Metagenomic Systems Biology of the Human Microbiome

Bonde, Ida; Nielsen, Henrik Bjørn; Sicheritz-Pontén, Thomas

Publication date:
2014

Document Version
Peer reviewed version

Metagenomic Systems Biology of the Human Microbiome

PhD Thesis

Ida Bonde

February, 2014

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

"If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat."

Douglas Adams (1952-2001)

Preface

This PhD was funded by the Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark. The work was mainly carried out at Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark under the supervision of Professor Thomas Sicheritz-Pontén and Associate Professor H. Bjørn Nielsen.

Ida Bonde
Lyngby, February 2014

Contents

Abstract

The human microbiome is an integrated part of the human body, outnumbering the human cells by approximately a factor 10. These microorganisms are very important for human health, hence knowledge about this, "our other genome", has been growing rapidly in recent years. This is manly due to the advances in next generation sequencing, which has allowed for large-scale metagenomics studies of different niches of the human microbiota. Especially the gut microbiota has been studied intensively. However, most studies have been purely descriptive, thus there is still a lot to learn regarding the interplay between species in the microbiota and also between the host and the inhabiting microorganisms. Additionally, the non-bacterial part of the microbiota, which includes bacteriophages, plasmids and micro-eukaryotes, is not very well described.

In this thesis, metagenomics data from the human gut, nose and oral cavity has been analyzed. The central method has been a co-abundance clustering method, which separates genes from metagenomics data under the assumption that genes originating from the same DNA (e.g. a bacterial genome, a phage or a plasmid) will co-vary across samples. Thus, co-abundance gene groups (CAGs) are obtained, which represent bacterial genomes, phages, plasmid or other genetic elements in the system. The ability to reassemble the metagenome in this way opens up new possibilities for analyzing the functional potential of species in the microbiota as well as the interactions in the system. Applying the CAG clustering method to data from the human gut microbiome, we identified dependency-associations between plasmids, phages and clone-specific gene sets to their bacterial host. Connections between CRISPR-elements and phages were also observed. Additionally, the persistence of some bacterial species in the human gut could be predicted based on absence or presence of specific genetic modules.

Based on the same CAG clustering of the human gut microbiome data, the link between bile acid degradation of bacteria in the gut and obesity was investigated. There seemed to be a slight correlation between the two. However, this remains to be a hypothesis for further studies. Furthermore, the prevalence of the parasite *Blastocystis* in the human gut microbiome data was analyzed. This is the first time a metagenomics approach has been applied to this problem and the results were similar to previous *Blastocystis* prevalence studies. Moreover, it was found that individuals with a *Bacteroides*-driven enterotype were less prone to harbor the *Blastocystis* parasite.

Finally, the CAG clustering method was applied to metagenomics data from the human nose- and oral-cavity. It was concluded that this method needs further improvement in order for it to be directly transferable to other datasets.

In summary this thesis presents co-abundance gene groups (CAG) clustering as a valuable tool for analyzing human microbiome data. Furthermore, results based on this method regarding important topics in relation to the human gut microbiota are described, including the interplay between bacterial species and other genetic elements in the system, factors that might influence development of obesity and prevalence studies of eukaryotes. Studies of other areas of the human microbiome might also benefit from CAG based analyses once the tool has been optimized.

Dansk Resumé

Det humane mikrobiom er en integreret del af den menneskelige krop og overstiger det humane celle antal med cirka en faktor 10. Disse mikroorganismer er vigtige i forbindelse med human sundhed og viden omkring dette, "vores andet genom", er vokset hurtigt de seneste år. Dette er hovedsageligt på grund af udvikling indenfor "next generation" sekventering, hvilket har muliggjort metagenomics studier i stor skala af forskellige nicher af det humane mikrobiom. Specielt tarmmikrobiomet er blevet studeret intensivt. Dog har de fleste studier udelukkende været deskriptive, og således er der stadig meget at lære i forbindelse med samspil mellem arter i mikrobiomet og mellem værten og de tilstedeværende mikroorganismer. Desuden er den ikke bakterielle del af mikrobiomet ikke særlig godt beskrevet, hvilket inkluderer bakteriofager, plasmider og mikroeukaryoter.

I denne ph.d.-afhandling er metagenomics data fra den menneskelige tarm, mund og næse blevet analyseret. Den centrale metode har været en "co-abundance clustering" metode, der separerer gener under den antagelse at gener der stammer fra det samme DNA (f.eks. et bakterie genom, en bakteriofag eller et plasmid) vil kovarierer henover alle prøver. Således opnås "co-abundance gene groups" (CAGs), som repræsenterer bakterielle genomer, bakteriofager, plasmider eller andre genomiske elementer i systemet. Evnen til på denne måde at samle metagenomet igen åbner nye muligheder for at analysere de forskellige arters funktionelle potentiale såvel som interaktioner i systemet. Ved at anvende denne CAG clustering metode på data fra det humane tarmmikrobiom kunne vi identificere afhængighed mellem plasmider, bakteriofager og klonspecifikke gen sæt og deres bakterielle vært. Også sammenhænge mellem CRISPR-elementer og bakteriofager blev oberveret. Ydermere kunne persistensen af nogle bakterie arter i den menneskelige tarm forudsiges baseret på tilstedeværelsen af specifikke genetiske moduler.

Baseret på den samme CAG clustering af det humane tarmmikrobiom data, blev sammenhængen mellem nedbrydning af galdesyrer udført af tarmbakterier og fedme undersøgt. Der forekom at være en svag sammenhæng mellem disse to, dog forbliver dette en hypotese til videre studier. Endvidere blev forekomsten af parasitten *Blastocystis* i det humane tarmmikrobiom data analyseret. Dette er første gang en metagenomics fremgangsmåde er blevet anvendt på dette problem og resultaterne stemte overens med tidligere prævalens studier af *Blastocystis*. Desuden blev det observeret at personer der havde en *Bacteroides* dreven enterotype havde en mindre tilbøjelig til at være bærere af *Blastocystis* parasitten.

Afslutningsvis blev CAG clustering metoden anvendt på metagenomics data fra human næse og mund. Det kunne konkluderes at denne metode kræver yderligere forbedringer før den kan overføres direkte til andre datasæt.

Sammenfattet præsenterer denne ph.d-afhandling ”co-abundance gene groups” (CAG) som et værdifuldt værktøj til analyse af humant mikrobiom data. Desuden bliver resultater, baseret på denne metode, der omhandler vigtige emner i forbindelse med det humane tarmmikrobiom beskrevet, hvilket inkluderer samspillet mellem bakteriearter og andre genetiske elementer i systemet, faktorer der kan have en indflydelse på udvikling af fedme og prævalens studier af eukaryoter. Studier af andre områder af det humane mikrobiom vil formentlig også kunne have gavn af CAG baserede analyser når metoden er blevet optimeret.

Acknowledgments

The three years I have just ended with this thesis is a time I will look back on with a lot of happy memories (and also with some relief that I am now out on the other side, as I guess most PhD students feel). These memories are very closely linked to a lot of fantastic people who work at CBS and CFB.

First of all, I would like to thank my supervisors Thomas Sicheritz-Pontén and H. Bjørn Nielsen for taking a chance on me when I came without any bioinformatics experience. You have very different perspectives on things, which luckily complement each other very well. Thanks to Nikolaj Blom for supervision, encouragement and for putting in a great effort in shaping our group. Thanks also to Søren Brunak for ideas along the way and especially for getting me in contact with Thomas when this PhD position opened up. Thanks to the Novo Nordisk Foundations Center for Biosustainability for funding my project.

Thanks to the rest of the Metagenomics group, Asli, Josef, Bent, Henrik, Aviaja, Dhany, Damian, Agnieszka, Thomas NP, Jacob, Simon and Jens. I think all of you have lent me a hand in one way or another through this process and I am very grateful for that.

Now on to the people without whom no scientist would ever get anywhere. This is of cause the secretariat and the system administration. Thanks to Kristoffer, John, Hans Henrik and Peter. To Marlene, Dorthe, Annette, Martin, Nanna, Karina, Lone and the office aids. You make all of our lives much easier. Thanks also to Susanne, Susanne, Helle and Mads at CFB.

I owe a thanks to Elin, Khoa and Jacob for sparing me the pain of having to put on a lab coat. To Piotr for implementing the canopy clustering. To Josef, Asli and Aviaja for proofreading my thesis. Thanks to Anders, Claus and Bjørn for being great office mates. To Jens K. for numerous chats in the kitchen. To Tammi for many hours of conversation.

In addition to having had all of you as great colleagues I feel like many of you have become my friends and I hope to keep in touch. Thanks to the rest of you at CBS, I wish I had room to mention all of you by name.

Additionally, I would like to thank my dad and my sister for always being there. I would also like send a grateful thought to my mother, who left us too soon. Thanks to the rest of my family and to all my friends for your patience. I have a lot of catching up to do. Thanks to Pia and Niels (and Santos) for taking care of me while I was writing, I owe you about a months rent. Finally a special thanks to my boyfriend Thomas. I could write something very emotional here, but you would hate it, so just thank you.

Papers included in the thesis

- Henrik Bjørn Nielsen*, Mathieu Almeida*, Agnieszka S. Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian Plichta, Laurent Gautier, Anders Gorm Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, **Ida Bonde**, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo Bertalan Quintanilha dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer Sølvsten Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf Sommer Kaas, Sean Kennedy, Karsten Kristiansen, Jens Roat Kultima, Pierre Leonard, Florence Levenez, Ole Lund, Bouziane Moumen, Denis Le Paslier, Nicolas Pons, Oluf Pedersen, Edi Prifti, Junjie Qin, Jeroen Raes, Søren Sørensen, Julien Tap, Sebastian Tims, David Wayne Ussery, Takuji Yamada, MetaHIT consortium, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, S. Dusko Ehrlich. *Variable gene modules predict persistence of microbes in the human gut.* Manuscript submitted for publication.

- Lee O'Brien Andersen*, **Ida Bonde***, Henrik Bjørn Nielsen, Christen Rune Stensvold. *A Metagenomic Approach to Studying Intestinal Microbial Eukaryotes.* Manuscript submitted for publication.

* These authors contributed equally.

Introduction

Chapter 1

General Introduction

The field of metagenomics is a fairly new discipline within genomics. It is the study of communities of microorganisms based on their DNA. The idea of studying communities is not new, but the advances in next generation sequencing within the last decade have made it possible to study bacterial communities as a whole, instead of limiting research to small subsections, as earlier culture dependent studies have been [25, 34, 149].

Some of the most studied microbial niches are the ones that are found on/inside the human body, as this is of great interest in relation to health and disease. This community is referred to as the human microbiota. It has also been named our second genome, since bacterial cells outnumber our own by approximately a factor 10. In these years it is becoming more and more clear how big an impact, both positive and negative, these organisms have on us, their host [58, 69].

This thesis focuses on the human microbiome, more specifically the human gut microbiome, nose microbiome and oral microbiome, with an emphasis on the gut microbiome. The analyzed data is whole metagenome sequencing data originating from two of the largest metagenome studies, namely the European METAgenomics of the Human Intestinal Tract (MetaHIT) project [104] and the American Human Microbiome Project (HMP) [129].

Metagenomics relies on direct sequencing, meaning the DNA is extracted from the samples without cultivation or amplification. This DNA is then fragmented before sequencing, resulting in chaotic datasets consisting of short DNA fragments originating from a whole range of species [18, 132]. Thus, some form of structuring of the data, determining which fragments originate from the same DNA, greatly aids the downstream analysis of

the taxonomy and functions of the metagenome in question. This work is centered around such a clustering method, which was named co-abundance gene groups (CAG) clustering. It was first employed on the MetaHIT data with great success and subsequently applied to the HMP data.

Chapter 2

Metagenomics

Metagenomics has revolutionized the field of microbial ecology in that it enables the study of whole microbial communities directly from a sample, rather than through cultivation of single organisms and amplification of their DNA [132, 149]. Microscopic life forms are by far the most common on the planet. These include billions of bacteria, archaea, micro-eukaryotes and viruses, and they have a huge impact on the ecosystems that they inhabit. The study of microorganisms is naturally interesting because they can cause disease, but this is definitely not the only reason. Bacteria and other microorganisms are important players in maintaining stable ecosystems in all thinkable and unthinkable places [130, 108, 141, 155]. In health care, they are interesting because of the impact on human health and disease. Furthermore, pathogenic and beneficial microorganisms in livestock have gotten some attention, since healthy animals yield better and larger quantities of products and consequently larger profit [16]. A whole other angle is the utilization of bacteria and other microorganisms for industrial purposes, either as expression hosts or as a source of interesting proteins and metabolites [25, 97, 149].

Classical methods for studying microorganisms were based on cultivation, i.e. growing bacteria as single cultures on (artificial) media under laboratory/standardized conditions [79, 149]. This approach has its limitations as the unculturable part of the microbes cannot be investigated. How big this part actually is depends on who have done the estimate and what type of samples has been investigated, however the scientific community seems to agree that it is the majority [6, 18, 29, 80, 120]. Another drawback of culture dependent studies is that it is not possible to measure the relative abundance of the species in a biological niche, as the cultivation process will

introduce major biases, due to differences in growth rate on the artificial media. Furthermore, it is not feasible to study a whole community in detail in a laboratory setting, as there are too many factors that need to be measured simultaneously. To some extent these problems can be overcome by use of metagenomic methods [25, 79, 149].

Metagenomics is a fairly new addition to the omics fields of study. The term was first described in 1998 by Handelsman and Rodon [43, 69, 160]. It is the study of all the microorganisms in a sample from a biological niche, including unculturable ones, based on their genetic material [111]. Such a community can be referred to as the microbiota, while the full genetic potential of this is called the microbiome [58], although the term microbiome tends also to be used when describing a community and not only the genetic pool. The term metagenomics can be used both as (1) targeted metagenomics, which is based on amplification and sequencing of a phylogenetic marker. This marker is typically the 16S ribosomal RNA gene for bacteria and the 18S gene for micro-eukaryotes [149]. And (2) shotgun metagenomics, where all DNA from the sample is fragmented and analyzed without amplification [59]. However, not all sources define the 16S approach as metagenomics [111]. The first large scale shotgun metagenomics studies were conducted in 2004 [18, 138, 141]. Since then, the field of metagenomics has gained momentum, mainly because of the development of next generation sequencing techniques, which has made it possible to sequence massive amounts of DNA at a reduced cost. Targeted approaches only provide information regarding the taxonomy, whereas the shotgun metagenomics method enables analysis of both the taxonomical makeup of the community as well as the functional potential [25, 80, 126].

Within shotgun metagenomics there are two main approaches, function-based and sequence-based [18]. The first focuses on investigating the biological function of gene products by cloning DNA from a sample into expression vectors and observing the effect on the expression organism. The other method relies on the sequencing of extracted DNA from a sample. The function and taxonomy of the sequenced DNA can, to some extent, be found using databases of known genes and genomes. The focus of this thesis is on whole genome shotgun metagenomics using sequence-based methods. From hereon the term metagenomics will refer to this method, unless stated otherwise. A typical workflow for such an analysis can be seen in Figure 2.1.

Figure 2.1: General workflow for shotgun metagenomics.

During the relatively short lifetime of metagenomics, an enormous number of samples has been sequenced. A large part of these have been made publicly available through resources like NCBI [147], IMG/M [83, 82], GOLD [99] and MG-RAST [89]. Table 2.1 shows the available datasets on the IMG-M system. This is only a small part of the big collection of publicly accessible data, however it still includes 1,957 samples[1]. It is clear from the table that the main interests of metagenomic studies have been on aquatic and terrestrial environments as well as on the human microbiome. This thesis will be focusing on the human microbiome.

Table 2.1: Counts of the datasets available at the IMG/M resource in December 2013. Source: http://img.jgi.doe.gov/cgi-bin/m/main.cgi

Engineered	84	Environmental	1265	Host-associated	880
Bioremediation	20	Air	2	Annelida	5
Biotransformation	9	Aquatic	824	Arthropoda	53
Solid waste	25	Terrestrial	423	Birds	6
Unclassified	1	Unclassified	16	Human	753
Wastewater	29			Mammals	18
				Microbial	1
				Mollusca	9
				Plants	32
				Porifera	2
				Tunciata	1

[1] http://img.jgi.doe.gov/cgi-bin/m/main.cgi, December 19th 2013

2.1 Next Generation Sequencing

Although the first metagenomic studies were performed using Sanger sequencing, it was not until the development of next generation sequencing (NGS) methods, with their lower cost and higher throughput, that the field really expanded. 454 sequencing, or pyrosequencing, has been, and still is, used extensively for 16S analyses of metagenomes. Whole metagenome sequencing studies have also been performed using this technique, but Illumina (Solexa) sequencing with its very high throughput and low cost is now the preferred method, despite the short read length of around 100-250bp (or 2x100-250bp if run as paired end) [59, 114, 132]. These methods are in some cases also referred to as 2$^{\text{nd}}$ generation sequencing, as so called 3$^{\text{rd}}$ generation methods have arisen, with the most popular being the Pacific Bio (PacBio) system. This method yields very long reads in the range of 5,000-30,000bp[1] (according to the manufacturer and users in the community), but the number of reads is low. When first introduced to the market the error rate was high, but according to the manufacturer, this has improved markedly. At present PacBio is not a tool that is very widely used on its own for metagenomics studies, but the read length makes it useful for gap closing of assemblies for example from Illumina data [32]. Other methods are, and have been, on the market, but have not be as important for the metagenomics filed as the ones mentioned here [59, 132]. These include SOLiD and Ion Torrent sequencing, which we have used for a pilot study described later in this thesis.

2.1.1 Quality Control and Data Processing

The quality of raw sequencing data will, in most cases, need to be assessed. The standard way is by use of the Phred score. Sanger and newer Illumina sequencing machines output this score directly (Solexa and older Illumina machines use another format). For SOLiD and 454 platforms tools exist for extracting quality score in this format [19]. The quality score is given to each base denoting the probability of the base calling being wrong [33]. The Phred score is defined as:

$$Q_{Phred} = -10log_{10}P(error)$$

Based on the Phred score, reads with an overall bad quality can be removed and reads with low quality ends are trimmed. A cutoff of 20 is often used. Phred score 20 means a probability of the base being called wrong is 1 in 100. This corresponds to an accuracy of 99%. Another important step is to remove adapter sequences from the reads, as these are sequencing artifacts and not biologically meaningful. Additionally, read length, GC-content, over represented sequences and k-mers might need to be assessed depending on the data and the analysis that will be performed afterwards.

[1] http://www.pacificbiosciences.com

2.1.2 Mapping and Assembly

Mapping is a key step in many studies, in some cases to annotate the metagenome and in other cases in order to construct an abundance matrix (see Chapter 4). Mapping methods have to be both fast and memory efficient, due to the amount of data produced by the sequencing platforms and, in many cases, also the size of the target database. Different methods have been developed, the most widely used is the Burrows–Wheeler Transform algorithm, which is employed in SOAPv2 [72], Bowtie [64] and BWA. For more information on the algorithm please refer to the BWA paper by Li *et al.* [70].

In many cases, the cleaned reads will have to be assembled into contigs. The complexity of the data complicates this process considerably, usually resulting in short contigs and many reads that do not assemble [95, 149]. This is important to have in mind when analyzing the data. A more thorough description of this step is beyond the scope of this thesis as all assembly of data used for this work has been done prior to my analysis.

2.2 Annotation of Metagenomes

Annotating metagenomes is not a trivial task due to the data sizes, the complexity and the fragmentation of the data [25, 101]. The aim of the annotation is to define which species are present in the samples and determine the functional potential of the community. Figure 2.2 shows a flowchart of common steps for metagenomic data processing. Taxonomical and functional annotation are in some cases done directly on the sequencing reads after quality filtering, simply by mapping to reference genomes or genes [101]. In most cases, the reads are assembled into contigs on which gene calling is performed to identify the coding genes. Additionally, tRNA and rRNA genes can be located on contigs as well as non-coding segments like Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) motifs [101]. The predicted genes are annotated functionally and in some cases taxonomically. The most common approach to functional annotation is by homology search to databases. The preferred algorithm is BLAST [5], but also BLAT [57], HMMER[1] and UBLAST [30] are frequently used. In addition, searches based on motif or patterns can also be performed using a range of methods, many of which are incorporated in the InterProScan tool [157]. Other less widely applied methods deal with annotation of specific functions and knowledge regarding the genetic structure, such as operons, co-expression and pathway structure [25, 101]. Most of the methods that exist have been designed for single genome annotation or even analysis of single genes. Thus, the performance of these are lacking, both with regard to speed and in handling the large unknown proportion of the genes.

[1] www.hmmer.org/

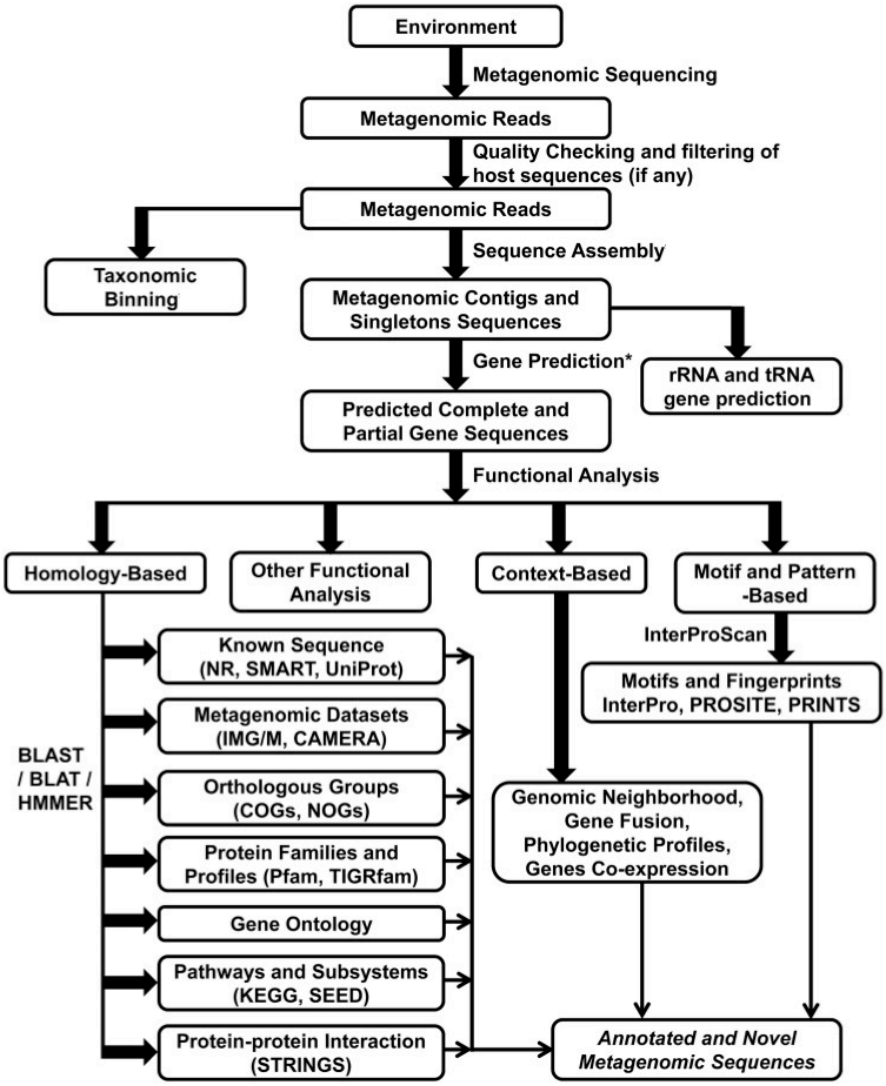


Figure 2.2: Flowchart showing the most common steps in data processing of metagenomic samples. Adapted from Prakash *et al.* [101]

A range of pipelines have been developed for metagenomic annotation, each having their own advantage. These integrate several methods to give an overall description of the metagenome under investigation [25, 101]. Among the web-based servers are IMG/M [81], MG-RAST [89], CAMERA [124], WebMGA [151] and CoMet [75]. Most of these have the advantage of access to a whole range of metagenomes that can be included in the analysis for comparison. However, web-based analyses are not feasible when the number of samples exceed a certain level. Annotation pipelines for local use include MEGAN4 [46], HUManN [2], RAMMCAP [73], Parallel-META [123] and SmashCommunity [7]. Additionally, the METAREP tool is available both as web-server and local tool [40]. Moreover, targeted pipelines like VIROME and VMGAP have been developed for analyzing the viral part of the metagenome, as this is particularly hard to annotate [77, 148].

