UNIVERSITY OF TRENTO - Italy

PhD Program in Biomolecular Sciences

Centre for Integrative Biology
XXIX Cycle

Predictive networks for multi meta-omics data integration

Tutor
Cesare FURLANELLO
Fondazione Bruno Kessler (FBK)

Advisor
Marco CHIERICI
Fondazione Bruno Kessler (FBK)

Ph.D. Thesis of
Alessandro ZANDONÀ
Centre for Integrative Biology (CIBIO)
University of Trento

Academic Year 2015-2016

Contents

Declaration of authorship

The undersigned Alessandro Zandonà declares that the work presented in this thesis is my own. The use of all material from other sources has been properly and fully acknowledged.

Alessandro Zandonà

Abstract

The role of microbiome in disease onset and in equilibrium is being exposed by a wealth of high-throughput omics methods. All key research directions, *e.g.,* the study of gut microbiome dysbiosis in IBD/IBS, indicate the need for bioinformatics methods that can model the complexity of the microbial communities ecology and unravel its disease-associated perturbations. A most promising direction is the “meta-omics” approach, that allows a profiling based on various biological molecules at the metagenomic scale (*e.g.,* metaproteomics, metametabolomics) as well as different “microbial” omes (eukaryotes and viruses) within a system biology approach. This thesis introduces a bioinformatic framework for microbiota datasets that combines predictive profiling, differential network analysis and meta-omics integration. In detail, the framework identifies biomarkers discriminating amongst clinical phenotypes, through machine learning techniques (Random Forest or SVM) based on a complete Data Analysis Protocol derived by two initiatives funded by FDA: the MicroArray Quality Control-II and Sequencing Quality Control projects. The biomarkers are interpreted in terms of biological networks: the framework provides a setup for networks inference, quantification of networks differences based on the glocal Hamming and Ipsen-Mikhailov (HIM) distance and detection of network communities. The differential analysis of networks allows the study of microbiota structural organization as well as the evolving trajectories of microbial communities associated to the dynamics of the target phenotypes. Moreover, the

framework combines a novel similarity network fusion method and machine learning to identify biomarkers from the integration of multiple meta-omics data. The framework implementation requires only standard open source computational biology tools, as a combination of R/Bioconductor and Python functions. In particular, full scripts for meta-omics integration are available in a GitHub repository to ease reuse (`https://github.com/AleZandona/INF`). The pipeline has been validated on original data from three different clinical datasets. First, the predictive profiling and the network differential analysis have been applied on a pediatric Inflammatory Bowel Disease (IBD) cohort (in faecal vs biopsy environments) and controls, in collaboration with a multidisciplinary team at the Ospedale Pediatrico Bambino Gesú (Rome, I). Then, the meta-omics integration has been tested on a paired bacterial and fungal gut microbiota human IBD datasets from the Gastroenterology Department of the Saint Antoine Hospital (Paris, F), thanks to the collaboration with “Commensals and Probiotics-Host Interactions” team at INRA (Jouy-en-Josas, F). Finally, the framework has been validated on a bacterial-fungal gut microbiota dataset from children affected by Rett syndrome. The different nature of datasets used for validation naturally supports the extension of the framework on different omics datasets. Besides, clinical practice can take advantage of our framework, given the reproducibility and robustness of results, ensured by the adopted Data Analysis Protocol, as well as the biological relevance of the findings, confirmed by the clinical collaborators. Specifically, the omics-based dysbiosis profiles and the inferred biological networks can support the current diagnostic tools to reveal disease-associated perturbations at a much prodromal earlier stage of disease and may be used for disease prevention, diagnosis and prognosis.

Chapter 1

Introduction

1.1 Background

Trillions of microbes inhabit human body and create complex, body-habitat-specific ecosystems: the communities formed by this complement of cells is called the human microbiota, and its genomic content is defined as microbiome. Microbiota contains almost ten times as many cells as are in the rest of the body and orders of magnitude more genes than are included in the human genome [1, 2]. It is generally accepted that humans are born with eukaryotic human cells only, but over the first years of life the oral cavity, the skin surface and gut are colonized by bacteria, archaea, fungi, and viruses [3]. Among the epithelial surfaces colonized by microbes, the gastro-intestinal (GI) microbiota is one of the most diverse communities consisting of hundreds of species which vary between individuals as well as across space and time within the same individual [4, 5, 1]. The colonization of the GI tract starts before birth with the fetus ingesting amniotic fluid containing microbes [6], and continues with aerobic and facultative anaerobic colonization during the first months of life, followed by obligate anaerobes and *Bifidobacteria*. The establishment of the gut microbiota is recognized as a complex process influenced by factors at the level of the host and of the microbes themselves [7]. The bacterial communities inhabiting the gut compete for a limited quantity of diet-derived or mucus-derived carbohydrate available for fermentation [8], with *Clostridia* and *Bacteroidia* among the most dominant obligate anaerobes over *Enterobacteriaceae*. Indeed, *Clostridia* and *Bacteroidia* use glycoside hydrolases (GHs) to degrade complex carbohydrates, binding proteins to concentrate carbohydrate at their surface and active transport systems to import substrates against a concentration gradient. By contrast, a paucity of GHs make *Enterobacteriaceae* illequipped to degrade complex carbohydrate, only relying on oligosaccharides passively transported across barrier. This might partially explain the difficulty of *Enterobacteriaceae* to compete with obligate anaerobic bacteria for high-energy nutrients to support their growth by fermentation, with a disavantage in acquiring fermentable nutrients during anaerobic growth [9, 10]. This microbial community structure is characterized by a dynamic steady-state undergoing changes due to genetic predispositions, external perturbations (*i.e.,* dietary input) and host-microbiota feedback interactions (*i.e.,* host nutrient requirements). As long as the host intestinal architecture and immune system are in complementary homeostasis with the commensal microflora, the system is considered as healthy. However, the baseline healthy homeostasis can be disrupted by perturbations affecting the host (*i.e.,* trauma, surgery, exposure to harsh chemicals), the microbiota (*i.e.,* the ingestion of toxins, drugs, tainted food) as well as a combination of perturbations through diet, antibiotic regimens, chemotherapy or radiation treatments. The disturbance of a balanced host-microbiota relationship leading to a prolongation, exacerbation, or induction of a detrimental health effect is defined as dysbiosis [11]. Several studies have associated a dysbiotic state of microbiota with numerous diseases, including allergies, asthma, autism, diabetes, multiple sclerosis, inflammatory bowel diseases and cancer [11, 12, 13, 14]. In particular, the impact of the gut microbiota on gut and systemic immune homeostasis has gained tremendous research interest over the last few years. Indeed, the intestinal epithelial barrier integrity is crucial for the maintenance of a correct intestinal absorption while shielding the body from the gut lumen content, including dietary antigens and microbial products [15, 16]. It is already known that enteric pathogens are able to strongly modify the intestinal permeability by affecting specific tight junction proteins [17, 18]; on the other hand, commensal and probiotic bacteria are known to improve the intestinal barrier functioning [19, 20]. However, the effects of the dysbiotic microbial community on cell permeability, junction complexes and in general on human health are still unclear: an important approach to investigate the microbiota and its interaction with the host is metagenomics.

Metagenomics is the study of the genomic DNA within a microbial community, mainly based on high-throughput sequencing, which determines the precise order of nucleotides within multiple DNA molecules in parallel, coupled with bioinformatic analyses. In detail, two main methods are used in metagenomics to sequence the microbiome: amplicon or targeted sequencing and shotgun or whole genome sequencing (WGS). The first one uses pooled sequencing of the PCR product of a specific marker gene (*i.e.,* 16S and 18S ribosomal RNA) followed by mapping the resulting uniquely identified sequences to a taxonomic database; on the other hand, the WGS method sequences the whole metagenomic content in a sample. Typically, the amplicon sequencing provides the composition of the sequenced microbial community, while WGS is commonly used to characterize the functional capability of the microbiota [21]. Some limitations of amplicon sequencing are: (1) lack of specificity in the taxonomic resolution, so that only taxonomic categories at the family or genus level can be well-characterized, (2) difficulties in cross-study comparisons, (3) reliance upon existing curated rRNA databases to align sequences for taxonomic assignment. Despite its limitations, amplicon sequencing is popular because it is cost-effective, it avoids non-bacterial contamination, it can potentially catch low abundance bacteria and mature analysis software are available. On the other hand, WGS is more expensive, can miss the low abundant bacteria and the resulting dataset is more challenging to process in terms of size. However, the choice of WGS as sequencing method is motivated by its capability of profiling the metabolic potential or virulence/antibiotic resistance, of surveying all domains of life simultaneously and by its high taxonomic resolution. Sequencing the microbiome is the initial step for metagenomics analysis, followed by a combination of data-driven bioinformatics with knowledge-driven computational modeling. Data-driven bioinformatics refers to the ensemble of statistical, mathematical and algorithmic methods that aim to discover meaningful patterns from biological data, such as metagenomic sequences from HTS technologies. Knowledge-driven computational modeling simulates the cause and effect relationships of biological mechanisms encoding knowledge about biological entities, processes and mechanisms into mathematical objects. Bioinformatics extracts insights from experimental data and suggests new hypotheses, but it is not capable of evaluating the causality behind these hypotheses, which computational models can provide. On the other hand, computational models design an abstraction of the real biological system and produce simulated data similar to experimental data, which requires bioinformatics to properly analyze [4].

In this thesis, bioinformatics and computational modeling are combined in order to characterize (1) microbiota communities composition and structure, (2) microbe-microbe and host-microbe interactions, and (3) model the dynamics of microbiota perturbations leading to a disease status in human. More in detail, this thesis adapts well-established machine learning and network analysis algorithms to predict phenotypes (*e.g.,* microbiota-relate disease), perform classification (*e.g.,* distinguish healthy from unhealthy microbiota), extract discriminative features and provide predictive dynamics of microbial communities. In this thesis such metagenomic analyses, from DNA sequence processing to machine learning profiling and differential network analysis, are implemented through Open Source Software tools embedded into two bioinformatics framework: PreMONet (Predictive Meta-Omics Networks) and its extended version I-PreMONet (Integrated Predictive Meta-Omics Networks). A peculiarity of these frameworks is the predisposition to handle not only microbiome data, but also other "meta-omics" data, intended as various biological molecules at the metagenomic scale (*e.g.*, metaproteomics, metametabolomics) as well as different "microbial" omes (eukaryotes and viruses). In detail, PreMONet provides a complete analysis of a specific meta-omics data type, while I-PreMONet implements PreMONet analyses on multiple meta-omics data simultaneously. It is clear that a comprehensive integrative modeling of "meta-omics" layers provides a much rich characterization of the complementary aspects of microbial communities [22, 23, 24]. Conversely, several meta-omics pipelines have been currently designed, but they are limited to a single meta-omics layer analysis. In particular, current computational tools either provide the characterization of microbiome taxonomic composition and functional potential (QIIME [25], UPARSE [26], MG-RAST [27], MICCA [28]), or the analysis of the RNA transcript pool expressed by microbes (SortMeRNA [29], Trinity [30]) or the study of the whole protein complement of microbiome (MetaProteomeAnalyzer [31], MASCOT [32]). Notably, I-PreMONet collects computational tools in a modular and customizable manner, avoiding performing these analyses separately, which usually requires the installation, integration, and tuning of multiple software packages, which is not always trivial even for groups with extensive bioinformatics expertise.

In summary, a modular framework to find meta-omics features that distinguish healthy from unhealthy microbial communities and modeling microbial environment could aid in the diagnosis of microbiota-related diseases and could potentially provide new means to prevent disease onset or to improve prognosis.

1.2 Publications

The central predictive profiling methods and differential networks analysis of microbiome data in this thesis have been implemented in PreMONet framework. The framework, its core components and an application to 16S rRNA-Seq dataset from gut microbiota of pediatric Inflammatory Bowel Disease (IBD) have been included in a computational biology conference:

- Zandonà A., Chierici M., Jurman G., Furlanello C., Cucchiara S., Del Chierico F., Putignani L. **A metagenomic pipeline integrating predictive profiling methods and complex networks for the analysis of NGS microbiome data.** NIPS Workshop - Machine Learning in Computational Biology, Montreal, Canada. December 13, 2014.

Further, the PreMONet was extended with algorithms for meta-omics integration, leading to a framework named I-PreMONet. The computational details and validation of I-PreMONet on a human IBD dataset (bacterial and fungal microbiota) are described both in a conference contribution and in a full paper:

- Zandonà A., Trastulla L., Jurman G., Agostinelli C., Furlanello C., Lavie-Richard M., Sokol H. **Integrated meta-omics for models of gut inflammatory disease.** NIPS Workshop - Machine Learning in Computational Biology, Barcelona, Spain. December 10, 2016.

- Zandonà A., Trastulla L., Jurman G., Agostinelli C., Furlanello C., Lavie-Richard M., Sokol H. **Integrative Network Fusion of bacterial-fungal microbiota for the identification of robust IBD biomarkers.** Submitted to PLOS Computational Biology (2017).

Components of the idea were already made available in a collaborative paper and in a conference contribution:

- Del Chierico F., Nobili V., Vernocchi P., Russo A., De Stefanis C., Gnani D., Furlanello C., Zandonà A., Paci P., Capuani G., Dallapiccola B., Miccheli A., Alisi A., Putignani L. **Gut microbiota profiling of pediatric NAFLD/obese patients unveiled by an integrated meta-omics based approach**. Hepatology, 2016.

- Zandonà A. **Complex networks for the analysis of microbiome structures.** Bringing Maths to Life (BMTL) Workshop, Naples, Italy. October 30, 2015.

The general machine learning setup for the identification of predictive biomarkers has been presented at different conferences:

- Zandonà A. **Choice of Training-Validation partitions impacts predictive performances.** 4th Italian Workshop on Machine Learning and Data Mining (#AI4-MLDM), Ferrara, Italy. September 22, 2015.

- Zandonà A. **A metagenomic pipeline integrating predictive profiling methods and complex networks for the analysis of NGS microbiome data.** 3S Biology Summer School, Center for Integrative Biology, University of Trento, Italy. September 9, 2015.

- Zandonà A. **From metagenomics to epigenetics: a bioinformatics pipeline.** KAUST-UCI Symposium-Epigenetics & Environment, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Date TBE.

1.3 Thesis outline

This thesis describes PreMONet and its extended version I-PreMONet, their core components (Chapters 2,3,4), their structure (Chapter 5) and their validation on well-phenotyped clinical datasets (Chapter 6). Throughout the thesis, the term (I-)PreMONet will be used to reference either of the frameworks.

Chapter 2 introduces the (I-)PreMONet module that implements machine learning algorithms to discriminate host phenotypes (*i.e.,* health conditions), based on meta-omics abundance. In detail, Paragraph 2.1 focuses on the microbiota as a tool for predicting host phenotypes, Paragraph 2.2 shows the mathematical details of machine learning algorithms and Paragraph 2.3 describes the protocol adopted to develop predictive models for meta-omics data.

Chapter 3 details the (I-)PreMONet component that models interactions among meta-omics as well as microbiota structure and dynamics, in association with host phenotype. Paragraph 3.1 describes the importance of ecological interactions within microbiota for the host health, while Paragraphs 3.2, 3.3 and 3.4 show the approaches adopted to model microbial communities as networks and analyze their perturbations associated to host phenotypes.

Chapter 4 focuses on I-PreMONet module, which offers a comprehensive integrative modeling of multiple meta-omics layers. Paragraph 4.1 presents Similarity Network Fusion (SNF), a state-of-the-art network-based method for meta-omics integration. Paragraph 4.2 introduces our *rSNF* extension, a feature ranking scheme on integrative features that extends SNF. Finally, Paragraph 4.3 describes *INF*, our framework combining a network fusion method with machine learning to identify robust biomarkers from the integration of meta-omics data. Note that PreMONet and I-PreMONet are equipped with the same modules for predictive modeling and network analysis, except for *INF*, which is a component exclusively embedded into I-PreMONet.

Chapter 5 is structured into two parts: a review of the bioinformatics frameworks for meta-omics data that are popular in literature is followed by the description of PreMONet and I-PreMONet design.

Chapter 6 reports the validation of our frameworks on three clinical metagenomic datasets: gut microbiota composition of (1) children and (2) adults with Inflammatory Bowel Disease as well as (3) children affected by Rett syndrome. These case studies are designed to identify meta-omics biomarkers discriminating amongst clinical phenotypes and to study the dynamics of microbial communities associated to the target phenotypes. In particular, dataset (1) provides the bacterial composition of gut microbiota, thus it is analyzed by PreMONet; conversely, datasets (2) and (3) are analyzed by I-PreMONet, since both bacterial and fungal microbiota are available.

General conclusions on (I-)PreMONet performance and biological significance, are summarized in Chapter 7.

Chapter 2

Predictive profiling

Early metagenomics studies on human microbiota mainly focused on its taxonomic profiling, the characterization of microbial community composition and structure, comparing samples from the same site among and between individuals [33, 2]. Subsequently, significant work has been devoted to reproducibly associate microbiome to specific diseases, such as inflammatory bowel disease [34, 23], diabetes [35, 36] and cancer [37, 38]. One of the main challenges now is the development of a microbiome-based diagnostic and possibly prognostic tool, able to estimate host phenotype for different illnesses or treatments, based on the current state of the microbiota [39, 40]. Microbial biomarkers might potentially facilitate more efficient and reliable clinical trials [39, 41] and confirm the diagnoses, as well as predict treatment outcomes. However, a microbial predictive signature may be complex: indeed, host phenotypes are commonly associated with changes in bacterial communities, not with a single specific biomarker. Machine learning offers several promising tools to deal with such complex high-throughput data and predict phenotypes, based on multivariate statistics, data mining and pattern recognition.

2.1 Microbial biomarkers

Biomarkers have been defined by Perlis [41] as “the measurable characteristics of an individual that may represent risk factors for a disease or outcome, or that may be indicators of disease progression or of treatment-associated changes”. An imbalance in the gut microbial community composition has been strongly related to disease, thus each microbial taxa can be potentially referred to as a biomarker. Thus, detecting microbial taxa that broadly distinguish healthy from “unhealthy” microbiomes or discriminate phenotypes could support the diagnosis of microbiome related diseases and aid in disease onset prevention [11]. Basically, microbial biomarkers discovery starts with the stratification of people on the basis of the microbiome, followed by the development of a predictive model that can then be used to predict the phenotype associated with specific microbial communities. Importantly, the traits leading to disease can be commonly related to simultaneous over- and under-representations of multiple taxa at multiple taxonomic levels and not limited to a single biomarker [42]. Owing to such a level of complexity, together with the large amount of metagenomic data, the bioinformatics support for biomarker discovery is crucial.

Several clinical studies and case reports have highlighted the potential benefits of microbiome-based diagnostic tools, complementing or improving traditional “gold standard” testing. For instance, Brown and colleagues [43] found that 63% of encephalitis cases go undiagnosed despite extensive testing, while metagenomics allowed the diagnosis of rare, novel, or atypical infectious etiologies for encephalitis, including cases of infection by *Leptospira* [44], astrovirus [45], and bornavirus [46]. Besides, microbial biomarkers could contribute to different precision medicine efforts, such as the deployment of prebiotics, probiotics, and targeted antibiotics. In fact, a better understanding of the individual microbiota in a patient could improve the effectiveness of treatment and avoid or mitigate adverse reactions. Another strength of microbiome-based assays is the ability to find many microorganisms without neither additional individual testing nor *a priori* knowledge of the type of pathogen.

Conversely, one of the major weaknesses of microbial biomarkers is that many of the potential clinical applications are still undefined and subject to continuous updating. For instance, the presence of a microorganism potentially pathogenic in one microenvironment, such as cerebrospinal fluid, typically considered sterile, may be normal in the mouth, skin or gut. Moreover, bacteria and viruses are continuously developing new resistance mechanisms, thus clinical laboratories should keep reference databases constantly updated to include novel resistance mutations and to avoid lowering the prediction effectiveness of microbiome-based diagnostic tools. Such challenging issues will probably slow down the phenotypic confirmation needed for systematic adoption in clinical practice.

In the next paragraphs, we will discuss the details of the predictive bioinformatics pipeline for microbial biomarker discovery, including the machine learning approach for associating individual features of the microbiome with phenotype.

2.2 Predictive models

Machine learning is the methodology of finding patterns and making predictions from data, based on a combination of tools from multivariate statistics, data mining and pattern recognition, with a focus on generalizing from given datasets to novel cases. Predictive modeling exploits patterns observed in datasets in order to identify an optimal model from a hypothesis space and estimate future outcomes. The main characteristic associated with a predictive model, and in general with Artificial Intelligence (AI) based on machine learning, is that it can improve with experience, by flexibly adapting to new domains and fine-tuning to the observed data. However, this flexibility requires an appropriate control of the risk of overfitting the available data and warrant adequate accuracy on novel data. This is typically achieved by introducing regularization mechanisms and adopting training schema that can balance bias and variance estimation.

Machine learning algorithms are often grouped into three categories: classification and regression (the outcome is a categorical or numerical function to fit), clustering (the aim is to optimally partition the input feature space in groups with respect to a criterion), dimensionality reduction (to achieve a lower-dimensional representation of the input data). Alternatively, learning algorithms can be characterized as *supervised* or *unsupervised*. Building a model from a set of labeled data points to predict the correct category of unlabeled future example is the goal of a supervised method, such as a classification or regression model. For instance, discriminating phenotypical traits based on metagenomics samples composition is a supervised problem. On the contrary, unsupervised methods do not aim to produce a labeled response directly, but rather to find the hidden structure of the data: cluster analysis and dimensionality reduction are the most common unsupervised methods. Here they will be used to identify subtypes of interest in the normal and pathophysiological conditions.

In this thesis, supervised methods will be used to associate subjects to phenotypes (*i.e.,* health status), based on features derived from the bioinformatics analysis of DNA fragments from human microbiome. The process will involve identifying the most discriminative variables, *i.e.,* the microbial taxa harboring those patterns in the microbiome that most contribute to the model accuracy. Basically, a dataset of example pairs of microbial communities with known phenotype labels will be used to train a learning algorithm, also selecting its optimal parameters for prediction of the phenotype of other unseen data describing microbial communities from similar datasets, in similar clinical setting. Notably, machine learning applied on metagenomics is a fruitful approach to associate phenotypes with unculturable microbial communities [47, 48].

More in detail, two well known supervised techniques are used in this thesis: Support Vector Machines (SVM) and Random Forest (RF). Before formally defining SVM and RF, the introduction of some basic concepts is required. In a typical supervised learning scenario, predictive model is built on a matrix X, consisting of n observations (*i.e.,* subjects) of p different measurements (*i.e.,* abundance of microbial taxa). X is commonly defined as *training* dataset. Moreover, each observation can be associated to a class (*i.e.,* a clinical phenotype, such as health status), denoted by Y. A learning algorithm exploits the training dataset (X,Y) to learn how to predict classes $\hat{Y}$ of observed and previously unseen data, defined as *validation* dataset. In this context, commonly one part of the training dataset is used to train the learning algorithm and develop a model, while another is used to test the algorithm performance, defined as *test* dataset. As we will see, resampling mechanisms are also adopted to ensure that the training and test splits are chosen in a way representative of the validation set or other novel data.

Support Vector Machines

Support vectors machines (SVM) are a set of supervised learning algorithms [49]; SVM searches for "optimal" hyperplane that linearly separates data, but can also be extended to patterns that are not linearly separable by transformations of data into a new space by appropriate maps [49].

To define SVM formally, the concept of *optimal separating hyperplane* is required. Consider a $(n \times p)$ data matrix X that consists of n training observations in p-dimensional space, $X = (x_1^T;\ldots;x_n^T)$ and that the observations belong to two classes, coded as -1 and 1, *i.e* $y_1,\ldots,y_n \in \{-1,1\}$. Moreover, suppose to have p-vectors of test observations. The optimal separating hyperplane divides the observations by the two classes and maximizes the distance to the closest point from either class.

Define an hyperplane H of $\mathbb{R}^p$ by

$$H = \{x \in \mathbb{R}^p : f(x) = \beta_0 + x^T\beta = 0\},$$

where $\beta \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$. Note that, for any two points x_1 and x_2 in H, $\beta(x_1 - x_2) = 0$ hence $\beta/\|\beta\|$ is the unit vector normal to the surface of H. Moreover, $x = x^p + r^x \dfrac{\beta}{\|\beta\|}$, since a vector x in $\mathbb{R}^p$ can be expressed by its projection on H plus its distance to H times the unit vector in that direction. This implies

$$f(x) = \beta_0 + \left((x^p)^T + r^x \frac{\beta^T}{\|\beta\|}\right)\beta = f(x^p) + r^x\|\beta\| = r^x\|\beta\| \implies r^x = \frac{f(x)}{\|\beta\|},$$

hence, the signed distance from x to the hyperplane H is $f(x)/\|\beta\|$. In conclusion, the classification rule induced by f is

$$G(x) = sign(\beta_0 + x^T\beta).$$

If the classes are linearly separable, then a function $f(x) = \beta_0 + x^T\beta$ can be found in order to correctly classify each sample, i.e. $y_i f(x_i) > 0$ for each $i = 1, \dots, n$. Hence, it is possible to find the hyperplane that traces the biggest *margin* (M) between the training points, through the optimization problem:

$$\begin{aligned} &\max_{\beta_0,\beta,\|\beta\|=1} \quad M \\ &\text{subject to} \quad y_i(\beta_0 + x_i^T\beta) \geq M \quad \text{for each} \quad i = 1, \dots, N. \end{aligned} \tag{2.1}$$

This set of conditions ensures that all the points are at least a signed distance M from the decision boundary defined by β_0 and β. We can get rid of the $\|\beta\| = 1$ constraint by replacing the condition with

$$\frac{1}{\|\beta\|} y_i(\beta_0 + x_i^T\beta) \geq M,$$

(which redefines β_0) or equivalently

$$y_i(\beta_0 + x_i^T\beta) \geq M\|\beta\|.$$

Since for any β and β_0 satisfying these inequalities, any positively scaled multiple satisfies them too, we can arbitrarily set $\|\beta\| = 1/M$. Thus (2.1) is equivalent to

$$\begin{aligned}&\min_{\beta_0,\beta}\frac{1}{2}\beta^T\beta\\&\text{subject to}\quad y_i(\beta_0 + x_i^T\beta) \geq 1 \quad \text{for each} \quad i = 1,\dots,n.\end{aligned} \tag{2.2}$$

The constraints define an empty margin around the linear decision boundary of thickness $1/\|\beta\|$, to be maximized tuning β_0 and β. (2.2) is a convex optimization problem since it is composed of a quadratic criterion with linear inequality constraints, for further details on the optimization problem see [49].

Suppose now that the classes overlap in the given feature space. The maximization of M is still possible, but some points will be placed on the wrong side of the margin, thus requiring a “soft margin” solution [49]. Define the slack variables $\xi = (\xi_1,\dots,\xi_n)$. Consequently, the constraints are modified with

$$y_i(\beta_0 + x_i^T\beta) \geq 1 - \xi_i \quad \text{for each} \quad i = 1,\dots,n$$

with $\xi_i \geq 0$ for each $i = 1,\dots,n$ and $\sum_{i=1}^n \xi_i \leq constant$. Value ξ_i in the constraint is proportional to the error of prediction $f(x_i)$. Hence by bounding $\sum_{i=1}^n \xi_i$, we bound the total proportional amount by which predictions fall on the wrong side of their margin. Problem (2.2) can be written for the soft margin case in compact form as

$$\begin{aligned}&\min_{\beta_0,\beta}\frac{1}{2}\beta^T\beta + C\sum_{i=1}^n \xi_i\\&\text{subject to}\quad \xi_i \geq 0,\ y_i(\beta_0 + x_i^T\beta) \geq 1 - \xi_i \quad \text{for each} \quad i = 1,\dots,n;\end{aligned} \tag{2.3}$$

where $C \geq 0$ is the regularization parameter and trades-off data fitting and the margin size.

Different choices on constraints type can be made, to further generalize, the problem can be expressed in the following form (see [50])

$$\underset{\beta,\beta_0}{min}\frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n}\xi(\beta,\beta_0;x_i,y_i), \tag{2.4}$$

where ξ is a loss function depending on β,β_0,x_i and the respective label y_i. Optimization problem (2.4) is known as L$_2$-regularized support vector classification problem, involving the squared L$_2$ norm of β.

Although different types of loss function can be considered, common choices are

$$\text{L}_1\,\text{loss}: \quad \xi(w;x_j,y_j) = (1 - y_i\left(x_i^T\beta + \beta_0\right))_+ \tag{2.5}$$

$$\text{L}_2\,\text{loss}: \quad \xi(w;x_j,y_j) = (1 - y_i\left(x_i^T\beta + \beta_0\right))_+^2. \tag{2.6}$$

Use L_1loss function in (2.4) is equivalent to solve (2.3). In this thesis, the SVM implementation will be mostly based on the scikit-learn v.0.17.1 Python module [51]. Details on the specific models will be given when describing the specific metagenomic models.

Random Forest

Random Forest (RF) is an ensemble method in which tree-based classifiers or regressors are combined after being developed over a resampling both over data and over features. This method generalizes bagging, the basic technique for reducing the variance of an estimated prediction function [52] by averaging over models each developed over bootstrap resampled versions of the data [49]. The basic component of Random Forest models is the tree-based method, which hierarchically partitions the feature space into a set of rectangles and then fits the simplest model (*i.e.*, a constant) in each one. Suppose that our data consist of p measurements and n observations, each one belonging to a different class: that is (x_i, y_i), $i = 1, \dots, n$ with $x_i = (x_{i1}, \dots, x_{ip})$ and $y_i \in \{1, \dots, K\}$ label class. The classification tree algorithm needs to automatically decide on the splitting variables and on split points, and on what topology the tree should have. To summarize, a classification tree can be built through two steps, repeated until a stop criterion is met:

1. data are partitioned into M distinct and non-overlapping regions $R_1, \dots, R_M$;
2. for every observation falling into the region R_m , we make the same prediction, which is simply the mode of the response values for the training observations in R_m.

Regions R_m are found by minimizing an error function. In a node m, representing a region R_m with $N_m = |\{x_i \in R_m\}|$ observations, let

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

the proportion of class k observations in node m. The observations in node m are associated to class $k(m) = \underset{k}{argmax}\, \hat{p}_{mk}$, the majority class in node m. Classification error rate can be defined in different ways, for example through misclassification error function

$$E_m = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)} \tag{2.7}$$

To find the best binary partition, a greedy algorithm is applied on the data. Consider a splitting variable j and split point s, and define the pair of half-planes

$$R_1(j,s) = \{X|X_j \leq s\} \qquad R_2(j,s) = \{X|X_j > s\}.$$

The splitting variable j and the split point s that solve $\underset{j,s}{min}(E_1(j,s)+E_2(j,s))$ identify the best split. Once the solution has been found, data are partitioned into the two resulting regions and the splitting process is repeated on both regions. Then the process is repeated on all of the resulting regions.

Tree size is a tuning parameter ruling the model's complexity and the optimal tree size should be chosen depending on the data. One approach would be to split tree nodes only if the decrease in misclassification error due to the split exceeds some threshold. However, according to [49], the most common strategy is growing a large tree T_0 , stopping the splitting process only when some minimum node size is reached. This procedure is strictly related to the concept of pruning. We define a subtree $T \subset T_0$ to be any tree that can be obtained by pruning T_0 , that is, collapsing any number of its internal (non-terminal) nodes. Let index the terminal nodes by m, with node m representing region R_m . Besides, let $|T|$ denote the number of terminal nodes in T. The node impurity measure can be defined as

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)},$$

while the cost complexity criterion can be defined as

$$C_\alpha(T) = \sum_{m=1}^{T} N_m Q_m(T) + \alpha|T|.$$

For each α, it can be found the subtree $T_\alpha \subset T_0$ to minimize $C_\alpha(T)$. The tuning parameter $\alpha \geq 0$ represents the tradeoff between tree size and the goodness of fitting the data. Large values of α result in smaller trees T_α and conversely for smaller values of α. Note that, if $\alpha = 0$ the solution is the full tree T_0. For each α it can be shown that there is a unique smallest subtree T_α minimizing $C_\alpha(T)$ [49]. Estimation of α is achieved by five- or ten-fold cross-validation, choosing the value $\hat{\alpha}$ that minimizes cross-validated $C_\alpha(T)$. $T_{\hat{\alpha}}$ is the final tree.

Another measure of node impurity and classification error rate is the Gini index, defined as

$$G_m = \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}), \tag{2.8}$$

Differently from misclassification error, Gini index is differentiable, and hence more amenable to numerical optimization. In addition, misclassification error is not sufficiently sensitive for tree-growing ([49]). Interestingly, rather than classifying observations to the majority class in the node, they can be classified to class k with probability $\hat{p}_{mk}$. Then the training error rate of this rule in the node is $\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'}$, *i.e.,* G_m. Similarly, if each observation is coded as 1 for class k and 0 otherwise, the variance over the node of this 0-1 response is $\hat{p}_{mk}(1-\hat{p}_{mk})$. Summing over classes k, it results again the Gini index.

Notably, trees can be non-robust, meaning that a small change in the data can cause a large change in the final estimated tree. However, by aggregating many decision trees, as in Random Forest, the predictive performance of trees can be substantially improved.

Thus, the same classification tree is fit many times to bootstrap-sampled versions of the training data and the result is averaged. Bootstrap methods randomly draw datasets with replacement from the original data, each sample same-sized as the training set. This is repeated B times, producing B bootstrap datasets.

The basic Random Forest algorithm can be summarized as follows

1. For $b = 1, \ldots, B$:

   - Draw a bootstrap sample Z^* of size n from the training data.
   - Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

  - Select m variables at random from the p variables;
  - Pick the best variable/split-point among the m;
  - Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

In order to classify a new point x, let $\hat{C}_b(x)$ be the class prediction of the bth Random Forest tree; then $\hat{C}_{rf}^B(x) = majority\ vote\,\{\hat{C}_b(x)\}_1^B$.

An average of B independent and identically distributed (i.i.d) random variables, each with variance σ^2, has variance $\frac{1}{B}\sigma^2$. If the variables are simply i.d. (identically distributed, but not necessarily independent) with positive pairwise correlation ρ, the variance of the average is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

As B increases, the second term disappears, but the first remains and hence the size of the correlation of pairs of trees limits the benefits of averaging [49]. The aim of Random Forest is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. Random selection of the input variables in the tree-growing process allows to achieve this. Specifically, when growing a tree on a bootstrapped dataset, before each split, $m \leq p$ of the input variables are selected at random as candidates for splitting. For classification, the default value for m is $\lfloor\sqrt{p}\rfloor$ and the minimum node size is one. In this thesis, Random Forest will be used in the implementation of the scikit-learn v.0.17.1 Python module [51]. Details on the specific models will be given when describing the specific metagenomic models.

2.3 Data Analysis Protocol

The development of predictive models can be affected by several sources of variability and bias effects, arising from choices hidden in modeling path. For instance, a model can perform well on training data, but achieve a poor predictive performance on unseen data (“overfitting”). This serious pitfall can be due to procedural errors in the Data Analysis Plan, such as testing the model on part or all the training data. Besides, training and test datasets should be representative of a generic sampling from a data population, *i.e.,* training and test are assumed to be independently and identically (iid) drawn from the same distribution. If this is not the case, and no corrections are performed, predictive modeling is affected by “’selection bias”. This is a key issue because predictive markers and, in general, the conclusions drawn from predictive profiling should be reproducible across different studies; this is crucial specifically for a valid clinical application of the biological findings.

In order to overcome these potential issues, the U.S. FDA-led initiatives MAQC-II [53] and SEQC [54] provided a set of guidelines for the development of predictive models on microarray and Next-Generation Sequencing (NGS) data, respectively. These projects established standardized steps in training, model selection and validation on novel data [53], summarized inside a Data Analysis Protocol (DAP). In this thesis, each predictive model is developed inside the FDA MAQC DAP, which is detailed step by step in the next paragraphs; a schematic representation is reported in Fig. 2.1.

2.3.1 DAP scheme

Consider a dataset composed of p variables (*i.e.,* microbial abundances) measured for n samples that can be associated to two classes, coded as label 0 (*i.e.,* healthy subjects) and label 1 (*i.e.,* diseased patients). The dataset is then split in *Training data* and *Validation data*, maintaining the same proportion of classes in each partition. In our analyses, training and validation were 70% and 30% of the entire dataset, respectively.

Stratified 10×5-fold Cross-Validation schema

1. First, the training data is split in $4/5$ as internal training set (int_tr) and $1/5$ as internal validation set (int_val) using a stratified 5-fold Cross-Validation (CV), this operation repeated 10 times. At the end, we will have 50 configurations of int_tr and int_val set. In detail, int_tr sets are used for classifier development, int_val sets are only used for classifier performance evaluation.

2. For each iteration, classifier parameters (i.e. hyperparameter C for SVM and number of trees for RF) are tuned through the following procedure:

   (a) for each possible parameter value to be tuned (the range is provided), use 10 Monte Carlo CV cycles to split original int_tr into two partitions (50%-50% tuning training-test proportions);

   (b) after the split, training and test tuning data are scaled through normalization;

   (c) build model (i.e. SVM or RF) on training tuning data and evaluate it on the test tuning data in terms of MCC (see Paragraph 2.3.3);

   (d) choose the parameter(s) maximizing the average MCC on the test tuning set (w.r.t. Monte Carlo CV cycles).

3. int_tr and int_val are scaled as in the tuning part.

4. The model (i.e. SVM or RF) is built on int_tr with the selected parameter(s)

and features are ranked according to weights computed either by the model or by other feature ranking methods such as Relief [55], ANOVA F-score, extraTrees [56].

5. Once the ranked list of discriminant features has been built, increasing sets of features are selected (feature steps: p = 1, 2, .., 10, 20, .., 100, 200, .., P_max), and for each set of features, a new model (*i.e.,* RF or SVM) is developed (using the same parameter coming from the tuning part).

6. This model is tested on the int_val and different performance metrics are computed (MCC, accuracy, specificity, sensitivity, etc.). In our analyses, we will consider MCC (see [57]) as a measure to evaluate the predictive performance of models.

7. At the end, for each metric, a matrix composed of 50 rows (from 510-CV splits) and $|Fs|$ (length of Fs) columns is produced. The mean w.r.t. rows of these metric matrices is computed, together with the 95% bootstrap confidence intervals; hence, for each metric the DAP provides an array of length $|Fs|$ corresponding to the mean value of that metric across the 50 set partitions.

8. The optimal method parameter is the one that occurred most frequently in the 50 cycles of the CV. Furthermore, the optimal number of features (n_opt_feat) is the one maximizing the mean MCC; the maximum MCC is denoted as MCC_int.

9. The rank of features over all CV cycles is computed through the Borda algorithm (see Paragraph 2.3.4) using the ranking matrix as input, composed of 50 rows (as many as the cycles of the 10×5-fold CV) and p columns (as

many as the features). In each row, the indexes of features inside int_tr are reported, sorted according to the weights computed at point 4.

10. Finally, for each feature, sorted according to the ranking given by Borda, the median value of that feature over all samples is computed, together with its median value over samples of first class ($medf\_1$) and over samples of second class ($medf\_2$). Besides, fold-change (ratio between $medf\_2$ and $medf\_1$) and its log base 2 are listed.

11. The "best" model is built by using the optimal parameter(s) and that features at the first n_opt_feat positions of the Borda list.

Validation schema

1. First, *Training* and *Validation data* are restricted considering only the "best" n_opt_feat features.

2. The *Training* and *Validation data* are scaled (as in tuning).

3. The model is built on the restricted training dataset.

4. The model is tested on the restricted validation dataset.

5. MCC is computed for the *Training* as well as for the *Validation* data if labels are available; MCC for the validation set is indicated with MCC_{val}.

2.3.2 Check through randomization

In order to detect possible bias effects (*i.e.,* overfitting, selection bias) of the predictive model, two approaches are implemented inside the DAP.

1. **Random ranking:** The model is trained on features that are ranked randomly rather than according to the model itself. Hence, consider mean MCC varying across Fs: if the model is not overfitting data, it must result as a growing

function with respect to Fs. Moreover, it must result that the best MCC, which probably will be similar to the one of the not randomized process, must be reached using all features.

2. **Random labels:** The associations between samples and their labels are randomly shuffled. Hence, if the model is not overfitting data, mean MCC varying across Fs must oscillate around zero, meaning that the prediction is actually random.

DAP is implemented in Open Source Software: it combines a suite of machine learning tools from the MLPY [58] and scikit-learn [51] Python libraries with in-house Python and R scripts.

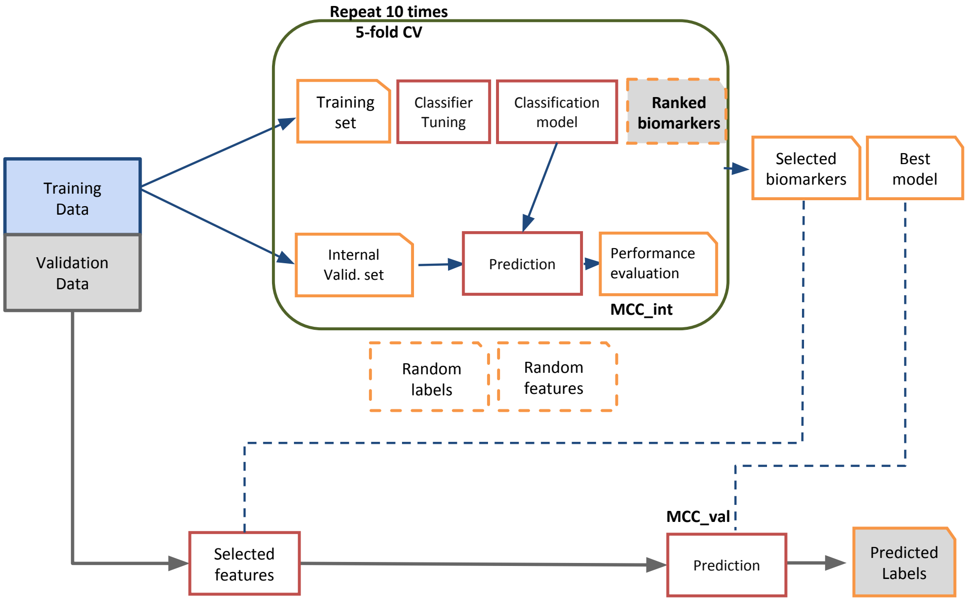


Figure 2.1. Data Analysis Protocol for predictive models development.

2.3.3 Matthews correlation coefficient

MCC is a metric that summarizes the confusion matrix into a single value, used as a reference performance measure on unbalanced data sets [57].

Let $\mathscr{S} = \{s_i : 1 \leq i \leq S\}$ the set of samples belonging to N classes $\{1,\ldots,N\}$. We define the two functions

$$tc, pc : S \to \{1,\ldots,N\}$$

indicating for each sample s its true class tc and its predicted class pc, respectively. Moreover, let $C \in \mathbb{N}^{N\times N}$ the confusion matrix, so that each C element is defined as

$$C_{ij} = |\{s \in S : tc(s) = i, pc(s) = j\}|.$$

Let $X, Y \in \mathbb{F}_2{}^{S\times N}$ two matrices defined as

$$X_{sn} = \begin{cases} 1 & \text{if} \quad pc(s) = n \\ 0 & \text{if} \quad pc(s) \neq n \end{cases}, \qquad Y_{sn} = \begin{cases} 1 & \text{if} \quad tc(s) = n \\ 0 & \text{if} \quad tc(s) \neq n \end{cases},$$

From the definition, it results

$$C_{kk} = \sum_{s=1}^{S} X_{sk} Y_{sk}, \qquad C_{kl} = |\{s \in S : X_{sk} = 1 \,\text{and}\, Y_{sl} = 1\}|$$

The covariance function between X and Y can be written as follows

$$\text{Cov}(X,Y) = \frac{1}{N} \sum_{s=1}^{S} \sum_{k=1}^{N} (X_{sk} - \bar{X_k})(Y_{sk} - \bar{Y_k}),$$

where $\bar{X_k}, \bar{Y_k}$ are the means of the $k-$column, that is

$$\bar{X_k} = \frac{1}{S} \sum_{s=1}^{S} X_{sk} = \frac{1}{S} \sum_{l=1}^{N} C_{kl}, \qquad \bar{Y_k} = \frac{1}{S} \sum_{s=1}^{S} Y_{sk} = \frac{1}{S} \sum_{l=1}^{N} C_{lk}.$$

Then, Matthews Correlation Coefficient can be written as

$$\begin{aligned}\text{MCC} &= \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \frac{\sum_{k,l,m=1}^{N}(C_{kk}C_{ml} - C_{lk}C_{km})}{\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{lk}\right)\left(\sum_{f,g=1,f\neq k}^{N} C_{gf}\right)\right]}\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{kl}\right)\left(\sum_{f,g=1,f\neq k}^{N} C_{fg}\right)\right]}}.\end{aligned}$$

MCC ranges in $[-1,1]$, where 1 means perfect classification, -1 is asymptotically reached in extreme misclassification case (all zeros but in two symmetric entries), 0 when C is all zeros but for one column or when all entries are equal (random classification).

Moreover,in case of binary classification, MCC can be written as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{2.9}$$

The formula (2.9) is the one used in this thesis, since all classification problems here considered are binary.

2.3.4 Borda algorithm

The Borda algorithm is a computational method for comparing sets of ranked biomarker lists developed by [59] based on concepts from permutation group theory.

Let D the dataset composed of n samples described by a set $\mathscr{F}$ of p features ($\{\mathscr{F}_j\}_j$ $j = 1 \dots p$). Let B the number of replicated experiments required by complete validation of profiling. They consist of instances of classification and feature ranking, they are also called runs. At each replicate ($i = 1 \dots B$) the ranking process sorts the features according to their importance in building the ith classifier and an ordered list L_i is produced. Let $\mathscr{L} = \{L_i\}_{i=1}^{B}$ the set of all the ordered list produced. Let L_i^k be its top-k list, i.e. the sublist consisting of the first k ranked elements from L_i. Let defined as $\tau_i(j)$ the rank (position) of feature F_j in L_i, as $\tau_i = (\tau_i(j))_{j=1}^{p}$ the dual list of L_i.

For each feature F_j, we define its top-k extraction set

$$E_k(j) = \{i = 1 \dots B : \tau_i(j) \leq k\}.$$

We define the *extraction number* of F_j as the number number of elements in $E_k(j)$

$$e_k(j) = |E_k(j)|,$$

which indicates how many top-k lists include a feature.

For each feature F_j, we define the (average) *position number* as

$$a_k(j) = \frac{1}{e_k(j)} \sum_{i \in E_k(j)} \tau_i(j).$$

For a given k, $e_k(j)$ and $a_k(j)$, induce a ranking of the features: high e_k and low a_k indicate features extracted often in top positions.

The purpose is to encode the ranking information coming from all the lists in $\mathscr{L}$ into a single optimal list. Then, for each k, consider the extraction number $e_k(j)$ in decreasing order as the ranking criterion; if for some $i \neq j$ it results $e_k(j) = e_k(i)$, then consider the position number $a_k(j)$ in increasing order. This criterion defines a dual list τ_o^k called *optimal top-k list* of $\mathscr{L}$. If $k = p$, the complete lists are considered, hence $e_p(j) = B$ for all F_j and $\tau_o = \tau_o^p$ is determined only by $a_p(j)$.

The optimal list definition is linked to the *Borda count*. Given a set of B ranked lists on p candidates, the Borda count associates to each candidate F_j a score $s(j)$ given by the total number of candidates with higher position over all lists. The Borda optimal list is then derived by ranking candidates with higher scores. Ranking according the increasing order of $a_k(j)$ is equivalent to ranking for the decreasing order of $s(j)$, indeed

$$\begin{aligned}
s(j) &= \sum_{i=1}^{B} |\{t : 1 \leq t \leq p \quad \tau_i(t) > \tau_i(j)\}| \\
&= \sum_{i=1}^{B} (p - \tau_i(j)) = Bp - B\frac{1}{B}\sum_{i=1}^{B} \tau_i(j) \\
&= B(p - a_p(j)).
\end{aligned}$$

Implementation

The DAP described in paragraphs 2.3.1, 2.3.2 and 2.3.4 has been implemented through Python functions organized in a system of scripts. It has been described in [53, 54] and originally directly implemented in MLPY [58]. In this thesis, the author has contributed by a general revision of several main features of the MAQC DAP. Specific adaptations for metagenomic are available for the paper “Integrative Network Fusion of bacterial-fungal microbiota for the identification of robust IBD biomarkers” (submitted 2017).

Chapter 3

Network analysis

A microbiota is a collection of microorganisms that do not live in isolation, but instead actively interact with one another and aggregate to form heterogeneous communities [60, 61]. Interspecies relationships reflect the overall function of the microbial communities and can be modulated by ecological competition/cooperation between the microbes [62, 60], direct secretion of substances such as bacteriocins [63], or indirect interactions through immune system modulation [64]. Besides, several studies found specific microbial interactions in healthy commensal microbiota conferring resistance against pathogens, thus providing overall stability of communities [65, 66, 67, 68]. Conversely, other microbial relationships, often due to perturbation-induced shifts in the commensal ecological networks, were associated with impaired macroscopic functionality and ill-health [69, 12, 60, 70].

