Figure selected to demonstrate that species are not separated from each other. Even in a simple model with only five selected bacterial species, the microorganisms are aggregated suggesting possible interaction among those species. Intestinal section from a gnotobiotic mouse model inoculated with selected bacterial species found in the human gut. Blue = *Bacteroides WH2*, green = *Bacteroides thetaiotamicron*, pink = *Bacteroides vulgatus*, yellow = *Collinsella aerofaciens*, red = *Ruminococcus torques*. Credit: Yuko Hasegawa/MBL Woods Hole.
“True friendship is like sound health, the value of it is seldom known until it is lost” — Charles Caleb Colton (1780–1832)

“Everything is everywhere: but the environment selects” — Martinus Willem Beijerinck (1851–1931)
Understanding the link between the human gut microbiome and human health is one of the biggest scientific challenges in our decade. Because 90% of our cells are bacteria, and the microbial genome contains 200 times more genes than the human genome, the study of the human microbiome has the potential to impact many areas of our health. This PhD thesis is the first study to generate a large amount of experimental data on the DNA and RNA of the human gut microbiome. This was made possible by our development of a human gut microbiome array capable of profiling any human gut microbiome. Analysis of our results changes the way we link the gut microbiome with diseases. Our results indicate that inflammatory diseases will affect the ecological system of the human gut microbiome, reducing its diversity. Classification analysis of healthy and unhealthy individuals demonstrates that unhealthy individuals have lower diversity microbiomes with incomplete functional capacity. Diversity is an important measurement linking microbiome variance to diseases. Our results suggest that diseases are linked to the microbiome not by the presence of “bad” bacteria, but mostly by the loss of the “good” bacteria. Finally, we show that bacterial adaptations explain the shift observed in the human gut microbiome.
I would like to thank everybody who worked at CBS on my PhD studies, the professors, staff, and students who make this environment an excellent place to work. The diversity and the expertise of those professionals generates a critical mass necessary to acquire the skills to successfully complete my thesis. I enjoyed many interesting scientific discussions with my colleagues, which, in my opinion, is an important part of scientific research. Therefore, the acknowledgement is a point where I appreciate the time they dedicated to my project.

I would like to start with the administration staff, which has helped me with all of the bureaucratic and administrative tasks. Without them, I would probably have lost an important part of my PhD trying to figure out how to fill out the forms. I would like to dedicate a special acknowledgement to Lone Boesen, who resolved many travel issues, and Annette Vibeke Uldall for everything related to Søren Brunak. I would also like to extend say a very special ‘thank you’ to Dorthe Kjærsgaard, who kindly helped me with everything related to the PhD programme.

The system administration also deserves special mention. As Kristoffer Rapacki once said, “Researchers only remember the system administration when the system has a problem”. I agree that when the system is working perfectly, we forget that a huge infrastructure must be set up in order for us to do our work. This was especially true for the MetaHIT project, which used a huge amount of disk space and computer power. I often contacted the system administration staff to find solutions for MetaHIT needs. Therefore, thanks to Peter Wad Sackett and Hans Henrik Stærfeldt for all the help. To John Damm Sørensen, I owe special thanks, as I annoyed him the most. Moreover, special thanks to Kristoffer Rapacki for always being able to manage our requirements and provide solutions to the projects, especially for the Saqqaq project.

For the scientific work, I would like to thank the whole Metagenomics group. I was their first PhD student and saw them grow. My knowledge of the field would have been significantly decreased without our group discussions. Special thanks to
Agnieszka S. Juncker and Simon Rasmussen for technical discussions and scientific comments on the MetaHIT project. Thanks to all the professors from other fields, especially the ones that I bothered the most with my problems, Chris Workman, Laurent Gautier, and Henrik Bjørn Nielsen, who provided and inspired most of the functions in the package. A very special thanks to Thomas Sicheritz Pontén and Henrik Bjørn Nielsen, as this thesis would not be possible without them. We spent many hours talking about biological ideas and the results for MetaHIT and other projects. It has been a pleasure to discuss those ideas with them. I am very glad to have Thomas Sicheritz Pontén as my chief supervisor, he always supported me during difficult times, managed problems, and gave me the freedom and flexibility to do my work. Thanks to David Wayne Ussery and Søren Brunak for always bringing new ideas and vision to our results. Your comments and suggestions played an important role in defining my work.

Thanks to Olivier Taboureau, Thomas Nordahl Petersen, Ramneek Gupta, Kasper Nielsen, and Bent Petersen, who were always present during our extracurricular activities, in particularly the football table section, for making the work environment a very friendly one. Moreover, I would like to thank the other PhD students, Josef Korbinian Vogt, Salvatore Cosentino, Ida Bonde, and Dhany Saputra, who always made our trips a lot of fun.

In MetaHIT project, I would like to thank all the participants. The internal meetings were very inspiring and important for my education. Being able to communicate the best researchers in Europe at an extraordinary place every six months was very motivating. Therefore, a special thanks for all the PIs who make MetaHIT possible: Joël Doré, Peer Bork, Søren Brunak, Dusko Ehrlich, Francisco Guarner, Julian Parkhil, Oluf Pedersen, Jeroen Raes, and Pierre Renault. Special thanks to the PhD students working on MetaHIT for the scientific comments and the excellent collaborative work. I would especially like to thank Falk Hildebrand who worked with me on the third article of this thesis.

I would like to thank the open source community, especially those related to Debian and Ubuntu, which who I have been using for many years. Thanks also to libre-office, open-office, and Latex. My thesis would not be possible without their work, which, in general, is not acknowledged. Finally, thanks to my family and friends who were always supportive. And special thanks to my daughter Isabella, who with every smile makes all my troubles go away, which was especially important during the PhD period.
Articles in the period of my PhD

Included in the Thesis


**Not included in the Thesis**


List of Figures

1.1 Sequencing technologies timeline. .......................... 3
1.2 First metagenomic diagram. ................................. 5
1.3 Activia an example of probiotic products. ................. 7

2.1 Bacterial behaviours in the presence of oxygen. ............. 14
2.2 Bacteroidetes relative abundance explains most of the variance in the data. .......................................................... 17
2.3 Interindividual variation and distribution of phyla composition. . . . 18
2.4 Increase in bacterial concentration in mice and human, from four studies 20
2.5 Factors shaping microbiome composition. ....................... 23
2.6 Diseases and diversity ............................................. 27
2.7 The importance of evenness. ....................................... 30
2.8 Example of diversity among three ecosystems. .................. 32
2.9 Steps to create species network. ............................... 35

3.1 Ecological network of the human gut microbiome. .......... 41
3.2 Main contributors genera of each enterotype ................. 52
3.3 Schematic representation of host–microbiota interactions in dysbiosis  62
Contents

Summary (English) i
Summary (Danish) iii
Acknowledgements v
Articles in the period of my PhD vii

1 Introduction 1

Introduction 1
1.1 Metagenomics and sequencing technologies . . . . . . . . . . . . . . . 1
1.2 MetaHIT: the project objectives . . . . . . . . . . . . . . . . . . . . . 7
1.3 Thesis objectives . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2 Background 11

Background 11
2.1 Microbiology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
2.1.1 Biological classification . . . . . . . . . . . . . . . . . . . . . . 11
2.1.2 Bacterial phenotype . . . . . . . . . . . . . . . . . . . . . . . 12
2.1.3 Bacterial fermentation: short-chain fatty acids . . . . . . . . . 13
2.2 Relevance of the gut microbiome . . . . . . . . . . . . . . . . . . . . 14
2.2.1 Gut microbiome composition . . . . . . . . . . . . . . . . . . . . 15
2.2.2 Human gut microbiome shifts related to disease . . . . . . . . 19
2.2.3 Evolution of the gut microbiome . . . . . . . . . . . . . . . . . 19
2.3 Factors shaping microbiome composition . . . . . . . . . . . . . . . . 21
2.3.1 Impact of genetics on gut microbiome composition . . . . . . . 22
2.3.2 Impact of diet on gut microbiome composition . . . . . . . . . 22
2.3.3 Impact of the immune system on gut microbiome composition 24
2.4 Applying ecology to explain the human gut microbiome . . . . . . . 26
Applying ecology to explain the human gut microbiome. .......................... 26
  2.4.1 Diversity ................................................................. 28
  2.4.2 Network ................................................................. 31
  2.4.3 Microbial ecological network ........................................... 33
  2.4.4 Properties of ecological networks .................................... 34

3 Results ................................................................................. 39

Results ..................................................................................... 39
  3.1 Article 1: “A human gut microbial gene catalogue established by
  metagenomic sequencing” ......................................................... 39
  3.2 Article 2: “Enterotypes of the human gut microbiome” .............. 51
  3.3 Article 3: “Ecological resilience of the human gut microbiome and
  human healthy” .................................................................... 60

4 Conclusion ............................................................................. 97

Bibliography ............................................................................ 105
Metagenomics is a new technique developed to study genomes directly in their natural environments. As with many techniques, metagenomics has evolved with other technologies, and it currently has a broad spectrum of applications. It is an important tool for academic research and industry, facilitating the identification of new species and enzymes. Moreover, metagenomics is elucidating the organization of microbial communities and the mechanisms through which species interact with one another. This introduction is a discussion of the development and use of metagenomics and its benefits for academic research and industry. A good example of its application in the food industry is the launch of Activia, a brand of yogurt containing a specific probiotic bacterium. Developed by Danone, a multinational food products corporation, Activia became one of the most successful product launches in recent history. The Metagenomics of the Human Intestinal Tract (MetaHIT) project and the objectives of this thesis are also described in this section.

1.1 Metagenomics and sequencing technologies

Major breakthroughs in science frequently follow the development of new investigative methods. The development of microscopes in the 17th century ushered in the discovery of bacteria, changing the treatment of diseases and the approach to human interactions with microorganisms. In the 20th century, the development of X-ray diffraction techniques paved the way for the discovery of the structure of DNA. Although these two techniques were not developed with the intention of increasing biological knowledge, they made significant contributions to the field nonetheless. Metagenomics is a new technique developed to overcome several limitations of ge-
nomic studies. It enables researchers to sequence genomes that were previously inaccessible, thereby providing important insights into community genomic studies.

However, the advancement of metagenomics has been made possible only by improvements in sequencing techniques. The structure of DNA was discovered in 1953 [1], but DNA sequencing technologies did not emerge until two decades later. Many sequencing techniques were developed in the early 1970s [2, 3]. However, the chain-termination method — developed by Frederick Sanger and co-workers in 1977 [4] — became the method of choice owing to its relative ease and reliability. Improvements in robotics, reagents, and computer software increased the efficiency of this method, and in 2001, it was applied to the human genome. Capillary electrophoresis instruments were developed that could detect 500–600 base pairs per reaction from individual fluorescently labelled samples in 10 hours. These machines were capable of performing 96 reactions per run, sequencing 115 kilobases in 24 hours of non-stop operation [5]. To put this output in perspective, approximately 4% of the Escherichia coli genome or 0.003% of the human genome could be sequenced per day. In 2005, a range of DNA sequencing techniques that were fundamentally different from capillary sequencing became available from several manufacturers (Figure 1.1). These new or ‘next-generation’ sequencing (NGS) technologies can be divided into four categories: emulsion polymerase chain reaction (Roche/454, Life/APG, Polonator), solid-phase amplification (Illumina/Solexa), single molecule: primer immobilised (Helicos Biosciences), and polymerase immobilised (Pacific Biosciences, Life/Visigen, LI-COR Biosciences). Compared with their predecessors, these new sequencing machines are superior by several orders of magnitude (0.45–50 gigabases per run). The basic mechanisms for data generation with the new instruments changed drastically, producing more sequence reads per instrument run at a significantly lower expense. Competition among different manufacturers has decreased the cost of the sequencing machines and also democratised sequencing technology [5].

The high amount of data produced and the low cost of operating the new technologies enable the profiling of multiple species simultaneously, thereby improving in situ analysis. Biologists have performed experimental analyses in vivo for a long time, in which a controlled environment permits the observation of a single stress factor in an organism. However, most microorganisms do not exist alone in nature, and their behaviour is influenced by many environmental factors. Thus, an improved understanding of the biology of an organism may be gained by studying the organism within its own environment. Improvements in sequencing have allowed biologists to collect large amounts of data from in situ experiments. The term ‘metagenomics’ first appeared in 1998 [6] to establish the idea that genes from uncultivable microorganism could be analysed directly from the environment in ways similar to those applied to the study of a single genome. An increasing interest in environmental genetics resulted in the broader use of the term metagenomics to describe the sequencing of any genetic material from environmental samples (Figure 1.2). Recently, Kevin Chen and Lior Pachter [7] defined metagenomics as “the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, by-passing the need for isolation and laboratory cultivation of
Figure 1.1: Figure 1 from perspective article [5] aggregating facts, which are important to understanding the impact of sequencing technology in biological research. The timeline shows how recently the new sequencing machines were introduced. Output increased exponentially between 2003 and 2006 and solved problems that would have been unthinkable to researchers just a decade ago.
individual species”.

The earth is a repository of resources that are important for human health and industry and provide food, clean air and water, shelter, and fuel. However, the vast number of resources available has not been fully explored. The total number of living organisms on the planet is currently unknown but is estimated to be more than $10^{30}$ bacteria alone [8]. Prokaryotes are thought to be the first organisms that evolved on Earth, appearing 3.5 billion years ago. Since then, these organisms have occupied the most extreme niches — from nuclear power plants to the human gut — by developing tools to survive and adapt to a wide variety of environments. Bacteria have frequently played an important role in life and industry; however, less than 1% of bacteria have been taxonomically classified, and only a few species (e.g. model organisms) have been physiologically characterised. Nevertheless, many important bacterial enzymes and antibiotics have been identified and are currently in use in health and industrial sectors. The potential for metagenomics applications in industry are many, and they are still in their infancy. In 2003, sales of industrial enzymes were estimated at $2.3$ billion [9], with the top four being detergents ($789$ million), food applications ($634$ million), agriculture/feed ($376$ million), and textile processing ($237$ million) [10]. The use of metagenomics and biotechnology has advantages over the industrial processes currently in use. Microorganisms are already producing many interesting industrial compounds, so the main challenge lies in developing large-scale processes. Large-scale biotechnology can replace or complement traditional industrial production, especially in chemical industries. The potential for biotechnology to replace traditional methods is enormous. For instance, biotechnology can produce an estimated 10–20% of all chemicals sold, with a market value of $160$ billion. The percentage of intermediary products within reach with biotechnology is up to 60% and covers a large range of products (e.g. detergents, agro-chemicals, and pharmaceuticals) [10]. Moreover, biocatalysis can introduce superior synthetic strategies in 10% of these processes, surpassing the effectiveness of classical chemical reactions [11].

Metagenomics may be a key factor in the identification of new enzymes to boost biotechnology. The metagenomics era began when sequencing was still expensive. Thus, before genomes from the whole environment could be sequenced, fragments of DNA had to be cloned into a host (e.g. *E. coli*), and the clones were screened for the enzyme of interest. After screening the clone for the target enzyme, the clone was isolated, extracted from the host, and completely sequenced. After assembly, bioinformaticians conducted a search for the enzyme in the cloned fragment. However, thousands or millions of genes may be present in some environments, many of which have biotechnological potential. Reductions in sequencing costs have enabled researchers to sequence multiple genes without cloning. Thus, metagenomics removes the cloning and screening steps, thereby speeding up the sequencing process and increasing the number of genes that can be analysed.

Metagenomics has the potential to reveal the immense uncultured microbial diversity present in the environment and offer new molecules for biotechnological and ther-
1.1 Metagenomics and sequencing technologies

Figure 1.2: Figure 3 from article [6] showing the extraction of genomic information from unknown soil organisms. This technique symbolises the beginning of metagenomic studies, and the report on this technique was the first to use the term metagenomics and describe the methodology.
apeutic applications. It can also be used to identify the species are present in an environment through 16S ribosomal RNA (rRNA) sequencing, highlight differences in bacterial communities, and evaluate the impact of external variables on a bacterial community — e.g. drugs, food, or bacterial species. These capabilities are particularly important when studying the association of the microbiome — a community of microbes within a given environment — with disease. Early examples of metagenomic applications include the analysis of previously sampled environments, such as the ocean [12, 13, 14], an acid mine site [15], soil [16, 17], and coral reefs [18]. However, these studies were limited by expensive sequencing processes incapable of revealing complex diversity. By contrast, NGS platforms, which provide fast, low-cost, and data-intensive output, have increased the application of metagenomic studies.

The human body contains a high number of ecosystems that are colonised by a diverse number of microorganisms, including the gastrointestinal tract, vagina, skin, and oral and nasal cavities. Variations in the microbial community of these ecosystems may be related to the health status of an individual. The human gut has been extensively studied, through 16S rRNA classification, capillary electrophoresis and 454 pyrosequencing [19, 20, 21, 22, 23, 24, 25]. The International Human Project is a large-scale endeavour aiming to develop a set of reference genomes from the human microbiome by sequencing hundreds of cultured and uncultured microbial genomes found in the human gut. The apparent relationship between human bodies and their microbiota communities combined with improvements offered by NGS technologies have opened the door for investigations of the interconnections among microbial populations and health status. This unique opportunity is being explored by two major projects, the Human Microbiome Initiative in the United States and MetaHIT in Europe. The MetaHIT study has the potential to enhance current knowledge about bacterial interactions with their hosts and with one another. The analysis of bacterial sub-products may help to reveal how the human microbiome influences human health. For example, the International Human Microbiome Consortium (IHMC), which includes MetaHIT, may provide information on the effects of probiotics and prebiotics in the human gut microbiome, thereby aiding the food industry. Moreover, the data produced by the consortium could be used to create tools to facilitate studies of the human microbiome.

The food industry is responding to consumer interest in healthier eating by producing new products. Probiotics are an important aspect of this initiative, and some companies are already using ‘bugs to create bucks’. In 2005, Danone launched a series of products containing the probiotic strain *Bifidobacterium animalis* DN173 010. This probiotic is the key ingredient in Activia, a yogurt marketed as an aid for constipation (Figure 1.3). Activia posted nearly $2 billion in worldwide sales in 2006, an increase of 30% over previous sales. Analysts have reported that the introduction of Activia in the United States in 2006 through the company’s Dannon division was one of the most successful product launches in the recent history of the food industry, with sales of $300 million in 2007. Until recently, most of the bacteria in the human gut — including *B. animalis* DN173010, which has been shown to help ease constipation — had not been fully analysed. However, the capability of modern
1.2 MetaHIT: the project objectives

Figure 1.3: Activia an example of a probiotic product. Activia was recently developed by Danone (known as Dannon in the United States) to fulfil the demand by consumers for healthier products. Activia is classified as functional food, and the company claims that it benefits human health by improving intestinal motility. The yogurt contains Bifidobacterium animalis DN 173 010, which bears the trade names Bifidobacterium Lactis, Bifidus Digestivum, Bifidus Actiregularis, and Bifidus Regularis.

sequencing techniques to map bacterial DNA in few days has expanded the exploration of gut flora. Company communications from Danone insist that, “The more we look, the more we are finding health benefit” [26]. Lactobacillus casei DN-114001, for example, appeared to boost the immune system in clinical trials. As a result, it was added to a yogurt drink introduced in the United States in 2007 under the brand name DanActive [26]. These products highlight the importance of human gut microbiome research and its applications in industry.

1.2 MetaHIT: the project objectives

MetaHIT was a project financed by the European Commission under the Seventh Framework Programme. It includes 13 academia and industry partners from eight countries. The total cost of MetaHIT has been evaluated at more than 21.2 million euros, and the funding requested from the European Commission had an upper limit
of 11.4 million euros. The project started in January 2008 and ended in June 2012. The central goal of the MetaHIT project is to find associations among the genes of the human intestinal microbiota and human health, especially as they pertain to inflammatory bowel disease (IBD) and obesity, disorders of growing importance in Europe. Cases of IBD have been increasing in recent decades in Western Europe, and this tendency is now being observed in Eastern Europe as well. Rises in obesity have been reported in many countries, and both diseases present an increasing challenge for public health services. To reach its principal objectives, the MetaHIT consortium has (1) established an extensive reference catalogue of microbial genes present in the human intestine, (2) developed bioinformatics tools to store, organise, and interpret this information, (3) determined the genes carried by cohorts of sick and healthy individuals, and (4) devised methods to study the function of bacterial genes associated with disease with the aim of understanding their underlying mechanisms and host–microbe interactions. To integrate this project into the real world, the MetaHIT consortium actively participates in the IHMC, transfers technology to industry, and presents information about the project to the general public.

1.3 Thesis objectives

The main objective of this thesis was to advance understanding of how the human gut microbiome affects health. However, to reach such a goal, we needed to create new tools and software to handle and analyse data from metagenomics, NGS, and metagenome arrays. During the period described herein, I developed a new tool to process empirical data on the human gut microbiome. As a result, I created the first human gut microbiome microarray, which generated DNA abundance and RNA expression data. This ‘Gut-Array’ includes 840,000 genes from more than 250 bacterial species commonly found in the human gut, and the results have been presented in the published scientific papers highlighted in this thesis. To design the microarray, I developed software to handle probe design for multiple species. This software, Meta-OligoWiz, is based on OligoWiz [27], which designs microarrays for single species.

Data analysis is an important component of a thorough exploration of the human gut microbiome. Therefore, I developed an R-package (MetaHIT-R) to investigate how the human gut microbiome affects human health. This thesis focuses on the scientific results produced by these tools and is composed of three articles. The first article focuses on the massive amount of empirical data produced by sequencing. The second article discusses data analysis, revealing an unexpected trend that could not be explained in the microbiome data distribution. The third article aims not only to explain the microbiome data distribution but also to define the constituents of healthy and unhealthy human gut microbiomes. It has been recently proposed that ecological theory must be applied to the human gut microbiome [28]. The third article applies many ecological concepts that have changed the way researchers associate
1.3 Thesis objectives

the microbiome with disease. This article is the first to apply ecological concepts to
four datasets using three different technologies. One of the goals of MetaHIT is to
find bacteria that are responsible for disease. The idea of ‘good’ and ‘bad’ bacteria
predominated until recently, when studies emerged that moved the perception of
the microbiota–disease relationship in a different direction. The three articles in
this thesis follow a timeline in the MetaHIT project and reveal how awareness of
this relationship changed. The first article provides the data necessary to study
the microbiome. The second demonstrates an unexpected result—that individual
microbiomes can be clustered into three groups, called ‘Enterotypes’. The third uses
ecological theory to explain the clusters and their link with disease. Changing the
way we link the human gut microbiome to our healthy.
The most relevant and recent knowledge about the human gut microbiome is discussed in this section. The biological relevance of the human gut microbiome is also examined, and the results of recent articles are summarised. Moreover, because the study of the human gut microbiome is a multidisciplinary field, I introduce concepts from scientific disciplines other than metagenomics that are central to explanations of the link between the gut microbiome and disease. These explanations will help readers from all scientific backgrounds comprehend the results and conclusions in this thesis. The background is divided into four sections. In the microbiology section, taxonomy classification is discussed with a focus on the link between phenotype characteristics and functional capacity. The gut microbiome section has three themes: bacterial composition, shifts observed in diseases, and evolutionary aspects of human gut microbiome composition that explain its co-evolution with humans. Another section of the background discusses the factors suggested to influence microbiome composition, and the final section discusses some of the ecological features of the human gut microbiome.

