IMT School for Advanced Studies, Lucca
Lucca, Italy

Network analysis of a complex disease: the gut microbiota in the inflammatory bowel disease case

PhD Program in Systems Science

Track in Computer Science and Systems Engineering

XXXIII Cycle

By

Mirko Hu

2022

The dissertation of Mirko Hu is approved.

PhD Program Coordinator: Prof. Rocco De Nicola, IMT School for Advanced Studies Lucca

Advisor: Prof. Tommaso Gili, IMT School for Advanced Studies Lucca

Co-Advisor: Prof. Guido Caldarelli, Ca' Foscari University of Venice

The dissertation of Mirko Hu has been reviewed by:

Prof. Paola Paci, Sapienza University of Rome

Dr. Giovanni Petri, ISI Foundation Institute for Scientific Interchange

IMT School for Advanced Studies Lucca
2022

To my family

Contents

List of Figures ix

List of Tables xvii

Listing xviii

Acknowledgements xix

Vita and Publications xx

Abstract xxiii

1 Introduction 1

2 The Gut Microbiota 6

3 Methods 13

3.1 General information 13

3.1.1 Recruitment 13

3.1.2 Database 14

3.2 Single-layer analysis of IBD 16

3.2.1 Gene Co-expression Network 18

3.2.2 Differential networking 22

3.2.3 Centrality measures in single-layer networks 24

3.2.4 Community detection in single-layer networks . . . 25

3.3 Bipartite network analysis of IBD 27

3.3.1 Binarisation 28

3.3.2 Randomisation of bipartite networks 28
3.3.3 Properties of bipartite networks 29
3.4 Multi-layer network analysis of IBD 32
3.4.1 Multiplex networks 33
3.4.2 Centrality measures in multiplex networks 37
3.4.3 Community detection in multiplex networks 38

4 Results **40**
4.1 Single-layer correlation network analysis results 40
4.1.1 Prevalent pathway correlation network 40
4.1.2 Common pathway correlation network 46
4.1.3 Uncommon pathway correlation network 56
4.2 Bipartite network analysis results 65
4.2.1 Projection on prevalent pathways 65
4.2.2 Projection on common pathways 67
4.2.3 Projection on uncommon pathways 70
4.3 Multilayer analysis results 73

5 Discussion **90**
5.1 Correlation networks 90
5.2 Projected networks 96
5.3 Multilayer networks 99

6 Conclusion **102**
6.1 Future directions 103

A Nodes Map **104**

List of Figures

1 Examples of taxonomic classification of the 90% composition of the bacteria of the gut microbiota. Adapted from Rinninella et al. (2019) 8

2 Representation of a single layer undirected network and its adjacency matrix. 16

3 Visualisation of the three centrality measures on the same network: a) degree centrality, b) betweenness centrality, and c) eigenvector centrality. 24

4 Representation of a bipartite graph. 27

5 Some of the properties of a bipartite network. On the left, motifs and, on the right, monopartite projections. 29

6 Representation of a multilayer network. 32

7 Example of a supra-adjacency matrix of a multilayer network with three layers and three nodes on each layer. The red blocks on the diagonal describe the intralayer edges, whereas the off-diagonal blocks describe the interlayer edges. If the network is undirected, the matrix is symmetric. . . . 34

8 Correlation networks of prevalent metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Nodes are proportional to the betweenness centralities. 41

9 Correlation network of the prevalent pathways in the CD metagenome, the red nodes depict the DE pathways between CD and NI subjects, and the light red nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are not differential nor in the DIAMOnD found group. 45

10 Different centralities measures with p-value < 0.05 for prevalent pathways comparing nodes in the CD correlation network and NI correlation network. Figure (**a**) compares the eigenvector centralities of the DE nodes calculated in the CD correlation network (red) with the eigenvector centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the eigenvector centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the eigenvector centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the eigenvector centralities in Figure (**a**) and Figure (**b**), respectively. 47

11 Different centralities measures with p-value < 0.05 for prevalent pathways comparing nodes in CD correlation network and NI correlation network. Figure (**a**) compares the degree centralities of the DE nodes calculated in the CD correlation network (red) with the degree centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the degree centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the degree centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the degree centralities in Figure (**a**) and Figure (**b**), respectively. 48

12 Correlation networks of common metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Nodes are proportional to the betweenness centralities. 49

13 Correlation network of the common pathways in the CD metagenome, the red nodes depict the DE pathways between CD and NI subjects, the light red nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are not differential nor in the DIAMOnD found group. 50

14 Different centralities measures with p-value < 0.05 for common pathways comparing nodes in CD correlation network and NI correlation network. Figure (**a**) shows the betweenness centralities of the DE nodes calculated in the CD correlation network (red) compared to the betweenness centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the degree centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the degree centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measure in Figure (**a**) and Figure (**b**), respectively. . 54

15 Correlation network of the common pathways in the UC metagenome, the yellow nodes depict the DE pathways between UC and NI subjects, and the light yellow nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are not differential nor in the DIAMOnD found group. 55

16 Different centralities measures with p-value < 0.05 for common pathways comparing nodes in UC correlation network and NI correlation network. Figure (**a**) shows the eigenvector centralities of the DE nodes calculated in the UC correlation network (yellow) compared to the eigenvector centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the degree centralities calculated for the nodes previously found through the DIAMOnD algorithm in the UC correlation network (yellow) with the degree centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measures in Figure (**a**) and Figure (**b**), respectively. 57

17 Correlation networks of uncommon metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Nodes are proportional to the betweenness centralities. . . 58

18 Correlation network of the uncommon pathways in the CD metagenome, the yellow nodes depict the DE pathways between CD and NI subjects, and the light yellow nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are the nodes that are neither differential nor DIAMOnD generated. 60

19 Correlation network of the uncommon pathways in the UC metagenome, the yellow nodes depict the DE pathways between UC and NI subjects, and the light yellow nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are the nodes that are neither differential nor DIAMOnD generated. 61

20 Different centralities measures with p-value < 0.05 for uncommon pathways comparing nodes in CD correlation network and NI correlation network. Figure (**a**) shows the eigenvector centralities of the DE nodes calculated in the CD correlation network (red) compared to the eigenvector centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the betweenness centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the betweenness centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measure in Figure (**a**) and Figure (**b**), respectively. 63

21 Different centralities measures with p-value < 0.05 for uncommon pathways comparing nodes in UC correlation network and NI correlation network. Figure (**a**) shows the degree centralities of the DE nodes calculated in the UC correlation network (yellow) compared to the degree centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the betweenness centralities calculated for the nodes previously found through the DIAMOnD algorithm in the UC correlation network (yellow) with the betweenness centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measure in Figure (**a**) and Figure (**b**), respectively. 64

22 Projected network representation of prevalent metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Networks are obtained through a bipartite projection, and the exclusion of an edge between two nodes is made through a comparison with a null model. Nodes are proportional to the betweenness centralities. 66

23 Projected network representation of common metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Networks are obtained through a bipartite projection, and the exclusion of an edge between two nodes is made through a comparison with a null model. Nodes are proportional to the betweenness centralities. 68

24 Projected network representation of uncommon metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Networks are obtained through a bipartite projection, and the exclusion of an edge between two nodes is made through a comparison with a null model ($\alpha < 0.05$). Node sizes are proportional to the betweenness centrality. 71

25 Colorbar plots for prevalent pathways comparing the degree of the different layers. In Figure (**a**), the first colorbar plot represented the weighted degree of the nodes in the overlap networks compared to the degree of the same nodes in the different layers. In Figure (**b**), the second colorbar plot represented the degree of the nodes of the aggregated networks compared to the degree of the same nodes in the different layers. In Figure (**c**), the heatmap shows the Kendall correlation of the degree sequences in the various networks (on the x-axis from left to right: overlap, aggregated, CD, UC, NI). 74

26 Comparison of the betweenness centrality distributions of the three layers (top to bottom: NI, CD, UC) for prevalent pathways. 75

27 Plots related to the betweenness centrality for prevalent pathways. In Figure (**a**) we analysed the centrality of the nodes across the three layers. In Figure (**b**), we explore in a more detailed way the betweenness centralities of the nodes in the different diagnoses. In Figures (**c**) and (**d**), we analysed the correlation of the multilayer centralities and single layer centralities. 76

28 Analysis of the community detection in the multiplex network composed of NI, CD, and UC layers for a range of different resolution parameters $\gamma = [0, 1]$ on the x-axis and three different coupling parameters $\omega = (0, 0.1, 1)$ in the case of prevalent pathways. 77
29 Comparison of the community detection with Infomap in the soft partitioning case (**a**) and in the hard partitioning case (**b**), for prevalent pathways. 78
30 Colorbar plots for common pathways comparing the degree of the different layers. In Figure (**a**), the first colorbar plot represented the weighted degree of the nodes in the overlap networks compared to the degree of the same nodes in the different layers. In Figure (**b**), the second colorbar plot represented the degree of the nodes of the aggregated networks compared to the degree of the same nodes in the different layers. In Figure (**c**), the heatmap shows the Kendall correlation of the degree sequences in the various networks (on the x-axis from left to right: overlap, aggregated, CD, UC, NI). 79
31 Comparison of the betweenness centrality distributions of the three layers (top to bottom: NI, CD, UC) for common pathways. 80
32 Plots related to the betweenness centrality for common pathways. In Figure (**a**) we analysed the centrality of the nodes across the three layers. In Figure (**b**), we explore in a more detailed way the betweenness centralities of the nodes in the different diagnoses. In Figures (**c**) and (**d**), we analyzed the correlation of the multilayer centralities and single layer centralities. 81
33 Analysis of the community detection in the multiplex network composed of NI, CD, and UC layers for a range of different resolution parameters $\gamma = [0, 1]$ on the x-axis and three different coupling parameters $\omega = (0, 0.1, 1)$ in the case of common pathways. 82

34 Comparison of the community detection with Infomap in the soft partitioning case (**a**) and in the hard partitioning case (**b**), for common pathways. 83
35 Colorbar plots for uncommon pathways comparing the degree of the different layers. In Figure (**a**), the first colorbar plot represented the weighted degree of the nodes in the overlap networks compared to the degree of the same nodes in the different layers. In Figure (**b**), the second colorbar plot represented the degree of the nodes of the aggregated networks compared to the degree of the same nodes in the different layers. In Figure (**c**), the heatmap shows the Kendall correlation of the degree sequences in the various networks (on the x-axis from left to right: overlap, aggregated, CD, UC, NI). 85
36 Comparison of the betweenness centrality distributions of the three layers (top to bottom: NI, CD, UC) for uncommon pathways. 86
37 Plots related to the betweenness centrality for uncommon pathways. In Figure (**a**) we analysed the centrality of the nodes across the three layers. In Figure (**b**), we explore in a more detailed way the betweenness centralities of the nodes in the different diagnoses. In Figures (**c**) and (**d**), we analyzed the correlation of the multilayer centralities and single layer centralities. 87
38 Analysis of the community detection in the multiplex network composed of NI, CD, and UC layers for a range of different resolution parameters $\gamma = [0, 1]$ on the x-axis and three different coupling parameters $\omega = (0, 0.1, 1)$ in the case of uncommon pathways. 88
39 Comparison of the community detection with Infomap in the soft partitioning case (**a**) and in the hard partitioning case (**b**), for uncommon pathways. 89

List of Tables

1 Representation of the database containing the gene expression data with example values. 18
2 Comparison of DIAMOnD and SWIM algorithms. 23

3 NMI and p-value for community detection in correlation networks with presence over 75%. 44
4 NMI and p-value for community detection in networks with 50% to 75% presence. 53
5 NMI and p-value for community detection in networks with presence between 25% and 50%. 62

6 Nodes map of the nodes being cited in the research project. 104

Listing

1 Extracting and combining the useful data from HMP2 Metadata with HMP2 Metagenomics database 14
2 Python example to reduce the size of the database 19
3 Code to find the percolation threshold. 21

Acknowledgements

For those that studied Mandarin, one of the first curiosities that are taught about Chinese culture is that in this language there is a specific word for 'human relationships', which is learnt right after 'hi' or 'how are you?', and this hard to translate word is guānxì. By studying Network Science, I have learnt that guānxì is not only about 'relationships' that seem to be exclusive one-to-one connections; guānxì is something closer to a network behaviour, in which willingly or unwillingly, we are all intrinsically connected. Long story short, I would thank everyone, by saying that I am very grateful to my Network, on the other hand, I feel the duty to fill the whole acknowledgements page. So, firstly, I would like to thank Guido, who introduced me to Network Science and has been guiding me through my studies since I found one of his books on this topic during my previous studies in Imperial College library. Secondly, I would like to thank Tommaso, who has been encouraging me since the first time he accepted me as his student. Thirdly, I would like to thank the whole Network Unit, which has been like a second family. Finally, I would like to thank the whole IMT community, which has been my extended family. Last but not least, I have to deeply thank my parents, my sister Alice and my friend Giacomo for their support and their encouragement. This work is dedicated to all of you.

Vita

August 30, 1992	Milan, Lombardy, Italy
2013	Double BSc degree project "PoliTong" in Information Technology Engineering Tongji University, Shanghai, China
2014	BSc in Automation and Control Engineering Final mark: 108/110 Politecnico di Milano, Milan, Italy
2017	MSc in Advanced Chemical Engineering with Process Systems Engineering Final mark: Pass with Distinction Imperial College London, London, UK

Publications

1. J. C. W. Billings, M. Hu, G. Lerda, A. N. Medvedev, F. Mottes, A. Onicas, A. Santoro, and G. Petri (2019). "Simplex2Vec embeddings for community detection in simplicial complexes". *arXiv preprint arXiv:1906.09068*.
2. M. Hu, G. Triulzi, and M. Sharifzadeh (2020). "Technological change in fuel cell technologies". *Design and Operation of Solid Oxide Fuel Cells*, pp. 3-41. Academic Press.

Presentations

1. M. Hu, "Network theory applications: from patent networks to biological networks" at *University of Parma*, Parma, Italy, 2022.

Abstract

The gut microbiota contains hundreds of types of microbes and dysbiosis can lead to inflammatory bowel diseases (IBD), which comprise Crohn's disease (CD) and ulcerative colitis (UC). Due to the complex nature of the IBD, it is interesting to understand the differences between a control (NI) and an IBD gut microbiome by using new tools offered by network science. In particular, when metagenomic data are considered, it is possible to build networks according to the covariance, the co-occurrence and multiple layers of networks (multilayer networks). In addition to the construction of the networks, an analysis of the differential expressed pathways is carried out, several centrality measures are calculated, and community detection is performed to explore the topological differences between the diagnosis networks. The analysis of the correlation network topology highlights that, in IBD networks, the pathway involving coenzyme A of the unclassified species becomes central. Furthermore, the modularity in the IBD networks is higher. In both the correlation network and the co-occurrence network, the modules belonging to B. ovatus and B. caccae are positioned differently in each diagnosis. Furthermore, the difference between the NI and the UC diagnosis networks lies in a change in the wiring that preserves the centralities. Moreover, the fundamental role of two of the Roseburia species in the NI is evidenced. A further step will consist of identifying the minimum number of pathways on which it would be ideal to intervene to drive the system back to a healthy state by the precision medicine way of operating.

Chapter 1

Introduction

> If you don't like bacteria, you are on the wrong planet.
>
> Brand (2010)

Microbes are ubiquitous. From radioactive waste to the human gastrointestinal tract, they can be found nearly everywhere. In and on the human body, they have evolved to co-exist with their host and it is estimated that the amount of microbes hosted by the human body is of the same order of magnitude as the number of human cells (Sender, Fuchs, and Milo, 2016). In particular, the 10^{14} commensal microbes living in the intestinal tract form the human gut microbiota, which has evolved to live in symbiosis with its host (Thursby and Juge, 2017). It is widely accepted that this symbiosis begins from birth and the microbial communities stabilize with age until the formation of an adult microbiota (Tanaka and Nakayama, 2017). Its genetic content (called the microbiome) characterizes each individual, rising also concerns about identity and privacy issues, specifically when the study and the manipulation of the microbiota are considered (Yonghui Ma et al., 2018). Since the 1840s, when the concept of gut microbiota first appeared, the topic has been studied for two centuries (Farré-Maduell and Casals-Pascual, 2019), and, at the

moment, it is known that the gut microbiota has a fundamental role in shaping the gut barriers (Natividad and Verdu, 2013), training the host immune system and regulating the metabolism (Sommer and Bäckhed, 2013). When the compositional and metabolic equilibrium of the commensal microbes living in the gut is disrupted, different types of diseases arise such as metabolic disorders or central nervous system disorders (Belizário and Faintuch, 2018). Historically, traditional medicine attempted to re-establish this equilibrium through remedies intervening on the digestive system, like fasting, diets, assumption of spring waters or laxatives. A quite recent procedure introduced to tackle the *C. difficile* infection is the faecal microbiota transplantation (FMT) (Borody and Khoruts, 2012) which consists in repopulating the intestinal tract of an ill subject with the microbiota of a healthy donor.

Inflammatory bowel diseases (IBDs), which comprise Crohn's disease (CD) and ulcerative colitis (UC), are an important class of diseases that arises from dysbiosis and are being treated with FMT. Typical symptoms of this class of diseases are chronic diarrhoea, abdominal pain, rectal bleeding, weight loss and fatigue (Singh et al., 2011). Although CD and UC are both characterized by the inflammation of the intestinal tract, there are several differences between the two diagnoses that span from the environmental factors that cause them, e.g. smoking or diet, to the clinical and endoscopic findings in the two diagnoses (Ananthakrishnan, Xavier, and Daniel K Podolsky, 2017). Overall, IBDs are becoming widespread in modern society because of the change in lifestyle, socioeconomic developments and environmental causes (Manichanh et al., 2012a). Until now, it is known that IBD is an exaggerated immune response to the gut microbiota of genetically predisposed subjects under the influence of the external environment. This complex interplay between genetics, the microbiota, the immune system and the environment makes it particularly hard to understand this class of diseases. Kirsner (1988) offered a complete historical review of the IBD until the 1980s, by quoting Hippocrates who described diarrhoea as a symptom of an infectious (or non-infectious) disease to a description of hypothetical pathogenesis of IBD, which the microbiota was not considered, though. A

more recent projection predicted the evolution of the disease between 2015 and 2025 and updated the possible origins of IBD including the action of antibiotics on the gut microbiota in Western society (Kaplan, 2015). Xavier and D K Podolsky (2007) summarized the findings of the origins of the IBD, mentioning the complexity of the disease. Another historical review focuses on the genetics of the IBD (Lees and Satsangi, 2009) identified *NOD2* as the first CD susceptible gene and then described the evolution of the IBD genetics with the coming of the modern genome-wide association study. One of the first most comprehensive work describing the interaction of all the aforementioned factors can be found in Y.-Z. Zhang and Y.-Y. Li (2014). Whereas, the systems biology approach to the study of IBD was presented by Fiocchi and Iliopoulos (2021), which proposed the creation of an IBD interactome, a complex system connecting all the potential agents interacting among them that derived from the combination of different omics (De Souza, Fiocchi, and Iliopoulos, 2017).

Our work starts from here and attempts to provide tools and methods from network science useful to build and to study the IBD interactome with a systems biology approach by commencing from the metagenomic data of the gut microbiome. This approach is typical of network medicine, a novel discipline that mixes network science with systems biology to tackle the challenges offered by the progress of personalized medicine (Barabási, Gulbahce, and Loscalzo, 2011), which opposes to the current effective yet drastic procedures like the aforementioned FMT. Network science is the discipline used to analyse complex systems and could be suited to understand a complex disease like IBD in which a complex system like the gut microbiota plays a fundamental role. Complexity in the intestinal microbial communities arises at different scales; from the macroscopic point of view, we have the ecological interactions (Faust et al., 2018; Bucci et al., 2016) that describe the relationships among the species in the gut microbiota; among these, we have three main different types of interactions (Coyte and Rakoff-Nahoum, 2019); positive interactions (cooperation, commensalism, cross-feeding), negative interactions (competition, ammensalism), and asymmetric interactions (exploitation,

predation, parasitism). Going towards a more microscopic scale, we can find the gene networks, often represented by gene co-expression networks (Vernocchi et al., 2020) and metabolic networks built by connecting the substances, known as metabolites, reacting in the same metabolic processes (Bauer and Thiele, 2018). When the interaction between bacterial proteins is considered, we are dealing with metaproteomics, which is a novel tool in the analysis of IBD; nonetheless, the data used is still scarce (Segal et al., 2019).

The application of network science for the study of the complexity of the gut microbiome is recent and one of the first research was in the case of *C. difficile* infection (Stein et al., 2013). The microbiome in this work was represented as a boolean network derived from binarized temporal data of the abundance of specific bacteria species in the gut. Although the study was able to capture the dynamics of the bacterial species, e.g. negative, positive or neutral interaction, it did not take into account the genetic expression of the microbiome (metagenome), which could explain better the complex interplay between the bacterial species. Our study, by contrast, gives a static screenshot of the microbial interactions through metagenomics. A more recent study (L. Chen et al., 2020) analysed the co-abundance network built with SparCC (Friedman and Alm, 2012), the need of this tool is due to the necessity of sparsifying the network that would have too many correlated nodes because of normalization and a p-value threshold too high (Weiss et al., 2016). Based on a topological property of the biological networks, the work by Vernocchi et al. (2020) portrays a weighted gene co-expression network analysis by building a network from metagenomic data and removing the weaker edges based on the assumption that the final network would be scale-free. In our work, we used thresholding methods that rely on the network topology such as the percolation threshold or the p-value for the projected edges, similarly to the later research. These methods should overcome the aforementioned problems.

In the next chapters, we will go through a review of the main concepts useful to understand the biology of the gut microbiota, then we will introduce the basics of network science describing the methods used in our

research after we will present the main results from our work and finally we will contextualize them in the light of the past works.

Chapter 2

The Gut Microbiota

In this chapter, we will briefly go through the microbiological fundamentals useful to shortly understand the results derived from the network analysis of the gut microbiota. We will learn the technical language used to describe the microbial world and explain the main bacterial species in the gut microbiota. After that, we will look at the main concepts behind metagenomics, what 16S rRNA gene sequencing is, and how it helped identify the bacterial species more quickly. Furthermore, we will deal with the foundations of the metabolic pathways present in the gut microbiota and analyse the primary metabolites involved in these processes.

The **microbiota** is the community of the microorganisms, such as bacteria, archaea, fungi, algae, and other eukaryotic organisms living in an environment. It differentiates from the **microbiome** as the latter includes the complete spectrum of the molecules produced by the microorganisms, such as their building blocks, the metabolites and other substances exchanged among the microbes (Berg et al., 2020). In the present work, we will focus on the bacteria that are single-celled microorganisms, also called **prokaryote**. The modern classification of the bacteria was possible thanks to the discovery of the 16S ribosomal RNA (**16S rRNA**), which is a specific section of the prokaryotic ribosome. A ribosome is a macromolecular machine whose function is to build proteins, and it is found inside all the living cells. Minimal changes during the evolutionary pro-

cesses characterise the 16S rRNA. Hence it is the most suitable to classify a bacterium.

The **taxonomic classification** of the bacteria is made by using the following levels (or ranks) in descending order: domain, phylum, class, order, family, genus, and species. The domain represents the highest hierarchical rank, and it could be Bacteria, Archaea, and Eucarya. Whereas, according to Yarza et al. (2014), two strains of bacteria are classified in the same phylum if they share 75% of the 16S rRNA. For the remaining ranks, we have 78.5% for class, 82.0% for order, 86.5% for family, and 94.5% for the genus. The definition of 'species' was given by Rosselló-Mora and Amann (2001) that defined it as a group of organisms that share a number of characteristics and can be differentiated through specific phenotypes.

Firmicutes and *Bacteroidetes* are the most common phyla, and they represent 90% of the bacteria in the gut microbiota. They have opposite **Gram staining**, the former are Gram-positive, whereas the latter are Gram-negative. Gram staining is a way to classify bacteria by the composition of the outer membrane; a pigment (crystal violet) is used to stain the bacterial membrane; if the cell wall retains the pigment and becomes pink or red, it is possible to deduce that they are Gram-positive, on the contrary, they are Gram-negative. There are also a class of bacteria that cannot be classified either way, and they are called gram-neutral.

In the following paragraphs, we will describe the main bacterial species that compose the gut microbiota. It is possible to recognise three main human enterotypes (Arumugam et al., 2011) (enterotypes are a way of classifying a living organism based on the composition of the gut microbiota), and they are *Bacteroides* (enterotype 1), *Prevotella* (enterotype 2), and *Ruminococcus* (enterotype 3). *Bacteroides* is a genus of bacteria characterised by being clinically pathogenic when outside the gut (Wexler, 2007). In the gut, they evolved to be symbiotic with the host. In fact, they help to digest complex polysaccharides (Salyers, Valentine, and Hwa, 1993) into nutrition and vitamins beneficial to both the host and the other members of the gut microbiota. They are Gram-negative obligate anaerobic bacteria, which means that they do not need oxygen to grow.

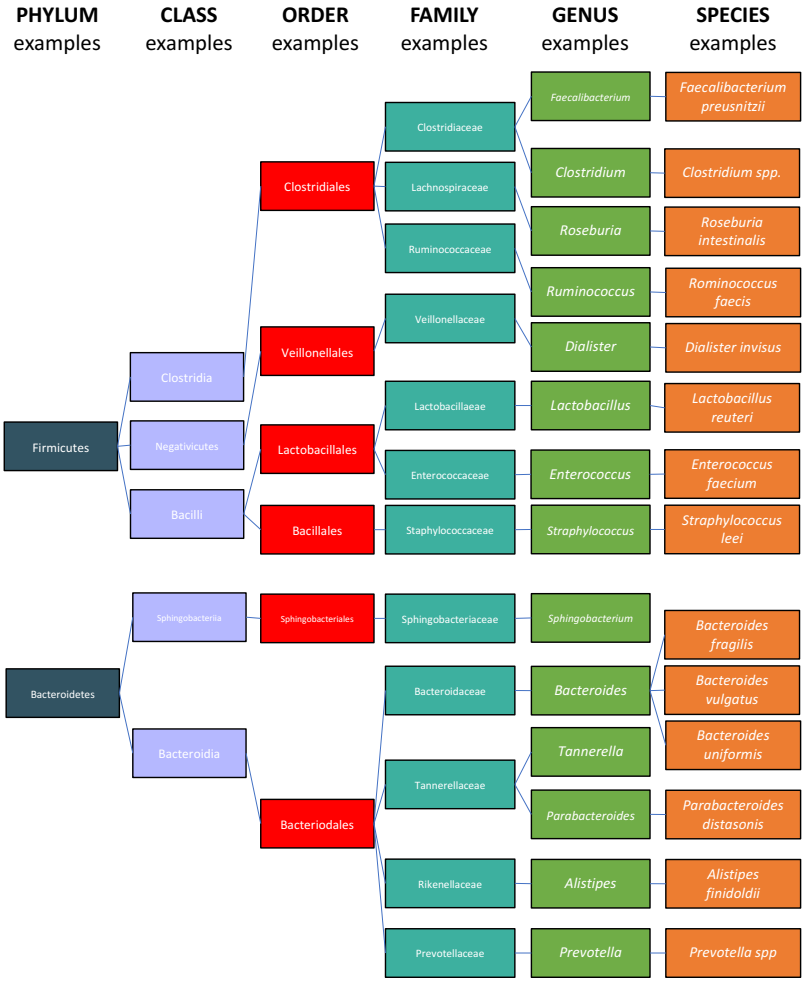


Figure 1: Examples of taxonomic classification of the 90% composition of the bacteria of the gut microbiota. Adapted from Rinninella et al. (2019)

Prevotella, similarly to *Bacteroides*, is a genus of Gram-negative bacteria; differently from the latter, *Prevotella* is prevalent in rural, preagricultural, isolated or vegetarian populations that consume a plant-rich diet, whereas the *Bacteroides* genus is more spread in Westernised populations (Precup and Vodnar, 2019). The presence of *Prevotella* is at the expense of *Bacteroides* and is also associated with chronic inflammatory status in the gut (Ley, 2016). Nevertheless, the two genera shared a common ancestor.

Ruminococcus is a genus of Gram-positive and obligate anaerobic gut bacteria. They were first isolated in ruminants, and they are known to digest cellulose. One of the most cited papers describes the transfer of molecular hydrogen produced through glucose fermentation by *R. albus*.