However, studies have initially focused more on alterations in abundance of microbial species, rather than on changes in larger-scale interspecies relationships. Only in recent years, several methods have been designed to investigate ecological organization and functional relationships of microbial communities. A widely used approach is to build mathematical models, graphs in particular, to evaluate relationships among microbes (Paragraphs 3.1 and 3.2) and to assess changes inside microbial communities due to disease-linked perturbations (see Paragraphs 3.3 and 3.4). Modelling microbiota community organization and dynamics may be a fruitful approach also for biomedical applications, allowing the simulation of the treatment with pre- and probiotics and investigating their impact on microbiota and host response [71, 72, 73].

3.1 Microbial networks

A microbiota is a complex ecosystem where species interact with one another [60], establishing different types of relationships that are categorized by their effect on the species involved, *i.e.,* positive, negative or neutral. Macroecology in this context defined by interactions as mutualism (positive-positive), commensalism (positive-neutral), antagonism (positive-negative), competition (negative-negative), amensalism (negative-neutral) and neutralism (neutral-neutral) [71].

Specifically, microbial mutualism is known to lead to biofilm development, increased levels of antibiotic resistance and adaptation to the environment [74, 75]. These synergic interactions are mediated by adhesins, which are membrane-bound structures recognizing specific receptors on microbial or host-associated surfaces [76]; microbiota components specialized in adhesin production are described as “bridging organisms” because they potentially aggregate community members that normally cannot bind to each other. A notable example is *Fusobacterium nucleatum*, known to attach to many different members of the oral community leading to the development of biofilms. These polymicrobial interactions during an infection result in worsened disease compared to infections involving an individual pathogenic microbe alone [74]. For instance, bacterium *Staphylococcus aureus* and pathogenic fungus *Candida albicans*, often co-isolated from both chronic and acute infections, represent a well-known example of synergy in disease. Murray and colleagues studied a mouse tongue epithelium *ex vivo* model and showed that only *C. albicans* could penetrate and colonize the subepithelium, while *S. aureus* was found in the subepithelium only when aggregated to *C. albicans*. The direct interaction between the bacterium and the fungus is mediated by Als3p, which is a specific *C. albicans* protein [74]. Other common forms of mutualism are cross-feeding interactions (also known as synthropy), in which nutrients excreted by some species are absorbed and metabolised by other species in the community [77]. An example of bi-directional cross-feeding interactions is provided by Moens and colleagues [78], who studied commensalism between *Faecalibacterium prausnitzii* and *Bifidobacteria*. *F. prausnitzii* is a colon bacterium that has been linked to health-promoting benefits for the host; this *Clostridium* cluster IV bacterium ferments complex carbohydrates, in particular oligofructose and inulin, with a consequent production of butyrate, which has a protective role on colon epithelial cells inducing the differentiation of regulatory T cells. On one hand, *F. prausnitzii* growth requires acetate as a mandatory co-substrate, mainly provided by acetate-producing *Bifidobaceria*; on the other hand, bifidobacterial strains need fructose as a substrate, but they are not capable of degrading oligofructose. Thus, *F. prausnitzii* cross-feeds fructose to *Bifidobacteria*, given its capacity of degrading oligofructose or performing a preferential degradation of short chain length fractions of oligofructose.

Antagonism is another common type of microbial interactions, which occurs when some species can exist only in absence of others. Consider, for instance, microbiota of breast-fed infants: it is dominated by *Bifidobacterium*, the principal consumer of human milk oligosaccharides (HMOs). Most strains of *Bifidobacterium* first import and then degrade HMOs by intracellular glycoside hydrolases. Consequently, the growth of competitor strains is limited by simple sequestration of available sugar substrates in the colon, protecting the neonate from possible pathogens [79]. Other classical antagonistic interactions are predator-prey (*i.e.*, ciliates feeding on bacteria) and host–parasite (*i.e.*, between bacteria and their bacteriophages) relationships [61].

Several studies have shown that specific ecological interactions within microbiota are crucial for community stability in the healthy commensal microbiota [66, 67, 68], while many others are involved in dysbiosis and disease [80, 81, 82]. Thus, a paradigm shift is needed, from a reductionist approach that focuses on individual microbes to more holistic approaches focusing on interactions among members of microbiota. Traditionally, the study of microbial relationships required the use of laboratory experiments such as growth and co-culture assays [83, 84], which could not be extended to large-scale applications. Computational methods to model microbial interactions alleviated this issue, by predicting candidates for experimental validation [85, 72]. Besides, computational approaches could provide knowledge-based databases with experimentally verified interactions from published literature. One of the most promising approach is network theory, which is specifically intended to represent and model the complexity of microbiota with multifaceted interactions between its members.

3.2 Networks inference

The construction of ecological network from presence-absence or abundance microbiome data is known as network inference; this method provides a ‘snapshot’ of the microbial community status at a given time. At a highly abstract level, a network is group of two or more objects (defined as nodes) connected to each other by links, with each link representing the interactions between two components. The nature of the interactions defines the network as directed or undirected. In directed networks, interactions between nodes have a well-defined direction, which can be used, for instance, to model direction of material flow from a substrate to a product in a metabolic reaction. In undirected networks, the links do not have an assigned direction; for example, in protein interaction networks, undirected link represents a mutual binding relationship [86].

Several methods have been proposed for network inference, depending on efficiency, accuracy, speed, and computational requirements, as well as on the specific microbial community aspect of interest. A popular approach focuses on co-occurrence or co-exclusion of species, modeling strong dependency (*i.e.*, positive interactions including mutualism and commensalism) or competition (i.e. negative interactions including competition, antagonism and amensalism) among them (see Paragraph 3.1). The detection of such patterns can be formulated into the computation of dependency measures among distributions of all species pairs. Commonly used measures include similarity (*e.g.*, mutual information), dissimilarity (*e.g.*, Kullback–Leibler) and correlation (*e.g.*, Pearson or Spearman).

In this thesis, co-occurence/co-exclusion networks are inferred by computing correlation between the abundance profiles of species, that is a widely used approach in literature [87, 88, 89]. Consider a table of microbial abundances, typical output of Next-Generation Sequencing (NGS) data analysis pipeline. The data are stored in a matrix $W \in \Upsilon^{n \times p}$, where $\mathbf{w}^j = [w_1^j, w_2^j, ..., w_p^j]$ denotes the p-dimensional row vector of microbial abundances from the j^{th} sample, $j = 1, ..., n,$; Υ denotes the set $\{0, 1, 2, ...\} \in \mathbb{N}$ or the set $[0, 1] \subseteq \mathbb{R}$, depending on the nature of abundances, being absolute or compositional, respectively. The objective is to build a network of pairwise associations, represented as an undirected graph $G = (V, E)$, where the node set $V = \{v_1, \ldots, v_p\}$ represents the p microbial taxa and the edge set $E \subset V \times V$ the possible associations among them. The graph G is inferred starting from abundances matrix W and computing the Pearson correlation coefficient (PCC) among each pair of taxa, resulting in a matrix $C^{p \times p}$. The PCC between two variables is defined as the covariance of the two variables divided by their standard deviations and it captures linear dependencies. In more detail, PCC between taxa w_i^j and w_k^j is:

$$PCC = \frac{\sum_{j=1}^{n} w_i^j w_k^j - \frac{(\sum_{j=1}^{n} w_i^j)(\sum_{j=1}^{n} w_k^j)}{n}}{\sqrt{(\sum_{j=1}^{n}(w_i^j)^2 - \frac{(\sum_{j=1}^{n} w_i^j)^2}{n})(\sum_{j=1}^{n}(w_k^j)^2 - \frac{(\sum_{j=1}^{n} w_k^j)^2}{n})}}, \tag{3.1}$$

where n is the number of samples.

Notably, the use of Pearson correlation to detect dependencies between members of a microbiome is common on absolute abundances, but it is sensitive to compositionality [81, 90]. Indeed, if the abundances of all taxa are constrained by a constant sum (*e.g.*, one), an increase in the relative abundance of one taxon will lead to a decrease in the abundance of all others, leading to spurious correlations. Therefore, PCC on relative abundances can lead to negative correlations and thus false interaction predictions; such a bias is known as the compositional effect [65, 91]. In order to successfully use correlations to infer interactions on compositional data, this bias is corrected by an approach based on Aitchison's centered log-ratio (clr) transformation [91, 92]. In detail, a clr-transformation involves computing the logarithm of the ratio between the relative abundance and the geometric mean of all relative abundances within sample j:

$$clr(\mathbf{w}^j) = [log(\frac{w_1^j}{g(\mathbf{w}^j)}), log(\frac{w_2^j}{g(\mathbf{w}^j)}), ..., log(\frac{w_p^j}{g(\mathbf{w}^j)})]^T = \mathbf{G} \cdot log(\mathbf{w}^j) \tag{3.2}$$

where: $\mathbf{w}^j = [w_1^j, w_2^j, ..., w_p^j]^T$ is a column vector representing the relative abundances of taxa in sample j; $g(\mathbf{w}^j) = (\prod_{i=1}^{p} w_i^j)^{\frac{1}{p}}$; $\mathbf{G} = \mathbf{I}_p - \frac{1}{p}\mathbf{J}_p$; $\mathbf{I}_p$ is the p-dimensional identity matrix; and $\mathbf{J}_p$ a p-dimensional matrix fill with 1s. Function in 3.2 removes the unit-sum constraint of compositional data, transforming data from a constrained space with p dimensions to a $(p-1)$-dimensional Euclidean space. In this thesis, CCLasso (Correlation inference for Compositional data through Lasso) is adopted to infer correlation network from compositional data [93]. CCLasso uses least squares with L1 penalty after a clr-transformation of raw compositional data to es-

timate the correlation matrix C for all pairs of taxa. The graph G of taxon-taxon associations is thus built from clr-transformed microbiome compositions $Z \in \mathbb{R}^{n \times p}$.
After co-occurrence/co-exclusion network has been inferred, microbial interactions associated to highest correlation values are analyzed, corresponding to the most ecologically informative associations. A null model for co-expression networks proposed by Gobbi and Jurman [94] is adopted, in order to compute a correlation threshold minimising the possible false positive links, paying a price in terms of false negative detected edges. Specifically, this is an a priori model based on the work of Fisher [95] and Bevington [96], that depends on the dimensions of the starting data matrix, assuming the skewness of the data distribution is compatible with the structure of abundances data.
In summary, network inference approach adopted in this thesis can be divided into three steps:

1. Compute Pearson Correlation Coefficient or CCLasso on absolute or relative microbiota abundances, respectively.

2. Select the taxa-taxa interactions with correlation higher than a threshold computed by Gobbi and Jurman model [94].

3. Build a graph with microbial taxa as nodes, correlation over Gobbi's threshold as edges.

In this thesis, network inference provides a static model of the interactions within microbiota at a given time and for a group of samples with a common phenotype (*i.e.*, health status). In order to study the microbial networks perturbations associated to different phenotypes, further analyses are required such as network distances and communities detection (see Paragraph 3.3 and 3.4, respectively).

3.3 Networks distance

The aim of microbial network analysis is unravelling the interactions among species that are either beneficial for the host or specifically linked to disease. To this purpose first network inference and then differential network analysis (netDA) are considered. In summary, netDA consists in the comparison of networks corresponding to different phenotypes or conditions. The best way to deal with similarity and dissimilarity between networks is to define a distance; the two most relevant families of graph distances are spectral measures and the edit distances. Edit distances are based on functions of insertion and deletion of matching links between the compared graph, evaluating the minimum cost of transformation of one graph into another; spectral measures define a suitable similarity measure on the topology of the underlying graphs, based on functions of the eigenvalues of one of the graph connectivity matrices.

In this thesis, netDA is based on the Hamming-Ipsen-Mikhailov (HIM) distance [97, 98], which linearly combines two distances, the Hamming [99, 100, 101] and the Ipsen-Mikhailov [102]; the first is an edit distance, while the latter is a spectral measure. The Hamming distance is the simplest member of the family of edit distances and it focuses on the links as independent entities, disregarding the overall structure. Conversely, the Ipsen-Mikhailov is a more reliable and stable global measure [97], evaluating the differences between the whole network structures, but it cannot discriminate between isospectral non-identical graphs. Thus, HIM overcomes the drawbacks of local (edit) and global (spectral) metrics when separately considered.

The HIM family of distances

Let $\mathscr{N}_1$ and $\mathscr{N}_2$ be two simple networks on N nodes, inferred from the corresponding adjacency matrices $A^{(1)}$ and $A^{(2)}$, with $a_{ij}^{(1)}, a_{ij}^{(2)} \in \mathscr{F}$, where $\mathscr{F}$=$\mathbb{F}_2$={0,1} for unweighted networks (links are unweighted) and $\mathscr{F}$=[0,1] for weighted graphs. Besides, let $\mathbb{I}_N$ be the $N \times N$ identity matrix, let $\mathbb{1}_N$ be the $N \times N$ unitary matrix with all entries equal to one and let $\mathbb{O}_N$ be the $N \times N$ null matrix with all entries equal to zero. Define then ε_N as the empty network with N nodes and no links (with adjacency matrix $\mathbb{O}_N$) and $\mathscr{F}_N$ as the clique (undirected full network) with N nodes and all possible $N(N1)$ links, whose adjacency matrix is $\mathbb{1}_N - \mathbb{I}_N$.

The Hamming distance. The (normalized) Hamming distance is the (local) simplest edit metric, counting the presence/absence of matching links on the two networks being compared:

$$H(\mathscr{N}_1,\mathscr{N}_2) = \frac{Hamming(\mathscr{N}_1,\mathscr{N}_2)}{Hamming(\varepsilon_N,\mathscr{F}_N)} = \frac{1}{N(N-1)} \sum_{1\le i\neq j\le N} |A_{ij}^{(1)} - A_{ij}^{(2)}|. \tag{3.3}$$

H ranges in the interval $[0,1]$, where the lower bound 0 is reached only for identical networks $A^{(1)} = A^{(2)}$, while the upper is attained whenever the two networks are complementary $A^{(1)} + A^{(2)} = \mathbb{1}_N - \mathbb{I}_N$. Note that, for H, all links are equivalent regardless of their position within the network.

The Ipsen-Mikhailov distance. The Ipsen-Mikhailov distance is the (global) L2 integrated difference of the Laplacian spectral densities:

$$IM(\mathscr{N}_1,\mathscr{N}_2) = \sqrt{\int_0^\infty [\rho_{\mathscr{N}_1}(\omega,\bar{\gamma}) - \rho_{\mathscr{N}_2}(\omega,\bar{\gamma})]^2 d\omega} \tag{3.4}$$

where $\rho_{\mathscr{N}_1}(\omega,\gamma)$ and $\rho_{\mathscr{N}_2}(\omega,\gamma)$ are defined as spectral densities of nodes $\mathscr{N}_1$ and $\mathscr{N}_2$, respectively [98]. By definition, IM too ranges between 0 and 1, with upper bound reached only for $\{\mathscr{N}_1,\mathscr{N}_2\} = \{\varepsilon_N,\mathscr{F}_N\}$. In fact, IM cannot distinguish isospectral (non isomorphic) networks, since it is a spectral measure.

The Hamming-Ipsen-Mikhailov distance. The normalized Cartesian product of H and IM defines the Hamming-Ipsen-Mikhailov (HIM) distance:

$$HIM_{\xi}(\mathscr{N}_1,\mathscr{N}_2) = \frac{1}{\sqrt{1+\xi}}\sqrt{H^2(\mathscr{N}_1,\mathscr{N}_2)+\xi\cdot IM^2(\mathscr{N}_1,\mathscr{N}_2)}, \tag{3.5}$$

where $\frac{1}{\sqrt{1+\xi}}$ is a normalizing factor, with $\xi \in [0,+\infty)$.

The normalization bounds the range of the HIM distance in the interval $[0,1]$, with lower bound reached for every couple of identical networks, and upper bound attained only on the pair $(\varepsilon_N,\mathscr{F}_N)$. Moreover, for non-identical isomorphic/isospectral graphs, all distances HIM_{ξ} will be nonzero.

Notably, network differential analysis based on HIM distance has been adopted in metagenomics [103], liver high-throughput oncogenomics [104], oncoimmunology [105], but also out of computational biology, *e.g.*, socioeconomics [98] or even in multiplex network theory [106].

The quantification of networks differences can associate shifts in microbial interactions with phenotypic changes; moreover, coupling network distance with a description of microbial community organization provides even a more comprehensive analysis. Indeed, microbiota is structured as a set of communities, so revealing the modular structure of microbial networks will provide invaluable insights into biologically relevant clusters characteristic of specific phenotype.

In this thesis networks inference and distance are implemented by in-house R scripts based on the *nettools* and *igraph* packages.

3.4 Community detection

Any graph can be decomposed into elementary units known as clusters (also defined as modules), which are sets of highly inter-connected nodes [107]. The identification of clusters within networks from microbiota may be seen, for example, as modeling groups of coexisting or coevolving microbes contributing towards a disease; besides, clusters detection unravels the local interaction patterns in the network and their contribution to the overall structure, connectivity, and function of the network. The key concept of community detection, which are clustering methods developed specifically for networks, is partitioning the graph into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected. [108]. In this thesis, the rationale behind community detection is that clusters of co-occurring species are commonly distorted in disease and alterations are most prominent in clusters containing a predominance of pathogenic organisms [109, 67, 12, 68].

Several types of community detection algorithms have been proposed in literature: agglomerative algorithms merge similar nodes/communities recursively [110], divisive algorithms detect inter-community links and remove them from the network [111, 112] and optimization methods are based on the maximisation of an objective function [113, 114]. The performance of these methods is often measured by the *modularity* of the partition: it is a scalar value in $[-1,1]$ that measures the density of links inside communities as compared to links between communities [111, 115]. In more detail, the modularity index Q is defined as [116]:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i,c_j), \tag{3.6}$$

where A_{ij} is the weight of the edge between i and j, $k_i = \sum_j A_i j$ is the sum of the weights of the edges attached to vertex i, c_i is the community where vertex i is included, the δ-function $\delta(u,v)$ is 1 if $u = v$ and 0 otherwise and $m = \frac{1}{2}\sum_{ij} A_{ij}$. In this thesis, community detection is performed by Louvain method [108], which is a greedy algorithm that finds partitions via modularity maximization (Equation 3.6). The algorithm is divided in two phases that are repeated iteratively; a scheme with the steps of the algorithm is shown in Fig. 3.1. Consider a weighted network of N nodes; first, network is partitioned into as many communities as nodes. Then, consider one node i and its neighbours j: node i is removed from its community and placed in the community of each j, in turn. For each assignment, the gain of modularity is evaluated and the node i is then placed in the community for which this gain is maximum, but only if this gain is positive. No positive gain, no assignment of i. This process is applied repeatedly and sequentially for all nodes, stopping when no further improvement can be achieved: the first phase is then complete. The communities found during the first phase are considered as the nodes of a new network, which is built during the second phase. The links connecting the new nodes are given by the sum of the weight of the links between nodes in the corresponding communities from the first phase. At the end of this second phase, it is possible to apply again the first phase of the algorithm to the resulting weighted network and to iterate. The algorithm is iterated until there are no more changes and a maximum of modularity is achieved. Louvain algorithm has been chosen for community detection because its steps are intuitive and easy to implement, and the outcome is unsupervised. Besides, the algorithm is fast, *i.e.*, the complexity is linear on typical and sparse data [108], since the number of communities decreases drastically after just a few passes so that most of the working time is spent on the first iterations. Moreover, the resolution limit problem of modularity (large network size) is overcome by the intrinsic multi-level nature of Louvain algorithm.

In this thesis, community detection is implemented through *community* and *networkx* Python modules.

Collectively, network differential analysis and community detection offer a comprehensive view of the conserved and variable architectures present in healthy and diseased microbiomes, such as complex polymicrobial interactions and co-occurrence patterns. From a clinical perspective, it would interesting to adopt these mathematical models of microbial communities to gain insight into the possible effects of a

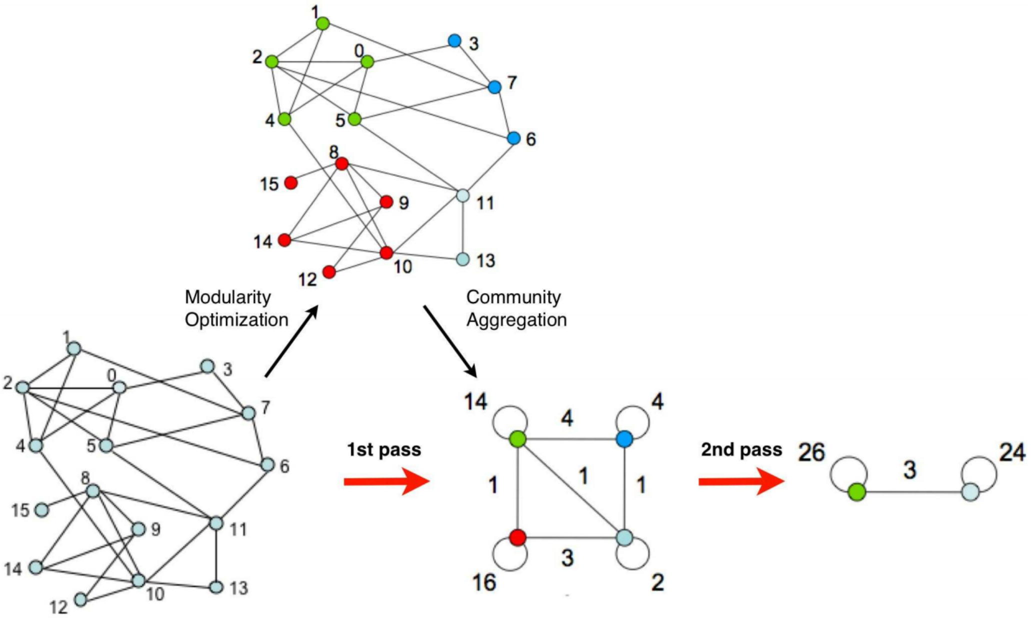


Figure 3.1. Steps of Louvain algorithm. Each pass consists of two phases: one for modularity maximization by local changes of communities, one for communities aggregation in order to build a new network. The passes are iterated until no more changes of modularity is attained. Figure extracted from [108]: the weights of the links are shown on the network after each phase.

disease-linked perturbation or a broad-spectrum antibiotic use.

Chapter 4

Meta-omics integration

Recent advances in high-throughput DNA sequencing, mass spectrometry and RNA-Seq methods, along with computational and algorithmic methods, have increased the accessibility of highly multivariate and heterogeneous datasets ("meta-omic" datasets) on biological systems. Meta-omics include various biological molecules at the metagenomic scale (metaproteomics, metatranscriptomics, metametabolomics) as well as different "microbial" omes (eukaryotes and viruses). Combining information from different biological components aims at a system-level understanding of microbiota-host ecosystem, unravelling complex processes involved in health and microbiota-related disease. The integrated analyses of meta-omics data can be performed through top-down or bottom-up approaches. Top-down modeling takes advantage of high-throughput meta-omics datasets aiming to predict novel biological hypotheses, which must be experimentally validated. Conversely, bottom-up approaches are formulated at the molecular level based on detailed mechanistic knowledge [117].

In this thesis, a top-down meta-omics data integration is adopted, as defined by Ritchie and colleagues [118]: the combination of multiple meta-omics datasets to develop classification models that are predictive of complex traits or phenotypes.

First, concatenation-based integration is analyzed, since it is the simplest framework for multi-omics data integration: it consists in the concatenation of normalized measurements into one joint matrix, followed by the development of a predictive model. Concatenation-based integration enables to identify multi-omics signatures by borrowing discriminatory strength from all information from datasets, but it dilutes the possibly low signal-to-noise ratio in each data type, affecting the understanding of the biological interactions at omics levels. Consequently, an alternative state-of-the-art integrative method is analyzed: Similarity Network Fusion (SNF, see Paragraph 4.1) [119]. Our top-down integration approach considers the development of predictive models on the integrated meta-omics: thus, a novel feature ranking scheme is developed as an extension of SNF (see Paragraph 4.2). The combination of the two integration approaches results into Integrative Network Fusion (INF, see Paragraph 4.3), a novel framework for the identification of robust meta-omics biomarkers.

4.1 Similarity Network Fusion

In a comparative review of scientific literature, SNF [119] emerged as one of the most reliable alternatives to concatenation-based integration. SNF is a non-Bayesian network-based method that can be divided into two main steps: the first step builds a sample-similarity network for each meta-omics data type (nodes as samples, edges as similarity measure), while the second step integrates these networks into a single similarity network, by using a nonlinear combination.

Consider L tables composed of different meta-omics (also defined as 'features') measured on the same n samples. Denote with M_l the lth $(n \times p_l)$ data matrix with $l = 1, \dots, L$.

Before SNF-integration, tables are normalized with respect to features. In particular, let $M_l^{h_l}$ the n-vector representing the h_lth column of lth data table. Then, for each $l = 1,\dots,L$, a new $(n \times p_l)$ matrix X_l is defined such that each column $X_l^{h_l}$, representing a feature, is obtained through:

$$X_l^{h_l} = \frac{M_l^{h_l} - \mathbb{E}(M_l^{h_l})}{\sqrt{Var(M_l^{h_l})}}, \tag{4.1}$$

where $\mathbb{E}(M_l^{h_l})$ and $Var(M_l^{h_l})$ represent the empirical mean and variance of $M_l^{h_l}$, respectively.

4.1.1 Similarity network inference

First, consider each meta-omics data table separately and indicate with $x_i^T = (x_{i,1},\dots,x_{i,p})^T$ for $i = 1,\dots,n$ the p-vector representing the ith row of the $(n \times p)$ data matrix X.
A patient similarity network is represented by a graph $G = (V,E)$ [120], where $V = \{x_1,\dots,x_n\}$ is the set of vertices corresponding to samples, while E is the set of edges modeling the similarity between patients. Thus, edge weights are represented by a $(n \times n)$ matrix W such that $W(i,j)$ indicates the similarity between samples x_i and x_j. In particular, similarity is defined as a scaled exponential distance kernel:

$$W(i,j) = exp\left(-\frac{d^2(x_i,x_j)}{\alpha\varepsilon_{i,j}}\right), \tag{4.2}$$

where $d(x_i,x_j) = \sqrt{\sum_{h=1}^{p}(x_{i,h} - x_{j,h})^2}$ is the Euclidean distance between samples x_i and x_j; α is a hyperparameter ranging in $[0.3, 0.8]$ as suggested in [119], which is tuned through a 5-fold cross validation repeated 10 times in our SNF implementation. Moreover, parameter $\varepsilon_{i,j}$ is set to eliminate the scaling problem and it is defined as follow:

$$\varepsilon_{i,j} = \frac{\frac{1}{K}\sum_{k\in N_i^K} d^2(x_i,x_k) + \frac{1}{K}\sum_{m\in N_j^K} d^2(x_j,x_m) + d^2(x_j,x_j)}{3}, \tag{4.3}$$

where N_i^K is the set of indexes of K nearest neighbors (KNN) samples of x_i with respect to the Euclidean distance d. Hence, N_i^K represents a set of K nearest neighbors for x_i, including x_i itself in G.

Once similarity networks are inferred, define a full and a sparse kernel, P and S respectively, on the vertex set V; P and S are necessary in the process that fuses together the similarity networks from multiple meta-omics. First, define a diagonal $(n \times n)$ matrix D such that $D(i,i) = \sum_{j=1}^{n} W(i,j)$. Then, the full kernel is a normalized weight matrix

$$P = D^{-1}W \tag{4.4}$$

so that $\sum_{j=1}^{n} P(i,j) = 1$.

On the other hand, S represents the local affinity of a graph G, measured by K nearest neighbors. Let N_i represent a set of x_i's neighbors including x_i in G; S is represented by a sparse $(n \times n)$ matrix such that:

$$S(i,j) = \begin{cases} \dfrac{W(i,j)}{\sum_{l\in N_i^K} W(i,l)}, & j \in N_i^K; \\ 0, & j \notin N_i^K. \end{cases} \tag{4.5}$$

K is a hyperparameter, set to 20 by Wang and colleagues [119]. The local affinity sets the similarities between non-neighboring points, in terms of the pairwise similarity values, to zero. Hence, it is assumed that local similarities with high values are more reliable than remote ones, as expected. Following the assumption adopted by other manifold learning algorithms, similarities to non-neighbors are assigned through graph diffusion on the network.

In summary, matrix P encodes the full information about the similarity of each sample to all others, whereas matrix S only accounts for the similarity to the K most similar samples for each sample.

The algorithm proposed in [119] for the fusion part considers P as initial status and uses S as the kernel matrix in the fusion process, both for the capacity of capturing local structure graphs and computational efficiency.

In our study a tuning procedure is implemented for parameters α and K from Equations (4.2) and (4.5), respectively (see Paragraph 4.2).

4.1.2 Similarity network fusion

Consider now the L meta-omics data tables; similarity matrices $W^{(l)}$, status matrices $P^{(l)}$ and local affinity matrices $S^{(l)}$ can be built for each $l = 1, \ldots, L$, through the equations (4.2), (4.4) and (4.5), respectively.

First, suppose $L = 2$. From two input similarity matrices, status matrices $P^{(1)}$ and $P^{(2)}$ are computed as in (4.4), and kernel matrices $S^{(1)}$ and $S^{(2)}$ as in (4.5).

Let $P_0^{(1)} = P^{(1)}$ and $P_0^{(2)} = P^{(2)}$ represent the initial two status matrices at $t = 0$. SNF iteratively updates status matrix corresponding to each of the data tables as:

$$P_{t+1}^{(1)} = S^{(1)} \cdot P_t^{(2)} \cdot \left(S^{(1)}\right)^T, \tag{4.6}$$

$$P_{t+1}^{(2)} = S^{(2)} \cdot P_t^{(1)} \cdot \left(S^{(2)}\right)^T, \tag{4.7}$$

where matrix $P_{t+1}^{(l)}$ is the status matrix of the lth data table ($l = 1, 2$) after t iteration. This procedure updates the status matrices each time generating two parallel interchanging diffusion processes. After T steps the overall status matrix is computed as

$$P^{(c)} = \frac{P_T^{(1)} + P_T^{(2)}}{2}. \tag{4.8}$$

Wang and colleagues [119] observed empirically that $T = 20$ leads to a SNF fast convergence. In detail, they kept track of the relative change in consecutive rounds for status matrices defined as

$$E_t^{(l)} = \frac{\|P_{t+1}^{(l)} - P_t^{(l)}\|}{\|P_t^{(l)}\|} \quad l = 1, \ldots, L$$

and set tol$= 10^{-6}$ as threshold: if the relative change was lower than the threshold for each $l = 1, \ldots, L$, they stopped the iteration. Hence, they noticed that $T = 20$ is enough to converge.

Besides, Wang and colleagues observed that SNF is robust to the noise in similarity measures due to KNN method used to compute S, which can reduce noise between instances. It can be observed from an equal formulation of (4.6):

$$P_{t+1}^{(1)}(i,j) = \sum_{k \in N_i} \sum_{m \in N_j} S^{(1)}(i,k) S^{(1)}(j,m) P_t^{(2)}(k,m), \tag{4.9}$$

similar for $P_t^{(2)}$. Note N_i^K is the neighborhood of x_i composed of K elements; hence, it is possible to observe from (4.9) that the similarity information is only propagated through the common neighborhood and this makes SNF robust to noise. Note that, if x_i and x_j have common neighbors in both similarity matrices, then they likely belong to the same cluster. Moreover, even if x_i and x_j are not very similar in one meta-omics table, their similarity can be expressed in another meta-omics type and this information can be propagated through fusion steps.

After each iteration, each $P_{t+1}^{(l)}$ ($l = 1, 2$) undergoes the following transformation:

$$P_{t+1}^{(l)} = P_{t+1}^{(l)} + I_n \tag{4.10}$$

This transformation ensures that, throughout SNF iterations, a sample is always more similar to himself than other patients; moreover, (4.10) makes the final network full rank, which is a crucial requirement for further analyses, such as clustering (see Paragraph 4.2). Wang and colleagues showed that this transformation leads to a quicker convergence of SNF.

The extension to the case $L > 2$ can be obtained substituting equation (4.6) and (4.7) with:

$$P^{(l)}_{t+1} = S^{(l)} \cdot \frac{\sum_{k \neq l} P^{(l)}_t}{L-1} \cdot \left(S^{(l)}\right)^T \qquad l = 1, \dots, L \tag{4.11}$$

and after T iterations the overall status matrix is computed as

$$P^{(c)} = \frac{1}{L} \sum_{l=1}^{L} P^{(l)}_T \tag{4.12}$$

In this thesis, the extension of Similarity Network Fusion starts from the overall status matrix $P^{(c)}$ (see Paragraph 4.2).

4.1.3 SNF algorithm

Similarity Network Fusion can be summarized into three steps: data preprocessing, networks inference and networks fusion. The detailed procedure is reported in Algorithm 1.

4.2 A network-based feature ranking scheme: rSNF

Similarity Network Fusion integrates multiple meta-omics datasets into a single comprehensive network in the space of samples rather than measurements (*e.g.,* bacterial DNA abundances). However, this thesis proposes meta-omics integration as an approach to identify robust biomarkers of samples phenotypes (*e.g.,* microbiota-related disease); consequently, it is necessary to extract measurements information from the SNF-fused network of samples. In our study, rSNF (ranked

Algorithm 1 SNF algorithm

1. *Normalization w.r.t features*
   For each $l = 1, \dots, L$ define a new $(n \times p_l)$ matrix X_l such that each column $X_l^{h_l}$, representing a feature, is obtained through Eq. (4.1).

2. *Similarity network inference*
   For each $l = 1, \dots, L$:
   Define $x_{i,l}$ the p_l-vector representing meta-omics variables for sample x_i in the lth normalized data table.

   (a) Infer the $(n \times n)$ similarity matrix W_l as defined in Eq. (4.2), supposed parameters α (scale coefficient for variance) and K (number of neighbors for each sample) fixed. α and K are tuned through a 10×5-fold cross-validation (see Algorithm 2)

   (b) Build the $(n \times n)$ full kernel matrix P_l using Eq. (4.4), which is normalized with respect to meta-omics variables.

   (c) Build the local affinity matrix S_l, which also depends on K as in Eq. (4.5).

3. *Similarity networks fusion*:
   For each $l = 1, \dots, L$:

   (a) define the initial status matrices $P_0^{(l)} = P_l$;

   (b) for $t = 1, \dots, 20$ update $P_{t+1}^{(l)}$ through Eq. (4.11), followed by the transformation $P_{t+1}^{(l)} = P_{t+1}^{(l)} + I_n$.

   Compute the overall status matrices $P^{(c)}$ through Eq. (4.12).

SNF) is thus designed: it is a feature ranking scheme, based on clustering performed on the fused similarity network.

4.2.1 Network Clustering

Suppose that samples modelled through the fused network can be grouped into C clusters, corresponding for example to known phenotypes. Each sample x_i, $i = 1,\dots,n$ can be associated to a label indicator vector $y_i \in \{0,1\}^C$, such that $y_i(k) = 1$ if sample x_i belongs to the kth cluster, otherwise $y_i(k) = 0$. Then, a $(n \times C)$ partition matrix $Y = (y_1^T;\dots;y_n^T)$ can be adopted to represent the clustering scheme. Let $P^{(c)}$ the fused graph matrix obtained by SNF as in Paragraph 4.1, then different types of clustering algorithm can be used to partition samples.

Spectral clustering is used in [119]. In general, spectral method aims to minimize the RatioCut, which is an objective function combining minimum cut of a graph and equipartitioning (see [121]), by solving the following optimization problem:

$$\begin{cases} \underset{Q \in \mathbb{R}^{n \times C}}{min} tr(Q^T L^+ Q), \\ Q^T Q = I_C, \end{cases} \tag{4.13}$$

where $Q = Y\left(Y^T Y\right)^{-\frac{1}{2}}$ is a scaled partition matrix and L^+ denotes the normalized Laplacian matrix

$$L^+ = I_n - D^{-\frac{1}{2}} P^{(c)} D^{-\frac{1}{2}}. \tag{4.14}$$

D is a network degree $(n \times n)$ matrix with degrees of each node from matrix $P^{(c)}$ on the diagonal ($D(i,i) = \sum_{j=1}^n P^{(c)}(i,j)$) and off-diagonal elements set to 0. Problem (4.13) can be solved using different algorithms, but in this thesis the choice is the one proposed in [122]. In particular, suppose samples can be partitioned in C clusters and define L^+ as in (4.14). Indicate with $\lambda_1,\dots,\lambda_C$ the C highest eigenvalues of L^+ and with $u_1,\dots,u_C$ the corresponding eiegenvectors and form the matrix

$U = [u_1|\dots|u_C]$. Rows of matrix U are then renormalized to have unit length yielding $(n \times C)$ matrix $\tilde{U}$ such that

$$\tilde{U}(i,j) = \frac{U(i,j)}{\sqrt{\sum_{j=1}^{C} U(i,j)^2}}. \tag{4.15}$$

Moreover, a rotation $(C \times C)$ matrix R is built, such that $Z = \tilde{U}R$ and for every row in Z there is at most one non-zero entry, *i.e.,* if $Z(i,j) = 1$, then sample i belongs to cluster j (see [122] for more details).

Estimated number of clusters

The number of samples clusters C can be known a priori or inferred from the similarity diffusion matrix $P^{(c)}$. In order to find the optimal number of cluster when it is unknown, [119] reports two main approaches.

1. The first method relies on eigengap, *i.e.,* the difference between subsequent ordered eigenvalues, to decide the best number of clusters based on the connectivity of the network. In particular, this is defined as follows:

$$eigengap : \mathbb{R}^n \to \mathbb{R} \quad \text{such that} \quad eigengap(i) = \lambda_{i+1} - \lambda_i \tag{4.16}$$

where λ_i is the i-th eigenvalue of the matrix L^+, defined in (4.14), sorted in ascending order ($\lambda_1 \leq \cdots \leq \lambda_n$). The best number of clusters C^* is

$$C^* = \max_{1<i\leq n} eigengap(i). \tag{4.17}$$

2. Another approach exploits the structure of eigenvector of L^+, as suggested in [122]. Assume $U = [u_1|\dots|u_C]$ is the orthogonal eigenvectors of L^+ corresponding to the eigenvalue $\lambda_1,\dots,\lambda_C$, renormalize it through (4.15) and build

$(n \times C)$ matrix $Z = UR$ describing the clustering. Denote with $M_i = \underset{1 \le j \le C}{max} Z_{ij}$ with $i = 1, \dots, n$; the optimal number of cluster C is the solution of the following problem

$$\underset{C \in \mathbb{N}}{min} \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{Z_{ij}^2}{M_i^2}. \tag{4.18}$$

A gradient descend method to solve this optimization problem is shown in [122].

Evaluation Metric

Several metrics can be adopted to evaluate the clustering performance, but Normalized Mutual Information (NMI) [123] emerged as the most reliable from a set of simulations performed by Wang and colleagues [119]. NMI has a crucial role in our feature ranking scheme (rSNF).

Let S be a set of n samples and define a clustering F on S as a way of partitioning S into non-overlap subsets $\{F_1, \dots, F_R\}$, where $\cup_{j=1}^{R} F_j = S$ and $F_j \cap F_i = \emptyset$ for $i \neq j$. The information on the overlap between two clustering $F = \{F_1, \dots, F_R\}$ and $G = \{G_1, \dots, G_C\}$ can be summarized in form of a $R \times C$ contingency table $N = (n_{ij})_{i=1,\dots,R; j=1,\dots,C}$ where n_{ij} denotes the number of objects that are common to clusters F_i and G_j. The outline of contingency table is illustrated in Table 4.1.

Table 4.1. Contingency Table N, $n_{ij} = |F_i \cap G_j|$

$F \setminus G$	G_1	G_2	$\cdots$	G_C	Sums
F_1	n_{11}	n_{12}	$\cdots$	n_{1C}	a_1
F_2	n_{21}	n_{22}	$\cdots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
F_R	n_{R1}	n_{R2}	$\dots$	n_{RC}	a_R
Sums	b_1	b_2	$\cdots$	b_C	$\sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij} = n$

Define $a_i = \sum_{j=1}^{C} n_{ij}$ for $i = 1, \dots, R$ the number of element in F_i and $b_j = \sum_{i=1}^{R} n_{ij}$ for $j = 1, \dots, C$ the number of element in G_j.

Given two clusterings F and G, their entropies and mutual information (MI) are defined naturally via the marginal and joint distributions of data items in F and G respectively as

$$H(F) = -\sum_{i=1}^{R} \frac{a_i}{n} log \frac{a_i}{N}, \tag{4.19}$$

$$H(G) = -\sum_{j=1}^{C} \frac{b_j}{n} log \frac{b_j}{n}, \tag{4.20}$$

$$I(F,G) = -\sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{n} log \frac{n_{ij}/n}{a_i b_j/n^2}. \tag{4.21}$$

The MI, a concept derived from information theory [124], measures the information that F and G share: basically, how much knowing one of these clusterings reduces the uncertainty about the other. The higher the MI, the more the information in F supports the prediction of cluster labels in G and viceversa.

Hence, NMI is defined as

$$NMI(F,G) = \frac{I(F,G)}{\sqrt{H(F)H(G)}}. \tag{4.22}$$

NMI ranges in $[0,1]$ and measures the concordance of two clustering results: the higher NMI the more similar the clusters.

Network clustering is a central step both for our novel feature ranking scheme (see Paragraph 4.2.2) as well as for tuning the parameters of Similarity Network inference (see Paragraph 4.1.1). A summary of parameters tuning procedure designed in this thesis is reported in Algorithm 2.

4.2.2 rSNF algorithm

rSNF is designed to rank key meta-omics variables for the identification of samples clustering (*e.g.,* phenotypes). Our feature ranking procedure is inspired by

Algorithm 2 Parameters Tuning Algorithm

:

1. *Normalization w.r.t features*
   Compute Step 1 of Algorithm 1
2. For each possible combination of $(K,\alpha) \in \{10,11,\dots,30\} \times \{0.3,0.35,\dots,0.8\}$
   (a) Compute Step 2 of Algorithm 1, which builds a similarity matrix $W_l^{(K,\alpha)}$, a full kernel matrix $P_l^{(K,\alpha)}$ and the local affinity $S_l^{(K,\alpha)}$ for each data table.
   (b) Compute Step 3 of Algorithm 1, which builds the overall status matrix $P^{(K,\alpha)}$.
   (c) Randomly group samples (columns or rows of $P^{(K,\alpha)}$) into 5 roughly equal-sized groups stratifying per classes and repeat this operation 10 times. Hence, for each $N = 1,\dots,10$, submatrices of $P^{(K,\alpha)}$ of dimension $(\lceil n/5 \rceil \times \lceil n/5 \rceil)$ are obtained, denoted as $P_{1,N}^{(K,\alpha)},\dots,P_{5,N}^{(K,\alpha)}$.
   (d) For each $N = 1,\dots,10$:
      i. for each $m = 1,\dots,5$, perform Step 4 of Algorithm 1 on $P_{m,N}^{(K,\alpha)}$ to obtain the sample clustering, denoted with $F_{m,N}^{(K,\alpha)}$, and evaluate it with respect to the known clustering for that subset of samples $(G_{m,N})$, through Step 5 of Algorithm 1. Then, it is computed $NMI_{m,N}^{(K,\alpha)} := NMI(F_{m,N}^{(K,\alpha)}, G_{m,N})$;
      ii. compute the median of $NMI_{m,N}^{(K,\alpha)}$ with respect to $m = 1,\dots,5$ which is indicated with $NMI_N^{(K,\alpha)}$.
   (e) Compute the median of $NMI_N^{(K,\alpha)}$ with respect to $N = 1,\dots,10$ and denote it with $NMI^{(K,\alpha)}$.
3. the optimal couple values of hyperparameters (K,α) is the one solving

$$\underset{(K,\alpha)}{max} NMI^{(K,\alpha)}$$

Wang and colleagues [119] and it is based on network clustering as described in Paragraph 4.2.1. Suppose a set of normalized data tables X_l of dimension $(n \times p_l)$, $l = 1, \dots, L$ are integrated by SNF (see Paragraph 4.1.3), into a fused network expressed through the matrix P. rSNF steps are summarized in Algorithm 3.

Algorithm 3 rSNF Algorithm

1. *Fused network clustering*:
   perform spectral clustering on fused network P, as described in Paragraph 4.2.1. In this thesis, a *non a priori* approach is adopted as suggested in [119], thus the number of clusters is not specified. The clustering of samples (*e.g.,* representing clinical phenotypes) is denoted with F.
2. *Feature ranking scheme*:
   For each feature $f_{h_l,l}$ with $l = 1, \dots, L$ and $h_l = 1, \dots, p_l$:
   (a) Build a sample network $P_{h_l,l}$ based on $f_{h_l,l}$ alone
   (b) Perform spectral clustering on $P_{h_l,l}$ to identify subtypes $F_{h_l,l}$
   (c) Measure the consistency between $f_{h_l,l}$ and the whole network P as $cs_{h_l,l} = NMI(F_{h_l,l}, F)$. Hence, if $cs_{h_l,l} = 1$, the network of samples based on $f_{h_l,l}$ leads to the same clusters as the fused network, therefore feature $f_{h_l,l}$ is determinant in the construction of sample network. On the other hand, if $cs_{h_l,l} = 0$, there is no real correspondence between the feature and the fused network. In conclusion, the higher is $cs_{h_l,l} = 0$, the more important is the $f_{h_l,l}$ to the fused network structure, allowing to rank all the features with respect to their importance for the fused network construction.

Clearly, rSNF is based on Similarity Network Fusion and network clustering (and consequently on NMI score). The motivation of the specific rSNF design is that it naturally extends a popular state-of-the-art method as SNF, which has inspired several studies in the recent scientific literature, specifically in cancer genomics [125, 126], in metagenomics [127, 128], as well as in precision medicine [129, 130]. rSNF exploits two main SNF advantages: integration of heterogeneous data and sample networks clustering. The main peculiarity of SNF integrative procedure is its robustness to noise [119], because weak similarities among samples (low-weight edges) disappear, except for low-weight edges supported by all networks that are conserved depending on how tightly connected their neighborhoods are across networks. Moreover, Wang and colleagues showed the advantage of combining SNF and network clustering: they performed spectral clustering on SNF-integrated data (DNA methylation, mRNA and miRNA expression) by identifying subtypes across a wide spectrum of cancers; best performance was found by evaluating the silhouette score [131], which is a measure of cluster coherence.

4.3 Integrative Network Fusion

This thesis introduces Integrative Network Fusion (INF), a bioinformatics framework for the identification of integrated meta-omics biomarkers. The framework is based on the predictive profiling of meta-omics data abundances (*i.e.,* bacterial and fungal DNA abundances) with a novel approach to their integration. In summary, INF implements and compares a standard naive and a novel integration approach: first, the standard method is considered by concatenating meta-omics data and training Random Forest (RF) or Support Vector Machine (SVM) classifiers on the combined dataset, finally obtaining a ranked list of biomarkers. This approach is referred as to ml-J. Secondly, meta-omics data are integrated by Similarity Network Fusion (see Paragraph 4.1); again, RF or SVM models are developed on the integrated dataset for the SNF-ranked list of meta-omics variables (see Paragraph 4.2). This approach is referred as to ml-rSNF. Finally, RF or SVM are trained on the dataset restricted on the intersection of the biomarkers lists from ml-J and ml-rSNF. Notably, predictive models are developed inside the Data Analysis Protocol described in Paragraph 2.3, ensuring reproducibility and avoiding overfitting or selection bias.
INF is structured as in Figure 4.1.