The main limitation of sequence annotation is that, one way or another, all types of annotation are based on previous knowledge, so the quality of the annotation is dependent on efforts to better describe the microbial world both functionally and taxonomically [25, 97, 101].

Chapter 3

The Human Microbiota

The human microbiota is the collection of microbes found on and inside the human body, where microbial communities are formed in numerous niches. These are some of the most studied microbiotas on the planet, especially the gut and the oral cavity have gotten a lot of attention. In fact we are more microbe than human if calculated by cell count, as the microbe cells outnumber the human cells by a factor of 10 [69]. In addition to that, a massive amount of viruses are also found all over the human body as well as a range of micro-eukaryotes [69]. Looking at functional potential the microbiome coding genes outnumber those of the human host by a factor of 100 [58, 126].

Most metagenomic studies concerning the human microbiota have used targeted methods focusing on the 16S marker gene [58, 111]. This has given a fairly clear picture of which organisms are present in the different sites of the body. However, in resent years the field has been shifting towards whole genome sequencing, which adds the functional layer to the information. Comparing the taxonomy annotation and the functional annotation of human microbiomes across a population has shown that the microbiotas are generally more stable with respect to the functional composition compared to the species composition (see Figure 3.1), at least when looking at the most abundant functions as in Figure 3.1(b). Hence, bacteria with similar functions can substitute each other in the ecosystem, hereby retaining a stable system [130, 126]. Figure 3.1(a) shows the most abundant phyla in the human microbiota. It is worth noticing that each site seems to be dominated by one phylum.

The microbiota has many beneficial functions that influence the immune system, the metabolism and aids in keeping the general homeostasis of the

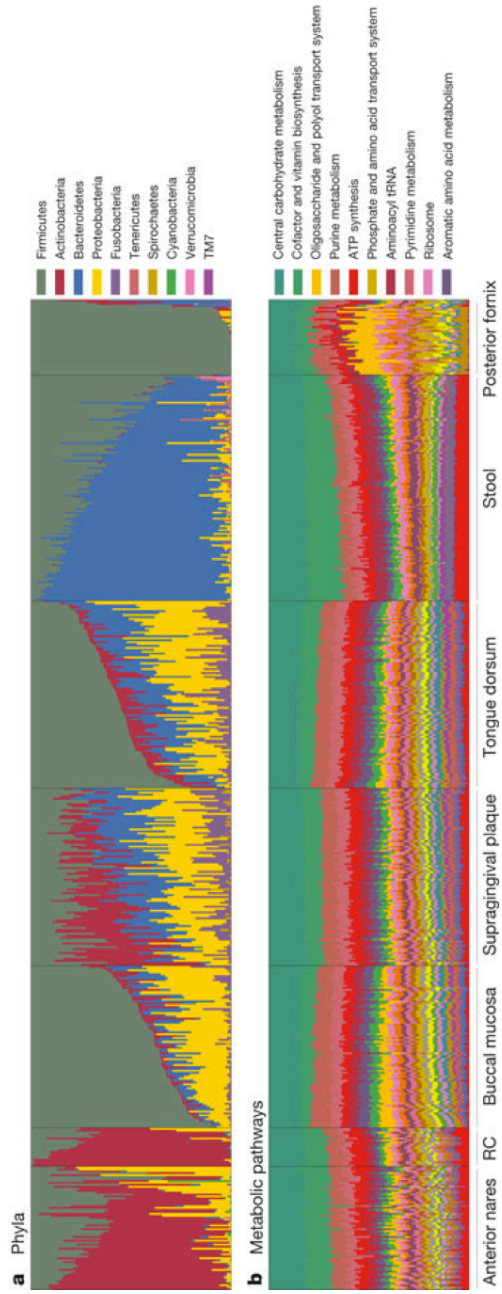


Figure 3.1: Phylum frequency and metabolic pathways in the human microbiota of healthy test subjects. The columns in both figures correspond to samples, each represented on both plots. (a) The phylum level annotation based on 16S data. (b) Metabolic pathway annotation based on whole genome sequencing. Seven sites are included: anterior nares (nose), RC (retroauricular crease -skin behind ear), buccal mucosa (oral soft tissue), supragingival plaque (teeth), tongue dorsum (oral soft tissue), stool (gut) and posterior fornix (vagina). Reprinted from The Human Microbiome Consortium [130].

body [58, 69, 111, 126]. Thus, humans and the human microbiota have co-evolved into a symbiotic state. The human body and the diet provide nutrition for the microorganisms, which in turn produce products that are essential for the human host, such as short chained fatty acids (acetate, butyrate and propionate) and vitamins. Furthermore, they aid in degradation of indigestible foods [69, 111, 126]. Accordingly, our microbiota is very important for a healthy body. Some of the health promoting properties of bacteria have been known for many years, like bacterial cultures in yogurt, and with the enormous focus on the human microbiota, we are getting a much better understanding of what characterizes a healthy microbiota [93].

Though more and more evidence emerge regarding the health benefit of the microorganisms we harbor, there are also studies showing that an imbalance, called dysbiosis, in the system might be involved in a whole range of illnesses including obesity, infant colicky, arthritis, type 2 diabetes, chronic inflammations, allergy and even mental disorders like autism, among others. These are all complex diseases and we still do not have the full picture of how microorganisms contribute to these phenotypes [31, 58, 94, 106, 146].

The focus of this thesis is primarily on the human gut microbiota and this will be introduced below. I have also in one project worked with data from the human nose and mouth. A brief introduction to these systems can be found in the chapter concerning that project (Chapter 8).

3.1 Human Gut Microbiota

The intestinal tract harbors the most diverse microbiota associated with the human body and it is competing with the oral microbiota for being the most studied human microbiota [17, 53, 126].

Each person harbors approximately 160 bacterial species [104]. Quite distinct variations can be seen on species level between individuals. However, on phylum level the diversity between people is fairly low [111]. Studies have shown that the most abundant phyla in the gut are Bacteroidetes and Firmicutes, constituting over 90% of the total bacterial pool [104]. Most gut bacteria are obligate anaerobes [111, 140], which is necessary in the oxygen poor environment of the gastrointestinal tract.

The gut microbiome does not seem to differ markedly between people of different nationalities. Rather, the gut microbiomes can be separated into so called enterotypes [69]. Although, the presence of enterotypes are highly debated in the scientific community [61]. The first report of enterotypes stated that it was possible to divide the test population into tree enterotypes, driven by *Prevotella*, *Bacteroides* and *Ruminococcus*, respectively [8]. However, further studies have shown that there is less support for the *Ruminococcus*

enterotype, and that it should be merged with the *Bacteroides* enterotype cluster [150]. There appears to be a connection between diet and at least the two well defined enterotypes. Hence, people with a diet high in animal fat and protein are more likely to have a *Bacteroides* driven enterotype and high intake of carbohydrates seem to correlate with the *Prevotella* enterotype. However, short term diet changes does not affect the enterotype [69].

One of the reasons for the immense interest in the human gut microbiota is the impact of these microorganisms on human health. The gut microbiota has been linked to inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), obesity and diabetes [26, 69] among others. Patients suffering from these diseases all seem to have a dysbiotic intestinal microbiota. This is an imbalance in the system, which otherwise, once matured, is quite stable unless it is disturbed. Factors like antibiotic treatments, which kill both the pathogenic organisms as well as the commensals, can cause dysbiosis. It can also be brought about by pathogenic bacteria, an example is *Clostridium difficile* infections. Western-diet, i.e. high fat and high sugar content, also seems to have undesired effects on the balance of the gut microbiota [63, 69, 91]. Whether the unhealthy microbiota is the cause or the effect of the illnesses observed is still unclear [56]. Despite the uncertainties that still exist regarding the health effect of the intestinal microorganisms, there seems to be a consensus about high richness of species and higher Bacteroidetes versus Firmicutes ratio being important for a healthy gut microbiota [31, 69, 111, 140].

It is important to mention that our microbial inhabitants of the gastrointestinal tract also perform very important functions. These include protection from pathogens, aiding in degradation of food, development of the immune system and production of short chain fatty acids and vitamins, just to name a few [24, 68, 69, 144]. Presently, there is a great interest in determining treatments that can restore a dysbiotic gut system to a healthy balance. Methods like functional foods such as probiotics, prebiotics and synbiotics have been used [140]. A rather new approach is to use fecal transplants, which among others have been successfully used for treating illnesses caused by *Clostridium difficile* infections and to combat the metabolic syndrome, which is related to various phenotypes, like type 2 diabetes, obesity and heart disease [55, 140, 142]. In time, a more controlled manipulation of the microbiome should be possible, which is somewhere between the controlled low complexity culture probiotics and the rather uncontrolled community infusion performed when using fecal transplants [91, 140]. Thus, introducing a desired function into the system by constructing a network of species that cooperate.

Although it is evident that the gut microbiota has a great influence on the human body, the interplay between host and microbiota and the driving forces of their co-evolution are still not very well described [69, 98]. In order

to study this we need a broader test population, preferably including remote populations. These people have a very different diet, which is perhaps more "ancient" than the western diet. This will hopefully be reflected in the microbiota. Accordingly, these individuals can function as a form of out group when studying diseases related to the gut microbiota [69]. A better understanding of the gut microbiome might aid in the development of personalized medicine, as the microbiome makeup might influence response to treatment [69, 98]. Furthermore, there is a need for biomarkers to detect various diseases. Even though several studies have set out with exactly that aim, no adequate markers have been discovered by studying the gut microbiota [69]. However, some species might be indicative of a low richness microbiota, thus indicating the risk of obesity. These species might be targets for biomarker development [66].

In addition to the cellular proportion of the gut microbiota, there are a vast amount of viruses inhabiting the intestinal tract, including bacteriophages, most of which are unknown [90]. These are thought to have a very important role in stabilizing the microbiota by killing bacteria. The virome seems to be much more individual than the bacterial part of the microbiome and it is very sensitive to changes in diet [69, 90].

The bacteria, viruses and micro-eukaryotes inhabiting the gastrointestinal tract constitutes an ecosystem in which each member has a part to play. Until now, most studies have been focused on determining who is there and what each individual contributes in terms of functional potential [24]. However, only a few studies concern the interplay of species, and the variability within species as well as the mechanisms driving persistence in the gut [107, 117, 135]. These questions are what we are addressing by using the co-abundance gene groups (CAG) clustering method on stool samples. This will be described in Chapter 5. This is a first step towards an understanding of the system that hopefully will aid in answering some of the pending questions that I have described in this overview.

Chapter 4

Metagenomic Binning

The purpose of metagenomic binning is to try and recreate the community under investigation, since in the sequencing process the data is fragmented and even with the best assembly methods it is not possible to recreate full genomes of the entire content of the sample from the short reads. In this context, metagenomic binning is to define which sequences originate from the same organism and cluster these together. Binning methods exist for data reduction by removing redundant sequences [132], but this is not the focus here. This chapter is devoted to give an overview of the existing methods for binning metagenomics data as well as introducing the method we have employed to construct co-abundance gene groups. It will only cover methods for shotgun metagenomics.

4.1 Existing Methods

Different approaches exist that rely on different information in order to cluster the data. Numerous methods have been developed, some of which will be described here.

4.1.1 Sequence Homology Based Binning

Homology based methods are based on similarity of the sequences (genes, contigs or reads) to known species or proteins. Thus, these are dependent on previous knowledge stored in a database. In this way, the data is binned based on the taxonomic or functional annotation [114, 149].

The LCA (Last Common Ancestor) method is a widely used approach which is implemented, in different forms, in SOrt-ITEMS [92], MTR [41] and CARMA3 [38] among others. These binning tools are based on alignment of reads or genes to known genomes and then binning the sequences according to the hits. A cutoff is set to define what level in the taxonomical tree to assign each sequence to. CARMA3 can also be used on data aligned to a protein database, thus giving a functional binning of the data rather than a taxonomical [38]. Others have also applied this method using databases like Pfam [103] or COG [127] or whole protein databases like UniProt [131] and NCBI nr [23, 102, 134]. In some cases, a binning method is applied to reduce the redundancy in the data and remove the shorter sequences before the actual annotation clustering [23, 38]. A fast and more precise method is the MetaPhiAn [116], which bin reads based on known clade-specific genes from reference genomes. The speed gain comes primarily from the reduced search database [59].

These types of methods work best when dealing with very well described ecosystems, as they are heavily dependent on already known species and proteins. However, in most cases metagenomics data contain sequences from unknown bacterial species. Additionally, existing databases are still biased towards model organisms and pathogens, even with the great efforts going on to fill in the missing gaps, like the Human Microbiome Project and the Earth Microbiome Project [39, 129, 114]. Even if an acceptable annotation is obtained for the bacterial part of the microbiome, the viruses and plasmids will cause problems, as these are generally very poorly annotated. The method cannot account for horizontal gene transfer and clonal differences. The methods based on alignment using BLAST or HMMER are computationally very challenging, especially if done on read level [85, 116].

4.1.2 Composition Based Binning

Another way to manage the problem is by clustering the data based on sequence composition. These methods are typically based on GC-content or k-mer abundance, which is conserved between closely related organisms. Additionally, codon usage has been used for binning [132, 149]. Many of these have the advantage of being independent of reference databases.

Most of the composition based methods rely on k-mer frequency. The method TETRA is based on tetranucleotide composition [128]. This has also been improved by combining it with codon-usage [139]. The PhyloPythia bin the sequences based on their k-mer frequency, for which it relies on a training set of genomes [88]. This works best when the data resembles the training set and with sequences longer than 1kb. S-GSOM also bin the sequences based on an analysis of the k-mer frequencies, but only in the region around 16S rRNA genes [15]. PCAHIER, TACOA and Phymm are additional methods

that cluster data based on k-mer frequencies [28, 159]. However, Phymm uses complete bacterial genomes for training, hence it is not independent of databases. It has the advantage of performing well on short sequences and the method was improved by combining it with a BLAST approach. This was named PhymmBL [12].

Generally, composition based methods are more computationally efficient than the homology based methods. However, the accuracy is often low, especially when dealing with short sequences, such as sequencing reads [116, 132].

4.1.3 Co-Abundance Based Binning

A third and, to my knowledge, rather new form of binning is based on co-abundance. Basically these methods work under the assumption that DNA sequences that originate from the same DNA will co-vary across samples. The method employed in the analyses described in this thesis is based on this principal. Thus, this will be explained in more detail than the methods mentioned above.

The input to the clustering algorithm is some form of abundance matrix. A simplified version of such a matrix is shown in Figure 4.1. It contains the abundance of each sequence in each sample. In this case, the sequences are genes, but they could also be contigs. The clustering bins genes that have a similar abundance profile across all samples. Hence, in Figure 4.1, gene 1 and 2 will be binned together in a cluster and gene 3, 4 and 6 in another.

This approach has been applied in two studies on the human gut microbiota in relation to type 2 diabetes. In these, the clusters were named metagenomic clusters (MGS) and metagenomic linkage groups (MLG) respectively [54, 105]. The aim of these studies was to investigate if any MGS or MLG were associated to diabetic traits. Using this method, they were able to better pinpoint the organisms responsible for the functional changes they observed in the metagenomes.

In another study, this method was applied to one waste water sample, which was sequenced twice using different DNA extraction methods, thus resulting in two samples with different abundance measures of the genes. In this case, other factors like tetranucleotide frequency and GC-content were also taken into account in addition to the co-abundance clustering. From the resulting bins, they were able to assemble 31 bacterial genomes, including four from the possibly new bacterial phylum TM7, thus markedly improving the assembly [4].

Carr *et al.* [14] performed a similar clustering, however, they chose a slightly different approach as the genes for constructing the abundance matrix were obtained from the KEGG database, whereas for the other studies

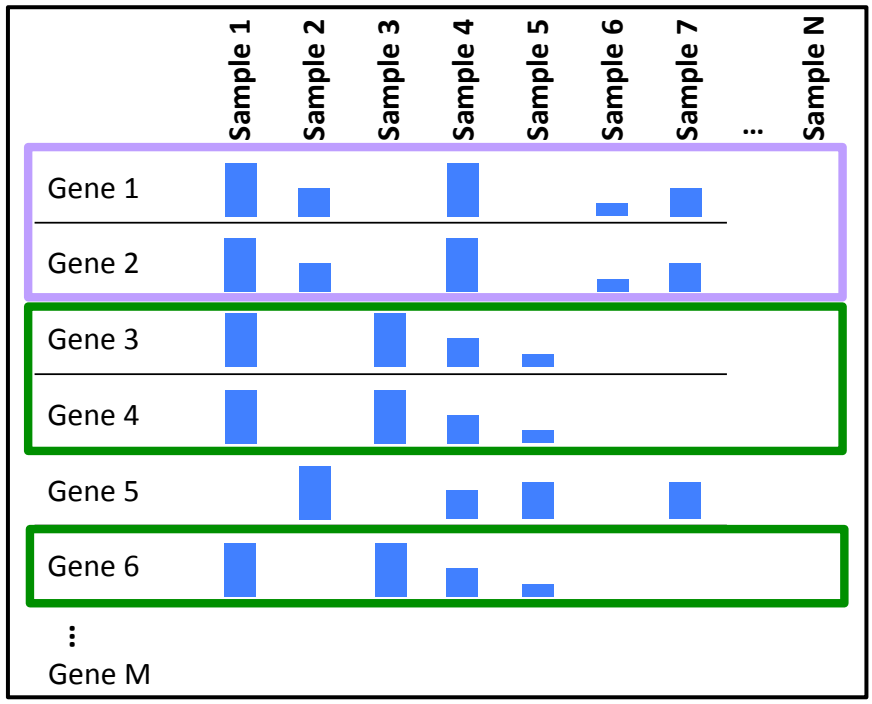


Figure 4.1: Example of an abundance matrix with N samples and M genes. The blue bars represent the abundance of each gene in each sample. Gene 1 and 2 have similar abundances, thus they are clustered together, as are gene 3, 4 and 6.

it was genes or scaffolds originating from assemblies of the metagenomic samples that were used. This makes this method dependent on database information.

The co-abundance approach seems to yield better results than any of the other methods for metagenomic binning in terms of accuracy [4, 14, 54, 105]. Further development of these tools and integration with composition based or homology based methods will lead to better annotation of metagenomes.

4.2 Co-Abundance Gene Groups

Co-abundance gene groups or CAGs were generated using the above described co-variance principal. Figure 4.2 depicts a work flow describing how we performed the analysis. The first step is to clean up the sequencing data

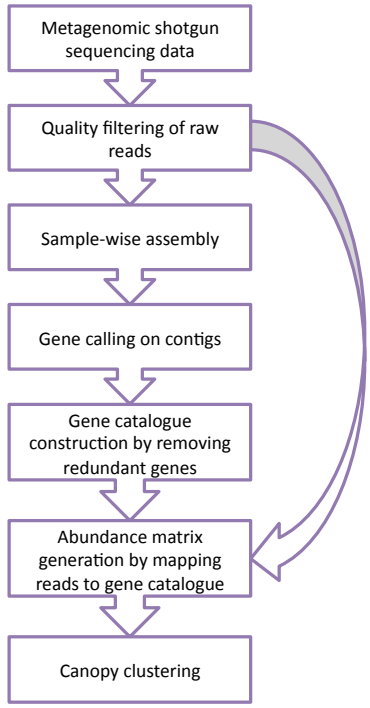


Figure 4.2: Work flow for constructing CAGs from metagenomic data.

as described in Section 2.1. The reads are used as input for assembly. Each sample is assembled on its own. Genes are predicted on the resulting contigs using a gene caller that can handle metagenomic data, e.g. MetaProdigal [47] or MetaGeneMark [161]. All genes are pooled and redundant genes are removed by applying a homology based clustering program, this could be CD-HIT-EST [74] or UCLUST [30]. This produces a gene catalogue to which all the trimmed reads are mapped in order to construct the abundance matrix. Thus, the raw count matrix contains counts for all genes in all samples. These will then have to be scaled to account for different gene lengths and then either scaled or downsized in order to correct for varying sequencing depth between samples, as comparing samples of very unequal quality, in terms of sequencing depth, cause false correlations in the data. The level of downsizing depends on the samples at hand. The processed abundance matrix serves as the input for the co-abundance gene groups clustering. The clustering could likely also be used directly on contigs.

The CAG clustering is based on the canopy clustering algorithm. This was originally intended as a pre-sorting of data before an actual clustering and, as such, it is developed for speed [87]. We use Pearson correlation to

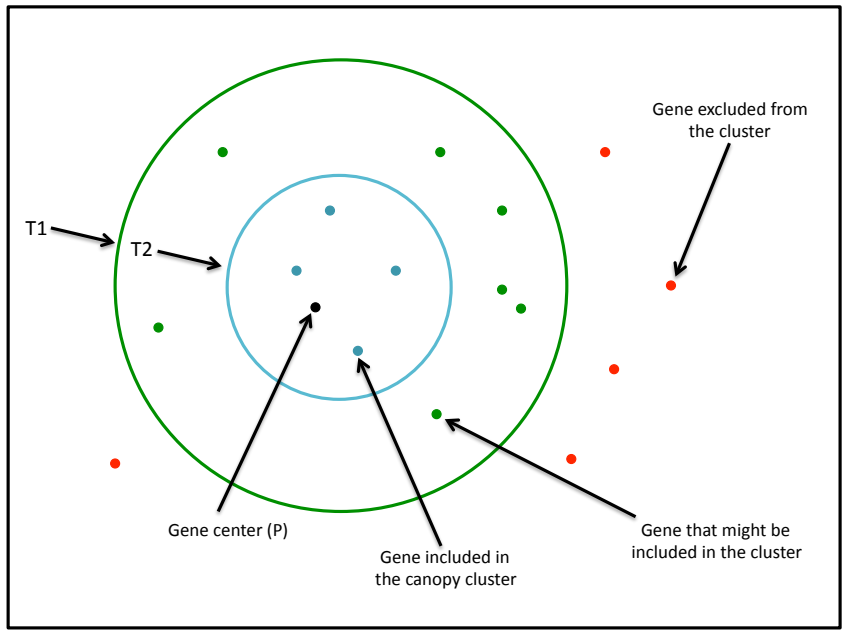


Figure 4.3: Illustration of the canopy clustering method. A seed gene is picked randomly as an initial cluster center (P). The genes that are within the inner circle with the limit T2 are members of the canopy. Genes between the circles could possible be part of the cluster. The center is recalculated based on the genes inside the T2 circle and the canopy is recalculated based on the genes inside the T1 circle, as some of the genes between the circles might now fall inside the T2 circle. All genes outside T1 are not included in the recalculation. Adapted from http://www.shahuwang.com/wp-content/uploads/2012/08/canopy.png, January 2014.

calculate the abundance correlation between genes. In Figure 4.3 is an illustration of the algorithm. A center is picked randomly and the correlations to all other genes in the data are calculated. All genes within the limit T2 (see Figure 4.3) are included in the cluster. All that are outside T2 but inside T1 (also in Figure 4.3) can possible be part of the canopy. A recalculation of the center is performed and the cluster is recalculated for all genes within the T1 limit. This is called canopy walks and is done until the optimal center is found (or a defined number of times) at this point all genes within the T2 limit are binned. Then a new center is picked from the gene pool, excluding the binned genes, and the process starts over. This is done until all genes are clustered.