2.1 Microbiology

2.1.1 Biological classification

The grouping of living and nonliving things that have similar characteristics is a common procedure in science. One of the first records of such classification can be found in the work of the philosopher Aristotle (384 BC – 322 BC) during the fourth century BC. Aristotle created biological classifications in his ‘Metaphysics’ book in
Microbiology

which he grouped animals by their characteristics — e.g. animals with blood and those without. Such separations help to reveal the common characteristics of organisms and therefore suggest a common origin for them. Carl Linnaeus (1707 – 1778) would later divide nature into three kingdoms in which species were grouped according to morphology in ‘Systema Naturae’. Linnaeus hypothesised that species with similar physical characteristics were more likely to be related and should therefore be placed more closely in a taxonomic rank. This form of classification is still used today to group species.

Prokaryote classification follows the same idea, applying morphology to define taxonomic rank. Rules can be applied to this process and are described in detail in Bergey’s Manual of Systematic Bacteriology. The overall characteristics include cell wall type, cell shape, oxygen requirements, endospore production, motility, and glucose fermentation activity, among others. However, some species groups are extraordinarily diverse, and morphological characterization is inadequate to separate them into subgroups. The advent of NGS technologies allows the use of new characteristics such as GC content to define similarities and group species. The increased capacity of modern sequencing machines permit the phylogenetic analysis of 16S rRNA genes, which can be used to cluster species into taxonomy groups. This methodology has confirmed many existing groupings, supporting the idea that species, which share morphology, are likely to have common ancestry. These techniques also provide interesting insights on bacterial evolution. Some groups with distinct morphological characteristics have been grouped together using 16S rRNA. Aristotle’s ancient biological classification grouped animals that could fly and animals of the sea. Current classifications reflect that even though bats and birds can fly, their wings have resulted from independent evolutionary events. The same principle can be applied to whales and fish. 16S rRNA adds similar evolutionary information to bacterial classification and is especially important for Clostridium species. Species from order Clostridiales have been shown to belong to clusters that disagree with the current scientific classification [29], suggesting that some of these species have different evolutionary origins. Because most of the species from the human gut microbiome belong to the order Clostridiales, correct taxonomy classification may have a critical impact on the understanding of their role in disease. In fact, species in the colon belonging to Clostridium clusters IV and XIVa induce T regulatory cells (Tregs) [30], which are important to immune homeostasis. Therefore, the use of phylogenetic 16S rRNA can enhance knowledge of the human gut microbiome. Furthermore, grouping organisms by characteristics is a valuable technique, which can further understanding of their function in the human gut microbiome.

2.1.2 Bacterial phenotype

Current biological classification is mostly based on morphology and biochemistry. The idea that species with shared characteristics are likely to have common ancestry is supported by 16S rRNA phylogenetic analysis. Therefore, organisms with simi-
lar phenotypes are likely to perform related functions or occupy analogous niches. Phylogeny classification of organisms in the human gut microbiome can help reveal conserved morphology and elucidate functional capacity, providing insights on the possible link between the microbiome and disease. Herein I discuss the morphological characteristics of the human gut microbiome relevant to this thesis.

In the late 19th century, Hans Christian Gram (1850–1938), a Danish scientist, invented a technique for identifying the presence of peptidoglycan in the bacterial envelope. The technique uses hexamethyl pararosaniline chloride dye, which is retained by peptidoglycan. Thus, high levels of peptidoglycan produce a dark violet colour and low levels result in red/pink colours. Today most bacteria can be divided into two groups by the composition of peptidoglycan in the cell wall. Gram-positive bacteria have large amounts of peptidoglycan (50-90%) and Gram-negative bacteria have small amounts (10%). Gram-positive and Gram-negative bacteria have other differences as well. The Gram-negative cell wall is characterised by the presence of lipopolysaccharide (LPS). LPS is composed of lipids and polysaccharides and is located at the outer membrane. It acts as an endotoxin in animals and activates a strong immune response. However, LPSs display diverse levels of toxicity. LPSs from *Bacteroides* are reportedly 10 to 5,000 times less toxic than hose from *E. coli* [31], indicating that Gram-negative bacteria affect their hosts in different ways.

Bacteria are also often classified by their capacity to use or survive in oxygen. Figure 2.1 depicts a classification of bacteria based on their behaviour in the presence of oxygen. In general, bacteria are grouped by their capacity to produce energy using oxygen. Those that can do so are classified as aerobes, and those that cannot are anaerobes. However, some microorganisms, classified as facultative, can generate energy with or without oxygen. In the presence of oxygen, they produce ATP via aerobic respiration, and without it, they produce energy via fermentation. *Staphylococcus* and *E. coli* are examples of facultative bacteria in the gut. Although some bacteria cannot produce energy from oxygen, they can tolerate its presence and are classified as aerotolerant. *Bacteroides* and *Lactobacillus* are aerotolerant species. Because the human gut is an anaerobic environment, the composition of the microbiome is dominated by anaerobic bacteria. However, increases in oxygen occasionally occur in the human gut—e.g. in cases of damage to the gut wall or inflammatory responses by reactive oxygen species (ROS) [32, 33]. Oxygen in the gut kills strict anaerobic bacteria and increases the abundance of species that tolerate or produce energy from oxygen.

### 2.1.3 Bacterial fermentation: short-chain fatty acids

Short-chain fatty acids (SCFAs) are products of bacterial fermentation under anaerobic conditions. SCFAs are classified as fatty acids with less than six carbons and aliphatic tails. Three SCFAs are integral for human health: acetate (with two carbons), propionate (with three carbons), and butyrate (with four carbons). SCFAs
Figure 2.1: Figure from Wikipedia [34]. *Bacteroides* species are consistently classified as strictly anaerobic, but they have demonstrated a capacity to tolerate oxygen. [35, 36]. Strict aerobic bacteria group close to the top of test tube (1), whereas strict anaerobes group at the bottom (2). Facultative bacteria (3) are more concentrated at the top because they are more energy efficient. Microaerophiles (4) require lower levels of oxygen (around 20%) to survive. Aerotolerant bacteria (5) do not use oxygen to produce energy but can tolerate its presence.

in the human intestine are used by the epithelial cells of the colon (colonocytes) as energy sources [37], and an estimated 5-10% of basal energy in humans is provided by SCFAs [38]. In addition to their roles as energy sources, SCFAs are involved in diverse processes related to the immune system [37]. SCFAs have been shown to activate inflammatory responses [39]. They induce apoptosis in lymphocytes [40, 41], neutrophils [42], and macrophages [43] and affect leukocytes. The role of SCFAs in apoptosis is not well understood but may be relevant to inflammatory response. Butyrate reduces the phagocytic activity of macrophages [44]. Thus, in general, SCFAs, especially butyrate [45], are considered to have anti-inflammatory action [46]. Because SCFAs are products of fermentation, the presence of oxygen in the gut may reduce their production.

### 2.2 Relevance of the gut microbiome

This section describes the importance of studying the gut microbiome and the ways in which it can influence human health. The most well-characterised effects of the gut microbiome on the host include energy absorption from food, production of vitamins, defence against pathogens, and interference with the immune system. Individual digestive capacity depends on microbiome composition [20, 22, 23, 47]. Mice with a
2.2 Relevance of the gut microbiome

High relative abundance of Firmicutes have an increased capacity to extract energy from food [23], and mice given diets that restrict carbohydrates or fats display an increased abundance of Bacteroidetes [48]. Microbiota metabolism of oxalate has been linked to a predisposition for kidney stones [49], and lipid metabolism is affected by microorganisms, which modify bile acids [50]. Interactions between the microbiota and immune cells have a demonstrably important impact in their host. Gut epithelial cells renew more slowly in germ-free mice than in colonised mice [51]. The presence of the microbiome can influence abnormal cell proliferation (neoplasia) [52] and alter the capacity of cells to repair damaged mucosal barriers [51]. In addition, the susceptibility of the gut to colonisation by enteropathogens is affected by the capacity of the microbiota to alter the expression of host genes encoding anti-microbial compounds [53, 54]. Moreover, the human gut microbiome affects the innate and adaptive immune systems [53, 55]. However, its impact is not restricted to the gut and can also be manifested in other organs; the incidence of asthma is correlated both with exposure to bacteria [56] and treatment with broad-spectrum antibiotics in childhood [57]. Taken together, those results demonstrate the importance of studying the human gut microbiome and its effect on health.

2.2.1 Gut microbiome composition

Although studies have demonstrated the importance of the human gut microbiome, its precise composition remained largely unknown owing to the low number of samples examined. However, superior sequencing technologies and a reduction in the cost of sequencing have recently allowed the sequencing of high numbers of samples, thereby expanding knowledge of human gut microbiome composition. The accurate characterisation of the gut microbiome and its variants is an important step in defining healthy and unhealthy microbiomes. After the profiling of hundreds of samples using various technologies, researchers have generally agreed that the human gut microbiome is mostly composed of four phyla: Actinobacteria, Firmicutes, Bacteroidetes, and Proteobacteria. Bacteroidetes and Firmicutes represent 90% of the relative abundance of samples. However, in the gut, these two phyla differ in their species diversity. Whereas Firmicutes has 12 families with more than 1% of the gene count, Bacteroidetes has only five. Moreover, members of Firmicutes have a greater distribution than those of Bacteroidetes. For instance, the most abundant genera from the Bacteroidetes phylum are Bacteroides and Prevotella, which represent 80% of all Bacteroidetes in most samples. The gene count is 71% and 7% for Bacteroides and Prevotella, respectively — because few Prevotella genes have very high relative abundance in some samples [58]. The phylum Bacteroidetes is composed of Gram-negative, non-endospore-forming, anaerobic bacteria; however, they are aerotolerant (Figure 2.1) — i.e. they tolerate the presence of oxygen but cannot use it for growth [59]. The name Firmicutes originates from the Latin ‘firmus’ meaning strong and ‘cutis’ meaning skin. The organisms in this phylum are characterised by a Gram-positive cell wall structure. They are mostly strictly anaerobic, and some produce endospores. Although Firmicutes is an extremely diverse phylum
in the gut microbiome, the orders Clostridiales and Lactobacillales contain 67% and 12% of all Firmicutes genes, respectively. The top four Firmicutes families in gene counts are Lachnospiraceae (23%), Clostridiaceae (19%), Ruminococcaceae (12%), and Erysipelotrichaceae (10%). This microbiome composition has been described using various methodologies, such as 16S rRNA, Sanger, 454, and Solexa/Illumina. Therefore, variations in the microbiome composition are not biased by the distinct technologies used. In 2009, the relative abundance of Bacteroidetes was demonstrated to explain a substantial part of the variation in data (Figure 2.2) [60]. However, the abundance of Bacteroidetes has not been associated with any metadata, and the cause of the shift from Bacteroidetes to Firmicutes remains unclear. The composition of the human gut microbiome is similar to that of the mouse microbiome. In both microbiomes, Bacteroidetes and Firmicutes compose more than 90% of the relative abundance [21, 48, 61]. This similarity suggests that the composition of the gut microbiome is defined by a factor that was present in the common ancestry of humans and mice.

As the abundance of Bacteroidetes can explain most of the variance in the gut microbiome it became common to plot the distribution sorting the samples by Bacteroidetes. This has been publishing in different studies (Figure 2.3), but the conclusion about the shifts have been different in each study and in some cases there were no association found among the shift of the microbiome with diseases or any metadata. Without a clear explanation this shift has been marginalized and most of the time present only as supplementary material. Therefore, even when the studies demonstrate the same microbiome distribution, the conclusion has been not similar. Until now Bacteroidetes (or only Bacteroides) have been related with obesity (low Bacteroidetes [60]), high-fiber diet [62] (high levels of Bacteroidetes), protein and animal fat [63] (high levels of Bacteroides), and no association with BMI [64].
Figure 2.2: Figure from the article [60] as figure 2b demonstrating that the relative abundance of Bacteroidetes correlates with PC1 (20%). Because Bacteroidetes and Firmicutes compose more than 90% of the relative abundance of the human gut microbiome, the ratio of their relative abundance explains most of the variance in the data.
Figure 2.3: Data from two studies of human gut microbiota. Although the results of these studies are interpreted in different ways, both demonstrate the same shift in human gut microbiota composition, demonstrating the importance of the Bacteroidetes/Firmicutes ratio. This abundance shift has also been observed in other studies. The figure 2.3a is the supplementary information as figure 7 from article [60]. The figure 2.3b is figure 2A from article [65].
2.2 Relevance of the gut microbiome

2.2.2 Human gut microbiome shifts related to disease.

The overall composition of the human gut microbiome is described in Section 2.2.1 based on the human gut microbiome catalogue [66]. To put the composition of the gut in perspective with health status, I describe herein some of the diseases that have been associated with shifts in gut microbiome composition. Reductions in the abundance of Firmicutes have been observed in IBD patients [67]. Three studies have documented that decreases in Firmicutes abundance and increases in Proteobacteria are associated with Crohn’s disease (CD) [68, 69, 70]. Moreover, decreases in Firmicutes abundance, specifically that of *Faecalibacterium prausnitzii*, has been described in CD [71, 72, 73]. A study of intestinal tissue from IBD patients has demonstrated that a *Prevotella* species was 3.6 times more abundant at inflamed than at non-inflamed sites, with total abundance of *Prevotella* reaching 25% of the total community in inflamed samples [68]. A high abundance of members of the genera *Bacteroides* and *Prevotella* in mucosal biopsy specimens from UC patients has also been reported [74]. In another study using fluorescence in situ hybridisation, mucosal bacteria were found at concentrations greater than $10^9$/mL in 90–95% of IBD patients compared with 35% of healthy controls [75]. The mean density of the mucosal biofilm was two powers higher in IBD patients than in controls, and bacteria were mostly adherent [75]. Increase in bacterial concentrations was also observed in IBD patients (Figure 2.4a) [76], T5KO mice (Figure 2.4b) [77], and in ileitis-induced by *Toxoplasma gondii* in mice (Figure 2.4c and 2.4d) [78, 79]. Although all studies do not agree completely and variation may occur owing to methodology and other effects, most have consistently reported that reductions in Firmicutes and increases in *Bacteroides*, *Prevotella*, and *E. coli* are associated with disease. This shift in composition observed in humans has also been observed in mice (Figure 2.4d) [79]. The similarity in both composition and shifts indicates a common mechanism, which can modulate the microbiome. In both cases the conclusion is the decrease of Gram-positive and increase of Gram-negative species [78]. In at least four different studies was demonstrated that the shift in the microbiome composition (from Firmicutes to Bacteroidetes) is associated with increase in bacterial concentration (Figure 2.4). With increase of Gram-negative, especially (*Bacteroides*, *Prevotella*, and *E. coli*). Bringing the question why increase in bacterial concentration? And can this increase have any impact on the diseases?

2.2.3 Evolution of the gut microbiome

Section 2.2.2 describes the composition of the gut microbiome and its shifts associated with disease, both of which are similar between humans and mice, alluding to their common origin. Moreover, the gut microbiomes of other organisms such as hoatzins (bird), cows, and macaques are also reportedly dominated by Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria [80, 81]. This composition suggests that the factor, which defines human gut microbiome composition was present
(a) Increase in bacterial concentrations in IBD patients [76].

(b) Increase in bacterial concentration in T5KO mice [77].

(c) Increase in bacterial concentration after ileitis-induced by *Toxoplasma gondii* and indomethacin in mice [78].

(d) Increase in bacterial concentration from mice with severe ileitis at day 8 after *Toxoplasma gondii* infection [79].

**Figure 2.4:** Figure composed of four different studies [76, 77, 78, 79]. In all studies the shift of the microbiota (from high Firmicutes to high Bacteroidetes and/or *E. coli*) is associated with increase in bacterial concentration. Moreover, they also report decrease in bacteria diversity. Figure 2.4a from study [76] as figure 1. Figure 2.4b in supplementary information as figure S13b [77]. Figure 2.4c from study [78] as figure 2a. Figure 2.4d from study [79] as figure 3.
Factors shaping microbiome composition

in a common ancestor. This commonality is not surprising given that the first eu-
karyote appeared on Earth approximately 2 billion years ago, and the environment
was dominated by prokaryotes for approximately 1.5 billion years. As eukaryotes sur-
vived and evolved in this environment, they found a means to coexist. The solution
probably occurred through various means, but its basic mechanism may still have
been the same. To defend against other organisms, the first eukaryotes must have
found a way to distinguish their own cells from external ones. Because the human
body is constantly exposed to a large number of microbes, it must recognise its own
cells, avoiding self-attack but also allowing the destruction of harmful organisms.

One important mechanism for the recognition of non-self organisms such as viruses,
bacteria, and fungi was discovered in 1985, when Christiane Nüsslein-Volhard identi-
fied a gene in Drosophila, which can activate immune response. The gene was named
Toll, and since then, Toll-like genes have been observed in many groups of eukaryotes
from plants to animals, suggesting a very early common ancestry [82]. In animals,
Toll-like receptors recognise features that are conserved among microbes, virus, and
fungi located mostly in cell compartments (e.g. flagellin, LPS, double-stranded RNA,
zymosan). The innate immune system is designed to recognise molecules shared by
groups of related microbes. These unique microbial molecules are called microbe-
associated molecular patterns [83]; they are not associated with mammalian cells and
are essential for the survival of microbes. Included in the innate immune response
is the capability to identify molecules shared by a large group of microorganisms
and distinguish them from self, activate a mechanism that eliminates most microbes
within hours, and induce and guide adaptive immune response [84]. Even if species
from the group Homo (e.g. Homo sapiens and Homo erectus) have mechanisms that
are completely different from those of other species to distinguish self and non-self
cells, they would have been defined through human evolution over 2 million years
[85]. Therefore, factors like lifestyle and nationality that are recent in human history
should not affect the microbiota-host relationship or shape gut composition, as has
been suggested before. The gut microbiome and its host form a mutualism relation-
ship where the microbiome provide the host at least two very important functions.
Protective function capable to defend against non-indigenous organisms and capacity
to extract more energy from food. These two functions were extremely important
over millions of years of evolution. Therefore, the evolutionary success of the human
species is associated with the gut microbiota capacity to perform these two functions
in the most optimal way. Thus in section 2.3, I discuss factors that have the most
relevance in defining the composition of the gut microbiome.

2.3 Factors shaping microbiome composition

Since the formation of the IHMC, many articles have discussed the possibility that the
human microbiome is determined by multiple factors — e.g. disease, diet, genotype,
physiology, environment, and immune system — which could shape it into various
forms. Because every individual has been thought to have a common or core micro-
biome [60], these factors would likely be the main influences in human microbiome variation (Figure 2.5). The core microbiome theory has lost strength over time as newer studies have concluded that individuals have unique microbiomes. However, the idea of multiple factors remains strong, and many researchers have attempted to explain the variance in gut microbiome composition. Although multiple factors are likely to contribute to microbiome composition, the relative strength of such factors remains unknown. Many studies have tried to identify the forces that shape the human gut microbiome, but most of the results have not been confirmed by additional studies, which may be due to the use of a low number of samples or lower sequencing depth. Moreover, some authors have drawn conclusions from their results without taking into account the influence of all possible factors. In this section, the factors that have been studied and their effects on the human gut microbiome are discussed.

2.3.1 Impact of genetics on gut microbiome composition

Genotype could contribute to the determination of the human microbiome. To isolate this factor, some studies have used samples from twins with obesity and IBD. Measurements of the distances between samples from obesity using UniFrac have revealed that samples from monozygotic twins are more similar to each other than to unrelated or family samples. However, self-sample comparison (i.e. samples from the same individual) has shown the smallest distance among all comparisons. In fact, the distance between self and monozygotic twins was greater than that between monozygotic twins and unrelated individuals. Moreover, the degree of similarity in the gut bacterial communities of monozygotic twin pairs was not significantly different from that between dizygotic twin pairs [60]. If genetics are the most influential factor in shaping the human gut microbiome, self samples and twin samples should be very similar in their composition. However, this similarity is absent. Therefore, other factors may have a greater influence than genetics. Similar results have been observed in IBD samples. Twin studies for IBD have established that concordance between both monozygotic and dizygotic twin pairs is 10%–15% for ulcerative colitis (UC) and 30–35% for CD in monozygotic twins [87]. These results demonstrate that genetics may have greater influence in some diseases than in others, suggesting, that a combination of environmental and genetic components are necessary to explain the observed results.

2.3.2 Impact of diet on gut microbiome composition

The effects of the modern Western diet on the human gut microbiome have been widely studied, particularly those related to obesity. One review on the influence of diet on human health has linked diet, gut microbiota, and immune responses to SCFAs and omega-3 fatty acids [46]. However, the mechanisms through which diet shapes the human gut microbiome remain unclear. Some authors have claimed that
2.3 Factors shaping microbiome composition

Figure 2.5: Figure from an insight article on the human gut microbiome [86] showing factors that may be important in defining its composition. The illustration also suggests that all individuals have a core microbiome, which provides basic functions necessary for the human body, and a variable microbiome that would be defined by the influence of various factors in their individual lives.
a high-fibre diet is related to high level of Bacteroidetes species, [62] but others have indicated that high levels of Bacteroides species are associated with diets based on protein and animal fat [63]. As previously mentioned, members of Bacteroidetes and Firmicutes are the most abundant phyla in the human gut microbiome, and both show high variance in samples. Thus, the Bacteroidetes and Firmicutes (B/F) ratio has been analysed in several studies. The B/F ratio in the mouse microbiome has been linked to obesity [25]. However, another study found no association between B/F ratio and body mass index, which is a measure of obesity [64]. In 2010, Alan Walker [88] analysed the impact of strict diets on the human gut microbiome and suggested that samples clustered more strongly by individual than by diet, confirming previous evidence of a stable microbiota. However, Alan Walker also observed an increase in groups of species after dietary change. For example, a restricted starch diet increases the abundance of species from Ruminococcus bromii and Eubacterium rectale [88].

The impact of diet on human health is well known and can be linked to immune response [46]. However, the claim that diet can shape microbiota composition has not yet been proved. Moreover, disagreements persist among the conclusions drawn from diet studies. Studies have interestingly presented data similar to those of previous microbiome studies (e.g. B/F ratio) but with different interpretations and conclusions. A recent article by Marcus J. Claesson [89] has suggested that gut microbiota composition is correlated with diet and health in the elderly. Because the gut microbiome can be affect by multiple factors, knowing whether the composition correlates with diet or health is difficult. In addition to the challenge of isolating dietary factors in his study, Claesson concluded that such factors “...would be difficult to correlate with health parameters, but become far more evident in the elderly who are immunophysiological compromised”. The association between health in the elderly and gut inflammation has been demonstrated previously [90]. In addition, the supplementary material for the Claesson study (figure 9; [89]) demonstrates that the groups analysed show significant differences in immune system molecules (e.g. interleukin [IL], IL-6, IL-8, IL-10, tumour necrosis factor-alpha [TNFα]). Together, these results indicate that diet does not play a central role shaping gut microbiome composition but is important to health. Moreover, combining the latest results from Claesson [89] and a review from Kendle Maslowski [46] suggests a larger influence of the immune system on the gut microbiota than previous thought.