The **metagenome** is the genome, i.e. the ensemble of the genes of uncultivated microorganisms in an ecosystem that could be from aquatic to terrestrial, but also the human skin or the human gut (Lindgreen, Adair, and Gardner, 2016). A metagenomic analysis requires six fundamental steps (Ladoukakis, Kolisis, and Chatziioannou, 2014), which are:

1. quality control;
2. assembly;
3. gene detection;
4. gene annotation;
5. taxonomic analysis;
6. comparative analysis.

Each step produces a different output that is processed by the data management side. Initially, the process starts from data called reads that are small nucleotide (or base) sequences, which are amplified portions of the same DNA molecules cut into smaller pieces in the process of shotgun sequencing. The generated text files containing the reads (e.g. CGGCCT) are called FASTA or FASTQ. Each read of the text files produced by the initial raw sequencing data is used in the quality control step. Depending on the settings applied, each read is divided into smaller sequences of

different lengths. This procedure is prone to bias; therefore, a score that estimates the quality of each sequenced base (**Phred**) is calculated. In the database used **Trim** are the number of reads that remain after trimming the bases with Phred > 20. The dataset is then filtered by comparing it with a database of human genome sequences (**hg38**). The sequences obtained can be considered high quality reads, and they are ready for the assembly step. From the assembly step, it is possible either to proceed directly with the taxonomic analysis or to go through gene prediction and annotation, which are useful to define their functions.

The main metabolic pathways found in the microbiota are:

- **1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis**: tetrahydrofolate is also known as vitamin B9, folate coenzymes are involved in the acceptance and donation of carbon-based molecules fundamental for the nucleotide synthesis, cellular methylation reactions, and amino acid involving pathways (West, Caudill, and Bailey, 2020). 10-formyl-tetrahydrofolates are used for tetrahydrofolate, purine and formate production (Nagy et al., 1995).

- **ANAGLYCOLYSIS-PWY: glycolysis III (from glucose)**: it is one of the most important pathways for the central metabolism, and it is fundamental for cellular growth. The main difference from glycolysis I is that glucose is phosphorylated inside the cell by the specific enzyme glucokinase.

- **ARO-PWY: chorismate biosynthesis I**: chorismate is a relevant intermediate that leads to the production of some fundamental metabolites, among which vitamins E and K (Bentley and Haslam, 1990).

- COA-PWY-1: coenzyme A biosynthesis II (mammalian) and **COA-PWY: coenzyme A biosynthesis I**: coenzyme A is a fundamental enzyme needed in different processes for all the organisms. It is fundamental for the production of fatty acids (Leonardi et al., 2005).

- **GLYCOGENSYNTH-PWY: glycogen biosynthesis I (from ADP-D-Glucose)**: glycogen is an extremely ramified glucose polymer

that is used as a form of energy storage. Glycogen particles are soluble in water.

- **DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I**: one of the metabolites in the pathways for dTDP-L-rhamnose biosynthesis is needed by many pathogenic bacteria, in particular the *Streptococci* (Beek et al., 2019).

- **GLCMANNANAUT-PWY: superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation**: N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate are also called amino sugars, and they are used to build the wall of the cells in *E. coli*. Moreover, these metabolites can be used as sources of carbon and nitrogen (Plumbridge and Vimr, 1999).

- **PWY-5097: L-lysine biosynthesis VI**: lysine is a key metabolite used to build proteins in humans and animals, (Wu, 2009) and depletion of this fundamental building block was found to be detrimental to the absorption of the nutrients. Therefore, it was found that different lysine levels influenced the metagenome on piglet models (Yin et al., 2017).

- **PWY-5659: GDP-mannose biosynthesis**: free mannose does not have the same role in all the organisms. In some cases, it can be useful to treat illnesses, and in other cases, like honeybees, it can kill the organism (Sharma, Ichikawa, and Freeze, 2014). Nevertheless, the metabolite produced in this pathway is fundamental for the process of glycoprotein formation.

- **PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I**: 5-aminoimidazole ribonucleotide is an intermediate metabolite fundamental for the production of purine nucleotides and thiamine. The pathway was studied in several bacteria.

- **PWY-6123: inosine-5′-phosphate biosynthesis I**: also known as inosinic acid, this metabolite is fundamental in the formation of

purines, which are used to synthesise nucleic acid constantly (Y. Zhang, Morar, and Ealick, 2008).

- **PWY-7111: pyruvate fermentation to isobutanol (engineered)**: pyruvate can be used as a substrate for growth by many bacteria (Wagner et al., 2005). In this case the final product of the fermentation is isobutanol, such as in Atsumi et al. (2010).
- **PWY-7219: adenosine ribonucleotides de novo biosynthesis**: adenosine ribonucleotides (AMP) is one of the products of the conversion of IMP.
- **VALSYN-PWY: L-valine biosynthesis**: L-valine is an amino acid used for synthesising other proteins; hence it has the property of inducing muscle growth and tissue repair.

Chapter 3

Methods

3.1 General information

3.1.1 Recruitment

The IBDMDB (*IBDMDB - Home | IBDMDB* n.d.) is one of the first comprehensive studies of the gut ecosystem's multiple molecular properties involved in the IBD dynamics. Some of the microbiome measurements offered by the study are metagenomics, metatranscriptomics and metabolomics. The data is related to 132 subjects approached in five medical centres: Cincinnati Children's Hospital, Emory University Hospital, Massachusetts General Hospital, Massachusetts General Hospital for Children, and Cedars-Sinai Medical Centre. The patients recruited for the study initially arrived at the hospitals either for routine age-related colorectal cancer screening, presence of other gastrointestinal symptoms, or suspected IBD. The latter could be a consequence of positive imaging, symptoms of chronic diarrhoea or rectal bleeding. A preliminary colonoscopy was performed to determine study strata if there were no conditions for exclusion right after enrolment. Based on initial analyses, the subjects not diagnosed with IBD were labelled as 'NI' controls. This group of subjects included the patients who arrived for routine screening and those with more benign or non-specific symptoms (Lloyd-Price et al., 2019).

3.1.2 Database

The IBDMDB website contains the raw and the final results of the processed information, the complete pipeline for producing the final results is:

1. Quality and error checking for completeness, producing `raw files`
2. Anadama pipeline, producing `products`

In particular, if we consider the pipeline for producing the metagenomic data, the samples for the quality control process go through the KneadData (*KneadData – The Huttenhower Lab* n.d.) and the Anadama pipelines. The former is a tool useful to exclude the reads, which are fragments of DNA related to the host or other contaminants from the metagenomic sequencing data. This separation step is made completely in silico. Whereas the latter, the Anadama pipeline, performs and produces documents from an automated scientific workflow, where a workflow is simply a succession of tasks, such as quantifying operational taxonomic units (OTU). The OTUs are classifications of groups of bacteria closely related to each other by sequence similarity. On the IBDMDB website, there are two versions of data `Version 2.0` and `Version 3.0`. `Version 3.0` has been uploaded with the new version of bioBakery (McIver et al., 2018). In the thesis, we use the `products` file related to the functional profiles `Version 2.0`. Moreover, we exploit the HMP2 Metadata file, containing the sample IDs, the subject IDs and the properties associated with each sample. See the Code 1. The `External ID` is the unique ID of the sample, `Participant ID` is the subject from where the sample has been taken, `diagnosis` is either ulcerative colitis (UC), Crohn's disease (CD) or control group (non-IBD), `week_num` points out the week number, when the sample has been taken and `data_type` is the type of sample (metagenomics, 16S, etc.).

```
import pandas as pd
# sample ID list contained in the metagenomic database
sample_id_list = list(pathabundances_mgx_red.columns[1:])
ID_diagnosis_db = pd.read_csv("./data/ibd/hmp2_metadata.csv",
    usecols=["External ID", "Participant ID","diagnosis", "
    week_num","data_type"])
```

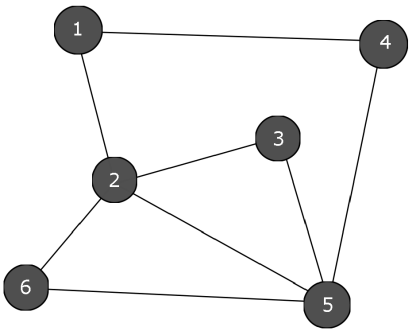
```
ID_diagnosis_db_red = ID_diagnosis_db[ID_diagnosis_db['External ID
    '].isin(list(sample_id_list))]
result = ID_diagnosis_db_red.sort_values(['week_num', 'Participant
     ID'])
result_metagenomics = result[result['data_type']=='metagenomics']
diagnosis_list = result_metagenomics[result_metagenomics['
    diagnosis']=='CD']
cd_mgx_week0 = diagnosis_list[diagnosis_list['week_num'] == 0.0]
cd_df = cd_mgx_week0.drop_duplicates("Participant ID")
cd_list = list(cd_df['External ID'])
columns_cd = ["# Pathway"]
columns_cd.extend(cd_list)
```

Listing 1: Extracting and combining the useful data from HMP2 Metadata with HMP2 Metagenomics database

In the Code 1, we extracted helpful information to avoid importing the whole database, and we selected only the samples from the first week (week 0). Moreover, the samples different from metagenomic ones were excluded. Finally, we dropped the samples from the same participant in week 0 and obtained a list of samples ID present in both the metagenomic database and the HMP2 Metadata. The metagenomic database contains the gene descriptors as row indexes; specifically, the descriptor is composed of the pathway, genus and species (e.g. "ARO-PWY: chorismate biosynthesis I |g__Alistipes.s__Alistipes_finegoldii"). The algorithm HUMAnN2 (Franzosa et al., 2018) has been used to generate the database. The algorithm can be divided into three phases; firstly, the metagenomic sample is quickly analysed to seek known species in the gut microbiome. The functional annotation of the identified pangenomes (i.e. the genome of a larger group of species) of the microbiome is concatenated to form a gene database of the sample. Secondly, using this database, the whole sample is aligned, meaning that statistics regarding the species and the genes are made, and unmapped reads are collected. Thirdly, the gene abundances are calculated, and they are combined with the metabolic network to determine the pathways in the microbial community.

3.2 Single-layer analysis of IBD

Networks are ubiquitous, and there exist several types of them, e.g. technological, financial, biological and so on. In particular, biological networks comprise a wide range of biological system scales, from the food web to the genome. Furthermore, the architecture of networks can be complicated and enriched by properties by modifying a single-layer network, such as by partitioning the nodes into several sets that have a specific relationship (multi-partite graphs) or by adding several layers that communicate among them (multilayer graphs). These modifications allow considering structures that are invisible to the single-layer network.



i \ j	1	2	3	4	5	6
1	0	1	0	1	0	0
2	1	0	1	0	1	1
3	0	1	0	0	1	0
4	1	0	0	0	1	0
5	0	1	1	1	0	1
6	0	1	0	0	1	0

Figure 2: Representation of a single layer undirected network and its adjacency matrix.

Networks generally simplify complex systems into structures made of nodes V and edges $E \subseteq V \times V$, called graphs $G = (E, V)$ (Caldarelli, 2007; Barabási, 2016). The nodes are entities depicted as circles representing elements in the actual complex systems, whereas the edges depicted as lines connecting the pairs of circles describe the relationships between

the nodes (see Figure 2). If the edges are associated with a positive real number $w \in \mathbb{R}^+$ that shows the importance of the connection, the graph is called weighted, whilst if the weights are binary $w \in \{0, 1\}$ then the graph is defined as unweighted. To summarize and facilitate the calculations on a graph, one can use the adjacency matrix, which is a matrix where the rows and columns indexes (respectively, i and j) are the nodes, and the entries are the weights connecting the corresponding nodes; in the case of undirected graphs that are the kind of graphs used in the thesis, adjacency matrices are symmetric. Now, it is possible to define several measures and important features that will be useful later; in particular, the degree is the total number of links that a node has, and it is expressed as k_i, where i is the node. By using the adjacency matrix A, we obtain:

$$k_i = \sum_{j=1}^{N} A_{ij}, \tag{3.1}$$

where N is the total number of nodes in the network. The average degree:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{3.2}$$

where $L = \frac{1}{2} \sum_{i=1}^{N} k_i$ is the total number of links in the network. The clustering coefficient:

$$C_i = \frac{2L_i}{k_i \left(k_i - 1\right)} \tag{3.3}$$

measures how much the neighbours in a network are connected to each other. The degree sequence is the non-increasing sequence of the degree of the nodes of the graph. Whilst, the degree distribution describes the probability that a node has to have a certain degree. They both contain a piece of similar information, and they characterise a network. The degree sequence is:

$$d = (k_1, k_2, \ldots, k_n) \tag{3.4}$$

whilst, the degree distribution is:

$$\left(p_k = \frac{n_k}{n}\right)_{k \geq 0} \tag{3.5}$$

where n is the number of nodes. For example, in Figure 2 the degree sequence is $\{2, 2, 2, 2, 4, 4\}$ and the degree distribution is $\{p_0 = 0, p_1 = 0, p_2 = \frac{4}{6}, p_3 = 0, p_4 = \frac{2}{6}\}$.

3.2.1 Gene Co-expression Network

According to Choobdar et al. (2019), the single-layer networks and the associated tools for the identification of the disease modules are only slightly less efficient than the multilayer counterparts for detecting the functional modules in a complex system. Therefore, for the sake of simplicity, we start to analyse the IBD from the single-layer perspective. The complexity of the relationships between the microbial genes in the gut microbiome led us to choose the gene co-expression network (GCN) as the tool to investigate the microbial complex system. In other words, we used co-expression networks as the mathematical tools to describe how much two genes are related to each other across the samples analysed. Nevertheless, we have to keep into account that the connection between two genes in the GCN does not always mean direct physical interaction between them (Buchanan et al., 2010). We started our construction of the GCN from the data downloaded from the IBDMDB, which is composed of the gene descriptors in the row indexes and the sample IDs in the column headers (see Table 1). GCNs are built from the following steps:

Table 1: Representation of the database containing the gene expression data with example values.

	SMPL1	SMPL2	SMPL3
PWY1	0.0435726	0.0377424	0.0118981
PWY2	0.0170328	0.0144735	0.0134886
PWY3	0.0145872	0.0177172	0.0121692
PWY4	0.0545121	0.0018744	0.0175601
PWY5	0.0881223	0.0111788	0.0163441

1. pairwise gene similarity score;
2. thresholding.

Normalisation methods, co-expression correlation (such as whether Pearson's or Spearman's correlation measure should be used) and significance and relevance are still debated (Tieri et al., 2019). In our project, we chose Pearson correlation similarly to the paper by MacMahon and Garlaschelli (2013). To reduce the number of genes present in the resulting network, we built the co-expression matrices for three different presence thresholds (namely, between 25% and 50%, between 50% and 75%, and more than 75%).

```
import pandas as pd
import numpy as np

def presence(data):
    data_aux = data.set_index('# Pathway')
    data_temp = data_aux.copy()
    # if path abundance is >0, genes marked as present
    data_aux[data_aux > 0] = 1
    # calculate relative presence in the respect of tot n. obs
    row_sum = data_aux.apply(np.sum, axis =1)
    temp = row_sum/len(data_aux.columns)
    # filter the genes with presence between 0.25 and 0.50
    result = data_temp[(temp > 0.25) & (temp < 0.50)]
    result = result.reset_index()
    return result
```

Listing 2: Python example to reduce the size of the database

We now introduce some helpful notations. Let us consider a complex system with M genetic pathways and S samples for each pathway. The single sample series

$$X_i \equiv \{x_i(SMPL_1); x_i(SMPL_2); \ldots; x_i(SMPL_S)\} \tag{3.6}$$

represents the activity of the ith pathway of the system over S samples. We do not care about the order of the samples because to apply the Pearson correlation measure, it is sufficient that all the series follow the same order. x_i is the normalised pathway abundance for each sample, and $SMPL_s$ is the sth sample. Then, we compute the correlation matrix $\mathbf{C}$, which measures the mutual dependencies over N sample series with a value between -1 and 1. The generic element of the correlation matrix $\mathbf{C}$ is:

$$C_{ij} \equiv Corr[X_i, X_j] \equiv \frac{Cov[X_i, X_j]}{\sqrt{Var[X_i] \cdot Var[X_j]}} \tag{3.7}$$

where

$$Cov[X_i, X_j] \equiv \overline{X_i X_j} - \overline{X_i} \cdot \overline{X_j} \tag{3.8}$$

is the covariance of X_i and X_j and

$$Var[X_i] \equiv {\sigma_i}^2 \equiv \overline{X_i^2} - \overline{X_i}^2 = Cov[X_i, X_i] \tag{3.9}$$

is the variance of X_i. In the above equations, the bar denotes an average across the samples, i.e.,

$$\overline{X_i} \equiv S^{-1} \sum_{s=1}^{S} x_i(s) \tag{3.10}$$

$$\overline{X_i^2} \equiv S^{-1} \sum_{s=1}^{S} x_i^2(s) \tag{3.11}$$

$$\overline{X_i X_j} \equiv S^{-1} \sum_{s=1}^{S} x_i(s) x_j(s) \tag{3.12}$$

Obviously, the diagonal elements of the correlation matrix are $C_{ii} = 1$. Even if we assume that the sample series are linearly independent, real biological systems violate this assumption. Nevertheless, we consider the assumption valid in the same way as it has been done for financial systems and other complex systems (MacMahon and Garlaschelli, 2013).

To transform a correlation matrix into the GCN, we used a thresholding method inspired by a brain network technique that was used to cut the least important edges and keep the significant relationships among the nodes; hence, we calculated the absolute value of the correlations, making the signs irrelevant. This method consists of increasing a cut-off threshold until the network connectivity breaks apart; because of this property, this cut-off threshold is also known as the percolation threshold. This method has been considered one of the most effective methods to maximise the information quantity kept by the network (Nicolini et al., 2020). In the thesis, we started from a cut-off threshold of $t = 0$, and we used a bisection method to get to the percolation threshold. In the bisection method, we flattened the absolute values in the weighted adjacency

matrix into a sorted array, we chose the median value and used it as the new cut-off threshold, we calculated the connectivity of the graph built from the adjacency matrix having this cut-off threshold, finally, if we obtained a connected graph with the median value as a cut-off threshold, we used as the sorted array the upper half array, on the contrary, we used the lower half. The procedure was iterative until convergence which corresponded to an array with zero length or with the same head and same tail.

```
def get_rho_threshold2(rho_matrix):
    rho_db = rho_matrix
    rho_np = rho_db.to_numpy()
    np.fill_diagonal(rho_np,0)

    a = rho_np[rho_np!=0].flatten()
    a_sorted=np.sort(a)

    N=len(a_sorted)
    L=N
    new_a = a_sorted[:L]
    th = 0
    # convergence condition
    while(new_a.size>1 and new_a[0]!=new_a[-1]):
        a_med = np.median(abs(new_a))
        th = a_med
        adj = np.array(abs(rho_np)>th*np.ones((rho_np.shape)),
    dtype=int)
        G=nx.from_numpy_matrix(adj)
        # iterative step
        L = len(new_a)
            # case of even number of elements in array
            if L%2==0:
            L=int(L/2)
            if nx.is_connected(G):
                new_a = new_a[L:]
            else:
                new_a = new_a[:L]
            # case of odd number of elements in array
        else:
            L=int((L-1)/2)
            if nx.is_connected(G):
                new_a = new_a[L:]
            else:
                new_a = new_a[:L]
    return th
```

Listing 3: Code to find the percolation threshold.

3.2.2 Differential networking

In the thesis, we considered the metagenomic data of the IBDMDB to find the differential genes and structures that underlie the GCN of CD, UC, and non-IBD controls. The GCN was generated from a total of 1,638 metagenomic samples. We divided the samples according to the diagnoses and kept only the samples taken in week number 0. We wanted to find the pathway abundances that were differential between the diseased and the control profiles. Therefore, for each pathway, we compared the abundances of the IBD profiles with those of the non-IBD profiles through permutation tests. We used 3000 as the number of Monte Carlo resampling, and we set the p-value threshold of significance to 0.05. We enriched the number of differentially expressed (DE) pathway pools by feeding the previously found DE pathways (as seeds) and the respective co-expression networks to the DIAMOnD algorithm (Ghiassian, Menche, and Barabási, 2015). Despite Paci et al. (2017) proposed a novel tool (SWIM), in our work, we decided to use the DIAMOnD algorithm over the SWIM algorithm mainly for two reasons. Firstly, DIAMOnD was more straightforward since it required one computational step less than SWIM (the community detection step). This allowed us to obtain satisfactory preprocessing without inducing data clustering in the sample. Secondly, since DIAMOnD was written in Python, while SWIM was in MATLAB code, DIAMOnD allowed us to optimally integrate results in the subsequent analysis pipeline, composed of Python functions. For additional information, see Table 2.

In our work, these nodes were found by the permutation tests; moreover, by combining the permutation tests with the DIAMOnD algorithm, we did not need the community detection step included by the SWIM algorithm. The DIAMOnD algorithm is based on the intuition that the connectivity significance, calculated as

$$p - value(k, k_s) = \sum_{k_i = k_s}^{k} p(k, k_i) \tag{3.13}$$

Table 2: Comparison of DIAMOnD and SWIM algorithms.

Method/Tool	Brief Description	Availability	References
DIAMOnD	It starts from a set of seed nodes to find disease module through 'connectivity significance' instead of local connection density.	Python	Ghiassian, Menche, and Barabási (2015)
SWIM	It integrates topological information of the network with gene expression data to identify a small pool of genes associated with drastic changes in cell phenotype.	MATLAB	Paci et al. (2017)

where

$$p(k, k_s) = \frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}} \tag{3.14}$$

is strongly distinctive for disease genes that we already know. Hence the algorithm tries to search for those disease genes that are not known yet in order to identify the disease module. DIAMOnD has an iterative procedure that comprises four steps:

1. Calculation of the connectivity significance for all the genes connected in the GCN to any of the s_0 seed genes.

2. Classification of the genes according to the p-values (listed from the lowest p-value to the highest p-value).

3. The gene at the top of the list is added to the set of the seed nodes. The number of the nodes increases from $s_0 \rightarrow s_1 = s_0 + 1$

4. it is possible to repeat steps 1-3 by using as seed set the extended set s_1. By repeating the procedure, the algorithm will increase the disease module one gene at a time.

We chose to obtain 50 ranked results from the algorithm for each disease. The intersection between the two results from the algorithm was found to obtain a list of common genes between CD and UC. The following network measurements were calculated and compared across the three diagnoses for the genes found by the algorithm: degree centrality, betweenness centrality, and eigenvector centrality. We performed a one-way ANOVA test for each measurement, and if the p-value was less than 0.05, we proceeded with a post-hoc t-tests with p-value adjusted with Holm (1979) to identify which measurements identify differential genes.

3.2.3 Centrality measures in single-layer networks

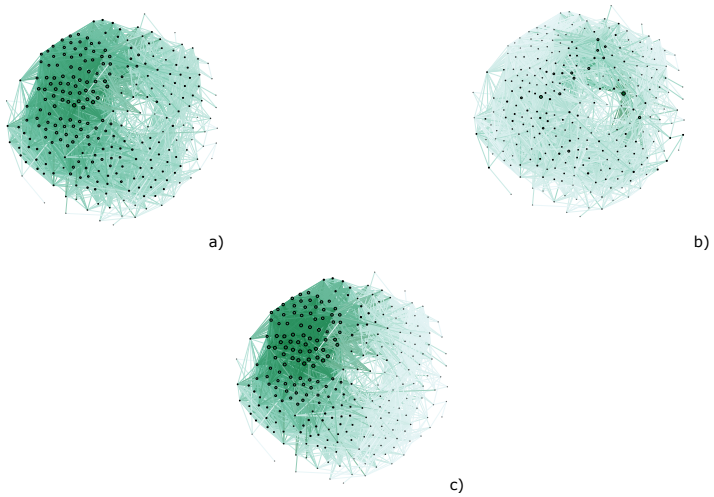


Figure 3: Visualisation of the three centrality measures on the same network: a) degree centrality, b) betweenness centrality, and c) eigenvector centrality.

There are many different centrality measures in network science; these measures describe the importance of a node in the network. The centrality measures used in the thesis are degree centrality, betweenness centrality, and eigenvector centrality (see Figure 3). The simplest one, degree

centrality, is calculated simply as the degree of the node,

$$C_D(i) = k_i, \tag{3.15}$$

the more connected the node is, the more important it is. The betweenness centrality was introduced by Freeman (1977), and it considers more important the nodes that behave as bridges in the network. It can be calculated as:

$$C_B(i) = \sum_{s \neq t \neq i \in V} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3.16}$$

where σ_{st} is the number of shortest paths connecting s and t, whilst $\sigma_{st}(i)$ is the number of shortest paths connecting s and t and going through i (Brandes, 2001). The eigenvector centrality estimates the influence of a vertex in a network, and it is measured by considering with a higher score those vertexes that are connected with vertexes with already high scores. The relative centrality measure in an unweighted and undirected graph can be defined as:

$$x_i = \frac{1}{\lambda} \sum_{t \in V} A_{it} x_t, \tag{3.17}$$

where λ is a constant (eigenvalue). All the centrality measures were calculated on the binarized adjacency matrices.

3.2.4 Community detection in single-layer networks

In the study of network science, both natural complex networks and artificial complex networks display a modular behaviour, i.e. groups of nodes are more densely connected within the members of the group than with the rest of the network. This phenomenon can also be described by a function called modularity (Newman, 2006), which can be used as a parameter for one of the several ways to perform community detection in complex networks. In our work, we used the Louvain method (Blondel et al., 2008) because it is suited to large complex networks. Louvain method is based on an optimisation problem that can be solved in a time $O(n \cdot log_2 n)$ where n is the number of nodes in the network (Lancichinetti

and Fortunato, 2009). The method is based on the aforementioned modularity optimisation. 'Modularity' is defined as (Newman, 2004),

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j). \tag{3.18}$$

The algorithm is based on two phases that repeat iteratively. In the first phase, each node is repeatedly moved individually between the communities to maximise modularity. The first phase stops when no further individual move can improve the modularity. In the second phase, each community formed in the first phase is considered a node of a weighted graph, where the weights of the edges are given by the sum of the edges connecting the nodes in the communities. The algorithm has a high efficiency partly because the gain modularity ΔQ, due to moving a node i into a community C, can be steadily calculated as:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right], \tag{3.19}$$

where $\sum_{in}$ is the sum of the weights of the edges inside C, $\sum_{tot}$ is the sum of the weights of the edges going from the outside to the nodes inside C, k_i is the sum of the weights of the edges going to node i, and $k_{i,in}$ is the sum of the weights of the edges going from i to the nodes in C and, finally, m is the sum of the weights of all the edges in the graph. One of the limitations of community detection based on the modularity is the resolution limit (Fortunato and Barthelemy, 2007). This limitation to modularity may be present when $l_s \approx \sqrt{2L}$, where l_s is the number of internal links in a module S and L is the total number of links in the network and it can be overcome through several methods, one of the most promising is Surprise maximization (Aldecoa and Marín, 2013). To compare the communities obtained in networks composed of the same nodes but different edges, we used the Normalized Mutual Information (NMI). NMI is a value comprised of 0, if the two clusterings are completely uncorrelated and 1, if the two clusterings are correlated. To evaluate the significance of the estimated value, we performed a permutation test with $N = 5000$ repetitions and calculated the p-value for each NMI.

3.3 Bipartite network analysis of IBD

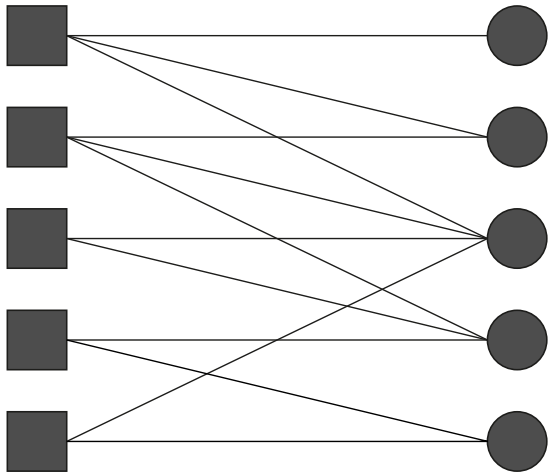


Figure 4: Representation of a bipartite graph.

Bipartite networks are graphs $G = (U, V, E)$ where the nodes can be divided into two disjoint sets, U and V, and every edge in E links an element in U to an element in V (see Figure 4). In the thesis, we designated with U the set of nodes representing the genes and with V the set of nodes representing the samples. We can find in E all the edges connecting the gene u to the sample v if the gene was over-expressed in the corresponding sample. We evaluated the over-expression of a gene through the binarisation of the data that led to the construction of a biadjacency matrix B of size $|U| \times |V|$ that described the bipartite network $G = (U, V, E)$ with the entries $(0, 1)$, where $B_{ij} = 1$ if the gene v_i is over-expressed in the sample u_j, and $B_{ij} = 0$ if it is not over-expressed. Biadjacency matrices are rectangular matrices where on one side, there are the nodes in U, and on the other side the nodes in V.