This binning method was first applied to the metagenomics data provided by the MetaHIT project. A short description of this project and the paper written on the analysis can be found in Chapter 5. Subsequently, the CAG clustering approach was attempted on the HMP oral and nose data. This is described in Chapter 8.

Projects

Chapter 5

Metagenomic analysis of the Human Gut Microbiota

5.1 Introduction

Co-abundance gene group (CAG) clustering is a highly valuable tool for studying metagenomes. It enables investigation of functions of specific organisms and interaction between elements (bacterial species, phages and plasmids etc.) in the system, which is a great improvement over the purely descriptive studies that have been the norm [24]. Thus, the first CAG clustering was performed on sequencing data from stool samples collected under the MetaHIT project. The fecal samples serve as a proxy for the human gastrointestinal tract. This, and the following two chapters, describe the results obtained from analyzing CAGs identified in the human gut microbiome data. The importance of studying the human gut microbiota has already been introduced in Chapter 3, hence in this chapter I will only describe the MetaHIT project and include the paper concerning the CAG clustering performed on the MetaHIT data.

5.1.1 The MetaHIT Project

MetaHIT is an abbreviation for METAgenomics of the Human Intestinal Tract. It was a large project including 13 academic and industrial partners from 8 countries. The funding was obtained from the European Commission for four years of work running between 2008 and 2012[1].

[1] www.metahit.eu/, February 2014

The aim of the study was to better understand the human gut microbiota in relation to human health. Accordingly, the first paper published described the human gut microbiota of 124 individuals with varying BMI from lean to obese and patients suffering from ulcerative colitis and Crohn's disease [104]. This was based on Illumina deep sequencing of fecal samples and it was the first study of this magnitude to be published on the human gut microbiome. The importance of the work can be illustrated by the 1,236 citations this paper has gotten according to Web of Science as of February 2014. The dataset included ~3.3 million non-redundant genes in total and approximately 1,100 bacterial species. Based on the gene count for each sample it was estimated that every individual harbored ~160 species, most of which were shared between all patients.

Two other studies have been published by the MetaHIT consortium. The first being the controversial paper describing how the MetaHIT samples could be separated into enterotypes, which has already been described in Chapter 3 [8]. The other study describes the difference in the microbiota of lean and obese individuals [66]. However, this is not performed on the original 124 samples but on 292 samples from individuals ranging in BMI from lean to obese. They found that the microbiome in lean and obese differ in relation to richness and species composition, but the difference is more significant if looking at high gene count versus low gene count of the microbiome. This could be due to differences in the severity of the obesity phenotype in the patients. Thus, individuals with a low gene count gut microbiota might have a higher risk of developing obesity related diseases than individuals with a high gene count microbiome. The sequencing data of the first 124 samples was made publicly available and has been used in many other studies. One of these concerned the metabolic potential of the microbiota and how this is involved in interactions with the host [48]. Another study compares the human and guinea pig gut microbiotas, which is interesting because guinea pigs are used as model organisms [44]. Others have included the samples in larger scale comparison studies [35, 51, 62, 113, 125].

The MetaHIT sequencing effort did not end with the 124 samples. In the present study we include 396 deeply sequenced samples from the MetaHIT collection. This data is clustered as described in Chapter 4. The analysis of the resulting CAG clustering is described in the manuscript included in the next section. Supplementary information for the paper can be found in Appendix A (supplementary methods) and Appendix B (other supplementary information).

5.2 Manuscript - Variable gene modules predict persistence of microbes in the human gut

Variable gene modules predict persistence of microbes in the human gut

Authors: Henrik Bjørn Nielsen[1,2]†, Mathieu Almeida[3,4]†, Agnieszka S. Juncker[1,2], Simon Rasmussen[1], Junhua Li[5,6,7], Shinichi Sunagawa[8], Damian Plichta[1], Laurent Gautier[1], Anders Gorm Pedersen[1], Emmanuelle Le Chatelier[3,4], Eric Pelletier[9,10,11], Ida Bonde[1,2], Trine Nielsen[12], Chaysavanh Manichanh[13], Manimozhiyan Arumugam[8], Jean-Michel Batto[3,4], Marcelo Bertalan Quintanilha dos Santos[1], Nikolaj Blom[2], Natalia Borruel[13], Kristoffer Sølvsten Burgdorf[12], Fouad Boumezbeur[3,4], Francesc Casellas[13], Joël Doré[3,4], Piotr Dworzynski[1], Francisco Guarner[13], Torben Hansen[12,14], Falk Hildebrand[15,16], Rolf Sommer Kaas[17], Sean Kennedy[3,4], Karsten Kristiansen[18], Jens Roat Kultima[8], Pierre Leonard[3], Florence Levenez[3,4], Ole Lund[1], Bouziane Moumen[3,4], Denis Le Paslier[9,10,11], Nicolas Pons[3,4], Oluf Pedersen[12,19,20,21], Edi Prifti[3,4], Junjie Qin[5,6], Jeroen Raes[15,16], Søren Sørensen[22], Julien Tap[8], Sebastian Tims[23], David Wayne Ussery[1], Takuji Yamada[8,24], MetaHIT consortium, Pierre Renault[3], Thomas Sicheritz-Ponten[1,2], Peer Bork[8], Jun Wang[6,12,18,25], Søren Brunak[1,2*], S. Dusko Ehrlich[3,4*]

[1] Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark.
[2] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark.
[3] Institut National de la Recherche Agronomique, UMR 14121 MICALIS, 78350 Jouy en Josas, France.
[4] Institut National de la Recherche Agronomique, US 1367 MGP, 78350 Jouy en Josas, France.
[5] BGI Hong Kong Research Institute, Hong Kong, China.
[6] BGI-Shenzhen, Shenzhen 518083, China.
[7] School of Bioscience and Biotechnology, South China University of Technology, Guangzhou 510006, China.
[8] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
[9] Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Institut de Génomique, Genoscope, 91000 Évry, France.
[10] Centre National de la Recherche Scientifique, UMR-8030, 91000 Évry, France.
[11] Université d'Évry Val d'Essonne, UMR-8030, 91000 Évry, France.
[12] The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark.
[13] Digestive System Research Unit, University Hospital Vall d'Hebron, Ciberehd, 08035 Barcelona, Spain.
[14] Faculty of Health Sciences, University of Southern Denmark, DK-5000 Odense, Denmark.
[15] Department of Structural Biology, VIB, Brussels, Belgium.
[16] Department of Bioscience Engineering, Vrije Universiteit Brussel, Brussels, Belgium.
[17] National Food Institute, Division for Epidemiology and Microbial Genomics, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark.
[18] Department of Biology, Ole Maaløes Vej 5, University of Copenhagen, DK-2200 Copenhagen, Denmark;
[19] Hagedorn Research Institute, DK-2820 Gentofte Denmark.
[20] Institute of Biomedical Science, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark.
[21] Faculty of Health Sciences, Aarhus University, DK-8000 Aarhus, Denmark.
[22] Section of Microbiology, Department of Biology, University of Copenhagen, DK 2100 Copenhagen, Denmark.
[23] Laboratory of Microbiology, Wageningen University, The Netherlands.
[24] Department of Biological Information, Tokyo Institute of Technology, Yokohama 226-850, Japan.
[25] King Abdulaziz University, Jeddah, Saudi Arabia.

*Correspondence to: E-mail: brunak@cbs.dtu.dk or dusko.ehrlich@jouy.inra.fr
†These authors contributed equally to this work.

ABSTRACT

The genetic diversity of the human gut microbiome includes numerous plasmids, phages and clonal differences and extends far beyond what is covered by reference genomes. Here we present the first microbiome-wide analysis of dependency-associations that link plasmids, phages and clone-specific genetic elements, to their bacterial hosts. The dependency-associations reveal specific CRISPR-phage relationships and longitudinal samplings show that the occurrence of some associated genetic elements predicts persistence probabilities of the microbes that carry these. The mapping of the dependency-association is made possible by an exhaustive co-abundance segregation of the gut metagenome into 7,381 highly correlated co-abundance gene groups (CAGs) including 741 metagenomic-species (MGS) across 396 human faecal samples.

INTRODUCTION

The human microbiome has recently gathered substantial attention because of its importance for human health and disease[1]. The gut microbiome is among the most complex microbial communities studied, with an estimated number of common microbial species about one thousand across humanity[2]. These microbes are generally believed to have co-evolved with the human host[3] and yet the composition of the microbiota varies considerably among human individuals[4].

In adults the species composition is relatively resilient, though changes over longer time periods have been reported. Genetic and environmental factors including diet and, not surprisingly, the use of antibiotics have been shown to influence the microbiota[5,6]. Finally, both the innate and the adaptive immune systems are likely to be important factors, though their impact is not fully known.

These factors are likely to shape the genetic makeup of the microbes that persist in the gut. Understanding of such genetic adaptations may be critical for modulating the microbial community. With the exception of a few well-studied species, the genetic makeup of the gut microbes is at best known from a small number of reference genomes. Consequently, the genetic variation of the microbial species is largely unknown. We recently described single nucleotide polymorphisms across the gut microbiome[7] and showed that comparison of genomes from different isolates of the same species indicates substantial differences at the level of gene composition[8]. Examples of such heterogeneity are genetic islands, bacteriophages and plasmids, but also clonal differences. While plasmids and phages have been identified in the human gut microbiome they have with a few exceptions[9–12] not been associated to specific microbial host species within the community. A first step towards unfolding such associations is a *de novo* description and profiling of the microbial species and their clone-specific and mobile genetic elements, across gut samples from a series of individuals.

Progress in assembly from metagenomics data utilizing sequence abundance and composition to segregate species, has resulted in genome assembly from several ecosystems[13,14]. Recently partial assembly of 47 microbial genomes based on co-abundance gene aggregation, has been reported from human faecal metagenomics data[1,15].

Here we exhaustively bin the genes of a new 3.9M gene catalogue from 396 deep sequenced human faecal samples by co-abundance clustering. In this way we aggregate 40% of the gene catalogue into 7,381 highly correlated co-abundance gene groups (CAGs), which range in size from small groups of only a few genes to what corresponds to complete prokaryote genomes. The latter we denote metagenomic-species (MGS). The smaller CAGs are likely to represent an ensemble of genetic variations associated with cognate species, including plasmids, phages, genomics islands and clonal differences. We link numerous smaller CAGs to MGS by determining dependency-associations between these. We also show that the persistence probabilities differ between bacterial populations with or without specific dependency-associated CAGs.

RESULTS

Exhaustive co-abundance gene segregation

In this study we use deep sequencing data from 396 human stool samples from Spain and Denmark, including 124 samples from a previous study[16] (see Methods and Supplementary Data 1 for details). 77 of the Spanish individuals were sampled twice, with, on average, six months between the samplings. The sequence reads were assembled separately for each sample and merged to form a non-redundant gene catalogue of 3.9M microbial genes (Supplementary Fig. 1).

An exhaustive and unsupervised co-abundance segregation of the entire gene catalogue was deployed to identify all CAGs (Supplementary Fig. 2 and 3). In brief, the profile of a randomly picked seed gene was used to capture groups of highly abundance-correlated genes (canopies, with Pearson correlation coefficient (PCC) > 0.9). The canopy profile was determined as the median abundance profile of the comprised genes and was used for recapturing the canopy until the profile stabilized. This process was iterated until all genes were assigned to a canopy. Canopies, with at least three genes and where the canopy abundance signal from any three samples constituted less than 90% of the total signal across all samples, were identified as CAGs (for additional details see Methods). The approach is ultra-fast because it is simple and because the large canopies tend to be extracted quickly and thereby reduces the computational cost of subsequent canopies.

Some 1.53 million genes (representing 68% of the mapped sequence reads) were assigned to 7,381 CAGs, ranging in size from 3 to 6,319 genes. Interestingly, the size distribution of the groups, in terms of genes contained, was bimodal with peaks around 50 genes and 1,700 genes, respectively (Fig. 1A). The 741 largest CAGs with more than 700 genes correspond approximately to complete genomes of bacteria or archaea, in terms of the number of genes they contain (Supplementary Fig. 4). Furthermore, the genes contained in these CAGs were highly consistent in base composition, had highly correlated abundance profiles in an independent set of 115 samples[17] and had consistent taxonomical annotation (Supplementary Fig. 5, 6 and 7C; Supplementary information; and Supplementary Data 2). Thus, for 115 of these more than 95% of the taxonomically annotated genes were similar to a reference genome from a single species. Finally, 238 of these could be assembled to the high quality draft genome assembly standard of the Human Microbiome Project (for details see Methods; Supplementary Information; Supplementary Fig. 7-10; and Supplementary Data 3 and 4). We therefore refer to these as metagenomic-species (MGS) or species.

Because 19 individuals consumed a defined fermented milk product containing the previously sequenced *Bifidobacterium animalis subsp. lactis CNCM I-2494*[18], this species may serve as a benchmark for the co-abundance clustering. Thus, although MGS:337 only constitute on average 0.3% of the 19 samples, it captures 95% of the *B. animalis* reference genes and the MGS augmented genome assembly covered 95% of the reference genome with 99.9% identity (Supplementary Fig. 10). Furthermore, sub-sampling of the data demonstrates that the *B. animalis* MGS can be segregated using as little as 700K sequence reads per sample or from a much smaller sample set consisting of only 18 samples (for details see Supplementary Fig. 11 and 12).

Functional characterization of small CAGs

While the majority of the abundance co-varying genes are contained in MGS, the 6,640 smaller CAGs with less than 700 genes, and on average 44 genes, show equally tightly correlated abundance profiles. With 848 small CAGs enriched for proteins characteristic for bacteriophages[10] or with consistent phage taxonomy[19] (for details see Supplementary Data and Methods), this is the most common type of annotation enrichment among the small CAGs. On average 113 (± 37) phage-like CAGs could be identified per sample. Although, bacteriophage taxonomy is relatively poor we observed consistent species or family level taxonomical annotation in 35 and 172 phage-like CAGs, respectively. In accordance with the many observed phage-like CAGs, transposase, integrase and recombinase encoding genes were primarily enriched in the smaller CAGs (Fig. 1B).

Another class of functions that were found primarily enriched in smaller groups may be described as functions that are important for biotic interactions. These include Clustered Regulatory Interspaced Short Palindromic Repeat (CRISPR) associated genes, which function in Bacteria and Archaea as a sequence dependent adaptive immune system directed against alien DNA[20]. In addition to core CRISPR associated genes, several CAGs were enriched for specific subtypes of these genes (Supplementary Fig. 13). Similarly, restriction endonucleases and DNA methylases, which are part of the non-adaptive defense system, were enriched in 120 small CAGs. Also, genes involved in modification of the bacterial exterior, important for bacterial identification and masking thereof, were enriched in a number of small CAGs. These included genes involved in modifications of the cell wall and, in particular, glycosyltransferases.

Dependency-associations affiliates small CAGs to MGS

Existence of small CAGs, representing mobile genetic elements and clone differences, implies that they depend on cellular organisms, for their proliferation. In relationships that are non-promiscuous, a dependent CAG should never occur independently of the hosting microorganism. Significant dependency-associations were indeed identified by comparing absence/presence profiles for all pairs of CAGs (incl. the MGS) across samples using Fisher's exact test and excluding relationships where a potential dependent CAG was observed independently of the hosting CAG (Fig. 2A). Notably, these relationships are directional, with one CAG being dependency-associated to the other CAG. The resulting network of the most significant dependency-associations is shown in Fig. 2B and contains 882 relationships between 1,205 CAGs (for details see Supplementary Data 6). The dependency-association network is dominated by sub-networks, most of which are centered on an MGS. However, the network also contains nine MGS-interconnecting small CAGs, which all connect MGS pairs of the same genus.

413 of the associations were supported by sample-specific sequence contigs that bridged between the dependency-associated and the hosting CAG (odds ratio 2,513, Fisher's exact test $P << 1 \times 10^{-100}$). This points to occasional genomic integration for some of the dependency-associated CAGs. As expected, the network is significantly overrepresented for small CAGs that associate to an MGS (odds ratio 12.7, Fisher's exact test: $P << 1 \times 10^{-100}$).

An important aspect of the dependency-associations lies in their ability to connect CAGs into sub-networks, which guide explorations and understanding of the parts. For example, the sub-network centred on *Sutterella wadsworthensis* (MGS:135, Fig. 2C) contains eight dependency-associations including the associations of the phage like CAG:3731 and the CRISPR associated genes and repeat region containing CAG:4011. Interestingly, the sample-wise detection of the CRISPR and phage like CAG were anti-correlated (Matthew's correlation coefficient -0.7) and one of the CRISPR spacers had a 15 bp sequence match to the phage. Observations that fit the interpretation that the CRISPR prevents the homologous phage from infecting the bacterium[20]. Obviously, all the features of this sub-network and of many others are not presently understood, but the description of the dependency-associations lays the ground for future studies.

Another example illustrating the non-syntenic nature of some dependency-associations is shown in Fig. 2D. Here sample-specific MGS augmented assembly of the *E. coli* MGS:4 and its dependency-associated CAGs (Supplementary Fig. 14; Supplementary Data 7) are shown. The sample-specific assemblies demonstrate strong sequence similarities throughout the majority of the chromosome, but also demonstrate differences. The largest of the *E. coli* associated CAGs (CAG:427, containing 345 genes) is indicated in red throughout the 11 best assemblies, and in agreement with its detection profile, it was absent in some sample-specific assemblies. Strikingly, the non-redundant integration of CAG:427 was found to be scattered across the genome. This lack of chromosomal continuity suggests that CAG:427 describes differences between *E. coli* strains that cannot be explained by a single mutational event.

Dependency-associated CAGs influence the persistence probability of their MGS host

To investigate the effect that dependency-associated CAGs may have on their host MGS, we analysed the 73 human individuals that where sampled at two different time-points. From each of these sample pairs it could be determined if a given MGS was present at the first time point, and whether it was still present at the second time point. Based on this information, it was possible to estimate how well a given MGS typically persisted, both with and without its dependency-associated CAGs present.

Some dependency-associated CAGs did indeed greatly influence the persistence of their hosting MGS. For instance, see the survival curve in Fig. 3A, where *Bifidobacterium adolescentis* can be seen to persist for much longer when co-observed with its dependency-associated CAG. To further analyse this phenomenon we employed logistic regression to infer annual persistence probabilities for MGS with and without their dependency-associated CAGs (Supplementary Data 6). The credibility of these estimates was quantified using Bayesian statistical methods[21], which in brief, outputs a so-called posterior probability distribution over the possible annual persistence probabilities (for details see Methods). From this analysis we identified 26 cases where the presence of a specific dependency-associated CAG

correlated with a substantially altered annual persistence probability of its hosting MGS. For example, the annual persistence probability of the aforementioned *B. adolescentis* (MGS:119) was estimated to be 88% in individuals where it was observed in association with CAG:2298, but only 18% in individuals where CAG:2298 was absent (posterior mean estimates). This corresponds to an increase of 70 percentage points in annual persistence probability (Fig. 3B, posterior probability that the effect is larger than zero = 99.94%). Similar positive effects were observed for *Prevotella copri*, *Escherichia coli, Faecalibacterium prausnitzii* and 12 additional MGS (see Supplementary Fig. 14). Additionally, 10 dependency-associated CAGs had a substantial negative effect on the persistence probability (Supplementary Data 6).

Across the dependency-associated CAGs which elevated the MGS persistence probability we observed a diverse set of functions including complexes of CRISPR associated genes (CAG:2720), collagen adhesion protein and gram-positive anchor proteins (CAG:2888) and 'thioredoxin family proteins' that may be important for the tolerance of reactive oxygen species (ROS). This is in line with the observation, that the most common species in the human gut microbiome are prone to have functions that mediate ROS tolerance (see Supplementary Information and Supplementary Data 8). Among the dependency-associated CAGs that contributed negatively to the MGS persistence probability we observed three phage-like CAGs.

DISCUSSION

We present a method for producing a complete co-abundance clustering of an entire microbiome. The resulting MGS and CAGs offer an unprecedented insight into the microbial species and their genetic makeup, and thereby elucidate details important for rationalizing the content of the gut microbial community. Importantly, the clustering is purely data driven and hence it circumvents the use of reference genomes and cultivation of microbial species and yet in a single analysis uncovers hundreds of microbial organisms and thousands of smaller genetic modules. The discrimination between strains of the same species indicates that the co-abundance is very powerful in segregating closely related biologically entities. In comparison, gene sets that are defined by sequence similarity to known reference genomes often displayed incoherent abundance profiles (Supplementary Fig. 15), which potentially could lead to false associations between clinical conditions and putative species. While we see a few cases of chimeric assemblies (see Supplementary information) we have no indication of CAGs constituting multiple species, however, such entities could in principle exist in very close co-abundance.

Interestingly, we found that genes involved in resistance to antibiotics had distinct single gene abundance profiles (except vancomycin resistance genes, for details see Supplementary Text). This is in line with the fact that most antibiotic genes, except vancomycin resistance genes, are known to single-handedly provide antibiotic resistance. It suggests that some genes may be highly dynamic and perhaps are best understood non-contextually, at the single gene level.

While the strong dependency-associations observed primarily between small CAGs and MGS are merely associations, they point to an occasional affiliation of the associated CAGs. This interpretation is strongly supported by sample specific sequence contig overlap, taxonomical consistency across associations, and the preferential direction from small CAGs to MGS. Furthermore, the occurrence of some

dependency-associations was found to discriminate between populations with different persistence probabilities. These discriminative CAGs contribute functions that suggest a role in tolerating ROS, a common innate immune response mediator and in anchoring to the intestinal epithelia. These CAGs may serve as adaptations to the conditions in the gut including the co-existence with other species or persistence in their absence (Supplementary Fig. 16). These findings are important because they generate insight on the selective pressures faced by the gut microbes and the genetic adaptations that these undergo. These findings also indicate factors that may prevent less adapted species from colonizing the system and suggest directions for future engineering of microbial communities, *e.g.* for substitution of faecal microbiome transplantation. It may even pinpoint critical genetic elements for this. However, the adaptations to species co-existence also indicate that there may not be one solution that will fit all communities.

METHODS SUMMARY

396 human faecal samples were deep sequenced (Illumina) and assembled into a non-redundant catalogue of 3,871,657 genes. Genes were clustered into co-abundant gene groups (CAGs) with a fixed inclusion criterion (PCC > 0.9 to the median profile). CAG augmented assembly (Velvet[22], SOAPdenovo GAPCloser[23]) was done on a subset of reads recruited by mapping sample specific reads to CAG gene containing contigs. Gene-wise species, genus and phylum level taxonomy were assigned using best-hit sequence similarity to a reference sequence (from NCBI, 3,048 reference genomes) that exceeded 95%, 85% and 75% over 100 bp, respectively. Genes were functionally annotated by sequence similarity to ACLAME[24], UniPro[25], VFDB[26], eggNOG database[27] and *Bacillus subtilis* essential genes (Supplementary Data 9). CRISPR repeat-spacer segments were identified using CRT[28]. Functional enrichment for CAGs was done using Fisher's exact test ($P < 0.001$). A CAG was called phage-like if it passed one of two criteria: a) If a CAG contained a minimum of 10 *phage-taxonomy annotated genes*[19] and 80% of these were consistent at species, genus or family level, or b) if a CAG encoded ≥ five distinct *characteristic phage functions*[10] and ≥ 40% of the CAG genes were *most similar to known phage genes*. A CAG was considered dependency-associated to another CAG if the pair co-occurred significantly (Fishers exact test, $P < 1 \times 10^{-10}$, after Bonferroni correction) and if the candidate CAG was not observed independently. Annual persistence probabilities for MGS with or without a given dependency-associated CAG were estimated from 73 individuals that were sampled twice using a probabilistic (Bayesian) model-based framework that explicitly accounts for time-dependence.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013): MetaHIT as well as the Novo Nordisk Foundation Center for Biosustainability. Work on the clustering concept has been supported by the OpenGPU FUI collaborative research projects, with funding from DGCIS. This work was granted access to the HPC resources of CCRT under the allocation 2011-036707 made by GENCI (Grand Equipement National de Calcul Intensif). The company Alliance Services Plus (AS+) has provided a helpful cooperation to scale up the process and especially Victor Arslan, Dany Tello, Vincent Ducrot, Tarik Saidani and Sébastien Monot. The authors affiliated to MGP are funded, in part, by the Metagenopolis ANR-11-DPBS-0001 grant. Ciberehd is funded by the Instituto de Salud Carlos III (Spain). A.M. was supported by a grant from the Ministère de la Recherche et de l'Education Nationale (France).