2.3.3 Impact of the immune system on gut microbiome composition

The human intestines have the largest surface area of any organ in the human body at \( \sim 200 \, \text{m}^2 \) (the skin has only \( \sim 2 \, \text{m}^2 \)). The intestines also harbour the most diverse and dense microbial communities in the body, with more than \( 10^{12} / \text{cm}^3 \) of intestinal content. A very thin epithelial layer (10 \( \mu \text{m} \)) and mucus separates this enormous bacteria community from body tissues. The layer is divided into inner and outer
layers, and total mucus thickness varies from 15 to 700 µm [91]. The large intestine has the thickest layer, which may be related to its higher concentration of bacteria [92]. The idea that the immune system influences gut microbiome composition is related to the fact that the intestine is the largest interface between the immune system and the environment. Just the presence of commensal bacteria in the gut is enough to induce low-grade inflammation which is absent in newborn humans and germ-free mice [93, 94, 95]. Some publications linking the immune system with inflammatory diseases like IBD have confirmed shifts in the microbiome composition in humans and mice. For instance, an increase in *Prevotella* abundance has been identified in inflamed intestinal tissue from IBD patients [68], and high abundance of *Bacteroides* and *Prevotella* have been observed in UC patients [74]. Similar shifts have been observed in other studies of patients with IBD [68] and obesity [96]. The same shift was observed in mice when immune system genes (e.g., *Nlrp6*−/−, *Nlrp3*−/−, *Rag*−/−, and T5KO) were knocked out [97, 98, 99, 77], and in non-obese diabetic mice the absence of MyD88 increases the abundance of the Bacteroidetes phylum [100]. Recently, a similar shift was discovered in a murine model of ileal CD, characterised by a decrease in Firmicutes (Gram-positive) and an increase in Proteobacteria (Gram-negative) with reduction in diversity [78].

Reduction in diversity has been extensively reported in many diseases under study. Moreover, increases in bacteria concentration have also been observed in IBD patients [76, 75]. Unfortunately, few studies have measured important variables such as microbiota composition, immune response, species diversity, and bacterial concentration. It may seem illogical that diversity decreases when bacteria concentration increases, but this observation has been confirmed not only in the gut microbial composition but also in other microbial ecological systems [101, 102]. A 2006 study by Markus M. Heimesaat [79] includes all important variables, explaining for probably the first time the dynamics of the human gut microbiome ecosystem. Heimesaat reported that “analysis of the intestinal microflora revealed that ileitis (inflammation of the ileum) was accompanied by increasing bacterial load, decreasing species diversity, and bacterial translocation. Gram-negative bacteria identified as *Escherichia coli* and *Bacteroides/Prevotella* spp. accumulated in inflamed ileum at high concentrations”. In the same report, he demonstrates that antibiotic treatment prevented ileitis. The decrease in Gram-positive bacteria and increase in Gram-negative bacteria tolerant of oxygen suggest an important role for the innate immune system in shaping human gut microbiome composition. Shifts in microbiota composition are not linked only to gut-related diseases. Other inflammation-related diseases have been accompanied by similar shifts in microbiota. A review by Mairi Noverr and Gary B. Huffnagle [103] explores whether gut microbiota regulate immune response outside of the gut. The results indicated that alterations in intestinal microbiota are linked to obesity, asthma, diabetes, IBD, and other inflammatory conditions. For instance, infants with atopy display fewer bifidobacteria in stool [104]. Rheumatoid arthritis is associated with a reduction in anaerobic bacteria [105] and a reduction in bifidobacteria, *Eubacterium rectale*, and Bacteroidetes [106]. This suggests the link microbiome-inflammation may have consequences in other parts of our body. Indicating a bigger importance of the human gut microbiome then previous though.
A recent review by Lora Hooper [91] aggregates conclusions about the impact of the immune system on gut microbiome composition and concludes that “it is now clear that the immune system plays a central role in shaping the composition of the microbiota as well as its proximity to host tissues”.

2.4 Applying ecology to explain the human gut microbiome.

One of the main goals of the IHMC is to understand the relationship between the microbiome and disease. Thus, IBD (e.g. CD and UC) have been extensively studied. The search for biomarker species, which can explain the links among gut microbiome and disease have been the focus of most of these studies. Using classic two-group analyses, these studies have separated data into healthy and unhealthy groups, looking for one bacterial species responsible for disease. However, multiple studies to date have been unable to find pathogenic species to explain diseases such as IBD and obesity. The lack of biomarkers has been surprising, because the influence of the microbiome in disease has been established. For example, in human and murine models of disease, the use of antibiotics reduces or prevents inflammation, suggesting an important role for the gut microbiome in disease [107, 75]. Moreover, a protective role for the microbiome has been indicated when germ-free mice develop disease (induced by dextran sulphate sodium) more severe than that in mice with normal flora [108]. Because studies have been unable to find disease-associated biomarkers, an alternative explanation for the microbiota–disease relationship has been gaining strength. Several opinion, perspective, and review articles have suggested that disequilibrium in the microbial ecosystem — known as dysbiosis — is a possible mechanism [109, 110, 28, 111, 112, 113]. At the same time, many researchers have identified differences in diversity (defined by the Shannon’s diversity index, including richness [or alpha-diversity] and evenness) between healthy individuals and those with disease. Unhealthy individuals with low-diversity microbiomes have been consistently reported (Figure 2.6) [58, 60, 68, 114].

These data imply that the role of the human gut microbiome in disease may be different than expected. The use of ecology theory has recently been proposed as a way to understand the human gut microbiome [28]. One of the interesting points of this suggestion is the idea of alternative stages of the microbiome in a community stage landscape. The theory of alternative stages suggests that a community can shift directly (e.g. antibiotics) or indirectly (e.g. environmental factors and immunity), changing from one stable stage to another. The third article presented in this thesis describes the application of ecological theory to explain the shifts observed in the human gut microbiome. Thus, in this section I explain ecology features used in the last article.
2.4 Applying ecology to explain the human gut microbiome.

(a) Measures of bacterial diversity in IBD patients. [68]

(b) Microbial biodiversity is reduced in children with severe UC [114].

(c) Distribution of nonredundant bacterial genes in IBD patients and healthy controls [66].

(d) Diversity of lean and obese individuals [60].

Figure 2.6: Data from four studies showing that unhealthy individuals have lower microbiome diversity than that of healthy individuals. Figure 2.6a is figure 3b from article [68]. Figure 2.6b is figure 1 from article [114]. Figure 2.6c is supplementary figure 8 from article [66]. Figure 2.6d is figure 1b from article [60].
2.4.1 Diversity

This section discusses the various types of diversity measurements often used in ecological studies. Many of these methods are used in the third study presented below and are important to explanations of shifts in human gut microbiome composition. This discussion will help readers who are unfamiliar with ecological terms and theory to understand the meaning of the study results and conclusions. Some differences exist between the traditional use of ecological indices and their use in metagenomics and microbial ecology. For instance, spatial analysis (i.e. topological, geometric, or geographic properties) is not used to investigate the human gut microbiome because most of the studies use stool samples. Another important point is the lack of absolute measurements. Most of the studies to date have not included values of absolute abundance; consequently, it is unknown whether 1 gram or 30 grams of stool were required to extract the necessary amount of DNA for sequencing. This information can be invaluable, as it indicates the concentration of bacteria in the ecosystem, which may be relevant to explanations of the microbiota–disease link. In the near future, this information may be more commonly included in the analysis. Currently, however, most studies of the human gut microbiome using metagenomic sequencing refer to their abundance as relative and not absolute. Thus, the abundance of a species will always be a proportion of the total, which allows the comparison of relative changes between samples. Moreover, the abundance of the species will be influenced by the richness of the sample. For instance, if one sample has 100 species and the distribution of the species is even, then each species should have 1% relative abundance. If another sample has 200 species with even distribution, then the expected abundance of each species is 0.5%. Consequently, richness affects the relative abundance of the species in a sample. Finally, a sample may contain only four species with the same absolute abundance as a 100 species sample and the same evenness, but the relative abundance of those species will be 25%. Therefore, results from relative abundance data must be interpreted carefully.

Diversity, according to ecologists, is the measurement of two features of an ecosystem: the total number of species present in an ecosystem (i.e. richness) and the evenness of their distribution (i.e. evenness). It is important to emphasise that richness is the total number of species found in one habitat and is unrelated to the total number of individuals present. Traditionally, ecology studies cannot measure the whole community at once and the analysis of alpha diversity is determined through multiple samples of the ecosystem. Then the mean of the samples is used to determine alpha diversity, including the proportional abundance of species. The use of metagenomics in microbial ecology allows the sampling of the whole environment at once and, in most studies, the sample is sequenced only once owing to high sequencing costs. Because multiple sampling does not take place and the entire habitat can be profiled at once, richness is used instead of alpha diversity. The index of diversity is the combination of richness and evenness. One of the most popular diversity indices currently used is the Shannon diversity index; however, many other methods are appropriate. One of the main differences among indices is the method for weighing rare species.
2.4 Applying ecology to explain the human gut microbiome.

The third article presented in this thesis and most of the studies published to date use the Shannon’s diversity index (Formula 1):

\[ H' = -\sum_{i=1}^{R} [p_i \times \log(p_i)] \]

Formula 1: Shannon’s diversity index. \( R \) is richness and \( p_i \) is the proportion of the species abundance in the habitat. Any log base can be used, but traditionally is has been used log base 2.

Evenness is an important measurement, because two ecosystems can have the same richness and total number of individuals but very different evenness. A simple explanation can be observed (Figure 2.7), in which each system has the same amount of richness (four species) and a total of 12 individuals. Intuitively, system ‘A’ is less diverse than system ‘B’ because system ‘A’ is composed mostly by one species. This formula can transform this intuition into a value that can be used for the evenness formula from the Shannon diversity index (Formula 2):

\[ E = \frac{H'}{\log(R)} \]

Formula 2: Evenness index from Shannon’s diversity index. \( E \) is the evenness index, \( H' \) is the Shannon’s diversity index and \( R \) is richness.

Beta diversity is used to calculate the difference in richness between two habitats and is often used to create a distance matrix among habitats to produce clusters of samples. However, it does not explain how different the communities are. To calculate beta diversity, we first need to explain gamma diversity. Gamma diversity is the total number of unique species observed in two habitats. In figure 2.8, the gamma diversity is 4 — the same value as richness — and both ecosystems have the same type of species. Beta diversity is calculated by dividing gamma diversity by alpha diversity (in this case, richness). However, beta diversity does not tell much about how different these ecosystems are, and one of the most important problems in studies of the human gut microbiome is determining whether one ecosystem is a subset of another or completely new. To solve this problem, I created a new diversity index called delta diversity, following the diversity indices created by Robert Harding Whittaker [115]. Delta diversity calculates the number of unique species that are shared by two ecosystems. To quantify how similar a lower-diversity ecosystem is to a higher-diversity ecosystem, I divided delta diversity by richness (or alpha diversity) from the ecosystem with the low Shannon diversity index (Formula 3):
Figure 2.7: Both ecosystems have a species richness value of 4 and a total of 12 individuals. The figure demonstrates how richness and total number of individuals in some cases incapable of explaining diversity. Each colour and shape represents a different bacterial species. Comparing the two biota reveals that ecosystem ‘A’ is dominated by one species, whereas ecosystem ‘B’ has even distribution of species. Therefore, ‘B’ has higher diversity than ‘A’.
2.4 Applying ecology to explain the human gut microbiome.

\[
\delta = R_H \cap R_L
\]

Formula 3: Delta diversity as the intersection of the richness from the high with richness from the low diversity ecosystem. \( \delta \) as delta diversity representing the number of unique species which are shared among two ecosystems, \( R_H \) the richness from the ecosystem with the highest diversity and \( R_L \) the richness of low diversity ecosystem.

\[
\Delta = \frac{\delta}{R_L}
\]

Formula 4: Proportion of delta diversity: \( \Delta \) is the similarity measurement among two ecosystems; \( \delta \) is the number of unique species shared among two ecosystems, \( R_L \) the richness of low diversity sample. The same formula can be calculated using alpha-diversity following the same logic.

Delta diversity is the proportion of species in the low-diversity species that is present in the high-diversity species (Formula 4). The assumption is that if two ecosystems are likely to be related, then most of the species in the low-diversity ecosystem should be present in the high-diversity system. The higher the proportion, the higher the likelihood that one ecosystem is related to the other. For instance, an ecosystem displays decreased diversity if it is under any perturbation, thereby losing species richness. Based on beta diversity, those ecosystems might be considered different, but delta diversity reveals that the low-diversity ecosystem is a subgroup of the high-diversity ecosystem. In this case, the main difference between the two ecosystems is the variance in abundance in addition to the loss of richness. When one system is a subgroup of another system, the overall system can be described as nested, a concept discussed in Section 2.4.2.

2.4.2 Network

In nature, most organisms do not exist alone and must interact with other species. Only rarely do samples from an environment contain cells from just one species. This generalization is true even for high eukaryotes, including humans: 90% of cells in the human body are actually bacteria from thousands of different species. The mechanisms of interactions among bacteria and between bacteria and their hosts are unknown. Therefore, some important questions should be addressed: Do species preferentially interact with a certain taxonomic group? Which species are dependent, and which are self-sufficient? Which microbes interact with one other? What are the dynamics of the ecosystem under pressure? Answering these questions is biologically
Figure 2.8: Diversity among three ecosystems. This example depicts how richness and beta, gamma, and delta diversity are calculated. Each diversity measurement explains some of the differences between the ecosystems. Many studies have used beta diversity to calculate the difference between ecosystems. However, beta diversity cannot explain whether one ecosystem is subset of another ecosystem or is a new ecosystem, composed completely of new species. In this example, ‘Blue’ is a subgroup of ‘Red’, and ‘Green’ is an ecosystem composed mostly by new species.
2.4 Applying ecology to explain the human gut microbiome.

relevant, and understanding ecosystem dynamics may have medical implications. In
this section, I explain three important concepts of microbial interaction.

2.4.2.1 Ecological relationship

In ecology, the interaction of any species with another can be classified as a type
of relationship. Relationship categories were introduced by William Lidicker [116]
and are based on the effect that occurs between the species. For instance, when
a relationship is beneficial for both species, it is called mutualism, and when it is
disadvantageous for both, it is called competition. In some cases, a relationship can
be positive for one species and negative for another — e.g. parasitism. The ecological
relationship model is important in this thesis to explain the interaction among the
species in a network. However, defining exactly which type of relationship occurs
between two species requires experimental analysis beyond the scope of this study.
A general categorization of the interaction was made instead, and relationships were
simplified in two types — positive and negative. Any positive interaction could be
positive-positive or positive-neutral, and the same possibilities applied to negative
relationships.

2.4.3 Microbial ecological network

An ecological network is a representation of the interaction among species in which
species are represented as nodes and associations among species are designated by
a link connecting them. The properties of these two features are, in general, used
to explain the nodes and interactions further. For example, node size is commonly
represented as the abundance of the organism, and its colour represents the taxono-
mic group. The associations are represented by arrows that indicate the direction
of the relationship. This information was unavailable for the thesis studies. Thus,
in microbial networks, colour is commonly used to represent positive and negative
interactions (blue and red, respectively). The transparency of the colour can also be
used to explain how strong associations are, with weak associations close to white
and the colour darkening with the growing strength of the association. These colour
assignments provide a visual representation of the relationship and emphasise the
strongest links in a complex network. To shed light on some of the questions above,
we built the first species network of the human gut microbiome using the species
abundance concepts presented in the first article below [66]. The approach is based
on the presumption that if samples of two species show co-abundance, they may have
an association. The assumption becomes stronger as more samples are analysed.
Figure 2.9 displays the steps necessary to build a microbial network. The micro-
bial abundance network starts with the genes abundance for each sample. Which
in sequencing is measured by count of reads and in microarray by signal. Genes
are combined to form species using taxonomy classification. The abundance of the
species is determined by the median abundance of the genes. Creating the species abundance matrix. Then a method to find the pairwise relationship is used. Several methods are appropriate, and they are thoroughly discussed in a recent review [117]. The methods create a matrix with pairwise scores, which are filtered by the user to determine which species are associated. Positive associations mean that both species have the same abundance profile and will therefore increase and decrease in abundance together. A negative association occurs when two species have opposite abundance profiles; therefore, when one increases in abundance the other decreases. Although this approach may be simple, it can aid the search for answers to the questions posed at the beginning of this section.

The microbial abundance network permits the identification of bacterial interactions from among hundreds of species. We can also observe whether a species from a taxonomic group will preferentially interact with the same group or another, often revealing some unexpected associations among different phyla. Another interesting insight is the level of independence that a bacterium or a group of bacteria display. For instance, a lack of association with other species or group of species may indicate that a species can be present in samples without the presence of other species, suggesting that it is self-sustaining. The network also reveals whether one group of species is more dependent than another and identifies the organism with the highest level of dependency. This species may be important in the ecosystem, and if it disappears, other species may be affected, in turn becoming lower in abundance or even disappearing completely.

2.4.4 Properties of ecological networks

Ecological networks are important tools with which to explain ecosystems. The representation of species interactions is just one of many relevant properties. Ecologists study networks using empirical data and model their dynamics with the goal of identifying common patterns among various ecosystems. Those patterns are basic mechanisms, which can explain the emergence of an ecosystem and its dynamics. Therefore, these properties can be used to describe an ecological network more thoroughly. Herein I discuss some of the properties that are used in the third article presented in this thesis. These properties are keys for understanding the study results and their relationship with other ecosystem studies.

Ecological networks were first developed by ecologists to explain food web interactions. In food webs, specialist and generalist species are defined by their diet ranges. Species, which fed on only a few species, are considered specialists and those with a broader dietary range are generalists. This concept was extended and can also be applied to the capacity of an organism to survive in a specific environment. For example, some plants can survive only within a narrow range of temperature, humidity, rainfall, and sun exposure. Changes in the ecosystem are more likely to influence specialist than generalists. Species degree is a property of the ecological network,
2.4 Applying ecology to explain the human gut microbiome.

Figure 2.9: Figure explaining the steps necessary to build a microbial ecological network. Figure 2.9A gene abundance matrix where colours represent the taxonomic classification. Figure 2.9B species matrix with the abundance of each species per sample. Figure 2.9C pairwise correlation matrix among all species. Blue colour represent positive association and red colour negative. The intensity of the colour represent how strong are the association. Figure 2.9D Network output from created from the association matrix.
which describes the number of associations a species has with other species in a network. In food web networks, generalists have higher species degrees (i.e. the species feeds on a broader range other species) and specialists have lower species degrees representing their limited diet. However, networks based on the capacity to adapt to changes in the environment reflect the opposite characteristics. Generalists have species degrees lower than those of specialists. An environment that remains stable generates specialization — e.g. a fish and its parasite will co-evolve. The co-evolution confers greater advantage to species living in that niche and introduces barriers for alien species (species that did not co-evolve in the environment) to colonise the niche. The stable and strict environment increases biodiversity through specialist species evolution. The relationship of specialist species in stable environments is, in general, mutualistic among the species and possibly with the host. Therefore, networks reflecting adaptation in a stable environment have greater species degrees in specialist species and lower degrees in generalists. Where the specialist species provides the most optimised function within an environment, the generalist provides less optimal functions when the ecosystem is under perturbation.

Two concepts of nestedness, which are important in this thesis, occur in ecological networks. The first concept originates in food web networks in which specialists nest with a subset of generalist species. The second concept is an extension of nestedness and is applied to comparisons of networks (from different habitats or in biodiversity conservation studies). Bruce Patterson [118] defined the principle of nestedness in 1987. A biota is considered nested if the species present (elements contained) in small biotas are also present in richer ones. In a non-nested biota, species in low-diversity ecosystems differ from those in high-diversity ecosystems. An ecosystem under pressure loses species (richness), transforming into a subset of the original biota; it cannot be considered a new ecosystem.

The stability of networks has been the subject of study for many years. In 1984, Stuard Pimm concluded in a review that complexity leads to instability [119]. The definition of complexity in a network is based on the number of species degrees divided by the number of species in the network. Thus, a complex network has many links among species. Pimm’s theory is that if one key member of a network disappears, all dependent species also die. Therefore, complexity leads to instability. However, new features such as compartmentalisation increase the stability of the network. Compartmentalisation occurs when a network consists of subnetworks or a large network composed of smaller networks connected by very few species. Moreover, an increase in interaction among species decreases interaction strength. For instance, if species ‘A’ has only one interaction with ‘B’, its interaction will be very strong and dependent. However, if A interacts with five species, the interactions will be weaker, and disturbances in the ecosystem will not affect the overall network. Therefore, increased interactions lead to a stable ecosystem. Network stability is an important feature because it can determine the capacity to withstand and recover from perturbations. Resilience is the capacity of an ecosystem to be disturbed and recover quickly. Most ecosystems have levels of stability and resilience because they are all exposed to stochastic events (e.g. fire, flooding, hurricane, earthquakes, insect plagues, and dis-
2.4 Applying ecology to explain the human gut microbiome.

ease). Long-term, high-magnitude disturbances can pass the threshold of resilience to a point at which the ecosystem can never recover, allowing a different biota to take over. Because most ecosystems are stable, their level of resilience is adequate to withstand some level of perturbation and restore original composition. Therefore, ecosystems are likely to display various stages in which species with higher capacities for adaptation to change (generalists) survive and specialists die.
3.1 Article 1: “A human gut microbial gene catalogue established by metagenomic sequencing”.

In 2010, few studies on the sequencing of metagenomic samples using NGS had been published. The studies that were available answered the question, ‘who is there?’ and showed that the number of species in the human gut microbiome is higher than that in other environments. Unfortunately, the low number of analysed samples compromised more complex analyses, which might have revealed the role of the microbiota in disease. Interestingly, every subject had a unique microbiome at a low taxonomy level, demonstrating the complexity of the human gut microbiome. Creating a reference genome is an important step in understanding the differences and functional capacities of the gut microbiome. Because most of the published articles used 16S rRNA, a human gut microbiome ‘genome’ (gene catalogue) was still far from available.