3.3.1 Binarisation

The binarization process is useful to highlight the edges in E that are over-expressed. By using the revealed comparative advantage (RCA) (Balassa, 1965), We highlighted the over-expressed genes for specific samples:

$$RCA_{ij} = \frac{E_{ij} / \sum_{j' \in V} E_{ij'}}{\sum_{i' \in U} E_{i'j} / \sum_{i' \in U, j' \in V} E_{i'j'}} \tag{3.20}$$

where E is the expression of the gene i in the sample j. When $RCA_{ij} > 1$, the quantity of gene i in sample j can be considered over-expressed and the entry $b_{ij} = 1$, in the other case $RCA_{ij} \leq 1$, then $b_{ij} = 0$.

3.3.2 Randomisation of bipartite networks

To generate a null model useful to calculate the statistically important properties of a real bipartite network, we randomised the bipartite networks by using the package BiCM (Bruno, 2020). In particular, the package is based on the works by Squartini and Garlaschelli (2011), Saracco, Di Clemente, et al. (2015), and Saracco, Straka, et al. (2017). In the aforementioned works, the Shannon entropy defined as

$$\mathcal{S} = -\sum_{\mathbf{M} \in \mathcal{G}} P(\mathbf{M}) \ln P(\mathbf{M}) \tag{3.21}$$

is maximised, where $\mathcal{G}$ is an ensemble of binary, undirected, bipartite networks, and $\overrightarrow{C}(M)$ is a given set of constraints. The result is:

$$P\left(\mathbf{M} | \overrightarrow{\theta}\right) = \frac{e^{-H\left(\mathbf{M}, \overrightarrow{\theta}\right)}}{Z\left(\overrightarrow{\theta}\right)} \tag{3.22}$$

Where $H\left(\mathbf{M}, \overrightarrow{\theta}\right) = \overrightarrow{\theta} \cdot \overrightarrow{C}(M)$ is the hamiltonian and $Z\left(\overrightarrow{\theta}\right) = \sum_{M \in G} e^{-H\left(\mathbf{M}, \overrightarrow{\theta}\right)}$ is the normalization. In the case of the bipartite extension of the configuration model (BiCM), the hamiltonian becomes:

$$H\left(\mathbf{M}, \overrightarrow{\theta}\right) = \overrightarrow{\alpha} \cdot \overrightarrow{d}(\mathbf{M}) + \overrightarrow{\beta}\, \overrightarrow{u}(\mathbf{M}) \tag{3.23}$$

because we have two layers of nodes and we constrained the degree sequences $\overrightarrow{d}(\mathbf{M})$ and $\overrightarrow{u}(\mathbf{M})$. $\overrightarrow{d}(\mathbf{M})$ is the degree sequence of the genes and $\overrightarrow{u}(\mathbf{M})$ is the degree sequence of the samples.

3.3.3 Properties of bipartite networks

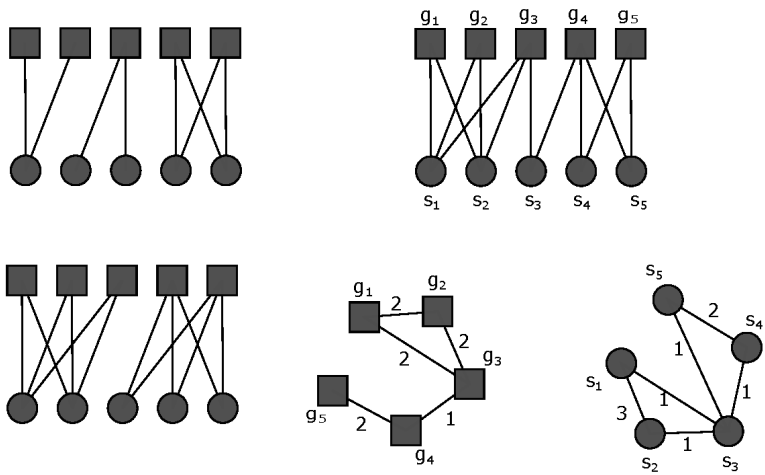


Figure 5: Some of the properties of a bipartite network. On the left, motifs and, on the right, monopartite projections.

Motifs

If, on the one hand, the usual triangular motifs that can be found in a standard single-layer network cannot be defined in bipartite networks because there are two separate partitions of nodes, and there cannot be edges between nodes in the same partition; on the other hand, in bipartite networks, we can find the simple V-motifs and Λ-motifs, or the higher-order correlations X-motifs, M-motifs and W-motifs. For V-

motifs, the number of V-motifs in the bipartite network is:

$$N_V(M) = \sum_{g=1}^{g} \binom{u_g}{2}, \tag{3.24}$$

where g is the generic gene or pathway; for Λ-motifs, the number of Λ-motifs in the bipartite network is:

$$N_\Lambda(M) = \sum_{s=1}^{s} \binom{d_s}{2}, \tag{3.25}$$

where s is the generic sample; for X-motifs, the number of X-motifs in the bipartite network is:

$$N_X(M) = \sum_{s<s'} \binom{\mathcal{S}_{ss'}}{2} = \sum_{g<g'} \binom{\mathcal{G}_{gg'}}{2}, \tag{3.26}$$

where,

$$\mathcal{S} = \mathbf{M} \cdot \mathbf{M^T}, \mathcal{G} = \mathbf{M^T} \cdot \mathbf{M}. \tag{3.27}$$

The entries $\mathcal{S}_{ss'}$ and $\mathcal{G}_{gg'}$ count respectively the number of genes over-expressed by both samples s and s' and the number of samples over-expressing both genes g and g'. For M-motifs, the number of M-motifs in the bipartite network is:

$$N_M(M) = \sum_{s<s'} \binom{\mathcal{S}_{ss'}}{3}, \tag{3.28}$$

for M-motifs, the number of M-motifs in the bipartite network is:

$$N_W(M) = \sum_{g<g'} \binom{\mathcal{G}_{gg'}}{3}. \tag{3.29}$$

Projections

One way to compress the information contained in a bipartite network is to project the bipartite network onto one of the two layers (gene/pathway layer or sample layer). We carried out the projection by connecting in the same layer the nodes that were linked by a common node in

the other layer. The projection leads to a loss of information itself, so to avoid further loss of information, we weighted the edges by the number of common nodes neighbouring the nodes in the same layer (Neal, 2014). The algorithm to perform the projection is:

1. select the partition on which the projection will be made

2. take two nodes of the selected partition, n and n', and calculate their similarity

3. by evaluating the corresponding p-value compute the statistical significance of the calculated similarity with respect to a properly-defined null model;

4. if, and only if, the p-value associated with the link n and n' is statistically significant, connect the selected nodes.

The similarity in the second step of the algorithm is evaluated by:

$$V_{nn'} = \sum_{c=1}^{N_c} m_{nc} m_{n'c} = \sum_{c=1}^{N_c} V_{nn'}^{c}, \tag{3.30}$$

where $V_{nn'}^{c} \equiv m_{nc} m_{n'c}$ and it is clear from the definition that $V_{nn'}^{c} = 1$ if, and only if, both n and n' are common neighbours of c. The third step of the algorithm passes through the calculation of the p-value of the Poisson–Binomial distribution, i.e. the probability of observing a number of V-motifs greater than, or equal to, the observed one (which will be indicated as $V_{nn'}^{*}$:

$$p - value(V_{nn'}^{*}) = \sum_{V_{nn'} \geq V_{nn'}^{*}} f_{PB}(V_{nn'}) = 1 - \sum_{V_{nn'} \leq V_{nn'}^{*}} f_{PB}(V_{nn'}). \tag{3.31}$$

Finally, in the last step of the algorithm, in order to understand which p-values were significant, a false-discovery rate or FDR has been adopted to take into account the fact that we were testing multiple hypotheses (Benjamini and Hochberg, 1995).

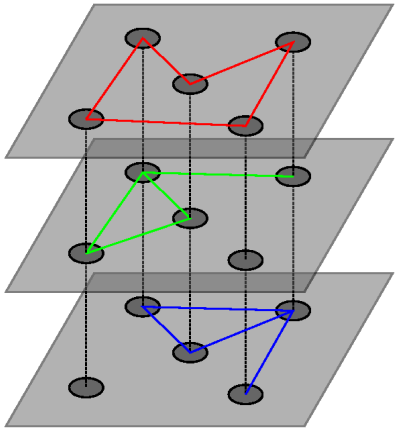


Figure 6: Representation of a multilayer network.

3.4 Multi-layer network analysis of IBD

Multilayer networks are graphs composed of several layers in order to represent complex systems that have many levels, or that need different types of edges connecting the same nodes (see Figure 6). Layers can be elementary layers or composite layers, according to the number of aspects that contribute to the layer. There exists a number d of aspects, and the Cartesian combination of the elementary layers $L_1 \times \ldots \times L_d$ generates the set of composite layers (or simply layers). A multilayer network can be defined with the quadruplet $M = (V_M, E_M, V, \mathbf{L})$, where $V_M \subseteq V \times L_1 \times \ldots \times L_d$ is the subset of nodes that exist in the multilayer network, $E_M = V_M \times V_M$ is the subset of edges connecting the combinations of nodes and the elementary layers and the sequence $\mathbf{L} = \{L_a\}_a^d$ of sets of elementary layers (Kivelä et al., 2014; Boccaletti et al., 2014). Whilst single layer networks are represented by simple adjacency matrices and bipartite networks by biadjacency matrices, and multilayer networks can be represented by supra-adjacency matrices. The supra-

adjacency matrix (see Figure 7) compared to the simple adjacency matrix, contains information about both the edges connecting the nodes in the same layer (intralayer edges) and those connecting the nodes in different layers (interlayer edges) (Cozzo et al., 2016). The supra-adjacency matrix is defined as:

$$\mathcal{A} = \bigoplus_k A^{(k)} + \mathbf{K}_m \bigoplus \mathbf{I}_n, \tag{3.32}$$

where $\bigoplus_k A^{(k)}$ is the direct sum of the intralayer adjacency matrices $A^{(k)}$, k is the layer and $\mathbf{K}_m$ is an interlayer adjacency matrix of n nodes, and $\mathbf{I}_n$ is an identity matrix of $n \times n$ (Toro, Mojica-Nava, and Rakoto-Ravalontsalama, 2019). An alternative to the supra-adjacency representation is the tensorial formalism that is useful to represent a multiplex network even if we do not have information about the block structure, which is the number of nodes that forms each layer of the multiplex (Bianconi, 2018). A tensor $\mathcal{T}$ is defined as the tensor product space:

$$\mathcal{T} : \mathcal{V} \otimes \hat{\mathcal{V}} \otimes \mathcal{V}^\star \otimes \hat{\mathcal{V}}^\star \tag{3.33}$$

with

$$T = \sum_{i,j=1,\ldots N} \sum_{\alpha,\beta=1,\ldots M} T_{i\alpha}^{j\beta} \mathbf{e}_j \otimes \hat{\mathbf{e}}_\beta \otimes \theta^i \otimes \hat{\theta}^\alpha \tag{3.34}$$

where $T_{i\alpha}^{j\beta}$ corresponds to the supra-adjacency matrix $\mathcal{A}$.

3.4.1 Multiplex networks

There exist several kinds of multilayer networks, but in the thesis, we focus on multiplex networks only. The main characteristic of multiplex networks is that the connections between layers are made by linking the replicas only, and some of the structural measures of multiplex networks are summarised in Battiston, Nicosia, and Latora (2014). The layers of the multiplex network considered in the thesis are $\alpha \in \{\text{CD}, \text{UC}, \text{non-IBD}\}$. There are different ways to study a multiplex network; one way is by analysing the aggregate overlapping adjacency matrix, and another way

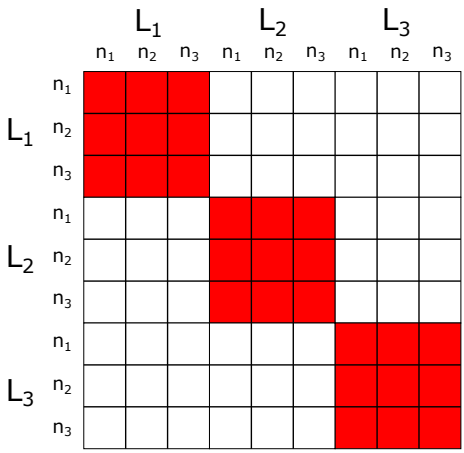


Figure 7: Example of a supra-adjacency matrix of a multilayer network with three layers and three nodes on each layer. The red blocks on the diagonal describe the intralayer edges, whereas the off-diagonal blocks describe the interlayer edges. If the network is undirected, the matrix is symmetric.

is by analysing the aggregate mean adjacency matrix. We define the former as the adjacency matrix formed by the edge overlap between nodes i and j:

$$o_{ij} = \sum_{\alpha} a_{ij}^{[\alpha]}, \tag{3.35}$$

where $a_{ij}^{[\alpha]}$ is the edge connecting the nodes i and j in layer α. The overlapping degree of nodes i can be defined as:

$$o_i = \sum_{j} o_{ij} = \sum_{\alpha} k_i^{[\alpha]}. \tag{3.36}$$

To analyse the correlations between degree sequences in the different layers and in the overlapping aggregate network, we calculate the Kendall rank correlation coefficient τ_k. The coefficient can assume values in the interval of $[-1, 1]$, where $\tau_k = 1$, if the two compared sequences are correlated, $\tau_k = -1$, if they are anti-correlated, and $\tau_k = 0$, if they are not correlated. We now introduce two measures that help to understand the distribution of information across the layers in the multiplex network, namely the entropy:

$$H_i = -\sum_{\alpha=1}^{M} \frac{k_i^{[\alpha]}}{o_i} \ln\left(\frac{k_i^{[\alpha]}}{o_i}\right), \tag{3.37}$$

and the participation coefficient:

$$P_i = \frac{M}{M-1}\left[1 - \sum_{\alpha=1}^{M}\left(\frac{k_i^{[\alpha]}}{o_i}\right)^2\right]. \tag{3.38}$$

The entropy is maximum when the information, i.e. the links, is uniformly distributed among the layers; similarly, the participation coefficient $P_i \in [0, 1]$ describes whether all the links of node i are contained in one layer ($P_i = 0$) or are equally distributed across the layers of the multiplex network ($P_i = 1$).

The procedure to organise the data into a multiplex network consisted of dividing the whole database, which contained all the samples,

into three smaller databases containing the samples labelled with one diagnosis each. The data in the databases was then transformed into GCNs as in the Subsection 3.2.1, and the genes contained in the GCNs (one for each diagnosis) were intersected to be sure that all the networks had the same nodes. Finally, to export a network file readable by the external software Muxviz (De Domenico, Porter, and Arenas, 2015), we created an edgelist file (`network.edge`), a layer file (`network_layers.txt`) and a layout file (`network_layout.txt`). In the first file, there is the description of all the connections among the nodes in the same layers (intra-layer connections) and among the nodes of different layers (inter-layer connections), the format of one edge in the list is 'start_nodeID'+' '+'start_layerID'+' '+'end_nodeID'+' '+'end_layerID'+' '+'weight'. In the second file, a unique ID is associated to each node, whereas in the third file, a unique ID is associated to each layer; these IDs were used in the edgelist file. A `config` file, which was used to show the directory positions of the files describing the network, was then uploaded to Muxviz and centrality measures and community detection were calculated. As it has been said at the beginning of the section, it is possible to investigate the mean aggregate form of the layers. In this case, we identified five layers for each diagnosis, corresponding to five metagenomic analyses every two weeks for ten weeks. We reduced the number of samples by keeping the samples of the set of participants that appeared in all the five analyses to make it possible for a correlation analysis similar to Subsection 3.2.1. Once we had obtained five correlation matrices, we performed the Fisher's z transformation of the correlation coefficients in order to make the sampling distribution of the coefficients approximately normal and to make the datasets comparable (Anderson, 1962). To perform the Fisher's z transformation, one can simply calculate:

$$z = \operatorname{arctanh}(r), \tag{3.39}$$

where r is the correlation coefficient. After applying the Fisher's z transformation of all the correlation coefficients in the matrices, we calculated the mean of the values in the correlation matrices, and then we took the module of the mean values. Then, we calculated once again the perco-

lation threshold of the weighted aggregate network with the same script in Code 3. Finally, we binarised the aggregate network and produced a file compatible with the open-source software Gephi (Bastian, Heymann, and Jacomy, 2009) for visualisation.

3.4.2 Centrality measures in multiplex networks

Centrality measures in multiplex networks are useful to highlight those nodes that are important across all the diagnoses. Nodes that have a low centrality in the entire multiplex but a high centrality in the single-layer network could correspond to a differential expression. We will now extend to the multiplex case the definitions of the same centrality measures listed in the single-layer case, namely degree centrality, betweenness centrality, and eigenvector centrality. The calculation of the degree centrality in multiplex networks corresponds to the extension of the concept of degree performed in Equation 3.36. Regarding betweenness centrality, one can calculate it by using a random walker or by using the concept of shortest paths (De Domenico, Solé-Ribalta, Omodei, et al., 2013). To be consistent with the monoplex definition of betweenness centrality, we will extend the shortest path based centrality measure to the multiplex case only. The sequence of nodes $\ell_{[o\sigma\to d\gamma]} \in \mathcal{P}_{[o\sigma\to d\gamma]}$ is a path joining a node origin o in layer σ and a node destination d in layer γ. A cost $c(\ell_{[o\sigma\to d\gamma]})$ can be associated to every path $\ell_{[o\sigma\to d\gamma]}$ and it takes into account the weights of the links and other parameters of interest. The shortest path is defined as:

$$\ell^*_{[o\sigma\to d\gamma]} = \underset{\ell'_{[o\sigma\to d\gamma]}\in\mathcal{P}_{[o\sigma\to d\gamma]}}{\arg\min} c(\ell'_{[o\sigma\to d\gamma]}), \tag{3.40}$$

that can be extended to every pair of origin and destination nodes:

$$\ell^*_{[o\to d]} = \underset{\sigma,\gamma\in\{\text{CD,UC,nonIBD}\}}{\arg\min} c(\ell^*_{[o\sigma\to d\gamma]}). \tag{3.41}$$

Finally, the betweenness centrality of a node j is proportional to the number of shortest paths between the pair of origin and destination nodes that contain the node j. The generalisation to the multiplex case of the

eigenvector centrality is carried out by using the supra-adjacency and supra-Laplacian matrices, the procedure described by De Domenico, Solé-Ribalta, Cozzo, et al. (2013) and De Domenico, Solé-Ribalta, Omodei, et al. (2015). By using the tensorial formalism, we can find the eigenvector centrality by solving the equation:

$$T_{j\beta}^{i\alpha}\Theta_{i\alpha} = \lambda_1 \Theta_{i\beta}, \tag{3.42}$$

where $T_{j\beta}^{i\alpha}$ is the multilayer adjacency tensor, λ_1 is the largest eigenvalue and $\Theta_{i\alpha}$ is the corresponding eigentensor containing the eigenvector centrality of each node.

3.4.3 Community detection in multiplex networks

The concept of community detection in the multiplex network is similar to the community detection in the single-layer analysis; nevertheless, we have to keep in consideration the addition of the interlayer links that connect the nodes of different layers. In the thesis, we use Louvain and Infomap to find the modules of the genes characterising a diagnosis. Louvain is a community detection method based on modularity maximisation as we have described in Subsection 3.2.4. The modularity function is modified to keep into account the interlayer connection and the resolution parameters ω and γ (Mucha et al., 2010):

$$\mathcal{Q}^M = \frac{1}{\mu} \sum_{i,j,\alpha,\beta} \left\{ \left(\mathcal{A}_{i\alpha,j\beta} - \gamma^{[\alpha]} p_{ij}^{[\alpha]} \right) \delta_{\alpha,\beta} + \omega \mathcal{A}_{i\alpha,j\beta} \delta_{ij} \right\} \delta \left(g_i^{[\alpha]}, g_j^{[\beta]} \right), \tag{3.43}$$

where $\mathcal{A}_{i\alpha,j\beta}$ is the supra-adjacency matrix, $\mu = \sum_{i,j,\alpha} \mathcal{A}_{i\alpha,j\alpha} + \omega \sum_{i,\alpha,\beta} \mathcal{A}_{i\alpha,i\beta}$ and $\delta(x, y)$ is the Kronecker delta.

$$p_{ij}[\alpha] = \frac{k_i^{[\alpha]} k_j^{[\alpha]}}{\langle k^{[\alpha]} \rangle N} \tag{3.44}$$

is the probability that two nodes i and j are connected by an edge in layer α.

Infomap is based on the modular flow of a random walker in a multiplex network (De Domenico, Lancichinetti, et al., 2015). One of the prop-

erties of the Infomap algorithm is the ability to find overlapping communities, i.e. communities that share some of the nodes (Bianconi, 2018). When the interlayer weights are not known, the Infomap algorithm is based on the relaxed transition probability:

$$\mathcal{P}_{ij}^{\alpha\beta}(r) = (1-r)\delta_{\alpha\beta}\frac{W_{ij}^{\beta}}{s_i^{\beta}} + r\frac{W_{ij}^{\beta}}{S_i}, \tag{3.45}$$

where r is the relax rate used to represent the movements from one layer to the other and $S_i = \sum_{\beta} s_i^{\beta}$ is the sum of the intralayer weights of node i not dependent on the layer.

Chapter 4

Results

4.1 Single-layer correlation network analysis results

We divided the gut metagenomic data into three groups; namely, a group containing prevalent bacterial pathways (over 75% of presence across the samples), a group containing common bacterial pathways (50% to 75%), and a group containing uncommon bacterial pathways (25% to 50%). Correlations between pathways meant co-variances of the pathway expressions. For each group and for each diagnosis, we performed a differential analysis comparing the pathways expression level across the samples between IBD (UC and CD) samples and NI samples. In the next paragraphs, the pathways are called with their node numbers for the sake of brevity and clarity. It is possible to consult the table mapping these correspondences in Appendix A.

4.1.1 Prevalent pathway correlation network

The prevalent bacterial pathways were 153 after filtering with a presence of over 75% samples. The number of edges in the NI correlation network was 3356 ($\text{th}_{\text{NI}} = 0.453$), in CD correlation network was 2905 ($\text{th}_{\text{CD}} = 0.357$), and in UC correlation network was 3160 ($\text{th}_{\text{UC}} = 0.364$).

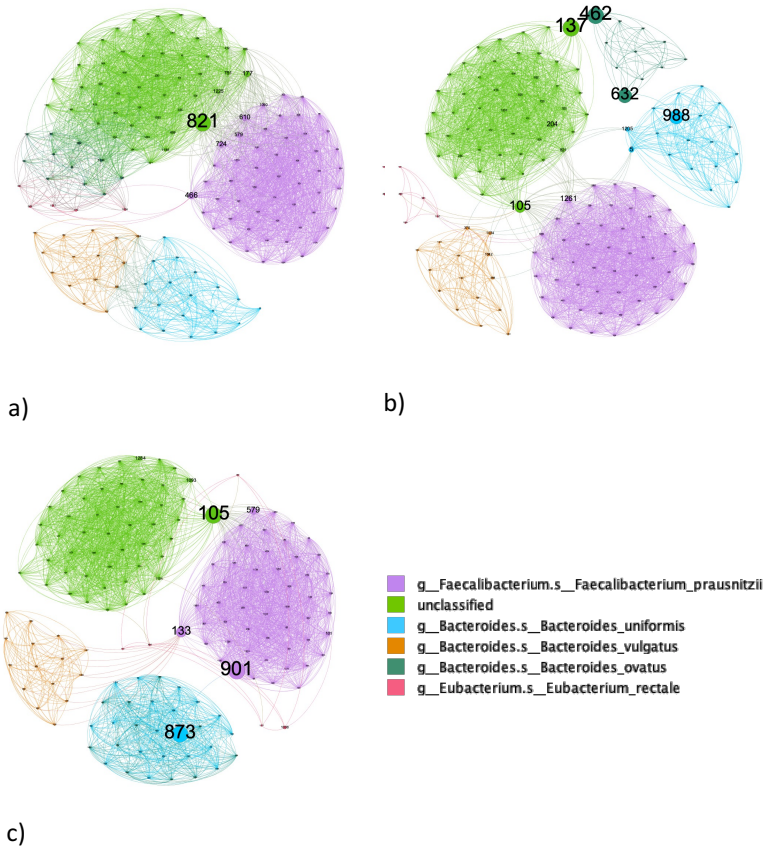


Figure 8: Correlation networks of prevalent metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Nodes are proportional to the betweenness centralities.

The results showed that the NI metagenome was more connected to prevalent pathways and the percolation threshold was higher compared to the percolation thresholds in CD and UC correlation networks, translating into more strongly correlated nodes in the NI correlation network. From the community detection algorithm, we obtained that the lowest modularity (0.538) can be found in the NI correlation network, meaning that it was not possible to completely separate some of the modules, and there would be interconnections between them, CD and UC correlation networks resulted in the modularity of 0.583 and 0.622, respectively.

As we can see in Figure 8, the pathways in the NI correlation network were divided into three large modules. One module was isolated; by contrast, the other two modules were communicating strictly through several nodes. The isolated module was composed of *Bacteroides vulgatus* and *Bacteroides uniformis* pathways; this meant that in control subjects, the two species co-variated and were interdependent through specific pathways (on the frontiers of the species modules, it was possible to find nodes 794, 658, 292, 477 on the *B. uniformis* side and nodes 261, 855, 1027, 1068 on the *B. vulgatus* side). The light purple module, on the other hand, contained *Faecalibacterium prausnitzii*, whereas the remaining large module contained the *B. ovatus* pathways, *E. rectale* pathways, and the unclassified species pathways. The nodes with the highest betweenness centrality among the unclassified pathways in the two connected modules were connected through:

1. node 821 (PWY-6527: stachyose degradation);
2. node 177 (GLYCOGENSYNTH-PWY: glycogen biosynthesis I (from ADP-D-Glucose));
3. node 1225 (PWY66-422: D-galactose degradation V (Leloir pathway));
4. node 751 (PWY-6317: galactose degradation I (Leloir pathway)).

Whereas the nodes with the highest betweenness centrality among the *F. prausnitzii* pathway module that was connected to the unclassified pathway module were:

1. node 466 (PWY-5659: GDP-mannose biosynthesis);
2. node 724 (PWY-6277: superpathway of 5-aminoimidazole ribonucleotide biosynthesis);
3. node 579 (PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I);
4. node 610 (PWY-6122: 5-aminoimidazole ribonucleotide biosynthesis II).

To notice that pathway of node 466 correlated with only 4 *E. rectale* pathways.

In the CD correlation network, there were fewer connections and the network was divided into 6 modules. Each module corresponded to the species groups of pathways. The smallest module was composed of *E. rectale* pathways. The largest (light purple) module comprising *F. prausnitzii* was connected to the unclassified (green) module by means of node 105, similarly to the UC correlation network. Moreover, an additional bridge connecting node was node 1261, which was linked to nodes 821 and 204, to mention two high betweenness centrality nodes among the unclassified pathways. High betweenness centrality nodes 137 and 462 connected unclassified species module with *B. ovatus* module, node 632, in turn, was connected to node 988 linking *B. ovatus* module to the *B. uniformis* module. Finally, similarly to the NI correlation network, *B. vulgatus* and *B. uniformis* were connected though (nodes 370, 1042, and 1074 on the former side and nodes 1205 and 5 on the latter side).

The pathways in the UC correlation network were divided into 5 modules. The smallest module (coral red) was composed of *E. rectale* pathways scattered around the network. The dark green nodes mixed to the turquoise nodes were *B. ovatus* pathways mixed with *B. uniformis* pathways, respectively. The green module comprised unclassified species pathways, whereas, the light purple module comprised *F. prausnitzii* pathways. The green and the light purple module were strictly connected similarly to the NI correlation network. The pathways connecting them were the node 105, which was linked to several nodes of both modules

and node 579, which was linked to nodes 1093 and 1284. Even *E. rectale* behaved as a bridge between the two aforementioned large modules through a few connections. Furthermore, there was one node of *F. prausnitzii* module that was deeply correlated with all the *B. vulgatus* pathways: node 133. Finally, two nodes with high betweenness centrality were node 901 in the light purple module and node 873 in the mixed dark green and turquoise module. Differently from the NI correlation network, *B. uniformis* did not correlate *B. vulgatus*, whereas it correlated with *B. ovatus*.