AUTHOR CONTRIBUTIONS

All authors are members of the Metagenomics of the Human Intestinal Tract (MetaHIT) Consortium. S.D.E, S.B. managed the project. F.C., N.B., F.G., T.H., K.S.B. and T. N. performed clinical sampling. F.L. and C.M. performed DNA extraction. J.L. performed sequencing. S.D.E., H.B.N, M.A., A.S.J., S.R. and P.B. designed the analyses. H.B.N., A.S.J., S.R., M.A., A.P., D.P., L.G., I.B, M.B., M.B.Q.S., M.A., J.H., S.S., T.Y and R.S.K performed the data analyses. H.B.N., S.B., A.S.J., S.R., A.P. and M.A. wrote the paper. S.D.E., S.B., P.B., E.P., O.P and D.W.U. revised the paper. The MetaHIT Consortium members contributed to the design and execution of the study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

ADDITIONAL INFORMATION

Sequence data was deposited at EBI with the accessions ERP002061, and CAG augmented assemblies were deposited at EBI under (PRJEB674 - PRJEB1046). 454 sequencing reads were added to the NCBI BioProjectID 32811. The 3.9 M gene catalogue and the CAGs are available for download from https://www.cbs.dtu.dk/projects/CAG/. Source code for the MGS canopy algorithm is available from: http://git.dworzynski.eu/mgs-canopy-algorithm

Correspondence and request for materials should be addressed to S.B. (brunak@cbs.dtu.dk) and S.D.E. (dusko.ehrlich@jouy.inra.fr).

REFERENCES

1. Qin, J. *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
2. Fodor, A. A. *et al*. The "most wanted" taxa from the human microbiome for whole genome sequencing. *PLoS One* **7,** e41294 (2012).
3. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–20 (2005).
4. Turnbaugh, P. J. *et al*. A core gut microbiome in obese and lean twins. *Nature* **457,** 480–4 (2009).
5. Rajilić-Stojanović, M., Heilig, H. G. H. J., Tims, S., Zoetendal, E. G. & de Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* (2012). doi:10.1111/1462-2920.12023
6. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–36 (2011).
7. Schloissnig, S. *et al*. Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).
8. Fitzsimons, M. S. *et al*. Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. *Genome Res.* gr.142208.112– (2013). doi:10.1101/gr.142208.112
9. Reyes, A. *et al*. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466,** 334–8 (2010).
10. Minot, S. *et al*. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21,** 1616–25 (2011).
11. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22,** 1985–94 (2012).
12. Zhang, Q., Rho, M., Tang, H., Doak, T. G. & Ye, Y. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biol.* **14,** R40 (2013).
13. Wang, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* **28,** i356–i362 (2012).
14. Albertsen, M. *et al*. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31,** 533–538 (2013).
15. Karlsson, F. H. *et al*. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498,** 99–103 (2013).
16. Qin, J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
17. Le Chatelier, E. *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–6 (2013).
18. Chervaux, C. *et al*. Genome sequence of the probiotic strain Bifidobacterium animalis subsp. lactis CNCM I-2494. *J. Bacteriol.* **193,** 5560–1 (2011).
19. *Virus Taxonomy: Classification and Nomenclature of Viruses : Ninth Report of the International Committee on Taxonomy of Viruses*. 1327 (Elsevier, 2012). at <http://books.google.com/books?hl=en&lr=&id=KXRCYay3pH4C&pgis=1>
20. Terns, M. P. & Terns, R. M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* **14,** 321–7 (2011).
21. Kruschke, J. K. Bayesian data analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* **1,** 658–676 (2010).
22. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–9 (2008).
23. Li, R. *et al*. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).
24. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.* **38,** D57–61 (2010).
25. Bairoch, A. *et al*. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33,** D154–9 (2005).
26. Chen, L., Xiong, Z., Sun, L., Yang, J. & Jin, Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* **40,** D641–5 (2012).

27. Powell, S. *et al*. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res*. **40,** D284–9 (2012).
28. Bland, C. *et al*. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).

FIGURES AND FIGURE LEGENDS

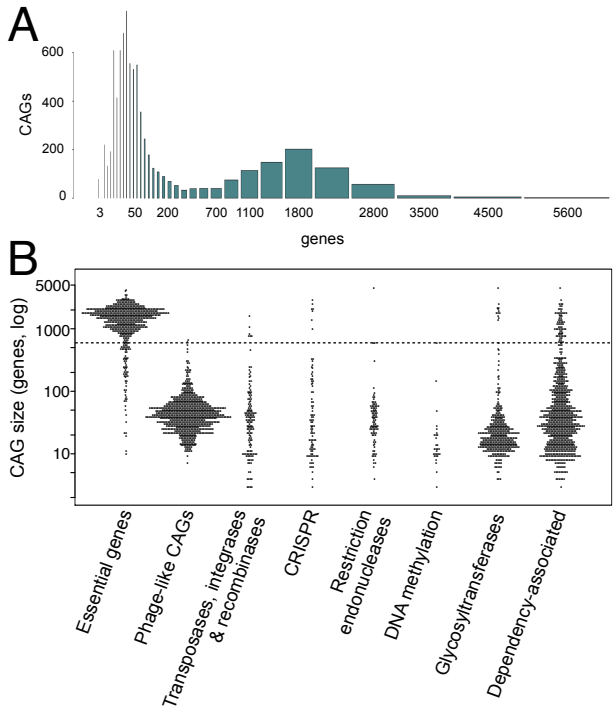


Fig. 1. Co-abundance gene groups (CAGs). **A**) Histogram showing the CAG size distribution in terms of gene content. The scale is logarithmic as indicated by the bar widths. **B**) Bee swarm plot showing CAGs that are significantly enriched for the indicated gene annotation (as dots) vs. the number of genes contained. Phage-like CAGs and dependency-associated CAGs are defined in Methods. The dashed line marks the small CAGs to MGS size threshold (700 genes).

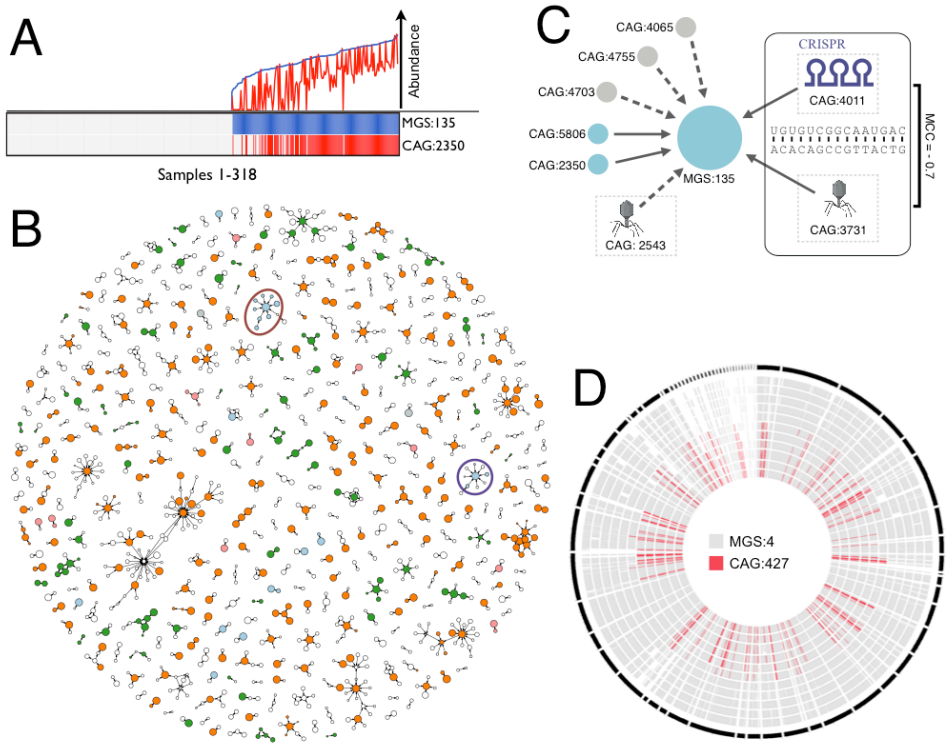


Fig. 2. Dependency-associations among MGS and CAGs. **A**) A typical example of a significant dependency-association. The abundance of the MGS:135 (S*utterella wadsworthensis*) and the small CAG:2350 across 318 faecal samples are shown as blue and red curves, respectively (upper panel, logarithmic scale). Below the sample-wise presence of the two CAGs is shown as bars. CAG:2350 is significantly co-occurring with MGS:135 and never detected independently ($P = 9 \times 10^{-74}$). **B**) A directional network showing all 882 significant dependency-associations among 287 MGS and 918 small CAGs. Arrows indicate the dependency-associations among CAGs (circles). The size of the circles indicates the number of genes in a specific CAG and the phylum level gene annotation is indicated by colour (green: Bacteroidetes, orange: Firmicutes, blue: Proteobacteria, pink: Actinobacteria). The blue circle indicates the S. *wadsworthensis* (MGS:135) centred sub-network shown in Fig. 2C and the red circle the *E. coli* (MGS:4) centred sub-network in Supplementary Fig. 14. **C**) The dependency-association sub-network of CAGs associated to S. *wadsworthensis* (MGS:135). Arrows show dependency-associations and solid arrows indicate that co-assembly of the MGS and the CAG in one or more samples supported the association. Blue colouring indicates CAGs dominated by genes with species level similarity to S*utterella wadsworthensis*. CAG:2543 and CAG:3731 are significantly enriched for phage like genes and CAG:4011 contains a series of CRISPR associated genes and a CRISPR cluster. The CRISPR (CAG:4011) and the one of the phages (CAG:3731) anti-correlate (Matthews correlation coefficient = - 0.7) and spacers of the CRISPR show sequence complementarity to the phage. **D**) The *E. coli* (MGS:4) and its nine dependency-associated CAGs were co-assembled to high quality draft genomes in each of 11 samples. The outer black circle represents the consensus assembly of the *E. coli* centred agglomerate and each of the gray circles represent alignment of the assembly from a particular sample. The positions and sequence coverage of CAG:427 are marked in red, across the assemblies.

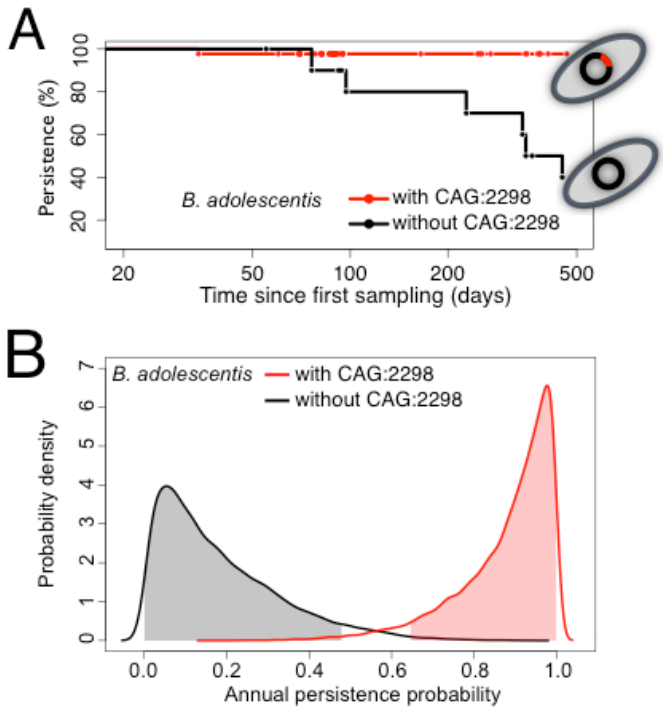


Fig. 3. *Bifidobacterium adolescentis* (MGS:119) has substantially higher persistence probability when the dependency-associated CAG:2298 is present. A) Persistence curves, shoving the cumulative loss of *B. adolescentis* over time as observed across 54 human individuals who had the bacterium in the first of two samples. The *B. adolescentis* containing individuals was stratified into two sub-populations, with or without the dependency-associated CAG:2298. Points indicate time (in days) between the first and second sample. The curve shows the "losses" when they are registered at the second time point (*i.e.* the data is interval-censored). B) Posterior probability densities of the annual persistence probability for *B. adolescentis* with or without the dependency-associated CAG. The shaded areas indicate the 95% probability distributions of the respective sub-populations.

Chapter 6

Bile Acid Degrading Bacteria in Relation to Obesity - A Hypothesis

6.1 Introduction

The relation between obesity and the gut microbiota was touched upon in Chapter 3. This seems to be a very complex phenotype, which cannot simply be described by absence or presence of a single species or gene [31]. There are trends pointing towards low richness of species and a skewed balance between Bacteroidetes and Firmicutes are more often found in obese than in lean individuals [136]. Although, there has been recent findings that seem to contradicting this [35].

It has been suggested that the difference in the obese microbiota may not be found in the taxonomical makeup, but rather at the functional level [35]. This was supported by a study by Turnbaugh *et al.* [137], which showed an increase in genes related to energy harvest from lipids and carbohydrates in obese individuals [133, 137].

The changes observed in the microbiota of obese individuals could be the effect of the obesity rather than the cause. However, in mouse studies it has been shown that transferring the microbiota from an obese mouse to a lean one will induce weight gain in the lean mouse, indicating that the gut microbiota directly affects the weight status of the host [31, 111, 137]. Additionally, the fact that obesity is associated with tissue inflammation hints at bacterial involvement, as these are able to produce inflammatory

agents like lipopolysaccharides and peptidoglycans [133].

One of the factors that might influence the progression of obesity is the ability of the intestinal bacteria to degrade bile acids [13]. These molecules are produced from cholesterol in the liver. The primary function of bile acids is in the digestion of fats and fat soluble vitamins from the diet, but they are also very important in regulation of lipid, glucose and cholesterol homeostasis as well as in immune cell regulation [3, 13, 106]. By degradation of the bile acids to deconjugated bile acids and secondary bile acids, the intestinal microbiota has a great influence on the bile acid functions and thus might also affect the development of obesity and other related diseases [50, 158].

Metagenomic analysis has shown that the degradation of bile acid is a conserved function in the gut microbiota. However, the abundance of bile acid degradation genes can vary considerably between individuals [50]. The reduction of species in a dysbiotic system can subsequently result in a reduction in bile acid degrading bacteria. This in turn, has an influence on the glucose and lipid balance as well as on cholesterol breakdown and excretion [52, 106]. Moreover, studies in mice have shown that obese test animals seem to have a reduction in bile acid degradation functions [52].

The first step in degradation of bile acids by bacteria is the deconjugation of the molecule. This is performed by the bile salt hydrolase (BSH) enzyme [110]. The deconjugated bile acid can subsequently be transformed to secondary bile acids or be excreted, the latter resulting in elevated cholesterol turnover [13, 52]. The primary pathway for bile acid degradation is shown in Figure 6.1.

In this study we wanted to investigate if any of the CAGs, obtained by clustering the MetaHIT data described in Chapter 5, could be related to obesity. As mentioned, no clear connection between obesity and species of the gut microbiota has previously been found. Thus, we wanted to investigate if any functional properties correlated with the BMI of the MetaHIT test population. Here, the focus is on the bile acid degradation potential.

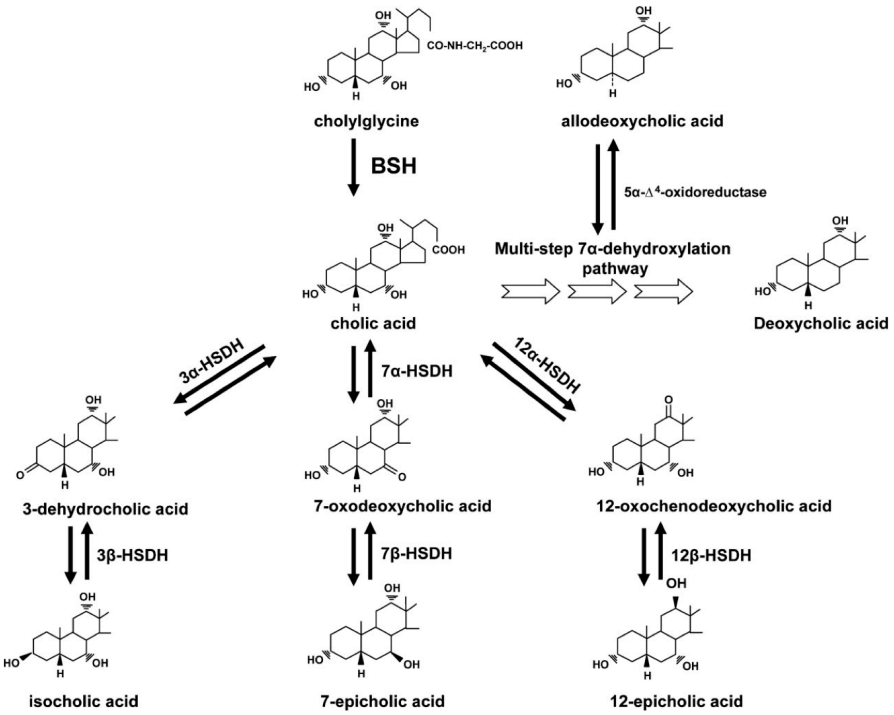


Figure 6.1: Biotransformation of bile acids by intestinal bacteria. The primary bile acids are cholic acid and chenodeoxycholic acid. These are conjugated to glycine or taurine in the liver thus resulting in bile salts. This figure shows the pathway for degradation of cholylglycine, the conjugated form of cholic acid. BSH: bile salt hydrolase, HSDH: hydroxysteroid dehydrogenase. Adapted from Ridlon *et al.* [110]

.

6.2 Methods

This project is also based on the CAG clustering of the the MetaHIT data, which was described in Chapter 5. Only the species sized CAGs, i.e. the ones containing 700 genes or more, were included. In the paper presented in Chapter 5, these are renamed MGS, however I keep the name CAG.

COG annotation

All CAGs were functionally annotated based on the COG database [127]. The annotation was performed as described by Qin *et al.* [104] for the first 124 MetaHIT samples. Essentially, a BLAST [5] search was performed with

an E-value cutoff of 10^{-5}.

BLAST annotation

Bile acid related genes were selected from the NCBI database based on the pathways described in the papers by Ridlon *et. al* [110] and Philipp [100]. In this way, a database was constructed containing 782 3-α-HSDH genes, 1829 3-β-HSDH genes, 2170 7-α-HSDH genes, 27-β-HSDH genes and 4008 BSH genes. Additionally one of each of the genes baiB, baiCD, baiE, baiA, baiF, baiG, baiH, baiI, baiA2, baiA3, acad, HsdA, tesH, tesI, KshA, KshB, HsaA, HsaB, HsaC, tesB, HsaD, tesD, ORF18, tesE, tesG and tesF were included. The full MetaHIT gene catalogue was aligned to this database using BLASTp [5] with a bit score cutoff of 60.

6.3 Results

Initially, it was investigated if any of the CAGs correlated significantly with BMI by means of Pearson correlation. The correlations were very weak, ranging between -0.24 and 0.20. This corresponds to the results described in the introduction, that no single species can be associated to obesity. A Wilcoxon signed-rank test between BMI and each COG annotation for all CAGs was performed to investigate if presence of any COG annotations were skewed towards either obesity or leanness. We identified 19 ($p<0.0001$ after Bonferroni correction) such COGs, 18 of which occurred mostly in leanness associated CAGs and 1 was primarily found in obesity correlated CAGs. Looking at the functional description of these COGs, COG4927 was biologically interesting, as it was annotated as "predicted choloylglycine hydrolase". Choloylglycine hydrolase is a bile salt hydrolase (BSH), which is, as mentioned in the introduction, the initial step in the bile acid degradation pathway of bacteria. The correlation to BMI of the 27 CAGs that include this COG is shown in Figure 6.2. 26 of these 27 CAGs are marginally correlated to leanness.

Although these are weak correlations, we take this as an indication that bile acid degradation could be important for the obesity phenotype. Next, we wanted to better annotate the bile acid pathways in the CAG set. This was done by BLAST search of the gene catalogue to a database of bile acid degradation pathway genes, as described in the methods. We did not identify any CAGs with a full degradation pathway. Most of the significant hits were to the three HSDH genes or BSH, which is most likely due to these genes being over represented in the database compared to the other genes. However, we did observe that the CAGs that were annotated as having COG4927, thus possessing the bile salt hydrolase function, had a higher number of different bile acid pathway genes on average than the rest of the

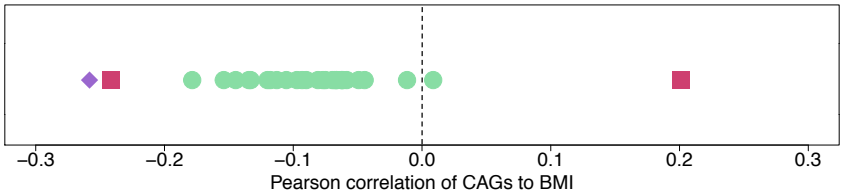


Figure 6.2: BMI correlation of the CAGs that are positive for COG4927 are shown in green circles. The red squares depict the minimum and maximum correlations of all CAGs to BMI and the purple diamond shows the correlation of all COG4927 CAGs (summarized abundance over all samples) to BMI (Pearson correlations).

CAGs (mean of 7 versus 5). This supports the annotation of these CAGs, as these being involved in bile acid degradation. The correlations between presence of the other bile acid degradation pathway genes and BMI were examined and we found that these do not correlate as well as the presence of COG4927. Additionally, the CAGs including most bile acid genes (by count of genes) did not correlate better to BMI than the COG4927 annotated CAGs.

There is also the possibility that the correlation could be stronger looking at the COG4927 annotation for each patient, leaving out the CAG clustering, but this was not the case.

6.4 Discussion and Perspectives

The weak correlation of BMI to CAGs possessing the choloylglycine hydrolase activity we see in this data is not enough to make any final conclusions. However, it could be an indication of a relationship between bile acid degradation of intestinal bacteria and obesity, which has also been suggested by others [50, 158]. To better assess this, improvement of the annotation is needed. By using, BLAST as we have done here, it is very easy to either over- or under-annotate the genes, depending on the number of genes in the selected database and the chosen cutoff.

A general limitation when analyzing metagenomics data is the fact that it only illustrates the potential of the system, not which genes are actually being expressed. In this case, it would be very valuable to integrate the metagenomics data with transcriptomics or metabolomics data to see if there are any correlations between the BMI of the test subjects and the bile acid

degradation activity of the intestinal bacteria.

Studies in mice might also be valuable in order to better understand the effect of bile acid degradation on the metabolism. This will hopefully be transferable to humans. In the future, perhaps bile acid degrading species could be used as probiotics against obesity.