Many species of bacteria in the human gut cannot be cultivated, so the creation of such a catalogue of the microbiome required metagenomic sequencing. Hence, the MetaHIT consortium undertook the sequencing of a large number of metagenomic samples from healthy individuals and those with obesity and IBD. The goal was to identify all of the genes that can exist and determine how many genes are present in any human gut microbiome, which is considered a second genome. The publication of this article was a major breakthrough. It was the largest dataset sequenced in a metagenomic project and demonstrated how NGS is presenting opportunities to study complex systems and diseases. At that time, a discussion was ongoing in the scientific community about whether short reads could be assembled into long contigs. This article demonstrates that short reads can generate contigs that are
long enough for the extraction of biological information — e.g. genes. A substantial portion of the article was dedicated to demonstrating that short reads could generate the same information as 454 and Sanger sequencing but with greater depth. With its high number of samples and great sequencing depth, this project was the first to produce sequence data adequate to saturate the number of unique genes, which was demonstrated using rarefaction curves. As a result, this project assembled large contigs, identified almost all of the possible genes in the human gut microbiome, and generated a human gut microbiome catalogue (a reference genome) with 3.3 million genes. Although the main aim of the study was to create a microbiome catalogue and prove that metagenomic sequencing could be carried out using short reads, some biological analyses for genomic and metagenomic data were also performed. As part of my work in MetaHIT, a network analysis was carried out to evaluate the data from a systems biology perspective. Using a technique similar to that of co-expressed genes, often used in transcription analysis to cluster genes that are co-expressed, we clustered species with co-abundance throughout the samples. We also applied the same principal to genes, clustering genes with co-abundance to find unknown species (metagenomic species; publication not included in this thesis). The main goal was to try to understand how bacteria interact with one another in such a complex environment. Because this type of analysis had never been performed on such a large and complex dataset, new tools were developed to undertake the analysis. My work included defining species abundance using gene abundance and generating taxonomy for each gene. Using the gene count for each sample and the taxonomy for each gene, I calculated species abundance for each organism. The species abundance correlation was then observed to define how species are connected (Figure 3.1). The network provides a first look at how species may be interacting in the human gut and revealed some unexpected results. The network shows that Bacteroides species have low numbers of species degrees (number of connections with other species) and Firmicutes species have high number of species degrees. Moreover, Bacteroides species have higher median abundance then that of most of the Firmicutes species. The result of this study had a significant impact on the scientific community. Microbiologists were particularly interested in the network because for the first time, they gained access to information about which bacteria are connected to one another. For example, one important result was the anti-correlated interaction of *F. prausnitzii* with two *Clostridium* species, which might have implications for human health because *F. prausnitzii* is an anti-inflammatory commensal bacterium that may play an important role in IBD.
### Figure 3.1: First human gut microbiome network revealing how species are connected and how the two major groups (Firmicutes and Bacteroidetes) are distributed. Figure from [66] as supplementary information figure 9.
A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin1*, Ruiqiang Li1*, Jeroen Raes2,3, Manimozhiyan Arumugam2, Kristoffer Solvsten Burgdorf4, Chaysavanh Manichanh5, Trine Nielsen6, Nicolas Pons6, Florence Levenez6, Takuji Yamada7, Daniel R. Mende7, Junhua Li1,2, Junming Xu1, Shaochuan Li1, Dongfang Li1,8, Jianjun Cao1, Bo Wang1, Huiqing Liang1, Huisong Zheng1, Yinlong Xie1,7, Julien Tap9, Patricia Lepage6, Marcelo Bertalan9, Jean-Michel Batto6, Torben Hansen4, Denis Les Paslier10, Allan Linneberg11, H. Bjørn Nielsen9, Eric Pelletier10, Pierre Renault6, Thomas Sicheritz-Ponten9, Keith Turner12, Hongmei Zhu1, Chang Yu1, Shengting Li1, Min Jian1, Yan Zhou1, Yingrui Li1, Xiaojing Zhang1, Songgang Li1, Nan Qin1, Huanming Yang1, Jian Wang3, Søren Brunak9, Joel Dore10, Francisco Guarner6, Karsten Kristiansen13, Oluf Pedersen6,14, Julian Parkhill12, Jean Weissenbach10, MetaHIT Consortium10, Peer Bork2, S. Dusko Ehrlich6 & Jun Wang1,13

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, tenfold the number of human cells, and suggested that they encode 100-fold more unique genes than our own genome. The majority of microbes reside in the gut, have a profound influence on human physiology and nutrition, and are crucial for human life. Furthermore, the gut microbes contribute to energy harvest from food, and changes of gut microbiome may be associated with bowel diseases or obesity.

To understand and exploit the impact of the gut microbes on human health and well-being it is necessary to decipher the content, diversity and functioning of the microbial gut community. 16S ribosomal RNA gene (rRNA) sequence-based methods revealed that two bacterial divisions, the Bacteroidetes and the Firmicutes, constitute over 90% of the known phylogenetic categories and dominate the distal gut microbiota. Studies also showed substantial diversity of the gut microbiome between healthy individuals and individuals from the United States or Japan. To get a broader overview of the human gut microbial genes we used the Illumina Genome Analyser (GA) technology to carry out deep sequencing of total DNA from faecal samples of 124 European adults. We generated 576.7 Gb of sequence, almost 200 times more than in all previous studies, assembled it into contigs and predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a broad view of the functions important for bacterial life in the gut and indicates that many bacterial species are shared by different individuals. Our results also show that short-read metagenomic sequencing can be used for global characterization of the genetic potential of ecologically complex environments.

Metagenomic sequencing of gut microbiomes

As part of the MetaHIT (Metagenomics of the Human Intestinal Tract) project, we collected faecal specimens from 124 healthy, overweight and obese individual human adults, as well as inflammatory bowel disease (IBD) patients, from Denmark and Spain (Supplementary Table 1). Total DNA was extracted from the faecal specimens and an average of 4.5 Gb (ranging between 2 and 7.3 Gb) of sequence was generated for each sample, allowing us to capture most of the
novelty (see Methods and Supplementary Table 2). In total, we obtained 576.7 Gb of sequence (Supplementary Table 3).

Wanting to generate an extensive catalogue of microbial genes from the human gut, we first assembled the short Illumina reads into longer contigs, which could then be analysed and annotated by standard methods. Using SOAPdenovo, a de Bruijn graph–based tool specially designed for assembling very short reads, we performed de novo assembly for all of the Illumina GA sequence data. Because a high diversity between individuals is expected\(^5,16,17\), we first assembled each sample independently (Supplementary Fig. 3). As much as 42.7% of the Illumina GA reads was assembled into a total of 6.58 million contigs of a length >500 bp, giving a total contig length of 10.3 Gb, with an N50 length of 2.2 kb (Supplementary Fig. 4) and the range of 12.3 to 237.6 Mb (Supplementary Table 4). Almost 35% of reads from any one sample could be mapped to contigs from other samples, indicating the existence of a common sequence core.

To assess the quality of the Illumina GA-based assembly we mapped the contigs of samples MH0006 and MH0012 to the Sanger reads from the same samples (Supplementary Table 2). A total of 98.7% of the contigs that map to at least one Sanger read were collinear over 99.6% of the mapped regions. This is comparable to the contigs that were generated by 454 sequencing for one of the two samples (MH0006) as a control, of which 97.9% were collinear over 99.5% of the mapped regions. We estimate assembly errors to be 14.2 and 20.7 per megabase (Mb) of Illumina- and 454-based contigs, respectively (see Methods and Supplementary Fig. 5), indicating that the short- and long-read-based assemblies have comparable accuracies.

To complete the contig set we pooled the unassembled reads from all 124 samples, and repeated the de novo assembly process. About 0.4 million additional contigs were thus generated, having a length of 370 Mb and an N50 length of 939 bp. The total length of our final contig set was thus 10.7 Gb. Some 80% of the 576.7 Gb of Illumina GA sequence could be aligned to the contigs at a threshold of 90% identity, allowing for accommodation of sequencing errors and strain variability in the gut (Fig. 1), almost twice the 42.7% of sequence that was assembled into contigs by SOAPdenovo, because assembly uses more stringent criteria. This indicates that a vast majority of the Illumina sequence is represented by our contigs.

To compare the representation of the human gut microbiome in our contigs with that from previous work, we aligned them to the reads from the two largest published gut metagenome studies (1.83 Gb of Roche/454 sequencing reads from 18 US adults\(^8\), and 0.79 Gb of Sanger reads from 13 Japanese adults and infants\(^17\)), using the 90% identity threshold. A total of 70.1% and 85.9% of the reads from the Japanese and US samples, respectively, could be aligned to our contigs (Fig. 1), showing that the contigs include a high fraction of sequences from previous studies. In contrast, 85.7% and 69.5% of our contigs were not covered by the reads from the Japanese and US samples, respectively, highlighting the novelty we captured.

Only 31.0–48.8% of the reads from the two previous studies and the present study could be aligned to 194 public human gut bacterial genomes (Supplementary Table 5), and 7.6–21.2% to the bacterial genomes deposited in GenBank (Fig. 1). This indicates that the reference gene set obtained by sequencing genomes of isolated bacterial strains is still of a limited scale.

### A gene catalogue of the human gut microbiome

To establish a non-redundant human gut microbiome gene set we first used the MetaGene\(^20\) program to predict ORFs in our contigs and found 14,048,045 ORFs longer than 100 bp (Supplementary Table 6). They occupied 86.7% of the contigs, comparable to the value found for fully sequenced genomes (~86%). Two-thirds of the ORFs appeared incomplete, possibly due to the size of our contigs (N50 of 2.2 kb). We next removed the redundant ORFs, by pair-wise comparison, using a very stringent criterion of 95% identity over 90% of the shorter ORF length, which can fuse orthologues but avoids inflation of the data set due to possible sequencing errors (see Methods). Yet, the final non-redundant gene set contained as many as 3,299,822 ORFs with an average length of 704 bp (Supplementary Table 7).

We term the genes of the non-redundant set ‘prevalent genes’, as they are encoded on contigs assembled from the most abundant reads (see Methods). The minimal relative abundance of the prevalent genes was \(\sim 6 \times 10^{-7}\), as estimated from the minimum sequence coverage of the unique genes (close to 3), and the total Illumina sequence length generated for each individual (on average, 4.5 Gb), assuming the average gene length of 0.85 kb (that is, \(3 \times 0.85 \times 10^7 / 4.5 \times 10^9\)).

We mapped the 3.3 million gut ORFs to the 319,812 genes (target genes) of the 89 frequent reference microbial genomes in the human gut. At a 90% identity threshold, 80% of the target genes had at least 80% of their length covered by a single gut ORF (Fig. 2b). This indicates that the gene set includes most of the known human gut bacterial genes.

We examined the number of prevalent genes identified across all individuals as a function of the extent of sequencing, demanding at least two supporting reads for a gene call (Fig. 2a). The incidence-based coverage richness estimator (ICE), determined at 100 individuals (the highest number the EstimateS\(^21\) program could accommodate), indicates that our catalogue captures 85.3% of the prevalent genes. Although this is probably an underestimate, it nevertheless indicates that the catalogue contains an overwhelming majority of the prevalent genes of the cohort.

Each individual carried 536,112 ± 12,167 (mean ± s.e.m.) prevalent genes (Supplementary Fig. 6b), indicating that most of the 3.3 million gene pool must be shared. However, most of the prevalent genes were found in only a few individuals: 2,375,655 were present in less than 20%, whereas 294,110 were found in at least 50% of individuals (we term these ‘common’ genes). These values depend on the sampling depth; sequencing of MH0006 and MH0012 revealed more of the catalogue genes, present at a low abundance (Supplementary Fig. 7). Nevertheless, even at our routine sampling depth, each individual harboured 204,056 ± 3,603 (mean ± s.e.m.) common genes, indicating that about 38% of an individual’s total gene pool is shared. Interestingly, the IBD patients harboured, on average, 25% fewer genes than the individuals not suffering from IBD (Supplementary Fig. 8), consistent with the observation that the former have lower bacterial diversity than the latter\(^22\).

### Common bacterial core

Deep metagenomic sequencing provides the opportunity to explore the existence of a common set of microbial species (common core) in...
the cohort. For this purpose, we used a non-redundant set of 650 sequenced bacterial and archaeal genomes (see Methods). We aligned the Illumina GA reads of each human gut microbial sample onto the integrated database by using EstimateS21 (version 8.2.0) on randomly chosen 100 samples (due to memory limitation). Coverage of genes from 89 frequent gut bacterial genome corresponds to an average length of about 40 kb, some 25-fold more than that of the 16S gene generally used for microbial identification. To visualize functions encoded by the prevalent gene set, we classified the predicted genes by aligning them to the integrated KEGG (Kyoto Encyclopedia of Genes and Genomes)24 pathways, COG (Clusters of Orthologous Groups)25 and eggNOG26 data-bases. There were 77.1% genes classified into phylogenetic, 57.5% to eggNOG clusters, 47.0% to KEGG orthology and 18.7% genes assigned to KEGG pathways, respectively (Supplementary Table 9).

Functions encoded by the prevalent gene set

We classified the predicted genes by aligning them to the integrated NCBI-NR database of non-redundant protein sequences, the genes in the KEGG (Kyoto Encyclopedia of Genes and Genomes)24 pathways, and COG (Clusters of Orthologous Groups)25 and eggNOG26 databases. The variability of abundance of microbial species in individuals can greatly affect identification of the common core. To visualize this variability, we compared the number of sequencing reads aligned to different genomes across the individuals of our cohort. Even for the most common 57 species present in at least one individual at ≥1% coverage Supplementary Fig. 9), Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota22. The principal component analysis, based on the same 155 species, clearly separates patients from one individual at ≥1% coverage Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and Dorea/Eubacterium/Ruminococcus groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.
Almost all (99.96%) of the phylogenetically assigned genes belonged to the Bacteria and Archaea, reflecting their predominance in the gut. Genes that were not mapped to orthologous groups were clustered into gene families (see Methods). To investigate the functional content of the prevalent gene set we computed the total number of orthologous groups and/or gene families present in any combination of $n$ individuals (with $n = 2–124$; see Fig. 2c). This rarefaction analysis shows that the ‘known’ functions (annotated in eggNOG or KEGG) quickly saturate (a value of 5,569 groups was observed): when sampling any subset of 50 individuals, most have been detected. However, three-quarters of the prevalent gut functionalities consists of uncharacterized orthologous groups and/or completely novel gene families (Fig. 2c). When including these groups, the rarefaction curve only starts to plateau at the very end, at a much higher level (19,338 groups were detected), confirming that the extensive sampling of a large number of individuals was necessary to capture this considerable amount of novel/unknown functionality.

**Bacterial functions important for life in the gut**

The extensive non-redundant catalogue of the bacterial genes from the human intestinal tract provides an opportunity to identify bacterial functions important for life in this environment. There are functionalities that are found in each of the 124 individuals (Fig. 5). The minimal gut metagenome includes a considerable fraction of functions (~45%) that are present in <10% of the sequenced bacterial genomes (Fig. 6c, inset). These otherwise rare functionalities that are found in each of the 124 individuals may be necessary for the gut ecosystem. Eighty per cent of these orthologous groups contain genes with at best poorly characterized function, underscoring our limited knowledge of gut functioning.

Of the known fraction, about 5% codes for (pro)phage-related proteins, implying a universal presence and possible important ecological role of bacteriophages in gut homeostasis. The most striking secondary metabolism that seems crucial for the minimal metagenome relates, not unexpectedly, to biodegradation of complex sugars present in fruits and vegetables, but which are not or poorly absorbed by humans. As some gut microorganisms were found to degrade both of them29,30, this capacity seems to be selected for by the gut ecosystem as a non-competitive source of energy. Besides these, capacity to ferment, for example, mannose, fructose, cellulose and sucrose is also part of the minimal metagenome. Together, these emphasize the
Figure 6 | Characterization of the minimal gut genome and metagenome. a, Projection of the minimal gut genome on the KEGG pathways using the iPath tool38. b, Functional composition of the minimal gut genome and metagenome. Rare and frequent refer to the presence in sequenced eggNOG genomes. c, Estimation of the minimal gut metagenome size. Known orthologous groups (red), known plus unknown orthologous groups (blue) and orthologous groups plus novel gene families (> 20 proteins; grey) are shown (see Fig. 2c for definition of box and whisker plot). The inset shows strong dependence of the gut ecosystem on complex sugar degradation for its functioning.

Functional complementarities of the genome and metagenome

Detailed analysis of the complementarities between the gut metagenome and the human genome is beyond the scope of the present work. To provide an overview, we considered two factors: conservation of the functions in the minimal metagenome and presence/absence of functions in one or the other (Supplementary Table 11). Gut bacteria use mostly fermentation to generate energy, converting sugars, in part, to short-chain fatty acid, that are used by the host as energy source. Acetate is important for muscle, heart and brain cells31, propionate is used in host hepatic neoglucogenic processes, whereas, in addition, butyrate is important for enterocytes32. Beyond short-chain fatty acid, a number of amino acids are indispensable to humans33 and can be provided by bacteria34. Similarly, bacteria can contribute certain vitamins35 (for example, biotin, phyloquinone) to the host. All of the steps of biosynthesis of these molecules are encoded by the minimal metagenome.

Gut bacteria seem to be able to degrade numerous xenobiotics, including non-modified and halogenated aromatic compounds (Supplementary Table 11), even if the steps of most pathways are not part of the minimal metagenome and are found in a fraction of individuals only. A particularly interesting example is that of benzoate, which is a common food supplement, known as E211. Its degradation by the coenzyme-A ligation pathway, encoded in the minimal metagenome, leads to pimeloyl-coenzyme-A, which is a precursor of biotin, indicating that this food supplement can have a potentially beneficial role for human health.
Discussion

We have used extensive Illumina GA short-read-based sequencing of total faecal DNA from a cohort of 124 individuals of European (Nordic and Mediterranean) origin to establish a catalogue of non-redundant human intestinal microbial genes. The catalogue contains 3.3 million microbial genes, 150-fold more than the human genome complement, and includes an overwhelming majority (>86%) of prevalent genes harbourd by our cohort. The catalogue probably contains a large majority of prevalent intestinal microbial genes in the human population, for the following reasons: (1) over 70% of the metagenomic reads from three previous studies, including American and Japanese individuals, can be mapped on our contigs; (2) about 80% of the microbial genes from 89 frequent gut reference genomes are present in our set. This result represents a proof of principle that short-read sequencing can be used to characterize complex microbiomes.

The full bacterial gene complement of each individual was not sampled in our work. Nevertheless, we have detected some 536,000 prevalent unique genes in each, out of the total of 3.3 million carried by our cohort. Inevitably, the individuals largely share the genes of the common pool. At the present depth of sequencing, we found that almost 40% of the genes from each individual are shared with at least half of the individuals of the cohort. Future studies of world-wide span, envisaged within the International Human Microbiome Consortium, will complete, as necessary, our gene catalogue and establish boundaries to the proportion of shared genes.

Essentially all (99.1%) of the genes of our catalogue are of bacterial origin, the remainder being mostly archaean, with only 0.1% of eukaryotic and viral origins. The gene catalogue is therefore equivalent to that of some 1,000 bacterial species with an average-sized genome, encoding about 3,364 non-redundant genes. We estimate that no more than 15% of prevalent genes of our cohort may be missing from the catalogue, and suggest that the cohort harbours no more than ~1,150 bacterial species abundant enough to be detected by our sampling. Given the large overlap between microbial sequences in this and previous studies we suggest that the number of abundant intestinal bacterial species may be not much higher than that observed in our cohort. Each individual of our cohort harbours at least 160 such bacterial species, as estimated by the average prevalent gene number, and many must thus be shared.

We assigned about 12% of the reference set genes (404,000) to the 194 sequenced intestinal bacterial genomes, and can thus associate them with bacterial species. Sequencing of at least 1,000 human-associated bacterial genomes is foreseen within the International Human Microbiome Consortium, via the Human Microbiome Project and MetaHIT. This is commensurate with the number of dominant species in our cohort and expected more broadly in human gut, and should enable a much more extensive gene to species assignment. Nevertheless, we used the presently available sequenced genomes to explore further the concept of largely shared species among our cohort and identified 75 species common to >50% of individuals and 57 species common to >90%. These numbers are likely to increase with the number of sequenced reference strains and a deeper sampling. Indeed, a 2–3-fold increase in sequencing depth raised by 25% the number of species that we could detect as shared between two individuals. A large number of shared species supports the view that the prevalent human microbiome is of a finite and not overly large size.

How can this view be reconciled with that of a considerable inter-personal diversity of innumerable bacterial species in the gut, arising from most previous studies using the 16S rRNA marker gene? Possibly the depth of sampling of these studies was insufficient to reveal common species when present at low abundance, and emphasized the difference in the composition of a relatively few dominant species. We found a very high variability of abundance (12- to 2,200-fold) for the 57 most common species across the individuals of our cohort. Nevertheless, a recent 16S rRNA-based study concluded that a common bacterial species ‘core’, shared among at least 50% of individuals under study, exists.

Detailed comparisons of bacterial genes across the individuals of our cohort will be carried out in the future, within the context of the ongoing MetaHIT clinical studies of which they are part. Nevertheless, clustering of the genes in families allowed us to capture a virtually full functional potential of the prevalent gene set and revealed a considerable novelty, extending the functional categories by some 30% in regard to previous work. Similarly, this analysis has revealed a functional core, conserved in each individual of the cohort, which reflects the full minimal human gut metagenome, encoded across many species and probably required for the proper functioning of the gut ecosystem. The size of this minimal metagenome exceeds several-fold that of the core metagenome reported previously. It includes functions known to be important to the host–bacterial interaction, such as degradation of complex polysaccharides, synthesis of short-chain fatty acids, indispensable amino acids and vitamins. Finally, we also identified functions that we attribute to a minimal gut bacterial genome, likely to be required by any bacterium to thrive in this ecosystem. Besides general housekeeping functions, the minimal genome encompasses many genes of unknown function, rare in sequenced genomes and possibly specifically required in the gut.

Beyond providing the global view of the human gut microbiome, the extensive gene catalogue we have established enables future studies of association of the microbial genes with human phenotypes and, even more broadly, human living habits, taking into account the environment, including diet, from birth to old age. We anticipate that these studies will lead to a much more complete understanding of human biology than the one we presently have.

METHODS SUMMARY

Human faecal samples were collected, frozen immediately and DNA was purified by standard methods. For all 124 individuals, paired-end libraries were constructed with different clone insert sizes and subjected to Illumina GA sequencing. All reads were assembled using SOAPdenovo, with specific parameter ‘-M 3’ for metagenomics data. MetaGene was used for gene prediction. A non-redundant gene set was constructed by pair-wise comparison of all genes, using BLAT under the criteria of identity >95% and overlap >90%. Gene taxonomic assignments were made on the basis of BLAST search (e-value <1 x 10^-3) of the NCBI-NR database and 126 known gut bacteria genomes. Gene functional annotations were made by BLASTP search (e-value <1 x 10^-10) with eggNOG and KEGG (v48.2) databases. The total and shared number of orthologous groups and/or gene families were computed using a random combination of n individuals (with n = 2 to 124, 100 replicates per bin).

Received 14 August; accepted 23 December 2009.

ARTICLES

19. Li, R. & Zhu, H.
20. Kent, W. J. BLAT—the BLAST-like alignment tool.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to the faculty and staff of Beijing Genomics Institute at Shenzhen, whose names were not included in the author list, but who contributed to large-scale sequencing of this team work. The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013): MetaHIT, grant agreement 201424–201702. The Old Farmer grant from the Danish National Science Research Council, the Solexa project (227–07–0196), the Shenzhen Municipal Government of China, the National Natural Science Foundation of China (30725008), the International Science and Technology Cooperation Project (No. 2008DFA02020), the Chinese Ministry of Science and Technology and the Chinese Academy of Sciences. The Danish Strategic Research Council grant No 2106–07–0021 (Seqnet), and the Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention and Care. Ciberehd is funded by Instituto de Salud Carlos III (Spain). We also thank W. Xiang from the School of Biosciences and Bioengineering, South China University of Technology, for his coordination on the Innovative Program for Undergraduate Students in which J.L. and Y.X. joined.


Author Information The raw illumina read data of all 124 samples has been deposited in the EBI, under the accession ERA000116. The contigs and gene set are available to download from the EMBL (http://www.bork.embl.de/~arumugan/Gin_et_al_2010/) and BGI (http://gutmeta.genomics.org.cn) websites. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share-Alike license, and is freely available to all readers at www.nature.com/supplement. Correspondence and requests for materials should be addressed to Ju.W. (wangj@genomics.org.cn) or S.D.E. (dusko.ehrlich@jouy.inra.fr).