After calculating the NMI of the community detection, it resulted that the partitions of the CD and UC correlation networks overlapped with an NMI of 0.955 (p-value = 0.173). The partitions of the CD and NI correlation networks overlapped with an NMI of 0.823 (0.916). Finally, the partitions of the UC and NI correlation networks overlapped with an NMI of 0.762 (0.427). The NMI are summed up in Table 3.

Table 3: NMI and p-value for community detection in correlation networks with presence over 75%.

	CD	UC	NI
CD	1	0.955 (0.173)	0.823 (0.916)
UC	0.955 (0.173)	1	0.762 (0.427)
NI	0.823 (0.916)	0.762 (0.427)	1

Using permutation tests, we discovered 31 pathways differentially expressed between CD and NI groups (p-value < 0.05), 29 of these pathways belong to the unclassified species module in the CD correlation network, whereas the remaining two pathways belonged to the *B. ovatus* module of the CD correlation network. One of the two pathways (node 462) was connected to the unclassified species module through node 137. According to the permutation tests, there were no significant differentially expressed pathways between NI and UC groups in this range of pathways.

Using the 31 pathways as seed nodes for the DIAMOnD algorithm, we expanded the set of differential pathways with 50 nodes (see Figure

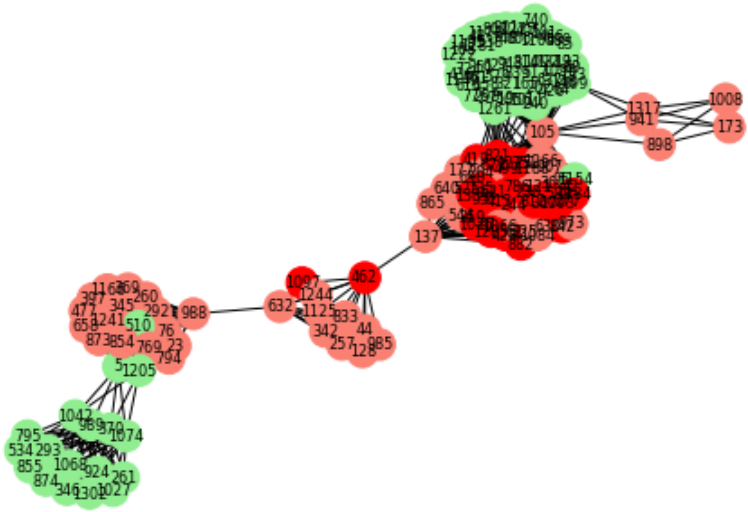


Figure 9: Correlation network of the prevalent pathways in the CD metagenome, the red nodes depict the DE pathways between CD and NI subjects, and the light red nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are not differential nor in the DIAMOnD found group.

9). We obtained in the order of significance 22 additional unclassified species pathways, 5 *E. rectale* pathways, and the remaining pathways were in *the B. ovatus* module and *B. uninformis* module.

Analysing the centralities of the differential pathways in the CD correlation network and in the NI correlation network, we noticed that the eigenvector centrality of the differential pathways in the latter group was smaller compared to the same measure of the former group with one-way ANOVA p-value $= 4.068 \times 10^{-23}$. The additional pathways found by the DIAMOnD algorithm can be divided into three groups, a group of pathways for which the eigenvector centrality was higher in the NI correlation network as for the seed differential expressed pathways, a group in which the measures were slightly higher in the NI correlation network, and a group where the centrality measures of the same nodes were higher in the CD correlation network. These three groups corresponded, namely to the pathways belonging to the unclassified species, the pathways belonging to the *E. rectale* module and *B. ovatus* module, and finally, the pathways belonging to *F. prausnitzii* module, as is possible to observe in Figure 10. The overall p-value that demonstrated the significance of the finding was 6.388×10^{-7} We did not find significant relationships between the betweenness centralities of the two groups either for the differentially expressed pathways or the additional DIAMOnD pathways. By contrast, there was a significant difference in the degree centrality of both the differentially expressed pathways in the CD correlation network compared to the same measurement in the NI correlation network (p-value = 4.416×10^{-4}) and the additional DIAMOnD pathways (p-value = 3.931×10^{-3}), see Figure 11.

4.1.2 Common pathway correlation network

As it is possible to observe in Figure 12, after filtering, the common bacterial pathways were 227 in each of the correlation networks. The number of edges among the common bacterial pathways in NI correlation network was 3410 ($\text{th}_{\text{NI}} = 0.429$), in CD correlation network was 2586 ($\text{th}_{\text{CD}} = 0.371$), and in UC correlation network was 5156 ($\text{th}_{\text{UC}} = 0.447$).

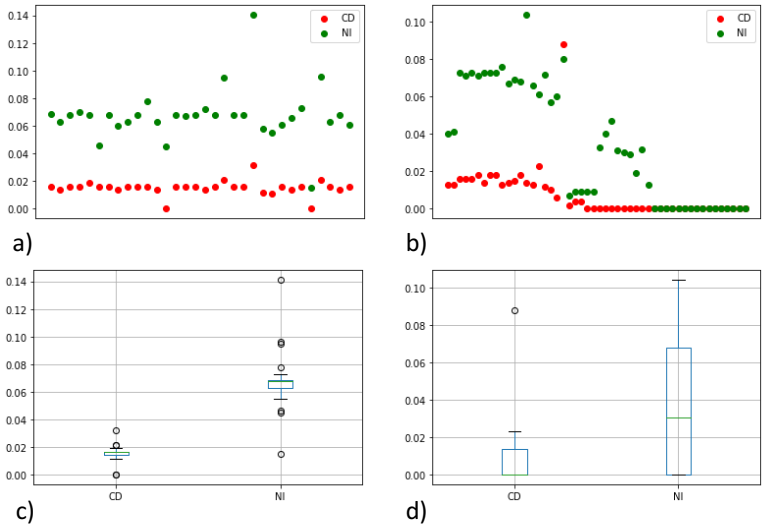


Figure 10: Different centralities measures with p-value < 0.05 for prevalent pathways comparing nodes in the CD correlation network and NI correlation network. Figure (**a**) compares the eigenvector centralities of the DE nodes calculated in the CD correlation network (red) with the eigenvector centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the eigenvector centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the eigenvector centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the eigenvector centralities in Figure (**a**) and Figure (**b**), respectively.

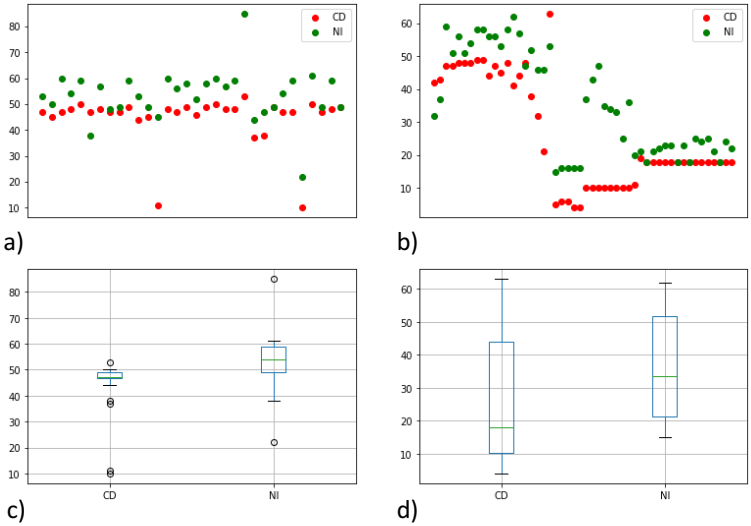


Figure 11: Different centralities measures with p-value < 0.05 for prevalent pathways comparing nodes in CD correlation network and NI correlation network. Figure (**a**) compares the degree centralities of the DE nodes calculated in the CD correlation network (red) with the degree centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the degree centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the degree centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the degree centralities in Figure (**a**) and Figure (**b**), respectively.

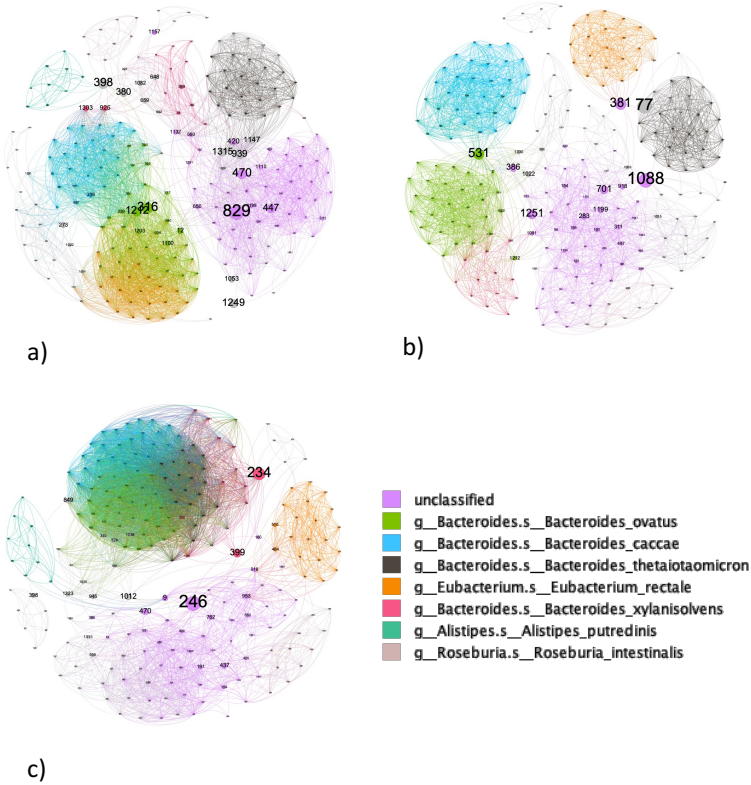


Figure 12: Correlation networks of common metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Nodes are proportional to the betweenness centralities.

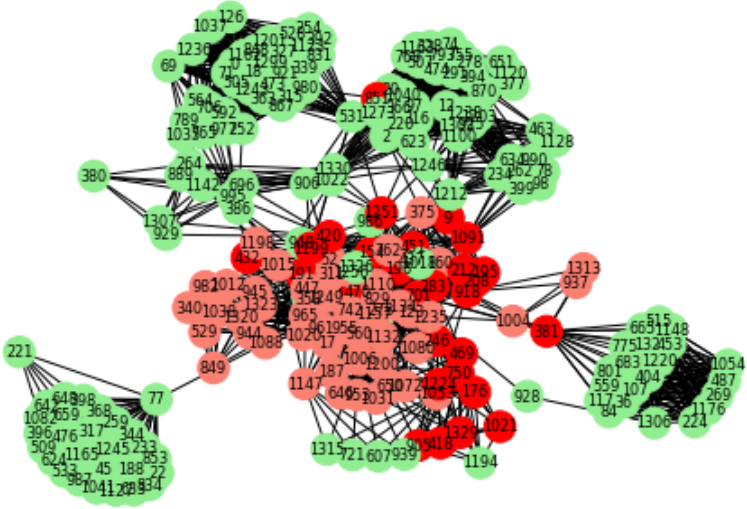


Figure 13: Correlation network of the common pathways in the CD metagenome, the red nodes depict the DE pathways between CD and NI subjects, the light red nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are not differential nor in the DIAMOnD found group.

From an overview of the networks, we could observe that the correlation network of UC subjects was more connected than the correlation network of NI or CD subjects. We performed the community detection of the networks representation of the different diagnoses.

The modularity obtained for the NI correlation network was 0.636, in which there were 8 modules. The largest community (pink) was entirely composed of pathways of unclassified species, and the second largest community we obtained comprised both the green and the orange nodes, namely, the *B. ovatus* and *E. rectale* pathways, which were densely connected between them. The third-largest module was the one containing the turquoise nodes, i.e. *B. caccae*, and the *R. torques* pathways. The fourth-largest module was the one containing *A. putredinis*, *B. xylanisolvens* and other species pathways. Finally, there was a module composed of *B. thetaiotamicron* and *E. eligens*. Node 398 had a high betweenness centrality and connected the fourth largest module with several pathways of the third-largest module. Two nodes (1212, 316) were at the intersection of pathways belonging to three different modules; the unclassified species, *B. caccae* module and the second-largest module. Node 829 had a high betweenness centrality because it was in the middle of the largest module and several edges were connecting it to nodes of other modules such as nodes 1315 and 939.

The modularities of the CD correlation network (0.748) and the NI correlation network were comparable with the striking difference that the module containing *B. ovatus* and *E. rectale* pathways in the NI correlation network was split into two distinct and separated modules in the CD correlation network; by re-arranging the connections, there were only 7 modules. *E. rectale* module in the CD correlation network was connected with the rest of the network only through the node 381, whereas in the NI correlation network, *E. rectale* module was deeply connected with *B. ovatus*. On the other hand, *B. xylanisolvens*, which in the NI correlation network was connected to the largest module composed of unclassified species pathways, were strongly interconnected with *B. ovatus*, and they formed a module connected with *B. caccae* through the high betweenness centrality node 531. In the NI correlation network, the same node had a

similar role, by connecting the two *Bacteroides* modules plus the *E. rectale*, which could be found in the same module as the *B. ovatus*. The largest module was again composed of unclassified species pathways and some additional pathways from other species such as *E. eligens*. Another relevant pathway could be found in node 77 in the *B. vulgatus* which was connected to the node 1088 of the unclassified species module, whereas in the NI correlation network the same species module was much better connected by means of three high centrality nodes belonging to the *E. eligens* (939, 1315, 1147).

Comparing the community detection of UC and NI correlation networks, we noticed that there were only 6 large modules in the UC correlation network, and it resulted in the modularity of 0.403. The largest module was formed by nodes of four species groups. Surrounding this module, there was a half-moon shaped module composed of 5 smaller modules. The second-largest module comprised unclassified species and several other species pathways. Node 246 had high betweenness centrality as it was connected to node 234, the central node in the largest ball-shaped module. Again, *E. rectale* pathways were separated from the other modules except for the unclassified species module. The relatively high betweenness centrality nodes connecting the two modules were nodes 404 and 515 from the *E. rectale* side and 918 and 953 from the unclassified species side.

After calculating the NMI of the community detection, it resulted that the partitions of the CD and UC correlation networks overlapped with an NMI of 0.373 (p-value = 0.339). The partitions of the CD and NI correlation networks overlapped with an NMI of 0.528 (0.764). Finally, the partitions of the UC and NI correlation networks overlapped with an NMI of 0.290 (0.427). The NMI are summed up in Table 4.

Using a permutation test with a significance threshold of 0.05, we found 28 differentially expressed common bacterial pathways between the CD group and NI group of pathways. Among these pathways, there were 19 pathways for unclassified species, 8 pathways for *R. intestinalis*, and one pathway for *B. ovatus*. Considering the differentially expressed pathways between the UC group and the NI group of pathways, we ob-

Table 4: NMI and p-value for community detection in networks with 50% to 75% presence.

	CD	UC	NI
CD	1	0.373 (0.339)	0.528 (0.764)
UC	0.373 (0.339)	1	0.290 (0.427)
NI	0.528 (0.764)	0.290 (0.427)	1

tained 22 for common bacterial pathways, 9 of these pathways belonging to *A. putredinis* species, 7 belonging to *B. xylanisolvens* species, and 6 for unclassified species. By expanding the set of differentially expressed pathways between CD and NI diagnoses through the DIAMOnD algorithm, we kept the first 20 significant pathways, among which there were 18 pathways of unclassified species, one pathway of *E. coli* (node 1194) and one pathway of *F. prausnitzii* (node 1249). By proceeding in the same way with the differentially expressed pathways between UC and NI, we expanded the set of pathways with 50 pathways, among which 9 were unclassified species pathways, 4 were *R. torques* pathways, 3 were *E. eligens* pathways.

Analysing the centrality measures of the CD differential pathways and the additional CD pathways computed with the DIAMOnD algorithm, we noticed that there were no significant differences between the eigenvector centralities of CD differential pathways in the CD correlation network and in the NI correlation network. Similarly, the CD pathways from the DIAMOnD algorithm did not show relevant differences in the eigenvector centralities in the considered networks. By contrast, the betweenness centralities of differentially expressed pathways were higher in the CD correlation network in the respect of the ones in the NI correlation network (p-value=0.0142). The betweenness centrality of the additional CD pathways found with the DIAMOnD algorithm did not show any significant difference in betweenness centrality. Finally, by considering the degree of centrality, we noticed that the differentially expressed CD pathways did not exhibit differential degree centrality, whereas the CD pathways from the DIAMOnD algorithm displayed a significantly

lower degree of centrality in the CD correlation network in the respect of the NI correlation network (p-value=0.00380), see Figure 14.

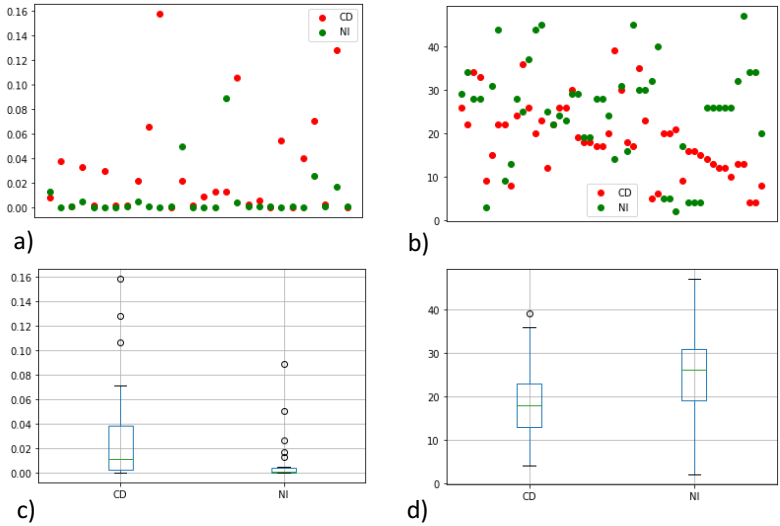


Figure 14: Different centralities measures with p-value < 0.05 for common pathways comparing nodes in CD correlation network and NI correlation network. Figure (**a**) shows the betweenness centralities of the DE nodes calculated in the CD correlation network (red) compared to the betweenness centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the degree centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the degree centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measure in Figure (**a**) and Figure (**b**), respectively.

This time, analysing the centrality measures of the UC differential pathways and the additional UC pathways by the DIAMOnD algorithm (see Figure 13), it was evident that the UC differential pathways exhibited a higher eigenvector centrality in the UC correlation network (p-value=0.00456). Also, the DIAMOnD nodes found had significant higher

Figure 15: Correlation network of the common pathways in the UC metagenome, the yellow nodes depict the DE pathways between UC and NI subjects, and the light yellow nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are not differential nor in the DIAMOnD found group.

eigenvector centralities compared to the same nodes in the NI correlation (p-value=0.0241). By contrast, the betweenness centralities of both the differential pathways and the DIAMOnD algorithm were not statistically differential between the NI correlation network and the UC correlation network. On the other hand, these pathways showed a higher degree of centrality in the UC correlation network (p-value = 0.0145), see Figure 16. On the other hand, the DIAMOnD added pathways did not exhibit significantly differential degree centralities compared to the same nodes in the NI correlation network.

4.1.3 Uncommon pathway correlation network

As it is possible to observe in Figure 17, the relatively uncommon pathways were 910 in total. There were 49203 edges in NI correlation network ($\text{th}_{\text{NI}} = 0.474$), 42617 edges in CD correlation network ($\text{th}_{\text{CD}} = 0.446$), 44523 edges in UC correlation network ($\text{th}_{\text{UC}} = 0.465$). The number of edges was comparable in the three networks. We could observe nine modules in the NI correlation network (modularity 0.623, nine modules in the CD correlation network (modularity 0.644), and six modules in the UC correlation network (modularity 0.695). In every network, it was possible to identify an approximately isolated ball-shaped module containing *E. coli* pathways. It was interesting to highlight the position of *B. fragilis* pathways in respect of *E. coli* pathways in the different diagnoses. In the NI correlation network, *B. fragilis* pathways were connected to the *E. coli* module through *V. parvula* pathways; by contrast, in the UC correlation network, *B. fragilis* pathways were incorporated and surrounded by the same module containing *E. coli* pathways, whereas in the CD correlation network the bacterial pathways of the two species were completely separated. In the NI correlation network, the module containing *E. coli* also included bacterial pathways of other species, notably *E. siraeum* and *R. gnavus* pathways that behaved as bridge nodes between *E. coli* containing module and the rest of the network. In the CD correlation network, this role was assumed by *R. intestinalis* and *V. parvula*, whereas, in the UC correlation network, we did not observe any pathways be-

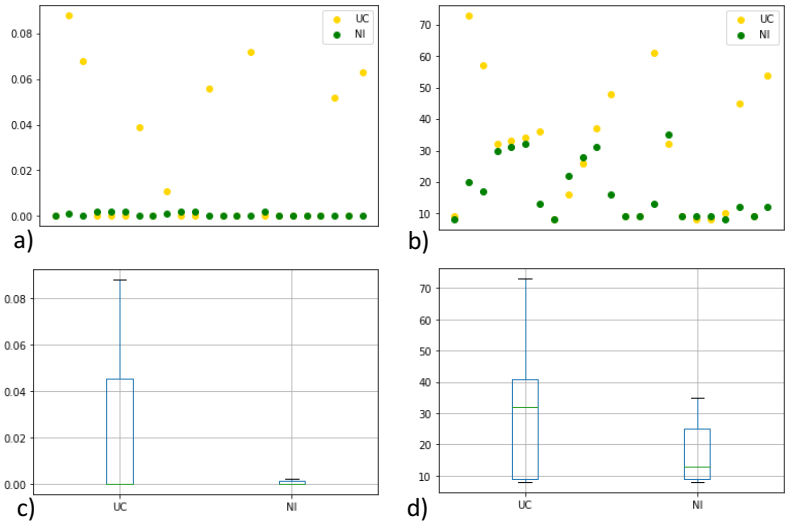


Figure 16: Different centralities measures with p-value < 0.05 for common pathways comparing nodes in UC correlation network and NI correlation network. Figure (**a**) shows the eigenvector centralities of the DE nodes calculated in the UC correlation network (yellow) compared to the eigenvector centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the degree centralities calculated for the nodes previously found through the DIAMOnD algorithm in the UC correlation network (yellow) with the degree centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measures in Figure (**a**) and Figure (**b**), respectively.

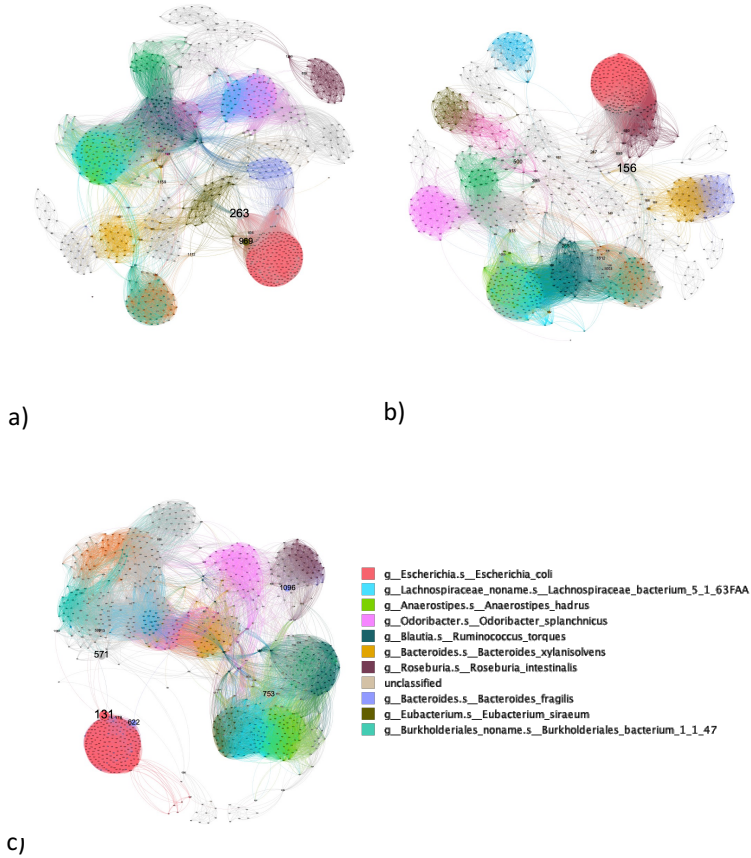


Figure 17: Correlation networks of uncommon metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Nodes are proportional to the betweenness centralities.

having as bridge nodes. An additional important module that was also the largest was the one mainly composed of *R. torques*, *A. hadrus*, *L. bacterium 5 1 63FAA*, plus other minor species. The pathways belonging to the last two mentioned species were strictly intertwined, forming the second-largest ball-shaped group of nodes. The same ball-shaped group of nodes was present in the CD correlation network; by contrast, in the UC correlation network, in the UC correlation network, the two species were in the same module, and the nodes were not mixed in the ball, but they laid separately.

Considering the betweenness centrality of the nodes, we could observe that in the NI correlation network, the ball-shaped group of nodes comprising *E. coli* was connected with the rest of the network through two nodes; node 969 stood in between the *E. coli* pathways and the *E. siraeum* pathways, whereas node 263 connected the module to *R. torques* pathways. In the CD correlation network, node 156 had a high betweenness centrality because it linked the pathways in the *E. coli* group to several minor species pathways in the rest of the network. In the UC correlation network, in ascending order of betweenness centrality, there were node 178, node 622, and node 131. These nodes were included in the same module containing the *E. coli* pathways. The first and the third nodes were connected to node 571 and several other *C. boltaea* pathways. By contrast, the second node was connected to node 1096, which in turn was connected to *R. intestinalis* pathways. Considering the NMI, the NMI estimated was maximum for the overlap of the CD and NI correlation network partitions (0.508 with p-value=0.398) and minimum for the overlap of the UC and CD correlation network partitions (0.399 with p-value=0.919). The NMI calculated for the overlap of the UC and NI correlation network partitions was close to the largest NMI (0.483 with p-value=0.402).

Using a permutation test with a significance threshold of 0.05, we discovered 52 differential pathways between CD and NI groups, see Figure 18. Among these 52 pathways, 31 were *R. intestinalis* pathways, 8 were unclassified species pathways, 4 *B. intestinihominis* pathways, 4 *D. longicatena* pathways, and 5 *C. aerofaciens* pathways. On the other

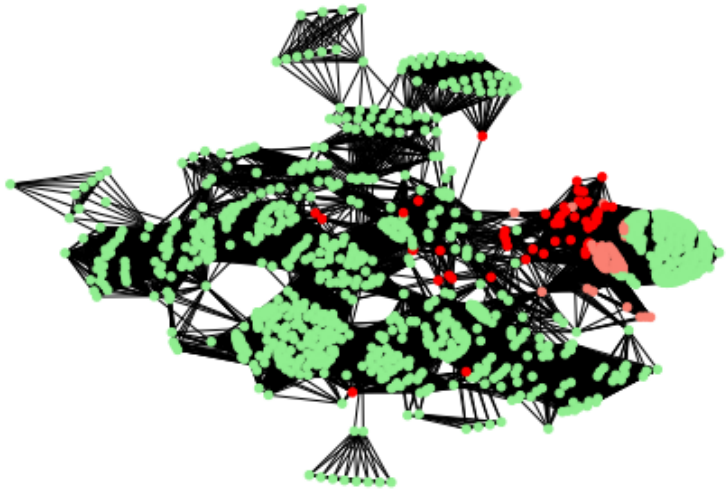


Figure 18: Correlation network of the uncommon pathways in the CD metagenome, the yellow nodes depict the DE pathways between CD and NI subjects, and the light yellow nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are the nodes that are neither differential nor DIAMOnD generated.

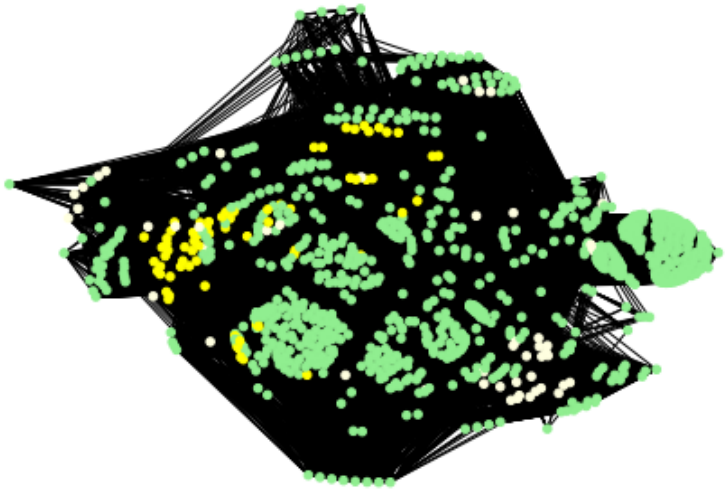


Figure 19: Correlation network of the uncommon pathways in the UC metagenome, the yellow nodes depict the DE pathways between UC and NI subjects, and the light yellow nodes are the first 50 nodes found through the DIAMOnD algorithm by using the DE nodes as seed nodes. The green nodes are the nodes that are neither differential nor DIAMOnD generated.