In conclusion, the hypothesis of bile acid degrading bacteria having a positive influence on the BMI still requires further investigation. However, this study illustrates one type of analysis that is enabled by the CAG clustering method. Thus, it was possible to extract data regarding bacteria possessing a certain function from the genomic data and analyze this in connection to the metadata.

Chapter 7

Blastocystis Occurrence in the Human Gut

7.1 Introduction

Blastocystis is a human gut parasite that is found in a large proportion of the general population. Worldwide, an estimated 1-2 billion people harbor this organism [112]. The parasite was first described in 1911 and named in 1912 as *Blastocystis hominis*. However, this name is rather misleading as the parasite also infects other animals, thus the name is at the moment changing to *Blastocystis spp.* [22]. Already in 1916, a study was conducted on the prevalence of this parasite in South Carolina [78]. It is striking that a parasite that has been known for about 100 years is still relatively poorly described with respect to its pathogenicity [22, 112, 122]. Some sources describe this parasite as part of the normal gut flora, whereas others contradict this and believe this to be an emerging pathogen [10, 22, 109, 112].

Symptoms suspected to be caused by *Blastocystis* include irritable bowel syndrome (IBS), diarrhea of varying severity, abdominal pain, nausea, anorexia and flatulence [9, 112]. However, studies describing *Blastocystis* as a pathogen generally suffer from limited sample sizes and detection methods. Furthermore, efforts to rule out other possible causes of the symptoms have been lacking in these works [22, 121]. Some results suggest that the severity of the infection depends on the subtype of the infecting agent. However, there is no consensus regarding which subtypes are found to be more pathogenic than others [22]. It is possible that symptomatic infections are caused by virulence genes expressed by some strains [112]. In most cases, the infection is asymptomatic, which is fortunate, as it is hard, if not impossible, to eradicate the parasite [112, 122]. *Blastocystis* has been described as the cause of

chronic diarrhea in patients suffering from other health issues, such as HIV infection and cancer [22]. This suggest that the parasite is an opportunistic pathogen, since a large part of the population, as mentioned, carries this parasite without any symptoms [112, 122].

In order to better describe the distribution of *Blastocystis* in the population and to identify possible connections to the remaining gut flora, metagenomic data from cohort studies like the MetaHIT project [104] can be very valuable [112, 145]. With this in mind we set out to describe the *Blastocystis* prevalence in the MetaHIT samples. To my knowledge this is the first time *Blastocystis* frequency in a test population has been described using metagenomics data. The analysis, which is presented in the paper included in the next section, is based on the CAG clustering described in Chapter 5.

7.2 Manuscript - A Metagenomic Approach to Studying Intestinal Microbial Eukaryotes

A Metagenomic Approach to Studying Intestinal Microbial Eukaryotes

Running title
Gut microbiota analysis of *Blastocystis* carriers

Authors
Lee O'Brien Andersen[1*], Ida Bonde[2,3*], Henrik Bjørn Nielsen[3], Christen Rune Stensvold[1]#

[*]Andersen and Bonde should both be considered first authors

Affiliations
[1]Unit of Mycology and Parasitology, Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark.
[2]The Novo Nordisk Foundation Center for Biosustainability, Scion-DTU, Hørsholm, Denmark.
[3]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark.

#Corresponding author:
Christen Rune Stensvold, Unit of Mycology and Parasitology, Department of Microbiology and Infection Control, Statens Serum Institut, Artillerivej 5, DK-2300 Copenhagen S, Denmark. Tel: +45 3268 8224. Email: run@ssi.dk

Abstract
Human fecal metagenomic data was screened for *Blastocystis*, a common single-celled parasitic protist of unsettled clinical significance, by searching for subtype-specific mitoribosomal proteins in co-abundance gene groups, which are groups of genes that co-vary across the 396 samples analyzed. We identified the prevalence of the four most common *Blastocystis* subtypes in healthy individuals and patients with inflammatory bowel disease. *Blastocystis* was absent in patients with Crohn's Disease (n=20). Individuals with the Bacteroides-driven enterotype were much less prone to having *Blastocystis*-positive stool than individuals with Ruminococcus- and Prevotella-driven enterotypes ($p<0.0001$). This is the first study to investigate the relationship between *Blastocystis* and communities of gut bacteria. The study serves as an example of how it is possible to investigate microbial eukaryotic communities in the gut using metagenomic datasets targeting the bacterial component of the intestinal microbiome and the interplay between these microbial communities.

Keywords: Metagenomics, Data mining, Gut ecology, *Blastocystis*, microbiota

THE HYPOTHESIS

Advances in Next Generation Sequencing (NGS) methods have led to the production of vast amounts of sequencing data. Metagenomic studies of relevance to human clinical microbiology have generally been limited to studying bacterial communities in various types of patient samples (1). Meanwhile, large pools of sequence data representing microbial eukaryotes are apparently being ignored despite eukaryotes being part of many habitats, including the intestinal tract. Long lasting intestinal colonisation with microbial eukaryotes is common in both healthy individuals and patients with common functional bowel diseases such as irritable bowel syndrome (2, 3). One of the most common micro-eukaryotes is *Blastocystis,* an anaerobic stramenopile of unsettled clinical significance (4-7). The genus comprises multiple ribosomal lineages, the so-called subtypes, which are arguably separate species, and of which nine have been found in humans (8, 9). In most countries, humans are primarily infected by ST3, followed in prevalence by ST1, ST2 and ST4, although clear geographical differences are seen (10). Here, we show that it is possible to extract data on *Blastocystis* from whole genome sequencing metagenomic data intended for analysis of bacterial communities. Such data represent a valuable source of information, enabling the linkage of data on microbial eukaryotes to bacterial microbiota profiles. At present, there are massive amounts of publicly available metagenomic datasets ready to be analyzed.

In this work we 'sift' metagenomics data generated by the MetaHIT consortium (Nielsen et al., submitted) for DNA signatures of *Blastocystis*. We suggest this method be applied to other microbiome datasets for elucidation of the micro-eukaryotic component.

***BLASTOCYSTIS* CARRIAGE – PREVALENCE AND ASSOCIATED INTESTINAL MICROBIOTA**

In the previous study (Nielsen et al., submitted), fecal genomic DNAs from 396 study individuals (177 Danish and 219 Spanish) were sequenced. Danish test subjects included only healthy individuals, whereas the Spanish cohort included 20 with Crohn's disease (CD), 127 with ulcerative colitis (UC), and 72 healthy individuals. The age of the study individuals ranged from 18 to 70 with a median of 49 (IQR: 40.25—59). All individuals had been assigned to one of three enterotypes as defined by Arumugam et al. (11) prior to this study (Nielsen et al., submitted).

We utilized the metagenomic binning method described by Nielsen et al. (submitted); genes were called in the metagenomics data for each sample, and the abundance of each gene in each sample was quantified. The abundance of each gene between all 396 samples would then result in an abundance profile, where genes from all samples were clustered according to co-abundance based on the assumption that genes originating from the same organism would have similar abundances profile between all samples. We refer to these clusters as CAGs (co-abundance gene groups). We searched our data for CAGs containing *Blastocystis* subtype-specific DNA signatures and found four CAGs that corresponded to *Blastocystis* ST1-ST4 respectively.

The distribution of the 4 subtypes in the different cohorts is summarized in Table 1. *Blastocystis* CAGs were detected in 20 % (81/396) of all study individuals, of whom 14 had ST1, 14 ST2, 17 ST3, and 36 had ST4. The prevalence of *Blastocystis* of any of the subtypes was not associated to gender, nationality, BMI or age of the sample donors. However, there were clear differences in *Blastocystis* prevalence among the enterotypes, commonly used for gut microbiome stratification. While *Blastocystis* colonized only 5 % of the 154 individuals with the *Bacteroides*-driven enterotype*, Blastocystis* was observed in 26 % (n=62) and 32 % (n=180) of the individuals with *Prevotella*- and *Ruminococcus*-driven enterotypes, respectively ($p<0.0001$, $\chi 2$-test). A possible explanation for this is that *Blastocystis* may be positively correlated to species richness; the

Bacteroides-driven enterotype appears to be negatively correlated to species richness and therefore possibly also to *Blastocystis.* The finding could also be explained by the possibility that *Blastocystis* colonization is dependent on the activity of certain types of fermenting bacteria; indeed *Blastocystis* is found in the cecum and colon, which is also where SCFA-producing bacterial communities exist.

We also analyzed the co-occurrence of the four *Blastocystis* subtypes and found that none of the investigated CAGs were found together in any of the patients. However, cases of mixed subtype colonization have been reported (12-15)

None of the 20 patients with CD had *Blastocystis*, consistent with previous findings (16). *Blastocystis* was common in all other groups included in the study, being most prevalent among healthy individuals (24 %). Although ST3 is the most common subtype world-wide, ST4 has previously been found particularly predominant in Danish and Spanish cohorts (17, 18); hence, the present results are comparable to findings for similar cohorts obtained by detection and differentiation of SSU rRNA genes, which is currently state-of-the-art for *Blastocystis* diagnosis and subtyping (19, 20). We believe that screening CAGs for *Blastocystis* has imminent potential as a valid method for detection and genetic differentiation of the parasite. Also, further investigations of the remaining genes in the CAGs, especially the genes with no similarity to known genes, might yield more knowledge about this parasite and its role in human health and disease.

CONCLUSION

We have shown that metagenomic data focused on prokaryotes can also be valuable for detection of eukaryotic DNA signatures. Other body sites and other poorly described species can be investigated in the same manner provided that there is data available from enough samples. Most metagenomic studies include both study and control groups, and so it should be possible to analyze the health impact of various micro-eukaryotes (some of which are still surrounded by conundrums (1)) in the context of the bacterial microbiota by phylogenetic interrogation of relevant genes. Utilizing existing metagenomic data from studies across continents could be used to produce standard microbiomes of healthy populations, which in turn could serve as reference enabling us to identify dysbiosis in both pro- and eukaryotic components of the intestinal microbiome.

Co-abundance gene groups clustering

The CAG construction that we utilize in this study was performed previously by Nielsen et al. (submitted). This is a summary of how it was done; for more information refer to the original paper. Sequencing data from 396 samples was assembled and a gene catalogue containing 3,871,657 genes was produced by gene calling and homology reduction (cut off 95 % identity over 90 % of the length of the shortest gene). An abundance matrix was generated by mapping back all the sequencing reads to the gene catalogue and counting the hits for each gene in each sample. The dataset was downsized to 3 million reads per samples, and samples with less than 3 million reads were discarded. Genes that were observed in less than 3 samples were removed and genes that had more than 90% of the signal coming from one sample were also ignored. The data was normalized based on the gene lengths and the total signal for each sample, and the resulting matrix was used as input for the clustering. The clustering method used canopy clustering that had a fixed inclusion criterion set to PCC >0.9 to the median profile. In this way, the genes that had similar abundance profiles were binned together in co-abundance gene groups (CAGs).

***Blastocystis* annotation of CAGs**

CAGs that had an overrepresentation of genes with similarity to *Blastocystis* ST7 (the only published complete *Blastocystis* genome at the time of analysis) when BLASTing to the UniProt database (http://www.uniprot.org/) were annotated as *Blastocystis*. Further annotation of these CAGs to *Blastocystis* subtypes were performed by comparing the CAG genes to genomes of mitochondrial-like organelles (MLO) of known *Blastocystis* subtypes (ST1–ST4) (21, 22) unpublished observations). Four CAGs could be assigned to a *Blastocystis* subtype based on the rps12 gene of the MLO, which was chosen as marker gene. For validation, the CAGs were blasted to draft genomes of *Blastocystis* subtypes 2, 3, 4, 6, 8, and 9 (unpublished data), using the CLC Genomics multi BLAST function (default settings and expect = 10; word size = 30). Of the 2041 sequences in the CAG that had been annotated as ST2 using the rps12 gene, 10 sequences matched the draft genome of *Blastocystis* ST9, 2 sequences matched ST8, 12 sequences matched ST6, 9 sequences matched ST4, 34 sequences matched ST3 and 1698 sequences matched the draft genome of *Blastocystis* ST2. The other CAGs had similar BLAST result patterns. This demonstrates that 1698/2041 (83.2%) of the genes combined into this specific CAG could be located in this crude draft genome of *Blastocystis* ST2 and that this CAG had a maximum similarity to any of the other five subtype draft genomes of 0.98%. The missing 16.8 % of the ST2 CAG is probably due to a cut-off of 2000 nt in minimum contig sizes when the draft genomes were assembled and could probably also be due to genomic differences within the different ST2 strains.

Acknowledgements:

Dr Graham Clark, London School of Hygiene and Tropical Medicine, is thanked for making the MLO genome of *Blastocystis* sp. ST2 available for analysis. Lee O'Brien Andersen's work is partly supported by the Lundbeck Foundation (Project no. R108-A10123).

Conflicts of interests:

All authors report of no conflict of interests.

References

1. **Andersen LO, Vedel Nielsen H, Stensvold CR.** 2013. Waiting for the human intestinal Eukaryotome. ISME J. **7:**1253-1255.
2. **Engsbro AL, Stensvold CR, Nielsen HV, Bytzer P.** 2012. Treatment of *Dientamoeba fragilis* in patients with irritable bowel syndrome. Am. J. Trop. Med. Hyg. **87:**1046-1052.
3. **Scanlan PD, Marchesi JR.** 2008. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. ISME J **2:**1183-1193.
4. **Stensvold CR, Nielsen HV, Mølbak K, Smith HV.** 2009. Pursuing the clinical significance of *Blastocystis*--diagnostic limitations. Trends Parasitol. **25:**23-29.
5. **Tan KS.** 2008. New insights on classification, identification, and clinical relevance of *Blastocystis* spp. Clin. Microbiol. Rev. **21:**639-665.
6. **Clark CG, van der Giezen M, Alfellani MA, Stensvold CR.** 2013. Recent developments in *Blastocystis* research. Adv. Parasitol. **82:**1-32.
7. **Scanlan PD, Stensvold CR.** 2013. *Blastocystis*: getting to grips with our guileful guest. Trends Parasitol. **29:**523-529.
8. **Stensvold CR, Suresh GK, Tan KSW, Thompson RCA, Traub RJ, Viscogliosi E, Yoshikawa H, Clark CG.** 2007. Terminology for *Blastocystis* subtypes - a consensus. Trends Parasitol. **23:**93-96.
9. **Alfellani MA, Taner-Mulla D, Jacob AS, Imeede CA, Yoshikawa H, Stensvold CR, Clark CG.** 2013. Genetic diversity of *Blastocystis* in livestock and zoo animals. Protist **164:**497-509.
10. **Alfellani MA, Stensvold CR, Vidal-Lapiedra A, Onuoha ES, Fagbenro-Beyioku AF, Clark CG.** 2013. Variable geographic distribution of *Blastocystis* subtypes and its potential implications. Acta Trop. **126:**11-18.
11. **Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J, Antolín M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Mérieux A, Melo Minardi R, M'rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P, Consortium M.** 2011. Enterotypes of the human gut microbiome. Nature **473:**174-180.
12. **Stensvold CR, Lewis HC, Hammerum AM, Porsbo LJ, Nielsen SS, Olsen KE, Arendrup MC, Nielsen HV, Mølbak K.** 2009. *Blastocystis*: unravelling potential risk factors and clinical significance of a common but neglected parasite. Epidemiol. Infect. **137:**1655-1663.
13. **Forsell J, Granlund M, Stensvold CR, Clark CG, Clark GC, Evengård B.** 2012. Subtype analysis of *Blastocystis* isolates in Swedish patients. Eur. J. Clin. Microbiol. Infect. Dis. **31:**1689-1696.

14. **Malheiros AF, Stensvold CR, Clark CG, Braga GB, Shaw JJ.** 2011. Molecular characterization of *Blastocystis* obtained from members of the indigenous Tapirapé ethnic group from the Brazilian Amazon region, Brazil. Am. J. Trop. Med. Hyg. **85:**1050-1053.
15. **Stensvold CR, Traub RJ, von Samson-Himmelstjerna G, Jespersgaard C, Nielsen HV, Thompson RC.** 2007. *Blastocystis*: subtyping isolates using pyrosequencing technology. Exp. Parasitol. **116:**111-119.
16. **Petersen AM, Stensvold CR, Mirsepasi H, Engberg J, Friis-Møller A, Porsbo LJ, Hammerum AM, Nordgaard-Lassen I, Nielsen HV, Krogfelt KA.** 2013. Active ulcerative colitis associated with low prevalence of *Blastocystis* and *Dientamoeba fragilis* infection. Scand. J. Gastroenterol. **48:**638-639.
17. **Stensvold CR, Christiansen DB, Olsen KE, Nielsen HV.** 2011. Blastocystis sp. subtype 4 is common in Danish *Blastocystis*-positive patients presenting with acute diarrhea. Am. J. Trop. Med. Hyg. **84:**883-885.
18. **Domínguez-Márquez MV, Guna R, Muñoz C, Gómez-Muñoz MT, Borrás R.** 2009. High prevalence of subtype 4 among isolates of *Blastocystis hominis* from symptomatic patients of a health district of Valencia (Spain). Parasitol. Res. **105:**949-955.
19. **Stensvold CR.** 2013. Comparison of sequencing (barcode region) and sequence-tagged-site PCR for *Blastocystis* subtyping. J. Clin. Microbiol. **51:**190-194.
20. **Stensvold CR, Ahmed UN, Andersen LO, Nielsen HV.** 2012. Development and evaluation of a genus-specific, probe-based, internal process controlled real-time PCR assay for sensitive and specific detection of *Blastocystis*. J. Clin. Microbiol. **50:**1847-1851.
21. **Pérez-Brocal V, Clark CG.** 2008. Analysis of two genomes from the mitochondrion-like organelle of the intestinal parasite *Blastocystis*: complete sequences, gene content, and genome organization. Mol. Biol. Evol. **25:**2475-2482.
22. **Stensvold CR, Alfellani M, Clark CG.** 2012. Levels of genetic diversity vary dramatically between *Blastocystis* subtypes. Infect. Genet. Evol. **12:**263-273.

Table 1. *Blastocystis* subtype (ST) prevalence in numbers (#) and percentages (%) of total according to cohort. A total of 396 individuals were included; each cohort group (seperated by horisontal lines) that does not add up to 396 is due to undetermined cohort status of one or more individuals. The bottom cohort group is based only on the Spanish participants. CAG = Co-abundance gene groups; CD = Crohn's Disease; UC = ulcerative colitis.

		CAG-ST1		**CAG-ST2**		**CAG-ST3**		**CAG-ST4**		**Total**	
Cohort[1]	**Size (N)**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Gender:											
Female	224	9	4.02	6	2.68	8	3.57	16	7.14	39	17
Male	171	5	2.92	8	4.68	9	5.26	20	11.70	42	25
Body mass:											
Lean	163	6	3.68	5	3.07	7	4.29	15	9.20	33	20
Obese	113	2	1.77	5	4.42	7	6.19	5	4.42	19	17
Overweight	70	5	7.14	3	4.29	1	1.43	10	14.29	19	27
Enterotype:											
Bacteroides	154	1	0.65	2	1.30	1	0.65	4	2.60	8	5
Prevotella	62	4	6.45	3	4.84	5	8.06	4	6.45	16	26
Ruminococcus	180	9	5.00	9	5.00	11	6.11	28	15.56	57	32
Nationality:											
Danish	177	8	4.52	9	5.08	10	5.65	15	8.47	42	24
Spanish	219	6	2.74	5	2.28	7	3.20	21	9.59	39	18
Spanish:											
CD	20	0	0	0	0	0	0	0	0	0	0
Healthy	72	0	0	2	2.78	4	5.56	11	15.28	17	24
UC	127	6	4.72	3	2.36	3	2.36	10	7.87	22	17

[1]The age of the 396 persons ranged from 18—70 with a median of 49 (IQR: 40.25—59); body mass index (BMI) ranged from 16—42 with a median of 25 (IQR: 22—31).

Chapter 8

Metagenomic Analysis of the Human Nose and the Human Oral Cavity Microbiotas

8.1 Introduction

After having generated the CAGs for the MetaHIT data we decided to apply the method to oral microbiome data. This is an interesting environment to study, as the oral microbiota has a great impact on human health and disease [27]. The mouth is connected to most of the internal part of the body, thus oral microorganisms do not only influence the site they inhabit, but also the rest of the human body [27]. Additionally, the oral cavity is characterized by extreme conditions, to which the microbiota has adapted [60, 118].

The initial plan for this project was to sequence saliva samples collected from the MetaHIT test subjects in order to investigate their oral microbiome. This could in turn be compared to the gut microbiome. The sequencing was never conducted. However, we did perform initial studies on DNA extraction for sequencing, which I have chosen to include in this chapter. We proceeded to apply the CAG clustering to metagenomics data from oral and nose samples made available by the Human Microbiome Project (HMP). The nose samples were included because this site is in very close proximity to the mouth. Thus, the hypothesis was that we might find a range of CAGs that recur in both environments, but maybe with some site specificity that could be interesting. Hence, there might be gene sets that are necessary for the survival in the oral cavity that are not found in the nose and vice versa. Furthermore, the nose data was considerably smaller in size than the oral data. Thus, it could function as a test set to investigate how the clustering

performed. This introduction includes a brief description of the microbiotas which were investigated as well as of the Human Micorbiome Project.

8.1.1 The Human Oral Microbiota

The human oral cavity is a habitat for some 500-1200 different bacterial species depending on the source of the estimate [1, 27, 84, 143]. Approximately 280 of these have been cultivated and about 600 species have been validated by 16S analyses [27]. These numbers are the total pool of species found across test populations. However, in one single individual the number of species is markedly lower, usually somewhere between 80-200, but is likely, in some cases, as high as 500 species [1, 60, 84]. The easy access to the site and the importance of the oral microbiome to human health have made this one of the most studied communities on the planet [60].

The human oral microbiota is actually not one uniform community. Rather, it is made up of multiple site-specific communities, inhabiting both hard and soft tissues. Very big variations between microbiotas of different niches can be observed [27, 118]. In addition to the tissue sites there are also planktonic cells in the saliva. The mouth is connected to the trachea, esophagus, nose and middle ear, thus functioning as a gateway to the internal parts of the body [27].

The mouth is an environment characterized by constant saliva flow, large but short term, temperature changes and availability of nitrates and various types of carbohydrates in addition to the saliva, which is a very complex, but energy poor nutrient source [60]. Properties like oxygen, redox potential, pH and nutrient accessibility differ between oral sites [60, 118].

The oral microbiome has a great impact on the health of the host. Oral microorganisms are the cause of a range of oral diseases, such as caries, periodontitis, root canal infections, tonsillitis and alveolar osteitis [27]. Additionally, oral bacteria have been found (with varying amount of evidence) to be linked to non-oral diseases including pneumonia, stroke, cardiovascular disease, preterm birth, diabetes, endocarditis, meningitis and spondylodiscitis among others [27, 143, 156].

When looking at the health impact of the oral bacteria, it is important to mention that the commensal bacteria perform many beneficial functions. First of all they form biofilms, which function as a barrier for pathogenic bacteria [143]. This is very important due to the flow of bacteria through the oral cavity, illustrated by the fact that people on broad spectrum antibiotics in many cases develop bacterial or yeast infections in the mouth [84]. Through the degradation of nitrate, the oral microbiota aids in maintaining a healthy gastrointestinal system as well as cardiovascular system [84, 143].

As mentioned, bacteria are found as free cells in the saliva, but the majority of bacteria are located as biofilms on the various surfaces in the mouth. These biofilms have been studied extensively, especially the biofilms formed on the teeth, which constitutes the hard tissue of the oral environment [60, 84, 143]. The organisms in the biofilm adapt to this cooperative state of living, hence behaving very differently than free cells [45, 60]. Thus, it is important to consider the oral microbiota as a community, rather than individual organisms.