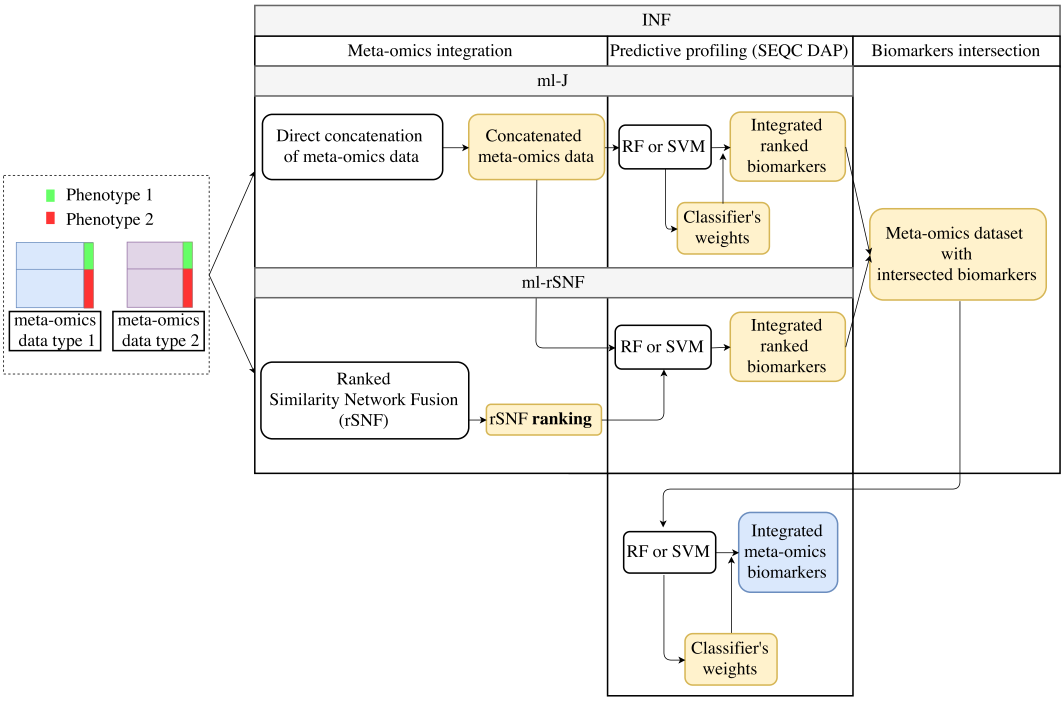

INF
Meta-omics integration
Predictive profiling (SEQC DAP)
Biomarkers intersection
ml-J
Direct concatenation of meta-omics data
Concatenated meta-omics data
RF or SVM
Integrated ranked biomarkers
Classifier's weights
🟩 Phenotype 1
🟥 Phenotype 2
meta-omics data type 1
meta-omics data type 2
Meta-omics dataset with intersected biomarkers
ml-rSNF
RF or SVM
Integrated ranked biomarkers
Ranked Similarity Network Fusion (rSNF)
rSNF **ranking**
RF or SVM
Integrated meta-omics biomarkers
Classifier's weights

Figure 4.1. INF workflow. The ml-J (classifier on juxtaposed datasets) and ml-rSNF (classifier on combined datasets with rSNF-ranked variables) are run in parallel. Integrated meta-omics signature is computed by training a classifier on the datasets restricted on the intersection of ml-J and ml-rSNF biomarkers.

The implementation requires only standard open source computational biology tools, as a combination of R/Bioconductor and Python functions: SNF and rSNF are implemented by in-house R scripts, extending R functions provided by Wang and colleagues [119]; predictive profiling combines in-house Python (for classifiers development) and R scripts (for graphical output). The code implementing INF and a clinical dataset for INF validation are available in the GitHub repository (`https://github.com/AleZandona/INF`).

Chapter 5

Bioinformatics workflows

Sequencing technologies have been rapidly improving, as well as the computational infrastructure needed to analyze the resulting volume of the data being generated. Indeed, computational tools have a central role in several meta-omics analyses, *i.e.,* the identification of associations between the microbiome and specific diseases, the deconstruction of the host-microbe-microbiome interactions as well as the integrative analysis of multiple meta-omics data. Diverse software applications have been designed to address these challenges, but performing these analyses separately usually requires the installation, integration, and tuning of multiple software packages, which is not always trivial even for groups with extensive bioinformatics expertise. Consequently, most studies rely on modular frameworks that collect computational tools in a modular and customizable manner, making it easier to reproduce or extend analysis results and encouraging collaboration. One of the most popular workflows for metagenomic data analysis is QIIME [25], which integrates several tools in a single framework: from the taxonomic classification of the DNA sequences (abundance of microbial taxa within microbiota) to the analysis of functional capacity of microbiome, including also the possibility to display the results in a graphical form. In particular, one of the software tools included in QIIME is mothur [132], which is a popular pipeline for amplicon metagenomic data integrating tools for sequence screening based on quality, Operational Taxonomic Units (OTUs) definition and estimation of ecological parameters (*i.e.,* α and β diversity). The UPARSE pipeline [26] is an alternative to QIIME and mothur: first, metagenomics sequences are cleaned, quality filtered and dereplicated, then reads are ordered according to their abundances considering that high abundance sequences are more likely to be correct before OTUs assignment. MG-RAST [27] is a webserver performing taxonomic analysis, functional profiling and gene calling using stat-of-the-art tools such as UCLUST [133] and custom protein databases (M5nr [134]) as well as M5rna (a combination of SILVA [135], GreenGenes [136] and RDP [137]) for rRNA analysis. The recently introduced MICCA pipeline [28] implements the processing of targeted metagenomic datasets combining quality filtering, chimera identification, taxonomic assignment, diversity analysis and phylogenetic tree inference.

This thesis introduces two modular and flexible bioinformatics frameworks, built around a collection of publicly available and in-house metagenomic analysis tools that can be tailored and extended to meet specific analysis needs.

5.1 PreMONet framework

Predictive Meta-Omics Networks (PreMONet) is a computational framework including a chain of tools for a complete quantitative analysis of meta-omics data. PreMONet is outlined in Fig. 5.1, both as conceptual workflow for metagenomics data and as its implementation for 16S and WGS data analysis. The application of PreMONet on other meta-omics data is out of the scope of this thesis, but can be achieved by the adaptation of modules A-C (see Fig. 5.1).

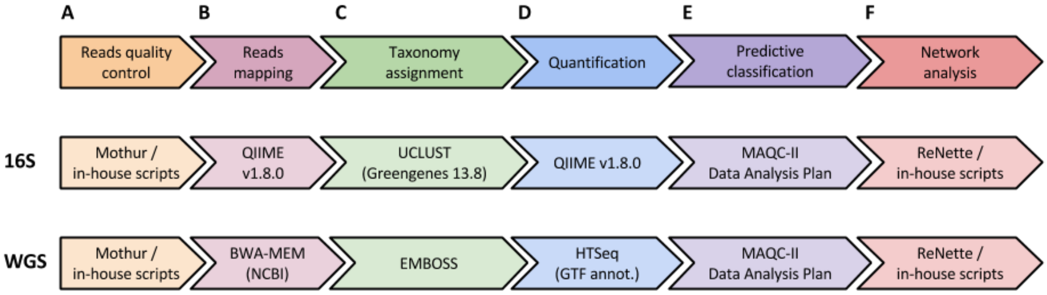


Figure 5.1. PreMONet framework for metagenomics data. Conceptual workflow and implementation for both 16S and WGS data.

Overall, the whole procedure can be split in three main modules, namely data preprocessing, machine learning profiling (see Chapter 2) and differential network analysis (see Chapter 3). The preprocessing step (module A, Fig. 5.1) implements quality filtering of metagenomics sequences (either in SFF or FASTQ/FASTA format) by Mothur [132]: removal of short sequences, reads with ambiguous bases, sequences with either low entropy or low Phred quality score. Reads are then aligned against a reference database (module B), which depends on the type of metagenomics data to analyze: it is a collection of NCBI annotated whole genome sequences in case of WGS data or Greengenes 16S rRNA gene database in case of 16S data. WGS reads are aligned by BWA-MEM [138], while amplicon reads undergo a different process implemented by QIIME. Sequences with a similarity level of 97% are clustered into Operational Taxonomic Units (OTUs) by PyNAST [139]. OTUs are commonly intended to represent some degree of taxonomic relatedness, depending on the sequence similarity; resulting clusters from a 97% threshold are typically considered of as representing a species. Each OTU may group many related sequences, thus a representative sequence from each OTU is picked (the OTU centroid sequence) by UCLUST [133] and aligned by PyNAST against Greengenes $v.\ 13.8$ database. After reads mapping, each organism matched in the metagenomic reference is taxonomically assigned (module C) at different taxonomic levels (Phylum, Genus, Species) by either EMBOSS (`http://emboss.sourceforge.net`), using the corresponding NCBI taxonomy ID, or UCLUST, if data are WGS or 16S, respectively. Taxonomy assigment is followed by quantification module (module D), which computes microbial reads abundance. Whilst 16S data are quantified by QIIME, WGS reads abundance is inferred by the Python module HTSeq, counting reads overlapping more than one specified genomic region (either the whole genomic sequences or coding sequences as well as structural RNA).

In summary, PreMONet preprocessing modules provide an abundance profile of the sequenced metagenomic reads, resulting in a sample-by-abundances count matrix, which is derived at each taxonomic level of interest. Count data can also be normalized in terms of the Trimmed Mean of M-values (TMM) normalization, implemented by the edgeR Bioconductor package, while compositional data are not normalized. Moreover, abundances matrix can be further processed, so that low abundant reads (present in less than a user-defined fraction of samples) are optionally filtered out before generating predictive models.

Machine learning profiling (module E) and network analysis (module F) are described in detail in Chapters 2 and 3, respectively.

5.2 I-PreMONet framework

PreMONet is equipped with additional functionality to implement a comprehensive integrative modeling of multiple meta-omics layers. Specifically, the integrative framework detailed in Paragraph 4.3 is embedded into PreMONet, right after the preprocessing step and before the machine learning profiling module. This extended version of PreMONet is referred as to Integrated Predictive Meta-Omics Networks, briefly as to I-PreMONet. The framework is structured as in Fig. 5.2.

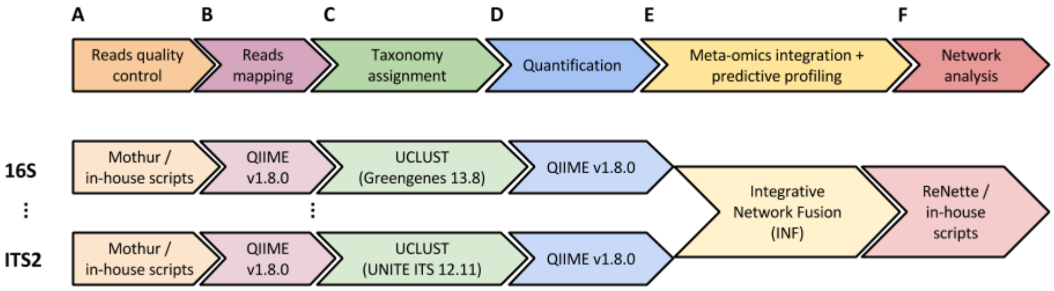


Figure 5.2. I-PreMONet framework for metagenomics data. Conceptual workflow and implementation for 16S and ITS2 data.

Notably, the implementation of both our frameworks requires only standard open source computational biology tools, as a combination of in-house scripts based on a suite of R/Bioconductor and Python functions. The code implementing (I-)PreMONet is thus accessible and customizable, making our frameworks easy-reusable also for researchers without bioinformatics expertise. Moreover, the modular design and open-source licensing model allow the extension of (I-)PreMONet to new applications beyond our initial focus on meta-omics data. Indeed, parameters (input, output, parameters either for classifiers training or networks inference) may be changed via the Bash command line; besides, users may directly edit the code of each (I-)PreMONet module to implement extensive changes to the pipeline, if required.

As validation, (I-)PreMONet have been tested on clinical metagenomics datasets; results are reported in detail in Chapter 6.

Chapter 6

Biological applications

In this Chapter we validate PreMONet (see Paragraph 5.1) and its extended version I-PreMONet (see Paragraph 5.2) on clinical datasets. The final objective is the prediction of disease phenotypes by the characterization and modeling of human microbiota, considering either metagenomic or multiple meta-omics data. In Sec. 6.1, we present an example of PreMONet application to a clinical context: the analysis of structural change of networks from bacterial communities to predict Inflammatory Bowel Disease (IBD) outcome and evolution in children. Sections 6.2 and 6.3 report the combined analysis of bacterial and fungal microbiota performed by INF (the core module of I-PreMONet), so to predict IBD outcome prediction in adulthood (see Sec. 6.2) as well as to relate gut microbiota composition to Rett syndrome (see Sec. 6.3), respectively. Our workflow aims at setting a new resource for predictive analysis that progressively exploits complex network methods, inspired to the general network medicine framework [140].

6.1 Pediatric Inflammatory Bowel Disease: P_IBD

Inflammatory Bowel Disease (IBD) is a broad term describing a set of complex chronic intestinal inflammatory disorders, Ulcerative colitis (UC) and Crohn's disease (CD) as the best known syndromes [141, 142]. Chronic diseases of the intestine, including IBDs, are a leading cause of morbidity and mortality in the developed world, *i.e.,* affecting approximately 1.4 million Americans, with a peak onset in people 15 to 30 years of age. CD is characterized by patchy and transmural inflammation that may affect any part of the gastrointestinal tract, including wall thickening, duct stricture, fistulas (abnormal passages between two organs, or between an organ and the outside of your body), abscesses and ulcers. UC is a chronic periodic inflammatory condition that involves only the large bowel at the mucosa level, leading to the loss of haustra (colon small pouches), rectal bleeding, pseudo-polyps formation and damage of mucosal lining. Although CD and UC are different disorders, both may manifest any of the following symptoms: abdominal pain, vomiting, diarrhea, rectal bleeding, severe internal cramps/muscle spasms in the region of the pelvis and weight loss. CD and UC may present extra-intestinal manifestations (*i.e.,* liver problems, arthritis, skin manifestations and eye problems, anemia, pyoderma gangrenosum, primary sclerosing cholangitis, and non-thyroidal illness syndrome) in different proportions [143, 144]. Diagnosis is generally achieved assessing blood as well as fecal inflammatory markers, followed by colonoscopy with biopsy of pathological lesions. The course of the disease is unpredictable, being characterized by periods of remission and recurrent active inflammation; moreover, IBD etiology has not been completely unraveled yet.

Accumulating evidence suggests that IBD involves dysregulation in the normally symbiotic relationship between mucosal immune system and intestinal commensal microbes, modulated by the genetic susceptibility of the host [145, 146, 147]. Studies profiling the gut microbiota in patients with IBD compared to controls have consistently shown changes in microbiota composition as well as reduction in overall biodiversity. The largest study to date in a treatment-naïve cohort of paediatric patients with CD [148], in whom analysis of mucosal and lumen-associated microbiota was performed, confirms that inflammation is strongly associated with an overall drop in species diversity and alterations in the abundance of several taxa. Moreover, the relative balance of beneficial vs. aggressive commensal enteric microflora has been advocated to determine mucosal homeostasis vs. inflammation [149]. This inflammation can determine tissue damage, cell proliferation and infiltration, potentially changing the metabolism between normal and diseased tissues. Besides, several pieces of evidence suggest that luminal commensal bacteria provide an antigenic stimulus, inducing immune response (dysregulation) and triggering the inflammation associated to IBD onset, in genetically susceptible individuals. IBD has a well-established genetic component and genome-wide association studies (GWAS) have been highly successful in identifying genes that contribute to the disease risk, underlining that characteristics of host defenses and their interplay with the enteric content are crucial in initiating the pathogenesis of IBD. GWAS have also identified a number of immune system pathways that are mutated in susceptible hosts, some of which are activated by infection-mediated syndromes [150] or by altered cellular responses [151]. A recent theory has shed light onto the gut ecology, which exerts concerted actions and synergic commensal responses to pathogens [152]. The gut microbiota is clearly the proximate environmental influence on the risk of IBD, even though it is unclear whether tissue damage results from an abnormal immune response to a normal microbiota or from a normal immune response against abnormal microbiota.

PreMONet contributes to the identification of new microbial biomarkers of dysbiosis predicting IBD outcome and, possibly, allowing targeted prevention.

6.1.1 Dataset

PreMONet was validated on the P_IBD dataset, generated by the Bambino Gesù Children's Hospital (Rome, Italy). P_IBD consists in bacterial composition of gut microbiota of 45 children with IBD and 47 healthy children (Tab. 6.1), determined using 16rRNA from Roche 454 platform. Microbiota was collected both from fecal samples and colon biopsies; in detail:

- 57 fecal samples: 27 healthy and 30 IBD children
- 15 biopsies from colon: matched normal/inflamed tissue
- 20 biopsies from healthy donors

Phenotype	Age: Year (mean ± SD)
Fecal healthy	10.6 ± 1.8
Biopsy healthy	12.0 ± 5.4
Biopsy IBD	13.2 ± 4.4
Fecal IBD	12.9 ± 4.5

Table 6.1. Age of subjects in P_IBD dataset.

Fecal samples were collected from either IBD and healthy subjects at visit (Pediatric Gastroenterology and Liver Unit, Sapienza University of Rome, Italy). Moreover, from one to two mucosal biopsies (1×2 mm/each), taken from the distal colon just above the rectosigmoid junction, were collected from IBD (two macroscopically inflamed and not inflamed tissue regions) and from healthy subjects (one not inflamed tissue). All biopsies were managed in the operating room according to standardized protocols for the preservation of nucleic acids.

Genomic DNA extraction from fecal samples. Stools were resuspended into $1.5\ ml$ PBS, homogenized by vortexing for 2 min and centrifuged at $20,800 \times g$. After supernatant removal, pellet was resuspended into $500\ \mu l$ of PBS added by $500\ \mu l$ of Beads/PBS $(1\ mg/\mu l, w/v)$ (Glass Beads, acid-washed SigmaAldrich). The $1:1$ mixture was homogenized by vortexing 2 min and centrifuged at $5200 \times g$ for 1 min. The supernatant was collected, and treated for one freeze-thaw cycle (-20℃/70℃) for 20 min each step. After centrifugation at $5200 \times g$ for 5 min, the supernatant was subjected to QIAamp DNA Stool Mini Kit (Qiagen, Germany) extraction, according to manufacturer's instructions. DNA was eluted into $50\ \mu l$ purified H_2O (Genedia, Italy) and its yield quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). DNA was adjusted to $10\ ng/\mu l$ concentration and used as template for successful 16S Metagenomic 454 Sequencing Analyses.

Genomic DNA extraction from biopsy samples. Biopsies were incubated for 1 h in $190\ \mu l$ Buffer G2 (Qiagen) (ensuring that the tissue sections are fully submerged in Buffer G2), incubated for 5 min at 75℃, with vigorous mixing. Sample were cooled to 56℃ and $10\ \mu l$ proteinase K solution $(600\ mAU/ml)$ were added, mixed and incubated for 1 h at 56℃ with continuous vigorous mixing. DNA extraction was performed by EZ1 automated procedures according manufacturer's procedures. DNA was eluted in $100\ \mu l$.

Amplicon library preparation and pyrosequencing. The gut microbiome was investigated by barcoded pyrosequencing V1-V3 regions of the 16S rRNA gene (amplicon size 520 bp), on a GS Junior platform (Roche 454 Life Sciences, Branford, USA), according to the pipeline described in [153].

Taxonomic analysis

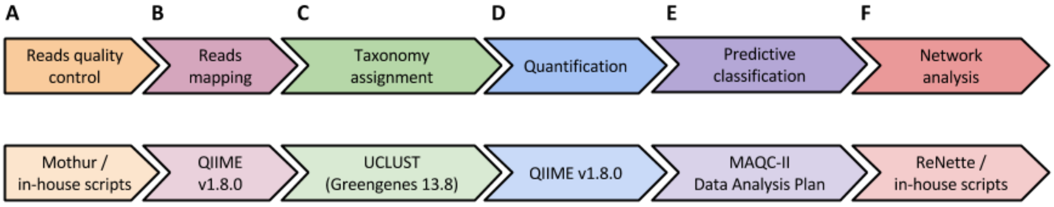


Figure 6.1. PreMONet applied on P_IBD dataset.

Pyrosequencing reads provided in SFF files were first processed via Mothur $v.\ 1.33.3$ [132] by module A of PreMONet (Fig. 6.1), filtering out reads with:

- Length less than 200 bp
- Homopolymers longer than 8 bp
- Ambiguous bases
- Average Phred quality score < 35, over windows of 50 bp each

After quality control process, the remaining reads were analyzed in the Quantitative Insights into Microbial Ecology (QIIME) $v.\ 1.8.0$ [25]. First, genomic reads from all samples with a sequence similarity level of 97% were clustered into Operational Taxonomic Units (OTUs) by PyNAST [139]. OTUs are commonly intended to represent some degree of taxonomic relatedness, depending on the sequence similarity; resulting clusters from a 97% threshold are typically considered of as representing a species. Each OTU may group many related sequences, thus a representative sequence from each OTU was picked (the OTU centroid sequence) by UCLUST [133] and aligned by PyNAST against Greengenes $v.\ 13.8$ database (module B in Fig. 6.1). Taxonomy was assigned to each representative sequence with the UCLUST consensus taxonomy assigner (module C in Fig. 6.1); a six-level taxonomy (from kingdom to species) was provided and both the unassigned OTUs and the unspecified levels were considered. Lastly, module D produced a table of OTU abundances in each sample with taxonomic identifiers for each OTU; the table was filtered by discarding unassigned OTUs and keeping taxa at genus level only.

A downstream filter was applied to the OTU table, keeping only that OTUs present in at least 20% of the samples; the resulting taxonomic units tables defined the classification problems associated to the dataset.

Statistical analysis

In order to discriminate dysbiotic microbial profiles associated with IBD from the normal ones, microbiota both from different environments (fecal samples vs. colon biopsies) as well as clinical phenotypes (healthy vs. IBD) was analyzed. In detail, five classification tasks were performed:

- 30 IBD vs. 27 healthy subjects, fecal samples (FEC_H_IBD)
- 27 fecal samples from healthy individuals vs. 15 not inflamed tissue biopsies from IBD patients ($FEC\_H\_B\_NORM$)
- 30 fecal samples from IBD patients vs. 15 inflamed tissue biopsies (FEC_B_IBD)
- 15 not inflamed vs. 15 inflamed tissue biopsies (B_NORM_IBD)
- 20 healthy vs. 15 inflamed IBD subjects, tissue biopsies (B_H_IBD)

Microbial biomarkers for each classification task were identified by machine learning methods, implemented in module E of the PreMONet pipeline (Fig. 6.1): Support Vector Machines and Random Forest classifiers were trained adhering to the Data Analysis Protocol presented in Sec. 2.3, so to ensure results reproducibility.

6.1.2 Network analysis

The PreMONet pipeline provided also the setup for quantitative analysis of microbial communities in terms of quantification of networks differences and evolution of microbial communities versus the dynamics of the target phenotypes (module F in

Fig. 6.1).

Co-occurence networks

Starting from predictive biomarkers, co-abundance undirected weighted networks were built using top-features as nodes from cohorts corresponding to patients phenotypes in terms of the (thresholded) absolute Pearson Correlation Coefficient (PCC). This approach aimed to highlight changes in links between OTUs, depending on either the health status (IBDs vs. controls) or the environment (fecal content vs. colon biopsies). Finally, the structures of the obtained microbiome networks were compared by quantifying network distances using the glocal HIM distance [98, 154]. The closer to zero the HIM distance, the more similar the compared networks. Graphical layout of networks was produced with CIRCOS [155], with genera being represented as arches of outer ring and PCC between genera represented as links connecting arches.

Network trajectories

Samples, regardless of the phenotype, were grouped by increasing level of calprotectin (range 10-370 mg/kg), which is commonly used in clinical settings as a non-invasive marker to assess the activity of IBD. The number of samples across sets were balanced and the co-occurrence networks were inferred on top-ranked genera for each sample subset. In order to measure changes in microbial communities structure in association to calprotectin levels, HIM distance was computed between the network corresponding to the lowest calprotectin range and the networks corresponding to increasing calprotectin levels.

Network communities

For each network, a community detection analysis was performed by the Louvain method [108], a quantitative technique for grouping nodes according to the network modularity, *i.e.*, the density of links inside communities as compared to links between communities.

6.1.3 Results

Analyses were performed at genus taxonomic level, thus the 3,510 OTUs table built by QIIME in the preprocessing steps was reduced to a 168 genera table; subsequently, one table for each classification task was built and low abundant genera were filtered out (Tab. 6.2).

Task	# samples	# genera
FEC_H_IBD	57	40
FEC_H_B_NORM	42	35
B_NORM_IBD	30	33
FEC_B_IBD	45	36
B_H_IBD	35	37

Table 6.2. P_IBD datasets dimension: number of samples and genera for each classification task.

For each classification task, main results are reported in Fig. 6.2 in terms of average MCC with 97.5% Student bootstrap ($1000\times$ resampling) confidence intervals (MCC_{min}, MCC_{max}), number of top-ranked features (Nf), and Canberra stability indicator (S). Top classification performance was achieved for FEC_H_B_NORM with $MCC = 0.81$ and $Nf = 30$ genera, and for FEC_B_IBD with $MCC = 0.74$ and $Nf = 36$ genera. IBD status in fecal samples (FEC_H_IBD) was predicted with $MCC = 0.61$ and 4 genera. A good classification performance was achieved also in biopsies from healthy donors and IBD patients (B_H_IBD) with $MCC = 0.61$ and $Nf = 9$. Remarkably, IBD status could not be predicted in matched biopsies (B_NORM_IBD), as the best model yielded $MCC = 0.01$ with 3 genera.

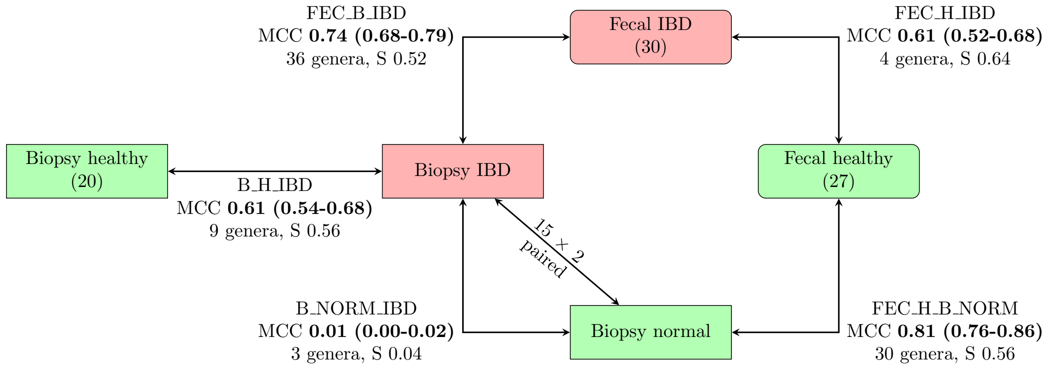
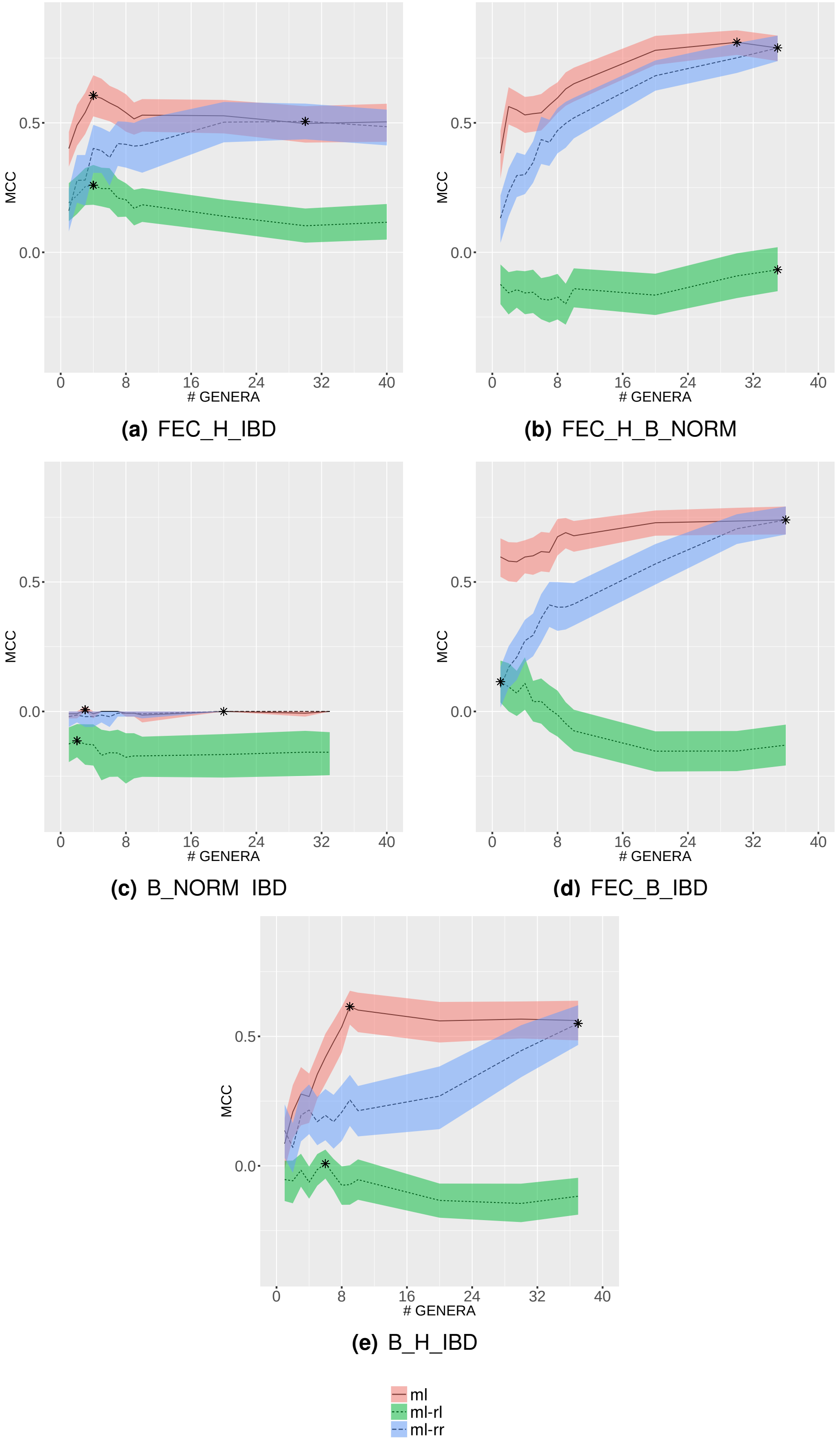


Figure 6.2. Best predictive performance for the five classification tasks.

Mean MCC values achieved by classifiers at increasing feature set sizes are shown in Fig. 6.3. As described in Sec. 2.3, first the classifier identified a list of genera ranked by their importance in the specific classification problem, then increasing sets of genera (referred as to “feature steps”) from the ranked list were used to build predictive model on training partition.

Classifiers did not overfit data: MCC curve resulting from random ranking experiment (ml-rr) increased with the feature steps, reaching the maximum by using the entire set of genera. Indeed along random ranking procedure, predictive power of genera for the specific classification problem was not considered (ranking was shuffled); thus, the classifier performance was expected to increase with genera set size, since the number of discriminant genera included in classifier training also increased.

Selection bias was also avoided, with evidence from ml-rl curve, which oscillated around zero. Thus, the prediction of models developed along random labels procedure was actually random, as the association between samples and phenotypes. As a first naïve evaluation of microbial communities structure, Pearson Correlation coefficient (PCC) was computed among the top-ranked genera of each classifica-



(a) FEC_H_IBD

(b) FEC_H_B_NORM

(c) B_NORM IBD

(d) FEC_B_IBD

(e) B_H_IBD

Figure 6.3. MCC plots for the 5 classification tasks. Solid curves in black indicate MCC, with corresponding 95% bootstrap confidence intervals, at increasing genera set sizes (STEP). MCC_{int} highlighted with ($*$). Random labels (RL) and random ranking (RR) experiments are indicated respectively in green and in blue (see text).

tion task; microbial interactions was modeled by networks, with nodes as genera and edges as correlation among them. CIRCOS plots in Fig. 6.4-6.6 were chosen as graphical layout for networks, arches on ring as genera and ribbons as correlation links over threshold, which was computed by the model proposed in [94].

In details, Fig. 6.4 reports the co-abundance networks on top-ranked genera in B_H_IBD task (task E). In biopsy from healthy subjects (green edges), *Oscillospira* was linked with *Ruminococcus*, while in biopsy from inflamed tissue (red edges) *Ruminococcus* was linked with *Dialister* and *Odoribacter*, which, in turn, was linked to *Coprococcus*.

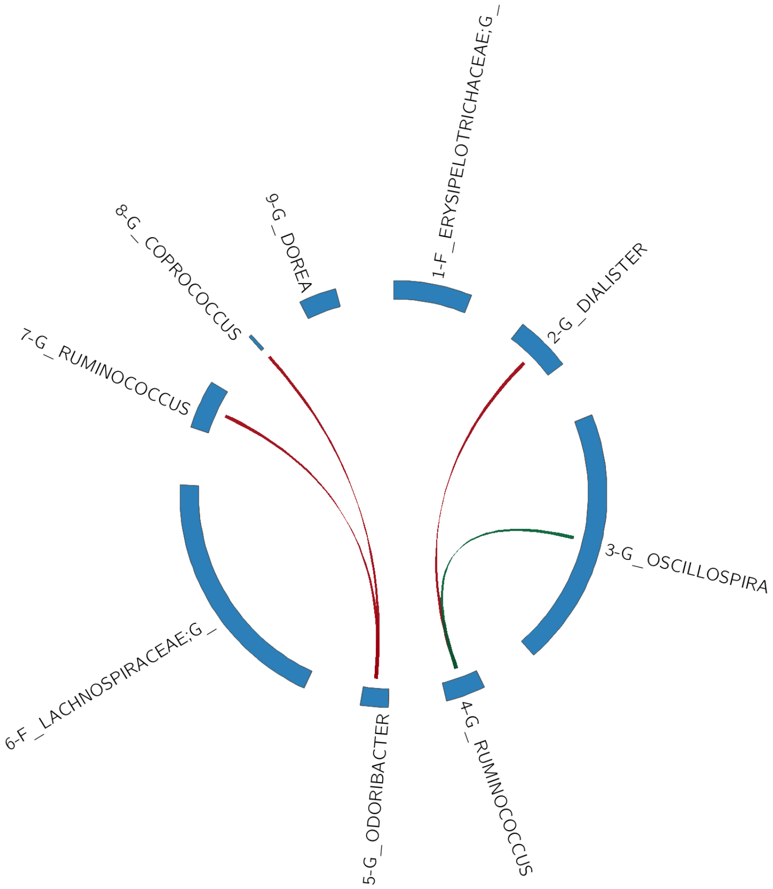


Figure 6.4. CIRCOS co-abundance networks on top-ranked genera (B_H_IBD). Red edges: links conserved in IBD only; green edges: links conserved in H only. Edge's color intensity is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.67$.

Analyzing the co-abundance network for task FEC_B_IBD (task C), reported in Fig. 6.5, the following links amongst OTUs were found only in feces from IBD subjects: *Streptococcus* to *Haemophilus*; *Prevotella, Oscillospira, Ruminococcus* and *Phascolarctobacterium* to *Clostridiales*; *Bacteroidales* to *Phascolarctobacterium*; *Sutterella* to *Akkermansia*; *Lachnospiraceae* to *Clostridium*; *Ruminococcus* to *Oscillospira*. In Biopsy inflamed samples, *Streptococcus* was linked to *Ruminococcus*; *Parabacteroides* to *Veillonella*; *Barnesiellaceae* to *Clostridiales* and *Blautia*; *Lachnospira* to *Lachnospiraceae*; *Prevotella, Fusobacterium* and *Bacteroidales* to *Akkermansia*; *Bacteroidales* to *Fusobacterium*; *Enterococcus* to *Veillonella*; *Blautia* to *Clostridiales*; *Clostridiaceae* to *Haemophilus*.

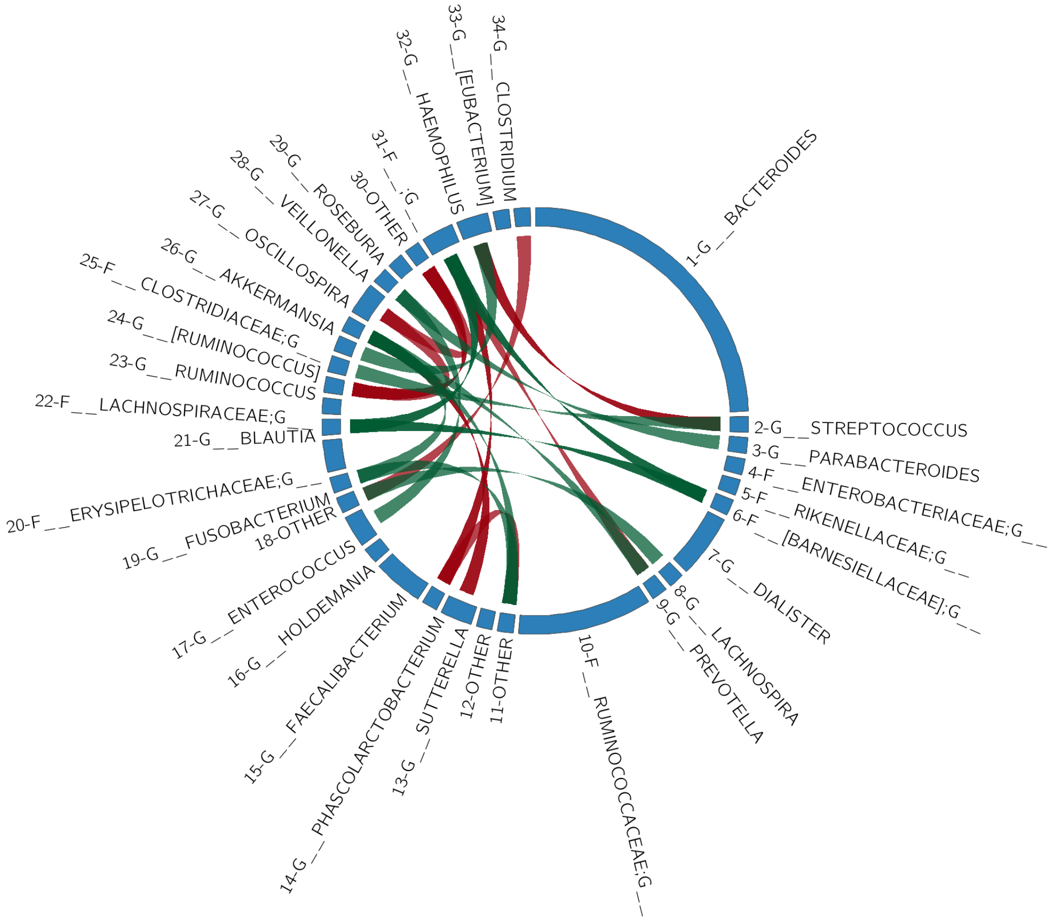


Figure 6.5. CIRCOS co-abundance networks on top-ranked features (FEC_B_IBD). Red edges: links conserved in FEC only; green edges: links conserved in B only. Edge color intensity thickness is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.8$.

The results from the co-abundance networks in FEC_H_B_NORM task (task B) are reported in Fig. 6.6. In fecal healthy samples, only one edge was found between *Anaerostipes* and *Erysipelotrichaceae*. In biopsy from inflamed samples, 22 links were found. In particular, *Bacteroides* was linked to *Lachnospiraceae* and *Lachnospira*; *Anaerostipes* to *Blautia* and *Roseburia*; *Clostridium* and *Ruminococcus* to *Lachnospiraceae*; *Lachnospira* to *Dialister, Lachnospiraceae, Ruminococcus*; *Odoribacter* to *Clostridiales*; *Rikenellaceae* to *Oscillospira*; *Erysipelotrichaceae* to *Veillonella, Parabacteroides, Ruminococcaceae*; *Ruminococcaceae* to *Veillonella* and *Parabacteroides*; *Prevotella* to *Akkermansia* and *Ruminococcaceae*; finally, *Ruminococcaceae* to *Akkermansia*.

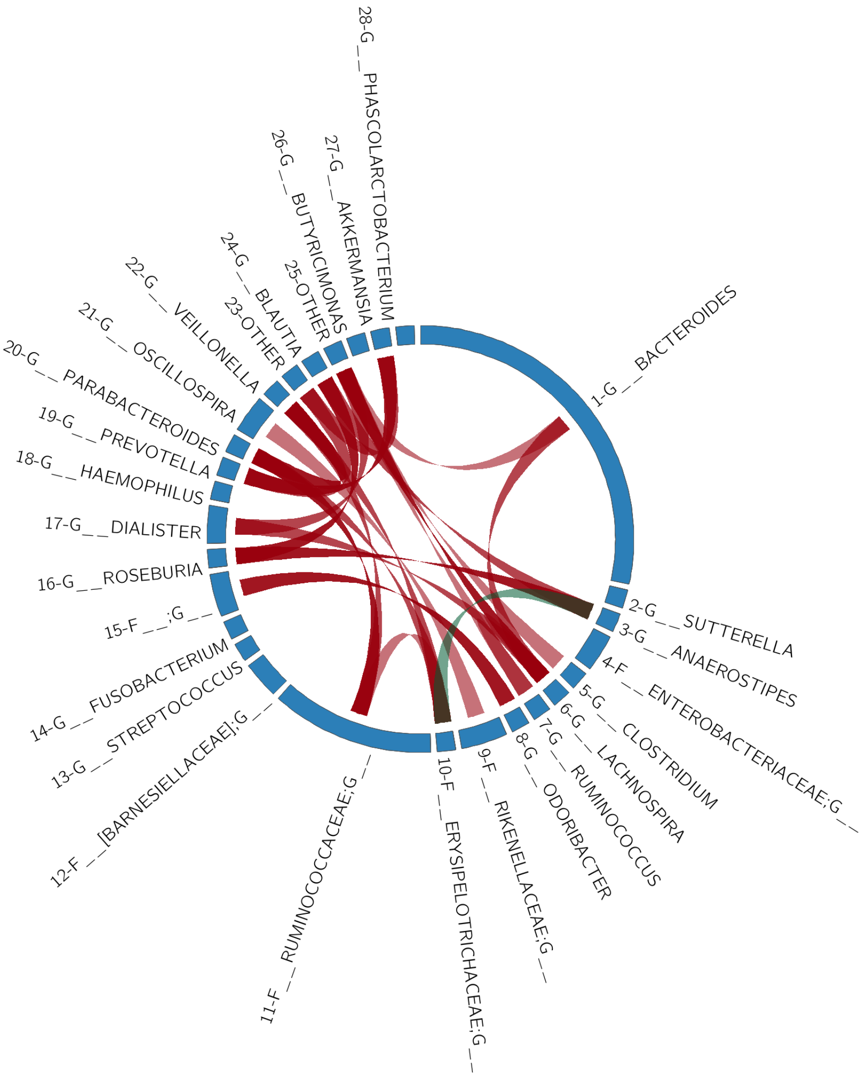
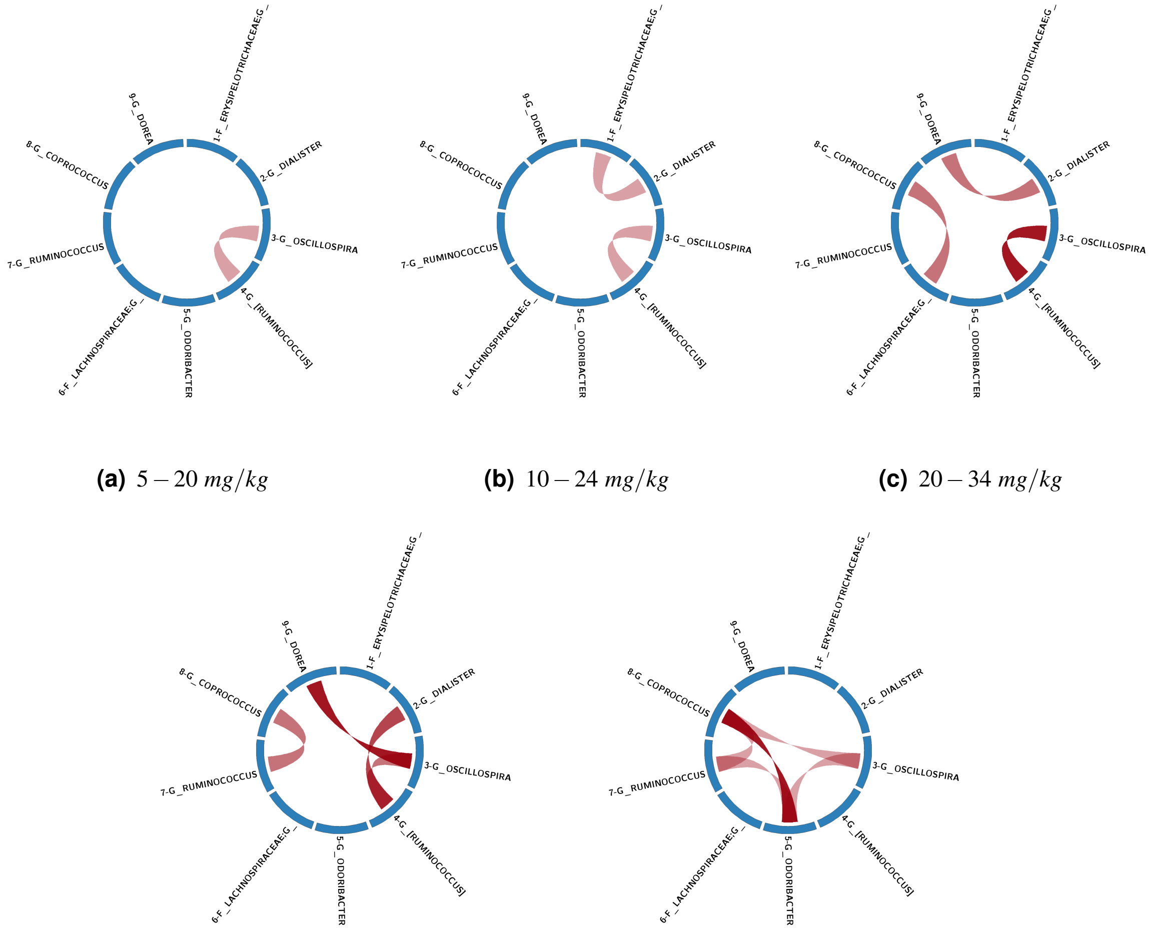


Figure 6.6. CIRCOS co-abundance networks on top-ranked features (FEC_H_B_NORM). Red edges: links conserved in B_NORM only; green edges: links conserved in FEC_H only. Edge's color intensity is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.75$.

Furthermore, analyses focused on network trajectories as a function of fecal calprotectin (Cp) concentration, which is one of the most popular non-invasive markers of IBD. It is commonly established that samples with $Cp < 50\ mg/kg$ can be regarded as healthy [156]. Calprotectin is a 36 kDa Calcium and Zinc binding protein expressed by neutrophils: neutrophil aggregation in the mucosa on inflamed intestine led to an increase of Cp concentration. Besides, it correlates well with fecal excretion of indium^{111}-labelled neutrophil granulocytes, the gold standard measure of gut inflammation [157]. The specificity of Cp for detecting IBD could be improved by considering another well-established risk factor of IBD, namely the gut microbiota. In detail, the structural changes of networks modeling interactions of gut microbial communities were analyzed in association to the levels of Cp. Thus, the B_H_IBD task was considered, given its best predictive performance in terms of MCC and stability of top discriminant genera list. First, Cp levels were divided into 5 consecutive ranges $(5-20, 10-24, 20-34, 25-113, 124-370\ mg/kg)$ grouping samples accordingly. Co-occurrence networks were inferred from each group of samples, so to compare microorganisms co-occurrence in relation to different calprotectin levels. The tendency of link number to increase proportionally to dysbiosis is evident (Fig. 6.7); it was noteworthy that for samples with $Cp > 50\ mg/kg$ correlation between bacteria was stronger than for samples with lower levels of dysbiosis. For instance, $PCC < 0.3$ between *Oscillospira* and *Odoribacter* for all groups with lower $(Cp < 113\ mg/kg)$ levels of calprotectin, while $PCC > 0.7$ for samples with the highest inflammation $(Cp > 124\ mg/kg)$. Correlation between *Odoribacter* and *Coprococcus* shows a similar trend, increasing with the raise in calprotectin levels; in particular, $PCC = 1$ in the range $124-370\ mg/kg$. On the contrary, links between *Erysipelotrichaceae* and *Dorea*, as well as *Erysipelotrichaceae* and *Coprococcus* tend to weaken with the increase of dysbiosis.



(a) $5-20\ mg/kg$ **(b)** $10-24\ mg/kg$ **(c)** $20-34\ mg/kg$

(d) $25-113\ mg/kg$ **(e)** $124-370\ mg/kg$

Figure 6.7. Networks for groups of samples with increasing ranges of calprotectin. CIRCOS plots of networks built on top ranked features from B_H_IBD task, considering samples grouped by increasing levels of calprotectin. Edge's color intensity is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.65$.