Table 5: NMI and p-value for community detection in networks with presence between 25% and 50%.

	CD	UC	NI
CD	1	0.399 (0.919)	0.508 (0.398)
UC	0.399 (0.919)	1	0.483 (0.402)
NI	0.508 (0.398)	0.483 (0.402)	1

hand, we found 71 differential pathways between UC and NI groups (see Figure 19), among which 30 were *B. xylanisolvens* pathways, 13 were *A. shahii* pathways, 14 were *O. splanchnicus*, and 5 *R. gnavus* pathways. By expanding the set of differential pathways with the DIAMOnD algorithm, we obtained 50 additional pathways for the CD differential pathways and 50 for the UC differential pathways. Among the CD pathways from the DIAMOnD algorithm, there were 26 *E. coli* pathways, 12 *C. bolteae* pathways, and 7 *V. parvula* pathways. By contrast, among the UC pathways from the DIAMOnD algorithm, we found 23 *R. obeum* pathways and 14 pathways of the *Bacteroides* genus. Analysing the centrality measures of the differential pathways and the additional pathways from the DIAMOnD algorithm, we noticed that the CD differential nodes exhibited a higher eigenvector centrality in the CD correlation network in respect of the same nodes in the NI correlation network (p-value = 4.43×10^{-11}), in fact, in NI correlation network the differential pathways had close to zero eigenvector centrality measures. Moreover, the betweenness centrality and the degree centrality calculated for the CD differentially expressed pathways were higher in the CD correlation network in respect of the NI correlation network (respectively, p-value = 0.00562 and p-value =0.00152), see Figure 20. We did not notice any significant difference in the centralities of the DIAMOnD enriched pathways between the CD correlation network and the NI correlation network. Considering the UC differential nodes, we obtained that some of them showed a slightly significant higher degree centrality in the NI correlation network compared to the UC correlation network (p-value = 0.036). Considering the additional DIAMOnD pathways for the UC case,

we obtained a slightly significant higher betweenness centrality of these pathways in the NI correlation network compared to the UC correlation network (p-value = 0.031), see Figure 21.

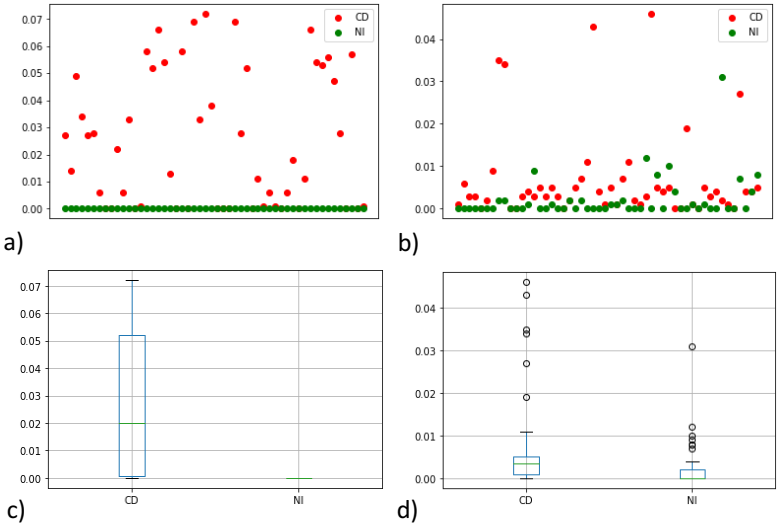


Figure 20: Different centralities measures with p-value < 0.05 for uncommon pathways comparing nodes in CD correlation network and NI correlation network. Figure (**a**) shows the eigenvector centralities of the DE nodes calculated in the CD correlation network (red) compared to the eigenvector centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the betweenness centralities calculated for the nodes previously found through the DIAMOnD algorithm in the CD correlation network (red) with the betweenness centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measure in Figure (**a**) and Figure (**b**), respectively.

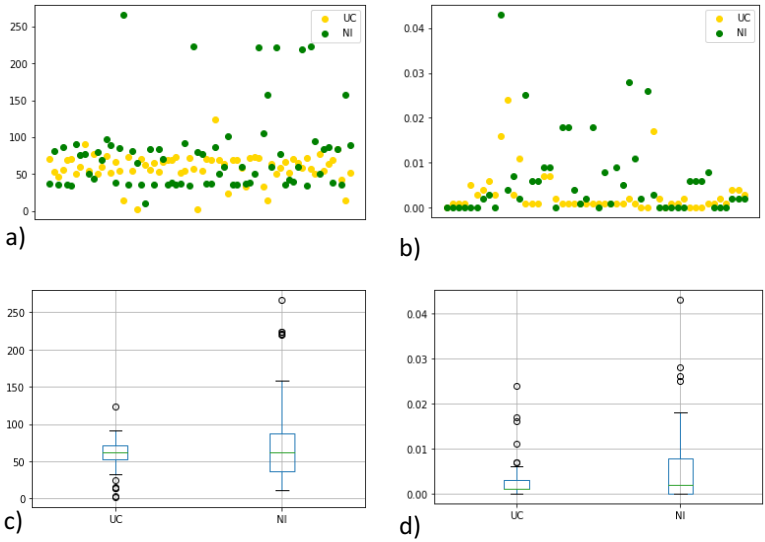


Figure 21: Different centralities measures with p-value < 0.05 for uncommon pathways comparing nodes in UC correlation network and NI correlation network. Figure (**a**) shows the degree centralities of the DE nodes calculated in the UC correlation network (yellow) compared to the degree centralities of the same nodes in the NI correlation network (green). Figure (**b**) compares the betweenness centralities calculated for the nodes previously found through the DIAMOnD algorithm in the UC correlation network (yellow) with the betweenness centralities calculated for the same nodes in the NI correlation network (green). Figure (**c**) and Figure (**d**) are the box plot showing the variability of the centrality measure in Figure (**a**) and Figure (**b**), respectively.

4.2 Bipartite network analysis results

Bipartite networks are useful when we have two sets of nodes, e.g. pathways on one side and samples on the other one. By computing the similarity of the pairs of samples, it is possible to project the bipartite network onto the sample set of nodes creating a single layer network. In our case, we obtain four networks containing the samples joined by an edge if the metagenomic profiles were similar and after being validated by a null model. We tried to project the bipartite networks by using the whole metagenome database, the pathways expressed in over 75% of the samples (prevalent pathways), the pathways expressed in 50% to 75% of the samples (common pathways), and the pathways expressed in 25% to 50% of the samples (uncommon pathways), similarly to the single-layer case.

4.2.1 Projection on prevalent pathways

We have projected the bipartite network onto the nodes of the bacterial pathways, and we validate the projection through a null model ($\alpha = 0.05$ and $fwer = none$). Again, we divided the pathways according to their presence along with the samples. We considered the case of 75% of the presence across the samples for NI subjects. We found 1715 edges for 153 nodes, and the community detection resulted in 6 large communities; namely, the unclassified species community, the *F. prausnitzii* community, the *E. rectale* community, the *B. uniformis* community, the *B. ovatus* community and *B. vulgatus* community. All the communities identified were isolated; the node 105 was connected to the unclassified module through the nodes 842, 362, and 1284. Also, nodes 419 and 277 were separated from the rest of the unclassified module. Considering *F. prausnitzii* module, nodes 1195, 740, and 466 were disconnected from the rest of the module.

In the CD case, there were 153 nodes and 1615 edges; the community detection algorithm identified 6 different modules, one for each bacterial species identified in the previous cases. All the communities were isolated without nodes connecting them. Differently from the NI case, node

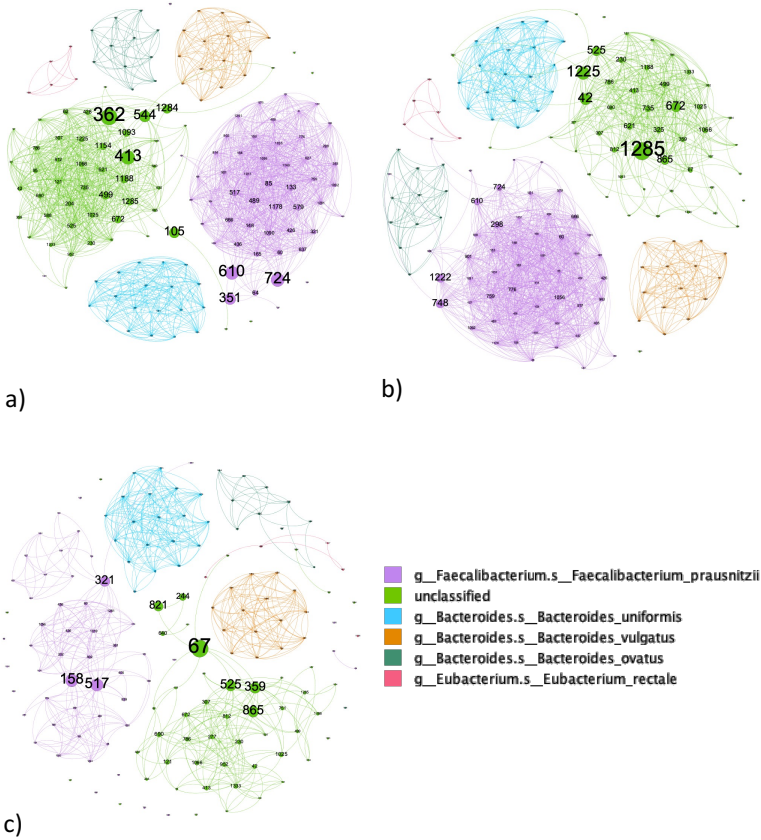


Figure 22: Projected network representation of prevalent metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Networks are obtained through a bipartite projection, and the exclusion of an edge between two nodes is made through a comparison with a null model. Nodes are proportional to the betweenness centralities.

105 does not result connected to the rest of the unclassified module. Similarly to the NI case, also node 419 was not isolated, whereas node 277 was well connected to the rest of the unclassified module. When *F. prausnitzii* was considered, we obtained that node 1195 was connected to nodes 748 and 1222, that both were well connected to the *F. prausnitzii* pathways. Also, node 740, differently from the NI projected network, is connected to two nodes of the bacterium module, namely 610 and 724, which were both pathways involving 5-aminoimidazole ribonucleotide. On the other hand, node 466 was connected to node 538, and to nodes 748 and 1222, where the former was involved in the biosynthesis of urate and the latter were involved in the degradation of galactose.

In the UC case, we could find 153 nodes and 669 edges. Differently from the previous cases, the community detection algorithm could not isolate modules corresponding to the 6 species of the prevalent pathway group. The *F. prausnitzii* module was split into two modules held together by the nodes 158 and 517. Similarly to the CD projected network, node 105 was isolated from the rest of the unclassified module, whereas node 419 was connected to the rest of the module by means of nodes 62, 204, 428, and 67. Differently from the NI projected network and similarly to the CD projected network, in the UC projected network, the node 277 was well connected to the rest of the unclassified module. When the *F. prausnitzii* module was considered, nodes 1195, 740 and 466 were isolated from the rest of the module in the same fashion as the NI case. Furthermore, the UC projected network, by displaying fewer connections, had more isolated pathways compared to the NI and the CD case. For instance, nodes 128 and 1097 detached from the *B. ovatus* module, node 1008 detached from the *E. rectale* module, and several more fragments of the module.

4.2.2 Projection on common pathways

In the NI network projection of common pathways, it was possible to observe 227 nodes and 952 edges. The largest modules were those collecting *B. caccae* pathways, *B. ovatus* pathways, *B. thetaiotaomicron* pathways

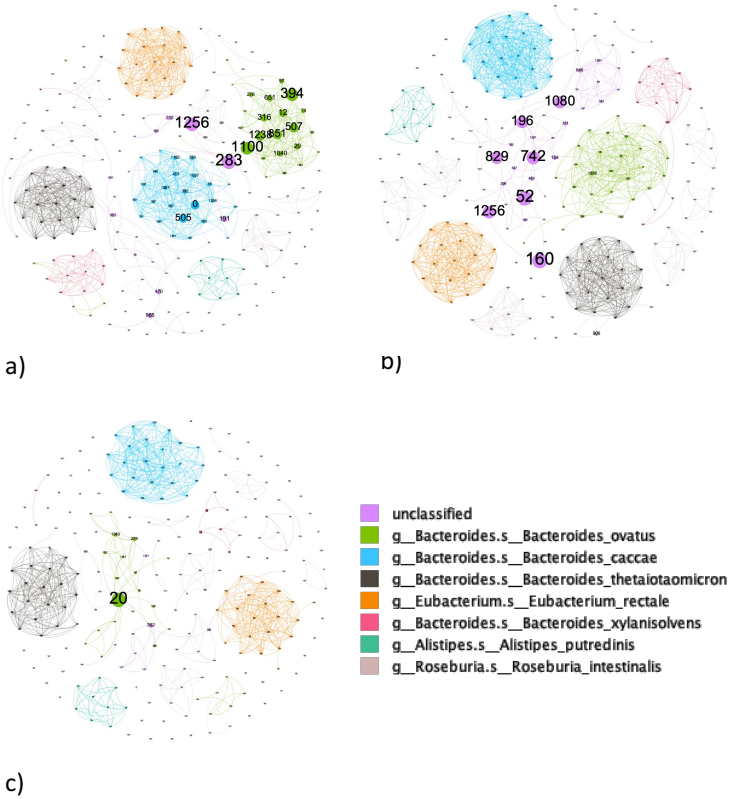


Figure 23: Projected network representation of common metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Networks are obtained through a bipartite projection, and the exclusion of an edge between two nodes is made through a comparison with a null model. Nodes are proportional to the betweenness centralities.

and *E. rectale* pathways. Although the unclassified community counts the largest number of pathways, the module was fragmented into several pieces. The second-largest group of pathways belongs to *B. ovatus*, which was split into two main modules, where the largest looked well-connected with the exception of node 870 connected to node 394 and node 1203 connected to node 1100, whereas the smallest was composed of nodes 755, 291, 768, and 793. Nodes 531, 1212, and 377 were completely isolated from the rest of the pathways. The pathways of *B. caccae* module, *B. thetaiotaomicron* module, *A. putredinis* module, *R. intestinalis* module and *E. rectale* module were well-connected to the remaining nodes of the same species without any isolated pathway for any species. The only particularity was node 1123, which had only 3 connections to the rest of the *B. caccae* module. Also, *B. xylanisolvens* had a large and well-connected module and a module composed of nodes 925 and 1303.

In the CD correlation network, we had 227 nodes and 1169 edges. Similarly to the NI case, the unclassified community was the one counting the most numerous pathways, although it was also the most fragmented community; nevertheless, it was more connected compared to the NI case. The other largest and most well-connected modules were *the B. caccae* module, *B. thetaiotaomicron* module, *B. rectale* module, the *B. xylanisolvens* module, the *A. putredinis* module and the *R. intestinalis* module. In the *B. ovatus* community, the second-largest community in the group, we had one large module of well-connected nodes with the exception of two nodes (node 651 connected to node 1120 and node 2 connected to node 394). It was possible to notice the differential connection of node 394 in the NI case and in the CD case. The isolated nodes were 1212 and 377. Node 531, which in the NI case was isolated, in the CD case, was well-connected to the rest of the *B. ovatus* module. Another striking difference appeared in the wiring of *B. xylanisolvens*; specifically, nodes 925 and 1303 were connected to the rest of the species module by means of node 1246.

Finally, in the UC network projection, it was possible to observe 227 nodes and 736 edges. Also, in this case, although the unclassified group of pathways was the largest, it was also even more fragmented than in

the previous two cases. The largest modules were *B. caccae* module, *B. thetaiotaomicron* module, *E. rectale* module, *A. putredinis* module, and *R. intestinalis* module. *B. ovatus* module was missing many edges compared to the CD and NI projected networks; therefore, it was even possible to notice a pathway with a betweenness centrality higher than the other nodes in the same module (node 20). Nodes 651 and 1120 were separated from the rest and connections between them. Nodes 1212, 377, and 531 were isolated from the other pathways, similarly to the NI case. Differently from the previous cases, also nodes 2 and 220 were not connected to the rest of the nodes. The second-largest module composed of *B. ovatus* pathways was formed by 5 pathways, which were nodes 1238, 623, 1100, 12, and 1203. The listed nodes in the NI projected network were well-connected to the rest of the *B. ovatus* main module. The *B. xylanisolvens* pathways were separated into different modules, which presented few edges. Similarly to the CD case, nodes 925 and 1303 were connected to the largest module, this time by means of node 98 that was connected via a chain of pathways to node 1246.

4.2.3 Projection on uncommon pathways

In the NI network projection, we obtained a network with 953 nodes and 15394 edges. As it is possible to notice in Figure 24a, there were several modules that emerged from the network. The most evident was the *E. coli* module which was one of the modules that composed the *E. coli* group in the uncommon pathways. This bigger module was connected to *F. plautii* via 5 nodes on the *E. coli* side and via 5 nodes on the *F. plautii* side. On the *E. coli* side:

1. node 146 (FUCCAT-PWY: fucose degradation);
2. node 896 (PWY-6737: starch degradation V);
3. node 838 (PWY-6609: adenine and adenosine salvage III);
4. node 877 (PWY-6703: preQ0 biosynthesis);
5. node 963 (PWY-7199: pyrimidine deoxyribonucleosides salvage).

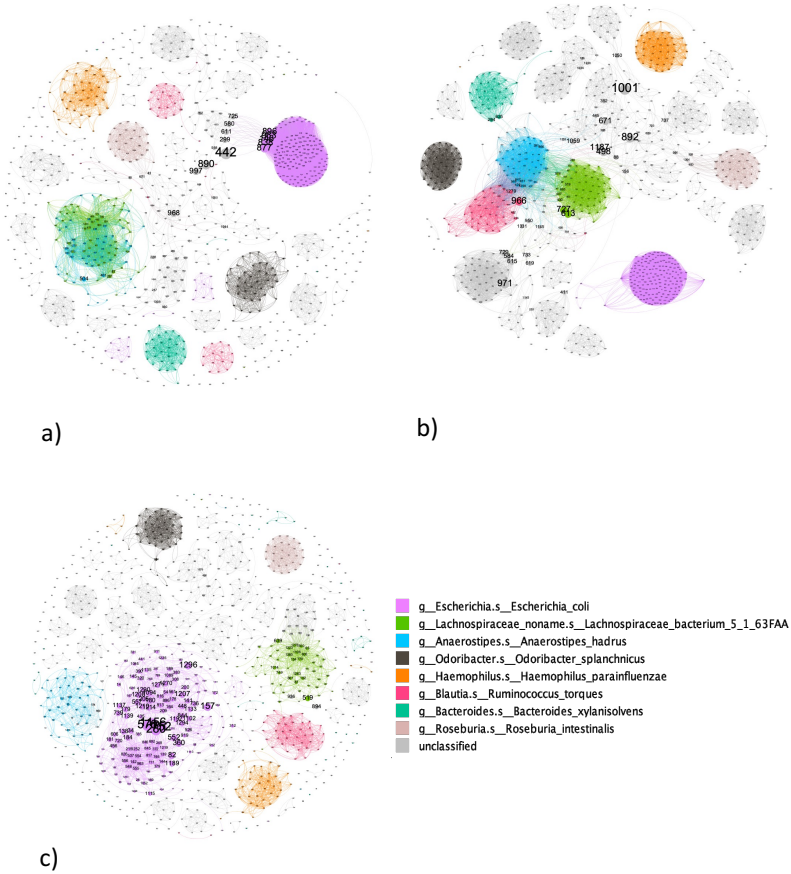


Figure 24: Projected network representation of uncommon metagenomic pathways in (**a**) NI subjects, (**b**) CD subjects, (**c**) UC subjects. Networks are obtained through a bipartite projection, and the exclusion of an edge between two nodes is made through a comparison with a null model ($\alpha < 0.05$). Node sizes are proportional to the betweenness centrality.

On the *F. plautii* side:

1. node 442 (PWY-5188: tetrapyrrole biosynthesis I (from glutamate));
2. node 611 (PWY-6122: 5-aminoimidazole ribonucleotide biosynthesis II);
3. node 725 (PWY-6277: superpathway of 5-aminoimidazole ribonucleotide biosynthesis);
4. node 299 (PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis (meso-diaminopimelate containing));
5. node 580 (PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I).

Curiously there was also a *C. bolteae* pathway well-connected to the *E. coli* main module (node 1143). The remaining part of the *E. coli* group was represented by fatty acid metabolism pathways (node 140) or pathways involving mannose biosynthesis, which were connected to other species' pathways. Apart from these nodes scattered around the network, the species group had two main ball-shaped modules, thanks to the well-connected nature of the nodes inside the module. Another interesting community was the module formed by the nodes of two different species, *Lachnospiraceae bacterium 5 1 63FAA* and *A. hadrus*. The edges connecting the nodes of the two species were so dense that the two groups formed a unique module.

In the CD network projection, there were 953 nodes and 24298 edges. One of the immediately visible properties of the projected network was the isolated module composed of *E. coli* pathways. Compared to the NI case, there were no connections to the other species nodes. *R. torques* module was connected to *A. hadrus* module, which in turn was connected to *L. bacterium 5 1 63FAA* module. 5 nodes of *the A. hadrus* community were connected to most of the nodes of the other two species exhibiting high betweenness centrality; they were:

1. node 1298 (VALSYN-PWY: L-valine biosynthesis);

2. node 429 (PWY-5104: L-isoleucine biosynthesis IV);
3. node 461 (PWY-5659: GDP-mannose biosynthesis);
4. node 847 (PWY-6700: queuosine biosynthesis);
5. node 565 (PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I).

Similarly to the NI case, node 155 was connected only to pathways of other species. Another difference with the NI projected network was the edges of *R. intestinalis*, which in this case were connected to *B. instestinihominis*, whereas, in the NI case, they were connected to nodes 671 and 41.

The UC network projection had 953 and only 5432 edges. The first striking property of this network was that there were much fewer edges compared to the previous two cases. It was not possible to say much about the module of *E. coli*. On the other hand, it was possible to observe that *L. bacterium 5 1 63FAA* and *A. hadrus* were completely separated. Furthermore, the *L. bacterium 5 1 63FAA* module was connected to the *R. hominis* module through several nodes. Related to *the R. torques* group, we had one main module similar to the other cases, two isolated nodes (node 1047 as the NI case and node 569) and two nodes connected between them (nodes 711 and 597), but separated from the rest of the pathways.

4.3 Multilayer analysis results

The overlap network was built considering the overlap of the three correlation networks in the previous results. We observed that the degree sequence of the overlapping network was highly correlated with all the three diagnosis layers, whereas the degree sequence in the aggregated network was correlated only with the degree sequence in the NI correlation network (0.807), whilst it had a low Kendall correlation with the degree sequences in the CD and UC correlation networks. Unsurprisingly, the degree sequence in the NI correlation network was uncorrelated to

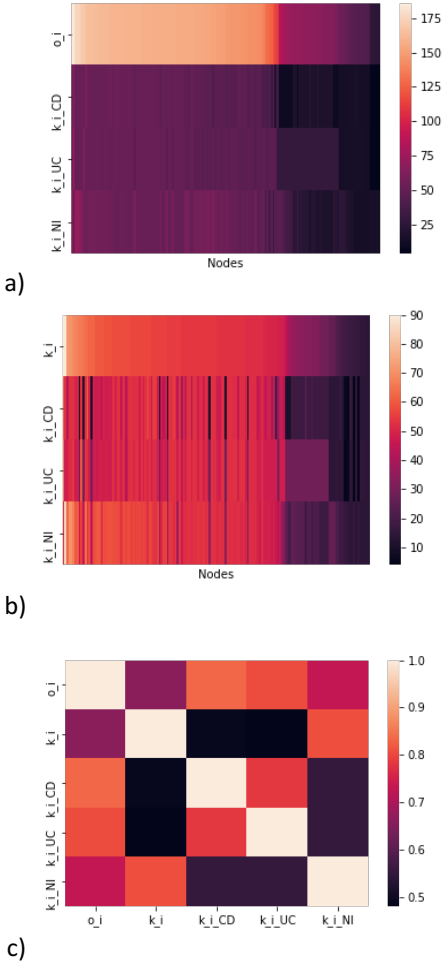


Figure 25: Colorbar plots for prevalent pathways comparing the degree of the different layers. In Figure (**a**), the first colorbar plot represented the weighted degree of the nodes in the overlap networks compared to the degree of the same nodes in the different layers. In Figure (**b**), the second colorbar plot represented the degree of the nodes of the aggregated networks compared to the degree of the same nodes in the different layers. In Figure (**c**), the heatmap shows the Kendall correlation of the degree sequences in the various networks (on the x-axis from left to right: overlap, aggregated, CD, UC, NI).

CD and UC correlation networks, which, in turn, were well correlated between them (0.782999), see Figure 25.

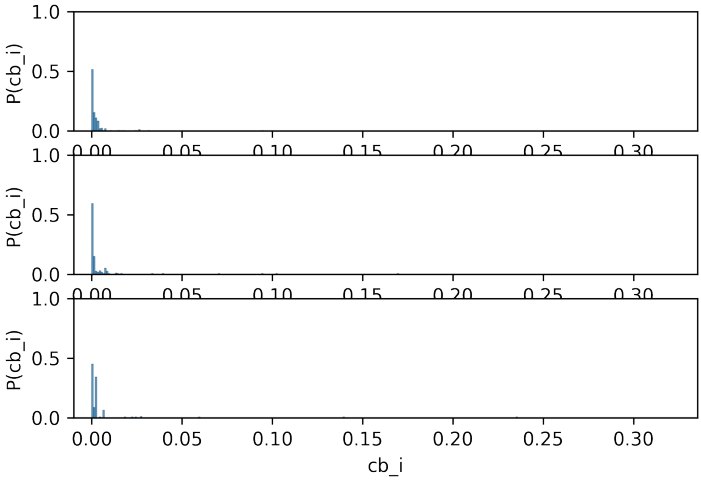


Figure 26: Comparison of the betweenness centrality distributions of the three layers (top to bottom: NI, CD, UC) for prevalent pathways.

For prevalent pathways, as it was possible to observe in Figure 26, the three distributions were quite different from each other. The CD distribution has the highest number of nodes having zero or close to zero betweenness centrality (about 60%), whereas the UC distribution had the fewest nodes with zero or close to zero betweenness centrality. It was possible to recognise a tail in the NI and CD distributions, wherein the former was more pronounced whilst the latter was flatter. In CD distribution, there were also several high centrality nodes, which were not present in the NI distribution. On the other hand, in the UC distribution, it seemed that the probabilities of each betweenness centrality were more irregular, with 30% of low centrality nodes (the second peak in the histogram) and about 10% of medium centrality nodes. These results show that, on average, nodes in networks representing microbiome of disease states have higher betweenness centralities (NI= 0.00309, CD=0.00518,

UC= 0.00503). Nevertheless, the CD distribution has values at the extremes, whereas the UC distribution has more medium centrality nodes.

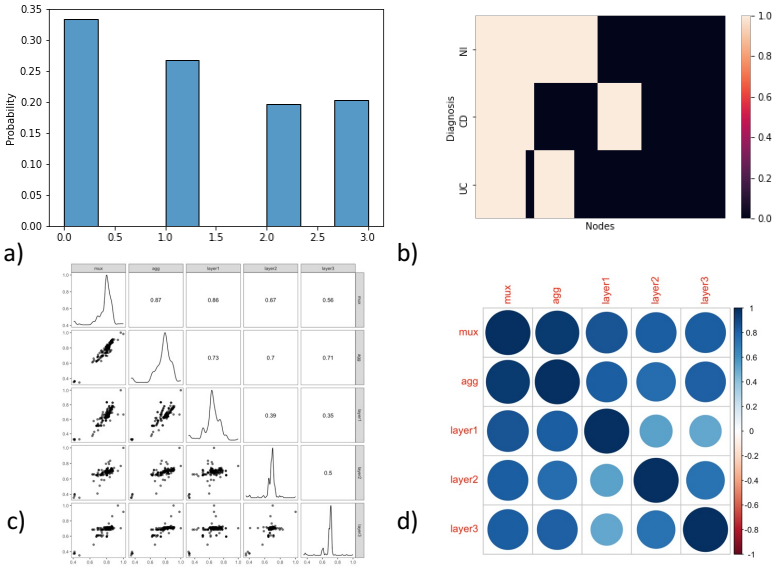


Figure 27: Plots related to the betweenness centrality for prevalent pathways. In Figure (**a**) we analysed the centrality of the nodes across the three layers. In Figure (**b**), we explore in a more detailed way the betweenness centralities of the nodes in the different diagnoses. In Figures (**c**) and (**d**), we analysed the correlation of the multilayer centralities and single layer centralities.