In addition to culture based methods, 16S and whole metagenome studies have been used to elucidate the composition and function of the oral microbiota [1, 11, 76, 96, 152, 154], with the biggest effort, in terms of number of samples, being the HMP [129]. Figure 8.1 shows the distribution of genera and phyla in oral sites (and the gut) in a healthy test population [115]. The most abundant phylum in the mouth is Firmicutes, of which *Streptococcus* is the most prevalent genus. Through the oral sites moving towards the gut, the ratio of Firmicutes to Bacteroidetes shifts towards the Bacteroidetes dominated state of the gut microbiota.

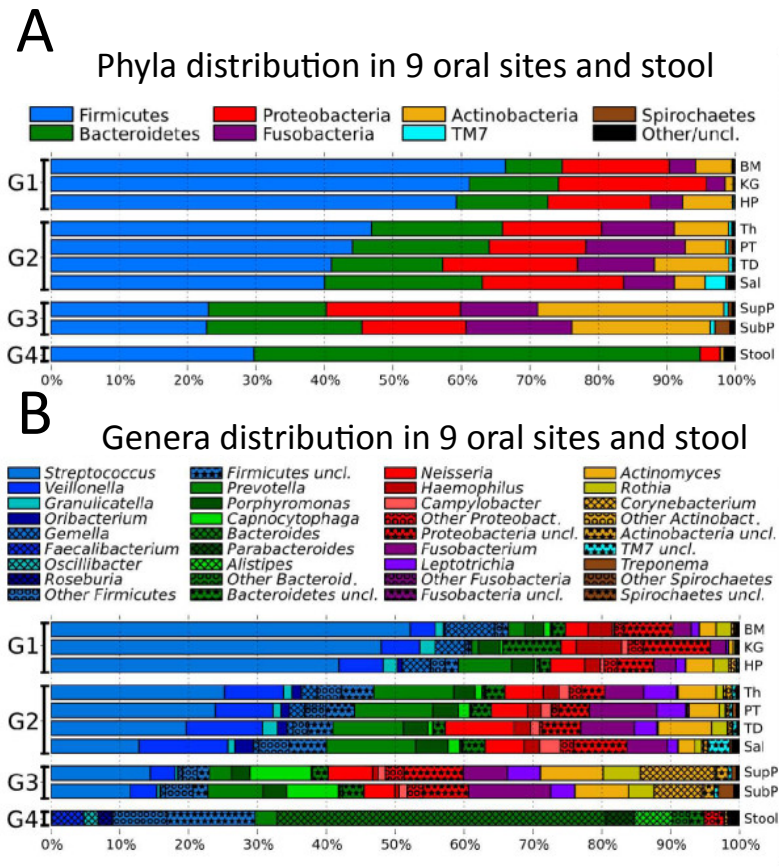


Figure 8.1: Distribution of phyla (A) and genera (B) in nine oral sites and the gut of a healthy test population. The sites are grouped in four groups (G1-G4) according to the Firmicutes/Bacteroidetes ratio. Group 1: Buccal mucosa (BM), keratinized gingiva (KG), hard palate (HP). Group 2: Throat (Th), palatine tonsils (PT), tongue dorsum (TD), saliva (Sal). Group 3: Supragingival plaque (SupP), subgingival plaque (SubP). Group 4: Stool. Adapted from Segata *et al.* [115]

8.1.2 The Human Nose Microbiota

The nose constitutes the upper part of the respiratory system. Here, the air is filtered before entering the airways and lungs [67]. Most of the studies regarding the nose microbiota have been focused on either one or a few species [67]. Especially, the carriage of the pathogen *Staphylococcus aureus* has received a great deal of attention, since people carrying this bacteria are at risk of server infections [42, 119, 153].

To the best of my knowledge, not many culture independent studies have been performed on the nose microbiota [67]. Although, samples from this site have been included in large scale studies like the HMP study [130] and the study by Costello *et al.* [20]. These and a few others have shown that healthy individuals mostly harbor Actinobacteria in the nose but also Proteobacteria and Firmicutes are commonly found in this site [36, 130].

8.1.3 The Human Microbiome Project

The Human Microbiome Project is funded by the US National Institute of Health (NIH) and is the largest resource of its kind. It includes samples from 242 healthy individuals across 18 body sites for women and 15 for men. Samples have been collected at up to three time points for each individual. Both 16S sequencing and whole metagenome sequencing have been performed. Additionally, single species have been isolated and sequenced [129].

The aim of this study is to define what constitutes a healthy human microbiota in order to use this as a baseline in studies regarding microbial impact on human health. Another part of it is to widen the reference databases for better annotation of metagenomes [129].

The sites that have been sampled include the gut (stool), oral cavity (buccal mucosa, hard palate, keratinized gingiva, palatine tonsils, saliva, subgingival plaque, supragingival plaque, throat, tongue dorsum), airway (anterior nares), skin (left antecubital fossa, left retroauricular crease, right antecubital fossa, right retroauricular crease) and vagina (mid-vagina, posterior fornix, vaginal introitus). All data is made publicly available [129].

8.2 DNA Extraction for Oral Microbiome Sequencing

The main concern about the DNA extraction from saliva samples was that we expected a substantial amount of the DNA to originate from the human host (80-90% [49, 65]). This proportion could possibly depend on the collection method of the saliva. This section describes how the DNA extraction procedure was optimized.

8.2.1 Methods

Sample Collection

Paraffin was used to stimulate saliva flow. The test subjects were asked to chew on the paraffin until it formed a uniform mass, which took approximately one minute. No saliva was collected at this stage. Subsequently, all saliva formed during the next 3 minutes of chewing on the paraffin, was collected. The samples were stored at -80°C. For all DNA extractions 1 ml of saliva was used.

Assessing Human DNA Content

DNA was extracted from a saliva sample using UltraClean Microbial DNA Isolation Kit supplied by MoBio Laboratories, Inc. This was sequenced in-house on the SOLiD platform, producing ~267 million singled end sequencing reads with an average read length of 75bp. The quality was assessed using the FastQC tool[1]. This was used for all quality checks for all the data in this project, and will not be described further. The reads were cleaned up using the genobox tool developed in-house. Bases with a quality below the threshold of Phred score 20 were trimmed of and whole reads with a quality average less than Phred score 20 were removed. Reads shorter than 25bp after trimming and reads containing "N"'s were also removed. This reduced the number of reads to ~152 million. These reads were mapped to the human reference genome build 37[2] using BWA [71] to calculate the abundance of human DNA in the sample. Subsequently, the reads were mapped to all bacterial reference genomes from NCBI (April 2011). Lastly the unmapped reads were mapped to a catalogue of oral microbiome genes obtained from the HMP data.

Removing Human DNA Prior to Sequencing

It was decided to test a method for depleting the samples of host DNA. The same sample was used as for assessing the human DNA content. We chose to apply the MolYsis® kit, Molzym GmbH & Co., which relies on differential lysis of prokaryotic and eukaryotic cells prior to DNA extraction. The applied DNA extraction kit was the same as above. This yielded approximately 200ng of DNA material. This was sequenced on the Illumina platform by Beijing Genome Institute (BGI). The data had been cleaned up by the provider and did not need trimming. This yielded ~29 million read pairs of a length of 90bp. All reads were mapped, using BWA [71], to the human genome build 37. The unmapped reads were subsequently mapped to bacterial genomes from NCBI (April 2011), then to the MetaHIT gene

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[2] http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

catalogue of 3.9million genes (see paper in Chapter 5) and lastly to the oral microbiome catalogue as described for the previous sample.

Testing DNA Extraction Methods

The UltraClean Microbial DNA Isolation Kit used for DNA extraction of the first samples did not to perform sufficiently stable (i.e. the resulting DNA concentration fluctuated between samples). Accordingly, we tested several protocols for DNA extraction, which can be seen in Table 8.1. The QIAamp DNA Blood Mini Kit provided by QIAGEN seemed to be the best performing method. Three samples from three individuals were collected, the DNA was extracted using the selected kit and sequenced in-house on the Ion Torrent platform. The samples will be referred to as Ion Torrent 1, 2 and 3 (IT-1, IT-2 and IT-3). This was first of all to validate that the DNA extraction kit performed stably and that the human versus non-human distribution was as expected and secondly to assess if the results were comparable between individuals, which is important when deciding on a sequencing depth at a later stage.

The read counts for the raw data were ~2.3 million, ~2.3 million and ~3.1 million for samples IT-1, IT-2 and IT-3, respectively. It was cleaned up in three steps. (1) Remove the first 9 bases from the 5' end using fastx_ trimmer provided in the FASTX-Toolkit[1]. (2) Quality trimming the reads to Phred score 20 using the DynamicTrim script provided with the SolexaQA software package [21]. (3) Hard trimming the ends to 200bp for one sample (IT-1) and 160bp for the last two (IT-2 and IT-3) and removing reads shorter than 40bp for all three samples. This reduced the read counts to ~1.4 million for IT-1, ~0.8 million for IT-2 and ~1.8 million for IT-3. The clean data was mapped to the human genome build 37, NCBI bacterial genomes (July 2012), the MetaHIT catalogue and the HMP oral microbiome catalogue like described for the other samples. However, BWA was not able to map the Ion Torrent data. Thus, the TMAP[2] mapper was chosen in this case.

[1] http://hannonlab.cshl.edu/fastx_ toolkit/
[2] https://github.com/iontorrent/TMAP

Table 8.1: DNA Extraction Methods Tested.

Method	DNA yield (μg/ml saliva)
MoBio Ultraclean	<0.01-1.3
Spin down + MoBio Ultraclean	<0.01
MoBio Powerlyzer	<0.01
Spin down + MoBio Powerlyser	<0.01
Freeze dry + MoBio Powerlyzer	<0.01
QiaAmp	10

8.2.2 Results

Based on the sequencing of a saliva sample on the SOLiD platform, we could determine that the human DNA content in our sample was within the range we expected, i.e. approximately 85%. The mapping results for this sample can be seen in Figure 8.2. The same sample was used when experimenting with removing human DNA prior to sequencing. It is evident from the results in Figure 8.3 that the host DNA depletion was successful.

Adding a DNA depletion step will inevitable introduce bias in the data. An indication of this bias can be seen in Figure 8.4, which displays the comparison of the non-human part of both the sample with and the sample without the depletion step. Hence, if there had been no bias, the bacterial fractions should have been the same, especially when the material originates from the same sample. It should be kept in mind that this is only based on one sample and the sequencing platforms differ.

Lastly, it was investigated if the DNA extraction method was reliable and how much the sample composition fluctuated between individuals in terms of human versus non-human DNA content. Figure 8.5 shows the mapping results for the three Ion Torrent samples from three individuals. The DNA was extracted successfully using the selected kit. As expected there is some variance between the three test subjects, especially IT-3 seems to have a larger proportion of non-human DNA compared to human DNA than the other two. IT-1 and IT-2 are very alike. Again, the human DNA content was within the expected range.

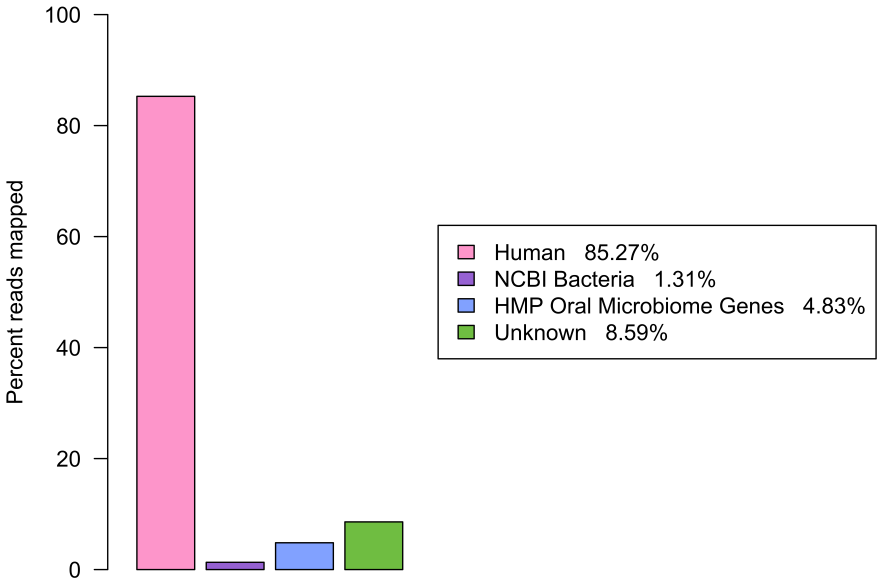


Figure 8.2: Mapping results for the saliva sample without adding a human DNA depletion step before sequencing.

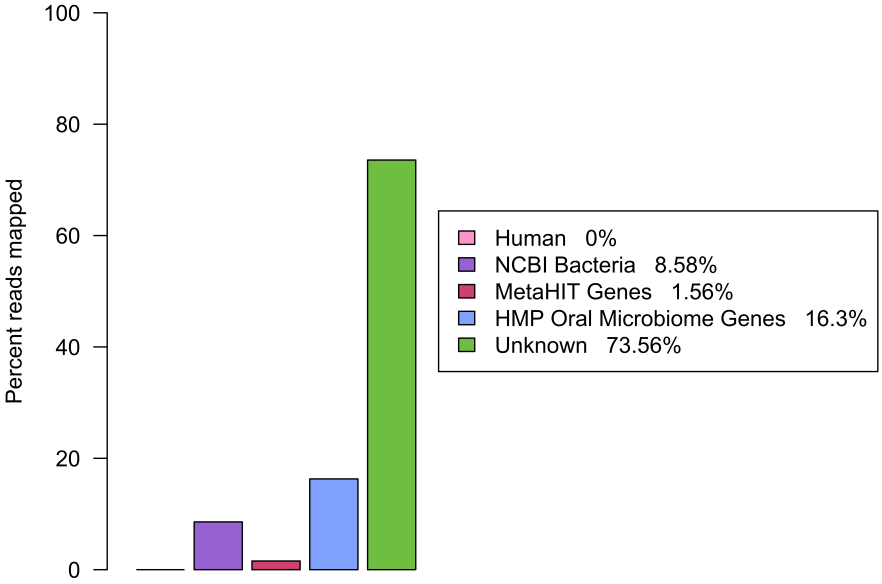


Figure 8.3: Mapping results for the saliva sample when adding a human DNA depletion step before sequencing.

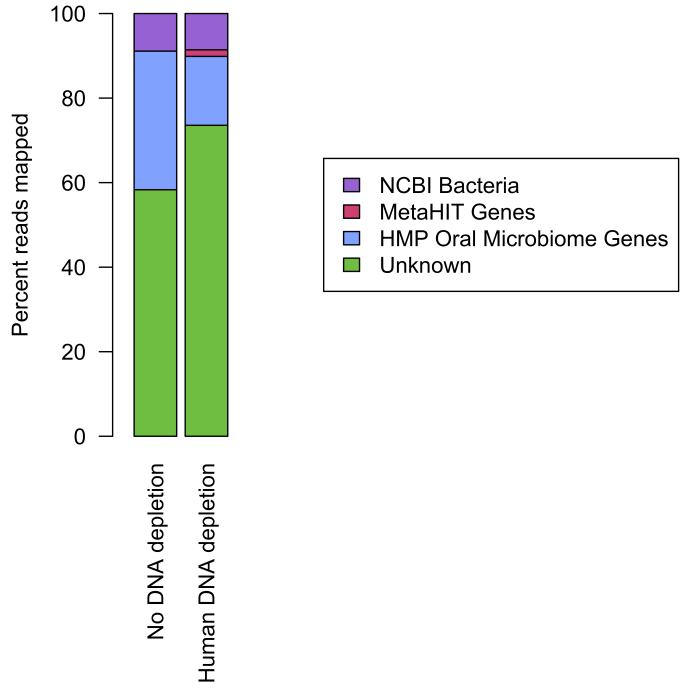


Figure 8.4: Comparison of the non-human mapping results with and without human DNA depletion.

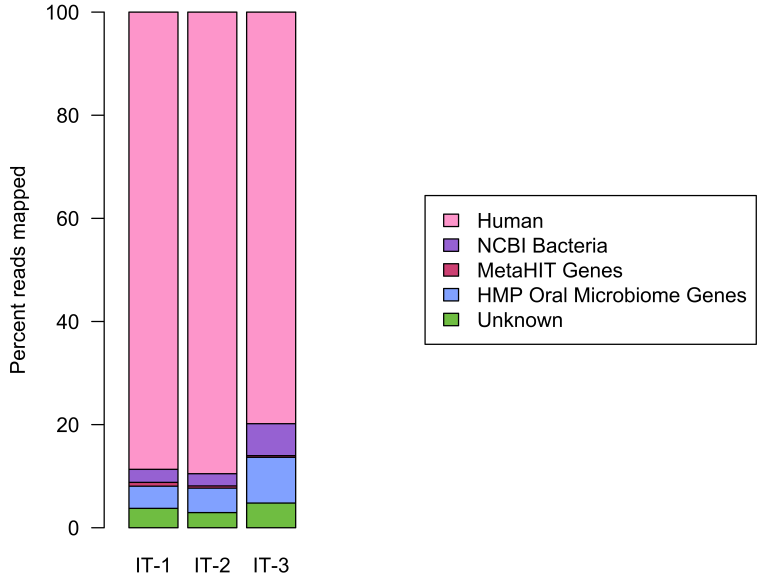


Figure 8.5: Mapping results comparing saliva samples from three individuals.

8.2.3 Conclusion

This experiment was a pilot study to investigate how best to handle the saliva samples we were planning to sequence. We verified that the human DNA content was within the 80-90% range we expected. We found that it was very efficient to remove human DNA using the MoLyzes® kit. However, we observe a bias in the non-human fraction of the data when applying this method (this might of course be partly due to differences between sequencing platforms). Lastly, we identified the QIAamp DNA Blood Mini Kit from QIAGEN to be the best option for DNA extraction.

If the sequencing of the saliva samples had been pursued, we would have omitted the DNA depletion step because we did not want the bias it introduces and since this DNA reduction result in very low amounts of DNA, which would cause problems in the downstream sequencing. The samples would then have had to be sequenced very deep in order to get an accurate detection of the microbiome. If aiming at an average depth of 4.5Gb per sample, which was the sequencing depth of the first MetaHIT fecal samples [104], it would require the sequencing depth to be somewhere close to 22Gb/sample. We might not have sequenced the samples that deep, thus accepting a lower coverage of the microbiome. Not removing the human DNA adds an extra layer of information, as it enables studying connections between the host genome and the microbiome.

8.3 Co-Abundance Gene Groups Clustering of Oral and Nose Metagenomic Samples

This section describes the results of the CAG clustering of the HMP samples. The method was applied successfully to the MetaHIT data and the hope was that it would be directly transferable the human nose and oral microbiome data.

8.3.1 Data

All available whole genome sequencing (WGS) samples from the nose and the oral cavity were downloaded from the HMP Data Analysis and Coordination Center[1]. The samples originated from one site in the nose, the anterior nares (i.e. nostrils) and from nine sites in the mouth including six soft tissue sites, two hard tissues sites and saliva. 90 nose samples were included out of 94 available. The excluded samples were technical replicates. All oral sites and the number of samples included in the analysis are listed in Table 8.2. Six additional samples were available, but they were excluded due to errors in the

[1] http://hmpdacc.org/, November 2012

Table 8.2: The sampling sites and the sample counts for samples included in the oral cavity CAG clustering.

Body site	Tissue Type	Number of samples
Attached keratinized gingiva	Soft	5
Buccal mucosa	Soft	121
Hard palate	Soft	1
Palatine tonsils	Soft	6
Saliva	Saliva	5
Subgingival plaque	Hard	8
Supragingival plaque	Hard	126
Throat	Soft	7
Tongue dorsum	Soft	136
Total		415

files. The number of samples was further reduced for some of the clustering runs, which will be described when presenting the results. Assemblies of most samples were also available. These were used for constructing the gene catalogue, which will be described later. Assemblies from stool samples were also included, because we wanted to compare the oral and nose results to the gut at a later stage and having a combined gene catalogue would make that easier.

8.3.2 Methods

The workflow for the data processing and clustering was introduced in Chapter 4. Here, only the specific details for this work will be described.

Pre-processing of Data

The sequencing data was downloaded as FASTQ files. Human DNA had already been removed and the reads had been quality trimmed to a quality score of 2, which was far from sufficient to remove the low quality data. FastQC[1] was run on all samples to assess the quality and identify possible problematic issues with the data. First step in the trimming of the data was to remove the adapters used for the sequencing. For this cutadapt [86] was applied. Adapter sequences were identified by use of FastQC, as no information regarding which adapters had been used was given in the data documentation. Different adapters seemed to have been applied, which was expected as the sequencing was performed at different labs. Only known adapter sequences were removed, as other repetitive sequences in the data

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

could be biologically important. In the same step as removing the adapters, the reads were also trimmed to a Phred score of 20 and only reads with a minimum length of 30bp were kept. Next step was to trim off 9 bases of the 5' end, as these bases generally were of very low quality, and to remove reads with "N"s in the sequence. After the trimming, some of the read pairs in the paired end files were not complete anymore. To correct for this a modified version of cmpfastq[1] was applied. The modification was to make it accept gzipped files, as all the files were gzipped. The read counts for the data after trimming can be seen in Table 8.3.

Gene Catalogue

The non-redundant gene catalogue was constructed by first predicting genes on the assemblies, provided by HMP, using MetaGeneMark [161] with the following settings, -a -d -f G -m MetaGeneMark_v1.mod. The modification file was provided by the developer. The genes called on all contigs from all samples were pooled and CD-HIT-EST [37, 74] was applied to remove redundancy. The cutoff for this clustering was 95% identity over 90% of the length of the shortest gene. The settings were, -c 0.95 -n 8 -M 102400 -l 100 -d 0 -aS 0.9 -B 1 -T 16 -g 1. The representative sequences from this clustering constituted the gene catalogue.

Abundance Matrix

All reads from all samples were mapped to the gene catalogue using the BWA aln program [70] with the default settings. Only reads that had one best hit were accepted, as reads mapping to several genes equally well would hinder proper downstream clustering. A count matrix was generated from the mapping results by counting the number of mapped reads to each gene for each sample. Single reads that mapped to a gene and a read pairs with both reads mapped to the same gene were counted as one observation. Whereas, if a read pair mapped to two genes this counted as two observations. The rational for this was that it is possible to hit two adjacent genes (e.g. in an operon), and this should be allowed for. The counts were normalized by dividing the counts by the length of the gene (excluding any "N"s in the sequence) and the total count of the sample. The matrix was modified so that genes observed in less than 3 samples were removed from the matrix, as well as genes for which more than 90% of the signal originated from three samples or less.

[1] http://compbio.brc.iop.kcl.ac.uk/software/cmpfastq.php

Table 8.3: Read count for the samples after trimming. Showing the minimum and maximum read counts, median read counts and total read counts. All counts are in 1000 reads.