Pairwise HIM distances were computed between the network on samples with the lowest Cp range ($5-20mg/kg$) and the networks for increasing Cp levels. Note that inflammation intensifies as the network distance increases, suggesting an impact of dysbiosis on both global and local structure of microbial communities (Fig. 6.8)

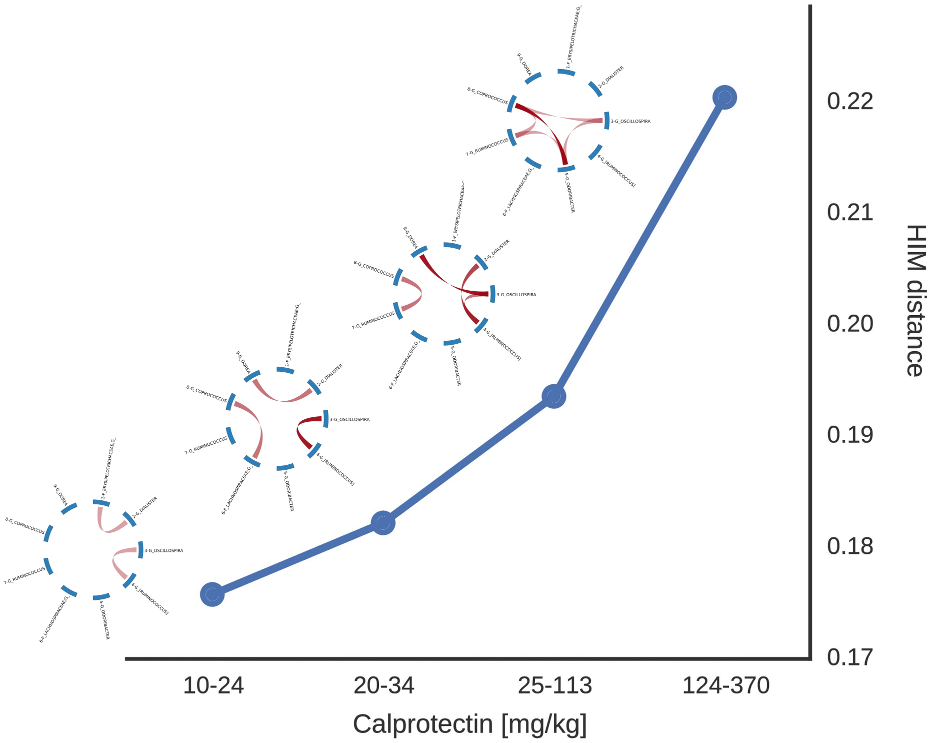
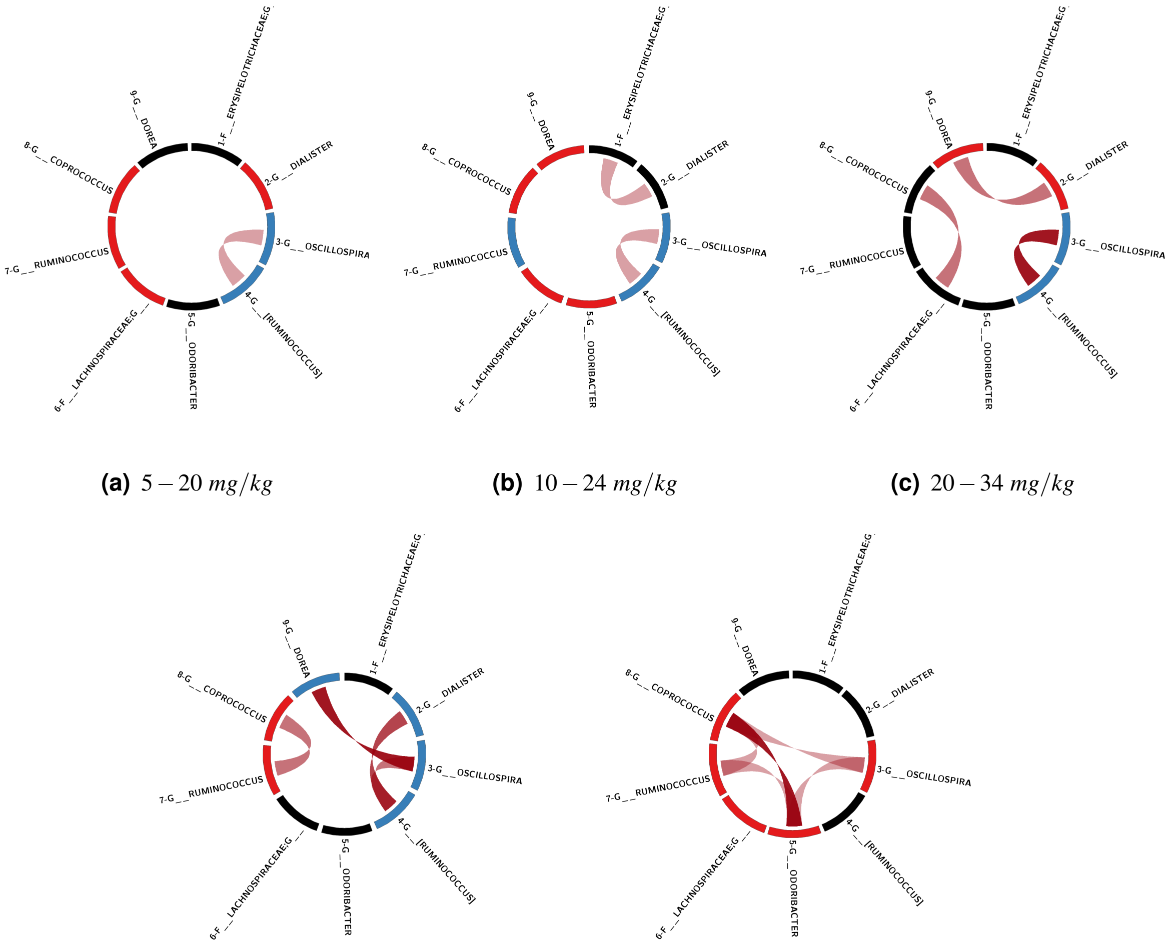


Figure 6.8. HIM distance between networks on samples with lowest vs networks on samples with increasing levels of calprotectin.

Community detection by Louvain method (Sec. 3.4) on the five networks highlighted that the number of communities decreases as inflammation intensifies (Fig. 6.9). Furthermore, we investigated how the composition of microbial communities change along Cp levels. In detail, we found that increasing dysbiosis breaks subcommunity *Ruminococcus-Oscillospira*, while *Odoribacter-Lachnospiraceae* is present for all calprotectin ranges but the lowest one.



(a) $5-20\ mg/kg$ **(b)** $10-24\ mg/kg$ **(c)** $20-34\ mg/kg$

(d) $25-113\ mg/kg$ **(e)** $124-370\ mg/kg$

Figure 6.9. Community detection on networks for groups of samples with increasing ranges of calprotectin. CIRCOS plot of networks on groups of samples with increasing levels of calprotectin, highlighting communities detected by Louvain method. Nodes with same color belong to the same community. Edge color intensity is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.65$.

6.1.4 Discussion

PreMONet identified the set of genera discriminating at best healthy children from IBD patients; this set includes *Lachnospira, Streptococcus, Dialister, Oscillospira, Ruminococcus* and genera belonging to *Rikenellaceae* and *Erysipelotrichaceae*.

Lachnospira, known to be butyrate-producing organism that can ferment dietary polysaccharides, was more abundant in healthy children, in accordance with [158], that proved its association with remission.

Besides, *Streptococcus* abundance was increased in IBD patients, evidence supported by Keshavarzian and colleagues [159]. Indeed, they showed an increase in glutathione transport and metabolism genes abundance in IBD; glutathione is a tripeptide of cysteine and glutamate, synthesized by a few streptococci and enterococci [160], which exerts an homeostatic function for bacteria during oxidative or acid stress. Notably, chronic inflammations lead to elevated oxygen levels in the intestine through increased blood flow and immunological responses associated with production of reactive oxygen and nitrogen metabolites. Thus, an increase in *Streptococcus*, the consequent increase in glutathione synthesis and metabolism may reflect a mechanism by which microbiome addresses the oxidative stress caused by inflammation [161]. The inflammation-related increase of oxygen levels leads to a perturbation of microbial composition by selecting subdominant facultative anaerobes and disrupting obligate anaerobic communities [162, 163]: interestingly, an anaerobic Gram-positive bacteria, *Ruminococcus*, was selected by PreMONet as one of the top discriminant biomarkers, decreased in IBD children.

Furthermore, *Erysipelotrichaceae* were identified as biomarkers of IBD both from fecal and biopsies samples, with an increased abundance in diseased children, in line with [164, 165, 166]. Indeed, specific taxa within *Erysipelotrichaceae* have been associated to inflammation [166], while others have been identified as highly immunogenic [165]. More in detail, Palm and colleagues [165] found that Immunoglobulin A (IgA)-coated members of intestinal microbiota conferred strong susceptibility to colitis in germ-free mice and showed that *Erysipelotrichaceae* are the highly coated by IgA relative to other bacteria. IgA is the predominant antibody isotype produced at mucosal surfaces, thus the recognition of enteric pathogens by the intestinal immune system results in the production of pathogen-specific IgA. The antibody exerts protection against infection by "coating" pathogens, which are neutralized and excluded. A dysbiotic microbiota can strongly influence IgA levels, as shown by Moon and colleagues [167] who found that *Sutterella* can degrade the secretory component of IgA, leading to reduced IgA levels. Interestingly, in our study *Sutterella* was more abundant in IBD children with respect to healthy subjects.

Furthermore, comparing healthy and IBD subjects, both from feces and biopsies (FEC_H_IBD and B_H_IBD tasks), *Roseburia* and *Phascolarctobacterium* were significantly reduced in IBD, as shown in [168, 161]; they are butyrate and propionate producers, respectively. Interestingly, it was proved that short-chain fatty acids (SCFAs) including butyrate and propionate play a protective role on epithelial cells and stimulate fluid absorption; butyrate is the major energy source for the epithelium, induces the differentiation of regulatory T cells (Tregs) and is important in the resolution of inflammation by signaling through G protein-coupled receptor 43 [162, 169, 170].

Another SCFA producer, *Odoribacter*, was selected by PreMONet as one of the top discriminant biomarkers in B_H_IBD task and it was found to colonize more healthy children than IBD ones. This is in accordance with [161], where *Odoribacter splanchus* was found to be reduced in patients with pancolitis and in patients with ileal CD. Although not a top discriminant IBD biomarker, our study highlighted an higher abundance of *Enterobacteriaceae* in IBD children, in line with [171]. Knights and colleagues found increased *Enterobacteriaceae* in subjects with higher NOD2 risk allele dosage. Interestingly, NOD2 was the first identified susceptibility gene for IBD [172, 173]. This gene codes for proteins sensing bacterial peptidoglycans (PGNs), which are essential components of bacteria cell wall; since they are not found in the host, they are recognized by specific host proteins called pattern recognition molecules, stimulating an immune response. Thus, an impaired function of NOD2 may cause an increase in bacteria producing PGNs or other pathogen-associated molecular patterns like bacterial lipopolysaccharides.

However, microbiota consists in diverse species interacting with one another, since microorganisms do not exist in isolation; the host health and well-being is critically influenced by the stability of microbial symbiotic relationships, rather than by individual species [174, 175]. Thus, the shifts in community composition should be analyzed together with the changes in microbial interactions, in order to better associate host illness to dysbiosis. In our PreMONet, microbial relationships were modeled by co-occurence networks in relation to calprotectin levels, which are commonly used in clinical settings to assess the activity of IBD. Consequently, our network analysis could be considered as a possible enhancement of calprotectin-based risk stratification tools.

In detail, severe inflammation (with consequent highest levels of calprotectin) was associated to some specific interactions between species. The strongest co-occurence was observed for *Odoribacter* and *Coprococcus*. Since they both are decreased in abundance in IBD children with respect to healthy ones, the host can not benefit of their anti-inflammatory activity, exerted by SCFA production. This is a consequence of cooperative metabolism between some species [7, 176], that commonly leads to system destabilization because the decrease in abundance of one species will tend to pull others down with it. The disruption of cooperative metabolism could be also observed from network communities analysis: *Ruminococcus* and *Oscillospira* belong to the same community only in healthy children.

Besides, *Coprococcus* and *Ruminococcus* were strongly correlated during inflammation, in accordance with Perez and colleagues [177], who performed a Bayesian network of microbial composition in Ulcerative Colitis, finding positive associations between *Lachnospiraceae* and *Ruminococcaceae*.

In conclusion, our study provides both a list of highly predictive biomarkers of IBD as well as an insight into the structure of microbial communities and its shifts from healthy to disease status. Certainly, a bigger cohort of samples and longitudinal dataset (microbiota composition and calprotectin levels of same samples along time) would confer even more robustness to our results.

6.2 Dataset A_IBD

IBD affects both adults and in children of all ages, with the peak age of onset before the age of 20 for 25%-30% of patients with Crohn's Disease (CD) and 20% of patients with Ulcerative Colitis (UC). The clinical course of disease differs between pediatric and adult patients, while the etiology may be similar, except for the effect of genetic factors. Adult patients show a positive family history for IBD less commonly than patients diagnosed before the age of 20; thus, risk factors for adults are mainly the environmental, such as smoking or enteric commensal bacteria [178, 179]. Since endogenous and external determinants modulate the child gut microbiota in a complex way, a recently adopted hypothesis is that the synergic meta-omics or systems biology approach could provide a comprehensive understanding of microbial communities perturbations in early IBD. A comprehensive integrative modeling of multiple meta-omics layers is thus required in disease-specific descriptions of the gut microbial community and its patho-physiologic evolution [22, 23, 24]. In this context, a clinical objective of integrative analysis is the discovery of multi-omic biomarkers to predict a phenotype of interest or drive the development of intervention protocols.

In particular, our I-PreMONet (see Chapter 5) provides INF (Integrative Network Fusion), which is a computational framework implementing the integration of bacterial and fungal datasets to identify an inter-kingdom biomarkers signature predictive of IBD phenotypes. Thus, I-PreMONet contributes to address two needs: developing a integrative meta-omics computational framework as well as investigating the role of fungi in IBD. Indeed, several human studies on the bacterial microbiota have observed imbalances of the gut microbiota in IBD patients, conversely only few studies have focused on the association between mycobiota and IBD [180, 181, 182]. Alterations in fungal biodiversity and composition in disease-specific gut environment have been reported, *i.e.,* modifying the *Ascomycota* to *Basidiomycota* ratio; furthermore, the existence of disease-specific inter-kingdom alterations has been suggested but not clarified yet.

6.2.1 Dataset

The dataset A_IBD was kindly provided by AVENIR Team “Gut Microbiota and Immunity” lab (MICALIS, Paris, France); all patients were recruited at the Gastroenterology Department of the Saint Antoine Hospital (Paris, France). A diagnosis of IBD was provided based on clinical, radiological, endoscopic and histological criteria. Only subjects that had not taken antibiotics or used colon-cleansing products for at least two months prior to enrolment were included in the study. Criteria to participate in the study included Crohn’s Disease (CD), ileal CD (iCD) or Ulcerative Colitis (UC), either in flare (f) or remission (r). Patient characteristics are presented Tab. 6.3.

Fecal samples were collected from 235 patients with IBD and 38 healthy (HS) individuals. Whole stools were collected in sterile boxes and immediately homogenised, and 0.2 g aliquots were frozen at -80°C for further analysis [180].

Patients affected by Pouchitis or with ileostomy have been subsequently excluded from the analyses, thus the study involved 222 IBD patients (60 CDf, 77 CDr, 41 UCf, 44 UCr, 44 iCDf, 59 iCDr) and 38 healthy subjects.

	IBD ($n = 235$)	HS ($n = 38$)
Age: Year (mean ± SD)	40.4 ± 14.6	35.8 ± 13.2
Male: n(%)	$94(40.0\%)$	$17(44.7\%)$
Flare/remission	$106(45.1\%)/129(54.9\%)$	NA

Table 6.3. A_IBD population characteristics

DNA extraction protocol and sequence data processing are reported in detail in [180], briefly described in the following paragraphs.

Genomic DNA extraction from fecal samples. Nucleic acids were precipitated by isopropanol for 10 minutes at room temperature, followed by incubation for 15 minutes on ice, and centrifugation for 30 minutes at $15,000\ g$ and 4℃. After the RNase treatment and DNA precipitation, nucleic acids were recovered by centrifugation at $15,000\ g$ and 4℃ for 30 minutes. The DNA pellet was finally suspended in $100\ \mu L$ of TE buffer.

Amplicon library preparation and sequencing. The sequence region of the 16S rRNA gene spanning the variable regions V3-V5 was amplified, subsequently a bidirectional library was prepared using the OneTouch2 Template Kit and sequenced on PGM Ion Torrent, using the Ion PGM Sequencing Kit (Life Technologies, Carlsbad, CA). Fungal diversity was determined for each sample via 454 pyrosequencing of Internal Transcribed Spacer 2 (ITS2), on a GS FLX Titanium Sequencing System (Roche Life Science Mannheim, Germany).

Taxonomic analysis

Sequences processing and quality control were performed in the Quantitative Insights into Microbial Ecology (QIIME) version 1.8.0 [25]. After barcodes and PCR primers were removed, reads were filtered to discard:

- 16S sequences shorter than 200 bp
- ITS2 sequences shorter than 150 bp
- reads with a base quality threshold < 25
- reads with homopolymers longer than 7 bp

Operational taxonomic units (OTUs) were binned at a sequence similarity level of 97% by using UCLUST algorithm [133]; their taxonomic assignments were obtained by mapping OTUs representative sequences against a reference database. Sequences were classified with UCLUST on Greengenes database [183] on 16S rDNA data, and against UNITE ITS database ($v.\ 12.11$) [184] on ITS2 data. In order to compare the OTUs abundances across samples, rarefaction analysis was performed (2,041-83,162 reads/sample for 16S and 540-5,648 reads/sample for ITS2) and samples with sequence depth lower than 10,000 sequences per sample for 16S data and 1,000 sequences per sample for ITS2 data were discarded.

Statistical analysis

In order to find biomarkers discriminating healthy subjects from IBD patients, six classification tasks were considered, comparing controls group with each IBD phenotype. Besides, UC and CD patients were compared. In summary:

- for healthy (HS) vs. all IBD subphenotypes: HS vs. CDf (HS_CDf , for short), HS_CDr, HS_UCf , HS_UCr, HS_iCDf and HS_iCDr
- for CD vs. UC comparison: CDf vs. UCf (CDf_UCf , for short) and CDr_UCr

For each task, Random Forest and Support Vector Machines were developed on 70% of the datasets (training partition) composed of bacterial and fungal genera

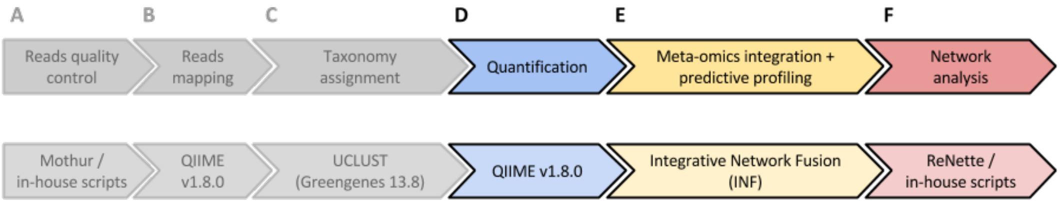


Figure 6.10. I-PreMONet applied on A_IBD dataset. Data not processed by grey modules.

from healthy individuals and patients belonging to one of the IBD phenotypes. The 30% of the datasets were used as blind validation set, testing the performance of trained models on unseen data. Integration of bacterial and fungal abundances and their predictive profiling were performed in module E (INF) of I-PreMONet pipeline (Fig. 6.10).

6.2.2 Meta-omics integration

The innovative core of I-PreMONet is the predictive profiling of bacterial and fungal DNA abundances with a novel approach to their integration (modules E defined as INF, Fig. 6.10). Omics data integration has been defined by Ritchie and colleagues [118] as the combination of multiple omics datasets so to develop classification models that are predictive of complex traits or phenotypes. INF module of I-PreMONet contributed both to datasets combination and predictive models, combining an improved version of a state-of-the-art integration technique [119] (Sec. 4.1) with predictive models developed inside a gold-standard Data Analysis Protocol [53] (see Sec. 2.3) for machine learning. In summary, in INF three integrative approaches were implemented and compared. First, the standard method was considered by concatenating bacterial and fungal features and training Random Forest (RF) classifiers on the combined dataset, finally obtaining a ranked list of biomarkers for the IBD tasks. This approach is referred as to ml-J. Secondly, bacteria and fungi were integrated by Similarity Network Fusion (SNF) [119], a non-Bayesian network-based method that computes a samples' similarity network for each data type and fuses them into one network. SNF technique was extended by adding a feature ranking procedure that sorts bacteria and fungi according to their contribution to the fused network structure (see Sec. 4.2). Again, RF models were developed on the integrated dataset for the SNF-ranked list of meta-omics variables. This approach is referred as to ml-rSNF. Finally, a compact model (INF) trained on the intersection of features from direct concatenation and rSNF features was derived. The details of integrative module are also summarized in Fig. 6.11. Compared to the state-of-the-art method ml-J, INF not only achieved comparable or better predictive performance, but also identified a compact list of inter-kingdoms biomarkers.

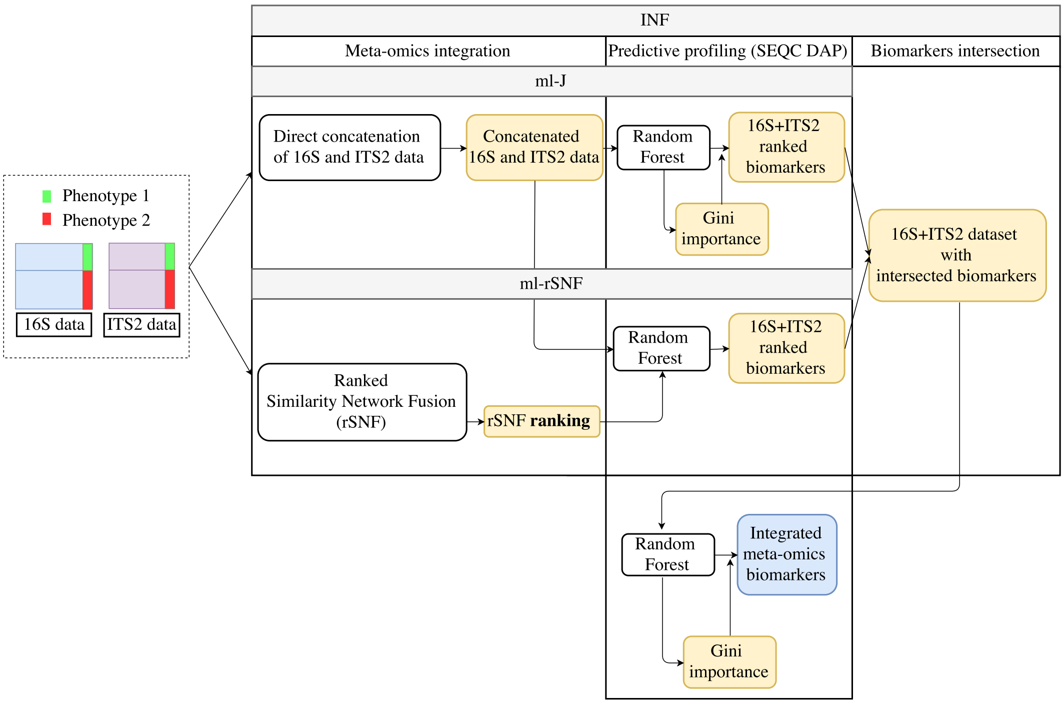

INF
Meta-omics integration
Predictive profiling (SEQC DAP)
Biomarkers intersection
ml-J
Phenotype 1
Phenotype 2
16S data
ITS2 data
Direct concatenation of 16S and ITS2 data
Concatenated 16S and ITS2 data
Random Forest
16S+ITS2 ranked biomarkers
Gini importance
16S+ITS2 dataset with intersected biomarkers
ml-rSNF
Random Forest
16S+ITS2 ranked biomarkers
Ranked Similarity Network Fusion (rSNF)
rSNF ranking
Random Forest
Integrated meta-omics biomarkers
Gini importance

Figure 6.11. INF module of I-PreMONet on the A_IBD dataset. The methods ml-J (RF on juxtaposed datasets) and ml-rSNF (RF on combined datasets with rSNF-ranked variables) were run in parallel. Integrated meta-omics signature was computed by RF on the datasets restricted on the intersection of ml-J and ml-rSNF biomarkers.

6.2.3 Results

Sequence processing and quality control led to 11,099,768 reads for 16S rDNA data and 755,350 for ITS2 data. For our further analyses the 16S and ITS2 OTUs tables (rows as taxonomically classified OTUs and columns as samples) were collapsed to the genus level, combining OTUs belonging to the same genus. In detail, matrixes with 308 16S and 187 ITS2 genera frequencies for each sample (rows as samples and columns as genera) were built. Samples were grouped according to the classification tasks (see Sec. 6.2.1), leading to eight tables, whose dimensions are detailed in Tab. 6.4.

Task	# samples
HS_CDf	98
HS_CDr	115
HS_UCf	79
HS_UCr	82
HS_iCDf	82
HS_iCDr	97
CDf_UCf	101
CDr_UCr	121

Table 6.4. A_IBD datasets dimension: number of samples for each classification task.

Predictive performance of the three integrative approaches in INF module (ml-J, ml-rSNF and INF) are reported in terms of best average MCC on training set (MCC_{int}) with 95% Student bootstrap ($1000\times$resampling) confidence intervals (MCC_{min} , MCC_{max}), MCC on validation set (MCC_{val}), number of features (Nf) leading to MCC_{int}. Both RF and SVM were developed, but in the main text only RF results will be reported (SVM results in Appendix C).

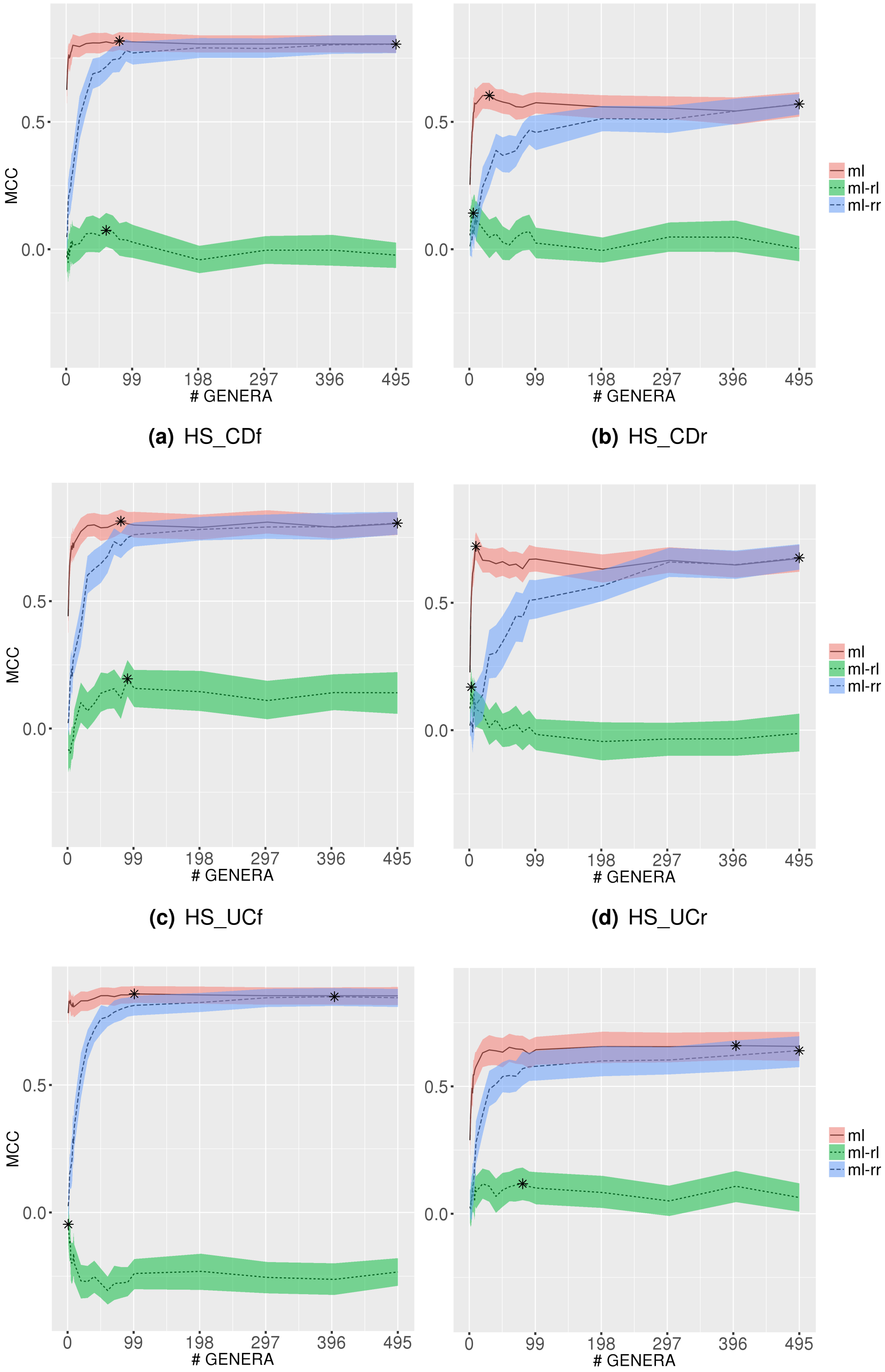
ml-J performance

The first approach to find inter-kingdoms IBD biomarkers consisted in concatenating bacterial and fungal datasets, and training predictive models on the joint dataset. Random Forest predictive results for each IBD classification task are reported in Tab. 6.5.

	ml-J		
	MCC_{int} (MCC_{min}, MCC_{max})	MCC_{val}	Nf
HS_CDf	0.82 (0.78, 0.85)	0.70	80
HS_CDr	0.60 (0.55, 0.65)	0.50	30
HS_UCf	0.81 (0.77, 0.86)	0.74	80
HS_UCr	0.72 (0.67, 0.78)	0.51	10
HS_iCDf	0.86 (0.82, 0.89)	0.66	100
HS_iCDr	0.66 (0.60, 0.71)	0.54	400
CDf_UCf	0.52 (0.48, 0.57)	0.14	20
CDr_UCr	0.33 (0.27, 0.39)	0.50	60

Table 6.5. Synopsis of RF accuracy for each classification task, on juxtaposed 16S and ITS2 datasets. (MCC_{min}, MCC_{max}): 95% bootstrap confidence interval. MCC_{int}: best mean MCC on training set; MCC_{val}: MCC on validation set; Nf: number of genera leading to MCC_{int}.

First, observe that remission cases and healthy subjects were more difficult to discriminate, as expected, since they are expected to have closer metagenomic profiles; on the contrary, flare conditions were more predictable. In detail, best performance balancing MCC_{int}, MCC_{val} and number of top discriminant features was achieved by comparing healthy subjects (HS) and patients with Ulcerative Colitis in flare (UCf). As expected, the comparison between CD and UC in remission, led to the worst predictive performances; indeed, CDr and UCr subjects usually have similar metagenomics profiles, since their microbiota tends to evolve towards a “healthier” composition. Moreover, by evaluating mean MCC values achieved by classifiers at increasing genera set sizes (shown in Fig. 6.12, 6.13), RF clearly did not overfit data: ml-rr curve increased with the feature steps, reaching the maximum by using the entire set of genera. The random ranking (ranking by shuffling) procedure does not select genera for prediction; thus, the classifier accuracy increases with the genera set size but much slower of the ranking methods, eventually reaching them once a number of discriminant genera is eventually included. Selection bias was also avoided; the ml-rl curve oscillated around $MCC = 0$, the expected accuracy of models trained with by a random labelling procedure disrupting the association between samples and phenotypes.



(a) HS_CDf

(b) HS_CDr

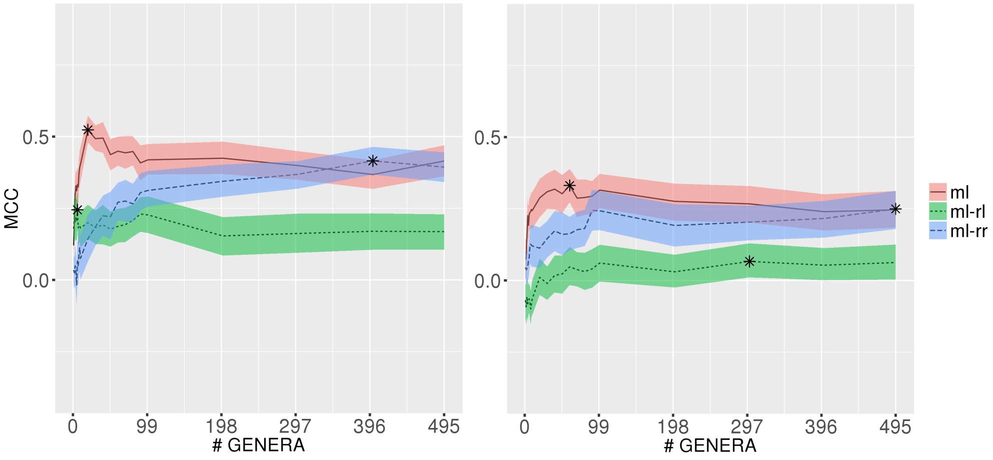
(c) HS_UCf

(d) HS_UCr

(e) HS_iCDf

(f) HS_iCDr

Figure 6.12. MCC plots for the healthy vs. IBD classification tasks. Solid curves in black indicate MCC, with corresponding 95% bootstrap confidence intervals, at increasing genera set sizes (# GENERA). MCC_{int} highlighted with ($*$). Random labels (RL) and random ranking (RR) experiments are indicated respectively in green and in blue (see text).



(a) CDf_UCf **(b)** CDr_UCr

Figure 6.13. MCC plots for the CD vs UC classification tasks. Solid curves in black indicate MCC, with corresponding 95% bootstrap confidence intervals, at increasing genera set sizes (# GENERA). MCC_{int} highlighted with ($*$). Random labels (RL) and random ranking (RR) experiments are indicated respectively in green and in blue (see text).

ml-rSNF performance

Datasets juxtaposition is the most naive integration technique, but it dilutes the possibly low signal-to-noise ratio in each data type, affecting the understanding of the biological interactions at omics levels. The refined SNF method (rSNF) accounts for common and correlated information between 16S and ITS2 data, producing a ranked list of features for each classification task. The machine learning DAP (see Sec. 2.3) was then applied on juxtaposed datasets, but biomarkers lists were built by using features weights from rSNF and not from RF classifiers.

MCC_{int}, MCC_{val} and Nf for each classification task are listed in Tab. 6.6.

	ml-rSNF		
	MCC_{int} (MCC_{min}, MCC_{max})	MCC_{val}	Nf
HS_CDf	0.83 (0.79, 0.87)	0.70	40
HS_CDr	0.60 (0.54, 0.65)	0.50	200
HS_UCf	0.81 (0.76, 0.85)	0.74	495
HS_UCr	0.67 (0.62, 0.73)	0.32	300
HS_iCDf	0.86 (0.83, 0.90)	0.75	100
HS_iCDr	0.65 (0.59, 0.70)	0.54	100
CDf_UCf	0.43 (0.39, 0.48)	0.28	300
CDr_UCr	0.27 (0.20, 0.33)	0.29	495

Table 6.6. Summarized best predictive performances of RF classifiers for each classification task, on juxtaposed 16S and ITS2 datasets, by using rSNF weights to rank features. (MCC_{min}, MCC_{max}): 95% bootstrap confidence interval. MCC_{int}: best mean MCC on training set; MCC_{val}: MCC on validation set; Nf: number of genera leading to MCC_{int}.

Higher performances were achieved when healthy subjects were compared to flare cases than remission cases, as for the most naive integration technique. Coherently with the underlying biology, discrimination between CDr and UCr samples led still to the lowest predictive performances. In general, as regards to MCC_{int}, predictive models using feature weights from rSNF obtained slightly better or comparable results (inside the confidence intervals) to models developed on juxtaposed datasets. However, Nf depends on the specific classification task: indeed, for CDr, UCf and UCr cases, ml-rSNF needed an higher number of features to obtain the best mean MCC score.

INF performance

For each classification task, the top discriminant feature lists found by ml-J (DAP on juxtaposed datasets) were intersected with the ones from ml-rSNF (DAP on juxtaposed datasets, rSNF as feature ranking method). The DAP was run on datasets reduced to the 16S and ITS2 genera common to ml-J and ml-rSNF, obtaining models denoted as INF.

Table 6.7 and Figure 6.14 report predictive performances of ml-J, ml-rSNF and INF in both a tabular and graphical representation.

	ml-J			**ml-rSNF**			**INF**		
	MCC_{int} (MCC_{min}, MCC_{max})	MCC_{val}	Nf	MCC_{int} (MCC_{min}, MCC_{max})	MCC_{val}	Nf	MCC_{int} (MCC_{min}, MCC_{max})	MCC_{val}	Nf
HS_CDf	0.82 (0.78, 0.85)	0.70	80	0.83 (0.79, 0.87)	0.70	40	0.84 (0.80, 0.87)	0.62	30
HS_CDr	0.60 (0.55, 0.65)	0.50	30	0.60 (0.54, 0.65)	0.50	200	0.65 (0.60, 0.69)	0.50	29
HS_UCf	0.81 (0.77, 0.86)	0.74	80	0.81 (0.76, 0.85)	0.74	495	0.81 (0.76, 0.86)	0.74	60
HS_UCr	0.72 (0.67, 0.78)	0.51	10	0.67 (0.62, 0.73)	0.32	300	0.75 (0.69, 0.79)	0.51	7
HS_iCDf	0.86 (0.82, 0.89)	0.66	100	0.86 (0.83, 0.90)	0.75	100	0.85 (0.80, 0.88)	0.84	20
HS_iCDr	0.66 (0.60, 0.71)	0.54	400	0.65 (0.59, 0.70)	0.54	100	0.65 (0.59, 0.70)	0.54	80
CDf_UCf	0.52 (0.48, 0.57)	0.14	20	0.43 (0.39, 0.48)	0.28	300	0.65 (0.60, 0.69)	0.14	20
CDr_UCr	0.33 (0.27, 0.39)	0.50	60	0.27 (0.20, 0.33)	0.29	495	0.37 (0.31, 0.43)	0.38	60

Table 6.7. RF predictive performances for ml-J, ml-rSNF and INF. MCC_{int}: best mean MCC on training set; (MCC_{min}, MCC_{max}): MCC_{int} 95% bootstrap confidence interval; MCC_{val}: MCC on validation set; Nf: number of genera leading to MCC_{int}.

Notably, INF obtained higher or comparable results in terms of MCC_{int} than ml-J or ml-rSNF, but with a significant reduction in number of top discriminant genera. Besides, INF results in terms of MCC_{val} were better or equal compared to ml-J or ml-rSNF, with the only exception of HS_CDf. In summary, the intersection of top discriminant features from ml-J and ml-rSNF led to a significant improvement in predictive performance. The best predictive performance as well as the best improvement of INF with respect to ml-J were achieved for HS_iCDf.

Remarkably, INF did not reduce Nf only for tasks comparing CD and UC samples, *i.e.,* is when the two IBD classes represent subjects with the same phenotype condition (either flare or remission) and samples were supposed to carry similar gut microbiota due to similar phenotypes.

Random labels and random ranking experiments ensured that the predictive models were not affected by selection bias or overfitting issues (Fig. 6.15).

The biological relevance of biomarkers from the different approaches was also assessed, comparing the top discriminant features lists with other studies in literature [148, 161, 180]. Detailed lists of biomarkers and their abundance are reported in Appendix B.

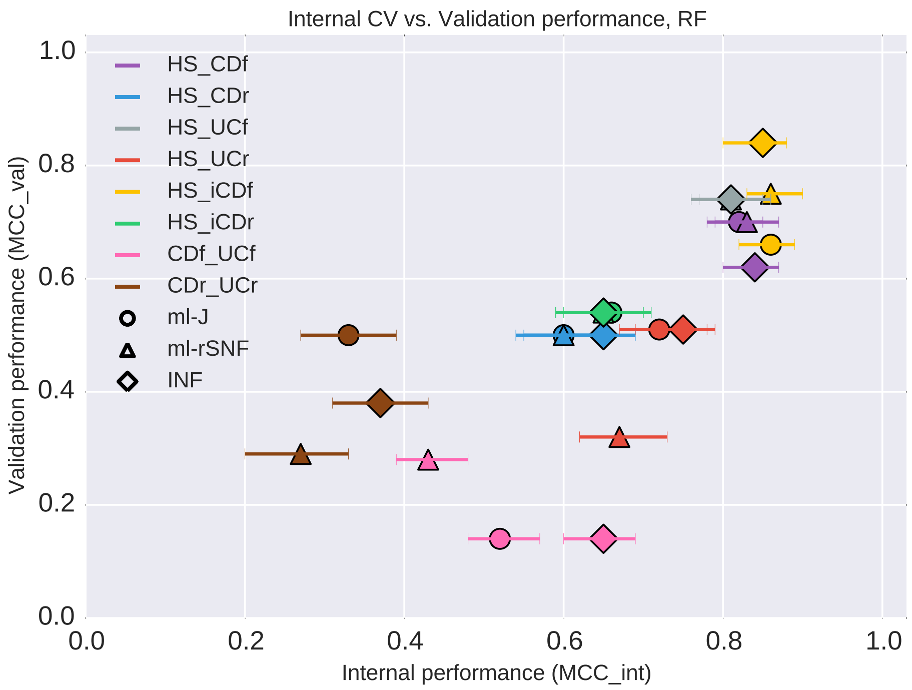
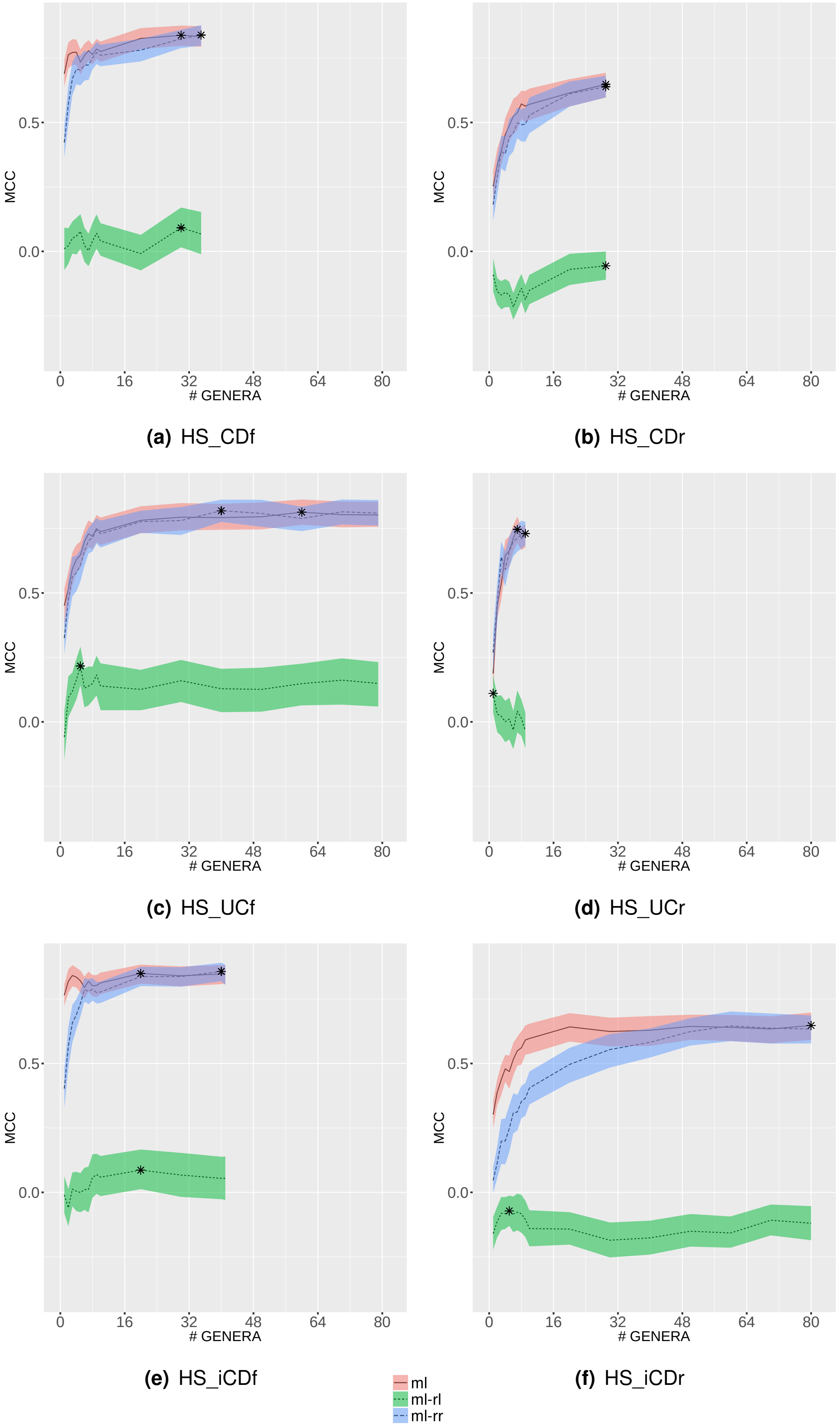


Figure 6.14. MCC on training and validation sets for ml-J, ml-rSNF and INF. Best average MCC on training set (MCC_{int}) vs MCC on validation set (MCC_{val}) for ml-J, ml-rSNF and INF by RF; horizontal bars correspond to 95% bootstrap confidence intervals of MCC values on training sets.



(a) HS_CDf

(b) HS_CDr

(c) HS_UCf

(d) HS_UCr

(e) HS_iCDf

(f) HS_iCDr

Figure 6.15. MCC plots for the healthy vs. IBD classification tasks. Solid curves in black indicate MCC, with corresponding 95% bootstrap confidence intervals, at increasing genera set sizes (# GENERA). MCC_{int} highlighted with ($*$). Random labels (RL) and random ranking (RR) experiments are indicated respectively in green and in blue (see text).

Full scripts and datasets are available in a GitHub repository to ease reuse (`https://github.com/AleZandona/INF`).

Furthermore, microbial communities structure were analyzed, by computing Pearson Correlation Coefficient (PCC) among the top-ranked genera of the classification tasks with the best performance. As in 6.1, CIRCOS plots were used as graphical layout for networks. In details, analyzing the co-abundance network for task HS_iCDf (Figure 6.16), healthy subjects showed a strong link between genera *Holdemania* and *Dehalobacterium* and family *Christensenellaceae*, while the strongest links found in iCDf patients only were: *Ruminococcaceae* to *Coriobacteriaceae*, *Holdemania* to *Ruminococcaceae* and *Coriobacteriaceae* to *Paraprevotella*.

Besides, Fig. 6.17 reports the co-abundance networks on top-ranked genera in HS_UCf task. In healthy individuals (green edges), *Pasteurellaceae* was linked to *Odoribacter* and *Sutterella*; in UCf patients, only two links were found: *Actinobacillus* was linked to both *Veillonella* and *Paraprevotella*.

As a further analysis, community detection was performed by Louvain method (Sec. 3.4); in detail, Fig. 6.18 and Fig. 6.19 compare the microbial communities in healthy individuals and in iCDf or UCf patients, respectively. In particular, we found that in iCDf subjects the number of communities decreases with respect to healthy individuals, and that the dysbiosis breaks the subcommunities *Ruminococcaceae-Ruminococcus* and *Streptococcus-Coriobacteriaceae*. Fig. 6.19 highlights that Ulcerative Colitis induces perturbation in communities structure; for instance, dysbiosis breaks some communities (*Coriobacteriaceae-Dehalobacterium-Roseburia* and *Parvimonas-Proteus-Firmicutes*), but also induces some others (*Roseburia-Phascolarctobacterium-Paraprevotella* and *Lactococcus-Aggregatibacter-Lactobacillus*).