Considering Figure 27, most of the nodes did not exhibit any relevant betweenness centrality in all the three networks. On the contrary, about 20% of the nodes were central in the three layers at the same time. The nodes that were central in only one layer but not in the other two were slightly above the 25%. The central nodes corresponding to a disease state were approximately the 45% of the nodes. In particular, we could observe that a few nodes that were central in the NI layer and the CD layer were not central in the UC layer. On the other hand, nodes that were central in both the NI layer and the UC layer were not central in

the CD layer. A great number of nodes were very central only in the CD layer, whereas we did not find nodes that were exclusively central in the UC layer. We obtained that the centralities in the multilayer network were highly correlated with the centralities in the aggregated network ($\rho = 0.87$); the multilayer network also has a high correlation with the NI layer ($\rho = 0.86$). On the contrary, the centralities of the NI layer were poorly correlated with both the CD layer and the UC layer, as expected.

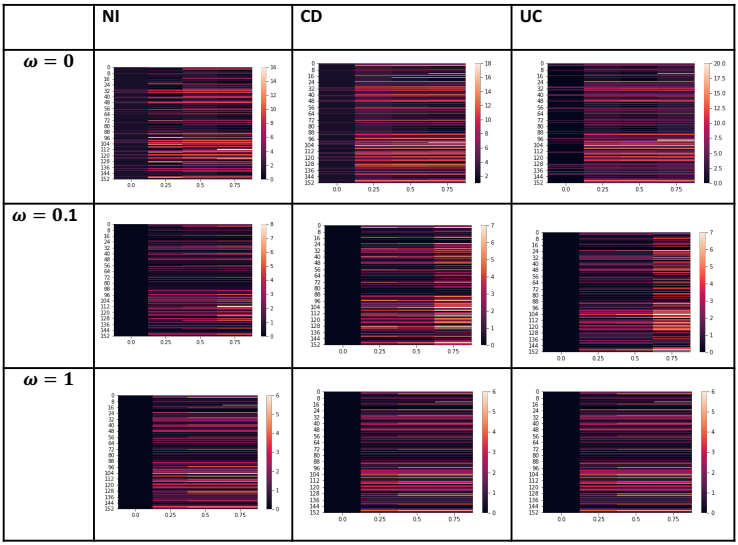


Figure 28: Analysis of the community detection in the multiplex network composed of NI, CD, and UC layers for a range of different resolution parameters $\gamma = [0, 1]$ on the x-axis and three different coupling parameters $\omega = (0, 0.1, 1)$ in the case of prevalent pathways.

For the coupling parameter $\omega = 0$, we found that there were up to 20 modules, whereas in the community detection with $\omega = 0.1$ there were up to 8 modules and $\omega = 1$ up to 6 modules. In particular, we noticed that for the low levels of resolution parameter ($\gamma = 0$) and for $\omega = 0$ there were two modules in all the layers, but for $\omega > 0$ and $\gamma = 0$, there was only one module. The community detection with $\omega = 1$ was the same for

all the layers, see Figure 28.

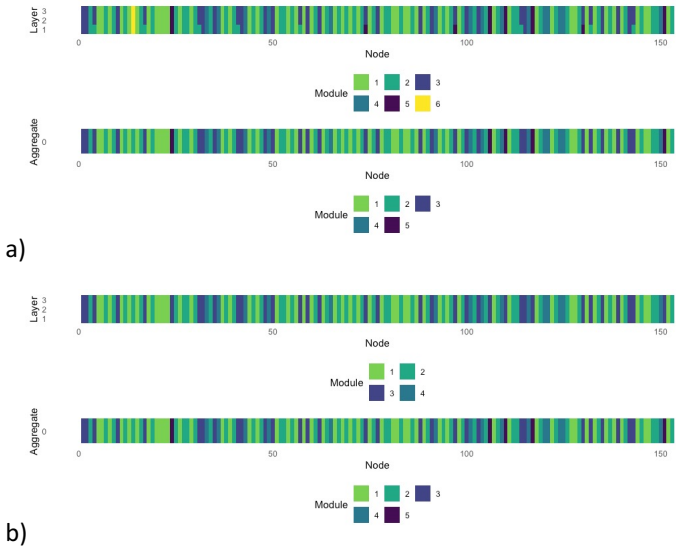


Figure 29: Comparison of the community detection with Infomap in the soft partitioning case (**a**) and in the hard partitioning case (**b**), for prevalent pathways.

In the multiplex network, we observed 6 modules in the soft partitioning case, whereas in the hard partitioning case, there were 4 modules only. The community detection in the aggregate networks was the same in both the hard partitioning case and the soft partitioning case. Focusing on the soft partitioning case, we noticed that only some of the nodes in layer 1 (NI) belonged to more than one module simultaneously (Figure 29).

For common pathways, in Figure 30, we could notice a very high correlation between the aggregated network degree sequence and the UC correlation network degree sequence (0.85). Similarly, there was a high correlation between the overlap network degree sequence and the UC correlation network degree sequence (0.78). By contrast, CD and NI correlation network degree sequences were poorly correlated between

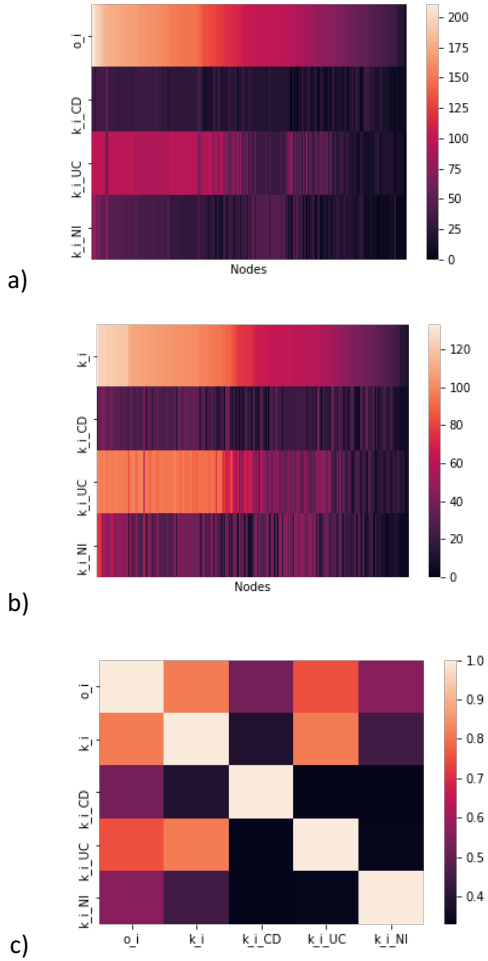


Figure 30: Colorbar plots for common pathways comparing the degree of the different layers. In Figure (**a**), the first colorbar plot represented the weighted degree of the nodes in the overlap networks compared to the degree of the same nodes in the different layers. In Figure (**b**), the second colorbar plot represented the degree of the nodes of the aggregated networks compared to the degree of the same nodes in the different layers. In Figure (**c**), the heatmap shows the Kendall correlation of the degree sequences in the various networks (on the x-axis from left to right: overlap, aggregated, CD, UC, NI).

them and with all the other networks.

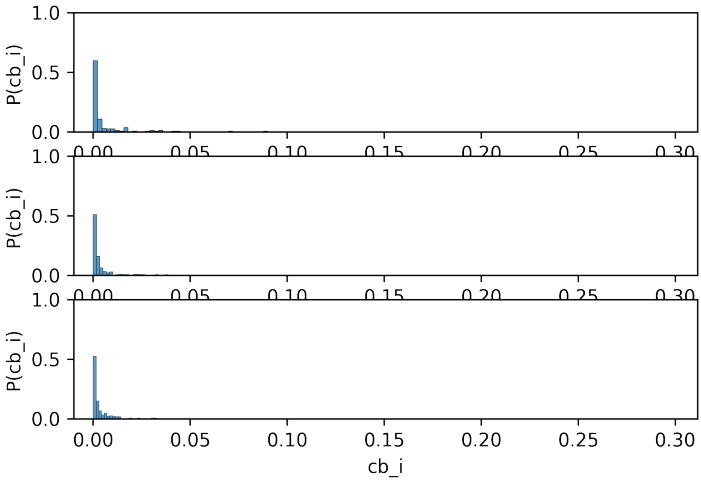


Figure 31: Comparison of the betweenness centrality distributions of the three layers (top to bottom: NI, CD, UC) for common pathways.

As can be seen in Figure 31. CD and UC profiles were quite similar, with approximately half of the nodes with 0.0 betweenness centrality. On the other hand, the NI correlation network had more than 0.0 betweenness centrality nodes but also several high centrality nodes at the expense of low-medium centrality nodes.

About 25% of the nodes in the common bacterial pathway multiplex network were central in all three layers at the same time, whereas approximately the 28% of nodes were peripheral in all three layers. The nodes with differential betweenness centrality (for example central in one layer and not in the other two, or central in two layers and not in the remaining one) were around the 45%; as it was highlighted in the Figure 32, except from the nodes central in all the three layers, those central in NI were central also in UC only, whereas those central in CD were not central either in NI or UC. It seemed that the centralities of some of the nodes in the CD network shifted to other nodes of the network. We

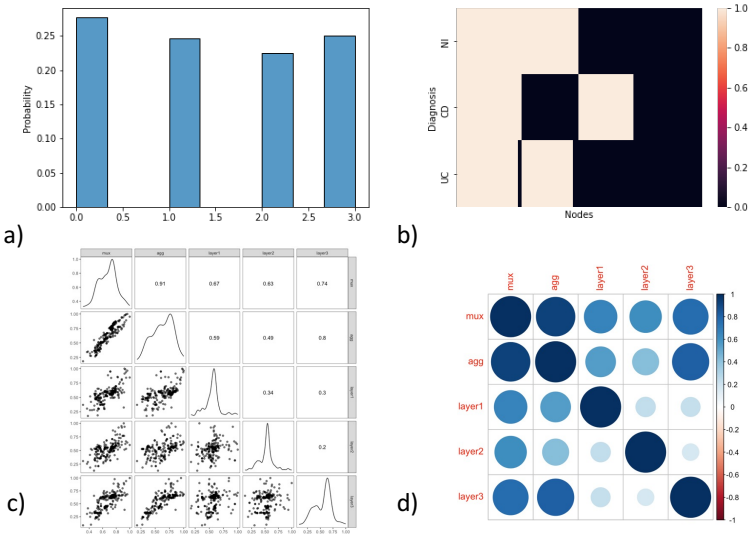


Figure 32: Plots related to the betweenness centrality for common pathways. In Figure (**a**) we analysed the centrality of the nodes across the three layers. In Figure (**b**), we explore in a more detailed way the betweenness centralities of the nodes in the different diagnoses. In Figures (**c**) and (**d**), we analyzed the correlation of the multilayer centralities and single layer centralities.

noticed that there was a poor correlation of the centralities among the nodes of the single layers. On the other hand, the centralities of nodes in the UC layers were strongly correlated to the multiplex centralities (0.74) and the aggregated centralities (0.8).

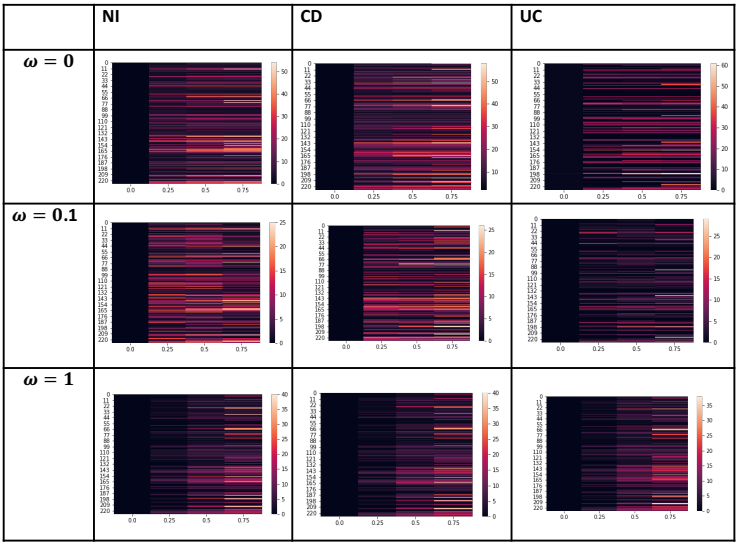


Figure 33: Analysis of the community detection in the multiplex network composed of NI, CD, and UC layers for a range of different resolution parameters $\gamma = [0, 1]$ on the x-axis and three different coupling parameters $\omega = (0, 0.1, 1)$ in the case of common pathways.

In Figure 33, we could find that for an $\omega = 0$, the NI multiplex network could be divided into fewer communities compared to the nodes in the CD and the UC layers. By increasing the resolution, it seemed that for all the ω's, the nodes in the UC layer remained in the same community as in the previous resolutions, i.e. less variability. On the other hand, NI and CD were similar.

In Figure 34, we compared the Infomap community detection in the soft partitioning case with the hard partitioning case for common bacterial pathways. We noticed that there were 9 modules in the soft parti-

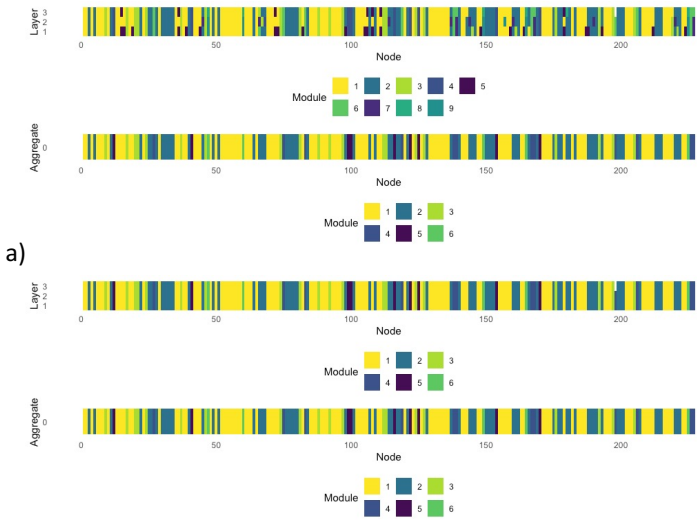


Figure 34: Comparison of the community detection with Infomap in the soft partitioning case (**a**) and in the hard partitioning case (**b**), for common pathways.

tioning case, whereas there were 6 modules in the hard partitioning case. The aggregated networks had the same Infomap community detection, as we used the same multiplex in both cases.

For uncommon pathways, as it is possible to see in Figure 35, the degree sequence in the NI correlation network was well-correlated with the degree sequences of the overlap and the aggregated network. By contrast, there was a low correlation between the degree sequences of the CD and UC correlation networks with the degree sequence of the NI correlation network. Surprisingly, there was a low correlation also between the degree sequences of the CD and UC correlation network, meaning that all the three layers had different degree structures. The highest correlation can be found between the degree sequences of the overlap network and the aggregated network.

The betweenness centrality distributions in the figures above showed that half of the nodes in the NI and CD correlation networks had a centrality of 0.0, whereas only about 40% had a zero betweenness centrality in the UC correlation networks. Furthermore, the centralities in the NI and CD correlation networks had distributions with heavier tails compared to the centrality distribution in the UC correlation network, see Figure 36.

As it is possible to see in Figure 37, comparing the centralities in the three layers at the same time, we noticed that around 30% of the nodes were peripheral in the three networks at the same time. Moreover, about 22% of the nodes were central in the whole multiplex. Again, about the 40% of the nodes were partially central in the multiplex, with a great portion of the nodes either central only in the CD layer or central in the NI and the UC layer. In the figures above, we observed that the correlations among the centralities of the nodes in the different single layers of the multiplex were very low (< 0.25). Centralities in the NI layer had a low correlation also with the aggregated network.

By observing Figure 38, we noticed that for $\omega = 0$ the community in the NI layer were fewer than the communities in the CD and the UC layers. By contrast, for $\omega = 0.1$ the number of the communities was comparable through all the resolution parameters and, finally, for $\omega = 1$

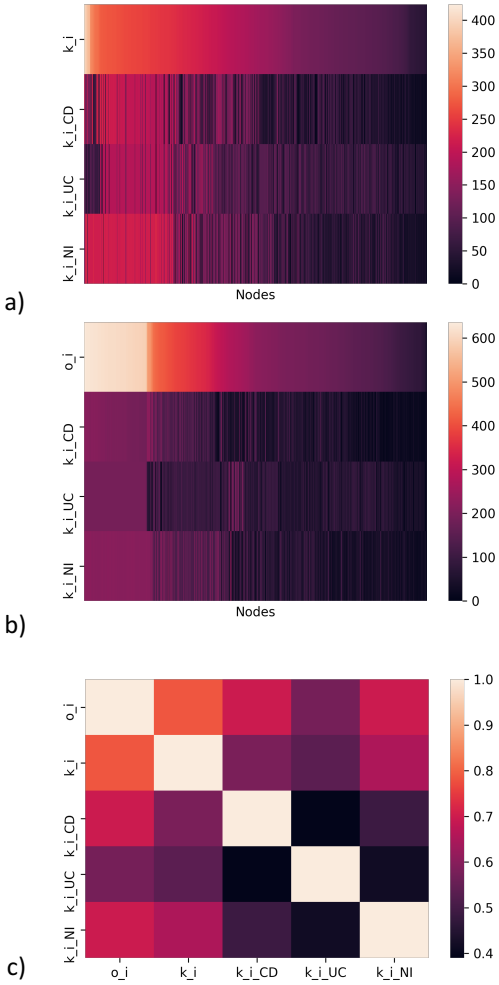


Figure 35: Colorbar plots for uncommon pathways comparing the degree of the different layers. In Figure (**a**), the first colorbar plot represented the weighted degree of the nodes in the overlap networks compared to the degree of the same nodes in the different layers. In Figure (**b**), the second colorbar plot represented the degree of the nodes of the aggregated networks compared to the degree of the same nodes in the different layers. In Figure (**c**), the heatmap shows the Kendall correlation of the degree sequences in the various networks (on the x-axis from left to right: overlap, aggregated, CD, UC, NI).

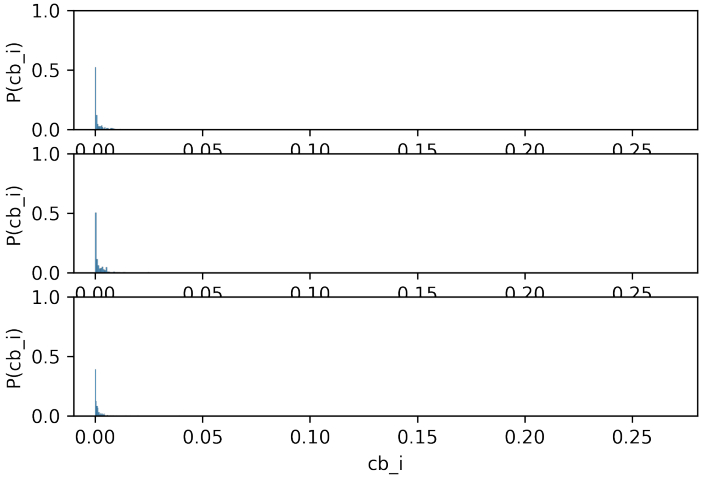


Figure 36: Comparison of the betweenness centrality distributions of the three layers (top to bottom: NI, CD, UC) for uncommon pathways.

the number of communities was much less for the nodes in the NI layer than the nodes in the other layers.

In Figure 39, we could observe 22 modules in the soft partitioning case and 9 modules in the hard partitioning case. In both cases, there were two regions in which the consecutive nodes belonged to the same community (around node 100 and around node 375).

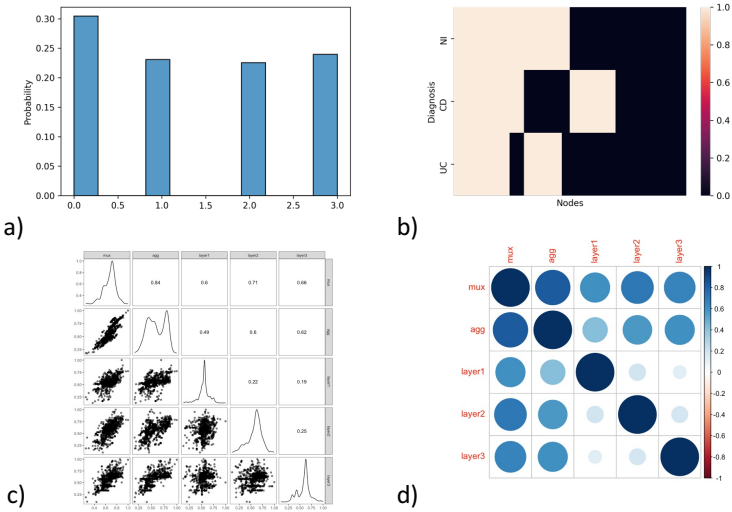


Figure 37: Plots related to the betweenness centrality for uncommon pathways. In Figure (**a**) we analysed the centrality of the nodes across the three layers. In Figure (**b**), we explore in a more detailed way the betweenness centralities of the nodes in the different diagnoses. In Figures (**c**) and (**d**), we analyzed the correlation of the multilayer centralities and single layer centralities.

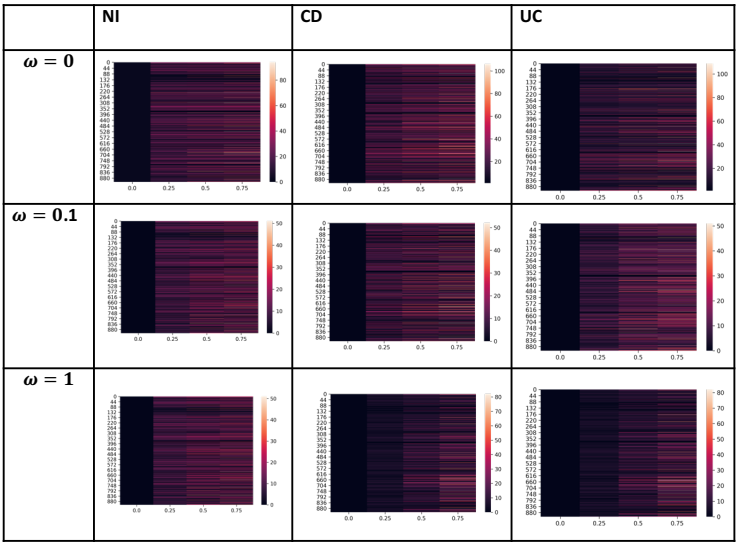


Figure 38: Analysis of the community detection in the multiplex network composed of NI, CD, and UC layers for a range of different resolution parameters $\gamma = [0, 1]$ on the x-axis and three different coupling parameters $\omega = (0, 0.1, 1)$ in the case of uncommon pathways.

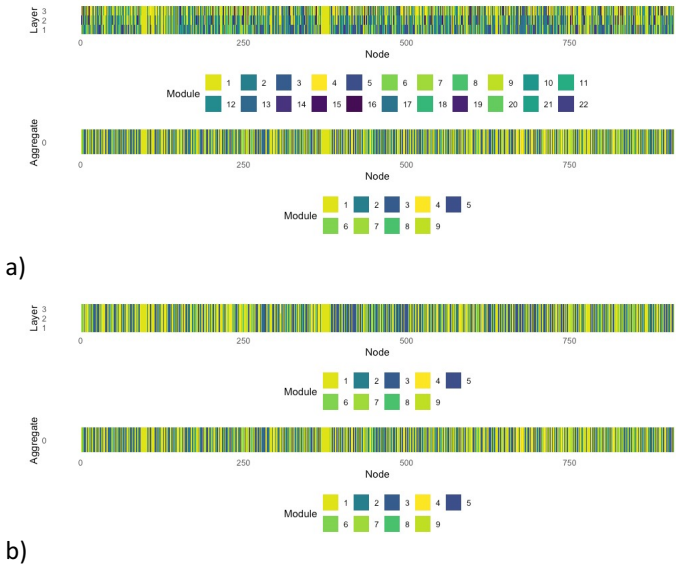


Figure 39: Comparison of the community detection with Infomap in the soft partitioning case (**a**) and in the hard partitioning case (**b**), for uncommon pathways.

Chapter 5

Discussion

5.1 Correlation networks

Considering the prevalent pathways resulted in the NI pathways being better connected (more correlated) to the other pathways than the same pathways in the IBD (CD and UC) correlation networks. That meant that there was a broader interaction between the nodes in the NI correlation network, where the modularity was lower, i.e., there were fewer but larger modules detected. In the literature, the reduction of *F. prausnitzii* has been associated with IBD; (Cao, Shen, and Ran, 2014), nevertheless, in our results, we recognised that instead of a decrease in the quantity of *F. prausnitzii* pathways expressed, there was a change in the wiring in the metagenomic network. In particular, what changed from the NI correlation network and the IBD correlation networks were the bridge pathways connecting the module of unclassified pathways and the module containing the *F. prausnitzii* pathways. For instance, in the NI correlation network, the bridge pathway was node 821, whereas, in the IBD networks, the bridge pathways between the two modules were node 105. The correlation between these bridge nodes and the nodes in the aforementioned modules meant that both modules relied on the pathway to function correctly. On the one hand, the central substance was the tetrasaccharide stachyose which in the pathway is degraded into UDP-alpha-D-glucose

and has been recognised as a potential probiotic against enterotoxigenic *E. coli* (Xi et al., 2020); on the other hand, there was the coenzyme A, which has a fundamental role in the metabolism and, in particular, it is important in the oxidation of fatty acids. Several studies linked the alteration of fatty acid production to the IBDs, hence, this change in the centrality of the pathway related to this substance could be investigated further to explain the origins of IBDs (Xi et al., 2020). The fact that the modules of *Bacteroides* in IBD networks corresponded to its species whilst in NI network are gathered in one module could demonstrate that in IBD the different *Bacteroides* species proliferates the gut independently. This could confirm the meta-analysis by Zhou and Zhi (2016), who showed that lower levels of *Bacteroides* were associated with IBDs. Other pathways; differentially wired between NI and IBD networks are those involving the bacterial metabolite 5-aminoimidazole ribonucleotide (nodes $579, 610, 724$), these nodes were behaving as bridges in the NI correlation network, by contrast, in the IBD correlation networks, they were substituted by a unique bridge node (node 105). Similar results were found in Yongshun Ma et al. (2021), where the same pathways were recognised as the differential between NI and IBD. The low NMI between NI and UC community detections was very significant. By using permutation tests, we highlighted several differential pathways, in particular, node 462 connected with node 137 was differential between NI and CD. The two main substances involved in these pathways are strictly linked to each other and rhamnose produced by *R. gnavus*, in particular, has been recognized as an inflammatory polysaccharide (Henke et al., 2019). The fact that there were not any differentially expressed pathways between NI and UC could mean that the main differences between the two diagnoses lay in the wiring scheme of the pathways rather than their expression levels. After feeding the DE pathways to the DIAMOnD algorithm, we obtained that also *B. ovatus* consistently with the literature was considered a disease module, in fact, it was linked to the antibody response that generated an IBD (Saitoh et al., 2002). Moreover, the DIAMOnD algorithm linked *E. rectale* to CD diagnosis as well, but the role of this species had not been explored in deep either in the current study or in

the past literature.