Site	No. of samples	Min. read count (x1,000)	Max. read count (x1,000)	Median read count (x1,000)	Total read count (x1,000)
Anterior nares (nose)	90	13	6,034	695	109,493
Attached keratinized gingiva	5	1,690	7,996	5,242	24,067
Buccal mucosa	121	2	14,085	681	158,055
Hard palate	1	58,477	58,477	58,477	58,477
Palatine tonsils	6	198	3,538	387	5,829
Saliva	5	2,445	13,680	8,207	40,823
Subgingival plaque	8	9,760	29,086	25,472	173,715
Supragingival plaque	126	520	60,272	5,677	830,920
Throat	7	82	2,013	865	6,143
Tongue dorsum	136	753	74,595	6,493	1,164,051
Total					2,571,573

Clustering

The Co-abundance Gene Groups (CAG) clustering was performed on the constructed abundance matrix. The clustering program is an in-house developed C implementation, which has been benchmarked to work as the one used in the paper included in Chapter 5. The clustering settings are all Pearson correlation and were set as follows: Genes included in a cluster should have a correlation of 0.9 or better to the cluster center. The closeness cutoff was set to 0.6 or better. These are the genes that are included in the recalculation when cluster centers are moved. Clusters were merged if the cluster centers correlated more than 0.9. The canopy walks, i.e. recalculation of centers, were done maximum 3 times or until the latest walk was between two genes that correlate more than 0.995. Only CAGs that including 3 or more genes and which was observed in minimum 3 samples were included. For more details on the method see Chapter 4.

Taxonomic annotation

The taxonomic annotation was done by aligning the gene catalogue to the NCBI bacterial genomes database (April 2013) using BLASTn [5]. Hits with an E-value <= 10^{-5} were considered significant and the best hit for each gene was selected.

8.3.3 Results

Clustering of Nose Samples

CAG clustering was first performed on the 90 nose samples. The abundance matrix included 91,997 genes after the filtering. The clustering resulted in 1,623 CAGs ranging in size between 3 and 2,117 genes pr. cluster. The size distribution was heavily skewed towards small CAGs. Thus, 75% of the CAGs contained less than 10 genes. Only 9 CAGs contained more than 700 genes, which was the cutoff previously set for being a bacterial species sized CAG (see Chapter 5). Most of the clusters did not separate out properly from the rest of the genes. An example of this can be seen in Figure 8.7 top panel. This is likely due to the very low sequencing depth of these samples. Even with the less than optimal clustering, there is still some taxonomic consistency within the CAGs, as can be seen in Figure 8.6.

To remove the problematic CAGs, a limit was set as to how many genes were allowed to fall between 0.8 and 0.9 Pearson correlation to the cluster center, which is right outside the cluster limit. This was set to maximum 50% of the CAGs size, i.e. if the CAG had 1000 genes, no more than 500 genes were allowed between 0.8 and 0.9 correlation to that CAG. This reduced the number of CAGs to 32 of which 7 CAGs included more than 100

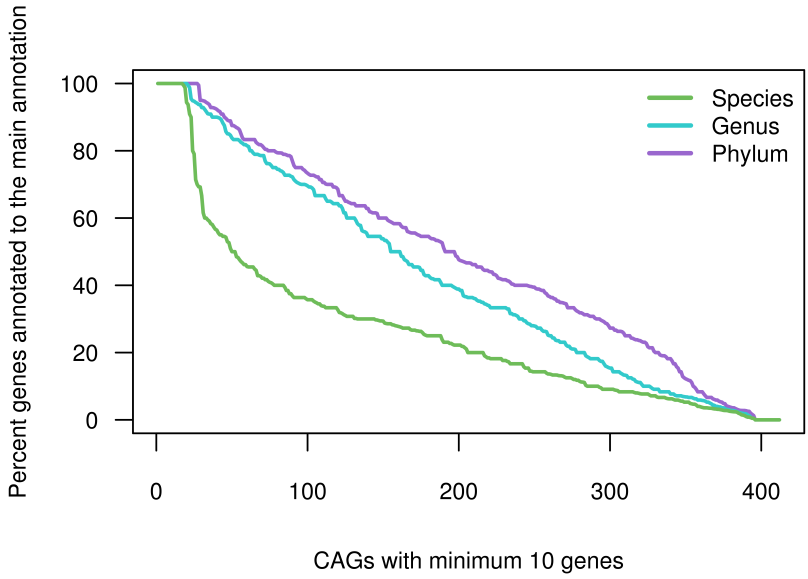


Figure 8.6: Plot showing the taxonomic consistency of the CAGs obtained by clustering anterior nares samples. On the x-axis are all the CAGs with a size larger than 10 genes. The y-axis denotes the percent genes that are annotated to the top annotation. This is shown for species, genus and phylum annotation.

genes and 21 contained 10 or less genes. These 32 CAGs had a much better separation and could be candidates for further investigation. An example of the separation of two such CAGs is shown in Figure 8.7 bottom panel.

The largest of all the CAGs, CAG:1, which is one of the clusters that separate out properly, could be annotated as *Staphylococcus aureus*, having 99.5% of the genes annotated to this species. As described in the introduction to this chapter, this is a bacteria commonly found in the anterior nares of humans. According to the NCBI genome database it encodes in the range of 2,600 to 2,900 genes depending on the strain. CAG:1 includes 2,117, hence, the clustering captures this bacteria quite well despite the issues we had with this data. It could be because this bacteria is on of the most abundant across the sampled individuals, which would make it more likely to sample it in these low depth samples.

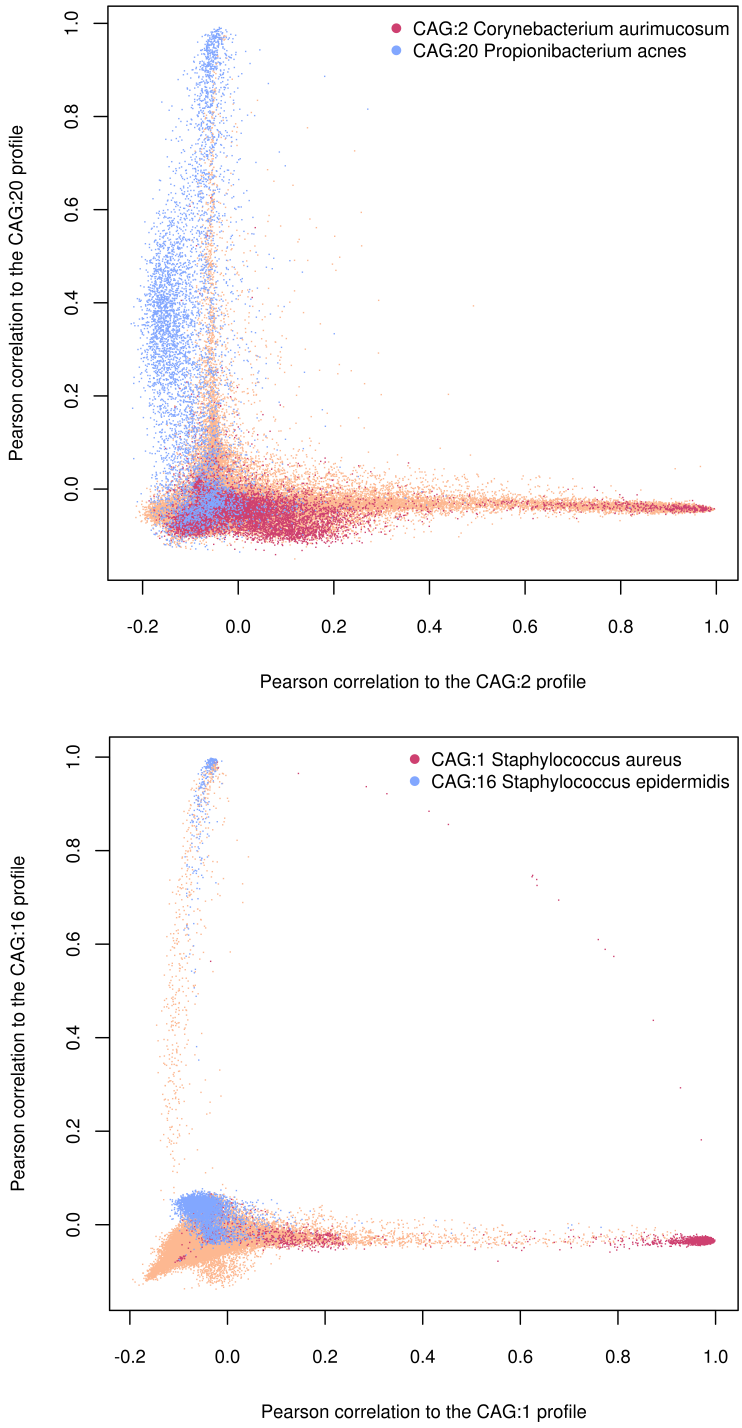


Figure 8.7: Scatter plots showing the Pearson correlation of all genes (one data point pr. gene) to the selected CAG centers. The dots are overlaid with color according to most prevalent annotation for both CAGs. Top: All genes versus the CAG:2 (size: 1420 genes) and CAG:20 (size: 320 genes) profiles. Bottom: All genes versus the CAG:1 (size: 2117 genes) and CAG:16 (size: 372 genes) profiles.

Clustering of Oral Samples

All of the oral samples were included in the input abundance matrix for the first clustering of the oral data, thus consisting of 415 samples and 2,967,572 genes after filtering. The clustering resulted in 11,297 CAGs in the size range of 3 to 8,250 genes. This was again dominated by small sized CAGs, hence 80% included less than 10 genes. The separation issues were more profound than what was observed for the clustering of the anterior nares samples. 133 of the CAGs that included more than 100 genes passed the criterion regarding the number of genes allowed to fall between 0.8 to 0.9 Pearson correlation, which was described for the nose samples. However, only a few of these were, by manual inspection, found to be separated properly. Experimenting with the settings for the clustering did not resolve the problem.

The poor clustering could be due to the substantial variation in sequencing depth of the samples (from approximately 2,000 to 74,000,000 reads pr. sample) and also a few of the samples were replicates. Thus, the replicate samples and samples having less than 10,000,000 reads mapping to the gene catalogue were removed from the dataset. Cutoffs at 1,000,000 and 2,300,000 mapped reads were also tried out, but this will not be described further. The reduced dataset included 275 samples and 2,880,105 genes after filtering. Clustering of this dataset resulted in 2,801 CAGs ranging in size between 3 and 9,688 genes. 44% of the clusters containing 10 genes or less. The reason for the big difference in CAG numbers compared to the clustering of all samples is that the clustering of all samples was allowed to run for a longer time, i.e. these datasets are too big for the clustering to complete fully within a reasonable time frame and the longer it runs the more small CAGs are picked up. Analyzing the CAGs with more than 100 genes, 151 passed the 50% criterion for the 0.8 to 0.9 Pearson correlation to the center. This was again not enough to properly correct for the clustering problems, although a few acceptable clusters were identified by further manual inspection.

We hypothesized that the clustering of samples from diverse sites simultaneously might not be possible. Hence, we continued to only clustering samples from one site at a time. 133 tongue dorsum samples were clustered as well as 111 supragingival plaque samples. These sites were chosen because of the number of samples for each site and the sequencing depth was reasonable. The results from clustering the two sites were comparable, thus only the results for the tongue samples will be described. The abundance matrix included 2,382,512 genes after filtering. These clustered into 5,543 CAGs, the largest CAG including 3,557 genes. 80% of the CAGs were small, thus containing less than 10 genes. After applying the 50% cutoff to how many genes were allowed between 0.8 and 0.9 Pearson correlation to the cluster center, 59 CAGs with a minimum size of 100 genes were left, of which 31 were of a species size, i.e. contained more than 700 genes. Again this criterion was not effective in distinguishing nicely clustered CAGs and as in the results above only a few reasonable clusters could be picked out by

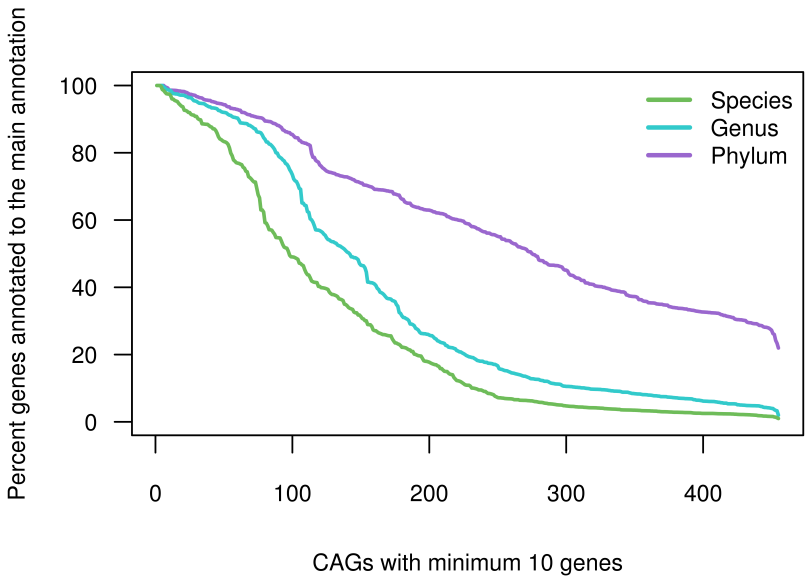


Figure 8.8: Taxonomic consistancy of the CAGs obtained from the clustering of tongue dorsum samples. On the x-axis are all the CAGs with a size larger than 10 genes. The y-axis denotes the percent genes that are annotated to the top annotation. This is shown for species, genus and phylum annotation.

visual inspection. As for the clustering of the nose samples, even with the non-optimal clustering, we still see some taxonomic consistency within the CAGs (see Figure 8.8).

8.3.4 Discussion and Perspectives

The clustering method for generating Co-Abundance Gene Groups showed great promises as a tool for deeper analysis of metagenomics datasets when applied to the MetaHIT dataset. However, in this case it was not possible to get the genes separated into CAGs to a satisfactory level. The settings of the clustering could not be adjusted to correct for poor separation. Removing low sequencing depth samples did not significantly improve the results and neither did only clustering samples originating from a single site.

We do not know exactly what gives rise to these problems. Maybe the method is sensitive to which genes are included when constructing the abundance matrix. Hence, in this study, all reads were mapped to a catalogue

including genes from all samples. It could be that this is not possible and that only genes from the samples in question should be included. We applied the full gene catalogue in the construction of all abundance matrices, as it would make downstream comparisons of the CAGs easier. It could also be that the settings for generating the non-redundant gene set is not optimized for these samples or the mapping needs to be more strict.

Better measures of how well the clustering has performed and how to distinguish a nicely separated cluster from the rest need to be established. In this work most of this assessment was done by visual inspection of different segments of the data and not on quantifiable values. These types of data are simply too big for this approach. We have tried to set up some guidelines, such as how many genes to allow close to a cluster, but this did not properly define ”good” versus ”bad” clusters.

The taxonomic consistency we see in the data and the fact that we capture the *Staphylococcus aureus* quite well in the nose data indicates that the clustering is working to some extent and it should be possible to improve on this. However, more work needs to be done in order to get the results we are looking for.

Epilogue

Chapter 9

Conclusion

This thesis presents projects which concern the human microbiota, with the main focus on the human intestinal tract microbiome. Furthermore, results regarding analyses of the nose- and oral-cavity microbiotas are described. The work was centered around a clustering method that bins genes from metagenomic samples according to their co-abundance under the assumption that genes located on the same DNA would co-vary in abundance across samples. These bins have been named co-abundance gene groups (CAGs). The ability to define which genes originate from the same organism, or other genetic element, is very valuable for obtaining a deeper understanding of the structure and function of metagenomes.

The CAG clustering method was first applied to the MetaHIT data. MetaHIT is a large-scale metagenomics study of the human intestinal tract. The dataset included 296 samples from which a gene catalogue of ~3.9 million non-redundant coding genes were obtained. These genes were clustered into 7,381 CAGs of which 741 had a size corresponding to that of a bacterial genome. Dependency-associations were observed between CAGs, some of which were phages and clone-specific elements that were dependent on the presence of their host. Relationships between CRISPR-elements and phages were also observed. Additionally, CAGs were identified that influenced the persistence of the host bacteria in the human gut. This effect could either be positive or negative. These results were described in the manuscript "Variable gene modules predict persistence of microbes in the human gut", which is included in this thesis.

Two additional projects which were based on the CAG clustering of human gut microbiome data are described in this thesis. The first study was an effort to investigate if bile acid degradation by intestinal bacteria was

associated to the BMI of the test population. Since bile acid degradation products are known to be involved in lipid and glucose homeostasis, this hypothesis seemed plausible. We identified a slight correlation between BMI and presence of CAGs encoding the first gene in the bile acid degradation pathway. This work does not give a final conclusion, but it indicates that the hypothesis could be true. The second project describes the prevalence of the gut parasite *Blastocystis* in the MetaHIT data. We identified four CAGs that could be annotated as four subtypes of *Blastocystis*. The occurrence of these four subtypes in our cohort was very similar to what has previously been observed. *Blastocystis* infection seemed to be less frequent in individuals with a *Bacteroides*-driven enterotype. To the best of my knowledge, this is the first time the presence of this parasite in a large test population has been investigated using metagenomics data. The results are described in the included manuscript "A Metagenomic Approach to Studying Intestinal Microbial Eukaryotes".

This thesis also describes the analysis of the human oral and nose microbiomes. The data was obtained from the Human Microbiome Project (HMP) and included 505 whole metagenome sequencing samples in total. The CAG clustering method was applied to this data in the attempt to obtain a deeper understanding of these ecosystems. Clustering of the nose samples resulted in 32 CAGs that were separated fairly well from the rest of the genes. One of these CAGs captured the bacterium *Staphylococcus aureus* with a high coverage. Three subsets of the oral samples were clustered. However, all three oral clustering runs only resulted in a few acceptable CAGs. We had expected this method to be directly transferable to these other datasets, but this was clearly not the case. We conclude that further optimization of the process is needed before CAG clustering can be used as a tool for analyzing all types of metagenomic datasets.

Lastly, preliminary experiments regarding DNA extraction from saliva samples for metagenomics sequencing were described. It was concluded that the best practice would be to use the QiaAmp DNA Blood Mini Kit without adding a human DNA depletion step. Deep sequencing would be necessary, as approximately 80-90% of the reads would be human.

In conclusion, this thesis describes CAG clustering of the human gut microbiome data as a valuable tool for better understanding the system. It was used for describing important topics like the interplay between species as well as other elements of the system, factors that might influence development of obesity and the prevalence of eukaryotic parasites in the gut. However, there are many other aspects of the human microbiome that could be studied based on CAG clustering, although it needs further improvement before it can be directly transferred to any given dataset.

Chapter 10

Future Perspectives

It was pointed out by Chistoserdova [18] in 2010 that the metagenomics field had yet to establish a golden standard for data analysis and this still seems to be the case. Development of a common practice will enable easier comparison of results and faster sample analysis throughput. With respect to sequencing, which is the corner stone in metagenomics, technologies are improving constantly and, with platforms such as PacBio, we are moving towards significantly longer reads[1] which will greatly simplify the data processing of metagenomics studies.

At present, metagenomic datasets are fragmented and imperfect. Thus, methods like co-abundance gene groups clustering are valuable tools in reassembling the ecosystem under investigation. This is an important step in the data analysis, as the time, in my opinion, is running out on large-scale metagenomics studies of the human microbiome that only define the overall species composition and functional potential of the system. The field is shifting towards trying to understand the dynamics of the various niches and defining which of the organisms are responsible for what functions and how that affects the host. To this end, methods like the CAG clustering are very valuable.

Numerous papers have, during the last decade, been published regarding the human microbiome [1, 3, 20, 36, 42, 130, 54, 65, 67, 104]. However, these have mostly focused on the bacterial fraction of the systems, while other parts of the microbiome such as plasmids, phages and micro-eukaryotes have been neglected [51, 90]. With better methods for binning the data, it will be possible to also describe this part of the data in more depth. Additionally,

[1] http://www.pacificbiosciences.com/

unknown species in the system can be picked out and studied further.

Annotation of metagenomic datasets is challenged by the lack of reference genomes. Even when studying the human microbiome, which has been the target for many microbiology studies through history, there is still a large part of the data that does not have a proper match in any database. It is improving all the time, especially with efforts like the HMP [129], but there is a lot of work to be done in this area. In this respect single genome sequencing is still immensely valuable.

The human microbiota is very important for human health. Better understanding of the mechanisms involved in this relationship will enable better treatment of many illnesses. Fecal transplants have already been performed, which was very effective in treating *Clostridium difficile* infections and metabolic syndrome [140]. In the future it will be possible to better modify the microbiota in order to treat various conditions, most likely in a more controlled manner than transplants of whole communities from one person to another. To this end, knowledge regarding the effects the microbial organisms have on each other and the host is necessary.

Through comprehensive studies like HMP and MetaHIT, we have gotten an insight into the genetic potential of the human microbiome. Hopefully future large-scale studies will emerge, which integrate metagenomics with metabolomics, metaproteomics and metatranscriptomics to also describe the activity of the genes [69].

The study of the human microbiome in health and disease has gotten off to a great start and a lot of knowledge has been collected, especially within the next generation sequencing era. However, there are still many pending questions to answer and I believe we can look forward to many exciting results in the years to come. I am sure these will lead to better understanding and treatment of numerous illnesses, thus improving the quality of life for many people.

Bibliography

[1] Aas J.r.A., Paster B.J., Stokes L.N., Olsen I., and Dewhirst F.E. (2005). Defining the Normal Bacterial Flora of the Oral Cavity. *Journal of clinical microbiology*, 43(11):5721–32.

[2] Abubucker S., Segata N., Goll J., Schubert A.M., Izard J., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology*, 8(6):e1002358.

[3] Ahmad N.N., Pfalzer A., and Kaplan L.M. (2013). Roux-en-Y gastric bypass normalizes the blunted postprandial bile acid excursion associated with obesity. *International journal of obesity*, 37(12):1553–9.

[4] Albertsen M., Hugenholtz P., Skarshewski A., Nielsen K.r.L., Tyson G.W., et al. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology*, 31(6):533–8.

[5] Altschul S.F., Madden T.L., Schäffer a.a., Zhang J., Zhang Z., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

[6] Amann R.I., Ludwig W., and Schleifer K.H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–69.

[7] Arumugam M., Harrington E.D., Foerstner K.U., Raes J., and Bork P. (2010). SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics (Oxford, England)*, 26(23):2977–8.

[8] Arumugam M., Raes J., Pelletier E., Le Paslier D., Yamada T., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–80.

[9] Bart A., Wentink-Bonnema E.M.S., Gilis H., Verhaar N., Wassenaar C.J.a., et al. (2013). Diagnosis and subtype analysis of Blastocystis sp. in 442 patients in a hospital setting in the Netherlands. *BMC infectious diseases*, 13:389.

[10] Basak S., Rajurkar M.N., and Mallick S.K. (2014). Detection of Blastocystis hominis: a controversial human pathogen. *Parasitology research*, 113(1):261–5.

[11] Belda-Ferre P., Alcaraz L.D., Cabrera-Rubio R., Romero H., Simón-Soro A., et al. (2012). The oral metagenome in health and disease. *The ISME journal*, 6(1):46–56.

[12] Brady A. and Salzberg S.L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods*, 6(9):673–6.

[13] Brestoff J.R. and Artis D. (2013). Commensal bacteria at the interface of host metabolism and the immune system. *Nature immunology*, 14(7):676–84.

[14] Carr R., Shen-Orr S.S., and Borenstein E. (2013). Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. *PLoS computational biology*, 9(10):e1003292.

[15] Chan C.K.K., Hsu A.L., Halgamuge S.K., and Tang S.L. (2008). Binning sequences using very sparse labels within a metagenome. *BMC bioinformatics*, 9:215.

[16] Chaucheyras-Durand F. and Durand H. (2010). Probiotics in animal nutrition and health. *Beneficial microbes*, 1(1):3–9.

[17] Chen T., Yu W.H., Izard J., Baranova O.V., Lakshmanan A., et al. (2010). The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database : the journal of biological databases and curation*, 2010:baq013.

[18] Chistoserdova L. (2010). Recent progress and new challenges in metagenomics for biotechnology. *Biotechnology letters*, 32(10):1351–9.

[19] Cock P.J.a., Fields C.J., Goto N., Heuer M.L., and Rice P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71.