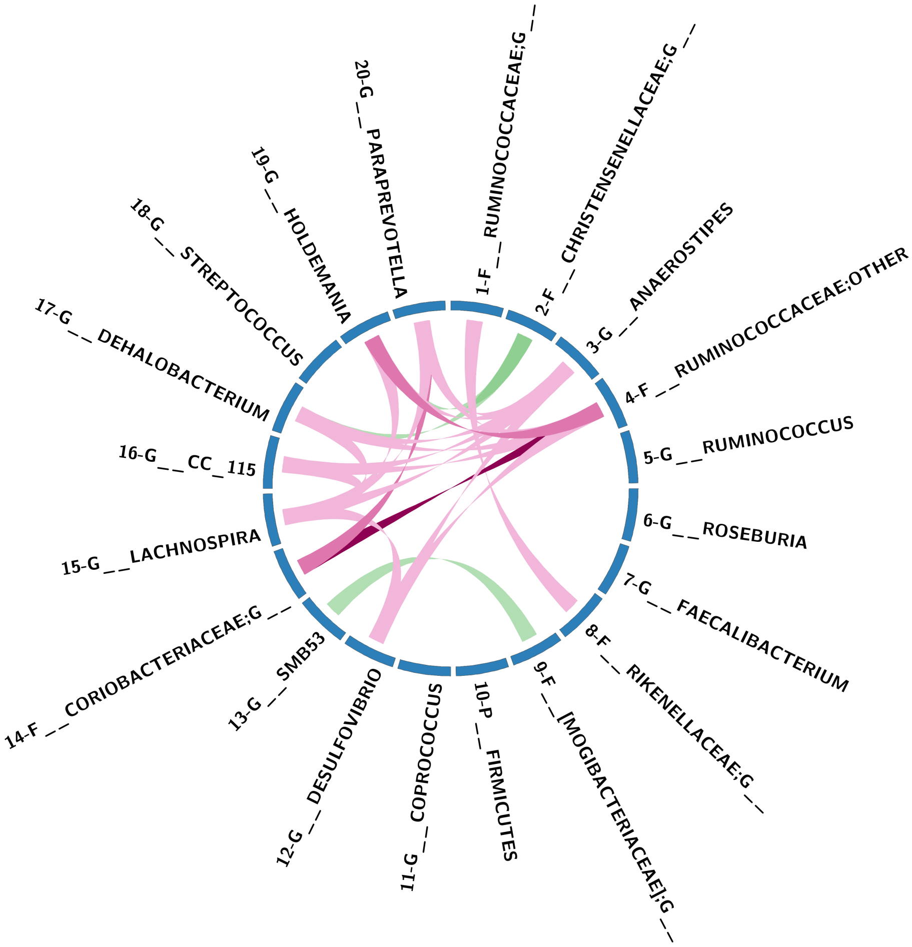


Figure 6.16. CIRCOS co-abundance networks on top-ranked genera (HS_iCDf). Pink edges: links conserved in iCDf only; green edges: links conserved in HS only. Edge's color intensity is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.7$.

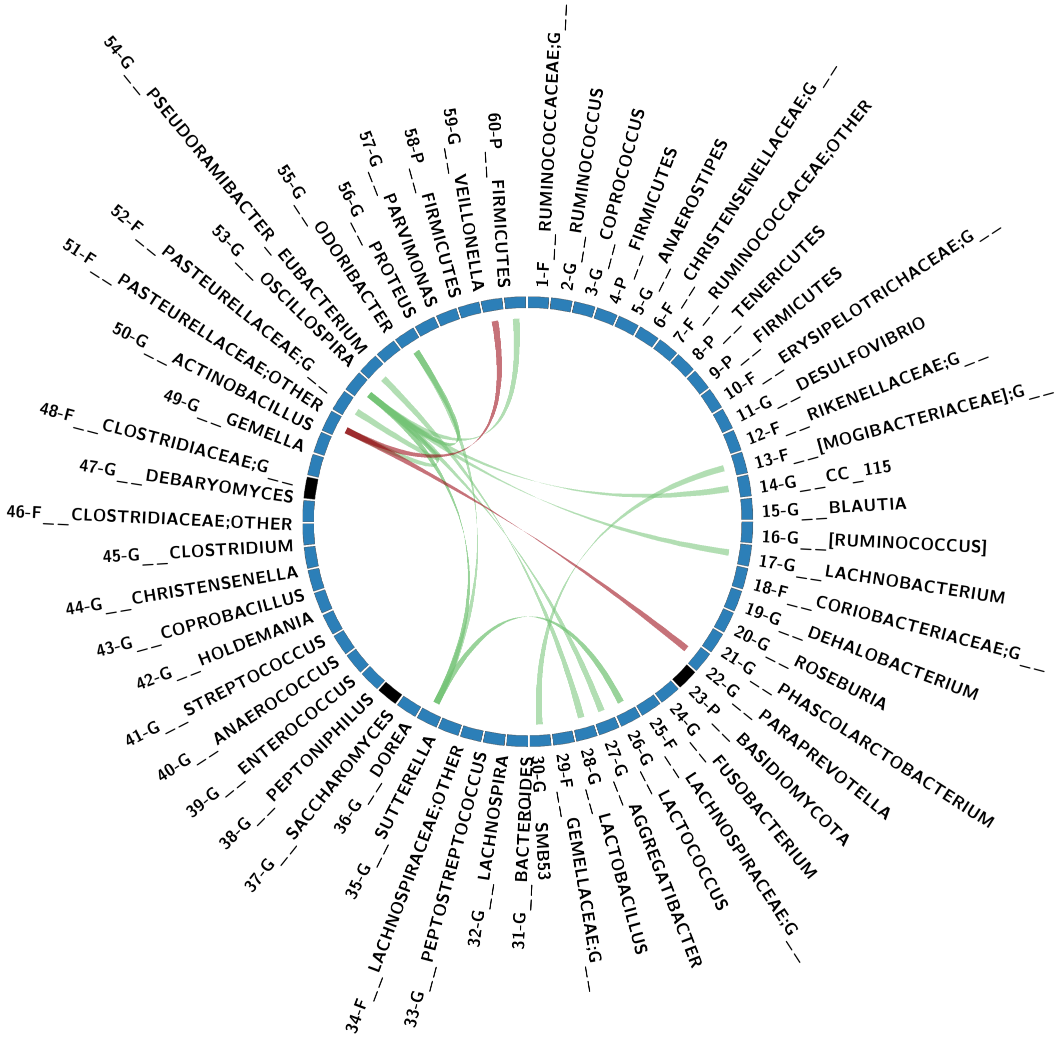
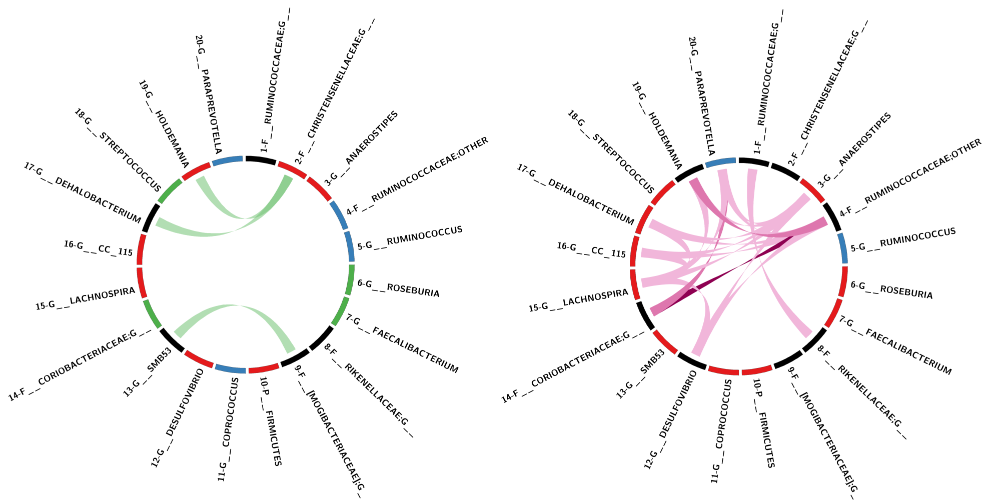


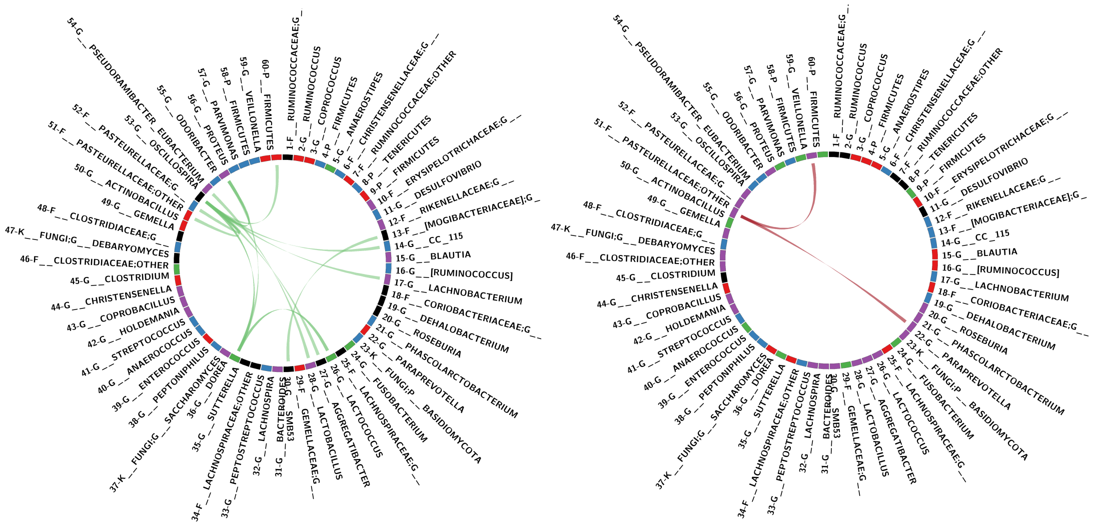
Figure 6.17. CIRCOS co-abundance networks on top-ranked genera (HS_UCf). Red edges: links conserved in UCf only; green edges: links conserved in HS only. Edge's color intensity is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.6$.



(a) HS

(b) iCDf

Figure 6.18. Community detection on networks on top-ranked genera (HS_iCDf). CIRCOS plot of networks on healthy (a) vs. iCDf (b) samples, highlighting communities detected by Louvain method. Nodes with same color belong to the same community. Edge color intensity is proportional to the absolute value of Pearson correlation coefficient (PCC). Edges are thresholded at $PCC = 0.7$.



(a) HS

(b) UCf

Figure 6.19. Community detection on networks on top-ranked genera (HS_UCf). CIRCOS plot of networks on healthy (a) vs. UCf (b) samples, highlighting communities detected by Louvain method. Nodes with same color belong to the same community. Edge color intensity is proportional to the absolute value of PCC. Edges are thresholded at $PCC = 0.6$.

Core biomarkers

From the previous analyses, lists of 16S and ITS2 genera discriminating between healthy subjects and each IBD phenotype (CD, UC, iCD) were identified. Furthermore, intersection of biomarkers lists was performed, in order to find sets of genera discriminating between healthy condition and more than one phenotype. First, INF discriminant biomarkers from three pairs of classification tasks were intersected: HS_CDf and HS_CDr, HS_UCf and HS_UCr, as well as HS_iCDf and HS_iCDr. The three sets of biomarkers could be associated to genera discriminating between healthy and CD, UC and iCD subjects respectively, independently from the flare or remission conditions (Tables 6.8, 6.9, 6.10).

HS vs CD biomarkers, INF		
f__[Barnesiellaceae];g__	g__Anaerostipes	g__Ruminococcus
f__Ruminococcaceae;Other	g__Faecalibacterium	g__Dehalobacterium
f__Christensenellaceae;g__	f__Ruminococcaceae;g__	o__Clostridiales;f__;g__
o__Clostridiales;Other;Other	o__RF39;f__;g__	g__Lachnobacterium
g__Clostridium	g__[Eubacterium]	f__Coriobacteriaceae;g__
g__Paraprevotella	f__[Mogibacteriaceae];g__	f__Erysipelotrichaceae;g__cc_115
g__Holdemania	g__Desulfovibrio	
f__Rikenellaceae;g__	g__Oscillospira	

Table 6.8. Bacterial genera discriminating healthy subjects from Crohn's patients both in flare or remission. Median of relative abundance of all biomarkers is higher in HS than in CD.

HS vs UC biomarkers, INF
g__Coprococcus
f__Ruminococcaceae;Other
o__RF39;f__;g__
g__Dehalobacterium
o__Clostridiales;f__;g__
f__[Mogibacteriaceae];g__
g__Desulfovibrio

Table 6.9. Bacterial genera discriminating healthy subjects from Ulcerative Colitis patients both in flare or remission. Median of relative abundance of all biomarkers is higher in HS than in UC.

HS vs iCD biomarkers, INF	
g__Anaerostipes	g__Dehalobacterium
g__Lachnospira	o__Clostridiales;f__;g__
f__Ruminococcaceae;g__	f__[Mogibacteriaceae];g__
g__Coprococcus	f__Rikenellaceae;g__
f__Ruminococcaceae;Other	g__Desulfovibrio
g__Ruminococcus	g__Holdemania
f__Christensenellaceae;g__	f__Erysipelotrichaceae;g__cc_115

Table 6.10. Bacterial genera discriminating healthy subjects from ileal Crohn's patients both in flare or remission. Median of relative abundance of all biomarkers is higher in HS than in iCD.

Furthermore, biomarkers from CDf_UCf and CDr_UCr were intersected, in order to find potential bacteria and fungi discriminating CD patients from UC. Table 6.11 reports the intersection result.

CD vs UC biomarkers, INF		
(B) g__Bacteroides	**(B) g__Faecalibacterium**	(B) f__Enterobacteriaceae;g__
(B) f__[Barnesiellaceae];g__	**(B) g__Oscillospira**	(B) g__Escherichia
(B) g__Parabacteroides	(B) g__Catenibacterium	(B) k__Bacteria;Unassigned
(B) g__Lactococcus	(B) f__Erysipelotrichaceae;g__cc_115	**(F) g__Debaryomyces**
(B) g__[Ruminococcus]	(B) g__Clostridium	
(B) f__Ruminococcaceae;g__	(B) g__Sutterella	

Table 6.11. Bacterial (B) or fungal (F) genera discriminating CD patients from UC. Bold: median of relative abundance is higher in CD than in UC. For other biomarkers, median of relative abundance is higher in UC than in CD.

Finally, the biomarkers potentially associated with IBD occurrence were identified by the intersection among INF top discriminant genera from all six classification tasks (Table 6.12) involving healthy subjects.

HS vs IBD, INF
o__Clostridiales;f__;g__
f__Ruminococcaceae;Other
g__Desulfovibrio
f__[Mogibacteriaceae];g__
g__Dehalobacterium

Table 6.12. Bacterial genera discriminating healthy subjects from IBD patients. Median of relative abundance of all biomarkers is higher in HS than in IBD.

6.2.4 Discussion

Disease-specific dysbiosis in fungal microbiota has been recently observed, as well as the interactions between bacteria and fungi, in the context of IBD. However, the associations between bacterial microbiota or mycobiota and disease phenotypes have always been investigated with separate analyses on bacteria and fungi. Moreover, inter-kingdom alterations have been studied by correlation analyses, but the prognostic power of these interactions has still to be unraveled. Our study was the first attempt to build a framework for IBD prognosis (referred as to INF), exploiting the interactions within and between kingdoms in the gut microbiota. Shared and complementary information from bacteria and fungi are first integrated, building a joint-model accounting for how informative each inter-kingdom interaction is to IBD phenotypes. Furthermore, classification algorithms are developed on bacterial and fungal abundances, taking into account the information combined in the first step. Hence, our framework results into a panel of inter-kingdom biomarkers correlated with IBD occurrence. The performance of INF could be assessed both in terms of statistical properties as well as biological interest. As regards the statistical aspect, INF was compared with predictive models developed on the concatenated bacterial and fungal datasets (ml-J technique). This framework commonly outperforms predictive performances of other techniques, since it exploits all the information from the combined datasets, although it dilutes the possibly low signal-to-noise ratio in each data type, affecting the understanding of the biological interactions at omics levels. Notably, for INF, less biomarkers were systematically needed to obtain comparable or even higher performance results both in classifying samples according to known phenotypes, as well as in predicting phenotypic groups in new test samples. Clearly, this is an added value for INF, since biomarkers need to be biologically validated before entering the clinical practice, thus an increased number of biomarkers is currently a limit. This result was mainly due to our rSNF, which increased the signal-to-noise ratio from the combined datasets, by prioritizing the most discriminative biomarkers prior to develop predictive models on them. Furthermore, compared to ml-J, INF identified biomarkers with more biologically plausible association with phenotypes. Indeed, INF produced panels of biomarkers with a higher ratio of biologically significant variables over total number of top discriminant features than ml-J.

As a summary: inter-kingdom biomarkers identified by INF are known in literature [180, 148, 161, 185]; concerning bacterial microbiota, genera *Coprococcus, Ruminococcus, Blautia, Anaerostipes* and genera from families *Gemellaceae* and *Coriobacteriaceae* were decreased in IBD. Conversely, *Streptococcus* was increased in IBD flare, while UC flare showed an increased abundance of *Fusobacterium, Aggregatibacter, Actinobacillus, Enterococcus, Peptoniphilus, Anaerococcus, Sutterella and Bacteroides*.

In particular, in ileal CD, the abundance of members of the family *Ruminococcaceae*, in particular *Faecalibacterium*, was decreased compared to healthy individuals, as reported in [34, 186, 161]. *Faecalibacterium prausnitizii* is a major producer of the SCFA butyrate, with anti-inflammatory effects in a colitis setting [187] and provides the first step in microbiome-linked carbohydrate metabolism by degrading dietary polysaccharides [188]. This bacterium elicits high levels of IL-10 production, enhancing ovalbumin-specific T cell proliferation, and reducing interferon gamma-positive T cells [189]. Besides, higher risk of postoperative recurrence of ileal CD [187] have been associated to the reduction in *F. prausnitizii* abundance [187], supported by the evidence that administration of *F. prausnitizii* reduces inflammation in mouse models.

Roseburia is reduced in all IBD subgroups, as shown also in [161], and it is connected to the family *Ruminococcaceae* as it relies on its members to produce acetate, which it uses to produce butyrate.

Another biomarker of inflammation found was *Fusobacteria*, with increased abundance in IBD, as confirmed by [190]. It is a phylum of adherent and invasive bacteria, inducing inflammatory responses through its unique FadA adhesin, which binds to E-cadherin and activates β-catenin signalling [191]. Potential mechanisms of function of *Fusobacterium* in IBD have not been described, but the invasive ability of *Fusobacterium* has been positively correlated with the IBD status of the host

[192]. *Fusobacterium* provides a potential theoretical link given the increased risk of CRC associated with IBD.

The association of mycobiome with IBD is a relatively unexplored area of research, but some studies confirmed our findings in terms of fungal microbiota [182, 34, 37]; in detail, *Basidiomycota* was increased in A_IBD UC flare samples, while *Ascomycota* (*Saccharomycetaceae*) was decreased in disease (see Appendix B). Several studies [193, 194, 195] detected a strong association between Crohn's disease and anti-*Saccharomyces cerevisiae* antibodies (ASCA) against yeast mannan; however the high prevalence of *Saccharomyces* in the fungal data can be due also to the ingestion of yeast-containing foods such as bread and beer. In addition, *Saccharomycetaceae* has been negatively correlated with total saturated fatty acid levels. Interestingly, among biomarkers discriminating CD from UC samples a slightly larger fungal component was found with respect to tasks discriminating healthy and IBD phenotypes (see Appendix B): *Debaryomyces* was increased in CD, *Saccharomyces* was increased in UCr with respect to CDr, *Malasseziales* were more abundant in UCf than in CDf.

To our knowledge, INF is the first method integrating bacteria and fungi interactions to build a panel of biomarkers predictive of IBD phenotypes. After a necessary biological validation, INF could be enter the clinical practice to track dysbiosis status before overt clinical disease, when healthy individuals can still benefit from a prevention strategy, or when the patient in remission can avoid further flare-ups.

6.3 Rett syndrome: RTT

I-PreMONet has been validated on a second dataset, with variables similar to A_IBD (bacterial and fungal microbiota abundances) but samples with different phenotype: Rett syndrome. The dataset was published by Strati and colleagues in 2016 [196].

This is a postnatal progressive neurological disorder, affecting almost exclusively females with an incidence of 1:10,000 live births [197]. For about 7 to 18 months after birth, infants with Rett syndrome develop normally, although before 6 months they can be affected by acquired microencephaly and low muscle tone. The disease is characterized by the loss of previously acquired skills (developmental regression), while, in some cases, development may continue but at a delayed rate. For example, an infant may learn to sit upright, but not to crawl. Affected children commonly develop autistic-like behaviours, such as panic attacks, teeth grinding (bruxism), purposeful hand movements and the ability to communicate, in addition to other abnormalities, including impaired control of voluntary movements (ataxia), forced expulsion of air and saliva, breathing problems and gastrointestinal (GI) issues. In particular, it has been shown that GI and nutritional dysfunctions occur frequently throughout life in Rett subjects [198, 199].

Approximately 90-95% of Rett syndrome cases are caused by random mutations (more than 200 identified) in X-linked methyl-CpG binding protein 2 gene (MECP2) on the X chromosome; the course and severity of Rett syndrome is determined by the location, type and severity of the MECP2 mutation. Recently, Wahba and colleagues showed that MeCP2 is expressed throughout the GI tract, exclusively within the enteric nervous system (ENS) of the GI tract, suggesting that GI dismotility observed in Rett may be mediated through ENS dysfunction secondary to MeCP2 mutation [200]. Interestingly, microbiota is known both to modulate central nervous system activities and to be stricly related to GI dysfunctions; thus, the relationship between gut microbiota dysbiosis and Rett is not a remote hypothesis [196]. As a matter of fact, several studies have suggested a disturbance in the gut microbiota as a potential contributor to pathogenesis of autism spectrum disorders (ASDs) [201, 202, 203].

6.3.1 Dataset

The dataset is publicly available [196] and includes bacterial and fungal gut microbiota abundances from healthy controls (HC) and Rett (RTT) syndrome subjects. In detail, stool samples were collected from 50 female subjects with RTT (average age 12±7.3) and 29 age-matched healthy subjects (average age 17±9.6). RTT individuals were genotyped for MECP2 and CDKL5 gene mutations, and both gastrointestinal symptoms (i.e. constipation) and intestinal inflammation were assessed. Only subjects that had not taken antibiotics, probiotics or prebiotics for at least three months prior to enrollment were included in the study. Whole fecal samples were collected, aliquoted as it us and stored at -80℃.
DNA extraction protocol and sequence data processing are reported in detail in [196], but briefly described in the following paragraphs.

Genomic DNA extraction from fecal samples. FastDNA™SPIN Kit for Feces (MP Biomedicals, Santa Ana, CA, USA) was used for total DNA extraction from fecal samples (200 mg, wet weight). A check of DNA integrity and quality was performed on 1% agarose gel and quantified with spectrophotometry.

Amplicon library preparation and pyrosequencing. Variable regions V3-V5 of the 16S rRNA gene and ITS1 were amplified in bacteria and fungi, respectively. Amplicon libraries were build using the FastStart High Fidelity PCR system (Roche, Basel, Switzerland) and the PCR products were pyrosequenced on the GS FLX+ system using the XL+ chemistry following the manufacturer's recommendations (Roche, Basel, Switzerland).

Taxonomic analysis

First, raw 454 files were demultiplexed using the Roche's GS Run Processor software, subsequently reads were pre-processed with the MICCA pipeline [28]. In detail, preprocessing included forward and reverse primer trimming and quality filtering, *de novo* sequence clustering, chimera filtering and taxonomy assignment. Sequences with a threshold of 97% pairwise identity were clustered into Operational taxonomic units (OTUs), and their representative sequences were classified using the RDP classifier version 2.7 on 16S rDNA data and using the RDP classifier version 2.8 [204] against the UNITE fungal ITS database [184] on ITS1 data. Bacterial sequences were aligned against the multiple alignment of Greengenes [136] (release 13_05) through Template-guided multiple sequence alignment (MSA); on the other hand, *de novo* MSA was performed by T-Coffee [205]. Fungal taxonomy assignments were also manually curated using BLASTn against the GenBank's database for accuracy.

Statistical analysis

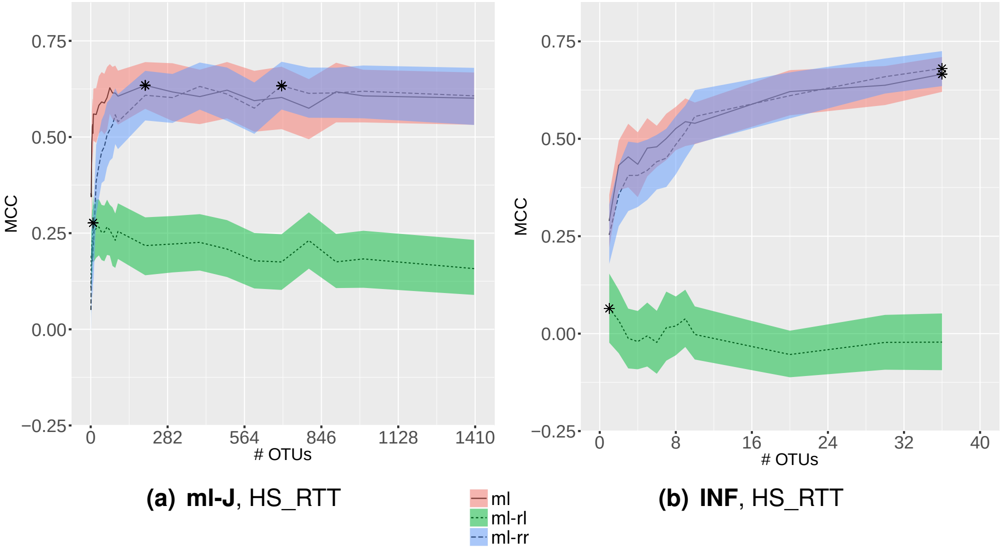
Our analyses started from OTU tables as provided by Strati and colleagues [196]. Tables with bacterial and fungal abundances were normalized in terms of the Trimmed Mean of M-values (TMM) normalization, implemented by the DESeq Bioconductor R package. OTUs with zero counts on all subjects were filtered out. Random Forest and Support Vector Machines were developed on 70% of the dataset (training partition) composed of bacterial and fungal genera from healthy individuals and Rett patients. The 30% of the datasets were used as blind validation set, testing the performance of trained models on unseen data. Integration of bacterial and fungal abundances, as well as the development of predictive models on integrated data were performed with I-PreMONet (see Sec. 5 and 6.2).

6.3.2 Results

Filtering low abundant OTUs led to a table with 1,155 bacterial and 251 fungal absolute abundances (rows as samples and columns as OTUs). Predictive performance of the three integrative techniques of INF module, ml-J, ml-rSNF and INF (see Sec. 6.2), are reported in terms of best average MCC on training set (MCC_{int}) with 95% Student bootstrap (1000×resampling) confidence intervals (MCC_{min} , MCC_{max}), number of features (Nf) leading to MCC_{int}. Differently from results in Sec. 6.2.3, MCC on validation set is not computed, since number of subjects did not allow a reliable classifier validation. As reported in Sec. 6.2.2, ml-J consisted in concatenating bacterial and fungal datasets, and training predictive models on the joint dataset; ml-rSNF developed predictive models on juxtaposed datasets, but biomarkers lists were built by using features weights from rSNF (see Sec. 4.2) and not from RF classifiers. Finally, classifiers were trained on the dataset restricted on the intersection of the biomarkers lists from ml-J and ml-rSNF: this approach was referred to as INF. In ml-J, RF built on 200 OTUs predicted Rett outcome with MCC=0.63 (C.I. $0.57-0.69$), performance slightly improved in ml-rSNF, that led to an MCC=0.65 (C.I. $0.60-0.69$) with 200 OTUs. Notably, for INF only 36 OTUs were needed to obtain an MCC=0.66 (C.I. $0.62-0.71$). MCC values along increasing sets of OTUs for ml-J and INF are reported in Figure 6.20, together with random labels and random ranking experiments performance.

Interestingly, as for the analysis of A_IBD dataset (see Sec. 6.2.3), INF obtained higher MCC_{int} than ml-J or ml-rSNF, but with a significant reduction in number of top discriminant OTUs. Thus, improvements introduced by INF over ml-J are confirmed to be independent from the biological problem considered.

OTUs selected as top discriminant biomarkers between healthy and Rett patients mostly belong to bacterial families *Ruminococcaceae* and *Lachnospiraceae*, as well as to fungal phyla *Basidiomycota* and *Ascomycota*.



(a) ml-J, HS_RTT

(b) INF, HS_RTT

Figure 6.20. MCC plots for the healthy vs. RTT classification tasks. Solid curves in black indicate MCC, with corresponding 95% bootstrap confidence intervals, at increasing OTUs set sizes (# OTUs). MCC_{int} highlighted with ($*$). Random labels (RL) experiments are indicated in green, random ranking (RR) experiments in blue (see text).

6.3.3 Discussion

Strati and colleagues discussed for the first time that RTT is characterized by a dysbiotic bacterial and fungal microbiota, thus literature lacks of reference studies to compare our results with. However, studies on ASD and microbiome can be considered as possible candidates, given the similarities between Rett and ASD phenotypes.

Biomarkers identified by our INF were in line with the ones found by Strati and colleagues [196]; in detail, the common biomarkers were 28 out of 30 bacteria and 4 out of 6 fungi. *Bifidobacterium* was found among the predictive biomarkers, increased in abundance in healthy individuals; several studies evidenced its health-promoting role [206] with potential probiotic properties [207].

Besides, one of the top discriminant biomarkers found by INF was *Bacteroides* genera, more abundant in healthy children than in Rett patients. This result is in accordance with Hsiao and colleagues [12], who studied microbiota alterations in the maternal immune activation (MIA) mouse model displaying features of ASD and investigated how these abnormalities impact ASD-related GI and/or behavioral alterations in MIA offspring. Interestingly, many of the adverse effects induced by MIA could be corrected by the *Bacteroides fragilis* treatment, lowering the concentration of potentially pathogenic metabolites and decreasing intestinal barrier permeability. Note that increased intestinal permeability is commonly associated with an altered immune response, given to leakage of gut-derived metabolites into the bloodstream [16, 208, 209]. Furthermore, several taxa belonging to *Clostridia* were selected by INF as predictive bacteria, according to other studies on ASD, showing an increased incidence of this bacteria [202, 210]. Yano and colleagues [211] highlighted that *Clostridia*, exerting high 7-dehydroxylation activity, increase the production of deoxycholate from cholate [212, 213], leading to an increase in the biosynthesis of colon serotonin (5-hydroxytryptamine [5-HT]). Mainly, the serotonin is a brain neurotransmitter, but it is also an important regulatory factor in the gastrointestinal (GI) tract and other organ systems. Thus, microbiota-mediated changes in colonic 5-HT promote GI motility and hemostasis in the host, by activation of enteric neurons.

Moreover, *Ruminococcaceae* and *Roseburia* genus, notable SCFAs producers [214], were identified by INF as the top two most discriminant biomarkers between healthy and Rett children. SCFAs could affect immune regulation and CNS function [215, 216], by translocating from intestinal mucosa to the systemic circulation. In detail, Erny and colleagues [215] found that SCFAs modulate maturation, morphology and function of microglia, the tissue resident macrophages of the brain and spinal cord, in germ-free mice. These results may support our finding of *Ruminococcaceae* and *Roseburia* as predictive biomarkers, given their contribution to microglia-mediated diseases of the CNS.

The results on RETT dataset confirmed that INF led to a compact list of highly predictive biomarkers not only on IBD clinical dataset, thus the extension of INF on different omics datasets is naturally supported.

Chapter 7

Conclusion

This thesis introduces two modeling frameworks (PreMONet and I-PreMONet) for the analysis of meta-omics data and their application on original clinical data of significant biomedical interest. The aim of these frameworks is to provide an integrated environment for predictive profiling and differential network analysis of meta-omics data, combined with a novel network-based approach to their integration. PreMONet and I-PreMONet rely on both existing and novel algorithms and Open Source software tools. In particular, I-PreMONet offers a first original algorithm (INF) combining a network fusion method with machine learning to solve the problem of heterogeneous integration of meta-omics data, obtaining shorter biomarker lists than those found by standard concatenation. Results on clinical metagenomics datasets demonstrate that our frameworks can provide an accurate characterization of microbiota composition and structural change of networks from microbiome communities associated to disease trajectories. Specifically, results highlighted the potential clinical role of PreMONet: it can model the specific interactions between species within microbiota that are associated with severe inflammation and the levels of calprotectin in chronic gut inflammation. Thus, PreMONet can be used as a possible enhancement of the calprotectin-based risk stratification tools, which is currently the most popular diagnostic approach. In addition, validation on clinical datasets demonstrated that I-PreMONet can integrate meta-omics datasets measured on the same set of samples and classify them according to known phenotypic traits; besides, I-PreMONet provides the identification of a small but robust multi-omics signature that can predict phenotypic groups in new test samples. Notably, to our knowledge, I-PreMONet is the first framework identifying inter-kingdom predictive biomarkers of Inflammatory Bowel Disease, by network-based integration of fungal and bacterial datasets.

It is also important to stress that clinical practice could take advantage of biomarkers identified by our frameworks. Indeed, clinical information is commonly sufficient to initiate a therapy, but biomarkers can assist the physicians to tailor subsequent therapy and decide on its duration. Moreover, meta-omics biomarkers can help tracking dysbiosis status before overt clinical disease, so that healthy individuals can benefit from a prevention strategy, or the patient in remission can avoid further flare-ups.

In addition to the possible biological and clinical impacts of both PreMONet and I-PreMONet, the modularity and open-source licensing can carry other several advantages. To researchers without bioinformatics expertise, our frameworks provide an accurate and reproducible solution for meta-omics analysis; to bioinformaticians, PreMONet and I-PreMONet provide a set of tools that can be adapted to meet specific analysis needs. Finally, the flexibility of our pipelines should encourage researchers to contribute their own analysis modules into a reproducible framework. The availability of our frameworks should reduce duplication of efforts and quicken developments in this field by encouraging scientists to focus on individual components without re-implementing all the modules of a meta-omics pipeline.

As a further development, the integrative technique implemented in I-PreMONet could be improved in order to handle data measurements with some samples missing a percentage of meta-omics features. Basically, the framework should enable integrative models to ‘reconstruct’ the missing information on some samples, based on the integrated knowledge acquired on other samples.

Bibliography

[1] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59–65, 2010.

[2] Human Microbiome Project Consortium et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[3] Xochitl C Morgan and Curtis Huttenhower. Human microbiome analysis. *PLoS Comput Biol*, 8(12):e1002808, 2012.

[4] Scott Christley, Chase Cockrell, and Gary An. Computational studies of the intestinal host-microbiota interactome. *Computation*, 3(1):2–28, 2015.

[5] Zhigang Zhang, Jiawei Geng, Xiaodan Tang, Hong Fan, Jinchao Xu, Xiujun Wen, Zhanshan Sam Ma, and Peng Shi. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME journal*, 8(4):881–893, 2014.

[6] Maka Mshvildadze, Josef Neu, Jonathan Shuster, Douglas Theriaque, Nan Li, and Volker Mai. Intestinal microbial ecology in premature infants assessed

with non–culture-based techniques. *The Journal of pediatrics*, 156(1):20–25, 2010.

[7] Fredrik Bäckhed, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon. Host-bacterial mutualism in the human intestine. *science*, 307(5717):1915–1920, 2005.

[8] Michael A Fischbach and Justin L Sonnenburg. Eating for two: how metabolism establishes interspecies interactions in the gut. *Cell host & microbe*, 10(4):336–347, 2011.

[9] AW Walker, J Ince, SH Duncan, LM Webster, G Holtrop, X Ze, D Brown, and MD Stares. 535 scott p, bergerat a, louis p, mcintosh f, johnstone am, lobley ge, parkhill j, flint 536 hj. 2011. dominant and diet-responsive groups of bacteria within the human colonic 537 microbiota. *ISME J*, 5:220–230.

[10] Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

[11] Jason Lloyd-Price, Galeb Abu-Ali, and Curtis Huttenhower. The healthy human microbiome. *Genome medicine*, 8(1):1, 2016.

[12] Elaine Y Hsiao, Sara W McBride, Sophia Hsien, Gil Sharon, Embriette R Hyde, Tyler McCue, Julian A Codelli, Janet Chow, Sarah E Reisman, Joseph F Petrosino, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, 155(7):1451–1463, 2013.

[13] Charisse Petersen and June L Round. Defining dysbiosis and its influence on host immunity and disease. *Cellular microbiology*, 16(7):1024–1033, 2014.

[14] Aurélien Trompette, Eva S Gollwitzer, Koshika Yadava, Anke K Sichelstiel, Norbert Sprenger, Catherine Ngom-Bru, Carine Blanchard, Tobias Junt, Laurent P Nicod, Nicola L Harris, et al. Gut microbiota metabolism of dietary fiber influences allergic airway disease and hematopoiesis. *Nature medicine*, 20(2):159–166, 2014.

[15] Åsa V Keita and Johan D Söderholm. Barrier dysfunction and bacterial uptake in the follicle-associated epithelium of ileal crohn's disease. *Annals of the New York Academy of Sciences*, 1258(1):125–134, 2012.

[16] Jerrold R Turner. Intestinal mucosal barrier function in health and disease. *Nature Reviews Immunology*, 9(11):799–809, 2009.

[17] Jerry M Wells, Oriana Rossi, Marjolein Meijerink, and Peter van Baarlen. Epithelial crosstalk at the microbiota–mucosal interface. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4607–4614, 2011.

[18] J Berkes, VK Viswanathan, SD Savkovic, and G Hecht. Intestinal epithelial responses to enteric pathogens: effects on the tight junction barrier, ion transport, and inflammation. *Gut*, 52(3):439–451, 2003.

[19] Stephanie Hummel, Katharina Veltman, Christoph Cichon, Ulrich Sonnenborn, and M Alexander Schmidt. Differential targeting of the e-cadherin/β-catenin complex by gram-positive probiotic lactobacilli improves epithelial barrier function. *Applied and environmental microbiology*, 78(4):1140–1147, 2012.

[20] Jane MM Natividad, Valerie Petit, Xianxi Huang, Giada de Palma, Jennifer Jury, Yolanda Sanz, Dana Philpott, Clara L Garcia Rodenas, Kathy D McCoy, and Elena F Verdu. Commensal and probiotic bacteria influence intestinal

barrier function and susceptibility to colitis in nod1-/-; nod2-/- mice. *Inflammatory bowel diseases*, 18(8):1434–1446, 2012.

[21] Vanni Bucci and Joao B Xavier. Towards predictive models of the human gut microbiome. *Journal of molecular biology*, 426(23):3907–3916, 2014.

[22] Eric A Franzosa, Tiffany Hsu, Alexandra Sirota-Madi, Afrah Shafquat, Galeb Abu-Ali, Xochitl C Morgan, and Curtis Huttenhower. Sequencing and beyond: integrating molecular'omics' for microbial community profiling. *Nature Reviews Microbiology*, 13(6):360–372, 2015.

[23] Mireia Valles-Colomer, Youssef Darzi, Sara Vieira-Silva, Gwen Falony, Jeroen Raes, and Marie Joossens. Meta-omics in inflammatory bowel disease research: applications, challenges, and guidelines. *Journal of Crohn's and Colitis*, page jjw024, 2016.

[24] Youssef Darzi, Gwen Falony, Sara Vieira-Silva, and Jeroen Raes. Towards biome-specific analysis of meta-omics data. *The ISME journal*, 2015.

[25] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

[26] Robert C Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature methods*, 10(10):996–998, 2013.

[27] Folker Meyer, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics rast server–a public resource for the auto-

matic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):1, 2008.

[28] Davide Albanese, Paolo Fontana, Carlotta De Filippo, Duccio Cavalieri, and Claudio Donati. Micca: a complete and accurate software for taxonomic profiling of metagenomic data. *Scientific reports*, 5:9743, 2015.

[29] Evguenia Kopylova, Laurent Noé, and Hélène Touzet. Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217, 2012.

[30] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nature biotechnology*, 29(7):644, 2011.

[31] Thilo Muth, Alexander Behne, Robert Heyer, Fabian Kohrs, Dirk Benndorf, Marcus Hoffmann, Miro Lehteva, Udo Reichl, Lennart Martens, and Erdmann Rapp. The metaproteomeanalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *Journal of proteome research*, 14(3):1557–1565, 2015.

[32] John S Cottrell and U London. Probability-based protein identification by searching sequence databases using mass spectrometry data. *electrophoresis*, 20(18):3551–3567, 1999.

[33] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804, 2007.

[34] Donal Sheehan, Carthage Moran, and Fergus Shanahan. The microbiota in inflammatory bowel disease. *Journal of gastroenterology*, 50(5):495–507, 2015.

[35] Mikael Knip and Heli Siljander. The role of the intestinal microbiota in type 1 diabetes mellitus. *Nature Reviews Endocrinology*, 2016.

[36] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[37] Chunguang Luan, Lingling Xie, Xi Yang, Huifang Miao, Na Lv, Ruifen Zhang, Xue Xiao, Yongfei Hu, Yulan Liu, Na Wu, et al. Dysbiosis of fungal microbiota in the intestinal mucosa of patients with colorectal adenomas. *Scientific reports*, 5, 2015.

[38] Wendy S Garrett. Cancer and the microbiota. *Science*, 348(6230):80–86, 2015.

[39] Nicola Segata, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6):1, 2011.

[40] Jun Yu, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao yi Liang, Youwen Qin, Longqing Tang, Hui Zhao, Jan Stenvang, Yanli Li, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, pages gutjnl–2015, 2015.

[41] RH Perlis. Translating biomarkers to clinical practice. *Molecular psychiatry*, 16(11):1076–1087, 2011.

[42] Dan Knights, Laura Wegener Parfrey, Jesse Zaneveld, Catherine Lozupone, and Rob Knight. Human-associated microbial signatures: examining their predictive value. *Cell host & microbe*, 10(4):292–296, 2011.

[43] Julianne R Brown, Sofia Morfopoulou, Jonathan Hubb, Warren A Emmett, Winnie Ip, Divya Shah, Tony Brooks, Simon ML Paine, Glenn Anderson, Alex Virasami, et al. Astrovirus va1/hmo-c: an increasingly recognized neurotropic pathogen in immunocompromised patients. *Clinical Infectious Diseases*, page ciu940, 2015.

[44] Michael R Wilson, Samia N Naccache, Erik Samayoa, Mark Biagtan, Hiba Bashir, Guixia Yu, Shahriar M Salamat, Sneha Somasekar, Scot Federman, Steve Miller, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *New England Journal of Medicine*, 370(25):2408–2417, 2014.

[45] Samia N Naccache, Karl S Peggs, Frank M Mattes, Rahul Phadke, Jeremy A Garson, Paul Grant, Erik Samayoa, Scot Federman, Steve Miller, Michael P Lunn, et al. Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. *Clinical Infectious Diseases*, 60(6):919–923, 2015.

[46] Bernd Hoffmann, Dennis Tappe, Dirk Höper, Christiane Herden, Annemarie Boldt, Christian Mawrin, Olaf Niederstraßer, Tobias Müller, Maria Jenckel, Elisabeth van der Grinten, et al. A variegated squirrel bornavirus associated with fatal human encephalitis. *New England Journal of Medicine*, 373(2):154–162, 2015.

[47] Hayssam Soueidan and Macha Nikolski. Machine learning for metagenomics: methods and tools. *Metagenomics*, 1(1):1–19, 2015.

[48] Jack A Gilbert, Robert A Quinn, Justine Debelius, Zhenjiang Z Xu, James Morton, Neha Garg, Janet K Jansson, Pieter C Dorrestein, and Rob Knight. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94–103, 2016.

[49] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

[50] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[52] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[53] MAQC Consortium et al. The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8):827–838, 2010.

[54] Wenqian Zhang, Ying Yu, Falk Hertwig, Jean Thierry-Mieg, Wenwei Zhang, Danielle Thierry-Mieg, Jian Wang, Cesare Furlanello, Viswanath Devanarayan, Jie Cheng, et al. Comparison of rna-seq and microarray-based models for clinical endpoint prediction. *Genome biology*, 16(1):1, 2015.

[55] Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.

[56] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[57] Giuseppe Jurman, Samantha Riccadonna, and Cesare Furlanello. A comparison of mcc and cen error measures in multi-class prediction. *PloS one*, 7(8):e41882, 2012.

[58] Davide Albanese, Roberto Visintainer, Stefano Merler, Samantha Riccadonna, Giuseppe Jurman, and Cesare Furlanello. mlpy: Machine learning python, 2012.

[59] Giuseppe Jurman, Stefano Merler, Annalisa Barla, Silvano Paoli, Antonio Galea, and Cesare Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, 24(2):258–264, 2008.

[60] Richard R Stein, Vanni Bucci, Nora C Toussaint, Charlie G Buffie, Gunnar Rätsch, Eric G Pamer, Chris Sander, and João B Xavier. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol*, 9(12):e1003388, 2013.

[61] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012.

[62] Vanni Bucci, Serena Bradde, Giulio Biroli, and Joao B Xavier. Social interaction, noise and antibiotic-mediated switches in the intestinal microbiota. *PLoS Comput Biol*, 8(4):e1002497, 2012.

[63] Vanni Bucci, Carey D Nadell, and Joao B Xavier. The evolution of bacteriocin production in bacterial biofilms. *The American Naturalist*, 178(6):E162–E173, 2011.

[64] Arya Khosravi and Sarkis K Mazmanian. Disruption of the gut microbiome as a risk factor for microbial infections. *Current opinion in microbiology*, 16(2):221–227, 2013.

[65] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*, 8(7):e1002606, 2012.

[66] Philippe J Sansonetti. War and peace at mucosal surfaces. *Nature Reviews Immunology*, 4(12):953–964, 2004.

[67] Egija Zaura, Bart JF Keijser, Susan M Huse, and Wim Crielaard. Defining the healthy" core microbiome" of oral microbial communities. *BMC microbiology*, 9(1):1, 2009.

[68] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, et al. Enterotypes of the human gut microbiome. *nature*, 473(7346):174–180, 2011.

[69] Sarkis K Mazmanian, June L Round, and Dennis L Kasper. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature*, 453(7195):620–625, 2008.

[70] Daniel N Frank, Wei Zhu, R Balfour Sartor, and Ellen Li. Investigating the biological and clinical significance of human dysbioses. *Trends in microbiology*, 19(9):427–434, 2011.

[71] Chenhao Li, Kun Ming Kenneth Lim, Kern Rei Chng, and Niranjan Nagarajan. Predicting microbial interactions through computational approaches. *Methods*, 102:12–19, 2016.

[72] Charlie G Buffie, Vanni Bucci, Richard R Stein, Peter T McKenney, Lilan Ling, Asia Gobourne, Daniel No, Hui Liu, Melissa Kinnebrew, Agnes Viale, et al. Precision microbiome reconstitution restores bile acid mediated resistance to clostridium difficile. *Nature*, 517(7533):205–208, 2015.

[73] Dan W Thomas, Frank R Greer, et al. Probiotics and prebiotics in pediatrics. *Pediatrics*, 126(6):1217–1231, 2010.

[74] Justine L Murray, Jodi L Connell, Apollo Stacy, Keith H Turner, and Marvin Whiteley. Mechanisms of synergy in polymicrobial infections. *Journal of Microbiology*, 52(3):188–199, 2014.

[75] Trevor Dalton, Scot E Dowd, Randall D Wolcott, Yan Sun, Chase Watters, John A Griswold, and Kendra P Rumbaugh. An in vivo polymicrobial biofilm wound infection model to study interspecies interactions. *PloS one*, 6(11):e27317, 2011.

[76] Paul E Kolenbrander and Jack London. Adhere today, here tomorrow: oral bacterial adherence. *Journal of bacteriology*, 175(11):3247, 1993.

[77] Björn Vessman, Philip Gerlee, and Torbjörn Lundh. Estimating the probability of coexistence in cross-feeding communities. *arXiv preprint arXiv:1512.02460*, 2015.

[78] Frédéric Moens, Stefan Weckx, and Luc De Vuyst. Bifidobacterial inulin-type fructan degradation capacity determines cross-feeding interactions between bifidobacteria and faecalibacterium prausnitzii. *International journal of food microbiology*, 231:76–85, 2016.

[79] Mark R Charbonneau, Laura V Blanton, Daniel B DiGiulio, David A Relman, Carlito B Lebrilla, David A Mills, and Jeffrey I Gordon. A microbial perspective of human developmental biology. *Nature*, 535(7610):48–55, 2016.

[80] A Rodrigues Hoffmann, LM Proctor, MG Surette, and JS Suchodolski. The microbiome the trillions of microorganisms that maintain health and cause disease in humans and companion animals. *Veterinary pathology*, page 0300985815595517, 2015.

[81] Mehdi Layeghifard, David M Hwang, and David S Guttman. Disentangling interactions in the microbiome: A network perspective. *Trends in Microbiology*, 2016.

[82] Tue H Hansen, Rikke J Gøbel, Torben Hansen, and Oluf Pedersen. The gut microbiome in cardio-metabolic health. *Genome medicine*, 7(1):1, 2015.

[83] Ahmad A Zeidan, Peter Rådström, and Ed WJ van Niel. Stable coexistence of two caldicellulosiruptor species in a de novo constructed hydrogen-producing co-culture. *Microbial cell factories*, 9(1):1, 2010.

[84] William Harcombe. Novel cooperation experimentally evolved between species. *Evolution*, 64(7):2166–2172, 2010.

[85] Shiri Freilich, Raphy Zarecki, Omer Eilam, Ella Shtifman Segal, Christopher S Henry, Martin Kupiec, Uri Gophna, Roded Sharan, and Eytan Ruppin. Competitive and cooperative metabolic interactions in bacterial communities. *Nature communications*, 2:589, 2011.