MetaHIT Consortium (additional members)

Maria Antolin1, Francois Artiguenave2, Hervé Blottiere3, Natalia Borouei3, Thomas Bruls1, Francesc Casellas3, Christian Cherveaux4, Antonella Cultore5, Christine Delorme6, Gérard Denariaz7, Rozenn Dervyn8, Miguel Forte5, Carsten Friss9, Maarten van de Gucht10, Eric Guendon11, Florence Haimet12, Alexandre Jamet13, Catherine Juste14, Ghalla Kaci15, Michel Klerebezem16, Jan Knol17, Michel Kristensen18, Severine Layez1, Karine Le Roux19, Marion Leclerc2, Emmanuelle Maguin20, Raquel Melo Minardi21, Raish Oozeer2, Maria Rescigno22, Nicolas Sanchez3, Sebastian Tims7, Toni Torrejon1, Encarna Varela23, Willem de Vos24, Yohanan Winogradsky25 & Erwin Zoetendal26

1Hospital Universitari Vall de Hebron, Cibererh, 08035 Barcelona, Spain. 2Comissariat à l’Energie Atomique, Gérosce, 91000 Evry, France. 3Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. 4Danone Research, 91120 Palaiseau, France. 5UCB Pharma SA, 28046, Madrid, Spain. 6Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark. 7Wageningen University, 6710BA Ede, The Netherlands. 8Hagedorn Research Institute, DK 2820 Copenhagen, Denmark. 9Istituto Europeo di Oncologia, 20130, Milan, Italy.
METHODS

Human faecal sample collection. Danish individuals were from the Inter-99 cohort40, varying in phenotypes according to BMI and status towards obesity/diabetes, whereas Spanish individuals were either healthy controls or patients with chronic inflammatory bowel diseases (Crohn’s disease or ulcerative colitis) in clinical remission.

Patients and healthy controls were asked to provide a frozen stool sample. Fresh stool samples were obtained at home, and samples were immediately frozen by storing them in their home freezer. Frozen samples were delivered to the Hospital using insulated polystyrene foam containers, and then they were stored at −80°C until use.

DNA extraction. A frozen aliquot (200 mg) of each faecal sample was suspended in 250 μl of guanidine thiocyanate, 0.1 M Tris (pH 7.5) and 40 μl of 10% N-lauroyl sarcosine. Then, DNA extraction was conducted as previously described37. The DNA concentration and its molecular size were estimated by nanodrop (Thermo Scientific) and agarose gel electrophoresis.

DNA library construction and sequencing. DNA library preparation followed the manufacturer’s instruction (Illumina). We used the same workflow as described elsewhere to perform cluster generation, template hybridization, iso-thermal amplification, linearization, blocking and denaturation and hybridization of the sequencing primers. The base-calling pipeline (version IlluminaPipeline-0.3) was used to process the raw fluorescent images and call sequences.

We constructed one library (clone insert size 200 bp) for each of the first 15 samples, and two libraries with different clone insert sizes (135 bp and 400 bp) for each of the remaining 109 samples for validation of experimental reproducibility.

To estimate the optimal return between the generation of novel sequence and sequencing depth, we aligned the Illumina GA reads from samples MH0006 and MH0012 onto 468,335 Sanger reads totalling to 311.7 Mb generated from the same two samples (156.9 and 154.7 Mb, respectively, Supplementary Table 2), using the Short Oligonucleotide Alignment Program (SOAP)41 and a match requirement of 95% sequence identity. With about 4 Gb of Illumina sequence, 94% and 89% of the Sanger reads (for MH0006 and MH0012, respectively) were covered. Further extensive sequencing, to 12.6 and 16.6 Gb for MH0006 and MH0012, respectively, brought only a moderate increase of coverage to about 95% (Supplementary Fig. 1). More than 90% of the Sanger reads were covered by the Illumina sequences to a very high and uniform level (Supplementary Fig. 2), indicating that there is little or no bias in the Illumina GA sequence. As expected, a large proportion of Illumina sequences (57% and 74% for MH0006 and M0012, respectively) was novel and could not be mapped onto the Sanger reads. This fraction was similar at the 4 and 12–16 Gb sequencing levels, confirming that most of the novelty was captured already at 4 Gb.

We generated 35.4–97.6 million reads for the remaining 122 samples, with an average of 62.5 million reads. Sequencing read length of the first batch of 15 samples was 44 bp and the second batch was 75 bp.

Public data used. The sequenced bacteria genomes (totally 806 genomes) deposited in GenBank were downloaded from NCBI database (http://www.ncbi.nlm.nih.gov/) on 10 January 2009. The known human gut bacteria genome sequences were downloaded from HMP database (http://www.hmpdcc-resources.org/gt-bin/hmp_catalyst/manip), GenBank (67 genomes), Washington University in St Louis (85 genomes, version April 2009, http://genome.wusl.edu/pub/organism/Microbes/Human_Gut_Microbiome/), and sequenced by the MetaHT project (17 genomes, version September 2009, http://www.sanger.ac.uk/pathogens/metahit/). The other gut metagenome data used in this project include: (1) human gut metagenomic data sequenced from US individuals4, which was downloaded from NCBI with the accession SRA002775; (2) human gut metagenomic data from Japanese individuals38, which was downloaded from P. Bork’s group at EMBL (http://www. bork.embibe.de). The integrated NR database we constructed in this study included NCBI-NR database (version April 2009) and all genes from the known human gut bacteria genomes.

Illumina GA short reads de novo assembly. High-quality short reads of each DNA sample were assembled by SOAPdenovo assembler42. In brief, we filtered the Illumina reads to low-abundance sequences according to frequency. The 17-mers with depth less than 5 were screened in front of these assembly, for these low-frequency sequences were very unlikely to be assembled, whereas removing them would significantly reduce memory requirement and make assembly feasible in an ordinary supercomputer (312 GB memory in our institute).

Then the sequences were processed one by one and the de Bruijn graph data format was used to store the overlap information among the sequences. The overlap paths supported by a single read were unrelined and removed. Short low-depth tips and bubbles that were caused by sequencing errors or genetic variations between microbial strains were trimmed and merged, respectively. Reads path were used to solve the tiny repetitions. Finally, we broke the connections at repeat boundaries, and outputted the continuous sequences with unambiguous connections as contigs. The metagenomic assembly model was chosen, and parameters ‘-K21’ and ‘-K23’ were used for 44 bp and 75 bp reads, respectively, to indicate the minimal sequence overlap required.

After de novo assembly for each sample independently, we merged all the unassembled reads together and performed assembly for them, as to maximize the usage of data and assemble the microbial genomes that have low frequency in each read set, but have sufficient sequence depth for assembly by putting the data of all samples together.

Validating Illumina contigs using Sanger reads. We used BLASTN (WU- BLAST 2.0) to map Sanger reads from samples MH0006 and MH0012 (156.9 Mb and 154.7 Mb, respectively) to Illumina contigs (single best hit longer than 75 bp and over 95% identity) from the same samples. Each alignment was scanned for breakage of collinearity where both sequences have at least 50 bases left unaligned at one end of the alignment. Each such breakage was considered an assembly error in the Illumina contig at the location where collinearity breaks. Errors within 30 bp from each other were merged. An error was discarded if there exists a Sanger read that agrees with the contig structure for 60 bp on both sides of the error. For comparison, we repeated this on a Newbler2 assembly of 454 Titanium reads from MH0006 (550 Mb reads). Supplementary Fig. 5a shows the number of errors per Mb of assembled Illumina/454 contigs. We estimate 14.12 errors per Mb of contigs for the Illumina assembly, which is comparable to that of the 454 assembly (20.73 per Mb). 98.7% of Illumina contigs that map at least one Sanger read were collinear over 99.55% of the mapped regions, which is comparable to 97.86% of such 454 contigs being collinear over 99.48% of the mapped regions.

Evaluation of human gut microbiome coverage. The Illumina GA reads were aligned against the assembled contigs and known bacteria genomes using SOAP41 by allowing at most two mismatches in the first 35-bp region and 90% identity over the read sequence. The Roche/454 and Sanger sequencing reads were aligned against the same reference using BLASTN with 1 × 10−2, over 100 bp alignment length and minimal 90% identity cutoff. Two mismatches were allowed and identity was set 95% over the read sequence when aligned to the GA reads of MH0006 and MH0012 to Sanger reads from the same samples by SOAP41.

Gene prediction and construction of the non-redundant gene set. We used MetaGene29—which uses di- and codon frequencies estimated by the GC content of a given sequence, and predicts a whole range of ORFs based on the anonymous genomic sequences—to find ORFs from the contigs of each of the 124 samples as well as the contigs from the merged assembly.

The predicted ORFs were then aligned to each other using BLAT30. A pair of genes with greater than 95% identity and aligned length covered over 90% of the shorter gene was grouped together. The groups sharing genes were then merged, and the longest ORF in each merged group was used to represent the group, and the other members of the group were taken as redundancy. Therefore, we organized the non-redundant gene set from all the predicted genes by excluding the redundancy. Finally, the ORFs with length less than 100 bp were filtered. We translated the ORFs into protein sequences using the NCBI Genetic Code 11.

Identification of genes. To make a balance between identifying low-abundance genes and reducing the error-rate of identification, we explored the impact of the threshold set for read coverage required to identify a gene in individual microorganisms. The number of genes decreased about twice when the number of reads required for identification was increased from 2 to 6, and changed slowly thereafter (Supplementary Fig. 6a). Nevertheless, to include the rare genes into the analysis, we selected the threshold of 2 reads.

Gene taxonomic assignment. Taxonomic assignment of predicted genes was carried out using BLASTP alignment against the integrated NR database. BLASTP alignment hits with e-values larger than 1 × 10−3 were filtered, and for each gene the significant matches which were defined by e-values ≤10−3 × e-value of the top hit were retained to distinguish taxonomic groups. Then we determined the taxonomical level of each gene by the lowest common ancestor (LCA)-based algorithm that was implemented in MEGAN2. The LCA-based algorithm assigns genes to taxa in the way that the taxonomical level of the assigned taxon reflects the level of conservation of the gene. For example, if a gene was conserved in many species, it was assigned to the LCA rather than to a species.

Gene functional classification. We used BLASTP to search the protein sequences of the predicted genes in the eggNOG database23 and KEGG database24 with e-values≤1 × 10−3. The genes were annotated as the function of the NOG or KEGG homologues with lowest e-value. The eggNOG database is an integ-
**Determination of minimal gut bacterial genome.** The number of non-redundant genes assigned to the eggNOG clusters was normalized by gene length and cluster copy number (Supplementary Fig. 8). The clusters were ranked by normalized gene number and the range that included the clusters encoding essential *Bacillus subtilis* genes was determined, computing the proportion of these clusters among the successive groups of 100 clusters. Analysis of the range gene clusters involved, besides iPath projections, use of KEGG and manual verification of the completeness of the pathways and protein machineries they encode.

**Determination of total functional complement and minimal metagenome.** We computed the total and shared number of orthologous groups and/or gene families present in random combinations of *n* individuals (with *n* = 2 to 124, 100 replicates per bin). This analysis was performed on three groups of gene clusters: (1) known eggNOG orthologous groups (that is, those with functional annotation, excluding those in which the terms [Uu]ncharacterized, [Uu]nknown, [Pp]redicted or[Pp]utative occurred); (2) all eggNOG orthologous groups; (3) all orthologous groups plus gene families constructed from remaining genes not assigned to the two above categories. Families were clustered from all-against-all BLASTP results using MCL with an inflation factor of 1.1 and a bit-score cutoff of 60.

**Rarefaction analysis.** Estimation of total gene richness was done using EstimateS on 100 randomly picked samples due to memory limitations. Because the CV value was >0.5, both chai2 (classic) and ICE richness estimators were calculated and the larger estimate of the two (ICE) was used. The estimate for this sample size was 3,621,646 genes (ICE) whereas *S*\_obs (Mao Tau) was 3,090,575 genes, or 85.3%. The ICE estimator curve did not completely saturate, (data not shown) indicating that additional samples will need to be added to achieve a final, conclusive estimate.

**Common bacterial core.** To eliminate the influence of very similar strains and assess the presence of known microbial species among the individuals of the cohort, we used 650 sequenced bacterial and archaean genomes as a reference set.

The set was composed from 932 publicly available genomes, which were grouped by similarity, using a 90% identity cutoff and the similarity over at least 80% of the length. From each group only the largest genome was used. Illumina reads from 124 individuals were mapped to the set, for species profiling analysis and the genomes originating from the same species (by differing in size >20%) curated by manual inspection and by using the 16S-based clustering when the sequences were available.

**Relative abundance of microbial genomes among individuals.** We computed the genome coverage by uniquely mapping Illumina reads and normalized it to 1 Gb of sequence, to correct for different sequencing levels in different individuals. The coverage was summed over all species of the non-redundant bacterial genome set for each individual and the proportion of each species relative to the sum calculated.

**Species co-existence network.** For the 155 species that had genome coverage by the Illumina reads ≥1% in at least one individual we calculated the pair-wise inter-species Pearson correlations between sequencing depths (abundance) throughout the entire cohort of 124 individuals. From the resulting 11,175 inter-species correlations, correlations less than −0.4 or above 0.4 (*n* = 342) were visualized in a graph using Cytoscape displaying the average genome coverage of each species as node size in the graph.

The Human Microbiome Project was launched in 2008 with the goal of identifying and characterising microorganisms associated with disease. The importance of the human gut microbiome to the host has been demonstrated by many studies describing the influence of the microbiome on gut development, the immune system, and even behaviour. Moreover, the transplant of microbiota from healthy individuals to IBD patients has led to unexpected positive results, thus supporting the importance of the gut microbiome for human health and indicating a possible division of so-called ‘good’ and ‘bad’ bacteria. The identification of bad bacteria could reveal the aetiology of many gut-related diseases, leading to the development of new drugs. However, the search for the ‘healthy microbiome’ is difficult, because many factors, including lifestyle, genotype, age, physiology, pathobiology, the environment, and immunity, influence the gut microbiome. Thus, we expected that individuals with the same genotype would be more likely than unrelated individuals would to have a similar gut microbiome. Moreover, the microbiome of individuals from the same country would likely be more similar than that between individuals from different countries owing to similarities in lifestyle and diet. Indeed, diet was expected to have a large impact on the human gut microbiome. However, because these factors were expectations based primarily on intuition rather than on specific knowledge of the relationships, researchers attempted to demonstrate which factors were most important in determining gut microbiome composition. Surprisingly, none of these factors was found to exert a strong enough influence to explain the variation of the human gut microbiome and its link to disease. Analysis of patients was considered as a better strategy for determining what constitutes healthy and unhealthy microbiomes. Therefore, the new goal of researchers was to find one or a few bacteria that could be the cause of disease. Previous articles had shown the importance of two phyla, Bacteroidetes and Firmicutes, of which Bacteroidetes has the highest variance in relative abundance. Moreover, the B/F ratio may be important for the gut microbiome, as 90% of the relative abundance of all samples is composed of Bacteroidetes and Firmicutes. Thus, while many groups were testing various factors of influence or looking for bacterial markers to explain variations in the gut microbiome, the MetaHIT consortium was doing the opposite. Peer Bork and co-workers, for example, have conducted studies to identify the main driving force behind the gut microbiome by analysing variance in the data. The publication of ‘Enterotypes of the human gut microbiome’ had a major impact on the field. For the first time, factors previously identified as influential in gut composition were demonstrated to have no effect on human gut microbiome variation. Factors such as body mass index, age, gender, nation, or continent could not explain the three robust clusters found in the data. Although the article offered no biological explanation for the cause of the data separation, it had a significant impact on the scientific view of the human gut microbiome. The article redefined the direction of research in the field, as the main factor in gut composition and variation, although of unknown cause, was now
clear and present in any microbiome dataset. The three clusters (Figure 3.2), called enterotypes, were confirmed by microarray results, showing that the results are consistent between technologies and reproducible. My work in this study was to develop a microarray chip, which could profile any human gut microbiome. Using the gene catalogue described in article one, I designed a microarray chip using 700,000 genes — the first microarray capable of profiling genes from the human gut microbiome. The 700,000 genes can represent 60% (Gut-Array V.4) to 80% (Gut-Array V.5) of the sequencing data from more than 600 samples, and the microarray has the versatility to profile DNA and RNA samples, overcoming the rRNA challenges in RNA sample sequencing. Using Gut-Array V.4, we profiled the samples used in the published paper, and I performed the analysis to find the clusters using the same methodology described in the paper. To perform the analysis, I developed a taxonomy pipeline to profile the high number of genes. This pipeline was used in the current article and in the third article, below. Moreover, to find the genus abundance used in the enterotypes study, I developed a package in ‘R’ to work with metagenomic data. To find genus abundance, I used the same methodology as that applied to find the clusters using the microarray samples. The results confirmed the three robust clusters dominated by Bacteroides, Prevotella, and Ruminococcus found in the sequencing data by Peer Bork and co-workers.

Figure 3.2: Abundance of the three main contributors for each enterotype. Figure from [58] as figure 2d. This figure is important as it shows not only the three main groups found in the human gut microbiome distribution, but also demonstrate the difference in diversity among the three groups. Prevotella group has the most uneven distribution. Ruminococcus with the most even distribution and the lowest relative abundance, suggesting high richness. Bacteroides average level of evenness but overall high relative abundance for the three groups, suggesting low richness.
Enterotypes of the human gut microbiome

Manimozhiyan Arumugam1*, Jeroen Raes1,2*, Eric Pelletier3,4,5, Denis Le Paslier3,4,5, Takjuji Yamada1, Daniel R. Mendé1, Gabriel R. Fernandes1,6, Julian Tap1,7, Thomas Bruls3,4,5, Jean-Michel Bato1, Marcelo Bertalan8, Natalia Borruel9, Francesc Casellas9, Leyden Fernandez9, Laurent Gautier9, Torben Hansen11,12, Masahira Hattori13, Tetsuya Hayashi14, Michiel Kleerebezem15, Ken Kurokawa16, Marion Leclerc17, Florence Levenez18, Chaysavanich Manichan18, H. Bjorn Nielsen19, Trine Nielsen20, Nicolas Pons21, Julie Poulain1,22, Thomas Sicheritz-Ponten23,4,5,6, Sebastian Tims23,4,6, David Torrents24,25, Gabriel R. Fernandes1,6, Julien Tap1,7, Thomas Bruls3,4,5, Jean-Michel Batto7, Marcelo Bertalan8, Natalia Borruel9, Edgardo Ugarte3, Erwin G. Zoetendal15, Jun Wang17,20, Francisco Guarner9, Oluf Pedersen11,21,22,23, Willem M. de Vos15,24, Francesc Casellas9, Leyden Fernandez10, Laurent Gautier8, Torben Hansen11,12, Masahira Hattori13, Tetsuya Hayashi14, Trine Nielsen20, Nicolas Pons21, Julie Poulain3, Junjie Qin17, Thomas Sicheritz-Ponten8,18, Sebastian Tims23,4,6, David Torrents24,25, Gabriel R. Fernandes1,6, Julien Tap1,7, Thomas Bruls3,4,5, Jean-Michel Batto7, Marcelo Bertalan8, Natalia Borruel9, Edgardo Ugarte3, Erwin G. Zoetendal15, Jun Wang17,20, Francisco Guarner9, Oluf Pedersen11,21,22,23, Willem M. de Vos15,24, Soren Brunak8, Joel Doré7, MetaHIT Consortium1, Jean Weissenbach14,5, S. Dusko Ehrlich7 & Peer Bork25.

Various studies of the human intestinal tract microbiome based on the 16S ribosomal-RNA-encoding gene reported species diversity within and between individuals1-3, and the first metagenomics studies characterized the functional repertoire of the microbiomes of several American4,5 and Japanese6 individuals. Although a general consensus about the phylum level composition in the human gut is emerging3-5,7, the variation in species composition1-2 and gene pools4,8 within the human population is less clear. Furthermore, it is unknown whether inter-individual variation manifests itself as a continuum of different community compositions or whether individual gut microbiota conglomerate around preferred, balanced and stable community compositions that can be classified. Studying such questions is complicated by the complexity of sampling, DNA preparation, processing, sequencing and analysis protocols9 as well as by varying physiological, nutritional and environmental conditions. To analyse the feasibility of comparative metagenomics of the human gut across cohorts and populations and to obtain first insights into commonalities and differences between gut microbiomes across different populations, we Sanger-sequenced 22 European metagenomes from Danish, French, Italian and Spanish individuals that were selected for diversity (Supplementary Notes section 1), and combined them with existing Sanger (13 Japanese6, 2 American7) and pyrosequencing (2 American7) gut data sets—totaling 39 individuals.

Global variation of human gut metagenomes

The vast majority of sequences in the newly sequenced 22 European samples belong to bacteria—only 0.14% of the reads could be classified as human contamination, all other eukaryotes together only comprised 0.5%, archaea 0.8% and viruses up to 5.8% (see Supplementary Notes section 2.1 for details).

To investigate the phylogenetic composition of the 39 samples from 6 nationalities, we mapped metagenomic reads, using DNA sequence homology, to 1,511 reference genomes (Supplementary Table 3) including 379 publicly available human microbiome genomes generated through the National Institutes of Health (NIH) Human Microbiome Project10 and the European MetaHIT consortium11 (Supplementary Methods section 4.1). To consistently estimate the functional composition of the samples, we annotated the predicted genes from the metagenomes using eggNOG12 orthologous groups (Supplementary Methods section 6.2). We ensured that comparative analysis using these procedures was not biased by data-set origin, sample preparation, sequencing technology and quality filtering (see Supplementary Notes section 1).
We also investigated whether the relatively low and somewhat arbitrary amounts of sequence per sample (between 53–295 Mb) bias our results: we assigned habitat information to 1,368 of the 1,511 reference genomes, distinguished between orthologous groups from gut and 'non-gut' species and conclude that our data set captures most of the functions from gut species even though functions from non-gut species accumulated with each additional sample (Fig. 1a; see Supplementary Notes section 1.3).

We then characterized the phylogenetic variation across samples at the genus and phylum levels, and functional variation at gene and functional class levels. As infants are known to have very heterogeneous, unstable and distinctive microbiota\(^6,13\), we excluded the four respective Japanese samples from the analysis. Using calibrated similarity cutoffs (Supplementary Fig. 1), on average, 52.8% of the fragments in each sample could be robustly assigned to a genus in our reference genome (Supplementary Table 5) can be assigned to orthologous groups, and orthologous group abundance patterns agree with previous observations\(^6,15\) (for example, histidine kinases make up the largest group; Fig. 1c and Supplementary Notes section 2.3).