In the single-layer analysis and considering the common pathways, we have that in the literature, less diversified bowel microbiota has been linked to the IBDs (Tannock, 2010); this was supported by the change in the relationship between the *B. caccae* module and the *B. ovatus* module between NI and CD networks. In the NI network, *B. caccae* module is connected to *B. ovatus* through nodes 1212 and 316, whereas, in the CD network, the only bridge is node 531. Another striking difference is the positioning of the nodes belonging to *Ruminococcus torques*. In the NI network, they are part of the module containing *B. caccae*; by contrast, in the CD network they are connected to the rest of the network through *R. inulinivorans* nodes. The aforementioned bacteria are known as butyrate-producing bacteria species, and they are known to be reduced in CD patients (Nishida et al., 2018; Takahashi et al., 2016). Node 77, in the CD correlation network, had a higher betweenness centrality; on the other hand, in the UC correlation network, the same exhibited a behaviour similar to the node 77 in the NI correlation network. This result went in contrast with the results in Y. Zhu et al. (2020), which found COA-PWY-1 differential in UC. The discussion of the results from the UC single-layer network is harder because the percolation threshold was very low due to two links with a very low correlation weight, as a consequence, the network resulted very connected and less modular in the respect of the previous two cases. In the UC correlation network, the node 829 is very central and several nodes in the three modules are correlated to it. Guanosine has been observed to have protective effects against substance-induced colitis in murine models (Zizzo et al., 2019) and a reduction of this metabolite certainly has a repercussion on the other bacterial pathways. *Ruminoccocus obeum* became central in the UC correlation network behaving as a bridge between two modules, in particular, nodes 928 and 1306. These two pathways are strictly intertwined to each other as L-valine is a substance involved in both pathways. *Flavonifractor plautii* is the bacteria species that in the UC correlation network linked the *E. rectale* module to the *Bacteroides* module, *F. plautii* plays a central role in the UC network and its central role is sup-

ported by a study which claimed that *F. plautii* together with other three species were increased in the mucosa of UC subjects, hence can be considered as a possible pathogen (W. Li et al., 2021). The three most central *F. plautii* pathways were nodes 1012, 944, and 1320, it is possible to recognize again the same two pathways PWY-7111 and VALSYN-PWY, which were central for *R. obeum*. The NMI calculated showed that there is a discreet overlap between NI and CD community detection, but the measure was not significant as it was not influenced by the change of the networks through the permutations of the subjects in the two diagnoses. On the other hand, the low NMI between NI and UC was quite significant. By considering the differential expression of the bacterial pathways between CD and NI, several *R. intestinalis* pathways resulted in differential pathways in accordance with (Vich Vila et al., 2018), in which the bacteria species resulted to be decreased. By contrast, *Alistipes putredinis* and *Bacteroides xylanisolvens* pathways were differential between UC and NI; a study observed that the former species is depleted in UC paediatric subjects (Meij et al., 2018), whereas there are no current studies linking the variation of the latter species to UC. By using the DIAMOnD algorithm to expand the number of differentially expressed pathways between NI and CD, there were mainly pathways of the unclassified species, whereas when NI and UC were considered, the algorithm showed the pathways belonging to *Ruminococcus torques*. This bacteria species resulted in being wired differently in the CD correlation network; by contrast in the UC network it is differentially expressed; this can lead to thinking that in one case, there could be a differential networking, whereas in the other case, there could be a differential expression of the pathways instead. From an analysis of the degree centralities, it was noticed that differential CD pathways and the additional DIAMOnD pathways had a lower degree centrality; this could mean that more isolated bacterial pathways have a different genetic expression in respect of the correspondent pathways in the other networks. A possible intervention to re-establish a healthy gut microbiome could be to examine the covariance between the differential pathway and the other pathways in the control subjects and re-establish local connectivity similar to the one in the NI network. Also, the differ-

ential UC pathways exhibited a different degree centrality, in particular, the degree centrality was higher in the UC correlation network in the respect of the same pathways in the NI network. This was mainly due to the greater link density of the UC network; the higher connectivity in the medium range of the bacterial pathways could be a sign of UC.

In the range of uncommon bacterial species, we can observe the *E. coli* module; in the literature, this bacterial species has a recognised role in developing IBD (Rhodes, 2007; Schirmer et al., 2019). It was possible to observe the different interplay between *E. coli*, *V. parvula* and *B. fragilis* across the different diagnoses. The increase of *E. coli* and *B. fragilis* in IBD was observed in a previous study (Keighley et al., 1978), but our results provide a piece of additional information about the differential wiring scheme of the aforementioned species. In particular, it seemed that *V. parvula* pathways mediated the connection of *E. coli* with the other module in the correlation network. In particular, in the NI correlation network, *V. parvula* pathways were in the same module of *B. fragilis* pathways which were connected to the rest of the correlation network. In the CD correlation network, *V. parvula* pathways were included in the *E. coli* module just to remark how close the two bacterial species were, but if, on one hand the relationship between *E. coli* and *B. fragilis* has been already studied, the effect of *V. parvula* on *E. coli* has to be investigated yet in the literature. In the UC correlation network, *V. parvula* formed an almost completely isolated module far from the *E. coli*; this result could differentiate the connectome of the UC microbiome from the connectome of the CD microbiome. The isolation of the *E. coli* module in the UC correlation network could represent further the peculiar features of the particular form of IBD. This isolation meant that there were no correlations with the other pathways, and the pattern of metagenomic expression across the samples is correlated only inside the same bacterial species. In the NI network, *E. siraeum* and *R. gnavus* pathways were the two main bridge pathways between *E. coli* and the rest of the network; it could be possible to hypothesise that re-establishing a connection between *E. coli* module with the aforementioned bacterial species could lead back to healthy gut microbiota. In the CD correlation network, *R. intestinalis* pathways

had the role of bridge pathways, and, in fact, by using the permutation tests between NI and CD samples, we obtained that the most differential pathways were *R. intestinalis* pathways. In the literature, this bacterial species, which has anti-inflammatory properties on the intestinal walls, was depleted in IBD subjects; nevertheless, the complete mechanisms underlying its protective action against IBD are still unknown (C. Zhu et al., 2018; Hoffmann et al., 2016). Extending the set of differentially expressed pathways, we obtained several *E. coli* pathways confirm that in CD, differential expression and differential wiring of the *E. coli*. With the permutation tests of NI and UC samples across all the pathways in this range, we found that mainly *B. xylanisolvens* pathways were differential. The bacteria; species is linked to *R. intestinalis* because of its capacity to degrade dietary fibres commonly assumed by humans; nonetheless, the two species are not in competition as they attach to different forms of substrates (Mirande et al., 2010); also the differential pathways belonging to *O. splanchnicus* showed that this bacterial species behaved differently between NI subjects and UC subjects, in fact, although it is pathogenetic in domestic animals, it resulted to be useful in the ecology of the human gut microbiota (Gomez-Arango et al., 2016), specifically the bacteria produce short-chain fatty acid, such as butyrate, which has a protective role against the IBD (J. Chen and Vitetta, 2020). By analysing the pathways from the DIAMOnD algorithm, there were several *R. obeum* pathways which is in accordance with a study that showed high percentage of this species in UC subjects (Lepage et al., 2011). The NMI calculation did not highlight any significant overlap between the three diagnoses, as the p-value were quite high.

By analysing the eigenvector centralities of the CD differential pathways, we noticed that these pathways were generally much central in the CD correlation network compared to the NI correlation network. Eigenvector centrality is the importance of a node calculated as the combination of the eigenvector centrality scores of the neighbouring nodes. In the same way, also betweenness centrality and degree centrality scores were high for the differential pathways in the CD correlation network compared to the NI network; this could mean that the pathways that

contributed to the CD diagnosis were connected to neighbouring pathways and behaved like bridges more than the correspondent pathways in the NI correlation network, where these pathways were more peripheral and less connected. The differentially expressed UC pathways did not show any strong differential networking that led to a relevant change in the centrality scores.

5.2 Projected networks

By building the projected networks from the bipartite networks and keeping the most significant edges, we obtained three networks of prevalent pathways with comparable modularity; this is a consequence of how we built the projected networks; in fact, the pathways that are joined by an edge are those appearing importantly expressed in the same samples, multiple times, more than it would happen in a random bipartite network. There is an important difference between the single-layer correlation network and the bipartite network as in the correlation network; we consider relevant in the network also the strong anti-correlations. In the projected bipartite network, it was natural that pathways of the same bacterial species were collected in the same community because they were linked to each other. Nevertheless, it was interesting that some modules were internally more connected than others; therefore, pathways centralities changed from one network to the other. This bipartite method was very effective for highlighting the bacterial modules present across the different samples; it seemed that all the diagnosis networks had the same number of modules for the projected bipartite networks. On the other hand, the betweenness centralities of the different pathways changed from disease network to disease network, because the density of the network was different in each case.

From a general point of view, for the prevalent pathways, the NI projected network was more densely connected than the CD and UC projected networks; by contrast, the UC projected network was the least densely connected. In this case, a less connected projected network corresponded to a less diversified microbiota in the subjects with UC and

a similar result was observed by Nemoto et al. (2012). The modularity and the community detection reflect the status of the microbiota in this range of pathways; the closer is the number of modules to the number of species, the better is the community detection because each group of nodes (and pathways) is functional to one species unless multiple species co-exist. In this case, the community detection applied to the projected networks seemed to perform well. The measurement of betweenness centrality showed that the centrality measures involving nodes 1222 and 748 in *F. prausnitzii* were higher in the CD projected network compared to the betweenness centralities of the same pathways in the NI projected network. This could mean that the aforementioned pathways are often highly expressed along with the neighbouring pathways in the CD subjects, and by a consequence, several metabolic pathways of the species relied on the galactose degradation I and D-galactose degradation V. Similarly, Jiang et al. (2021) observed that carbohydrate degradation metabolic pathways were depleted for bacterial species beneficial to the gut. In the NI projected network, the most central nodes in the *F. prausnitzii* module were nodes 724 and 610. It is important to notice the fundamental role of the substance 5-aminoimidazole ribonucleotide, which also appeared in the single-layer analysis, and, in the past studies, it was found differential between healthy control subjects, IBD subjects and colon-rectal cancer subjects (Yongshun Ma et al., 2021). In the UC projected network, nodes 158 and 517 had a high betweenness centrality; these two nodes are so central because they held together with the two groups of the *F. prausnitzii*. In the UC case, many pathways are dependent on these two nodes, and the literature agrees that the metabolite produced in the former pathway could have a potential medical effect. Moreover, also node 67 exhibited a high betweenness centrality measure that could be related to a change in the digestion of sugars observed in patients with irritable bowel disease syndrome (Mack et al., 2020).

In the UC bipartite network of the common bacterial pathways, the density of the connections was again lower compared to the densities of the NI and the CD projected networks; this could be translated into less co-expression of the bacterial pathways in the same samples, hence,

less interaction among bacterial species and pathways as in the previous case. The change in the relationship between *B. caccae* module and the *B. ovatus* module was already observed in the single-layer analysis of the correlation networks. Nevertheless, the change was even clearer in the bipartite analysis, where it was possible to observe that in the NI and UC projected networks, *B. caccae* module communicated with *the B. ovatus* module, whereas in the CD projected networks, it was linked to *A. putredinis* module through 789, which was joint to ketodeoxyoctonate and co-enzyme A biosynthesis pathways in *B. caccae* module. The aforementioned substances are fundamental for the functioning of a bacterium; the former is used to form the outer membrane, whereas the latter is needed as a co-factor in many enzymatic processes. Another important difference that was not shown in the single-layer analysis was the fact that the *E. eligens* module, which was connected to the unclassified species module in the NI projected network through node 829, whilst in the CD and UC projected networks, the same bacterial module resulted isolated from the rest of the network. This setting could be typical of healthy gut microbiota; in fact, *E. eligens* is a butyrate-producing bacteria (Mukherjee et al., 2020) and its metabolic processes like nodes 1315 and 939 could be facilitated by node 829.

In the uncommon bacterial range of pathways, we could observe that the difference in the edge densities was even more evident in the respect of other bacterial ranges; in fact, the UC projected network counted half the edges compared to the NI projected network and one-third of the edges of the CD projected network. As in the first case, this could be explained by the reduced diversity of the gut microbiota of the UC subjects (Nemoto et al., 2012). In the case of the CD projected network, we had a greater complexity in the bacterial interactions for this range of pathways; in this range there were no pathways of *Firmicutes* species, which had a lower diversity in a previous study (Manichanh et al., 2012b). In the uncommon bacterial pathways of the NI projected network, the *E. coli* module was connected to *the C. boltae*, which, in turn, was linked to *B. longum* module. *B. longum* is a bacterial species that can have anti-inflammatory properties in the human gut (Singh et al., 2011). By con-

trast, in the CD projected network, *E. coli* was connected to the rest of the network through two *Dorea longicatena* pathways, which were nodes 680 and 81 and were connected to node 697. On the other hand, in the UC projected network, two *E. coli* nodes were connected to six *C. comes* nodes, showing the existence of an interaction between the two species in the UC diagnosis.

5.3 Multilayer networks

For the prevalent pathways, the multilayer analysis of the metagenomic data of the microbiome of the NI subjects and the subjects with IBD gave a wider view of the difference between the IBD correlation networks and the NI correlation network. From the point of view of the degree centrality, we obtained that there were remarked differences between the degree of the nodes in the IBD correlation networks and the NI correlation network, these differences were highlighted by the Kendall correlation coefficient, which showed the high correlation between the degree sequence in the CD and the UC correlation network. It could be possible to deduce that the IBD correlation networks for prevalent bacterial pathways had specific degree sequences that were different from the NI network. In particular, it seemed that there were several nodes in the IBD correlation networks with medium degree centrality measures in the overlapping network that exhibited lower degree centralities in the IBD correlation networks compared to the NI correlation network. The low correlation of the overlap degree sequence and the aggregated degree sequence highlighted the fact that the nodes in the different layers had similar neighbouring and compressing the layers in an overlapping way, we obtained a much higher (weighted) degree centrality in the nodes. Furthermore, the comparison of the betweenness centrality showed a shift in the centralities of the nodes in the CD diagnosis, whereas the centralities in the UC diagnosis were similar to the ones in the NI diagnosis. This result could be interpreted as that the correlation network in UC diagnosis preserved the roles of core and periphery of the nodes in the NI network, hence, the pathogenesis of the disease could not be identi-

fied in a shift of the importance of the nodes, whereas the CD condition could be linked to the differential betweenness centrality. The multilayer betweenness centrality did not show any relevant result as it resulted correlated to the betweenness centralities of each other network at the same level. The community detection for each inter-coupling parameter had many similarities in the three layers. This could be explained by the fact that communities corresponded to the species group of pathways. For higher resolution parameters γ we obtained more diversified community detection, which highlighted the internal structure of the communities.

The analysis of the common bacterial pathways highlighted that for this range of pathways, the UC network exhibited nodes with higher degree centralities compared to the CD and NI correlation networks. This result showed that this range of nodes in the UC correlation network was clearly connected to more neighbours compared to the same nodes in the other layers. Also, for this range of bacterial pathways, the betweenness centrality of the nodes in the UC and NI correlation networks were similar; hence, they had the same core and the same periphery. It could be possible to explain this result by saying that the UC diagnosis could be based on a change in the wiring of the network and a change in the number of neighbours, whereas the CD diagnosis could be based on a shift of importance of the pathways. Considering the community detection in the multiplex, we know that the inter-coupling parameter ω was hard to set Didier, Brun, and Baudot (2015). Therefore we used three different parameters; for lower inter-coupling parameters, we obtained more communities, whereas, for $\omega = 1$, we obtained a less diversified community detection, with more similarity among the layers. Furthermore, we could observe that UC was already composed of smaller modules; since varying the resolution parameter, the change in the communities was minor compared to the other two layers.

The results of the uncommon bacterial pathways led to the understanding that the degree sequences of the UC correlation network were different from the degree sequences of the NI and CD correlation networks. The same nodes of the UC correlation matrix had a lower degree compared to the nodes in the other two networks. This result was in

line with the finding that the UC correlation network was less densely connected. Also, the betweenness centralities were differential, in fact, it was generally higher for the nodes in the NI and CD correlation networks. This could be translated into the fact that in these two networks the structure was developed around a few important nodes, on which the rest of the pathways relied. About the community detection, we found out that the algorithm did not differentiate much varying inter-layer coupling parameters or layers. Only for $\omega = 1$, we found half the number of modules in the NI layer.

On the other hand, in the Infomap community detection, we could observe that in the soft partitioning Infomap algorithm some nodes were shared among multiple communities. The most important nodes were those in which the module was the same in the IBD layers (second and third) and different in the NI layer.

Chapter 6

Conclusion

We analyzed and built the correlation network, the projected bipartite network and the multiplex network of the gut microbiotas of three types of samples, namely in CD subjects, UC subjects, and NI subjects. We obtained that the correlation networks and the projected bipartite networks offered the richest information, the former gave information about the co-variation of the pathways, whereas the latter gave information about the co-expression of the pathways. These two types of networks could highlight the important role of *Bacteroides* and of *F. prausnitzii*, in particular the connection of co-enzyme A with the IBD diagnosis. The multilayer network was built using as layers only the correlation networks of the different diagnoses. We could have built a multilayer by overlapping the correlation network and the projected bipartite network for each group of pathways. One of the drawbacks of our method was to subdivide the metagenomic dataset into several sections instead of using it as a whole. In fact, there could be a connection between pathways belonging to different sections. Moreover, it was possible to subdivide the data differently, e.g. by dividing the samples based on the severity of the disease or the location of the biopsy, but this different way of dividing them would reduce the number of samples in each group influencing the robustness of the statistics.

6.1 Future directions

This work can represent the starting point for several future projects, the main question is whether it is possible to create a digital twin of the gut microbiome and to get there we should go through several steps:

- we should understand the direction of the use of resources between bacterial pathways, i.e. the directions of the reactions;
- we should analyse the complexity of the gut microbiome by building the corresponding multilayer network;
- we should study the dynamics of the metabolic networks composed of the metabolic pathways found in the metagenomic or metabolomics and the dynamics of the multilayer network;
- we should investigate the control (the action of steering the system from a dysbiosis state to a healthy state) of the biological system represented by the metabolic network through prebiotics, probiotics, and antibiotics.

Appendix A

Nodes Map

Table 6: Nodes map of the nodes being cited in the research project.

Node	Pathway
2	1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis|g__Bacteroides.s__Bacteroides_ovatus
5	1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis|g__Bacteroides.s__Bacteroides_uniformis
12	ANAGLYCOLYSIS-PWY: glycolysis III (from glucose)|g__Bacteroides.s__Bacteroides_ovatus
20	ARGININE-SYN4-PWY: L-ornithine de novo biosynthesis|g__Bacteroides.s__Bacteroides_ovatus
41	ARO-PWY: chorismate biosynthesis I|g__Roseburia.s__Roseburia_inulinivorans
62	BRANCHED-CHAIN-AA-SYN-PWY: superpathway of branched amino acid biosynthesis|unclassified
67	CALVIN-PWY: Calvin-Benson-Bassham cycle|unclassified
77	COA-PWY-1: coenzyme A biosynthesis II (mammalian)|g__Bacteroides.s__Bacteroides_vulgatus
91	COA-PWY-1: coenzyme A biosynthesis II (mammalian)|g__Parabacteroides.s__Parabacteroides_merdae

(To be continued)

Node	Pathway
98	COA-PWY: coenzyme A biosynthesis I|g__Bacteroides.s__Bacteroides_xylanisolvens
105	COA-PWY: coenzyme A biosynthesis I|unclassified
128	DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I|g__Bacteroides.s__Bacteroides_ovatus
131	DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I|g__Escherichia.s__Escherichia_coli
133	DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I|g__Faecalibacterium.s__Faecalibacterium_prausnitzii
137	DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I|unclassified
146	FUCCAT-PWY: fucose degradation|g__Escherichia.s__Escheri
155	GLCMANNANAUT-PWY: superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation|g__Blautia.s__Ruminococc
156	GLCMANNANAUT-PWY: superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation|g__Dorea.s__Dorea_longic
158	GLCMANNANAUT-PWY: superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation|g__Faecalibacterium.s__F
177	GLYCOGENSYNTH-PWY: glycogen biosynthesis I (from ADP-D-Glucose)|unclassified
178	GLYCOL-GLYOXDEG-PWY: superpathway of glycol metabolism and degradation|g__Escherichia.s__Escherichia
204	ILEUSYN-PWY: L-isoleucine biosynthesis I (from threonine)|unclassified
220	NONMEVIPP-PWY: methylerythritol phosphate pathway I|g__Bacteroides.s__Bacteroides_ovatus
234	NONOXIPENT-PWY: pentose phosphate pathway (non-oxidative branch)|g__Bacteroides.s__Bacteroides_xyl
236	NONOXIPENT-PWY: pentose phosphate pathway (non-oxidative branch)|g__Blautia.s__Ruminococcus_torque
246	P161-PWY: acetylene degradation|unclassified

(To be continued)

Node	Pathway
253	PANTO-PWY: phosphopantothenate biosynthesis I|g__Anaerostipes.s__Anaerostipes_hadrus
261	PANTO-PWY: phosphopantothenate biosynthesis I|g__Bacteroides.s__Bacteroides_vulgatus
277	PANTO-PWY: phosphopantothenate biosynthesis I|unclassified
291	PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing)|g__Bacteroides.s__Bacteroides_ovatus
292	PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing)|g__Bacteroides.s__Bacteroides_uniformis
299	PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing)|g__Flavonifractor.s__Flavonifractor_plautii
316	PWY-1042: glycolysis IV (plant cytosol)|g__Bacteroides.s__Bacteroides_ovatus
362	PWY-3001: superpathway of L-isoleucine biosynthesis I|unclassified
370	PWY-3841: folate transformations II|g__Bacteroides.s__Bacteroides_vulgatus
377	PWY-4242: pantothenate and coenzyme A biosynthesis III|g__Bacteroides.s__Bacteroides_ovatus
381	PWY-4242: pantothenate and coenzyme A biosynthesis III|unclassified
394	PWY-5097: L-lysine biosynthesis VI|g__Bacteroides.s__Bacteroides_ovatus
398	PWY-5097: L-lysine biosynthesis VI|g__Bacteroides.s__Bacteroides_vulgatus
404	PWY-5097: L-lysine biosynthesis VI|g__Eubacterium.s__Eubacterium_rectale
419	PWY-5100: pyruvate fermentation to acetate and lactate II|unclassified
428	PWY-5103: L-isoleucine biosynthesis III|unclassified

(To be continued)

Node	Pathway
429	PWY-5104: L-isoleucine biosynthesis IV|g__Anaerostipes.s__Anaerostipes_hadrus
442	PWY-5188: tetrapyrrole biosynthesis I (from glutamate)|g__Flavonifractor.s__Flavonifractor_plautii
461	PWY-5659: GDP-mannose biosynthesis|g__Anaerostipes.s__A
462	PWY-5659: GDP-mannose biosynthesis|g__Bacteroides.s__Ba
464	PWY-5659: GDP-mannose biosynthesis|g__Escherichia.s__Es
466	PWY-5659: GDP-mannose biosynthesis|g__Faecalibacterium.
477	PWY-5667: CDP-diacylglycerol biosynthesis I|g__Bacteroides.s__Bacteroides_uniformis
515	PWY-5686: UMP biosynthesis|g__Eubacterium.s__Eubacteriu
517	PWY-5686: UMP biosynthesis|g__Faecalibacterium.s__Faeca
531	PWY-5695: urate biosynthesis/inosine 5'-phosphate degradation|g__Bacteroides.s__Bacteroides_c
565	PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I|g__Anaerostipes.s__Anaerostipes_hadrus
569	PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I|g__Blautia.s__Ruminococcus_torques
571	PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I|g__Clostridium.s__Clostridium_bolteae
579	PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I|g__Faecalibacterium.s__Faecalibacterium_prausnitzii
580	PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I|g__Flavonifractor.s__Flavonifractor_plautii
597	PWY-6122: 5-aminoimidazole ribonucleotide biosynthesis II|g__Blautia.s__Ruminococcus_torques
610	PWY-6122: 5-aminoimidazole ribonucleotide biosynthesis II|g__Faecalibacterium.s__Faecalibacterium_prausnitzii

(To be continued)

Node	Pathway
611	PWY-6122: 5-aminoimidazole ribonucleotide biosynthesis II|g__Flavonifractor.s__Flavonifractor_plautii
622	PWY-6123: inosine-5'-phosphate biosynthesis I|g__Bacteroides.s__Bacteroides_fragilis
623	PWY-6123: inosine-5'-phosphate biosynthesis I|g__Bacteroides.s__Bacteroides_ovatus
632	PWY-6124: inosine-5'-phosphate biosynthesis II|g__Bacteroides.s__Bacteroides_ovatus
651	PWY-6147: 6-hydroxymethyl-dihydropterin diphosphate biosynthesis I|g__Bacteroides.s__Bacteroides
658	PWY-6151: S-adenosyl-L-methionine cycle I|g__Bacteroides.s__Bacteroides_uniformis
671	PWY-6151: S-adenosyl-L-methionine cycle I|g__Roseburia.s__Roseburia_inulinivorans
680	PWY-6163: chorismate biosynthesis from 3-dehydroquinate|g__Dorea.s__Dorea_longicatena
711	PWY-6277: superpathway of 5-aminoimidazole ribonucleotide biosynthesis|g__Blautia.s__Ruminococcus_torques
724	PWY-6277: superpathway of 5-aminoimidazole ribonucleotide biosynthesis|g__Faecalibacterium.s__Faecalibacterium_pra
725	PWY-6277: superpathway of 5-aminoimidazole ribonucleotide biosynthesis|g__Flavonifractor.s__Flavonifractor_plautii
740	PWY-6305: putrescine biosynthesis IV|g__Faecalibacterium.s__Faecalibacterium_prausnitzii
748	PWY-6317: galactose degradation I (Leloir pathway)|g__Faecalibacterium.s__Faecalibacterium_prausni
751	PWY-6317: galactose degradation I (Leloir pathway)|unclassified
755	PWY-6385: peptidoglycan biosynthesis III (mycobacteria)|g__Bacteroides.s__Bacteroides_ovatus
768	PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)|g__Bacteroides.s__Ba

(To be continued)

Node	Pathway
793	PWY-6387: UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing)|g__Bacteroides.s__Bacteroides_ovatus
794	PWY-6387: UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing)|g__Bacteroides.s__Bacteroides_uniformis
821	PWY-6527: stachyose degradation|unclassified
829	PWY-6608: guanosine nucleotides degradation III|unclassified
838	PWY-6609: adenine and adenosine salvage III|g__Escherichia.s__Escherichia_coli
842	PWY-6609: adenine and adenosine salvage III|unclassified
847	PWY-6700: queuosine biosynthesis|g__Anaerostipes.s__Ana
855	PWY-6700: queuosine biosynthesis|g__Bacteroides.s__Bact
870	PWY-6703: preQ0 biosynthesis|g__Bacteroides.s__Bacteroi
873	PWY-6703: preQ0 biosynthesis|g__Bacteroides.s__Bacteroi
877	PWY-6703: preQ0 biosynthesis|g__Escherichia.s__Escheric
896	PWY-6737: starch degradation V|g__Escherichia.s__Escherichia_coli
897	PWY-6737: starch degradation V|g__Eubacterium.s__Eubacterium_hallii
901	PWY-6737: starch degradation V|g__Faecalibacterium.s__Faecalibacterium_prausnitzii
918	PWY-6936: seleno-amino acid biosynthesis|unclassified
925	PWY-7111: pyruvate fermentation to isobutanol (engineered)|g__Bacteroides.s__Bacteroides_xylanisolvens
928	PWY-7111: pyruvate fermentation to isobutanol (engineered)|g__Blautia.s__Ruminococcus_obeum
939	PWY-7111: pyruvate fermentation to isobutanol (engineered)|g__Eubacterium.s__Eubacterium_eligens
944	PWY-7111: pyruvate fermentation to isobutanol (engineered)|g__Flavonifractor.s__Flavonifractor_plauti
953	PWY-7184: pyrimidine deoxyribonucleotides de novo biosynthesis I|unclassified

(To be continued)

Node	Pathway
963	PWY-7199: pyrimidine deoxyribonucleosides salvage|g__Escherichia.s__Escherichia_coli
969	PWY-7208: superpathway of pyrimidine nucleobases salvage|g__Eubacterium.s__Eubacterium_siraeu
988	PWY-7219: adenosine ribonucleotides de novo biosynthesis|g__Bacteroides.s__Bacteroides_uniformis
1008	PWY-7219: adenosine ribonucleotides de novo biosynthesis|g__Eubacterium.s__Eubacterium_rectale
1012	PWY-7219: adenosine ribonucleotides de novo biosynthesis|g__Flavonifractor.s__Flavonifractor_plautii
1027	PWY-7220: adenosine deoxyribonucleotides de novo biosynthesis II|g__Bacteroides.s__Bacteroides_vulga
1042	PWY-7221: guanosine ribonucleotides de novo biosynthesis|g__Bacteroides.s__Bacteroides_vulgatus
1047	PWY-7221: guanosine ribonucleotides de novo biosynthesis|g__Blautia.s__Ruminococcus_torques
1068	PWY-7222: guanosine deoxyribonucleotides de novo biosynthesis II|g__Bacteroides.s__Bacteroides_vulga
1074	PWY-7228: superpathway of guanosine nucleotides de novo biosynthesis I|g__Bacteroides.s__Bacteroides_vulgatus
1088	PWY-7237: myo-, chiro- and scillo-inositol degradation|unclassified
1093	PWY-724: superpathway of L-lysine, L-threonine and L-methionine biosynthesis II|unclassified
1096	PWY-7282: 4-amino-2-methyl-5-phosphomethylpyrimidine biosynthesis (yeast)|g__Bacteroides.s__Bacteroides_fragi
1097	PWY-7282: 4-amino-2-methyl-5-phosphomethylpyrimidine biosynthesis (yeast)|g__Bacteroides.s__Bacteroides_ovatu
1100	PWY-7323: superpathway of GDP-mannose-derived O-antigen building blocks biosynthesis|g__Bacteroides.s__Bacteroides_ovatus
1120	PWY-7539: 6-hydroxymethyl-dihydropterin diphosphate biosynthesis III (Chlamydia)|g__Bacteroides.s__Bacteroides_ovatus

(To be continued)

Node	Pathway
1123	PWY-7663: gondoate biosynthesis (anaerobic)|g__Bacteroides.s__Bacteroides_caccae
1147	PWY0-1296: purine ribonucleosides degradation|g__Eubacterium.s__Eubacterium_eligens
1155	PWY0-1297: superpathway of purine deoxyribonucleosides degradation|g__Blautia.s__Ruminococcus_torques
1194	PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing)|g__Escherichia.s__Esch
1195	PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing)|g__Faecalibacterium.s_
1203	PWY0-845: superpathway of pyridoxal 5'-phosphate biosynthesis and salvage|g__Bacteroides.s__Bacteroides_ovatus
1205	PWY0-845: superpathway of pyridoxal 5'-phosphate biosynthesis and salvage|g__Bacteroides.s__Bacteroides_uniformis
1212	PWY66-400: glycolysis VI (metazoan)|g__Bacteroides.s__B
1222	PWY66-422: D-galactose degradation V (Leloir pathway)|g__Faecalibacterium.s__Faecalibacterium_prausni
1225	PWY66-422: D-galactose degradation V (Leloir pathway)|unclassified
1238	PYRIDOXSYN-PWY: pyridoxal 5'-phosphate biosynthesis I|g__Bacteroides.s__Bacteroides_ovatus
1246	RHAMCAT-PWY: L-rhamnose degradation I|g__Bacteroides.s__Bacteroides_xylanisolvens
1249	RHAMCAT-PWY: L-rhamnose degradation I|g__Faecalibacterium.s__Faecalibacterium_prausnitzii
1253	SALVADEHYPOX-PWY: adenosine nucleotides degradation II|g__Blautia.s__Ruminococcus_torques
1261	SER-GLYSYN-PWY: superpathway of L-serine and glycine biosynthesis I|g__Faecalibacterium.s__Faecalibacterium_prausnitzii
1284	THRESYN-PWY: superpathway of L-threonine biosynthesis|unclassified
1298	VALSYN-PWY: L-valine biosynthesis|g__Anaerostipes.s__Ana

(To be continued)

Node	Pathway
1303	VALSYN-PWY: L-valine biosynthesis|g__Bacteroides.s__Bact
1306	VALSYN-PWY: L-valine biosynthesis|g__Blautia.s__Ruminoco
1315	VALSYN-PWY: L-valine biosynthesis|g__Eubacterium.s__Euba
1320	VALSYN-PWY: L-valine biosynthesis|g__Flavonifractor.s__F

Bibliography

Aldecoa, Rodrigo and Ignacio Marín (2013). "Surprise maximization reveals the community structure of complex networks". In: *Scientific reports* 3.1, pp. 1–9. ISSN: 2045-2322.