[86] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[87] Noriko Maruyama, Fumito Maruyama, Yasuo Takeuchi, Chihiro Aikawa, Yuichi Izumi, and Ichiro Nakagawa. Intraindividual variation in core microbiota in peri-implantitis and periodontitis. *Scientific reports*, 4:6602, 2014.

[88] Feng Ju, Yu Xia, Feng Guo, Zhiping Wang, and Tong Zhang. Taxonomic relatedness shapes bacterial assembly in activated sludge of globally distributed wastewater treatment plants. *Environmental microbiology*, 16(8):2421–2432, 2014.

[89] David Berry and Stefanie Widder. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in microbiology*, 5:219, 2014.

[90] Eric Z Chen and Hongzhe Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, page btw308, 2016.

[91] John Aitchison. A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13(2):175–189, 1981.

[92] John Aitchison. The statistical analysis of compositional data. 1986.

[93] Huaying Fang, Chengcheng Huang, Hongyu Zhao, and Minghua Deng. Cclasso: correlation inference for compositional data through lasso. *Bioinformatics*, page btv349, 2015.

[94] Andrea Gobbi and Giuseppe Jurman. A null model for pearson coexpression networks. *PloS one*, 10(6):e0128115, 2015.

[95] Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[96] Philip R Bevington. Data reduction and error analysis for the physical sciences mcgraw hill book co. *New York*, 1969.

[97] Neural Nets WIRN10 B Apolloni et al. An introduction to spectral distances in networks. In *Neural Nets WIRN10: Proceedings of the 20th Italian Workshop on Neural Nets*, volume 226, page 227. IOS Press, 2011.

[98] Giuseppe Jurman, Roberto Visintainer, Michele Filosi, Samantha Riccadonna, and Cesare Furlanello. The him glocal metric and kernel for network comparison and classification. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.

[99] Richard W Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[100] Koji Iwayama, Yoshito Hirata, Kohske Takahashi, Katsumi Watanabe, Kazuyuki Aihara, and Hideyuki Suzuki. Characterizing global evolutions of complex systems via intermediate network representations. *Scientific reports*, 2, 2012.

[101] Martina Morris, Mark S Handcock, and David R Hunter. Specification of exponential-family random graph models: terms and computational aspects. *Journal of statistical software*, 24(4):1548, 2008.

[102] Mads Ipsen and Alexander S Mikhailov. Evolutionary reconstruction of networks. *Physical Review E*, 66(4):046109, 2002.

[103] Alessandro Zandoná, Marco Chierici, Giuseppe Jurman, Cesare Furlanello, Salvatore Cucchiara, Federica Del Chierico, and Lorenza Putignani.

[104] Giuseppe Jurman, Michele Filosi, Roberto Visintainer, Samantha Riccadonna, and Cesare Furlanello. Stability indicators in network reconstruction. *arXiv preprint arXiv:1209.1654*, 2012.

[105] Marco Mina, Renata Boldrini, Arianna Citti, Paolo Romania, Valerio D'Alicandro, Maretta De Ioris, Aurora Castellano, Cesare Furlanello, Franco Locatelli, and Doriana Fruci. Tumor-infiltrating t lymphocytes improve clinical outcome of therapy-resistant neuroblastoma. *Oncoimmunology*, 4(9):e1019981, 2015.

[106] Giuseppe Jurman. Metric projection for dynamic multiplex networks. *arXiv preprint arXiv:1601.01940*, 2016.

[107] Santo Fortunato and Claudio Castellano. Community structure in graphs. In *Computational Complexity*, pages 490–512. Springer, 2012.

[108] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[109] Steven N Baldassano and Danielle S Bassett. Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease. *Scientific reports*, 6, 2016.

[110] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218, 2006.

[111] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[112] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[113] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[114] Fang Wu and Bernardo A Huberman. Finding communities in linear time: a physics approach. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):331–338, 2004.

[115] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[116] Mark EJ Newman. Analysis of weighted networks. *Physical review E*, 70(5):056131, 2004.

[117] Luigi Nibali and Brian Henderson. *The Human Microbiota and Chronic Disease: Dysbiosis as a Cause of Human Pathology*. John Wiley & Sons, 2016.

[118] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

[119] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

[120] J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 290. Citeseer, 1976.

[121] Y.C. Wei and C.K. Cheng. Towards efficient hierarchical designs by ratio cut partitioning. In *International Conference Computer-Aided Design*, pages 298–301, 1989.

[122] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2004.

[123] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[124] N.X. Vinh, J. Epps, and Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal Machine Learning Research*, 11:2837–2854, December 2014.

[125] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, 2015.

[126] Lieven PC Verbeke, Jimmy Van den Eynden, Ana Carolina Fierro, Piet Demeester, Jan Fostier, and Kathleen Marchal. Pathway relevance ranking for tumor samples through network-based data integration. *PloS one*, 10(7):e0133503, 2015.

[127] Yong Zhang, Xiaohua Hu, and Xingpeng Jiang. Multi-view clustering of microbiome samples by robust similarity network fusion and spectral clustering. 2015.

[128] Xingpeng Jiang and Xiaohua Hu. Microbiome data mining for microbial interactions and relationships. In *Big Data Analytics*, pages 221–235. Springer, 2016.

[129] Francisco Azuaje. Computational models for predicting drug responses in cancer research. *Briefings in Bioinformatics*, page bbw065, 2016.

[130] Song He, Haochen He, Wenjian Xu, Xin Huang, Shuai Jiang, Fei Li, Fuchu He, and Xiaochen Bo. Icm: a web server for integrated clustering of multidimensional biomedical data. *Nucleic acids research*, page gkw378, 2016.

[131] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[132] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

[133] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[134] Andreas Wilke, Travis Harrison, Jared Wilkening, Dawn Field, Elizabeth M Glass, Nikos Kyrpides, Konstantinos Mavrommatis, and Folker Meyer. The m5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC bioinformatics*, 13(1):141, 2012.

[135] Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, 35(21):7188–7196, 2007.

[136] Todd Z DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7):5069–5072, 2006.

[137] James R Cole, Benli Chai, Terry L. Marsh, Ryan J Farris, Qiong Wang, SA Kulam, S Chandra, Donna M McGarrell, Thomas M. Schmidt, George M. Garrity, et al. The ribosomal database project (rdp-ii): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic acids research*, 31(1):442–443, 2003.

[138] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[139] J Gregory Caporaso, Kyle Bittinger, Frederic D Bushman, Todd Z DeSantis, Gary L Andersen, and Rob Knight. Pynast: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26(2):266–267, 2010.

[140] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[141] Sandra Macfarlane, Helen Steed, and George T Macfarlane. Intestinal bacteria and inflammatory bowel disease. *Critical reviews in clinical laboratory sciences*, 46(1):25–54, 2009.

[142] B.B. Crohn, L. Ginzburg, and G. Oppenheimer. Regional Ileitis a Pathologic and Clinical Entity. *Journal of American Medical Association*, 99(16):1323–1329, 1932.

[143] Jürgen Stein, Franz Hartmann, and Axel U Dignass. Diagnosis and management of iron deficiency anemia in patients with ibd. *Nature Reviews Gastroenterology and Hepatology*, 7(11):599–610, 2010.

[144] Yan-mei WANG, Mei-qing KANG, Yan-bing CUI, Zhi-xia XING, and Rui-fang WANG. Clinical study of internal and external treatment of gegen qinlian wu-

tan decoction in active ulcerative colitis [j]. *Chinese Journal of Experimental Traditional Medical Formulae*, 17:079, 2012.

[145] Nabeetha A Nagalingam and Susan V Lynch. Role of the microbiota in inflammatory bowel diseases. *Inflammatory bowel diseases*, 18(5):968–984, 2012.

[146] Andrew W DuPont and Herbert L DuPont. The intestinal microbiota and chronic disorders of the gut. *Nature Reviews Gastroenterology and Hepatology*, 8(9):523–531, 2011.

[147] RJ Xavier and DK Podolsky. Unravelling the pathogenesis of inflammatory bowel disease. *Nature*, 448(7152):427–434, 2007.

[148] Dirk Gevers, Subra Kugathasan, Lee A Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, et al. The treatment-naive microbiome in new-onset crohn's disease. *Cell host & microbe*, 15(3):382–392, 2014.

[149] R Balfour Sartor. Therapeutic manipulation of the enteric microflora in inflammatory bowel diseases: antibiotics, probiotics, and prebiotics. *Gastroenterology*, 126(6):1620–1633, 2004.

[150] Andre Franke, Dermot PB McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–1125, 2010.

[151] Altin Gjymishka, Roxana M Coman, Todd M Brusko, and Sarah C Glover. Influence of host immunoregulatory genes, er stress and gut microbiota on

the shared pathogenesis of inflammatory bowel disease and type 1 diabetes. *Immunotherapy*, 5(12):1357–1366, 2013.

[152] Silvia Caballero and Eric G Pamer. Microbiota-mediated inflammation and antimicrobial defense in the intestine. *Annual review of immunology*, 33:227, 2015.

[153] Danilo Ercolini, Francesca De Filippis, Antonietta La Storia, and Michele Iacono. “remake” by high-throughput sequencing of the microbiota involved in the production of water buffalo mozzarella cheese. *Applied and environmental microbiology*, 78(22):8142–8145, 2012.

[154] Michele Filosi, Shamar Droghetti, Ernesto Arbitrio, Roberto Visintainer, Samantha Riccadonna, Giuseppe Jurman, and Cesare Furlanello. Renette: a web-infrastructure for reproducible network analysis. *bioRxiv*, page 008433, 2014.

[155] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[156] Craig Mowat, Jayne Digby, Judith A Strachan, Robyn Wilson, Francis A Carey, Callum G Fraser, and Robert JC Steele. Faecal haemoglobin and faecal calprotectin as indicators of bowel disease in patients presenting to primary care with bowel symptoms. *Gut*, pages gutjnl–2015, 2015.

[157] Taina Sipponen. Diagnostics and prognostics of inflammatory bowel disease with fecal neutrophil-derived biomarkers calprotectin and lactoferrin. *Digestive diseases*, 31(3-4):336–344, 2013.

[158] Eliseo Papa, Michael Docktor, Christopher Smillie, Sarah Weber, Sarah P Preheim, Dirk Gevers, Georgia Giannoukos, Dawn Ciulla, Diana Tabbaa, Jay Ingram, et al. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PloS one*, 7(6):e39242, 2012.

[159] A Keshavarzian, A Banan, A Farhadi, S Komanduri, E Mutlu, Y Zhang, and JZ Fields. Increases in free radicals and cytoskeletal protein oxidation and nitration in the colon of patients with inflammatory bowel disease. *Gut*, 52(5):720–728, 2003.

[160] Christopher Sherrill and Robert C Fahey. Import and metabolism of glutathione bystreptococcus mutans. *Journal of bacteriology*, 180(6):1454–1459, 1998.

[161] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology*, 13(9):1, 2012.

[162] Marta Wlodarska, Aleksandar D Kostic, and Ramnik J Xavier. An integrative view of microbiome-host interactions in inflammatory bowel diseases. *Cell host & microbe*, 17(5):577–591, 2015.

[163] Lionel Rigottier-Gois. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *The ISME journal*, 7(7):1256–1261, 2013.

[164] Nadeem O Kaakoush. Insights into the role of erysipelotrichaceae in the human host. *Frontiers in cellular and infection microbiology*, 5, 2015.

[165] Noah W Palm, Marcel R De Zoete, Thomas W Cullen, Natasha A Barry, Jonathan Stefanowski, Liming Hao, Patrick H Degnan, Jianzhong Hu, Inga

Peter, Wei Zhang, et al. Immunoglobulin a coating identifies colitogenic bacteria in inflammatory bowel disease. *Cell*, 158(5):1000–1010, 2014.

[166] Duy M Dinh, Gretchen E Volpe, Chad Duffalo, Seema Bhalchandra, Albert K Tai, Anne V Kane, Christine A Wanke, and Honorine D Ward. Intestinal microbiota, microbial translocation, and systemic inflammation in chronic hiv infection. *Journal of Infectious Diseases*, page jiu409, 2014.

[167] Clara Moon, Megan T Baldridge, Meghan A Wallace, Carey-Ann D Burnham, Herbert W Virgin, and Thaddeus S Stappenbeck. Vertically transmitted faecal iga levels determine extra-chromosomal phenotypic variation. *Nature*, 521(7550):90–93, 2015.

[168] Rosario Lucas López, María José Grande Burgos, Antonio Gálvez, and Rubén Pérez Pulido. The human gastrointestinal tract and oral microbiota in inflammatory bowel disease: a state of the science review. *APMIS*, 2016.

[169] Yukihiro Furusawa, Yuuki Obata, Shinji Fukuda, Takaho A Endo, Gaku Nakato, Daisuke Takahashi, Yumiko Nakanishi, Chikako Uetake, Keiko Kato, Tamotsu Kato, et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory t cells. *Nature*, 504(7480):446–450, 2013.

[170] Kendle M Maslowski, Angelica T Vieira, Aylwin Ng, Jan Kranich, Frederic Sierro, Di Yu, Heidi C Schilter, Michael S Rolph, Fabienne Mackay, David Artis, et al. Regulation of inflammatory responses by gut microbiota and chemoattractant receptor gpr43. *Nature*, 461(7268):1282–1286, 2009.

[171] Dan Knights, Mark S Silverberg, Rinse K Weersma, Dirk Gevers, Gerard Dijkstra, Hailiang Huang, Andrea D Tyler, Suzanne van Sommeren, Floris Imhann, Joanne M Stempak, et al. Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome medicine*, 6(12):1, 2014.

[172] Yasunori Ogura, Denise K Bonen, Naohiro Inohara, Dan L Nicolae, Felicia F Chen, Richard Ramos, Heidi Britton, Thomas Moran, Reda Karaliuskas, Richard H Duerr, et al. A frameshift mutation in nod2 associated with susceptibility to crohn's disease. *Nature*, 411(6837):603–606, 2001.

[173] Jean-Pierre Hugot, Mathias Chamaillard, Habib Zouali, Suzanne Lesage, Jean-Pierre Cézard, Jacques Belaiche, Sven Almer, Curt Tysk, Colm A O'Morain, Miquel Gassull, et al. Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease. *Nature*, 411(6837):599–603, 2001.

[174] Reetta Satokari, Susana Fuentes, Eero Mattila, Jonna Jalanka, Willem M de Vos, and Perttu Arkkila. Fecal transplantation treatment of antibiotic-induced, noninfectious colitis and long-term microbiota follow-up. *Case reports in medicine*, 2014, 2014.

[175] Wendy S Garrett, Carey A Gallini, Tanya Yatsunenko, Monia Michaud, Andrea DuBois, Mary L Delaney, Shivesh Punit, Maria Karlsson, Lynn Bry, Jonathan N Glickman, et al. Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell host & microbe*, 8(3):292–300, 2010.

[176] Gerard Eberl. A new vision of immunity: homeostasis of the superorganism. *Mucosal immunology*, 3(5):450–460, 2010.

[177] Ana E Pérez-Cobas, Alejandro Artacho, Stephan J Ott, Andrés Moya, María J Gosalbes, and Amparo Latorre. Structural and functional changes in the gut microbiota associated to clostridium difficile infection. *Frontiers in microbiology*, 5:335, 2014.

[178] Judith Kelsen and Robert N Baldassano. Inflammatory bowel disease: the difference between children and adults. *Inflammatory bowel diseases*, 14(S2):S9–S11, 2008.

[179] Christian Jakobsen, Jiri Bartek, V Wewer, I Vind, P Munkholm, Randi Grøn, and A Paerregaard. Differences in phenotype and disease course in adult and paediatric inflammatory bowel disease–a population-based study. *Alimentary pharmacology & therapeutics*, 34(10):1217–1224, 2011.

[180] Harry Sokol, Valentin Leducq, Hugues Aschard, Hang-Phuong Pham, Sarah Jegou, Cecilia Landman, David Cohen, Giuseppina Liguori, Anne Bourrier, Isabelle Nion-Larmurier, et al. Fungal microbiota dysbiosis in ibd. *Gut*, pages gutjnl–2015, 2016.

[181] I Mukhopadhya, R Hansen, C Meharg, JM Thomson, RK Russell, SH Berry, EM El-Omar, and GL Hold. The fungal microbiota of de-novo paediatric inflammatory bowel disease. *Microbes and Infection*, 17(4):304–310, 2015.

[182] Qiurong Li, Chenyang Wang, Chun Tang, Qin He, Ning Li, and Jieshou Li. Dysbiosis of gut fungal microbiota is associated with mucosal inflammation in crohn's disease. *Journal of clinical gastroenterology*, 48(6):513–523, 2014.

[183] Daniel McDonald, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610–618, 2012.

[184] Urmas Kõljalg, R Henrik Nilsson, Kessy Abarenkov, Leho Tedersoo, Andy FS Taylor, Mohammad Bahram, Scott T Bates, Thomas D Bruns, Johan Bengtsson-Palme, Tony M Callaghan, et al. Towards a unified paradigm

for sequence-based identification of fungi. *Molecular ecology*, 22(21):5271–5277, 2013.

[185] Kelly A Shaw, Madeline Bertha, Tatyana Hofmekler, Pankaj Chopra, Tommi Vatanen, Abhiram Srivatsa, Jarod Prince, Archana Kumar, Cary Sauer, Michael E Zwick, et al. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Medicine*, 8(1):75, 2016.

[186] Seungha Kang, Stuart E Denman, Mark Morrison, Zhongtang Yu, Joel Dore, Marion Leclerc, and Chris S McSweeney. Dysbiosis of fecal microbiota in crohn's disease patients as revealed by a custom phylogenetic microarray. *Inflammatory bowel diseases*, 16(12):2034–2042, 2010.

[187] Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G Bermúdez-Humarán, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of crohn disease patients. *Proceedings of the National Academy of Sciences*, 105(43):16731–16736, 2008.

[188] Harry J Flint, Edward A Bayer, Marco T Rincon, Raphael Lamed, and Bryan A White. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews Microbiology*, 6(2):121–131, 2008.

[189] Oriana Rossi, Lisette A Van Berkel, Florian Chain, M Tanweer Khan, Nico Taverne, Harry Sokol, Sylvia H Duncan, Harry J Flint, Hermie JM Harmsen, Philippe Langella, et al. Faecalibacterium prausnitzii a2-165 has a high capacity to induce il-10 in human and murine dendritic cells and modulates t cell responses. *Scientific reports*, 6, 2016.

[190] Toshifumi Ohkusa, Nobuhiro Sato, Tatuo Ogihara, Koji Morita, Masayuki Ogawa, and Isao Okayasu. Fusobacterium varium localized in the colonic mucosa of patients with ulcerative colitis stimulates species-specific antibody. *Journal of gastroenterology and hepatology*, 17(8):849–853, 2002.

[191] Mara Roxana Rubinstein, Xiaowei Wang, Wendy Liu, Yujun Hao, Guifang Cai, and Yiping W Han. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating e-cadherin/β-catenin signaling via its fada adhesin. *Cell host & microbe*, 14(2):195–206, 2013.

[192] Jaclyn Strauss, Gilaad G Kaplan, Paul L Beck, Kevin Rioux, Remo Panaccione, Rebekah DeVinney, Tarah Lynch, and Emma Allen-Vercoe. Invasive potential of gut mucosa-derived fusobacterium nucleatum positively correlates with ibd status of the host. *Inflammatory bowel diseases*, 17(9):1971–1978, 2011.

[193] Christian Hoffmann, Serena Dollive, Stephanie Grunberg, Jun Chen, Hongzhe Li, Gary D Wu, James D Lewis, and Frederic D Bushman. Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. *PloS one*, 8(6):e66019, 2013.

[194] Cynthia H Seow, Joanne M Stempak, Wei Xu, Hui Lan, Anne M Griffiths, Gordon R Greenberg, A Hillary Steinhart, Nir Dotan, and Mark S Silverberg. Novel anti-glycan antibodies related to inflammatory bowel disease diagnosis and phenotype. *The American journal of gastroenterology*, 104(6):1426–1434, 2009.

[195] Sofie Joossens, Walter Reinisch, Séverine Vermeire, Boualem Sendid, Daniel Poulain, Marc Peeters, Karel Geboes, Xavier Bossuyt, Peggy Vandewalle, Georg Oberhuber, et al. The value of serologic markers in indeter-

minate colitis: a prospective follow-up study. *Gastroenterology*, 122(5):1242–1247, 2002.

[196] Francesco Strati, Duccio Cavalieri, Davide Albanese, Claudio De Felice, Claudio Donati, Joussef Hayek, Olivier Jousson, Silvia Leoncini, Massimo Pindo, Daniela Renzi, et al. Altered gut microbiota in rett syndrome. *Microbiome*, 4(1):41, 2016.

[197] Maria Chahrour and Huda Y Zoghbi. The story of rett syndrome: from clinic to neurobiology. *Neuron*, 56(3):422–437, 2007.

[198] Helen Leonard, Madhur Ravikumara, Gordon Baikie, Nusrat Naseem, Carolyn Ellaway, Alan Percy, Suzanne Abraham, Suzanne Geerts, Jane Lane, Mary Jones, et al. Assessment and management of nutrition and growth in rett syndrome. *Journal of pediatric gastroenterology and nutrition*, 57(4):451, 2013.

[199] Kathleen J Motil, Erwin Caeg, Judy O Barrish, Suzanne Geerts, Jane B Lane, Alan K Percy, Fran Annese, Lauren McNair, Steven A Skinner, Hye-Seung Lee, et al. Gastrointestinal and nutritional problems occur frequently throughout life in girls and women with rett syndrome. *Journal of pediatric gastroenterology and nutrition*, 55(3):292, 2012.

[200] G Wahba, SC Schock, E Claridge, M Bettolli, D Grynspan, P Humphreys, and WA Staines. Mecp2 in the enteric nervous system. *Neurogastroenterology & Motility*, 27(8):1156–1161, 2015.

[201] Emeran A Mayer, David Padua, and Kirsten Tillisch. Altered brain-gut axis in autism: Comorbidity or causative mechanisms? *Bioessays*, 36(10):933–939, 2014.

[202] Caroline GM de Theije, Harm Wopereis, Mohamed Ramadan, Tiemen van Eijndthoven, Jolanda Lambert, Jan Knol, Johan Garssen, Aletta D Kraneveld, and Raish Oozeer. Altered gut microbiota and activity in a murine model of autism spectrum disorders. *Brain, behavior, and immunity*, 37:197–206, 2014.

[203] Q Li and J-M Zhou. The microbiota–gut–brain axis and its potential therapeutic role in autism spectrum disorder. *Neuroscience*, 324:131–139, 2016.

[204] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

[205] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, 2000.

[206] Marco Ventura, Francesca Turroni, Mary O'Connell Motherway, John MacSharry, and Douwe van Sinderen. Host–microbe interactions that facilitate gut colonization by commensal bifidobacteria. *Trends in microbiology*, 20(10):467–476, 2012.

[207] Saranna Fanning, Lindsay J Hall, Michelle Cronin, Aldert Zomer, John MacSharry, David Goulding, Mary O'Connell Motherway, Fergus Shanahan, Kenneth Nally, Gordon Dougan, et al. Bifidobacterial surface-exopolysaccharide facilitates commensal-host interaction through immune modulation and pathogen protection. *Proceedings of the National Academy of Sciences*, 109(6):2108–2113, 2012.

[208] Laura de Magistris, Valeria Familiari, Antonio Pascotto, Anna Sapone, Alessandro Frolli, Patrizia Iardino, Maria Carteni, Mario De Rosa, Ruggiero Francavilla, Gabriele Riegler, et al. Alterations of the intestinal barrier in patients with autism spectrum disorders and in their first-degree relatives. *Journal of pediatric gastroenterology and nutrition*, 51(4):418–424, 2010.

[209] S Boukthir, N Matoussi, A Belhadj, S Mammou, SB Dlala, M Helayem, F Rocchiccioli, S Bouzaidi, and M Abdennebi. [abnormal intestinal permeability in children with autism]. *La Tunisie medicale*, 88(9):685–686, 2010.

[210] Dae-Wook Kang, Jin Gyoon Park, Zehra Esra Ilhan, Garrick Wallstrom, Joshua LaBaer, James B Adams, and Rosa Krajmalnik-Brown. Reduced incidence of prevotella and other fermenters in intestinal microflora of autistic children. *PloS one*, 8(7):e68322, 2013.

[211] Jessica M Yano, Kristie Yu, Gregory P Donaldson, Gauri G Shastri, Phoebe Ann, Liang Ma, Cathryn R Nagler, Rustem F Ismagilov, Sarkis K Mazmanian, and Elaine Y Hsiao. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell*, 161(2):264–276, 2015.

[212] Maki Kitahara, Fusae Takamine, Teisuke Imamura, and Yoshimi Benno. Clostridium hiranonis sp. nov., a human intestinal bacterium with bile acid 7alpha-dehydroxylating activity. *International journal of systematic and evolutionary microbiology*, 51(1):39–44, 2001.

[213] Seiko Narushima, Kikuji Itoh, Yukiko Miyamoto, Sang-Hee Park, Keiko Nagata, Kazuo Kuruma, and Kiyohisa Uchida. Deoxycholic acid formation in gnotobiotic mice associated with human intestinal bacteria. *Lipids*, 41(9):835–843, 2006.

[214] Virginia García-Cañas, Alejandro Cifuentes, and Carolina Simó. *Applications of Advanced Omics Technologies: From Genes to Metabolites*, volume 64. Elsevier, 2014.

[215] Daniel Erny, Anna Lena Hrabě de Angelis, Diego Jaitin, Peter Wieghofer, Ori Staszewski, Eyal David, Hadas Keren-Shaul, Tanel Mahlakoiv, Kristin Jakobshagen, Thorsten Buch, et al. Host microbiota constantly control maturation and function of microglia in the cns. *Nature neuroscience*, 18(7):965–977, 2015.

[216] Yuliya E Borre, Gerard W O'Keeffe, Gerard Clarke, Catherine Stanton, Timothy G Dinan, and John F Cryan. Microbiota and neurodevelopmental windows: implications for brain disorders. *Trends in molecular medicine*, 20(9):509–518, 2014.

Appendices

Appendix A

Additional results on P_IBD

A.1 Biomarkers of P_IBD dataset

For each classification task, Figures A.1-A.5 report heatmaps with median abundance of predictive biomarkers stratified by phenotypes. Figures A.6-A.10 show the fold change of biomarkers abundance between phenotypes. Biomarkers are sorted in descending order according to their importance in discriminating phenotypes; the most discriminant feature on the top. Tables A.1-A.5 report the p-values of Wilcoxon test on biomarkers abundance, for each classification task.

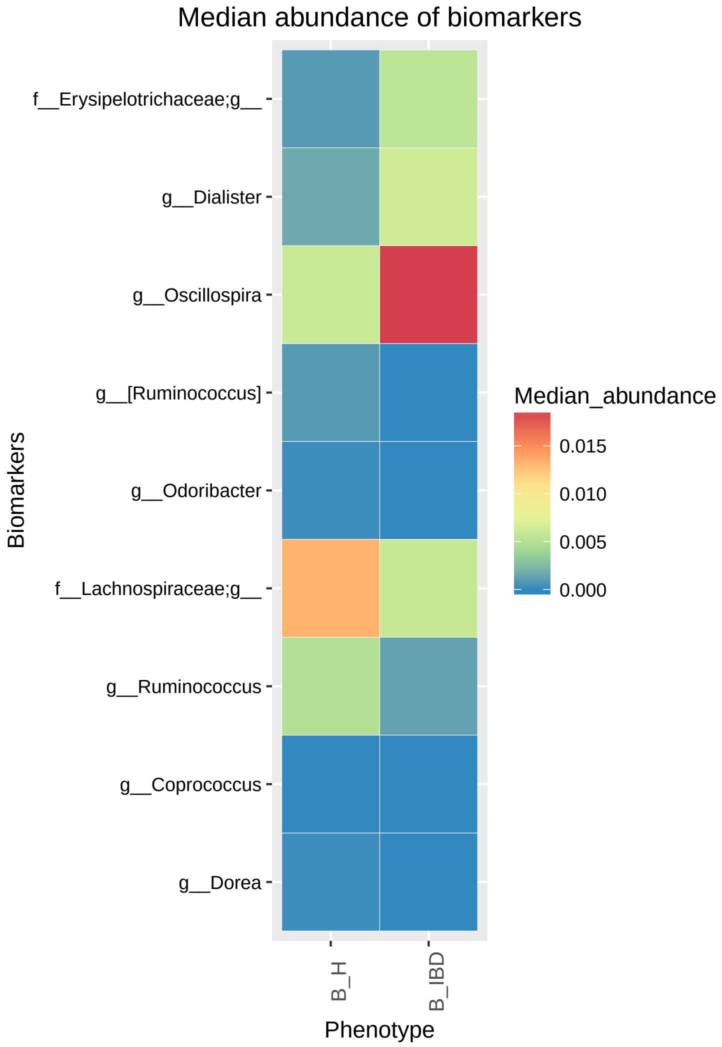


Figure A.1. Heatmap of median abundance of biomarkers for healthy biopsies vs. IBD biopsies.

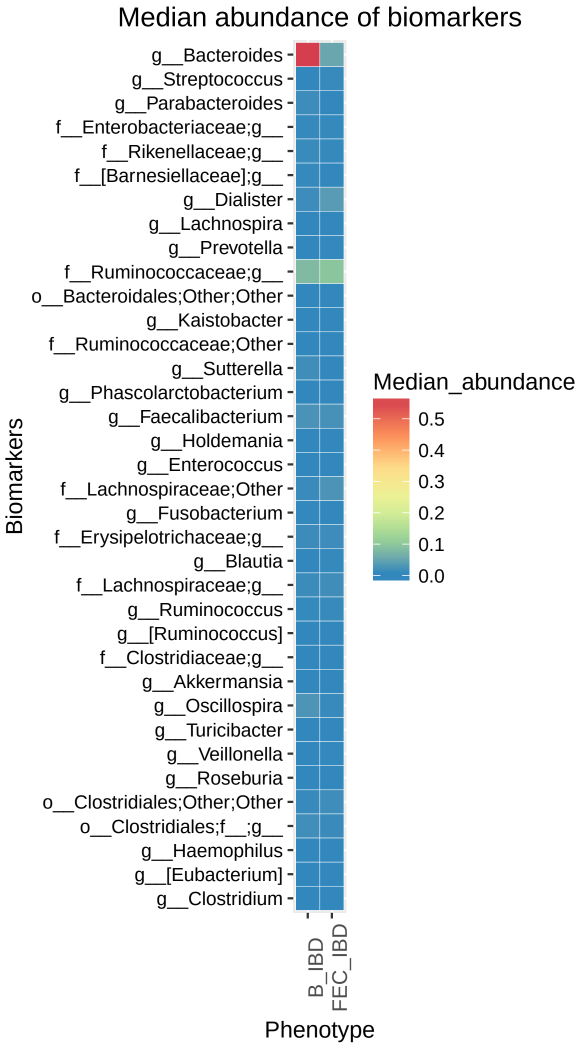


Figure A.2. Heatmap of median abundance of biomarkers for IBD feces vs. IBD biopsies.

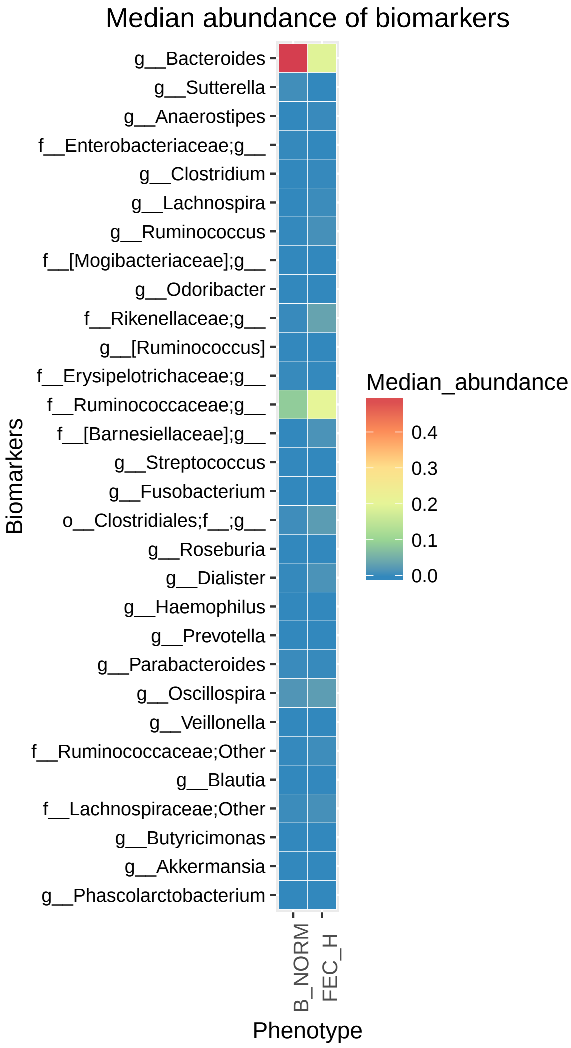


Figure A.3. Heatmap of median abundance of biomarkers for healthy feces vs. IBD biopsies, not inlamed tissue.

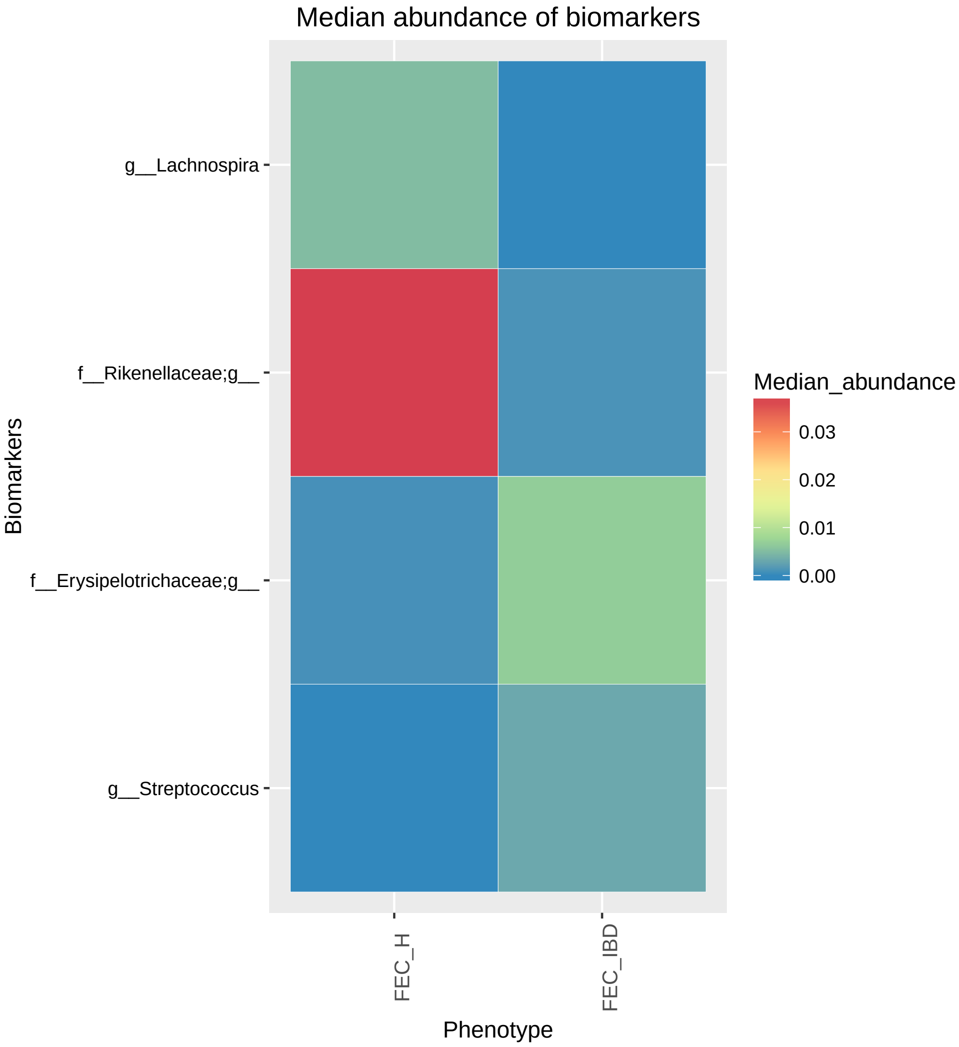


Figure A.4. Heatmap of median abundance of biomarkers for healthy feces vs. IBD feces.

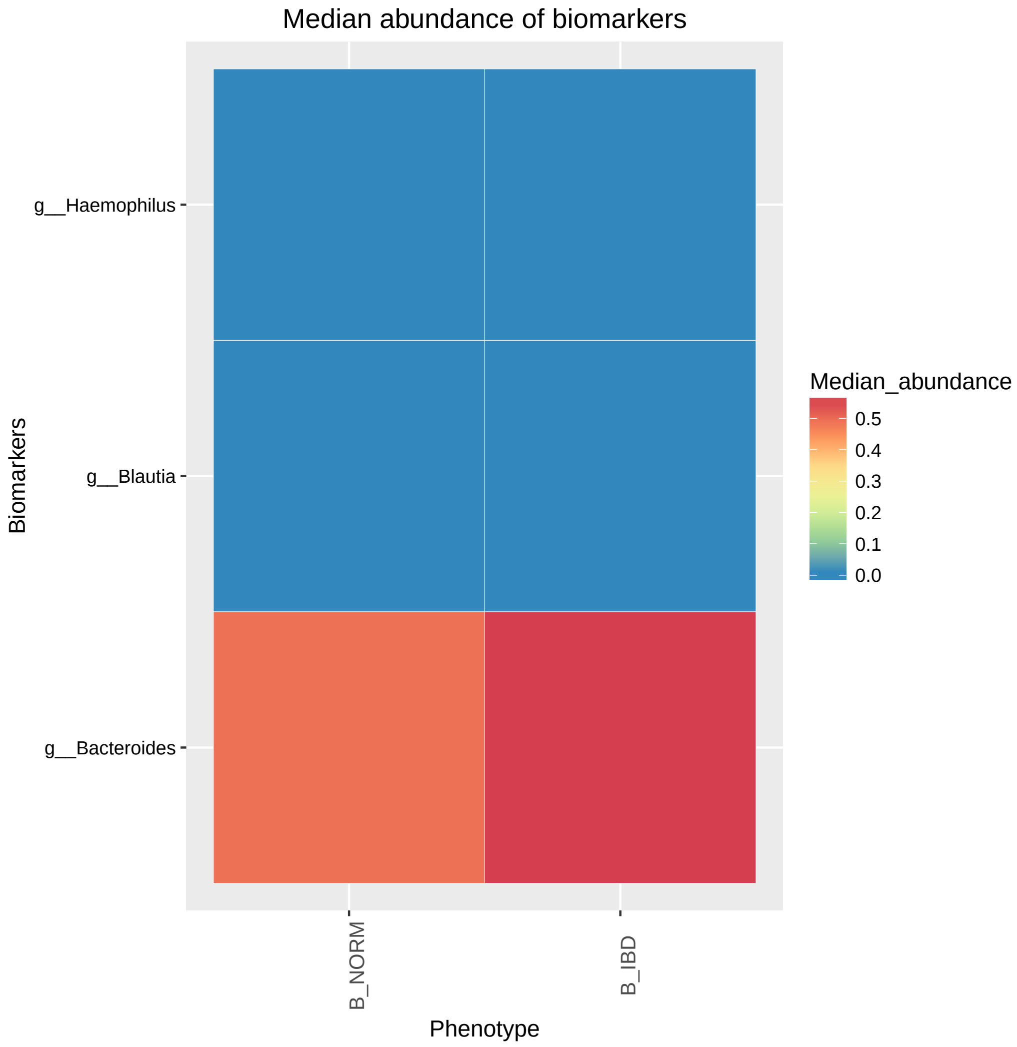


Figure A.5. Heatmap of median abundance of biomarkers for matched IBD biopsies, normal vs. inflamed tissue.

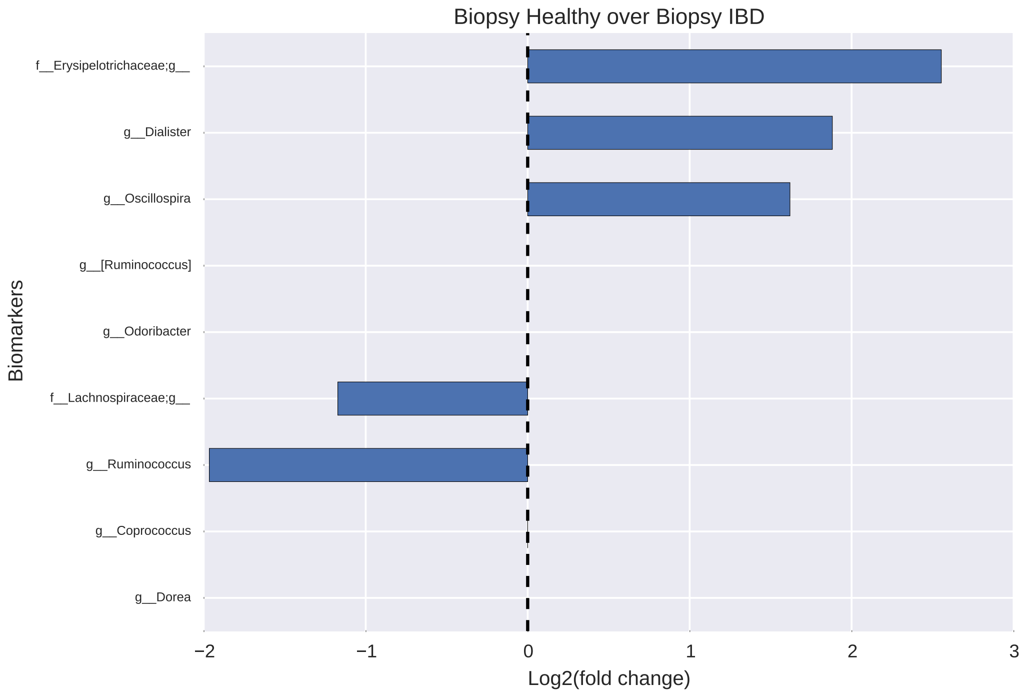


Figure A.6. Log base 2 of [median abundance in organisms increased in IBD biopsies] over [median abundance of organisms increased in healthy biopsies].

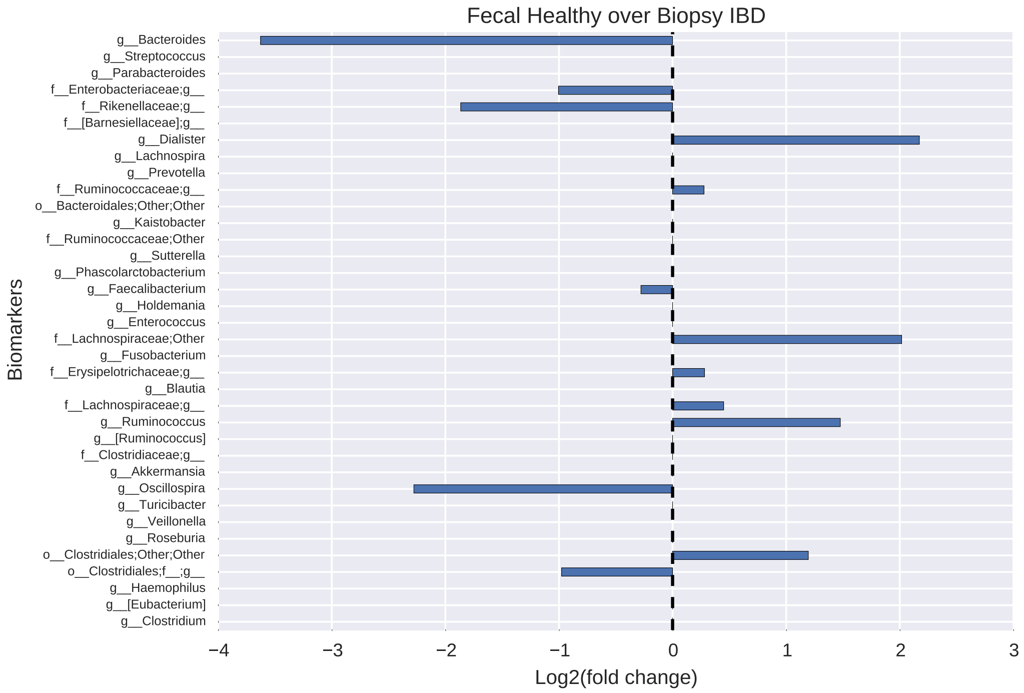


Figure A.7. Log base 2 of [median abundance of organisms increased in IBD feces] over [median abundance in organisms increased in IBD biopsies.

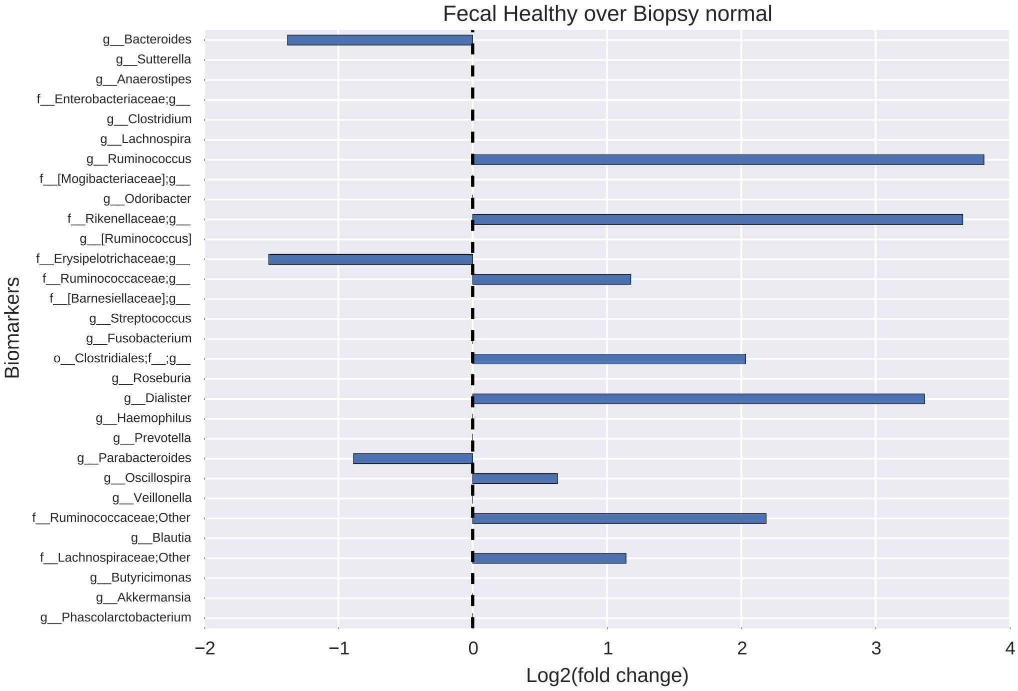


Figure A.8. Log base 2 of [median abundance of organisms increased in healthy feces] over [median abundance in organisms increased in IBD biopsies, not inflamed tissue].

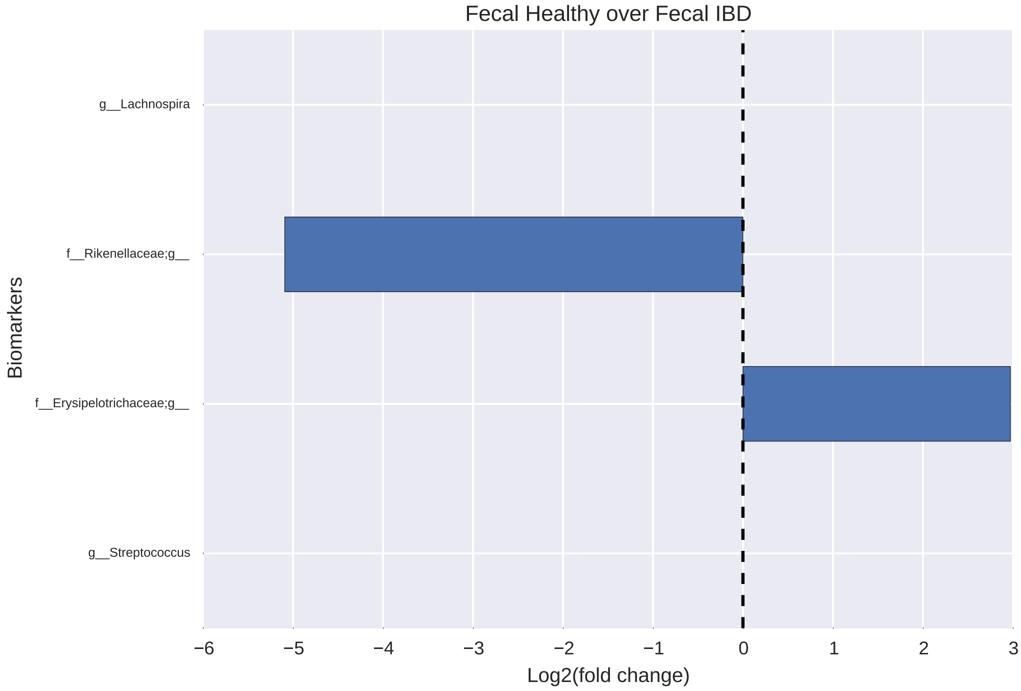


Figure A.9. Log base 2 of [median abundance in organisms increased in IBD feces] over [median abundance of organisms increased in healthy feces].