**Abundant functions from low-abundance microbes**

Microbes in the human gut undergo selective pressure from the host as well as from microbial competitors. This typically leads to a homeostasis of the ecosystem in which some species occur in high and many in low abundance\(^14\) (the 'long-tail' effect, as seen in Fig. 1b), with some low-abundance species, like methanogens\(^15\), performing specialized functions beneficial to the host. Metagenomics enables us to study the presence of abundant functions shared by several low-abundance phylum level abundances were measured using reference-genome-based mapping with 85% and 65% sequence similarity cutoffs. Unclassified genera under a higher rank are marked by asterisks. b. Orthologous group abundance variation box plot for the 30 most abundant orthologous groups as determined by assignment to eggNOG\(^2\). Orthologous groups are coloured by their respective functional category (see inset for colour key). Inset shows abundance box plot of 24 functional categories. Boxes represent the interquartile range (IQR) between first and third quartiles and the line inside represents the median. Whiskers denote the lowest and highest values within 1.5 \(\times\) IQR from the first and third quartiles, respectively. Circles represent outliers beyond the whiskers.
species, which could shed light on their survival strategies in the human gut. In the samples analysed here, the most abundant molecular functions generally trace back to the most dominant species. However, we identified some abundant orthologous groups that are contributed to primarily by low-abundance genera (see Supplementary Fig. 2, Supplementary Table 6 and Supplementary Notes section 3). For example, low-abundance *Escherichia* contribute over 90% of two abundant proteins associated with bacterial pilus assembly. FimA (COG3539) and PapC (COG2188), found in one individual (IT-AD-5). Pili enable the microbes to colonize the epithelium of specific host organs; they help microbes to stay longer in the human intestinal tract by binding to human mucus or mannose sugars present on intestinal surface structures. They are also key components in the transfer of plasmids between bacteria through conjugation, often leading to exchange of protective functions such as antibiotic resistance. Pili can thus provide multiple benefits to these low-abundance microbes in their efforts

Figure 2 | Phylogenetic differences between enterotypes. a–c. Between-class analysis, which visualizes results from PCA and clustering, of the genus compositions of 33 Sanger metagenomes estimated by mapping the metagenome reads to 1,511 reference genome sequences using an 85% similarity threshold (a). Danish subset containing 85 metagenomes from a published Illumina data set (b) and 154 pyrosequencing-based 16S sequences (c) reveal three robust clusters that we call enterotypes. IBD, inflammatory bowel disease. Two principal components are plotted using the ade4 package in R with each sample represented by a filled circle. The centre of gravity for each cluster is marked by a rectangle and the coloured ellipse covers 67% of the samples belonging to the cluster. IBD, inflammatory bowel disease. d. Abundances of the main contributors of each enterotype from the Sanger metagenomes. See Fig. 1 for definition of box plot. e. Co-occurrence networks of the three enterotypes from the Sanger metagenomes. Unclassified genera under a higher rank are marked by asterisks in b and e.

Macmillan Publishers Limited. All rights reserved ©2011
to survive and persist in the human gut. This example illustrates that abundant species or genera cannot reveal the entire functional complexity of the gut microbiota. More reference genomes will facilitate better taxonomic assignment from samples and thus the detection of more low-abundance species. However, there is not much room for as yet undetected, abundant genera. Even with our limited genus assignment rate of 52.8% of all reads, we estimate that we miss another 30.7% of the already classified genera owing to our strict assignment criteria (Supplementary Fig. 1); that is, only 16.5% of all reads are likely to belong to hitherto unknown genera.

**Detection of enterotypes, cross-national clusters**

To get an overview of species variation we used phylogenetic profile similarities obtained by mapping metagenomic reads to the 1,511 reference genomes (Fig. 2a; see Supplementary Methods section 4.1). We excluded the top American Sanger-sequenced samples from further analysis because of an unusual, very low fraction of Bacteroidetes and suspected technical artefacts. Multidimensional cluster analysis and principal component analysis (PCA) revealed that the remaining 33 samples formed three distinct clusters that we designate as enterotypes (see Supplementary Notes section 4.1, Supplementary Fig. 3a and Supplementary Table 8). Each of these three enterotypes is identifiable by the variation in the levels of one of three genera: *Bacteroides* (enterotype 1), *Prevotella* (enterotype 2) and *Ruminococcus* (enterotype 3) (Fig. 2a, d), which was reproduced using independent array-based HiTChip data in a subset of 22 European samples (Supplementary Fig. 4 and Supplementary Notes section 4.5). The same analysis on two larger published gut microbiome data sets of different origins (16S pyrosequencing data from 154 American individuals, and Illumina-based metagenomics data from 85 Danish individuals; Supplementary Methods section 5) shows that these data sets could also be represented best by three clusters (Supplementary Fig. 3b, c and Supplementary Tables 9, 10). Two of these are also driven by *Bacteroides* and *Prevotella*, whereas the third cluster is mostly driven by related groups of the order Clostridiales, *Blautia* and unclassified Lachnospiraceae in the 16S rDNA and Illumina data, respectively (Fig. 2b, c). This can be explained by a different reference data set in the instance of the 16S rDNA data, different mapping behaviour of short reads in the case of the Illumina data or current taxonomic uncertainties in the Lachnospiraceae and Ruminococcaceae clades (see Supplementary Notes section 4.2). The differences might also hint at community subpopulations within this enterotype, which might only be detectable with substantially more samples. Correlation analysis of the Sanger data revealed that abundances of each of the three discriminating genera strongly correlate (that is, they co-occur or avoid each other) with those of other genera (Fig. 2d; see Supplementary Methods section 11), indicating that the enterotypes are in fact driven by groups of species that together contribute to the preferred community compositions.

We demonstrate further the robustness of the enterotypes using two distinct statistical concepts. First, we used the silhouette coefficient to validate that the three clusters are superior to clusterings obtained from various randomizations of the genus profile data, indicating a potential role for the interactions between co-occurring genera (see Supplementary Fig. 5 and Supplementary Notes section 4.3). Second, we used supervised learning and cross-validation to establish that these clusters have non-random characteristics that can be modelled and subsequently used to classify new samples (learning on clusters from randomized genus profiles led to considerably worse classification performance; see Supplementary Fig. 6 and Supplementary Notes section 4.4). These consistent results indicate that enterotypes will be identifiable in human gut metagenomes also from larger cohorts.

We then clustered the 33 samples using a purely functional metric: the abundance of the assigned orthologous groups (Fig. 3a). Remarkably, this clustering also showed a similar grouping of the samples with only minor differences (five samples placed in different clusters compared to Fig. 2a), indicating that function and species composition roughly coincide with some exceptions such as Spanish sample ES-AD-3, whose genus composition belongs to enterotype 2 whereas its functional composition is similar to members of enterotype 1. This individual has high levels of phage-related genes compared to the other samples (see Supplementary Fig. 7), hinting at partial temporal variability and dynamics of the microbiota, and perhaps indicating phage or virus bursts.

The robustness and predictability of the enterotypes in different cohorts and at multiple phylogenetic and functional levels indicates that they are the result of well-balanced, defined microbial community compositions of which only a limited number exist across individuals. These enterotypes are not as sharply delimited as, for example, human blood groups; they are, in contrast, densely populated areas in a multi-dimensional space of community composition. They are nevertheless likely to characterize individuals, in line with previous reports that gut

---

**Figure 3 | Functional differences between enterotypes.**

**a.** Between-class analysis (see Fig. 2) of orthologous group abundances showing only minor disagreements with enterotypes (unfilled circles indicate the differing samples). The blue cloud represents the local density estimated from the coordinates of the gut microbiota. More reference genomes will facilitate better taxonomic assignment from samples and thus the detection of as yet undetected, abundant genera. Even with our limited genus assignment rate of 52.8% of all reads, we estimate that we miss another 30.7% of the already classified genera owing to our strict assignment criteria (Supplementary Fig. 1); that is, only 16.5% of all reads are likely to belong to hitherto unknown genera.

**b.** Four enzymes in the biotin biosynthesis pathway (COG0132, COG0156, COG0161 and COG0502) are overrepresented in enterotype 1. c. Four enzymes in the thiamine biosynthesis pathway (COG0422, COG0351, COG0352 and COG0611) are overrepresented in enterotype 2. d. Six enzymes in the haem biosynthesis pathway (COG0007, COG0276, COG0407, COG0408, COG0716 and COG1648) are overrepresented in enterotype 3.
Variation between enterotypes

To determine the phylogenetic and functional basis of the enterotypes, we investigated in detail their differences in composition at abundance of co-occurring genera (Figs 2, 3; also see Supplementary Methods sections 10, 11 and 12). Enterotype 1, containing eight samples, is enriched in Bacteroides (P < 0.01; Supplementary Fig. 8), which co-occurs, for example, with Parabacteroides (see Supplementary Table 11 for enriched genera and Fig. 2e for correlation networks of co-occurring genera in each enterotype). The drivers of this enterotype seem to derive energy primarily from carbohydrates and proteins through fermentation, as these closely related genera have a very broad saccharolytic potential26 and because genes encoding enzymes involved in the degradation of these substrates (galactosidases, hexosaminidases, proteases) along with glycolysis and pentose phosphate pathways are enriched in this enterotype (see Supplementary Tables 12, 13). Enterotype 2 contains six samples and is enriched in Prevotella (P < 0.01; Supplementary Fig. 9) and the co-occurring Desulfovibrio, which can act in synergy to degrade mucin glycopolymers present in the mucosal layer of the gut: Prevotella is a known mucin-degrader and Desulfovibrio may enhance the rate-limiting mucin desulfation step by removing the sulphate27. Enterotype 3 is the most frequent and is enriched in Ruminococcus (P < 0.01; Supplementary Fig. 10) as well as co-occurring Akkermansia, both known to comprise species able to degrade mucins28. It is also enriched in membrane transporters, mostly of sugars, indicating the efficient binding of mucin and its subsequent hydrolysis as well as uptake of the resulting simple sugars by these genera. The enriched genera indicate that enterotypes use different routes to generate energy from fermentable substrates available in the colon, reminiscent of a potential specialization in ecological niches or guilds. In addition to the conversion of complex carbohydrates into absorbable substrates, the gut microbiota is also beneficial to the human host by producing vitamins. Although all the vitamin metabolism pathways are represented in all samples, enterotypes 1 and 2 were enriched in biosynthesis of different vitamins: biotin (Fig. 3b), riboflavin, pantothenate and ascorbate in the former, and thiamine (Fig. 3c) and folate in the latter. These phylogenetic and functional differences among enterotypes thus reflect different combinations of microbial trophic chains with a probable impact on synergistic interrelations with the human hosts.

Functional biomarkers for host properties

Enterotypes do not seem to differ in functional richness (Supplementary Fig. 11), and virtually none of several measured host properties, namely nationality, gender, age or body mass index (BMI), significantly correlates with the enterotypes (with the exception of enterotype 1, which is enriched in Japanese individuals). However, some strong correlations do occur between host properties and particular functions, at the genes or module level (a module is a part of a pathway that is functionally tightly interconnected; see Supplementary Methods sections 6, 13 and Supplementary Notes section 6). The only significant correlation between a host property and a taxonomic group is a negative one between age and the abundance of an unknown Clostridiales genus (P < 0.02) containing three obligate anaerobes (Supplementary Fig. 12a; see Supplementary Notes section 6.2). It should be noted that age is not constant across the nationalities (in our data set, Italians are relatively old and Japanese young), but that individuals did not stratify by nationality, indicating that this is not a confounding factor. Our data did not reveal any correlation between BMI and the Firmicutes/Bacteroidetes ratio and we thus cannot contribute to the ongoing debate on the relationship between this ratio and obesity29,30.

In contrast to the minor phylogenetic signal, we found several significant functional correlations with each of the host properties studied (after correcting for multiple testing to avoid artefacts; see Supplementary Methods section 13), indicating that metagenomic-derived functional biomarkers might be more robust than phylogenetic ones. For example, the abundance of ten orthologous groups varies more than between within nationalities (Supplementary Table 14), although overall, the functional composition in total was remarkably similar among the nations (also with respect to the functional core; see Supplementary Fig. 13). For gender, we find five functional modules and one orthologous group that significantly correlate (P < 0.05; for example, enriched aspartate biosynthesis modules in males; see Supplementary Table 16). In addition, twelve orthologous groups significantly correlate with age (Supplementary Table 17). For instance, starch degradation enzymes such as glycosidases and glucan phosphorlyases increase with age (which could be a reaction to decreased efficiency of host breakdown of dietary carbohydrates with age31) and so does the secA preprotein translocase (Supplementary Fig. 14). Conversely, an orthologous group coding for the facilitative σ54 subunit of RNA polymerase, which drives expression under various stress responses and is linked to intestinal survival32, decreases with age (Fig. 4a). One explanation for this could be the reduced need for stress response in the gut due to the age-associated decline in host immune response33 (immunosenescence). Our analyses also identified three marker modules that correlate strongly with the hosts’ BMI (Supplementary Table 19 and Supplementary Fig. 14), two of which are ATPase complexes, supporting the link found between the gut microbiota’s capacity for energy harvest and obesity in the host34. Interestingly, functional markers found by a data-driven approach (derived from the metagenomes without previous knowledge) gave much stronger correlations than genes for which a link would be expected (for example, susC/susD, involved in starch utilization26).
Fig. 4b. Linear models combining the abundance of only a few functional modules correlate even better with host properties (Fig. 4c, d). It should be noted that given the possibility of many confounding variables owing to the heterogeneity and size of our cohort, these observations will need to be substantiated using larger, independent cohorts in the future. Furthermore, patterns in metagenomics data can (partly) reflect indirect factors such as genome size (the smaller the average genome size of a sample, the higher the relative fraction of single copy genes therein), which, however, does not matter for diagnostics.

Although individual host properties do not explain the enterotypes, the latter might be driven by a complex mixture of functional properties, by host immune modulation or by hitherto unexplored physiological conditions such as transit time or pH of luminal content. Furthermore, the three major enterotypes could be triggered by the three distinct pathways for hydrogen disposal (Supplementary Notes section 6.4). Indeed, despite their low abundance, Methanobrevibacter (a methanolithoautotrophic) and Desulfovibrio (a known sulphate-reducer) are enriched in enterotypes 3 and 1, respectively.

Taken together, we have demonstrated the existence of enterotypes in the human gut microbiome and have identified three of them that vary in species and functional composition using data that spans several nations and continents. As our current data do not reveal which environmental or even genetic factors are causing the clustering, and as faecal samples are not representative of the entire intestine, we anticipate that the enterotypes introduced here will be refined with deeper and broader analysis of individuals’ microbiomes. Presumably, enterotypes are not limited to humans but also occur in animals. Their future investigation might well reveal novel facets of human and animal symbiotic biology and lead to the discovery of those microbial properties correlated with the health status of individuals. We anticipate that they might allow classification of human groups that respond differently to diet or drug intake. Enterotypes appear complex, are probably not driven by nutritional habits and cannot simply be explained by host properties such as age or BMI, although there are functional markers such as genes or modules that correlate remarkably well with individual features. The latter might be utilizable for diagnostic and perhaps even prognostic tools for numerous human disorders, for instance colorectal cancer and obesity-linked co-morbidities such as metabolic syndrome, diabetes and cardiovascular pathologies.

METHODS SUMMARY

Sample collection. Human faecal samples from European individuals were collected and frozen immediately, and DNA was purified as described previously. Sequencing was carried out by Sanger-sequencing random shotgun DNA libraries of 3 kb using standard protocols established at Genoscope. For sequence processing, cloning vector, sequencing primers and low-quality bases were end-trimmed from raw Sanger reads, and possible human DNA sequences were removed. Reads were processed by the SMASH comparative metagenomics pipeline for assembly and gene prediction. Informed consent was obtained from the 22 European subjects. Sample collection and experiments were approved by the following ethics committees: MetaHIT (Danish), ethical committee of the Capital Region of Denmark; MetaHIT (Spanish), CEC, Hospital Vall d‘Hebron; MicroObes, Ethical Committee for Studies with Human Subjects of Cochin Hospital in Paris, France; MicroAge, Joint Ethical Committee of the University of Camerino.

Phylogenetic annotation. Phylogenetic annotation of samples was performed by (1) aligning reads (Sanger/Illumina) against a database of 1,511 reference genomes (listed in Supplementary Table 3), or (2) classifying 16S rDNA reads using RDP classifier. Genus and phylum abundance was estimated after normalizing for sequence read counts (Sanger/Illumina) against a database of 1,511 reference genomes. For sequence processing, cloning vector, sequencing primers and low-quality bases were end-trimmed from raw Sanger reads, and possible human DNA sequences were removed. Reads were processed by the SMASH comparative metagenomics pipeline for assembly and gene prediction. Informed consent was obtained from the 22 European subjects. Sample collection and experiments were approved by the following ethics committees: MetaHIT (Danish), ethical committee of the Capital Region of Denmark; MetaHIT (Spanish), CEC, Hospital Vall d‘Hebron; MicroObes, Ethical Committee for Studies with Human Subjects of Cochin Hospital in Paris, France; MicroAge, Joint Ethical Committee of the University of Camerino.

Functional annotation. Genes were functionally annotated using BLASTp against eggNOG (v2) and KEGG (v50) databases. Protein abundances were estimated after normalizing for protein length. Functional abundance profiles at eggNOG, KEGG orthologous group, functional module and pathway level were created.

Clustering and classification. Samples with identical Shannon–Shannon distance and partitioning around medoid (PAM) clustering. Optimal number of clusters was estimated using the Calinski–Harabasz (CH) index. We used the silhouette validation technique for assessing the robustness of clusters. Additionally, within a cross-validation scheme, we trained predictive decision tree models on clusters obtained using the same clustering method and evaluated the classification of hold-out samples by accuracy, average precision and average precision gain. Statistics. Correlations between metadata and feature abundances were computed as described previously, based on multiple-testing corrected pairwise Spearman correlation analysis and stepwise regression for multi-feature model building. For categorical metadata and enterotype comparisons, samples were pooled into bins (male/female, obese/lean, one enterotype/rect, specific nationality/rect etc) and significant features were identified using Fisher’s exact test with multiple testing correction of P-values.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors are grateful to C. Creevey, G. Falony and members of the Bork group at EMBL for discussions and assistance. We thank the EMBL IT core facility and Y. Yuan for managing the high-performance computing resources. The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013); MetaHIT, grant agreement HEALTH-F4-2007-201052, EMBL, the Lundbeck Foundation Centre for Applied Medical Genomics in Personalized Disease Prediction, Prevention and Care (LUCAMP), Novo Nordisk Foundation and the International Science and Technology Cooperation Project in China (0806). Obese/non-obese volunteers for the MicroObes study were recruited from the EU-VIVA cohort study coordinated by S. Silvi and A. Cresci, and metagenome sequencing was funded by Genoscope. Ciberehd is funded by the Instituto de Salud Carlos III (Spain). J.R. is supported by the Institute for the encouragement of Scientific Research and Innovation of Brussels (ISRIB) and the Odysseus programme of the Fund for Scientific Research (Flanders). We are thankful to the Human Microbiome Project for generating the reference genomes from human gut microbes and the International Microbiome Consortium for discussions and exchange of data.


Author Information Raw Sanger read data from the European faecal metagenomes have been deposited in the NCBI Trace Archive with the following project identifiers: MH6 (33049), MH13 (33053), MH12 (33055), MH50 (33057), C01 (33059), C02 (33061), UC4 (33113), UC6 (33063), N01 (33305), N03 (33307), N04 (33309), N08 (33311), OB2 (33313), OB1 (38231), OB6 (38233), OBB (45929), A (60703), B (60705), C (60707), D (60709), E (60801), G (60803). Contigs, genes and annotations are available to download from http://www.bork.embl.de/Docu/Arumugam_et_al_2011/. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.B. (bork@embl.de) or S.D.E. (dusko.ehrlich@jouy.inra.fr).

MetaHIT Consortium (additional members)

Marie-Antoinette2, François Artigue4,2, Hervé M. Blottière5, Mathieu Almeida2,3, Christian Brechot2,3, Carlos Cara2,3, Christian Chervaux2,3, Antoinette Cütrom2,3, Christine Delorme2,3, Gérard Denariaz2,3, Rozenn Dervyn2,3, Konrad U. Foerstner6,7, Carsten Friss8, Maarten van de Guchtte2,3, Eric Guadon2,3, Florence Haimet6, H. Johann van Hylckama-Vlieg2,3, Alexandre Jarnet1,2, Catherine Justet2, Ghiaia Kaci2,3, Jan Knol5,10, Karsten Kristiansen13, Omar Lakhdari2,3, Severine Layec2,3, Karine Le Roux2,3, Emmanuelle Maquin2,3, Alexandre Mérieux12,10, Raquel Melo Minardi2,3, Christine M’rini2,3, Jean Müller8,2, Reash Oozeer5,10, Julien Parkhill10, Pierre Prud’homme2,3, Maria Rescigno11, Nicolas Sanchez2,3, Shinichi Sunagawa2,3, Antonio Torrejon1,2, Katherin Tran2,3, Gaetana Vandermeulebroeck2,3, Encarna Varela2,3, Yohanian Winogradsky3,9 & Georg Zeller9

1Digestive System Research Unit, University Hospital Vall d’Hebron, Ciberer, 08035 Barcelona, Spain. 2Commissariat à l’Energie Atomique, Genoscope, 91000 Evry, France. 3Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. 4UCB Pharma SA, 28046 Madrid, Spain. 5Danone Research, 91120 Palaiseau, France. 6European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. 7Heidelbergstrasse 24, 64285 Darmstadt, Germany. 8Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark. 9Institute of Genetics and Molecular and Cellular Biology, CNRS, INSERM, University of Strasbourg, 67404 Illkirch, France. 10The Welcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. 11Istituto Europeo di Oncologia, 20100 Milan, Italy. 12Institut Mérieux, 17 rue Burgelat, 69002 Lyon, France. 13Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark.

©2011 Macmillan Publishers Limited. All rights reserved
3.3 Article 3: “Ecological resilience of the human gut microbiome and human healthy”

Many recent studies have attempted to determine the mechanism through which the human gut microbiome affects human health. The goal was to identify so-called ‘bad’ bacteria that could cause disease or influence human health. Unfortunately, the search for such bacteria was unfulfilled, as each human gut microbiome harbours distinct bacterial strains. The idea that human gut microbiome composition would be defined by expected factors — e.g. disease, lifestyle, and age — also lost strength after the discovery of enterotypes described in article two. However, during 2011, several articles related to IBD and model organisms demonstrated that some commensal bacteria can regulate the innate immune system. Moreover, subproducts of the gut microbiome (e.g. SCFAs) are known to be important for intestinal cell function and the control of inflammatory responses. This knowledge changed the scientific perspective on the human gut microbiome and indicated that defining the constituents of a healthy microbiome is key to understanding what constitutes an unhealthy microbiome. Although homeostasis was defined in 1929 by Walter Bradford Cannon, it had not been explored by many microbiome studies. Homeostasis is the property of a system that regulates its internal environment and tends to maintain a stable, constant condition of properties [120]. Because many articles had failed to identify ‘bad’ bacteria, the theory of homeostasis gained force to explain the dynamics of unhealthy gut microbiomes. A microbial imbalance in humans is defined as dysbiosis and is characterised by the overgrowth of a few species, which can damage smaller, beneficial species. Dysbiosis had been explored in several articles as a possible explanation for the dynamics of an unhealthy human gut microbiome. Thus, a healthy microbiome is defined as a high-diversity environment (many bacterial species with low abundance and even distribution), whereas an unhealthy microbiome is a low-diversity environment (overgrowth of a few species and uneven distribution). After exploring the MetaHIT data, I believe that dysbiosis may play a prominent role in shaping the human gut microbiome. Although the cause of dysbiosis in the gut is currently unknown, its effects on the gut microbiome can be observed. Since publishing the human microbiome gene catalogue, we have been interested in how microorganisms interact with each other. Because few organisms found in nature exist alone, especially bacteria, describing a bacterial community in full is a critical component of understanding the role of each species in the community and its relationship with the host. Therefore, the human gut microbiome species network published in the first article was the initial step in elucidating the microbiome ecology of the human gut. Many studies have been unable to find biomarkers to explain the relationship among microbiome variation and expected factors such as diet, health status, age, lifestyle, and nationality. However, many others have reported that differences in diversity could explain health status, supporting the idea that ecology theory could be applied to the human gut microbiome to find the link between gut microbiome and disease. This third article presents a biological explanation for human gut microbiome variation. Furthermore, it defines both healthy and unhealthy gut microbiomes. In
3.3 Article 3: “Ecological resilience of the human gut microbiome and human healthy”

Contrast to general beliefs, the unhealthy microbiome is not composed of ‘bad’ bacteria but rather is composed of fragments of a healthy microbiome with an overgrowth of *Bacteroides*, *Prevotella*, and *Escherichia* species, suggesting an imbalance in the ecosystem. These characteristics match those that define dysbiosis. The loss of some species and the growth of others likely stimulates immune response and aggravates inflammation (Figure 3.3).