Ananthakrishnan, Ashwin N, Ramnik J Xavier, and Daniel K Podolsky (2017). *Inflammatory Bowel Diseases: A Clinician's Guide*. John Wiley & Sons. ISBN: 1119077605.

Anderson, Theodore Wilbur (1962). *An introduction to multivariate statistical analysis*. Tech. rep.

Arumugam, Manimozhiyan et al. (2011). "Enterotypes of the human gut microbiome". In: *nature* 473.7346, pp. 174–180. ISSN: 1476-4687.

Atsumi, Shota et al. (2010). "Engineering the isobutanol biosynthetic pathway in Escherichia coli by comparison of three aldehyde reductase/alcohol dehydrogenase genes". In: *Applied microbiology and biotechnology* 85.3, pp. 651–657. ISSN: 1432-0614.

Balassa, Bela (1965). "Trade liberalisation and "revealed" comparative advantage 1". In: *The manchester school* 33.2, pp. 99–123. ISSN: 1463-6786.

Barabási, Albert-László (2016). *Network science*. Cambridge university press. ISBN: 1107076269.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo (2011). "Network medicine: a network-based approach to human disease". In: *Nature reviews genetics* 12.1, p. 56. ISSN: 1471-0064.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009). "Gephi: an open source software for exploring and manipulating networks". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 3. 1.

Battiston, Federico, Vincenzo Nicosia, and Vito Latora (2014). "Structural measures for multiplex networks". In: *Physical Review E* 89.3, p. 32804.

Bauer, Eugen and Ines Thiele (2018). "From network analysis to functional metabolic modeling of the human gut microbiota". In: *MSystems* 3.3, pp. 00209–17. ISSN: 2379-5077.

Beek, Samantha L van der et al. (2019). "Streptococcal dTDP-L-rhamnose biosynthesis enzymes: functional characterization and lead compound identification". In: *Molecular microbiology* 111.4, pp. 951–964. ISSN: 0950-382X.

Belizário, José E and Joel Faintuch (2018). "Microbiome and gut dysbiosis". In: *Metabolic interaction in infection*. Springer, pp. 459–476.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300. ISSN: 0035-9246.

Bentley, Ronald and E Haslam (1990). "The shikimate pathway—a metabolic tree with many branche". In: *Critical reviews in biochemistry and molecular biology* 25.5, pp. 307–384. ISSN: 1040-9238.

Berg, Gabriele et al. (2020). "Microbiome definition re-visited: old concepts and new challenges". In: *Microbiome* 8.1, pp. 1–22. ISSN: 2049-2618.

Bianconi, Ginestra (2018). *Multilayer networks: structure and function*. Oxford university press. ISBN: 0191068500.

Blondel, Vincent D et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10, P10008. ISSN: 1742-5468.

Boccaletti, Stefano et al. (2014). "The structure and dynamics of multilayer networks". In: *Physics reports* 544.1, pp. 1–122. ISSN: 0370-1573.

Borody, Thomas J and Alexander Khoruts (2012). "Fecal microbiota transplantation and emerging applications". In: *Nature reviews Gastroenterology & hepatology* 9.2, pp. 88–96. ISSN: 1759-5053.

Brand, Stewart (2010). *Whole Earth Discipline: Why Dense Cities, Nuclear Power, Transgenic Crops, RestoredWildlands, and Geoeng ineering Are Necessary*. Penguin. ISBN: 1101483415.

Brandes, Ulrik (2001). "A faster algorithm for betweenness centrality". In: *Journal of mathematical sociology* 25.2, pp. 163–177. ISSN: 0022-250X.

Bruno, Matteo (2020). *BiCM: Python package for the computation of the Bipartite Configuration Model.* URL: `https://github.com/mat701/BiCM`.

Bucci, Vanni et al. (2016). "MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses". In: *Genome biology* 17.1, pp. 1–17. ISSN: 1474-760X.

Buchanan, M et al. (2010). *Networks in Cell Biology*. Cambridge University Press. ISBN: 9780521882736. URL: https://books.google.it/books?id=ojMhR2pq7qIC.

Caldarelli, Guido (2007). *Scale-free networks: complex webs in nature and technology*. Oxford University Press. ISBN: 0199211515.

Cao, Yuan, Jun Shen, and Zhi Hua Ran (2014). "Association between Faecalibacterium prausnitzii reduction and inflammatory bowel disease: a meta-analysis and systematic review of the literature". In: *Gastroenterology research and practice* 2014. ISSN: 1687-6121.

Chen, Jiezhong and Luis Vitetta (2020). "Butyrate in inflammatory bowel disease therapy". In: *Gastroenterology* 158.5, p. 1511. ISSN: 0016-5085.

Chen, Lianmin et al. (2020). "Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity". In: *Nature communications* 11.1, pp. 1–12. ISSN: 2041-1723.

Choobdar, Sarvenaz et al. (2019). "Assessment of network module identification across complex diseases". In: *Nature methods* 16.9, pp. 843–852. ISSN: 1548-7105.

Coyte, Katharine Z and Seth Rakoff-Nahoum (2019). "Understanding competition and cooperation within the mammalian gut microbiome". In: *Current Biology* 29.11, R538–R544. ISSN: 0960-9822.

Cozzo, Emanuele et al. (2016). "Multilayer networks: metrics and spectral properties". In: *Interconnected Networks*. Springer, pp. 17–35.

De Domenico, Manlio, Andrea Lancichinetti, et al. (2015). "Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems". In: *Physical Review X* 5.1, p. 11027.

De Domenico, Manlio, Mason A Porter, and Alex Arenas (2015). "MuxViz: a tool for multilayer analysis and visualization of networks". In: *Journal of Complex Networks* 3.2, pp. 159–176. ISSN: 2051-1310.

De Domenico, Manlio, Albert Solé-Ribalta, Emanuele Cozzo, et al. (2013). "Mathematical formulation of multilayer networks". In: *Physical Review X* 3.4, p. 41022.

De Domenico, Manlio, Albert Solé-Ribalta, Elisa Omodei, et al. (2013). "Centrality in interconnected multilayer networks". In: *arXiv preprint arXiv:1311.2906*.

— (2015). "Ranking in interconnected multilayer networks reveals versatile nodes". In: *Nature communications* 6.1, pp. 1–6. ISSN: 2041-1723.

De Souza, Heitor S P, Claudio Fiocchi, and Dimitrios Iliopoulos (2017). "The IBD interactome: an integrated view of aetiology, pathogenesis

and therapy". In: *Nature Reviews Gastroenterology & Hepatology* 14.12, pp. 739–749. ISSN: 1759-5053.

Didier, Gilles, Christine Brun, and Anaïs Baudot (2015). "Identifying communities from multiplex biological networks". In: *PeerJ* 3, e1525. ISSN: 2167-8359.

Farré-Maduell, Eulàlia and Climent Casals-Pascual (2019). "The origins of gut microbiome research in Europe: From Escherich to Nissle". In: *Human Microbiome Journal* 14, p. 100065. ISSN: 2452-2317.

Faust, Karoline et al. (2018). "Signatures of ecological processes in microbial community time series". In: *Microbiome* 6.1, pp. 1–13. ISSN: 2049-2618.

Fiocchi, Claudio and Dimitrios Iliopoulos (2021). "IBD Systems Biology Is Here to Stay". In: *Inflammatory Bowel Diseases* 27.6, pp. 760–770. ISSN: 1078-0998.

Fortunato, Santo and Marc Barthelemy (2007). "Resolution limit in community detection". In: *Proceedings of the national academy of sciences* 104.1, pp. 36–41. ISSN: 0027-8424.

Franzosa, Eric A et al. (2018). "Species-level functional profiling of metagenomes and metatranscriptomes". In: *Nature methods* 15.11, pp. 962–968. ISSN: 1548-7105.

Freeman, Linton C (1977). "A set of measures of centrality based on betweenness". In: *Sociometry*, pp. 35–41. ISSN: 0038-0431.

Friedman, Jonathan and Eric J Alm (2012). "Inferring correlation networks from genomic survey data". In: ISSN: 1553-734X.

Ghiassian, Susan Dina, Jörg Menche, and Albert-László Barabási (2015). "A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome". In: *PLoS Comput Biol* 11.4, e1004120. ISSN: 1553-7358.

Gomez-Arango, Luisa F et al. (2016). "Increased systolic and diastolic blood pressure is associated with altered gut microbiota composition and butyrate production in early pregnancy". In: *Hypertension* 68.4, pp. 974–981. ISSN: 0194-911X.

Henke, Matthew T et al. (2019). "Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide". In: *Proceedings of the National Academy of Sciences* 116.26, pp. 12672–12677. ISSN: 0027-8424.

Hoffmann, Thomas W et al. (2016). "Microorganisms linked to inflammatory bowel disease-associated dysbiosis differentially impact host

physiology in gnotobiotic mice". In: *The ISME journal* 10.2, pp. 460–477. ISSN: 1751-7370.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics*, pp. 65–70. ISSN: 0303-6898.

IBDMDB - Home | IBDMDB (n.d.). URL: https://ibdmdb.org/.

Jiang, Puzi et al. (2021). "Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models". In: *Msystems* 6.3, pp. 00112–21. ISSN: 2379-5077.

Kaplan, Gilaad G (2015). "The global burden of IBD: from 2015 to 2025". In: *Nature reviews Gastroenterology & hepatology* 12.12, pp. 720–727. ISSN: 1759-5053.

Keighley, M R et al. (1978). "Influence of inflammatory bowel disease on intestinal microflora." In: *Gut* 19.12, pp. 1099–1104. ISSN: 0017-5749.

Kirsner, Joseph B (1988). "Historical aspects of inflammatory bowel disease." In: *Journal of clinical gastroenterology* 10.3, pp. 286–297. ISSN: 0192-0790.

Kivelä, Mikko et al. (Sept. 2014). "Multilayer networks". In: *Journal of Complex Networks* 2.3, pp. 203–271. ISSN: 2051-1310. DOI: 10.1093/comnet/cnu016. URL: https://doi.org/10.1093/comnet/cnu016.

KneadData – The Huttenhower Lab (n.d.). URL: http://huttenhower.sph.harvard.edu/kneaddata.

Ladoukakis, Efthymios, Fragiskos N Kolisis, and Aristotelis A Chatzioannou (2014). "Integrative workflows for metagenomic analysis". In: *Frontiers in cell and developmental biology* 2, p. 70. ISSN: 2296-634X.

Lancichinetti, Andrea and Santo Fortunato (2009). "Community detection algorithms: a comparative analysis". In: *Physical review E* 80.5, p. 56117.

Lees, Charlie W and Jack Satsangi (2009). "Genetics of inflammatory bowel disease: implications for disease pathogenesis and natural history". In: *Expert review of gastroenterology & hepatology* 3.5, pp. 513–534. ISSN: 1747-4124.

Leonardi, Roberta et al. (2005). "Coenzyme A: back in action". In: *Progress in lipid research* 44.2-3, pp. 125–153. ISSN: 0163-7827.

Lepage, Patricia et al. (2011). "Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis". In: *Gastroenterology* 141.1, pp. 227–236. ISSN: 0016-5085.

Ley, Ruth E (2016). "Prevotella in the gut: choose carefully". In: *Nature reviews Gastroenterology & hepatology* 13.2, pp. 69–70. ISSN: 1759-5053.

Li, Wendy et al. (2021). "Ecological and network analyses identify four microbial species with potential significance for the diagnosis/treatment of ulcerative colitis (UC)". In: *BMC microbiology* 21.1, pp. 1–12. ISSN: 1471-2180.

Lindgreen, Stinus, Karen L Adair, and Paul P Gardner (2016). "An evaluation of the accuracy and speed of metagenome analysis tools". In: *Scientific reports* 6.1, pp. 1–14. ISSN: 2045-2322.

Lloyd-Price, Jason et al. (May 2019). "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases". In: *Nature* 569.7758, pp. 655–662. ISSN: 14764687. DOI: 10.1038/s41586-019-1237-9.

Ma, Yonghui et al. (2018). "Help, hope and hype: ethical considerations of human microbiome research and applications". In: *Protein & cell* 9.5, pp. 404–415. ISSN: 1674-8018.

Ma, Yongshun et al. (2021). "Metagenome Analysis of Intestinal Bacteria in Healthy People, Patients With Inflammatory Bowel Disease and Colorectal Cancer". In: *Frontiers in Cellular and Infection Microbiology* 11.

Mack, A et al. (2020). "Changes in gut microbial metagenomic pathways associated with clinical outcomes after the elimination of malabsorbed sugars in an IBS cohort". In: *Gut microbes* 11.3, pp. 620–631. ISSN: 1949-0976.

MacMahon, Mel and Diego Garlaschelli (2013). "Community detection for correlation matrices". In: *arXiv preprint arXiv:1311.1924*.

Manichanh, Chaysavanh et al. (2012a). "The gut microbiota in IBD". In: *Nature reviews Gastroenterology & hepatology* 9.10, pp. 599–608. ISSN: 1759-5053.

— (2012b). *The gut microbiota in IBD*. DOI: 10.1038/nrgastro.2012.152. URL: https://pubmed.ncbi.nlm.nih.gov/22907164/.

McIver, Lauren J. et al. (Apr. 2018). "BioBakery: A meta'omic analysis environment". In: *Bioinformatics* 34.7, pp. 1235–1237. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx754.

Meij, Tim G J de et al. (2018). "Variability of core microbiota in newly diagnosed treatment-naïve paediatric inflammatory bowel disease patients". In: *PloS one* 13.8, e0197649. ISSN: 1932-6203.

Mirande, Caroline et al. (2010). "Dietary fibre degradation and fermentation by two xylanolytic bacteria Bacteroides xylanisolvens XB1AT and Roseburia intestinalis XB6B4 from the human intestine". In: *Journal of applied microbiology* 109.2, pp. 451–460. ISSN: 1364-5072.

Mucha, Peter J et al. (2010). "Community structure in time-dependent, multiscale, and multiplex networks". In: *science* 328.5980, pp. 876–878. ISSN: 0036-8075.

Mukherjee, Arghya et al. (2020). "Gut microbes from the phylogenetically diverse genus Eubacterium and their various contributions to gut health". In: *Gut Microbes* 12.1, p. 1802866. ISSN: 1949-0976.

Nagy, Peter L et al. (1995). "Formyltetrahydrofolate hydrolase, a regulatory enzyme that functions to balance pools of tetrahydrofolate and one-carbon tetrahydrofolate adducts in Escherichia coli". In: *Journal of Bacteriology* 177.5, pp. 1292–1298. ISSN: 0021-9193.

Natividad, Jane M M and Elena F Verdu (2013). "Modulation of intestinal barrier by intestinal microbiota: pathological and therapeutic implications". In: *Pharmacological research* 69.1, pp. 42–51. ISSN: 1043-6618.

Neal, Zachary (2014). "The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors". In: *Social Networks* 39, pp. 84–97. ISSN: 0378-8733.

Nemoto, Hideyuki et al. (2012). "Reduced diversity and imbalance of fecal microbiota in patients with ulcerative colitis". In: *Digestive diseases and sciences* 57.11, pp. 2955–2964. ISSN: 1573-2568.

Newman, Mark E J (2004). "Analysis of weighted networks". In: *Physical review E* 70.5, p. 56131.

— (2006). "Modularity and community structure in networks". In: *Proceedings of the national academy of sciences* 103.23, pp. 8577–8582. ISSN: 0027-8424.

Nicolini, Carlo et al. (2020). "Scale-resolved analysis of brain functional connectivity networks with spectral entropy". In: *NeuroImage* 211, p. 116603. ISSN: 1053-8119.

Nishida, Atsushi et al. (2018). "Gut microbiota in the pathogenesis of inflammatory bowel disease". In: *Clinical journal of gastroenterology* 11.1, pp. 1–10. ISSN: 1865-7265.

Paci, Paola et al. (2017). "SWIM: a computational tool to unveiling crucial nodes in complex biological networks". In: *Scientific reports* 7.1, pp. 1–16. ISSN: 2045-2322.

Plumbridge, Jacqueline and Eric Vimr (1999). "Convergent pathways for utilization of the amino sugars N-acetylglucosamine, N-acetylmannosamine, and N-acetylneuraminic acid by Escherichia coli". In: *Journal of bacteriology* 181.1, pp. 47–54. ISSN: 1098-5530.

Precup, Gabriela and Dan-Cristian Vodnar (2019). "Gut Prevotella as a possible biomarker of diet and its eubiotic versus dysbiotic roles: a

comprehensive literature review". In: *British Journal of Nutrition* 122.2, pp. 131–140. ISSN: 0007-1145.

Rhodes, Jonathan M (2007). "The role of Escherichia coli in inflammatory bowel disease". In: *Gut* 56.5, pp. 610–612. ISSN: 0017-5749.

Rinninella, Emanuele et al. (2019). "What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases". In: *Microorganisms* 7.1, p. 14.

Rosselló-Mora, Ramon and Rudolf Amann (2001). "The species concept for prokaryotes". In: *FEMS microbiology reviews* 25.1, pp. 39–67. ISSN: 0168-6445.

Saitoh, Shin et al. (2002). "Bacteroides ovatus as the predominant commensal intestinal microbe causing a systemic antibody response in inflammatory bowel disease". In: *Clinical and diagnostic laboratory immunology* 9.1, pp. 54–59. ISSN: 1071-412X.

Salyers, Abigail A, Peter Valentine, and Vivian Hwa (1993). "Genetics of polysaccharide utilization pathways of colonic Bacteroides species". In: *Genetics and Molecular Biology of Anaerobic Bacteria*. Springer, pp. 505–516.

Saracco, Fabio, Riccardo Di Clemente, et al. (June 2015). "Randomizing bipartite networks: The case of the World Trade Web". In: *Scientific Reports* 5.1, pp. 1–18. ISSN: 20452322. DOI: 10.1038/srep10595.

Saracco, Fabio, Mika J Straka, et al. (2017). "Inferring monopartite projections of bipartite networks: an entropy-based approach". In: *New Journal of Physics* 19.5, p. 53022. ISSN: 1367-2630.

Schirmer, Melanie et al. (2019). "Microbial genes and pathways in inflammatory bowel disease". In: *Nature Reviews Microbiology* 17.8, pp. 497–511. ISSN: 1740-1534.

Segal, Jonathan P et al. (2019). "The application of omics techniques to understand the role of the gut microbiota in inflammatory bowel disease". In: *Therapeutic advances in gastroenterology* 12, p. 1756284818822250. ISSN: 1756-2848.

Sender, Ron, Shai Fuchs, and Ron Milo (2016). "Revised estimates for the number of human and bacteria cells in the body". In: *PLoS biology* 14.8, e1002533. ISSN: 1544-9173.

Sharma, Vandana, Mie Ichikawa, and Hudson H Freeze (2014). "Mannose metabolism: more than meets the eye". In: *Biochemical and biophysical research communications* 453.2, pp. 220–228. ISSN: 0006-291X.

Singh, Sunny et al. (2011). "Common symptoms and stressors among individuals with inflammatory bowel diseases". In: *Clinical Gastroenterology and Hepatology* 9.9, pp. 769–775. ISSN: 1542-3565.

Sommer, Felix and Fredrik Bäckhed (2013). "The gut microbiota—masters of host development and physiology". In: *Nature reviews microbiology* 11.4, pp. 227–238. ISSN: 1740-1534.

Squartini, Tiziano and Diego Garlaschelli (2011). "Analytical maximum-likelihood method to detect patterns in real networks". In: *New Journal of Physics* 13.8, p. 83001. ISSN: 1367-2630.

Stein, Richard R et al. (2013). "Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota". In: *PLoS Computational Biology* 9.12. Ed. by Christian von Mering, e1003388. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003388. URL: https://dx.plos.org/10.1371/journal.pcbi.1003388.

Takahashi, Kenichiro et al. (2016). "Reduced abundance of butyrate-producing bacteria species in the fecal microbial community in Crohn's disease". In: *Digestion* 93.1, pp. 59–65. ISSN: 0012-2823.

Tanaka, Masaru and Jiro Nakayama (2017). "Development of the gut microbiota in infancy and its impact on health in later life". In: *Allergology International* 66.4, pp. 515–522. ISSN: 1323-8930.

Tannock, Gerald W (2010). "The bowel microbiota and inflammatory bowel diseases". In: *International journal of inflammation* 2010. ISSN: 2090-8040.

Thursby, Elizabeth and Nathalie Juge (2017). "Introduction to the human gut microbiota". In: *Biochemical Journal* 474.11, pp. 1823–1836. ISSN: 0264-6021.

Tieri, Paolo et al. (2019). "Network inference and reconstruction in bioinformatics". In:

Toro, Vladimir, Eduardo Mojica-Nava, and Naly Rakoto-Ravalontsalama (2019). "Multiplex Networks Centrality Measurements for Microgrids". In: *IFAC-PapersOnLine* 52.4, pp. 24–29. ISSN: 2405-8963.

Vernocchi, Pamela et al. (2020). "Network analysis of gut microbiome and metabolome to discover microbiota-linked biomarkers in patients affected by non-small cell lung cancer". In: *International journal of molecular sciences* 21.22, p. 8730.

Vich Vila, A et al. (2018). *Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. Sci Transl Med 10: eaap8914.*

Wagner, Nicole et al. (2005). "Pyruvate fermentation by Oenococcus oeni and Leuconostoc mesenteroides and role of pyruvate dehydrogenase in anaerobic fermentation". In: *Applied and environmental microbiology* 71.9, pp. 4966–4971. ISSN: 0099-2240.

Weiss, Sophie et al. (2016). "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision". In: *The ISME journal* 10.7, pp. 1669–1681. ISSN: 1751-7370.

West, Allyson A, Marie A Caudill, and Lynn B Bailey (2020). "Folate". In: *Present Knowledge in Nutrition*. Elsevier, pp. 239–255.

Wexler, Hannah M (2007). "Bacteroides: the good, the bad, and the nitty-gritty". In: *Clinical microbiology reviews* 20.4, pp. 593–621. ISSN: 0893-8512.

Wu, Guoyao (2009). "Amino acids: metabolism, functions, and nutrition". In: *Amino acids* 37.1, pp. 1–17. ISSN: 0939-4451.

Xavier, Ramnik J and D K Podolsky (2007). "Unravelling the pathogenesis of inflammatory bowel disease". In: *Nature* 448.7152, pp. 427–434. ISSN: 1476-4687.

Xi, Menglu et al. (2020). "Effects of stachyose on intestinal microbiota and immunity in mice infected with enterotoxigenic Escherichia coli". In: *Journal of Functional Foods* 64, p. 103689. ISSN: 1756-4646.

Yarza, Pablo et al. (2014). "Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences". In: *Nature Reviews Microbiology* 12.9, pp. 635–645. ISSN: 1740-1534.

Yin, Jie et al. (2017). "Lysine restriction affects feed intake and amino acid metabolism via gut microbiome in piglets". In: *Cellular Physiology and Biochemistry* 44.5, pp. 1749–1761. ISSN: 1015-8987.

Zhang, Yang, Mariya Morar, and Steven E Ealick (2008). "Structural biology of the purine biosynthetic pathway". In: *Cellular and molecular life sciences* 65.23, pp. 3699–3724. ISSN: 1420-9071.

Zhang, Yi-Zhen and Yong-Yu Li (2014). "Inflammatory bowel disease: pathogenesis". In: *World journal of gastroenterology: WJG* 20.1, p. 91.

Zhou, Yingting and Fachao Zhi (2016). "Lower level of bacteroides in the gut microbiota is associated with inflammatory bowel disease: a meta-analysis". In: *BioMed research international* 2016. ISSN: 2314-6133.

Zhu, Changxin et al. (2018). "Roseburia intestinalis inhibits interleukin-17 excretion and promotes regulatory T cells differentiation in colitis". In: *Molecular medicine reports* 17.6, pp. 7567–7574. ISSN: 1791-2997.

Zhu, Yunzhen et al. (2020). "High-throughput sequencing analysis of differences in intestinal microflora between ulcerative colitis patients with different glucocorticoid response types". In: *Genes & Genomics* 42.10, pp. 1197–1206. ISSN: 2092-9293.

Zizzo, Maria Grazia et al. (2019). "Preventive effects of guanosine on intestinal inflammation in 2, 4-dinitrobenzene sulfonic acid (DNBS)-

induced colitis in rats". In: *Inflammopharmacology* 27.2, pp. 349–359. ISSN: 1568-5608.

Unless otherwise expressly stated, all original material of whatever nature created by Mirko Hu and included in this thesis, is licensed under a Creative Commons Attribution Noncommercial Share Alike 3.0 Italy License.

Check on Creative Commons site:

https://creativecommons.org/licenses/by-nc-sa/3.0/it/legalcode/

https://creativecommons.org/licenses/by-nc-sa/3.0/it/deed.en

Ask the author about other uses.