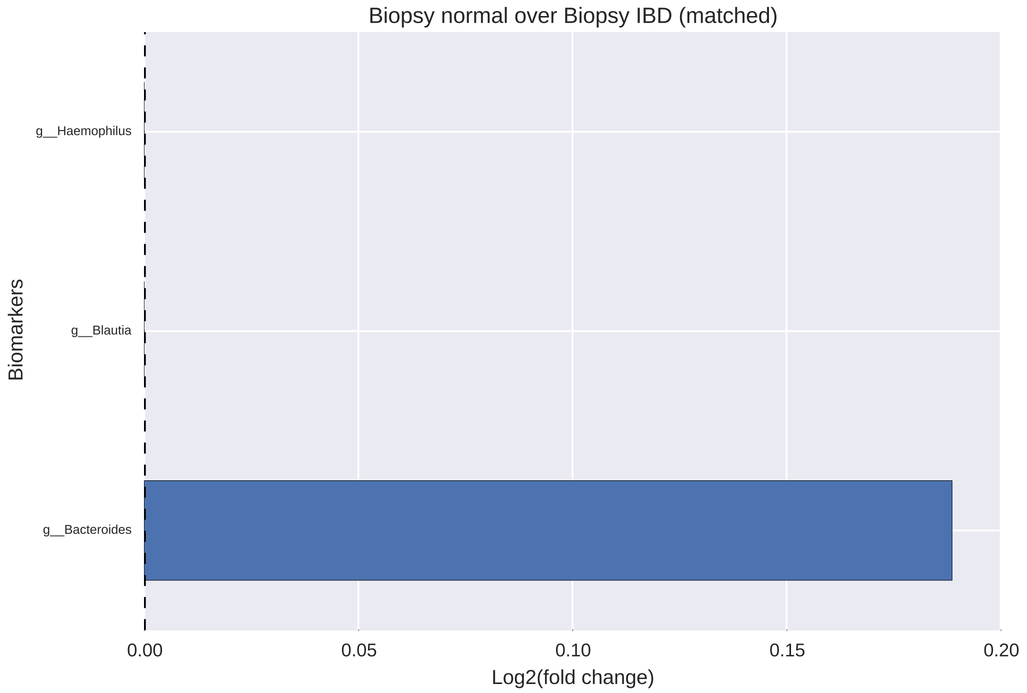


Figure A.10. Log base 2 of [median abundance in organisms increased in IBD biopsies, inflamed tissue] over [median abundance of organisms increased in IBD biopsies, not inflamed tissue].

B_H_IBD	
Top biomarkers	**Wilcoxon p-values**
f__Erysipelotrichaceae;g__	1.77881E-02
g__Dialister	2.61558E-01
g__Oscillospira	9.55575E-02
g__[Ruminococcus]	2.17119E-02
g__Odoribacter	4.18383E-01
f__Lachnospiraceae;g__	7.96954E-02
g__Ruminococcus	2.93979E-03
g__Coprococcus	6.25527E-02
g__Dorea	4.25594E-02

Table A.1. P-values of Wilcoxon test on biomarkers abundance for healthy biopsies vs. IBD biopsies.

FEC_H_IBD	
Top biomarkers	**Wilcoxon p-values**
g__Lachnospira	1.71376E-03
f__Rikenellaceae;g__	1.38976E-07
f__Erysipelotrichaceae;g__	4.26684E-04
g__Streptococcus	2.05802E-04

Table A.4. P-values of Wilcoxon test on biomarkers abundance for healthy feces vs. IBD feces.

FEC_B_IBD	
Top biomarkers	**Wilcoxon p-values**
g__Bacteroides	4.80859E-04
g__Streptococcus	4.61166E-04
g__Parabacteroides	7.9805E-03
f__Enterobacteriaceae;g__	4.49528E-01
f__Rikenellaceae;g__	6.75826E-02
f__[Barnesiellaceae];g__	1.05914E-02
g__Dialister	2.6979E-02
g__Lachnospira	1.04275E-01
g__Prevotella	9.28555E-01
f__Ruminococcaceae;g__	2.66239E-01
o__Bacteroidales;Other;Other	8.25216E-02
g__Kaistobacter	5.62926E-03
f__Ruminococcaceae;Other	8.30254E-01
g__Sutterella	3.31742E-02
g__Phascolarctobacterium	9.88294E-01
g__Faecalibacterium	6.46956E-01
g__Holdemania	5.85148E-01
g__Enterococcus	8.92591E-02
f__Lachnospiraceae;Other	2.1441E-02
g__Fusobacterium	1.30504E-02
f__Erysipelotrichaceae;g__	6.12137E-01
g__Blautia	5.70265E-02
f__Lachnospiraceae;g__	9.61383E-01
g__Ruminococcus	2.28666E-02
g__[Ruminococcus]	9.64892E-01
f__Clostridiaceae;g__	2.50455E-01
g__Akkermansia	2.40207E-01
g__Oscillospira	1.41452E-01
g__Turicibacter	1.37116E-02
g__Veillonella	5.13455E-01
g__Roseburia	6.43443E-01
o__Clostridiales;Other;Other	4.3209E-01
o__Clostridiales;f__;g__	8.9458E-01
g__Haemophilus	2.83861E-01
g__[Eubacterium]	4.22703E-01
g__Clostridium	1.17965E-01

Table A.2. P-values of Wilcoxon test on biomarkers abundance for IBD feces vs. IBD biopsies.

FEC_H_B_NORM	
Top biomarkers	**Wilcoxon p-values**
g__Bacteroides	2.76595E-03
g__Sutterella	3.71264E-05
g__Anaerostipes	9.44149E-05
f__Enterobacteriaceae;g__	1.88115E-03
g__Clostridium	1.95763E-02
g__Lachnospira	3.98538E-03
g__Ruminococcus	1.02281E-04
f__[Mogibacteriaceae];g__	5.38087E-03
g__Odoribacter	5.38541E-01
f__Rikenellaceae;g__	1.00361E-02
g__[Ruminococcus]	9.62147E-06
f__Erysipelotrichaceae;g__	1.52362E-01
f__Ruminococcaceae;g__	2.59742E-04
f__[Barnesiellaceae];g__	1.10526E-02
g__Streptococcus	2.16156E-01
g__Fusobacterium	1.65858E-03
o__Clostridiales;f__;g__	5.16186E-03
g__Roseburia	5.34096E-01
g__Dialister	1.01919E-02
g__Haemophilus	3.3895E-02
g__Prevotella	2.61541E-01
g__Parabacteroides	7.89102E-01
g__Oscillospira	4.62267E-01
g__Veillonella	9.08573E-01
f__Ruminococcaceae;Other	7.31038E-03
g__Blautia	5.11699E-02
f__Lachnospiraceae;Other	8.08641E-02
g__Butyricimonas	8.98006E-01
g__Akkermansia	6.77637E-02
g__Phascolarctobacterium	3.27037E-01

Table A.3. P-values of Wilcoxon test on biomarkers abundance for healthy feces vs. IBD biopsies, not inflamed tissue.

B_NORM_IBD	
Top biomarkers	**Wilcoxon p-values**
g__Haemophilus	8.33087E-01
g__Blautia	1.98723E-01
g__Bacteroides	7.3991E-01

Table A.5. P-values of Wilcoxon test on biomarkers abundance for matched IBD biopsies, normal vs. inflamed tissue.

Appendix B

Additional RF results on A_IBD

B.1 RF biomarkers of A_IBD dataset

For each classification task, Figures B.1-B.8 report heatmaps with median abundance of predictive biomarkers stratified by phenotypes. Figures B.9-B.16 show the fold change of biomarkers abundance between phenotypes.Biomarkers are sorted in descending order according to their importance in discriminating phenotypes; the most discriminant feature on the top.

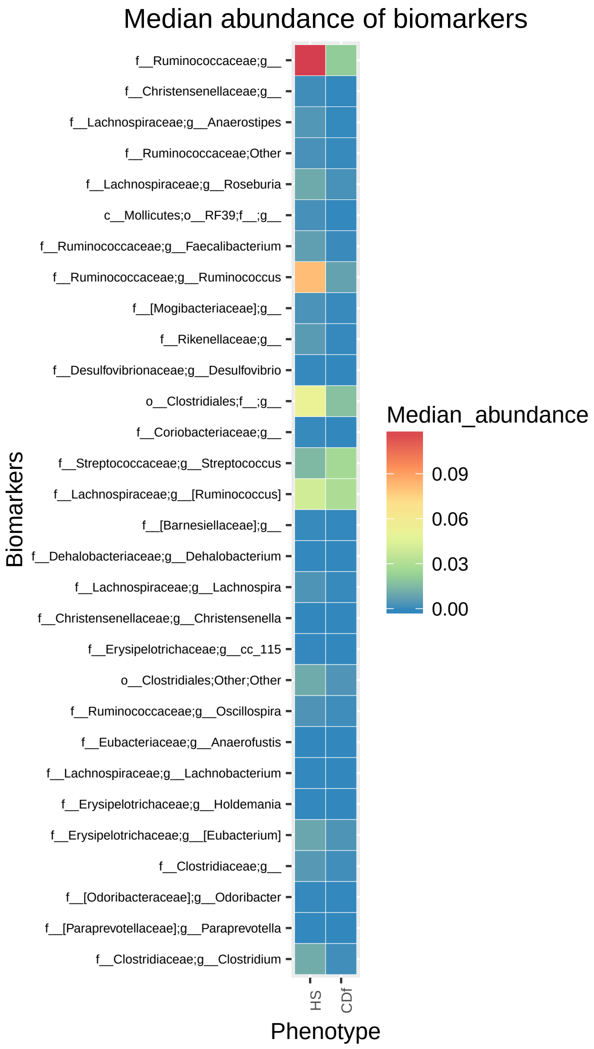


Figure B.1. Heatmap of median abundance of biomarkers for healthy vs. Crohn’s disease patients in flare.

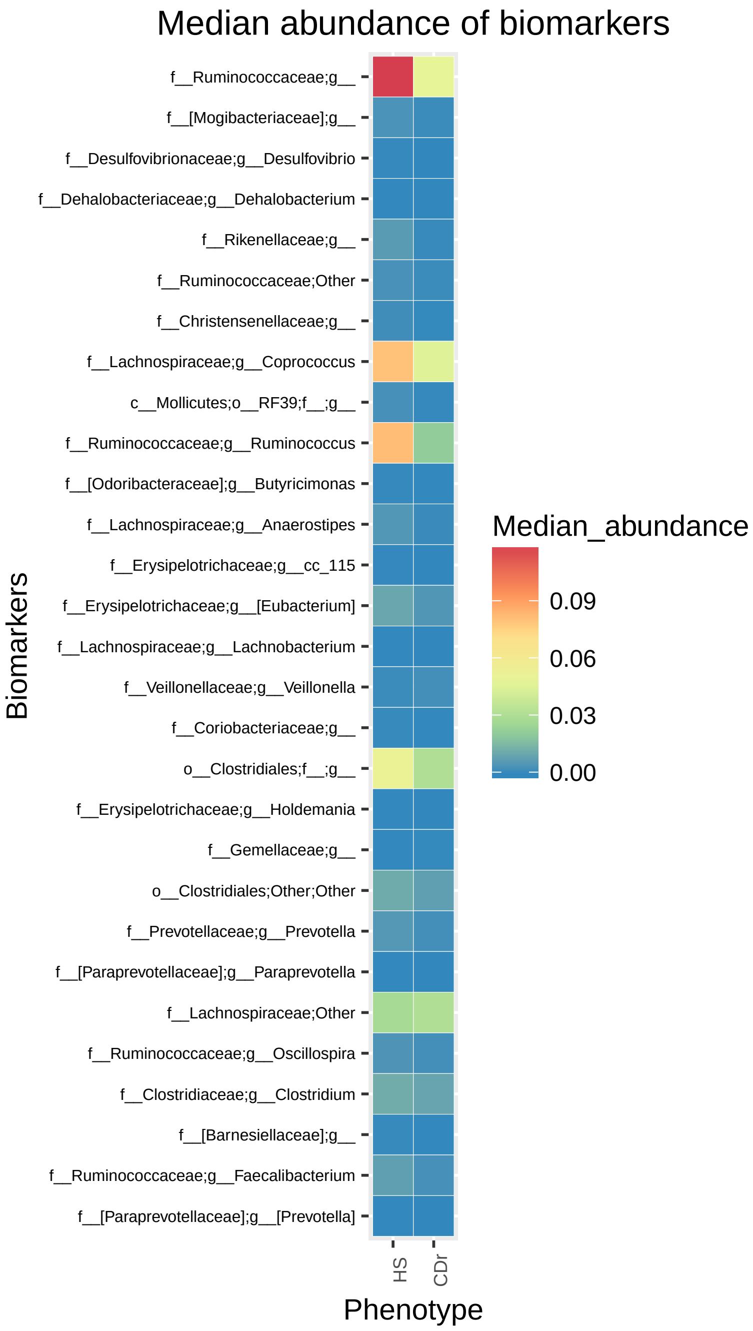


Figure B.2. Heatmap of median abundance of biomarkers for healthy vs. Crohn's disease patients in remission.

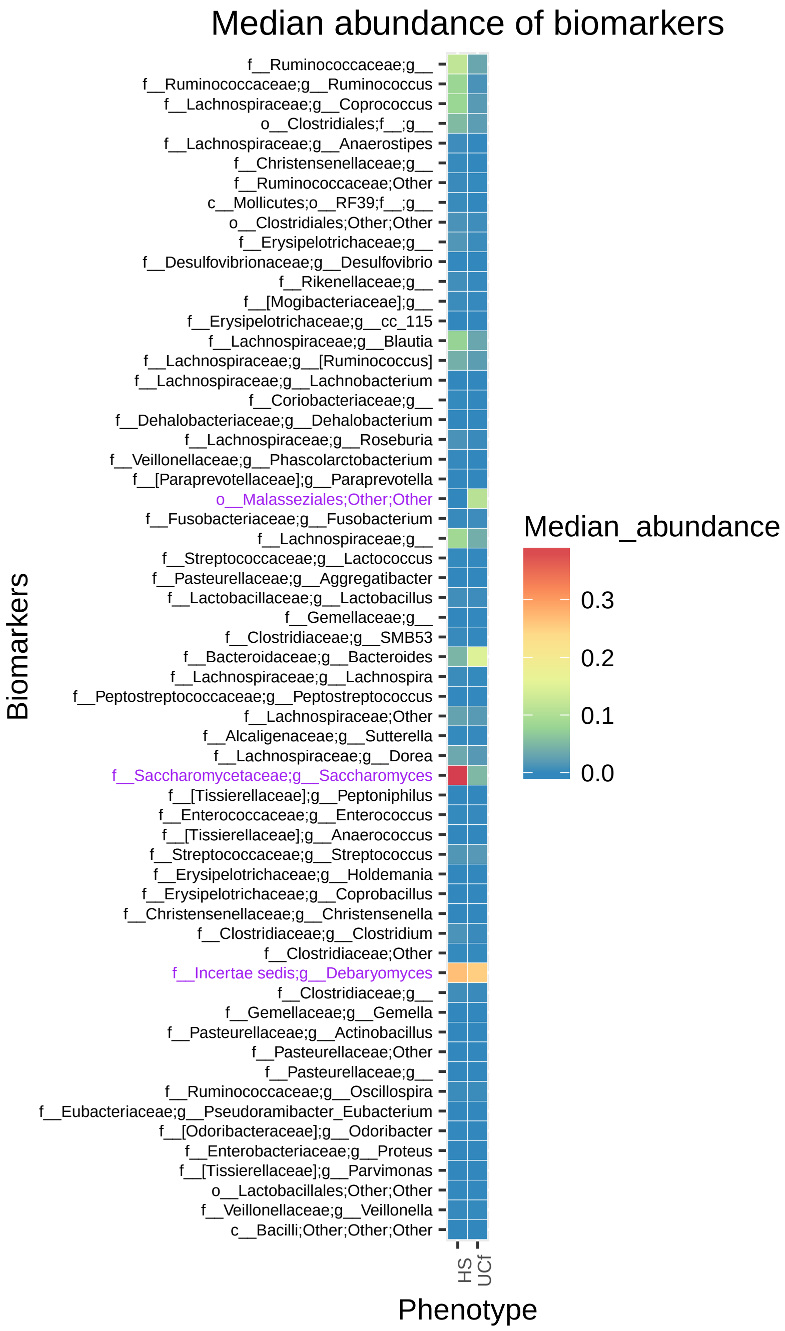


Figure B.3. Heatmap of median abundance of biomarkers for healthy vs. Ulcerative Colitis patients in flare. Black: bacteria; purple: fungi.

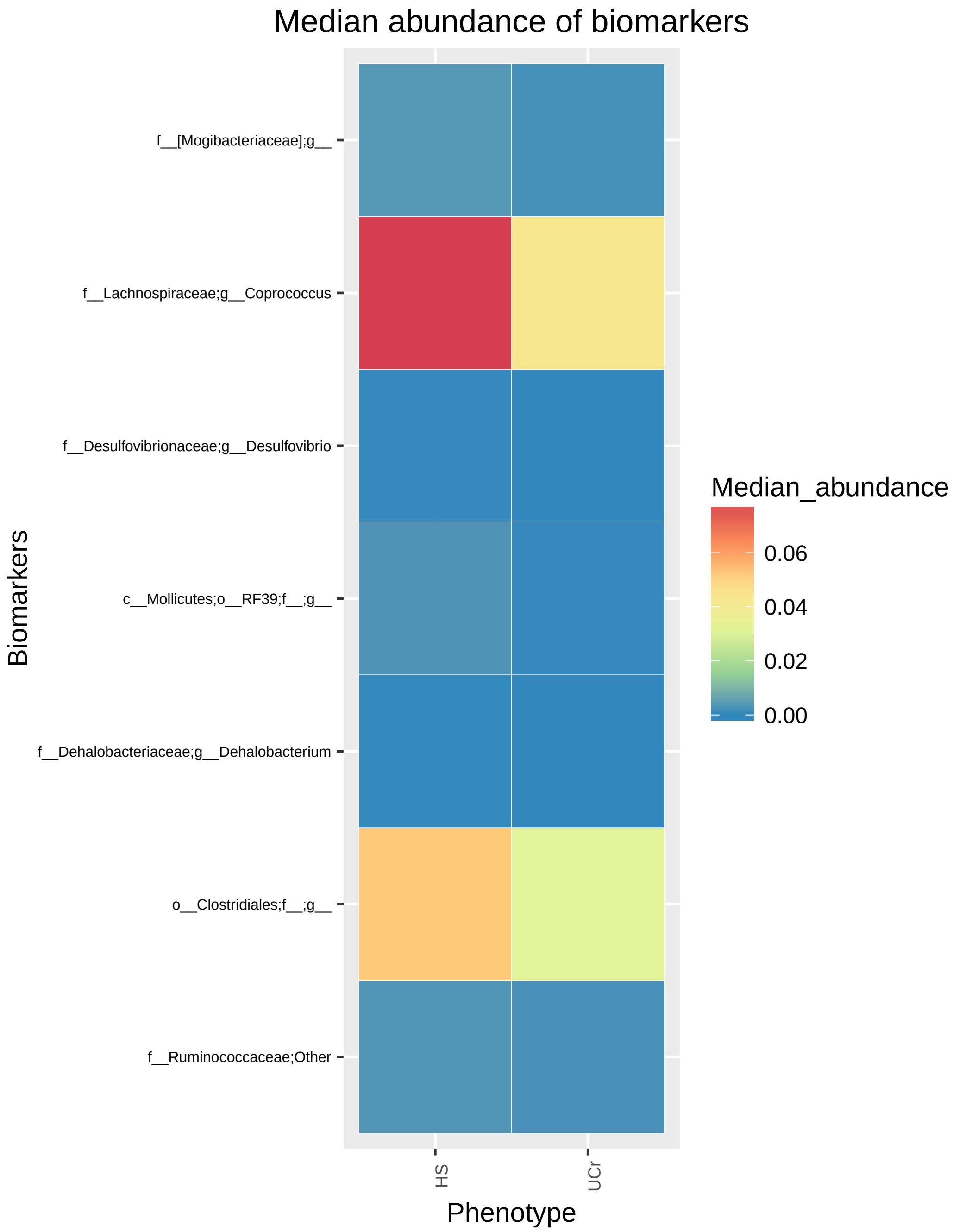


Figure B.4. Heatmap of median abundance of biomarkers for healthy vs. Ulcerative Colitis patients in remission.

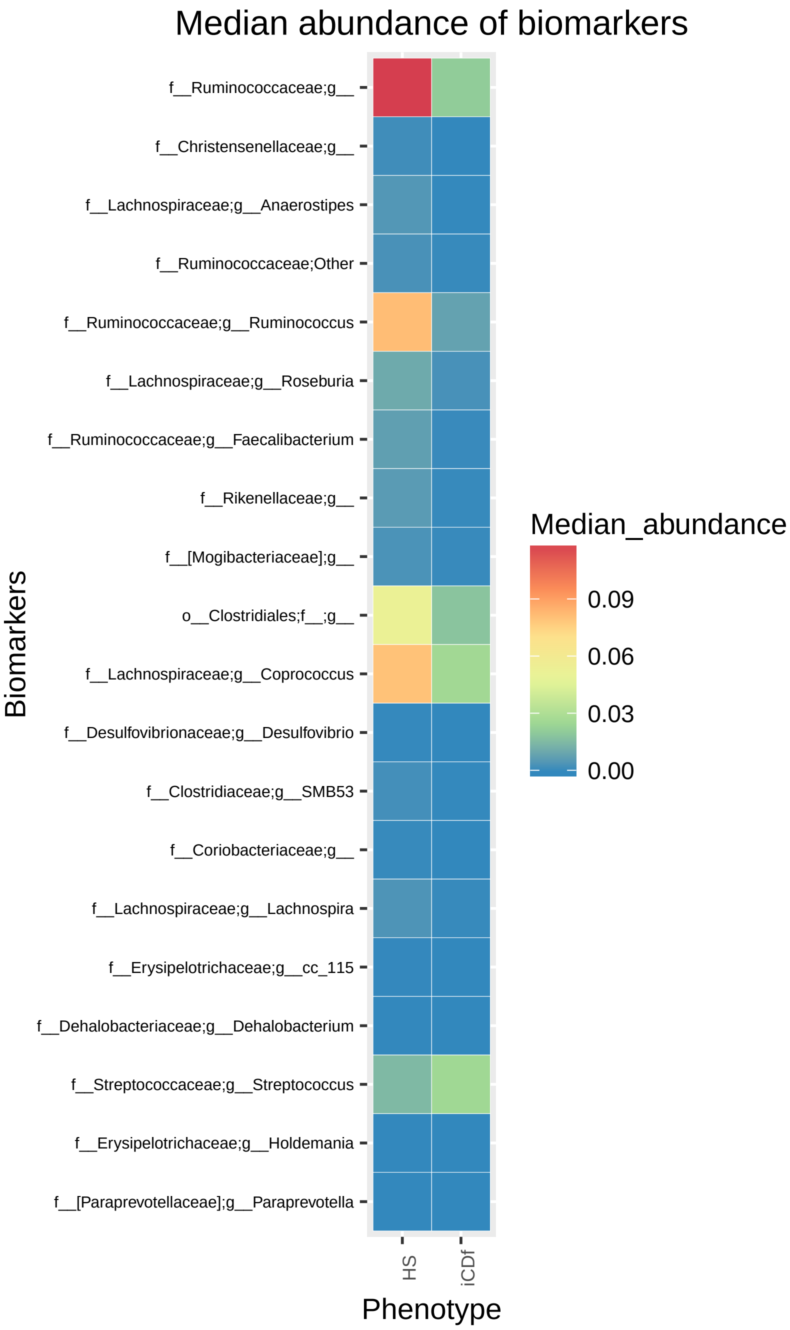


Figure B.5. Heatmap of median abundance of biomarkers for healthy vs. ileal Crohn's disease patients in flare.

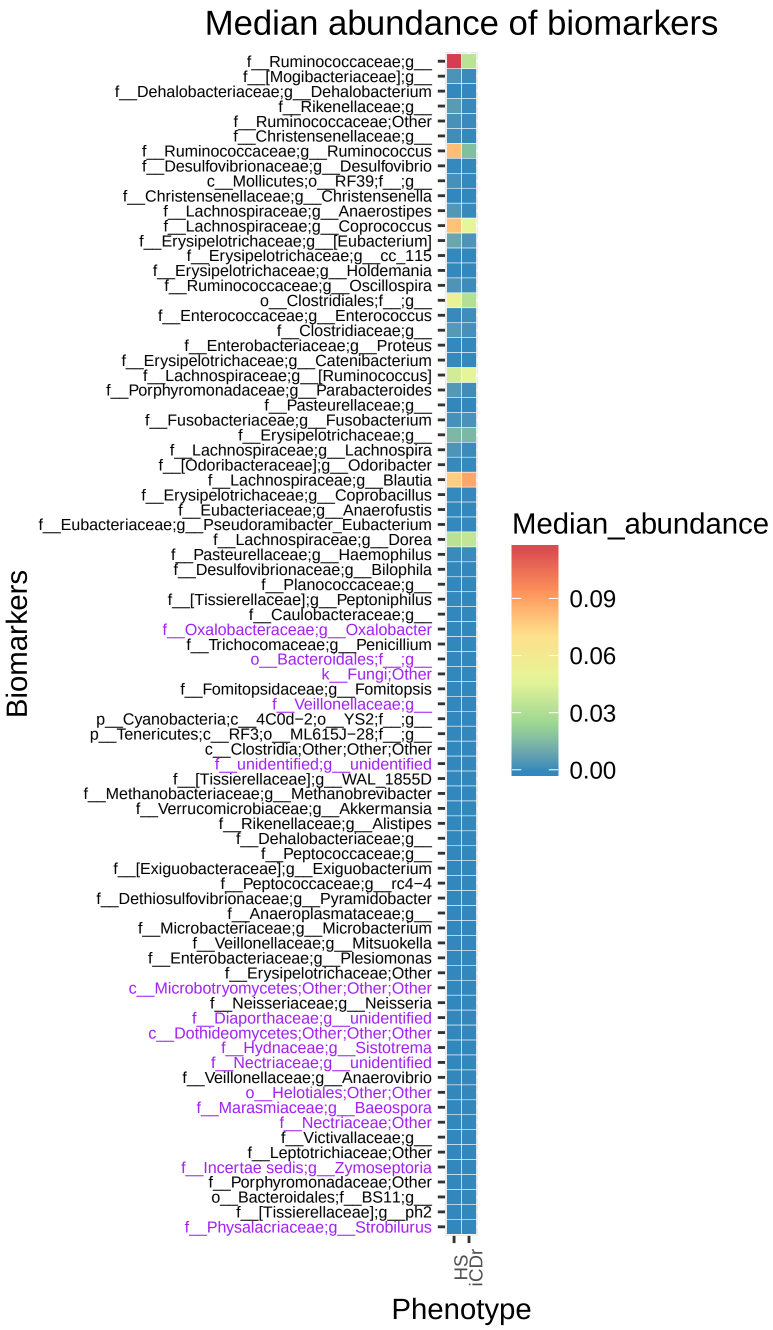


Figure B.6. Heatmap of median abundance of biomarkers for healthy vs. ileal Crohn's disease patients in remission. Black: bacteria; purple: fungi.

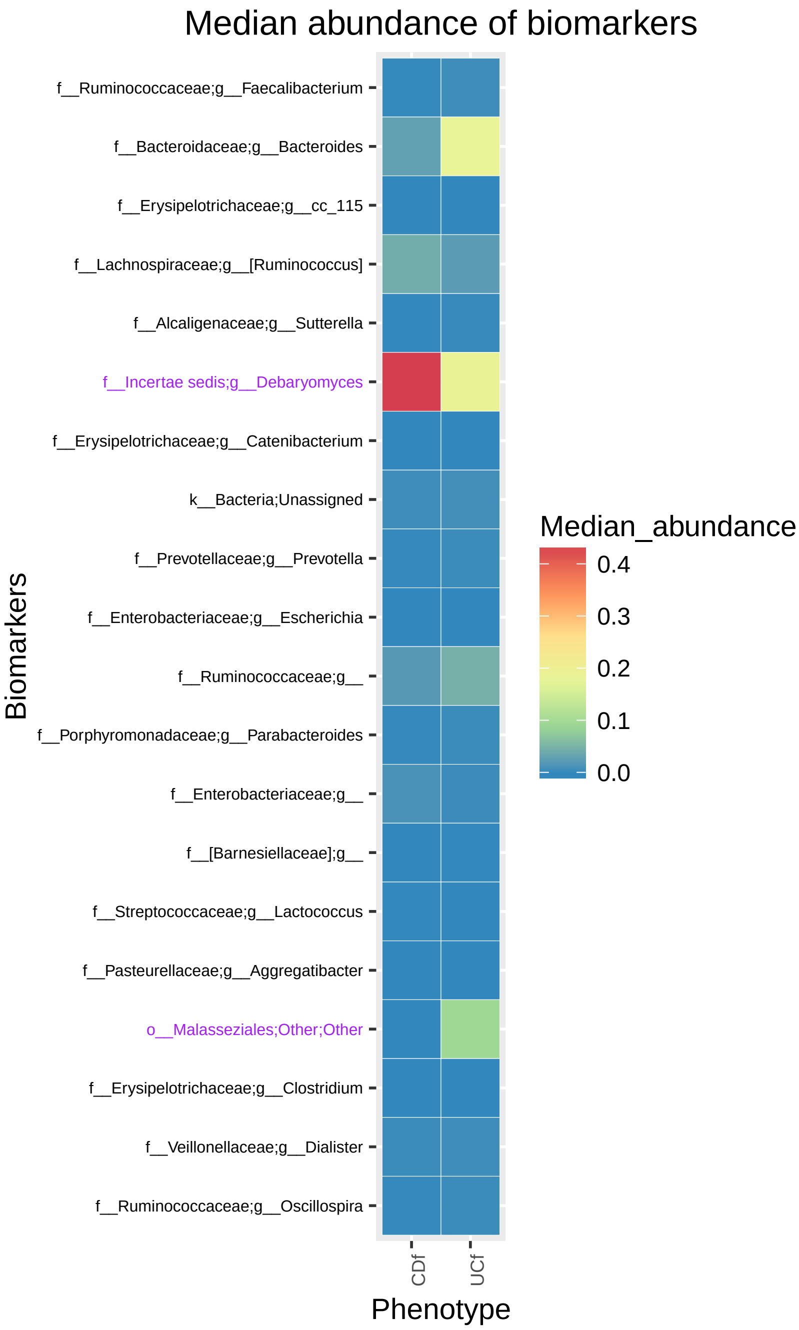


Figure B.7. Heatmap of median abundance of biomarkers for Crohn vs. Ulcerative Colitis patients in flare. Black: bacteria; purple: fungi.

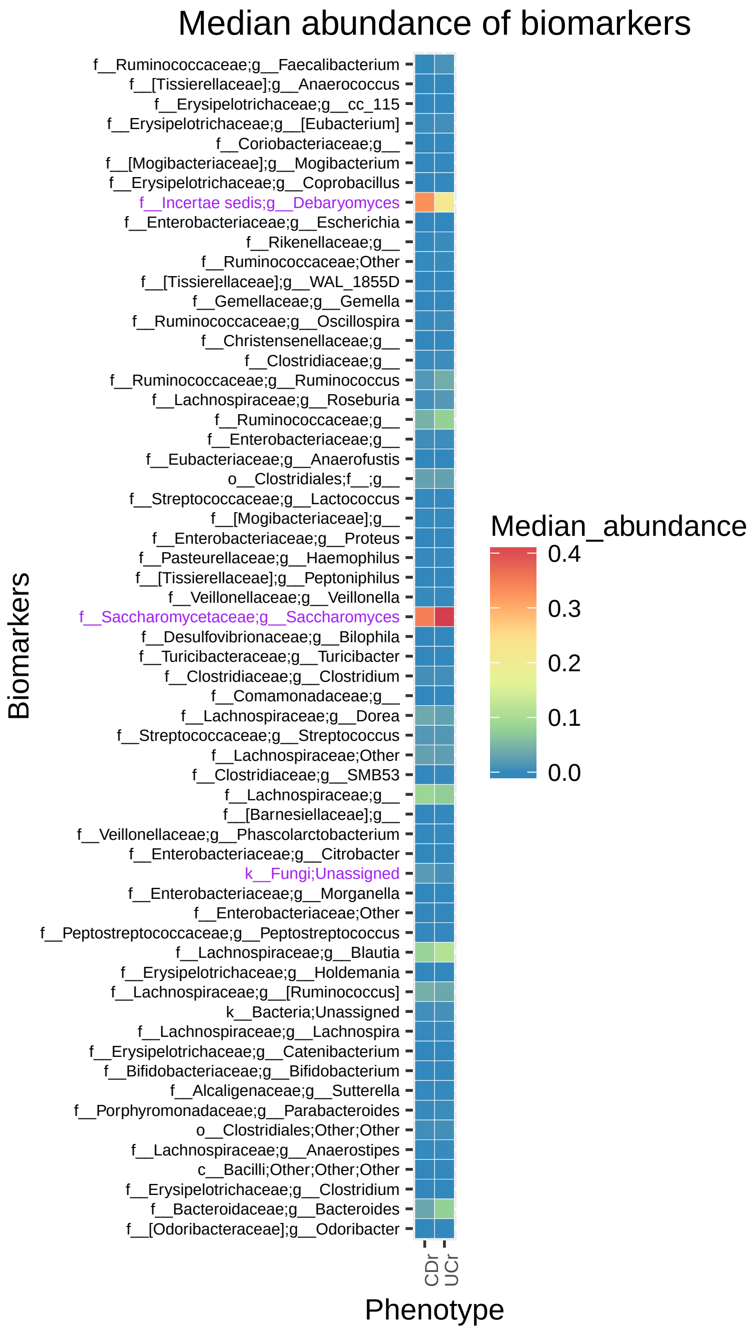


Figure B.8. Heatmap of median abundance of biomarkers for Crohn vs. Ulcerative Colitis patients in remission.

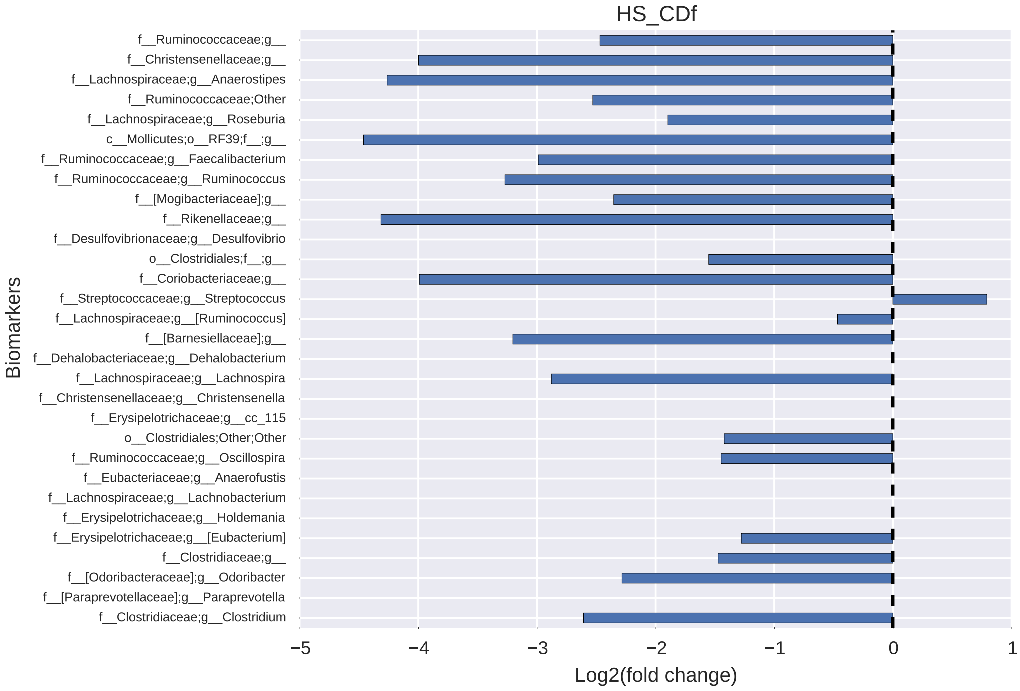


Figure B.9. Log base 2 of [median abundance in organisms increased in Crohn's disease patients in flare] over [median abundance of organisms increased in healthy subjects].

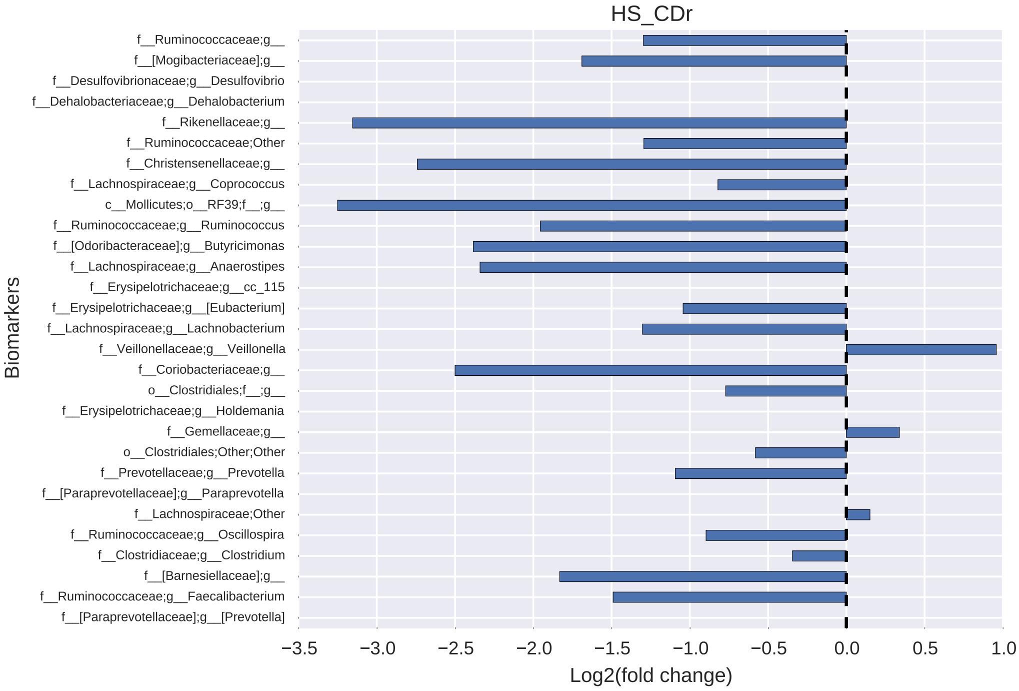


Figure B.10. Log base 2 of [median abundance in organisms increased in Crohn's disease patients in remission] over [median abundance of organisms increased in healthy subjects].

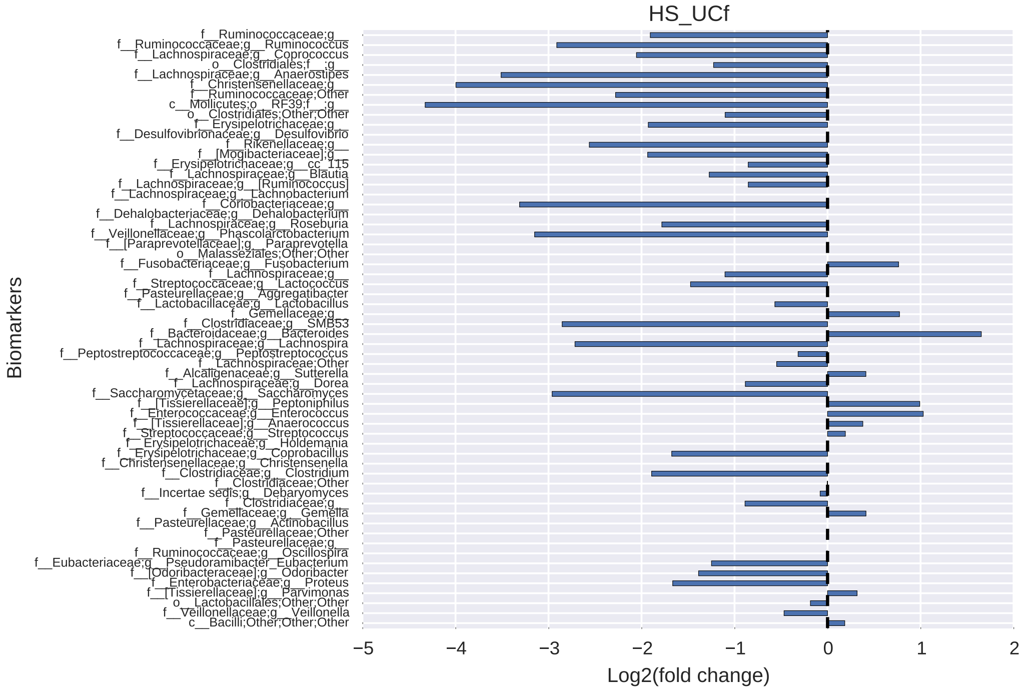


Figure B.11. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in flare] over [median abundance of organisms increased in healthy subjects].

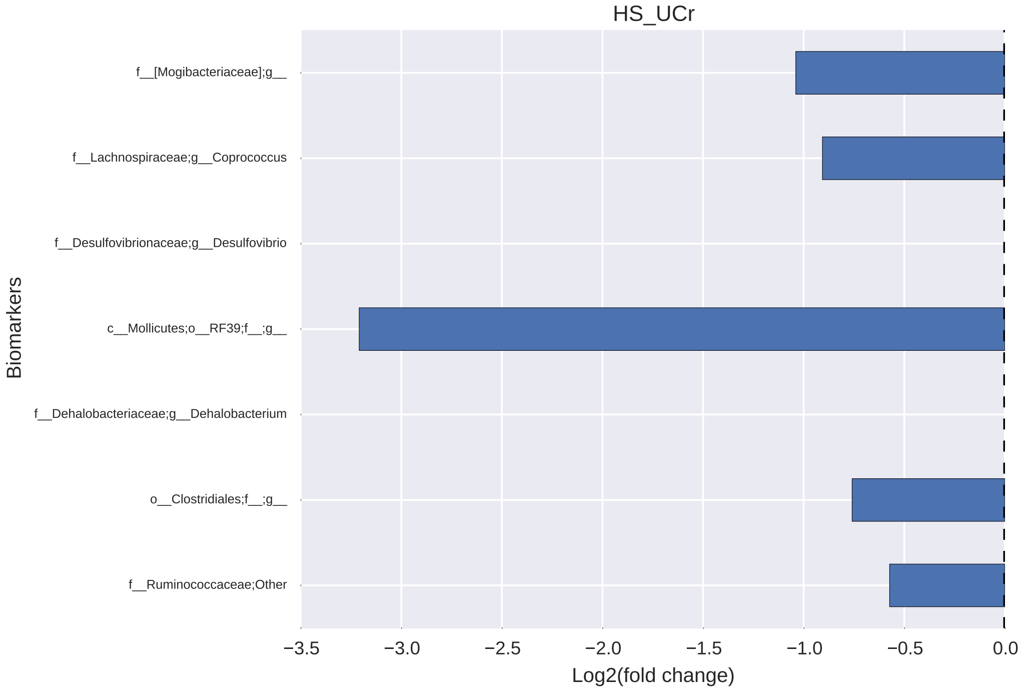


Figure B.12. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in remission] over [median abundance of organisms increased in healthy subjects].

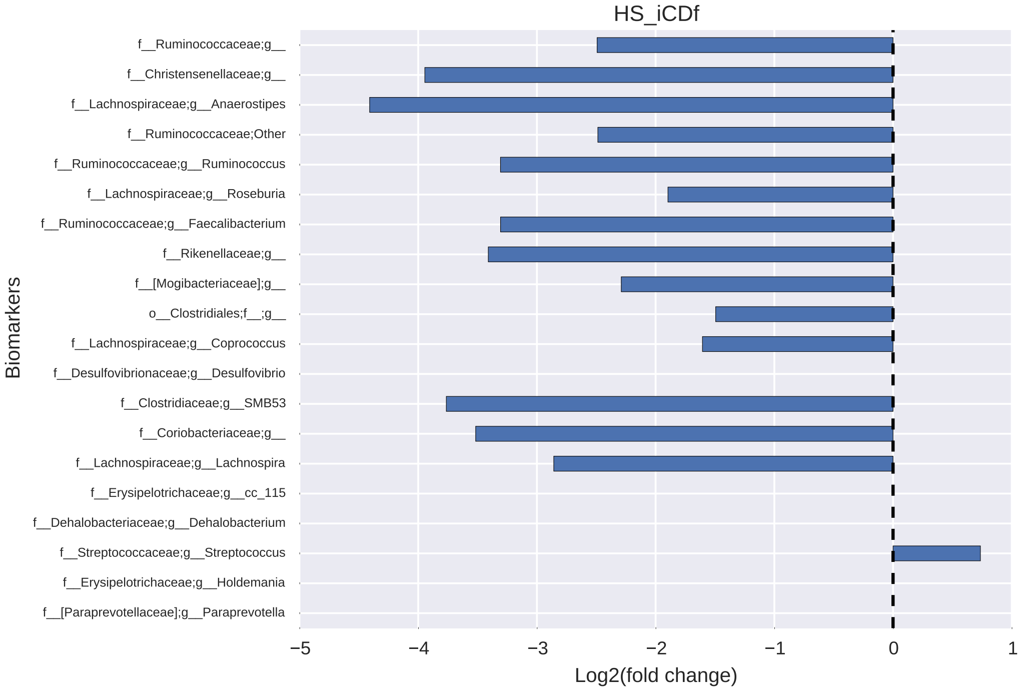


Figure B.13. Log base 2 of [median abundance in organisms increased in ileal Crohn's disease patients in flare] over [median abundance of organisms increased in healthy subjects].

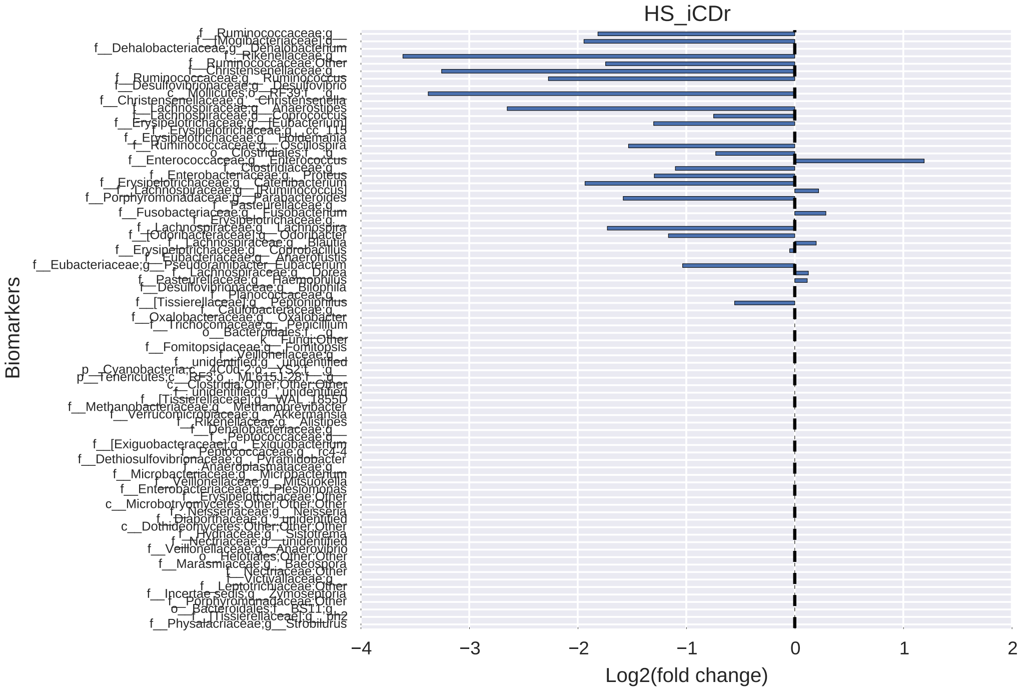


Figure B.14. Log base 2 of [median abundance in organisms increased in ileal Crohn's disease patients in remission] over [median abundance of organisms increased in healthy subjects].