This article was the first to apply theories of microbial ecology to explain the relationship between variation in composition of the human gut microbiome and disease. It was also the first to use four datasets generated using three technologies — metagenomic sequencing with Illumina, metagenomic microarray, and 16S rRNA sequencing — demonstrating that the results were not biased by methodology. Moreover, sample preparation was carried out by other groups using various extraction methods. The analyses are also true for different levels of taxonomy annotation, including 16S rRNA operational taxonomic unit 100%, strain, species, and genus. All three datasets contain more than 100 samples, which provide adequate result robustness.

Since my group defined its first species network, I have been interested applying microbial ecology theory to explain the microbiota-disease link and integrating it with published results. The challenge was to combine ecology theories with known gut parameters and report results in the contexts of enterotypes, diversity, immune system, B/F ratio, and diet, creating one model which can explain not only the results observed in the third article but also results from other studies. The article provides answers for three main questions related to the human gut microbiome. First, can any biomarker identify an unhealthy human gut microbiome? This question is important because its answer can be used to create medical tools for rapid profiling of the gut microbiome and determination of whether intervention is necessary. Such tools might also be used in the prevention of IBD relapse. Second, is the human gut microbiome composed of one or many ecosystems? The second article suggests the existence of three completely independent gut microbiomes. Answering this question may help to develop medical treatments tailored to various gut microbiome types. Finally, one of the most relevant questions about the gut microbiome is whether the shift in gut microbiome composition a cause or consequence of disease? Many articles thus far have been unable to answer this question. In addition to answering these questions, we generated a model capable of integrating many of the results published thus far, putting together small pieces of this puzzle to form a big picture.
Figure 3.3: Figure from review article [109] as figure 4. Dysbiosis model explains how an healthy gut environment become unhealthy. Instead of ‘bad’ bacteria the model suggests lack of ‘Peace-keeping’ bacteria by altered gut environment may be the cause of gut diseases. The model also suggests that inflammation can favour the selection of pathogenic bacteria [109].
Explanations of the association between the human gut microbiome and disease from an ecological viewpoint are gaining increasing acceptance, as many studies have reported that unhealthy samples have a species diversity that is lower than that in healthy samples. However, the impact of low diversity in the gut microbiome ecosystem is still unknown. By analysing four independent datasets from the human gut microbiome, we observed that the B/F ratio is a strong indicator of microbial gut composition. This ratio is negatively correlated with diversity and can explain much of the variance in our datasets. High-diversity samples comprise almost all species found in the gut microbiome, whereas low-diversity samples are subsets of high-diversity samples (i.e. nested). Using a species network, we show that species with a high number of species degrees (number of links among species) are the first to disappear when diversity decreases owing to a perturbation in the ecosystem. Furthermore, intra-sample variance is negatively associated with diversity. Our results resemble dysbiosis, with overgrowth of Bacteroidetes species and a decrease in Firmicutes and Actinobacteria. Species networks and functional analysis indicate that the microbiome shift
is related to the fitness of species under inflammatory stress. Therefore, an unhealthy microbiome is defined by the absence of specialist species and an increase in generalist species, which can thrive under a wide variety of environmental conditions. Here, we propose the first model capable of explaining human gut microbiome variance and its relationship to disease.

1 Introduction

The gut microbial flora contributes to the health of the human body. It performs several useful functions, including fermenting otherwise indigestible carbohydrates to SCFAs, forming the immune system in early childhood, preventing the growth of pathogenic bacteria, regulating the development of the gut lining, and producing vitamins for the host. The gut microbiome influences the amount of energy an individual can obtain via diet [1], and the microbial modification of bile acids affects lipid metabolism in the host [2]. The gut microbiota also plays a role in both innate [3] and adaptive [4] immune responses. In the gut of germ-free mice, epithelial cells renew at a slower rate than those in their bacterially colonised counterparts [5]. Studies have demonstrated the impact of the human gut microbiota on its host. Therefore, characterisation of variations in the human gut microbiome and the identification of key microbial species are important steps in defining healthy and unhealthy gut microbiomes.

In general, the composition of the human gut microbiome is dominated by two phyla, Bacteroidetes and Firmicutes, which display distinct phenotypes. Whereas most Firmicutes species in the gut microbiome are Gram-positive, strictly anaerobic, and endospore forming, Bacteroidetes species are Gram-negative and aerotolerant and do not form endospores. Firmicutes is a diverse taxonomic group in the gut microbiome, with many families present as commensal bacteria, whereas Bacteroidetes is dominated by two genera, Bacteroides and Prevotella [6]. Members of Bacteroidetes also produce high levels of acetate and propionate, whereas Firmicutes species produce high amounts of butyrate [7].

Complex interactions among multiple factors such as diet, genotype, physiology, environment, and the immune system have been proposed to play an important role in shaping gut microbiome composition [8]; however, the exact mechanisms of their effects remain unknown, and modelling approaches have proven unsuccessful in uncovering them. Despite extensive studies, the link between the microbiome and disease remains unclear, and no single species
has been identified to explain IBD or obesity. However, some observations have been consistently reported. For example, the microbiome of IBD patients is lower in species richness (the number of unique species in a sample) than that of healthy individuals, and dramatic reductions in the relative abundance of members of Firmicutes have been observed in IBD patients [9, 10, 11, 12]. Members of Firmicutes are known producers of acetate and butyrate, which are important SCFA metabolites with potent anti-inflammatory properties; thus, members of this group have garnered considerable interest [13, 14, 15]. Clostridial clusters IV and XIVa exhibit lower relative abundance in patients with IBD than they do in healthy controls, suggesting that this cluster may have an important anti-inflammatory role in the gut microbiome [16]. Analysis of the intestinal tissue from IBD patients has revealed that *Prevotella* species was 3.6 times more abundant at an inflamed site than at a non-inflamed site, accounting for 25% of the total community present in inflamed tissue [10]. In addition, high abundance of members of the genera *Bacteroides* and *Prevotella* has been observed in mucosal biopsy specimens from ulcerative colitis patients [17].

Dysbiosis, or disequilibrium of the microbiota, may explain the association of the microbiome with inflammatory disease. Because the gastrointestinal tract forms the largest interface between the immune system and the environment, it may have an effect on the microbiome. The normal mucosa of the small and large intestine experiences continuous low-grade (physiologic) inflammation that is absent from the gut of adult germ-free mice and the intestines of newborn humans [18, 19, 20], suggesting that the presence of microorganisms in the gut is enough to create low-grade inflammation. Regulatory T cells (Tregs), which express the *Foxp3* transcription factor, play a critical role in immune homeostasis [21] and in the distinction of self and non-self cells [22]. Bacterial species — particularly Gram-positive, spore-forming species — are prominent players in the induction of colonic Tregs. Specifically, *Clostridium* species belonging to clusters IV and XIVa are potent inducers of Tregs in the colon [23]. Thus, microbial flora is crucial to intestinal homeostasis.

In the present study, we analysed four human gut metagenomic datasets obtained using three technologies (microarray, sequencing, and 16S rRNA). We observed a strong pattern resembling dysbiosis within all four datasets. By combining our results with those of previous studies, we proposed definitions of healthy and unhealthy ecological states of the gut microbiome. A healthy community is characterised by high diversity, low B/F ratio, high number of species degrees, and consistent microbiota composition. An unhealthy microbiome is defined as a subset of the healthy microbiome with disrupted species networks caused by a disturbance in the ecosystem. Because such
disturbances affects the gut microbiome in the various ways, unhealthy samples display inconsistent composition and great variation, which may explain the difficulty in identifying biomarker species with strong associations to disease states.

2 Results

2.1 Diversity and species composition.

Defining human gut microbiome composition is an important step in understanding its relationship with disease. Bacteroidetes and Firmicutes phyla are reportedly the most abundant in the human gut microbiome, and we confirmed this observation in all four datasets (Sup. Figure 1). These two phyla represented more than 90% of the species abundance in all samples. Moreover, these phyla displayed the highest variance across all samples, although variance in Bacteroidetes was high than that in Firmicutes (Sup. Figure 2). To visualise the change in relative abundance from Firmicutes to Bacteroidetes, we plotted species abundance coloured by phyla and sorted by Bacteroidetes abundance (Sup. Figure 3). The results revealed a pattern in which species abundance is lower in samples with high B/F ratio and higher in those with low B/F ratio. The Shannon’s diversity index was used to confirm this observation. In all datasets, diversity was negatively associated with Bacteroidetes abundance and positively associated with Firmicutes abundance (Figure 1). Thus, samples with a high abundance of Bacteroidetes species have lower diversity, whereas those with a high abundance of Firmicutes display high diversity.

2.2 One or multiple ecosystems

The diversity of a sample is strongly associated with the shift in species composition (Bacteroidetes and Firmicutes). However, whether low-diversity samples are a completely new ecosystem or a subset of a high-diversity system is unclear. To study this question, we used beta diversity (a measure of the number of different species between two samples) and delta diversity (a measure of the number of common species between two samples). We observed that delta diversity was higher than beta diversity for these ecosystems, which indicates that a greater proportion of the differences between these two groups is determined by the variation in abundance and not by new species. This finding supports the idea that low-diversity samples are subsets (or nested) of a high-diversity ecosystem (Figure 2) and not a new ecosystem, although they
Figure 1: Association among diversity and Bacteroidetes/Firmicutes (BF) ratio for each sample. The x-axis shows the log of relative abundance of B/F. The y-axis shows the Shannon’s diversity index. Samples were classified into enterotype by the most abundant group in each sample represented by different colours and shapes. The size of the sample represents its richness. The linear model in blue describes the association among diversity with $R^2$. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
may contain few ‘new’ species — species that are not present in high-diversity samples. However, those are present in low abundance.

2.3 Diversity and Enterotypes
Species abundance (Sup. Figure 3) indicates a continuous distribution of Bacteroidetes abundance, not a discrete distribution as has been suggested [6]. Moreover, our results indicated that the human gut microbiome is composed of just one ecosystem, which means that most of the difference observed is determined by the loss of species and variation in relative abundance. A previous study has established that individuals can have three enterotypes of gut microbiome [6], and we used the enterotype algorithm to cluster four independent datasets. The overall result revealed that each enterotype has a different median value according to the Shannon’s diversity index, confirming our previous observation that the high relative abundance of Bacteroidetes is associated with low diversity. The Firmicutes enterotype had the highest diversity followed by the Prevotella and Bacteroides enterotypes. Therefore, diversity can also explain the enterotype clusters (Figure 3). Combining these results (Shannon’s diversity index, delta diversity, continuous distribution and enterotype cluster) with results from published studies demonstrating the shift of the clusters (from Firmicutes to Bacteroidetes), we propose that the three enterotypes are one ecosystem, and the clusters are states and therefore should be labelled as ‘Enterostages’.

2.4 Diversity and species degree
Species degree can be important in explaining diversity effects on ecosystems. Therefore, using the species network, we investigated the species degree for all datasets. We found that organisms have a higher number of species degrees with species from the same phyla. Bacteroidetes species have the lowest number of degrees, whereas Firmicutes has the highest number of degrees with itself and other phyla (Sup. Figure 4). The number of degrees in a species can be interpreted as an indicator of how dependent bacteria are on other organisms in the ecosystem. Therefore, if the community is unbalanced with decreasing diversity, it is likely that species with high degrees (more dependent) will disappear first, as the loss of one species in the highly connected part of the network affects all associated species.

To test this theory, we first correlated species abundance with diversity to determine which species increased in abundance when diversity decreased and vice versa. Many of the species that were positively correlated with diversity
Figure 2: Beta and delta diversity. Beta diversity is the number of different species between two groups. Delta diversity is the number of similar species between two groups. Samples were ordered using the Shannon’s diversity index and grouped with a window size of 20 samples and a sliding window of 1. Beta and delta diversity were calculated against the reference group (the 20 samples with the highest diversity) and moved in the direction of low diversity. Delta diversity in red represents the percentage of similar species, and beta diversity in blue represents the percentage of different species. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
Figure 3: Shannon’s diversity index for each sample classified in enterotype clusters. In all four datasets, the Firmicutes enterotype had the highest median diversity index. Although the *Prevotella* cluster had some of the highest evenness, it had medium richness. Overall, *Prevotella* had the second lowest diversity index. The *Bacteroides* cluster had the lowest diversity index in all datasets. These data are important because a healthy microbiota reportedly has diversity higher than that in an unhealthy microbiota. Therefore, unhealthy individuals are more likely to display the *Prevotella* or *Bacteroides* cluster. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
belonged to Firmicutes (Sup. Figure 5), whereas most of the species that were negatively correlated with diversity belonged to Bacteroidetes. Therefore, species analysis also confirmed previously observed associations between diversity and phyla. Subsequently, we correlated this measurement with the number of species degrees. The results suggested that species with a high number of degrees are the first to disappear from an ecosystem when diversity decreases (Sup. Figure 6). The number of species degrees increased with diversity in all datasets (Figure 4). Therefore, species are less dependent in the setting of low diversity and more dependent in high diversity.
2.5 Interindividual variance and ecosystem stability

Although diversity may explain most of the variance in human gut microbiome data, the impact of diversity on the ecosystem is unknown. Therefore, to observe the impact of evenness, we selected the top 100 species from each enterotype cluster to observe which species were more abundant and how bacterial abundance changes with diversity. The cluster with the highest diversity displayed a low B/F ratio, a high number of species, low species variation, low abundance, and an even distribution. The cluster with the next highest diversity showed an increased B/F ratio and a higher variance across samples. The cluster with the third highest diversity had the most extreme profile, displaying very high abundance of a few species as well as species with the highest variance across samples and the highest B/F ratio. Therefore, samples with high diversity had lowered variance and were more similar to each other in terms of species composition than were samples with low diversity (Figure 5). This result is important, as it may explain the lack of species markers associated with disease because low-diversity samples have the highest variance, and therefore, do not display similar species prevalence. Furthermore, the effect of diversity also changed the rankings of the most abundant species in the three diversity groups. In cluster 1, the top Firmicutes species are mostly Clostridium species. Clostridium has been reported to regulate immune homeostasis, and a high number of species have been shown necessary for efficient Treg induction [23]. Therefore, high-diversity ecosystems have lower interindividual variance and thus seem more stable.

2.6 Adaptation: Generalist and specialist species

The fact that low-diversity samples are a subset of high-diversity samples suggests that low-diversity samples would not be able to fulfil the requirements for normal gut function. To test this hypothesis, we compared the functional capacity of high- and low-diversity samples. We observed that low-diversity samples are rich in endotoxins (Sup. Figure 8). This characteristic supports the taxonomic analysis, as most of the high-abundance species in low-diversity samples belong to Bacteroidetes (Gram-negative). Moreover, low-diversity samples are rich in catalase and peroxidase genes. Peroxidase genes are overrepresented in aerotolerant species such as Bacteroides, and catalase was observed in facultative anaerobic species such as E. coli. The capacity of those species to survive in the presence of oxygen has been previously reported [26, 27], and the enrichment of those genes in low-diversity samples suggests the presence of oxygen in the ecosystem, which could be caused by the inflammatory response.
Figure 5: Variance of species abundance over diversity. Samples were ordered by diversity with a window of 20 samples and a slide window of 1. The sum of variance of all species in each group was measured and compared with the mean Shannon diversity index for the same group. The figure shows a strong negative correlation among species abundance variance and diversity. Species in low-diversity samples have high variance in abundance, and those in high-diversity samples have low variance in abundance. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
of ROS. The presence of oxygen gives aerotolerant/facultative anaerobes an advantage by increasing their fitness, thereby increasing their abundance in the environment. A high abundance of Enterobacteriaceae has been reported in IBD patients [28, 29, 30]. Furthermore, genes responsible for the production of the anti-inflammatory compounds butyrate and acetate are enriched in high-diversity samples, confirming the taxonomic distribution as Firmicutes in these samples. Butyrate decreases pro-inflammatory cytokine expression via inhibition of nuclear factor kappa B [15], suggesting that high-diversity microbiomes have a greater capacity to reduce inflammatory response.

Spore genes are also enriched in high-diversity samples. Clostridium is the genus with the highest number of spore-related genes in the gut microbiome, confirming the known capacity of this group to produce spores. Analysis of carbohydrate enzymes revealed that 50% of all genes are found in only three genera: Bacteroides, Lactobacillus, and Clostridium. The distribution of carbohydrate enzymes is not equal for all genera. Bacteroides has 80 families, Lactobacillus 40, and Clostridium 51, although Lactobacillus has two times more carbohydrate enzymes than Clostridium. Comparing Bacteroides and Lactobacillus, we observed that glycoside hydrolase (GH) GH13, GH1, GH73, and GH25 are enriched for Lactobacillus, and GH2, GH3, and GH43 are enriched for Bacteroides, suggesting different roles for these genera in the gut. Carbohydrate enzyme capacity separated the genera into three clusters, indicating three major functional groups of carbohydrate enzymes (Sup. Figure 9). The fact that Bacteroides species can increase abundance when the ecosystem is under stress suggests a greater capacity to adapt to changes than that of other species in the gut microbiome.

Bacterial genome size has been associated with adaptation, as free-living bacteria have genomes that are larger than those of host-restricted bacteria [31, 32]. To confirm this association, we analysed the bacterial genome size for the most abundant genera found in the gut microbiome. We found that Bacteroides and Escherichia genera have large genomes, supporting the idea that they are more capable of adapting to environmental changes (Sup. Figure 7). Both of these genera have been shown to be overrepresented in gut microbiome samples from IBD patients. Although these genera belong to different phyla, both are Gram-negative, aerotolerant, and facultative. To test this idea further, we analysed the distribution of average genome size within a sample from the MetaHIT dataset. The results showed that average genome size decreases with increasing sample richness. The overall functional analysis indicated the capacity of each genus to adapt to changes in the environment. Some species are specialists, being present only under optimal conditions and producing spores when the ecosystem changes, whereas others were like generalists with
3 Discussion

The shift observed in our results is similar to that exemplified by dysbiosis, with overgrowth of *Bacteroides, Prevotella* and *E. coli*. These species are known for their capacity to survive under oxidative stress. Interestingly, the same shift in species has been observed in IBD [10], obese [33], and knockout mice (e.g. *Nlrp6*−/−, *Nlrp3*−/−, *Rag*−/−, and T5KO) [34, 35, 36, 37], demonstrating that the immune system plays a central role in shaping the composition of the microbiota [38]. The observed shift affects the production of SCFAs, especially butyrate, which may in turn influence colonic epithelial cells because butyrate is used as a primary energy source [39]. Moreover, G-protein-coupled receptor 43-SCFA interactions affect inflammatory responses, establishing a link among diet, gastrointestinal microbiome, and immune and inflammatory responses [40].

Phenotypic characteristics are properties of their common ancestors, and ecological interactions have been shown to be phylogenetically conserved across the tree of life, with closely related species interacting with similar partners [41]. We confirmed this assumption in the four datasets we analysed, in which most of the species interacted with other species from the same phyla in the ecological network. In ecological communities, species can have negative or positive interactions with one another (e.g. competition/predation and mutualism). Our ecological network revealed that most of the interactions among species are positive, suggesting mutualism. Recently, mutualistic ecological networks have been found to have minimal competition and increased alpha diversity [42, 43]. We observed increased richness in samples with high relative abundance of Firmicutes and a high number of species degrees. It has been said that “nestedness in mutualistic ecological networks is the tendency for ecological specialists to interact with a subset of species that also interact with more generalist species” [44]. Firmicutes species display a high number of interactions with species from different phyla in the network we developed, suggesting that in the gut microbiome, some Firmicutes species function as intermediaries among specialists and generalists. Nestedness reduces effective interspecific competition and enhances the number of coexisting species [42], and nested networks naturally emerge if new species are more likely to enter a community.
in which they will have a minimal competitive load [42]. A highly connected
and nested architecture promotes community stability in mutualistic networks
[43]. Intra-species variation supports the stability of a community with high-
diversity samples. Therefore, we suggest that high Firmicutes samples may
begin as a nested mutualism ecological network but, with an increased number
of species, becomes nested and stable. Of interest is the fact that most of
the unknown species are mainly found in high-diversity samples. Taxonomic
prediction methods suggest that most of those species belong to the Firmicutes
phylum.

To explain the variation in the human gut microbiome and inflammatory
responses, we propose the following model (Figure 6). Healthy individuals
have a high-diversity microbiome [10, 11, 6, 45], and our results associate high
diversity with a high abundance of Firmicutes. Disturbances in the balance (e.g.
viruses, pathogens, antibiotics, diet, allergy, and ageing) can lead to dysbiosis
and reduce diversity. The first species to disappear have high species degrees
and small genome size. Firmicutes — more specifically, Clostridium species
— predominate in this group. This disappearance abrogates the inflammatory
response that is mediated by these organisms [23]. To survive oxidative stress,
Clostridium species produce endospores, whereas species with higher tolerances
to inflammatory stress continue to grow, decreasing the production of SCFAs
and increasing LPS levels, as most of the species capable of surviving ROS in
the gut are Gram-negative.

For example, the role of Bifidobacteria may be related to the spacing of tight
junctions of the cells lining the gut [46]. A decrease in the abundance of Bifi-
dobacteria would increase the space between the cells, increasing permeability.
In case of high abundance and concentration of Gram-negative bacteria, the
loss of Bifidobacteria species in turn enhances the amount of LPS passing out of
the gut wall. This change causes endotoxaemia and inflammation, which may
induce a number of metabolic disorders [46]. A high abundance of Actinobac-
teria is found in high-diversity samples, and Bifidobacteria is one of the first
to disappear from the network when diversity decreases. The increase in LPS
worsens the inflammatory response, especially that of E. coli, because the LPS
of Bacteroides is 10 to 1,000 times less toxic than that of E. coli [47]. When the
relative abundance of Bacteroidetes and Escherichia species increase, bacterial
density also increases [37]. If the immune system is capable of controlling the
overgrowth of these bacteria and decreasing bacterial concentration, the inflam-
mation will be reduced. If not, the individual will be diagnosed with severe
inflammation and treated with antibiotics and anti-inflammatory drugs [48],
thereby reducing these bacterial concentrations. With no inflammatory signals,
Clostridium species will grow again, inducing Tregs and restoring homeostasis.
Figure 6: Model combining the results from this article and published results. The model explains the dynamics of the human gut microbiome from high diversity to low diversity. It also offers an explanation for the mechanism underlying the restoration of the microbiome composition, as reported in antibiotic experiments and studies in mice. The transition from Gram-positive to Gram-negative has been also previously reported. Our results placed the enterotypes into three ecological stages, which can be related to diseases.
This model predicts how environmental changes will impact the microbial ecosystem and how overgrowth of certain groups can reinforce these changes, creating a positive feedback loop. The model also aids understanding of dysbiosis and offers an explanation for its transmissibility [34], because low-diversity samples have a high abundance of LPS, active inflammatory response, and high bacterial concentration, which is likely to induce inflammation in the new host. The opposite is untrue, however, as specialist species only colonise the ecosystem under optimal conditions. Our model helps to explain whether the shift in the human gut microbiome is a cause or a consequence of disease, and the results suggest that is both. The microbiome shifts as consequence of inflammatory response, but the species capable of survival in the stressed environment may worsen inflammation via LPS, increased gut permeability, and reduced capacity to regulate the immune-system, giving the human gut microbiome a causative role in disease.

