features between biopsies IBD vs. healthy

Pearson Correlation

Co-occurrence networks inference

Links conserved in healthy only

Links conserved in IBD only

Edge’s color intensity $\propto$ absolute value of Pearson correlation coefficient (PCC)

Shown links with PCC > 0.65
Calprotectin level is associated to increasing dysbiosis in Biopsy Networks

Phenotype: Cucchiara Lab - Rome

Healthy: Calprotectin < 50 mg/kg
Calprotectin level is associated to increasing dysbiosis in Biopsy Networks

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

Correlation between links 3 and 5

Pearson_correlation

Calprotectin

5-20 10-24 20-34 25-113 124-370
Markers patterns vs Calprotectin

Features abundance vs Calprotectin range

Feats
- f_Erysipelotrichaceae;g_
- g_Dialister
- g_Oscillospira
- g_[Ruminococcus]
- g_Odoribacter
- f_Lachnospiraceae;g_
- g_Ruminococcus
- g_Coprococcus
- g_Dorea

Classes
- biopsy_healthy
- biopsy_infl

Value

Calprotectin ranges [mg/Kg]

Median OTUs abundance
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,

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-occurrence.

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]

---

**Calprotectin [5-20 mg/kg] vs other ranges**

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

**Calprotectin [mg/kg]**

**HIM distance**

0.17 - 0.23
Networks: healthy vs IBD

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

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 Disorders (UniTN-ODFLab, OPBG, FBK)

IN PROGRESS: METHODS

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

D. **Functional Metagenomics Features** (with N. Segata, UniTN-CiBiIo)
HOW-TO ... NOW?

Bioinform. Framework  Machine Learning  Network Analysis

Microbiota in Health & Disease
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

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

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

Pushing pipelines on the Cloud
- Completely scalable infrastructure
- Use of computing resources as a service
- Pay-as-you-go
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**
Acknowledgments

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

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. Putilignani, A. Alisi, F. Del Chierico, P. Vernocchi, D. Fruci (OPBG), S. Cucchiara, S. Isoldi (Uni Sapienza), P. Venuti (UniTN), P. Zanini - Unifarm