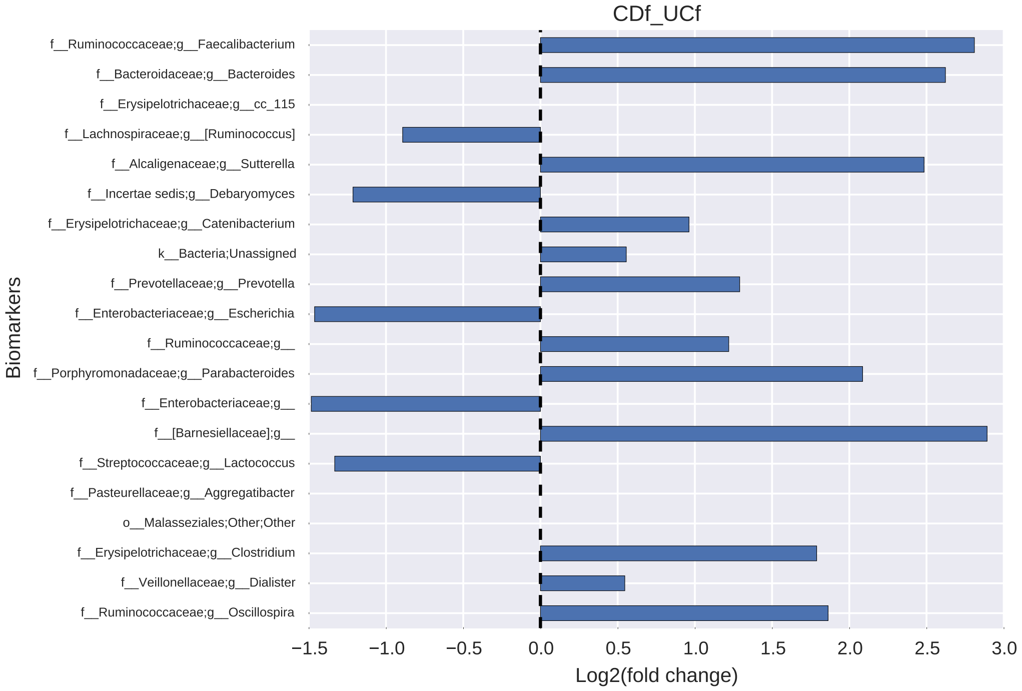


Figure B.15. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in flare] over [median abundance of organisms increased in Crohn's disease patients in flare].

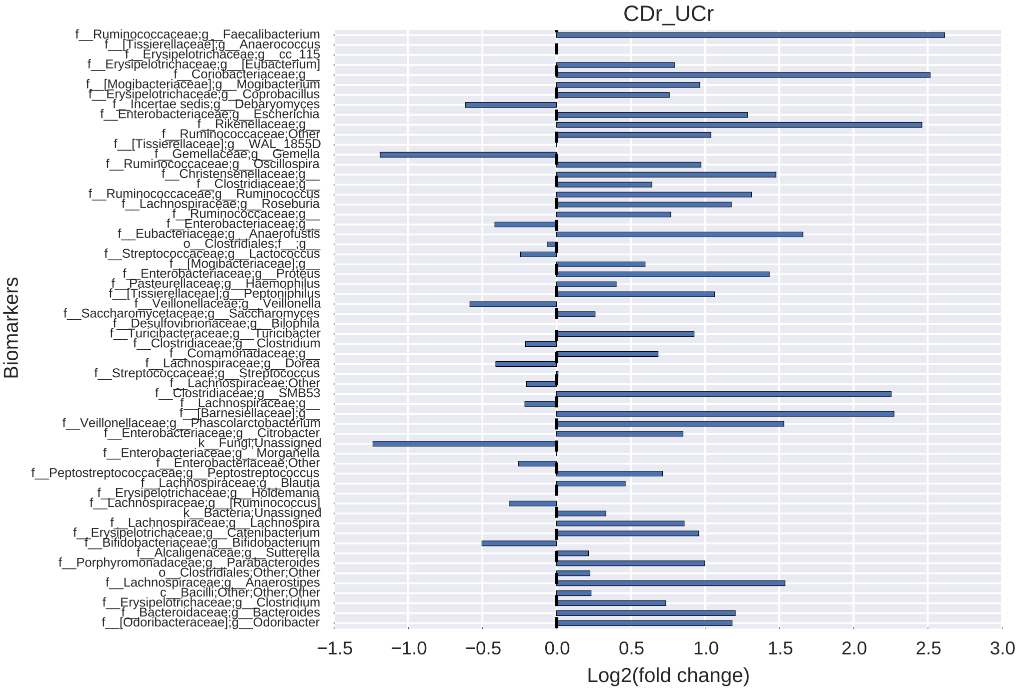


Figure B.16. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in remission] over [median abundance of organisms increased in Crohn's disease patients in remission].

Appendix C

SVM results on A_IBD

C.1 SVM predictive performance on A_IBD dataset

Predictive performance of SVM in INF module are reported in Table C.1 in terms of best average MCC on training set (MCC_{int}) with 95% Student bootstrap (1000×resampling) confidence intervals(MCC_{min} , MCC_{max}), MCC on validation set (MCC_{val}), number of features (Nf) leading to MCC_{int}.

	ml-J			**INF**		
	MCC_{int} (MCC_{min}, MCC_{max})	MCC_{val}	Nf	MCC_{int} (MCC_{min}, MCC_{max})	MCC_{val}	Nf
HS_CDf	0.73 (0.69, 0.76)	0.57	5	0.79 (0.75, 0.82)	0.55	2
HS_CDr	0.49 (0.43, 0.55)	0.59	20	0.78 (0.74, 0.82)	0.59	20
HS_UCf	0.78 (0.73, 0.83)	0.48	60	0.83 (0.78, 0.87)	0.48	6
HS_UCr	0.58 (0.52, 0.65)	0.43	10	0.78 (0.73, 0.83)	0.24	6
HS_iCDf	0.72 (0.68, 0.77)	0.58	7	0.80 (0.76, 0.85)	0.66	1
HS_iCDr	0.56 (0.50, 0.62)	0.47	40	0.79 (0.74, 0.83)	0.47	33
CDf_UCf	0.31 (0.24, 0.38)	0.16	50	0.66 (0.61, 0.70)	0.25	38
CDr_UCr	0.40 (0.32, 0.46)	0.17	20	0.74 (0.70, 0.79)	0.13	19

Table C.1. SVM predictive performances for ml-J and INF. MCC_{int}: best mean MCC on training set; (MCC_{min}, MCC_{max}): MCC_{int} 95% bootstrap confidence interval; MC_val: MCC on validation set; Nf: number of genera leading to MCC_{int}.

C.2 SVM biomarkers of A_IBD dataset

For each classification task, Figures C.1-C.8 report heatmaps with median abundance of predictive biomarkers stratified by phenotypes. Figures C.9-C.16 show the fold change of biomarkers abundance between phenotypes. Biomarkers are sorted in descending order according to their importance in discriminating phenotypes; the most discriminant feature on the top.

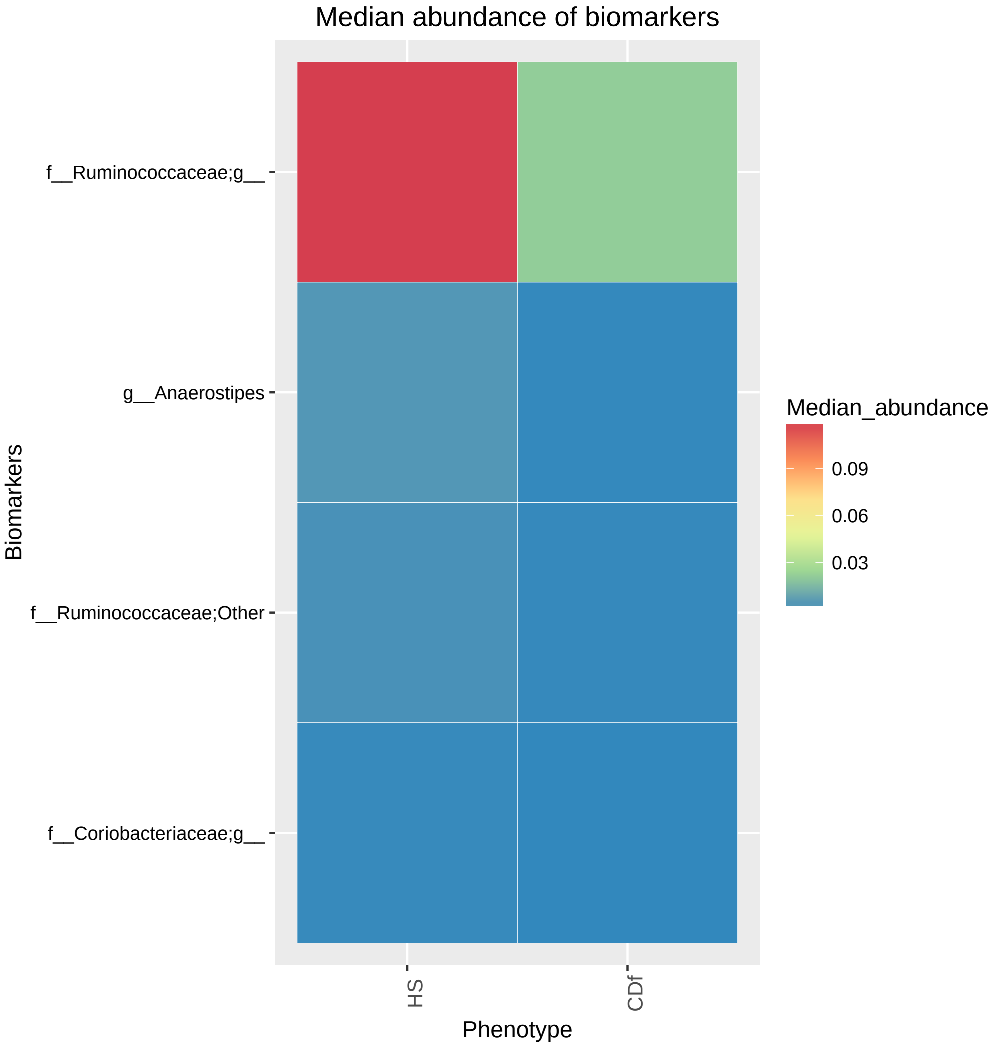


Figure C.1. Heatmap of median abundance of biomarkers for healthy vs. Crohn's disease patients in flare.

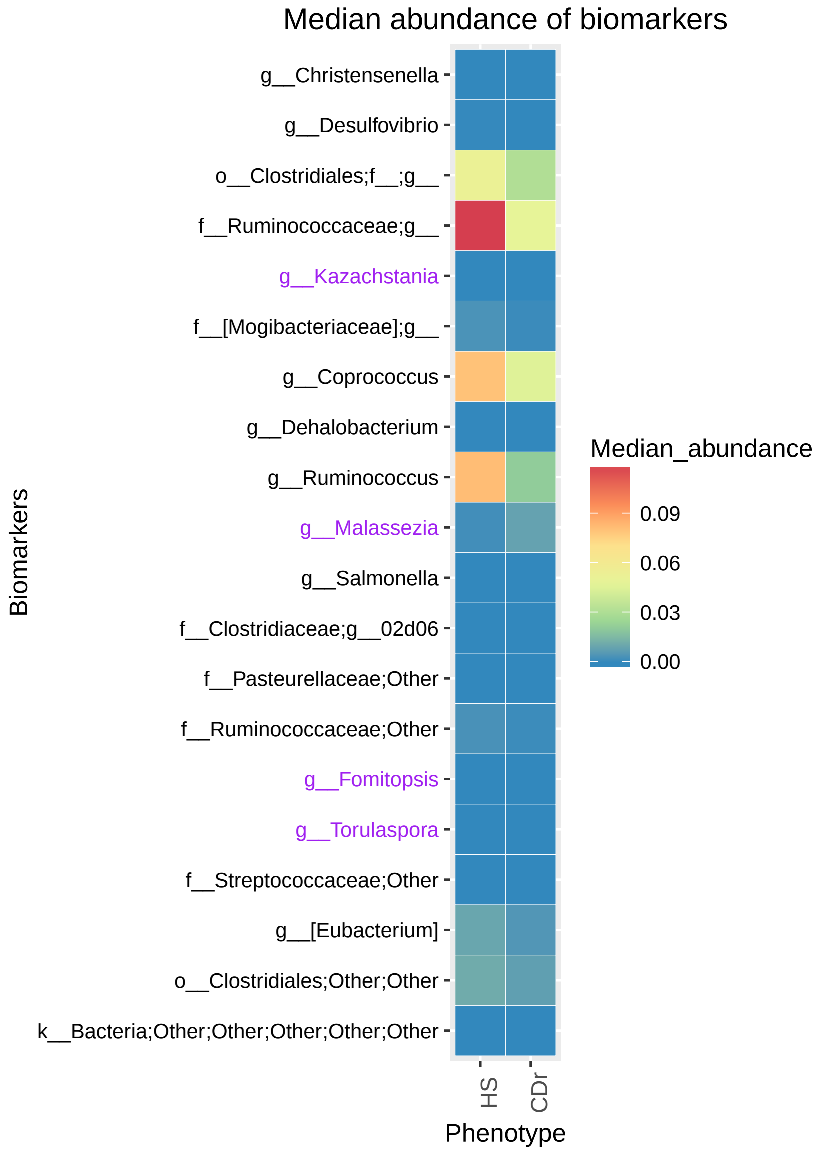


Figure C.2. Heatmap of median abundance of biomarkers for healthy vs. Crohn’s disease patients in remission. Black: bacteria; purple: fungi.

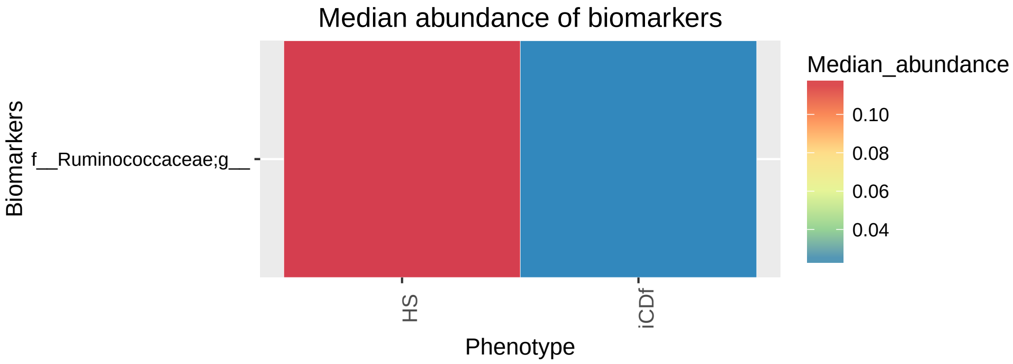


Figure C.3. Heatmap of median abundance of biomarkers for healthy vs. ileal Crohn's disease patients in flare.

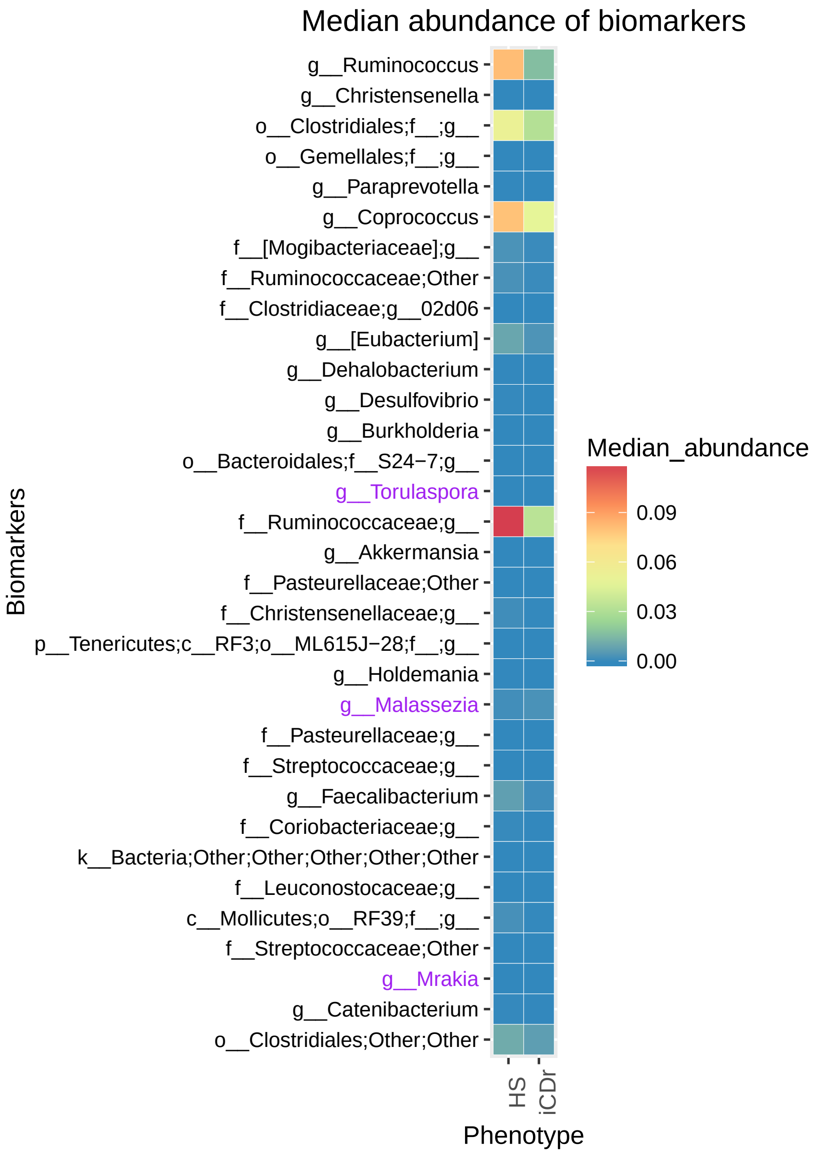


Figure C.4. Heatmap of median abundance of biomarkers for healthy vs. ileal Crohn's disease patients in remission. Black: bacteria; purple: fungi.

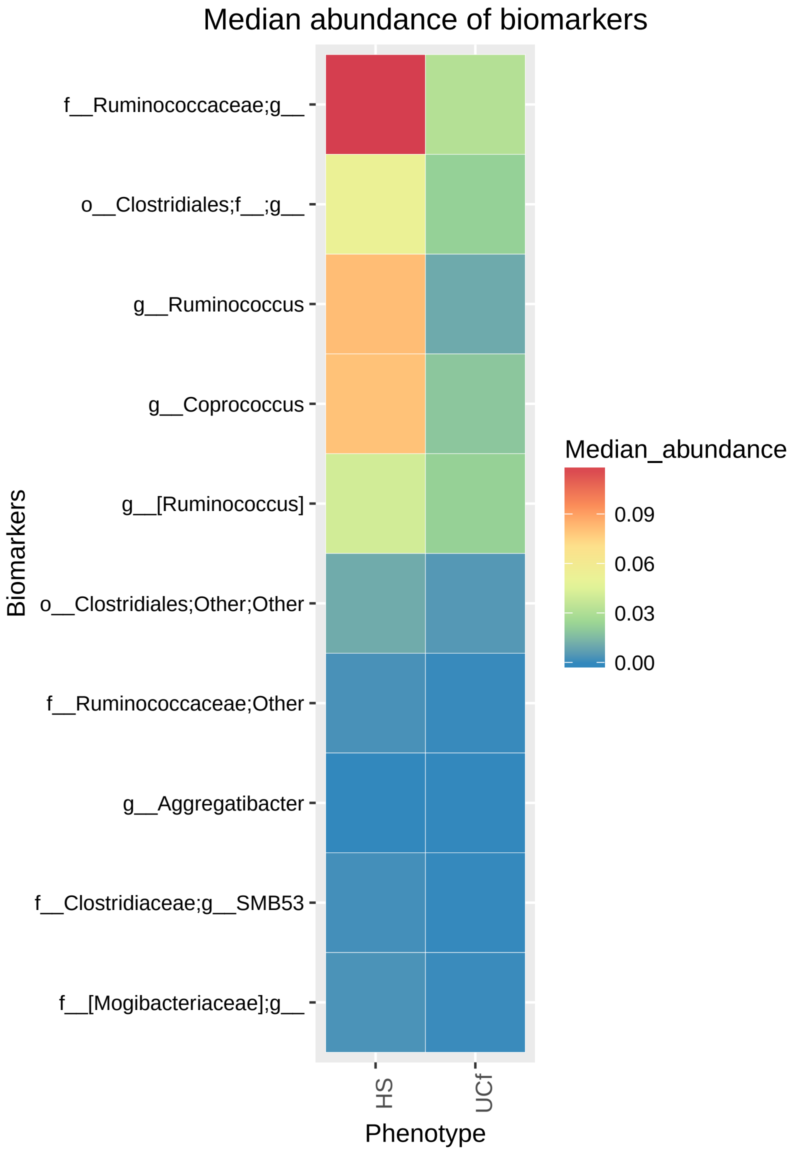


Figure C.5. Heatmap of median abundance of biomarkers for healthy vs. Ulcerative Colitis patients in flare.

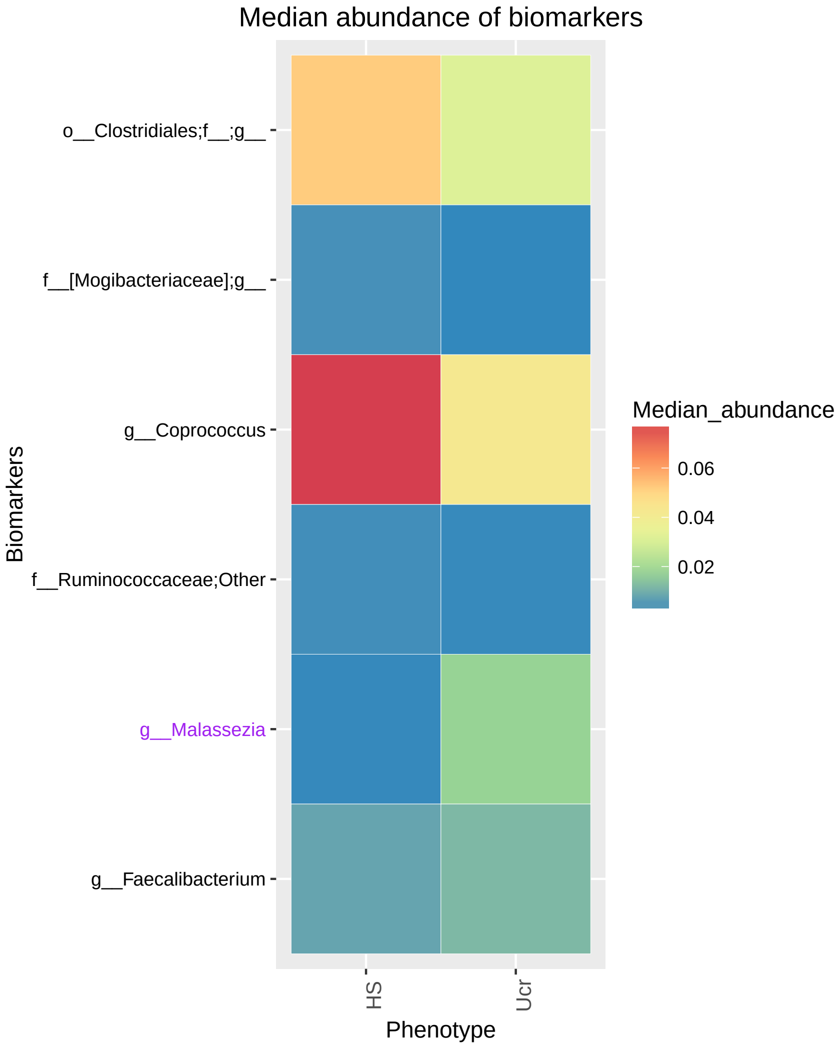


Figure C.6. Heatmap of median abundance of biomarkers for healthy vs. Ulcerative Colitis patients in remission. Black: bacteria; purple: fungi.

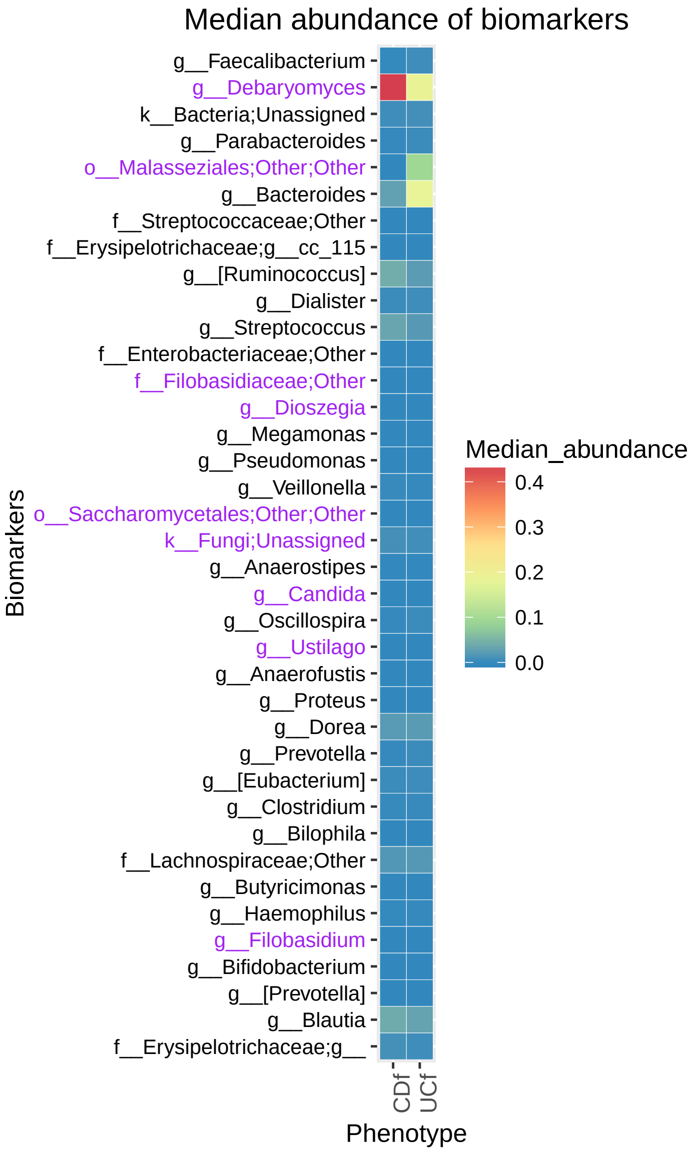


Figure C.7. Heatmap of median abundance of biomarkers for Crohn vs. Ulcerative Colitis patients in flare. Black: bacteria; purple: fungi.

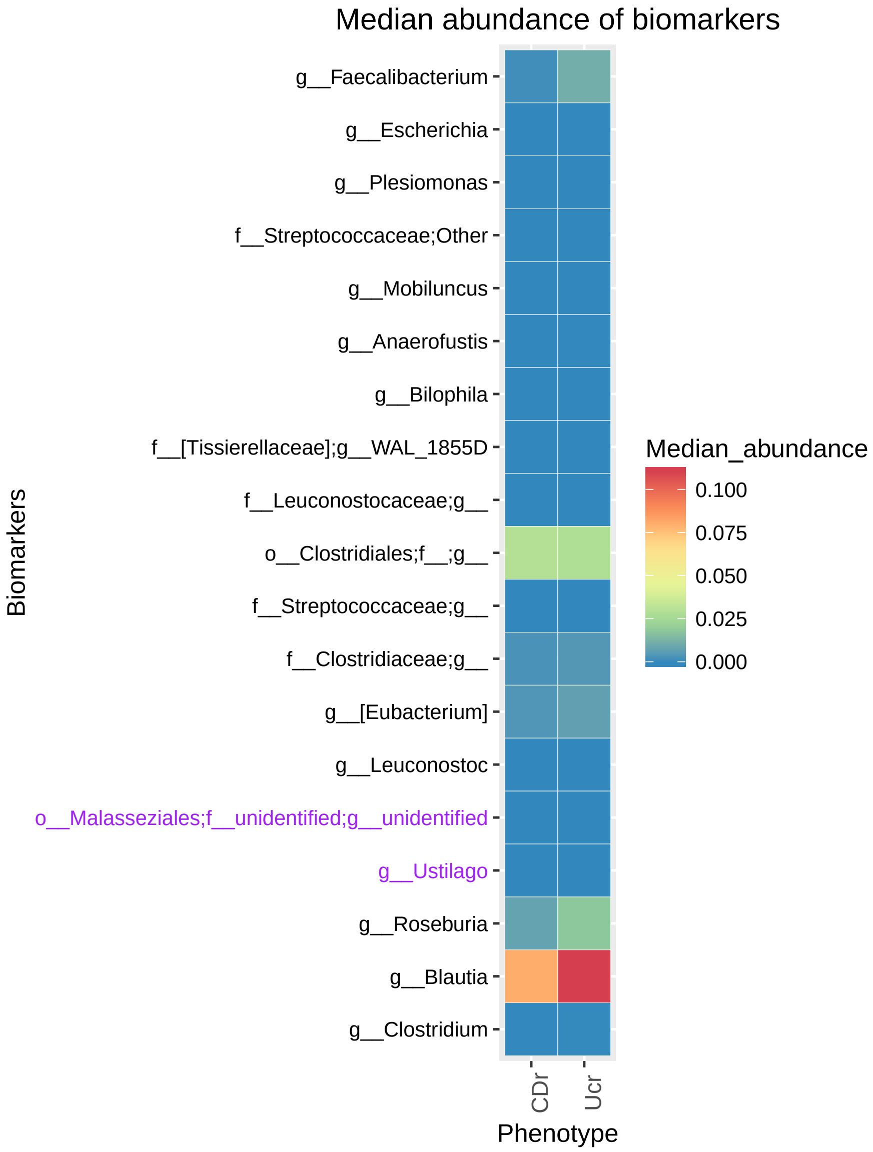


Figure C.8. Heatmap of median abundance of biomarkers for Crohn vs. Ulcerative Colitis patients in remission. Black: bacteria; purple: fungi.

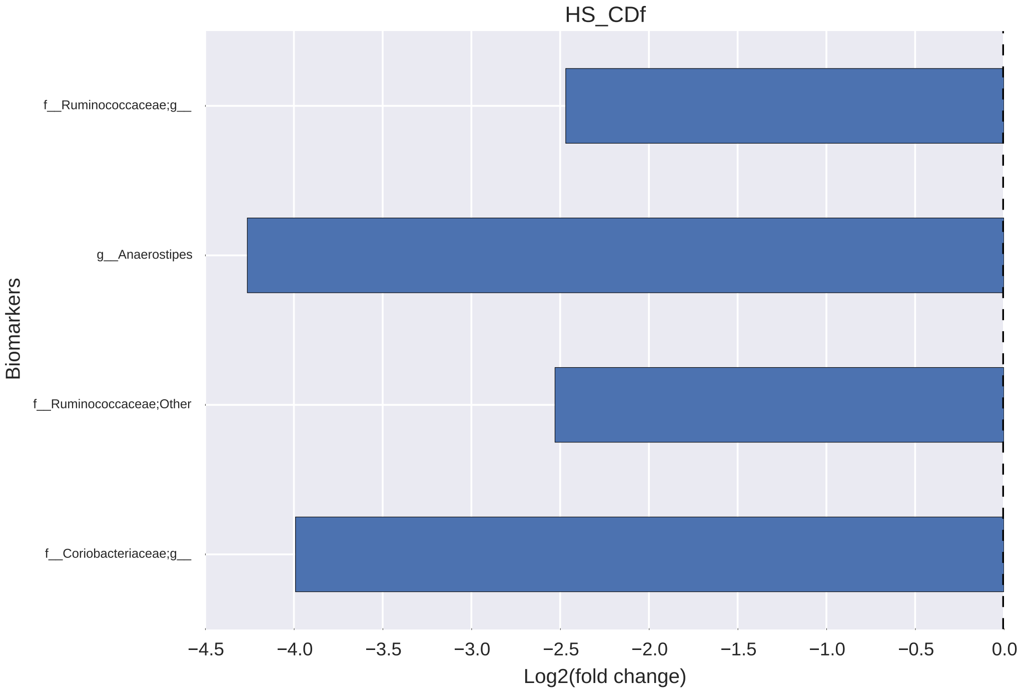


Figure C.9. Log base 2 of [median abundance in organisms increased in Crohn's disease patients in flare] over [median abundance of organisms increased in healthy subjects].

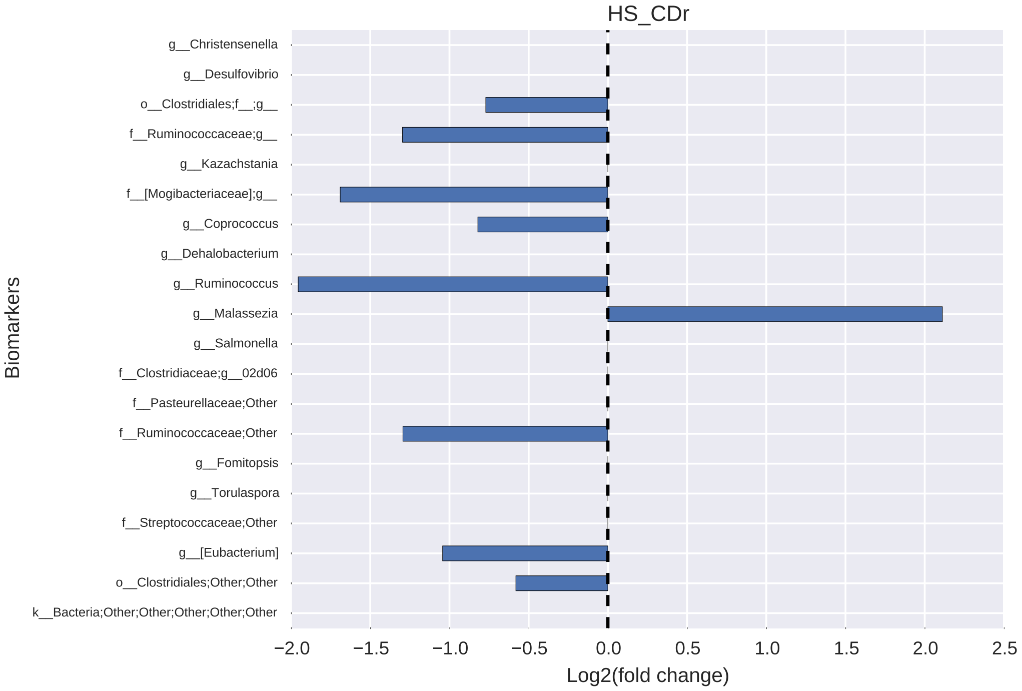


Figure C.10. Log base 2 of [median abundance in organisms increased in Crohn’s disease patients in remission] over [median abundance of organisms increased in healthy subjects].

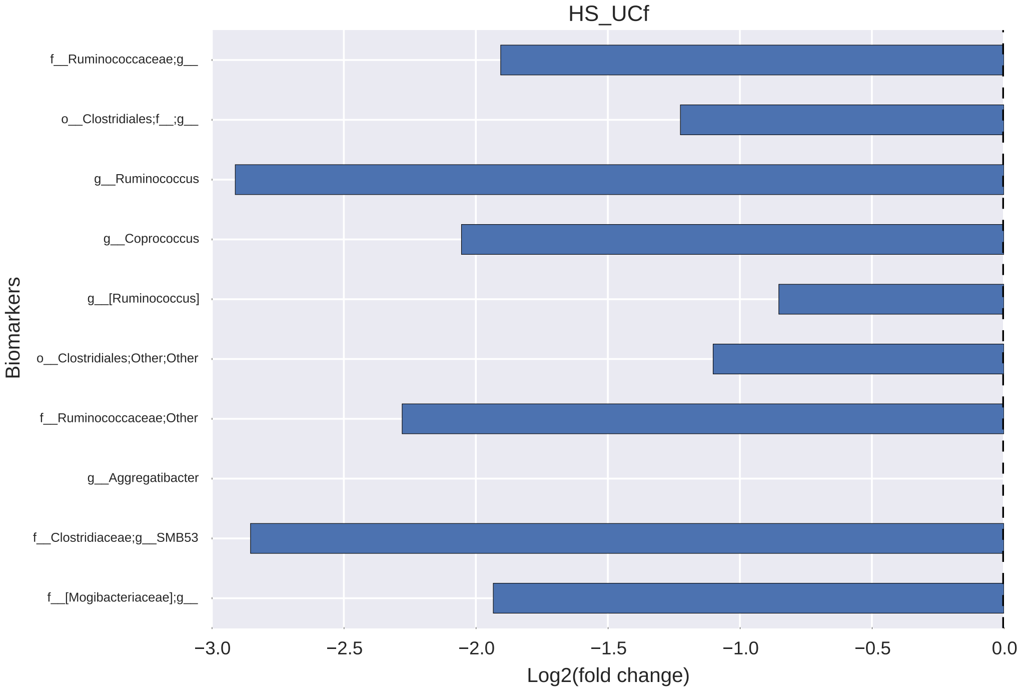


Figure C.11. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in flare] over [median abundance of organisms increased in healthy subjects].

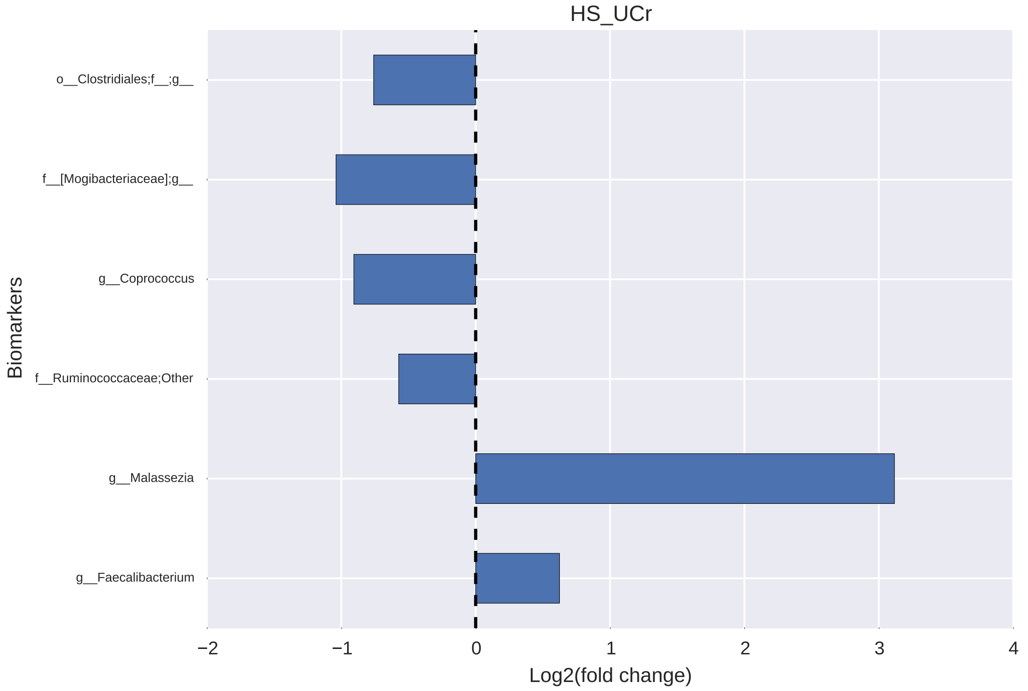


Figure C.12. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in remission] over [median abundance of organisms increased in healthy subjects].

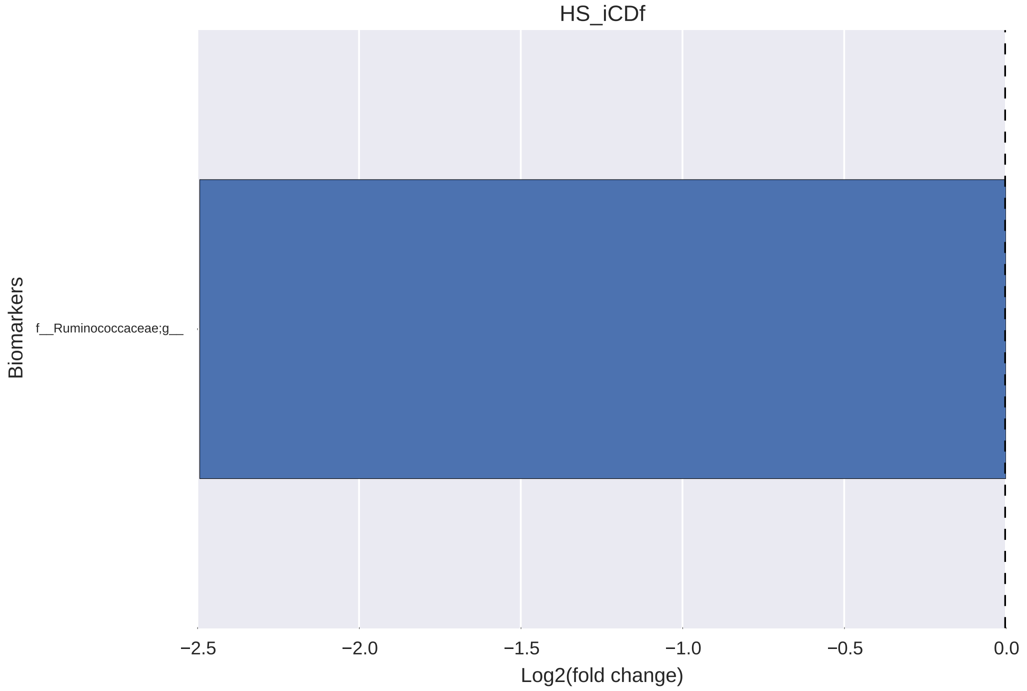


Figure C.13. Log base 2 of [median abundance in organisms increased in ileal Crohn's disease patients in flare] over [median abundance of organisms increased in healthy subjects].

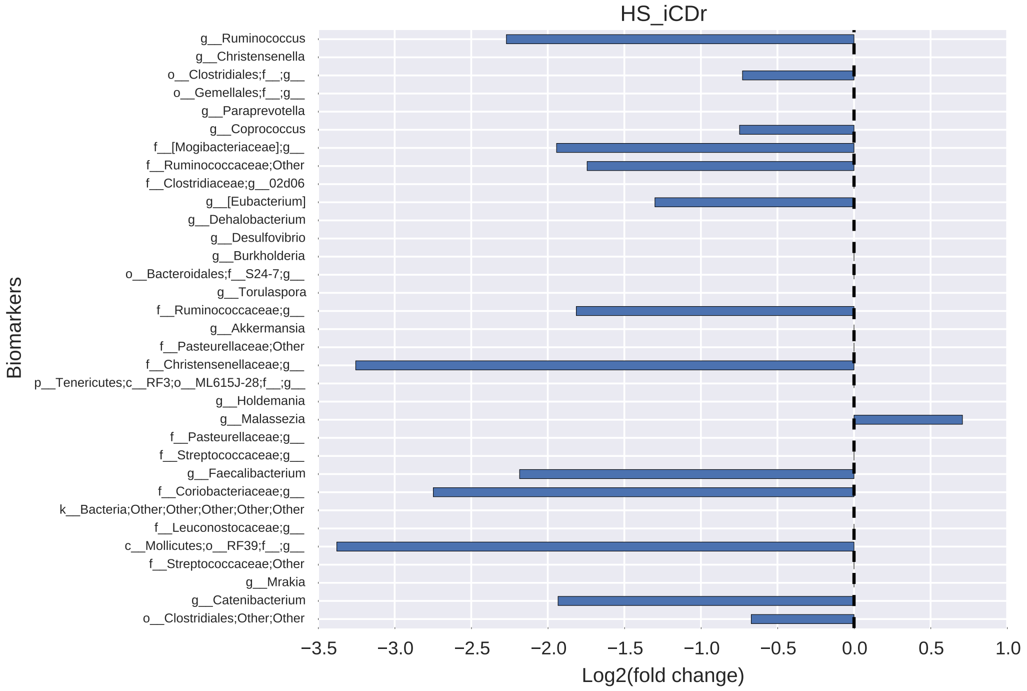


Figure C.14. Log base 2 of [median abundance in organisms increased in ileal Crohn's disease patients in remission] over [median abundance of organisms increased in healthy subjects].

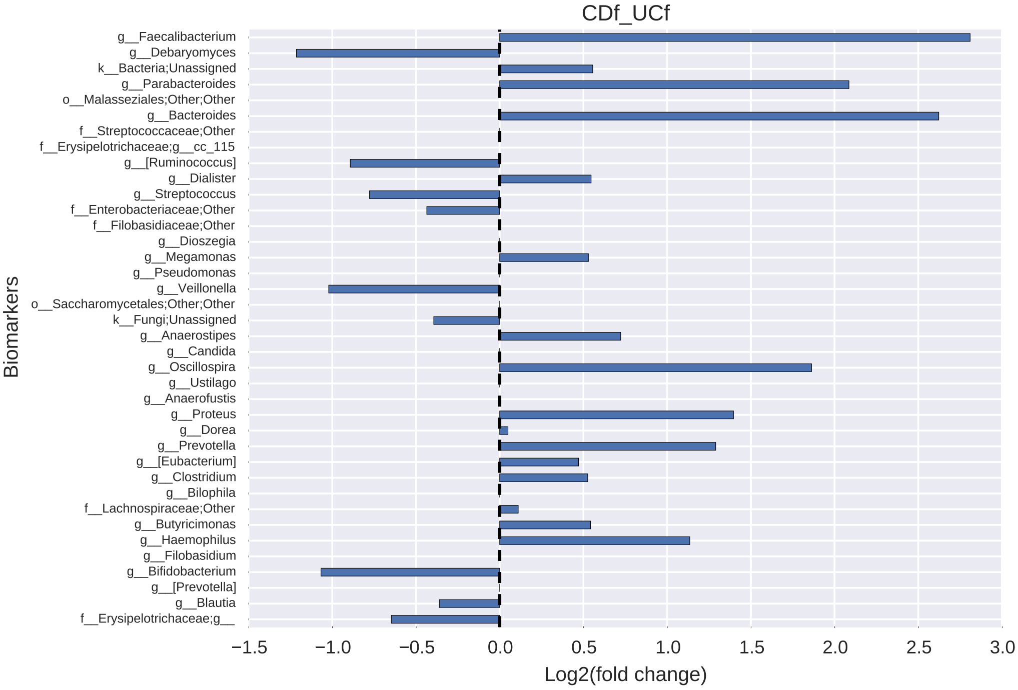


Figure C.15. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in flare] over [median abundance of organisms increased in Crohn's disease patients in flare].

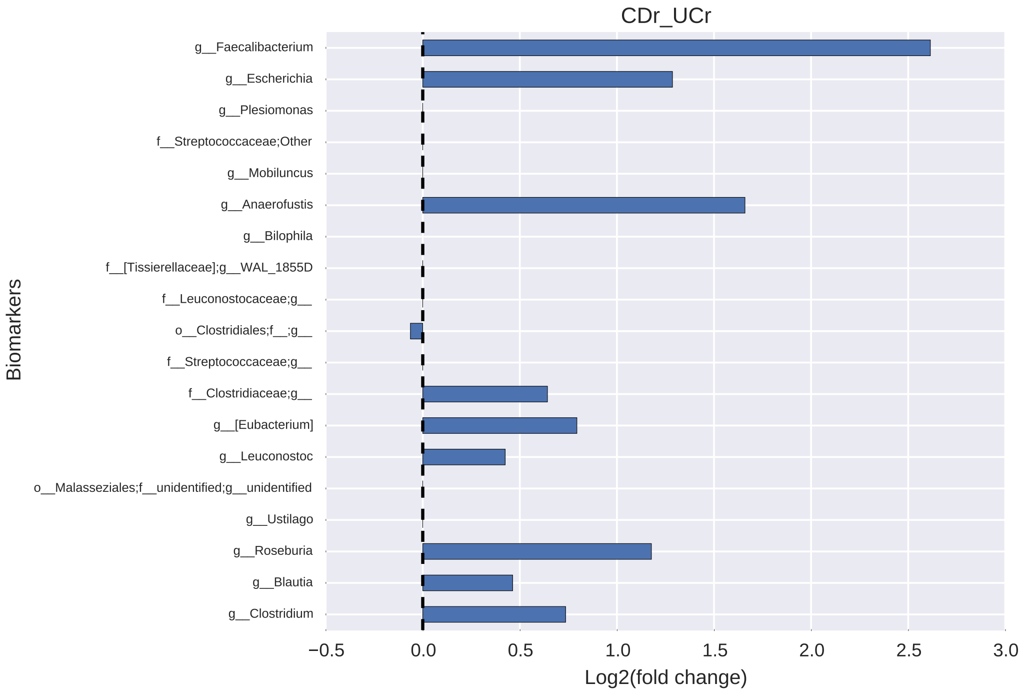


Figure C.16. Log base 2 of [median abundance in organisms increased in Ulcerative Colitis patients in remission] over [median abundance of organisms increased in Crohn's disease patients in remission].

List of Figures

List of Tables