Moreover, our model helps to explain observations of obesity in mice. Obese mice reportedly have a high abundance of Firmicutes, and mice on a high-carbohydrate diet display an increase in the relative abundance of Bacteroidetes. Our results suggest that Firmicutes are the specialists and therefore will grow in the ecosystem that accompanies a high-carbohydrate diet. Therefore, mice on low-carbohydrate diets display a decrease in Firmicutes abundance, which increases the relative abundance of Bacteroidetes. Together, these data allow us to propose an ecological model that describes the community changes that go along with richness decrease and hypothesises its underlying causes (inflammation/ROS) and the effects it has on host interplay. In addition, this theory allows the linkage of well-known gut parameters and reported results (enterotypes, B/F ratio, richness, immune-system, and diet), bringing us one step closer to a full understanding of the ecological underpinnings of gut health.

**Methods**

Taxonomy: Taxonomy for genes in the microarray and sequencing data was carried out using nucleotide Basic Local Alignment Search Tool (BLAST), selecting the best hit. Only alignments with more than 95% identity and alignment length longer than 100 base pairs were accepted. The full-path taxonomy information (strain to kingdom) was collected from the National Center for Biotechnology Information (NCBI) taxonomy database.

Species abundance: Species abundance was calculated using the mean, count (from sequencing), or signal (from microarray) from all genes for each species.
The abundance was then normalised for relative abundance, with all samples having 100%.

Phylum abundance: Phylum abundance was determined by summing the relative abundance of all species that belonged to each phylum.

Phylum variance: Variance (\(\sigma^2\)) for each phyla was calculated in R using the function `var`.

Species distribution: Species abundance and phylum abundance were calculated for each sample (see species abundance and phylum abundance), and samples were sorted according to Bacteroidetes. Each phylum received a colour from the RColorBrewer package using palette Set3.

Enterotype clusters: Enterotype clusters were created using JensenShannon distance to measure the distance among samples and partitioning around medoid to generate the clusters, with three clusters as total number of clusters [6].

Diversity: Diversity was calculated using the Shannon’s diversity index from package vegan, function: `diversity` and method ‘shannon’.

B/F ratio: B/F ratio was calculated by dividing the phylum abundance for Bacteroidetes by the phylum abundance for Firmicutes.

Richness: Species richness or alpha diversity was calculated by the number of species with abundance higher than 0.2% in relative abundance for all four datasets.

Beta diversity: Beta diversity was calculated by counting the number of species that were not shared by the two communities. The relative beta diversity is calculated as beta diversity divided by the gamma diversity from the communities.

Delta diversity: Delta diversity was calculated by counting the number of common species between two communities. The relative delta diversity was calculated as delta diversity divided by the number of species in the community with the lowest alpha diversity.

Gamma diversity: Gamma diversity was calculated by counting the total number of unique species found between two communities.
Relative delta and relative beta diversity in samples: To calculate the delta diversity in samples, we first sorted all samples according to Shannon diversity index values. The 20 samples (window size: 20) with the highest diversity were used as a representation of a high-diversity community. Then we moved one sample down and selected the next 20 samples (sliding window: 1) to represent the community with low diversity. The two communities are compared using delta and beta diversity calculations. The process continued until the last 20 samples were analysed.

Species network: To create the species network, we used the Spearman correlation index. We only accepted the association between two species if the correlation index was higher than an absolute value of 0.4.

Species degrees: Species degrees were calculated by the number of associations one species had with other species in a network (see species network).

Species abundance versus diversity: The association among species abundance (see species abundance) and diversity (see diversity) was calculated with the Spearman correlation test from R package using the stats function cor.test method = Spearman. Only correlations with p values of < 0.001 were accepted.

Species degrees in samples: Samples are sorted by diversity (see diversity), and the species degrees were calculated (see species degrees) for only species with a relative abundance higher than 0.2%, a window size of 20, and a sliding window of 1.

Species variance: Samples were sorted by diversity (see diversity), and the species variance was calculated in R using the var function with a window of 20 samples and sliding window of 1. The sum of variance of all species in each group (group of 20 samples) was calculated. The mean Shannon diversity index was also calculated for each group.

Functional analysis: Genes were functionally annotated using protein BLAST against the Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (v2), Kyoto Encyclopedia of Genes and Genomes (v50) and Carbohydrate Active Enzymes Database databases. The best hit was used only if it had more than 25% identity and alignment length longer than 100 amino acids.

Genome size: Genome size and taxonomy identification information was ob-
References


Figure 1: Relative phylum abundance distribution from four datasets assembled using three technologies. The x-axis shows the phyla Bacteroidetes, Firmicutes, Proteobacteria, Actinobacteria, and other (a group of all other phyla). The y-axis shows the relative abundance. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
Figure 2: Relative phyla abundance variance in four datasets assembled using three technologies. The x-axis shows the phyla Bacteroidetes, Firmicutes, Proteobacteria, Actinobacteria, and other (the sum of variance from all other phyla). The y-axis shows the variance. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
Figure 3: Relative species abundance distribution from four datasets assembled using three technologies. The x-axis shows the samples. The y-axis shows the relative abundance for each species in each sample coloured by phyla (Bacteroidetes = green, Firmicutes = yellow, Proteobacteria = purple, Actinobacteria = red, and other = blue). (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
Figure 4: Using the species network, we measured the number of species degrees (link between two species) for each phyla. In all four datasets, Firmicutes displayed the highest number of species degrees within Firmicutes and with other phyla. The remaining phyla have greater degrees among themselves than with other phyla. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
Figure 5: Species abundance versus diversity. We identified which species increase and decrease in relative abundance when diversity increases. In all four datasets, more species increased in abundance when diversity increased. Most of the species with increased abundance belonged to Firmicutes and Actinobacteria. The species decreasing in abundance were mostly from Bacteroidetes. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
Figure 6: To understand how diversity affects the species network, we correlated species degree with diversity. The y-axis shows the number of species degrees measured for each species. The x-axis shows the correlation of species abundance versus diversity divided into three groups. Group 1: Species significantly negatively correlated with diversity. Group 2: Species with no significant correlation with diversity. Group 3: Species significantly positively correlated with diversity. Each circle represents a species, coloured by phylum. The size of the circle represents the mean abundance over all samples. (A) 16S rRNA dataset with 190 samples [24]. (B) 16S rRNA dataset with 297 samples [25]. (C) MetaHIT microarray with 121 samples. (D) MetaHIT sequencing with 124 samples [9].
Figure 7: We used taxonomic identification to find the genome size of each taxonomic organism in the dataset. The genome size is grouped by genus to provide an estimation of overall size for each genera, as individual species can vary in size. *Bacteroides* and *Escherichia* have the highest median genome size in the most abundant genera in the human gut microbiome.
Figure 8: Functional annotation of the most abundant families in the human gut microbiome. For each taxonomic group, the number of annotated genes was counted. The functional annotation analysis confirmed what is, in the most cases, well described in the literature: Gram-negative bacteria with high lipopolysaccharide and *Bacteroides* and *Escherichia coli* with genes related to oxygen. With respect to short-chain fatty acids (SCFAs), most of the genes were found in Clostridiaceae (butyrate) and Lactobaciliaceae (acetate).
Figure 9: Similarity in carbohydrate enzyme capacity per genera. The result divided the most abundant genera found in the human gut microbiome into three clusters. With the exception of Prevotella, the species in the clusters are related with the most abundant species in the enterotype clusters. Species in cluster one are more similar to each other in the use of carbohydrate enzymes than are those in clusters two and three.
Chapter 4

Conclusion

The understanding of how the human gut microbiome affects health has changed during recent years. The MetaHIT project and the results of this thesis are partially responsible for that change. The connection between the microbiome and disease is not as direct as once thought — it is complex, and the microbiota may play an indirect role in disease, probably through the loss of so-called ‘good’ bacteria. Moreover, variation in the human gut microbiome cannot be explained by factors such as food, genotype, nationality, or pathogenic bacteria. Thus, the idea that so-called ‘bad’ bacteria cause disease is losing strength. Our results indicated that diversity defines the human gut microbiome, and decreases in diversity resemble the effects of dysbiosis. In particular, the innate immune system may play a more influential role in intestinal homeostasis than previously perceived. The equilibrium between the immune system and immunoregulatory bacteria appears to be a delicate balance in which the loss of a specific species can lead to an overreaction of the innate immune system. The mechanism through which the imbalance starts is unclear, but once it begins, it can lead to strong inflammatory responses that may cause tissue damage and disease.

This thesis accomplished three technical milestones. A Gut-Array was developed with which to profile hundreds of DNA and RNA samples rapidly. The current version of the array (version 6.0) includes 840,000 genes with five probes per gene, representing more than 80% of all data from the 700 samples sequenced. This versatile tool can be used to explore overall tendencies as well as to determine how one variable can affect an ecosystem. The array provides a singular opportunity to profile the effects of drugs or bacteria on the gut microbiome at the DNA and RNA levels. It may also be used to test various industrial products or compounds in clinical testing. The Gut-Array tool makes the analysis of high-throughput data affordable to small laboratories, hospitals, and industry. Moreover, we generated a large amount of
empirical data from the human gut microbiome for the first time in this thesis — specifically, a profile of 350 RNA samples and 350 DNA samples using the Gut-Array.

The development of software to analyse metagenomic data was another important part of this thesis. I created an R-package, MetaHIT-R, for rapid profiling of 16S rRNA, DNA, and RNA data obtained through sequencing or array. This package has many functions that transform data from basic information to biological information, and all of the figures presented in the third article described above were generated using the program. Biological annotation of the data is required to fully utilise MetaHIT-R. Therefore, I created a pipeline to annotate more than 4 million bacterial genes with taxonomy and functional information. For taxonomy annotation, I extracted the exact taxonomy or used prediction models in cases in which exact taxonomy was unknown. For functional annotation, the genes were annotated using the COG, eggNOG, KEGG, CAZy, and Pfam databases. This infrastructure can be easily applied to other metagenomic projects.

The scientific results can be found in the three published articles outlined above. The articles represent the evolution of our view of the gut microbiome. The first article provided us with the highest number of empirical metagenome sequence data from the human gut microbiome. The species network by co-abundance demonstrated an unexpected result, showing that Bacteroides species have few species degrees and form a separated network, whereas Firmicutes species are more connected with other species. Moreover, the overall abundance of those species was very different. Whereas Firmicutes had a very low median abundance, most of the Bacteroidetes had high abundance. We wondered whether the species abundance and its species degrees were connected, suggesting different levels of independence which could explain their role in the gut. The third article is an extension of our results from the first paper in which we applied an ecological approach and explored dysbiosis to explain the dynamic of the gut microbiome.

The second article marks the beginning of our new perspective on the microbiome and disease. It demonstrates three major compositions in the human gut microbiome but cannot explain them. The publication of the paper had a major impact because it did not offer biomarkers to explain the microbiome–diseases link. This move away from a search for biomarkers represents a new approach to understanding the microbiome. The introduction of these ideas was followed by many reviews (including insights and perspective articles). One of these reviews indicates that the way to understand the microbiome is through the use of ecology. As I worked on this project, I began to see many features that were incompatible with the conclusions of other researchers. For instance, to design a microarray capable of evaluating any human gut microbiome, I needed to find genes that were common among many individuals. Although the gene catalogue reported 3.3 million genes, my analysis indicated that more than 2 million genes were rare, suggesting that the real number of genes in the human gut microbiome is lower than what has been suggested. With 700 thousand genes, the microarray is capable of profiling nearly the full range of the gut microbiome, as the number of genes in samples varies from approximately 200,000 to 1 million.
Moreover, I observed that most of the rare genes were present in samples with high Firmicutes abundance, and most of them had no taxonomy classification. Early in this research, the results were already indicating characteristics of the human gut microbiome that are explored in more detail in the third article.

The third article describes in detail the characteristics of the gut microbiome dynamic using an ecological approach. We showed that 90% of the relative abundance of the human gut microbiome consists of two phyla (Bacteroidetes and Firmicutes). These two phyla have the highest variance in all datasets, thus demonstrating their importance in the gut of both healthy and unhealthy individuals. Moreover, we found that the B/F ratio forms a strong pattern that defines the composition of the human gut microbiome. This trend was associated with diversity and led to biological changes in the ecosystem. Thus, we concluded that a high abundance of Firmicutes is associated with high diversity, whereas a high abundance of Bacteroidetes is associated with low diversity. These observations are important, because the ecosystem may be examined by measuring the overall abundance of just one phylum. In clinical tests, an overabundance of Bacteroidetes signifies unhealthy gut microbiota, and monitoring Bacteroidetes abundance in patients could alert doctors before severe disease occurs. We also propose that high-diversity samples are healthier than low-diversity samples. This classification supports previous observations that high-diversity samples are healthy [58, 60, 68, 114]. This classification may be the first step in understanding how the microbiota influences inflammatory disease. Therefore, The first important result is the association of B/F ratio with diversity. Many studies to date have demonstrated the composition of these two phyla with various interpretations, and many other studies have demonstrated differences in diversity among healthy and unhealthy individuals. We demonstrated for the first time the strong association of diversity and shifts in gut microbiome composition. The species network demonstrates some interesting characteristics which have never been reported, such as species degrees, and describes which species interact more frequently with species from other taxonomy groups. However, the two most important results of the third article are related to the data obtained through delta diversity calculations and determinations of variance among diversity. Delta diversity clearly demonstrated that the human gut microbiome is high diversity and composed of all possible species, and low-diversity samples are subgroups (nested). The same idea is often used for bioconservation studies to measure the impact of disturbances in an ecosystem. The variance showed that as an ecosystem loses diversity, samples become less similar to one another, which offered for the first time a possible explanation for the lack of biomarkers in the human gut microbiome.

Because the network suggested various levels of independency among species, I also explored that issue. Using genome size and functional analysis, I separated the two major phyla in the gut microbiome as generalists and specialists; most of the Firmicutes were specialists, and most of the Bacteroidetes were generalists. This view is completely different from previous perspectives of the gut microbiome and does not relay on so-called ‘good’ or ‘bad’ bacteria. The idea of stages is also becoming more common in the scientific community. I demonstrated that the gut microbiome
enterotypes have different diversity levels and are nested. Therefore, the enterotypes are in fact ‘Enterostages’, losing diversity when perturbation occurs in the microbiome. Putting the results of the third paper in context with studies that have been published thus far, I generated the first model that explains the dynamics of the gut microbiome and how they can be restored to the original stage after perturbation (i.e. resilience and ‘Enterostages’). Moreover, I offered an explanation to integrate the many independent conclusions of other studies that addressed diet (obese patients have high abundance of Firmicutes), enterotype, transmissible dysbiosis, disease-related microbiome shifts, lack of biomarkers, probiotics, and immunity.

To conclude, I believe the co-evolution of the human gut microbiome and its host is the most important factor in shaping its function. The two major roles of the gut microbiome are to protect the host against alien species and provide energy from food. Therefore, the shifts we observed are not related to ‘good’ or ‘bad’ bacteria or biomarkers but are an evolutionary response of the microbiome to disturbances. Because the gut is an anaerobic environment colonised by indigenous species (species that co-evolved with the host), aerobic pathogenic species are unlikely to invade and colonise the gut, as the lack of oxygen is a major barrier. However, anaerobic pathogenic species can invade and cause disease. The first host response is to activate the innate immune system, and one of the characteristics is to induce the production of ROS. This reaction eliminates anaerobic pathogen species, thereby eliminating the disease. However, if the gut were invaded by facultative pathogenic species — facultative species that can outcompete anaerobic species — it would have an advantage when the innate immune system activated ROS, removing the competition and increasing energy capacity, thus increasing its fitness. One feature that has been demonstrated before is the increase of bacterial concentration in the gut when the shift in composition occurs at the same time that diversity decreases. This mechanism may be one that developed in the microbiome to outcompete pathogenic bacteria in the gut, especially facultative species. Having bacteria that are more self-sustaining and have a high range of functional capacity is important when the human gut microbiome is under stress.

My results from the third paper suggest that Bacteroides have this characteristics and may play a protective role in the gut when the ecosystem is under stress. Specialists, mostly Firmicutes, have an energetic role when the system is under optimal conditions. However, the species network also showed that generalists are not restricted to Bacteroides. We also found some Firmicutes species with high abundance and low species degrees. Coincidently, those species are also aerotolerant/facultative, like some species from the Bacilli class. Specialists are not restricted to Firmicutes. The results in the third paper demonstrated that some Actinobacteria species could also be considered specialists and disappear at the first sign of ecosystem disturbance.

The model may also explain the role of probiotics in diseases and gut microbiome composition. As was discussed earlier, some species (specialists) do not survive when the ecosystem is under pressure. Thus, the use of probiotic specialists is likely to have no effect if the ecosystem is under stress. Coincidently, many probiotic organisms
are aerotolerant/facultative and capable of growing independently of other species. This capability is found in generalists in the gut microbiome, suggesting a role for some probiotics when the system is disturbed. The capacity to grow bacteria in the laboratory suggests a level of independence. One of the first bacteria to be isolated and easily cultivated in the laboratory was \textit{E. coli}, which has become a ‘model’ organism because it is so easy to work with in that setting. Its level of independence and its facultative capacity to grow with or without oxygen also suggest a role during ecosystem stress. High levels of \textit{E. coli} have been reported in many studies, as mentioned in the background section. In our model, \textit{E. coli} may represent the last stand of the gut microbiome against facultative pathogenic species. However, the same co-evolutionary system, which was designed to protect the host, can also be used against it, indicating a gap in the microbiome–host relationship. An \textit{E. coli} organism with a pathogenic island has a strong fitness advantage because of the dynamic of the gut microbiome and the immune system. The same tools developed by indigenous \textit{E. coli} to avoid an immune response can be used by alien \textit{E. coli} to infect a host, making avoidance of infection impossible for the host microbiome. Coincidently or not, many cases of serious illness originating in the human gut have been related to specific strains of \textit{E. coli}.

The results of the third article demonstrate that the human gut microbiome has characteristics of resilience. The definition of resilience in ecology is the capacity of the ecosystem to experience disturbance and recover to its original stage. Using the species network, species degree, and functional analysis, we demonstrated that some species (generalists) have an important role when the system is under pressure. However, how does it recover? Our model offers an explanation. Many specialist species, such as \textit{Clostridium}, produce endospores, which guarantee that they will survive as spores and return when the ecosystem is more favourable for their growth. This strategy is used widely in nature by a large number of species in addition to other survival strategies when an environment is unfavourable. Another characteristic of the gut microbiome is the large tail in its distribution, which indicates that even though it may be composed of many species, few species are actually highly abundant, and the majority have low abundance. Survival with very low abundance may be another bacterial strategy for surviving an unfavourable environment.

Bacterial filamentation is yet another interesting survival strategy that has been well documented but not well understood. Bacteria such as \textit{E. coli} can change shape under stress, forming long cells up to 50 times its normal size. An interesting discussion about how the morphologies of the bacteria change under stress has been published [121]. As has been discussed before, an increase in Gram-negative bacteria is observed when immune response increases, and Gram-negative bacteria contain bacterial LPS recognised by the immune system. \textit{Escherichia coli} do not filament if Toll-like receptor 4 is knocked out [122], suggesting that the receptor is responsible for activating the change in the bacteria morphology. This change in morphology helps bacteria survive the immune response specifically by evading Europhile phagocytes [122]. This phenomenon suggests a link between the gut microbiome and immune response.
The profiling of DNA abundance in the human gut microbiome was an important step toward a better understanding of the distribution of the microbiome in the gut and the dynamics of the gut microbiome. The Gut-Array also allowed us to profile RNA expression. Now that we know ‘who is there’, we want to know ‘what are they doing’. However, RNA analysis in an environmental system is complex. By using DNA, we analysed two main variables: the species present in a sample and the abundance of those species. DNA is considered very stable over time, even when the diet of an individual changes. However, RNA is not expected to be stable, and bacteria likely express distinct genes at different points in time. The first attempt to characterise the functional human gut microbiota occurred in 2011, when María Gosalbes and co-workers [123] used messenger RNA sequencing data from 10 healthy individuals to show that the composition of the active microbiota is more uniform than the taxonomic composition, and our results confirmed this observation. RNA analyses revealed that the variance of Firmicutes was greater than that of Bacteroidetes. This result differs from that in DNA and reflects the importance of Firmicutes in the functional gut microbiome. Analysis of the most abundant species expressed in each sample reveals a very interesting trend. The expression of species seems to be divided into three groups (similar to the division in DNA), in which the first group is dominated by *Clostridium* species and is characterised by high diversity. The second group expresses mostly *Eubacterium* species and has intermediate diversity. The third group is composed of *Ruminococcus* species with low diversity.

These three groups of highly expressed species could be interpreted as the enterostages of RNA in the gut microbiome. Surprisingly, Bacteroidetes are not the most expressed species, although they are the most abundant at the DNA level. The variance in the expression in our data is smaller than that in abundance, which reflects the results of other studies. Thus, RNA may tell a different story than DNA does. Whereas DNA results appear to be related to survival by allowing species to adapt to a stressful environment, RNA results may be related to the expression of proteins with important functions in the gut. Although DNA analysis did not reveal any species-specific association with disease, RNA expression shows more promising results. Examination of the expression of phyla revealed that Firmicutes are positively associated with TNF-α, which is involved in the regulation of immune cells. Deregu-lation of TNF-α production has been implicated in a variety of human diseases, including Alzheimer’s disease [124], cancer [125], and IBD [126]. Thus, this result supports the theory of dysbiosis and immune-response observed for the DNA data. Using genus expression, I observed that *Clostridium*, *Blautia*, and *Oribacterium* are positively associated with lean fat, whereas *Dorea*, *Ruminococcus*, and *Coprococcus* are positively associated with fat and body mass index. Moreover, the same trend was observed using higher taxonomic levels such as order. Specifically, Clostridiales was the main order displaying a positive association with obesity indicators, which agrees with results reported by Ruth Ley [48]. This preview of RNA results supports the DNA results found in the third paper in this thesis but adds another layer of complexity, improving our view of the relationship between the microbiome and its host.
Combining all of the results presented in this thesis shows how the scientific view of the gut microbiome has changed. This view is becoming more common and accepted by the scientific community, and a large number of related reviews and original studies have been published in the field as of 2012. It is important to remember that the view of the gut microbiome is always limited by the technology and information available at the time of a study. Therefore, I expect to see this view, including the models I created, to change and improve within the next few years. I believe that the overall observations in the model will stay, as they are based on results from many studies. Detailed information about precisely what is happening in each stage is likely to emerge soon as well.
Bibliography