[20] Costello E., Lauber C., and Hamady M. (2009). Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–7.

[21] Cox M.P., Peterson D.a., and Biggs P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC bioinformatics*, 11:485.

[22] Coyle C.M., Varughese J., Weiss L.M., and Tanowitz H.B. (2012). Blastocystis: to treat or not to treat... *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 54(1):105–10.

[23] Dalevi D., Ivanova N.N., Mavromatis K., Hooper S.D., Szeto E., et al. (2008). Annotation of metagenome short reads using proxygenes. *Bioinformatics (Oxford, England)*, 24(16):i7–13.

[24] Dave M., Higgins P.D., Middha S., and Rioux K.P. (2012). The human gut microbiome: current knowledge, challenges, and future directions. *Translational research: the journal of laboratory and clinical medicine*, 160(4):246–57.

[25] De Filippo C., Ramazzotti M., Fontana P., and Cavalieri D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in bioinformatics*, 13(6):696–710.

[26] Devaraj S., Hemarajata P., and Versalovic J. (2013). The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clinical chemistry*, 59(4):617–28.

[27] Dewhirst F.E., Chen T., Izard J., Paster B.J., Tanner A.C.R., et al. (2010). The human oral microbiome. *Journal of bacteriology*, 192(19):5002–17.

[28] Diaz N.N., Krause L., Goesmann A., Niehaus K., and Nattkemper T.W. (2009). TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics*, 10:56.

[29] Dinis J.M., Barton D.E., Ghadiri J., Surendar D., Reddy K., et al. (2011). In search of an uncultured human-associated TM7 bacterium in the environment. *PloS one*, 6(6):e21280.

[30] Edgar R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, 26(19):2460–1.

[31] Eloe-Fadrosh E.a. and Rasko D.a. (2013). The human microbiome: from symbiosis to pathogenesis. *Annual review of medicine*, 64:145–63.

[32] English A.C., Richards S., Han Y., Wang M., Vee V., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one*, 7(11):e47768.

[33] Ewing B. and Green P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3):186–94.

[34] Filoche S., Wong L., and Sissons C.H. (2010). Oral biofilms: emerging concepts in microbial ecology. *Journal of dental research*, 89(1):8–18.

[35] Finucane M.M., Sharpton T.J., Laurent T.J., and Pollard K.S. (2014). A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. *PloS one*, 9(1):e84689.

[36] Frank D.N., Feazel L.M., Bessesen M.T., Price C.S., Janoff E.N., et al. (2010). The human nasal microbiota and Staphylococcus aureus carriage. *PloS one*, 5(5):e10598.

[37] Fu L., Niu B., Zhu Z., Wu S., and Li W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23):3150–2.

[38] Gerlach W. and Stoye J. (2011). Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic acids research*, 39(14):e91.

[39] Gilbert J.A., Meyer F., Antonopoulos D., Balaji P., Brown C.T., et al. (2010). Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Standards in genomic sciences*, 3(3):243–8.

[40] Goll J., Rusch D.B., Tanenbaum D.M., Thiagarajan M., Li K., et al. (2010). METAREP: JCVI metagenomics reports–an open source tool for high-performance comparative metagenomics. *Bioinformatics (Oxford, England)*, 26(20):2631–2.

[41] Gori F., Folino G., Jetten M.S.M., and Marchiori E. (2011). MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics (Oxford, England)*, 27(2):196–203.

[42] Grice E., Kong H., Conlan S., and Deming C. (2009). Topographical and Temporal Diversity of the Human Skin Microbiome. *science*, 324(5931):1190–92.

[43] Handelsman J., Rondon M.R., Brady S.F., Clardy J., and Goodman R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):R245–9.

[44] Hildebrand F., Ebersbach T., Nielsen H.B.r., Li X., Sonne S.B., et al. (2012). A comparative analysis of the intestinal metagenomes present in guinea pigs (Cavia porcellus) and humans (Homo sapiens). *BMC genomics*, 13:514.

[45] Huang R., Li M., and Gregory R.L. (2011). Bacterial interactions in dental biofilm. *Virulence*, 2(5):435–44.

[46] Huson D.H., Mitra S., Ruscheweyh H.J., Weber N., and Schuster S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome research*, 21(9):1552–60.

[47] Hyatt D., LoCascio P.F., Hauser L.J., and Uberbacher E.C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics (Oxford, England)*, 28(17):2223–30.

[48] Jacobsen U.P., Nielsen H.B.r., Hildebrand F., Raes J., Sicheritz-Ponten T., et al. (2013). The chemical interactome space between the human host and the genetically defined gut metabotypes. *The ISME journal*, 7(4):730–42.

[49] James C., Iwasiow R.M., and Birnboim H.C. (2011). Human genomic DNA content of saliva samples collected with the Oragene ® self-collection kit. *DNA Genotek, Inc.; Ottawa*.

[50] Jones B. and Begley M. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36):13580–13585.

[51] Jones B.V. (2010). The human gut mobile metagenome: a metazoan perspective. *Gut microbes*, 1(6):415–31.

[52] Jones M.L., Martoni C.J., Ganopolsky J.G., Labbé A., and Prakash S. (2014). The human microbiome and bile acid metabolism: dysbiosis, dysmetabolism, disease and intervention. *Expert opinion on biological therapy*, 14(4).

[53] Karlsson F., Tremaroli V., Nielsen J., and Bäckhed F. (2013). Assessing the human gut microbiota in metabolic diseases. *Diabetes*, 62(10):3341–9.

[54] Karlsson F.H., Tremaroli V., Nookaew I., Bergström G., Behre C.J., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103.

[55] Kelly C.R., de Leon L., and Jasutkar N. (2012). Fecal microbiota transplantation for relapsing Clostridium difficile infection in 26 patients: methodology and results. *Journal of clinical gastroenterology*, 46(2):145–9.

[56] Kelsen J. and Wu G. (2012). The gut microbiota, environment and diseases of modern society. *Gut microbes*, 3(4):374–82.

[57] Kent W.J. (2002). BLAT–the BLAST-like alignment tool. *Genome research*, 12(4):656–64.

[58] Kim B.S., Jeon Y.S., and Chun J. (2013). Current status and future promise of the human microbiome. *Pediatric gastroenterology, hepatology & nutrition*, 16(2):71–9.

[59] Kim M., Lee K.H., Yoon S.W., Kim B.S., Chun J., et al. (2013). Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era. *Genomics & informatics*, 11(3):102–13.

[60] Kolenbrander P.E., Palmer R.J., Periasamy S., and Jakubovics N.S. (2010). Oral multispecies biofilm development and the key role of cell-cell distance. *Nature reviews. Microbiology*, 8(7):471–80.

[61] Koren O., Knights D., Gonzalez A., Waldron L., Segata N., et al. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS computational biology*, 9(1):e1002863.

[62] Lan Y., Kriete A., and Rosen G. (2013). Selecting age-related functional characteristics in the human gut microbiome. *Microbiome*.

[63] Landy J., Al-Hassi H.O., McLaughlin S.D., Walker A.W., Ciclitira P.J., et al. (2011). Review article: faecal transplantation therapy for gastrointestinal disease. *Alimentary pharmacology & therapeutics*, 34(4):409–15.

[64] Langmead B., Trapnell C., Pop M., and Salzberg S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

[65] Lazarevic V., Whiteson K., Gaïa N., Gizard Y., Hernandez D., et al. (2012). Analysis of the salivary microbiome using culture-independent techniques. *Journal of clinical bioinformatics*, 2(4).

[66] Le Chatelier E., Nielsen T., Qin J., Prifti E., Hildebrand F., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–6.

[67] Lemon K., Klepac-Ceraj V., and Schiffer H. (2010). Comparative analyses of the bacterial microbiota of the human nostril and oropharynx. *MBio*, 1(3).

[68] Leone V., Chang E.B., and Devkota S. (2013). Diet, microbes, and host genetics: the perfect storm in inflammatory bowel diseases. *Journal of gastroenterology*, 48(3):315–21.

[69] Lepage P., Leclerc M.C., Joossens M., Mondot S., Blottière H.M., et al. (2013). A metagenomic insight into our gut's microbiome. *Gut*, 62(1):146–58.

[70] Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

[71] Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

[72] Li R., Yu C., Li Y., Lam T.W., Yiu S.M., et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England)*, 25(15):1966–7.

[73] Li W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC bioinformatics*, 10:359.

[74] Li W. and Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–9.

[75] Lingner T., Asshauer K.P., Schreiber F., and Meinicke P. (2011). CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic acids research*, 39:W518–23.

[76] Liu B., Faller L.L., Klitgord N., Mazumdar V., Ghodsi M., et al. (2012). Deep sequencing of the oral microbiome reveals signatures of periodontal disease. *PloS one*, 7(6):e37919.

[77] Lorenzi H.a., Hoover J., Inman J., Safford T., Murphy S., et al. (2011). TheViral MetaGenome Annotation Pipeline(VMGAP):an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Standards in genomic sciences*, 4(3):418–29.

[78] Lynch K.M. (1916). BLASTOCYSTIS HOMINIS: ITS CHARACTERISTICS AND ITS PREVALENCE IN INTESTINAL CONTENT AND FECES IN SOUTH CAROLINA. *Journal of bacteriology*, 2(4):369–77.

[79] Maccaferri S., Biagi E., and Brigidi P. (2011). Metagenomics: key to human gut microbiota. *Digestive diseases (Basel, Switzerland)*, 29(6):525–30.

[80] Maillet N., Lemaitre C., Chikhi R., Lavenier D., and Peterlongo P. (2012). Compareads: comparing huge metagenomic experiments. *BMC bioinformatics*, 13(Suppl 19):S10.

[81] Markowitz V.M., Chen I.M.a., Chu K., Szeto E., Palaniappan K., et al. (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic acids research*, 40(Database issue):D123–9.

[82] Markowitz V.M., Chen I.M.a., Palaniappan K., Chu K., Szeto E., et al. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic acids research*, 40(Database issue):D115–22.

[83] Markowitz V.M., Ivanova N.N., Szeto E., Palaniappan K., Chu K., et al. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic acids research*, 36(Database issue):D534–8.

[84] Marsh P.D. (2012). Contemporary perspective on plaque control. *British dental journal*, 212(12):601–6.

[85] Martin J., Sykes S., Young S., Kota K., Sanka R., et al. (2012). Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PloS one*, 7(6):e36427.

[86] Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12.

[87] McCallum A., Nigam K., and Ungar L.H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00*, pages 169–178. ACM Press, New York, New York, USA.

[88] Mchardy A.C. (2007). Accurate phylogenetic classification of variable-length DNA fragments. 4(1):63–72.

[89] Meyer F., Paarmann D., D'Souza M., Olson R., Glass E.M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9:386.

[90] Minot S., Bryson A., Chehoud C., Wu G.D., Lewis J.D., et al. (2013). Rapid evolution of the human gut virome. 110(30):12450–5.

[91] Mondot S., de Wouters T., Doré J., and Lepage P. (2013). The human gut microbiome and its dysfunctions. *Digestive diseases (Basel, Switzerland)*, 31(3-4):278–85.

[92] Monzoorul Haque M., Ghosh T.S., Komanduri D., and Mande S.S. (2009). SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics (Oxford, England)*, 25(14):1722–30.

[93] Morgan X.C., Segata N., and Huttenhower C. (2013). Biodiversity and functional genomics in the human microbiome. *Trends in genetics : TIG*, 29(1):51–8.

[94] Mulle J.G., Sharp W.G., and Cubells J.F. (2013). The gut microbiome: a new frontier in autism research. *Current psychiatry reports*, 15(2):337.

[95] Nagarajan N. and Pop M. (2013). Sequence assembly demystified. *Nature reviews. Genetics*, 14(3):157–67.

[96] Nasidze I., Li J., Quinque D., Tang K., and Stoneking M. (2009). Global diversity in the human salivary microbiome. *Genome research*, 19(4):636–43.

[97] Nelson K.E. (2013). Microbiomes. *Microbial ecology*, 65(4):916–9.

[98] Nicholson J.K., Holmes E., Kinross J., Burcelin R., Gibson G., et al. (2012). Host-gut microbiota metabolic interactions. *Science (New York, N.Y.)*, 336(6086):1262–7.

[99] Pagani I., Liolios K., Jansson J., Chen I.M.a., Smirnova T., et al. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 40(Database issue):D571–9.

[100] Philipp B. (2011). Bacterial degradation of bile salts. *Applied microbiology and biotechnology*, 89(4):903–15.

[101] Prakash T. and Taylor T.D. (2012). Functional assignment of metagenomic data: challenges and applications. *Briefings in bioinformatics*, 13(6):711–27.

[102] Pruitt K.D., Tatusova T., and Maglott D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33(Database issue):D501–4.

[103] Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., et al. (2012). The Pfam protein families database. *Nucleic acids research*, 40(Database issue):D290–301.

[104] Qin J., Li R., Raes J., Arumugam M., Burgdorf K.S., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

[105] Qin J., Li Y., Cai Z., Li S., Zhu J., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

[106] Rajilić-Stojanović M. (2013). Function of the microbiota. *Best practice & research. Clinical gastroenterology*, 27(1):5–16.

[107] Rakoff-Nahoum S., Coyne M.J., and Comstock L.E. (2013). An Ecological Network of Polysaccharide Utilization among Human Intestinal Symbionts. *Current biology: CB*, 24(1):40–49.

[108] Ram R., VerBerkmoes N., and Thelen M. (2005). Community proteomics of a natural microbial biofilm. *Science*, 1915(2005):1915–20.

[109] Rao K., Sekar U., Iraivan K.T., Abraham G., and Soundararajan P. (2003). Blastocystis hominis–an emerging cause of diarrhoea in renal transplant recipients. *The Journal of the Association of Physicians of India*, 51:719–21.

[110] Ridlon J.M., Kang D.J., and Hylemon P.B. (2006). Bile salt biotransformations by human intestinal bacteria. *Journal of lipid research*, 47(2):241–59.

[111] Robles Alonso V. and Guarner F. (2013). Linking the gut microbiota to human health. *The British journal of nutrition*, 109(Suppl 2):S21–6.

[112] Scanlan P.D. and Stensvold C.R. (2013). Blastocystis : getting to grips with our guileful guest. *Trends in Parasitology*, 29(11):523–29.

[113] Schloissnig S., Arumugam M., Sunagawa S., Mitreva M., Tap J., et al. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430):45–50.

[114] Segata N., Boernigen D., Tickle T.L., Morgan X.C., Garrett W.S., et al. (2013). Computational meta'omics for microbial community studies. *Molecular systems biology*, 9:666.

[115] Segata N., Haake S.K., Mannon P., Lemon K.P., Waldron L., et al. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome biology*, 13(6):R42.

[116] Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O., et al. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–4.

[117] Shoaie S., Karlsson F., Mardinoglu A., Nookaew I., Bordel S., et al. (2013). Understanding the interactions between bacteria in the human gut through metabolic modeling. *Scientific reports*, 3:2532.

[118] Simón-Soro a., Tomás I., Cabrera-Rubio R., Catalan M.D., Nyvad B., et al. (2013). Microbial geography of the oral cavity. *Journal of dental research*, 92(7):616–21.

[119] Sivaraman K., Venkataraman N., and Cole A. (2009). Staphylococcus aureus Nasal Carriage and its Contributing Factors. *Future microbiology*, 4(8):999–1008.

[120] Stanghellini V., Barbara G., Cremon C., Cogliandro R., Antonucci A., et al. (2010). Gut microbiota and related diseases: clinical features. *Internal and emergency medicine*, 5(Suppl 1):S57–63.

[121] Stensvold C.R., Nielsen H.V., Mø lbak K.r., and Smith H.V. (2009). Pursuing the clinical significance of Blastocystis–diagnostic limitations. *Trends in parasitology*, 25(1):23–9.

[122] Stensvold C.R., Smith H.V., Nagel R., Olsen K.E.P., and Traub R.J. (2010). Eradication of Blastocystis carriage with antimicrobials: reality or delusion? *Journal of clinical gastroenterology*, 44(2):85–90.

[123] Su X., Xu J., and Ning K. (2012). Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC systems biology*, 6(Suppl 1):S16.

[124] Sun S., Chen J., Li W., Altintas I., Lin A., et al. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic acids research*, 39(Database issue):D546–51.

[125] Sunagawa S., Mende D.R., Zeller G., Izquierdo-Carrasco F., Berger S.a., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, 10(12):1196–9.

[126] Sweeney T.E. and Morton J.M. (2013). The human gut microbiome: a review of the effect of obesity and surgically induced weight loss. *JAMA surgery*, 148(6):563–9.

[127] Tatusov R.L. (1997). A Genomic Perspective on Protein Families. *Science*, 278(5338):631–637.

[128] Teeling H., Waldmann J., Lombardot T., Bauer M., and Glöckner F.O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics*, 5:163.

[129] The Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature*, 486(7402):215–221.

[130] The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–14.

[131] The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic acids research*, 42(1):D191–8.

[132] Thomas T., Gilbert J., and Meyer F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3.

[133] Tremaroli V. and Bäckhed F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature*, 489(7415):242–9.

[134] Tringe S.G., von Mering C., Kobayashi A., Salamov A.a., Chen K., et al. (2005). Comparative metagenomics of microbial communities. *Science*, 308(5721):554–7.

[135] Trosvik P.l., Stenseth N.C., and Rudi K. (2010). Convergent temporal dynamics of the human infant gut microbiota. *The ISME journal*, 4(2):151–8.

[136] Turnbaugh P.J., Hamady M., Yatsunenko T., Cantarel B.L., Duncan A., et al. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–4.

[137] Turnbaugh P.J., Ley R.E., Mahowald M.a., Magrini V., Mardis E.R., et al. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31.

[138] Tyson G.W., Chapman J., Hugenholtz P., Allen E.E., Ram R.J., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43.

[139] Tzahor S., Man-Aharonovich D., Kirkup B.C., Yogev T., Berman-Frank I., et al. (2009). A supervised learning approach for taxonomic classification of core-photosystem-II genes and transcripts in the marine environment. *BMC genomics*, 10:229.

[140] Van den Abbeele P., Verstraete W., El Aidy S., Geirnaert A., and Van de Wiele T. (2013). Prebiotics, faecal transplants and microbial network units to stimulate biodiversity of the human gut microbiome. *Microbial biotechnology*, 6(4):335–40.

[141] Venter J.C., Remington K., Heidelberg J.F., Halpern A.L., Rusch D., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

[142] Vrieze A., Van Nood E., Holleman F., Salojärvi J., Kootte R.S., et al. (2012). Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology*, 143(4):913–16.

[143] Wade W.G. (2013). The oral microbiome in health and disease. *Pharmacological research : the official journal of the Italian Pharmacological Society*, 69(1):137–43.

[144] Wallace T.C., Guarner F., Madsen K., Cabana M.D., Gibson G., et al. (2011). Human gut microbiota and its relationship to health and disease. *Nutrition reviews*, 69(7):392–403.

[145] Wawrzyniak I., Poirier P., Viscogliosi E., Dionigia M., Texier C., et al. (2013). Blastocystis, an unrecognized parasite: an overview of pathogenesis and diagnosis. *Therapeutic Advances in Infectious Disease*, 1(5):167–178.

[146] Weerth C.D., Fuentes S., and de Vos W. (2013). Crying in infants: On the possible role of intestinal microbiota in the development of colic. *Gut microbes*, 4(5):416–21.

[147] Wheeler D.L., Barrett T., Benson D.a., Bryant S.H., Canese K., et al. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 36(Database issue):D13–21.

[148] Wommack K.E., Bhavsar J., Polson S.W., Chen J., Dumas M., et al. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in genomic sciences*, 6(3):427–39.

[149] Wooley J.C., Godzik A., and Friedberg I. (2010). A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667.

[150] Wu G.D., Chen J., Hoffmann C., Bittinger K., Chen Y.Y., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–8.

[151] Wu S., Zhu Z., Fu L., Niu B., and Li W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC genomics*, 12:444.

[152] Xie G., Chain P.S.G., Lo C.C., Liu K.L., Gans J., et al. (2010). Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Molecular oral microbiology*, 25(6):391–405.

[153] Yan M., Pamp S.J., Fukuyama J., Hwang P.H., Cho D.Y., et al. (2013). Nasal Microenvironments and Interspecific Interactions Influence Nasal Microbiota Complexity and S. aureus Carriage. *Cell host & microbe*, 14(6):631–40.

[154] Yang F., Zeng X., Ning K., Liu K.L., Lo C.C., et al. (2012). Saliva microbiomes distinguish caries-active from healthy human populations. *The ISME journal*, 6(1):1–10.

[155] Yooseph S., Sutton G., Rusch D.B., Halpern A.L., Williamson S.J., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS biology*, 5(3):e16.

[156] Zbinden A., Mueller N.J., Tarr P.E., Spröer C., Keller P.M., et al. (2012). Streptococcus tigurinus sp. nov., isolated from blood of patients with endocarditis, meningitis and spondylodiscitis. *International journal of systematic and evolutionary microbiology*, 62(Pt 12):2941–5.

[157] Zdobnov E. and Apweiler R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–48.

[158] Zhao Y., Wu J., Li J., and Zhou N. (2013). Gut microbiota composition modifies fecal metabolic profiles in mice. *Journal of proteome research*, 12:2987–99.

[159] Zheng H. and Wu H. (2010). Short Prokaryotic Dna Fragment Binning Using a Hierarchical Classifier Based on Linear Discriminant Analysis and Principal Component Analysis. *Journal of Bioinformatics and Computational Biology*, 08(06):995–1011.

[160] Zhu B., Wang X., and Li L. (2010). Human gut microbiome: the second genome of human body. *Protein & cell*, 1(8):718–725.

[161] Zhu W., Lomsadze A., and Borodovsky M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 38(12):e132.

Appendices

Appendix A: Methods - Variable gene modules predict persistence of microbes in the human gut

METHODS (ONLINE)

Sample description

396 stool samples from 177 Danish and 141 Spanish human individuals were collected (Supplementary Data 1). 124 of the samples were sequenced and used previously[16]. The Spanish samples include 13 individuals with Crohn's disease and 69 with ulcerative colitis. 78 of the Spanish individuals were sampled twice with, on average, 6 months between the samplings. The Danish samples include healthy individuals ranging in body mass index from 18 to 42. All were subjected to Illumina deep sequencing resulting in 4.5Gb sequence per sample on average, and a total of 23.2 billion high quality sequencing reads with an average length of 77 bp.

Construction of a non-redundant metagenomic gene catalogue

Illumina raw sequencing reads from 396 metagenomic samples (Supplementary Data 1) were processed using the MOCAT software package[29]. In brief, >23.2 Billion raw sequencing reads were filtered using the FastX software (http://hannonlab.cshl.edu/fastx_toolkit) with quality cutoff 20 and reads shorter than 30 bp were discarded. High-quality reads (92% of raw reads) were assembled into scaftigs using SOAPdenovo (version 1.05)[23]. Genes were predicted on 18.5 M scaftigs longer than 500 bp (35 Gbp in total) using MetaGeneMark[30]. Predicted genes from all samples (45.4 M in total) were clustered using BLAT[31] by single linkage. Any two genes with greater than 95% identity and covering more than 90% of the shorter gene were clustered together. Finally, cluster representatives shorter than 100 bp were discarded resulting in a set of 4,201,877 non-redundant genes. From this set, we removed genes that were considered spurious or likely originated from human, animals or plants were removed to yield a final set of 3,871,657 genes that formed the reference gene catalogue. For a comparison to our previous gene catalogue[12] see Supplementary Data 10.

Quantification of reference gene abundances

High-quality reads were mapped to the reference gene catalogue using the screen function in MOCAT[29]. Briefly, reads were mapped with SOAPaligner (version 2.21)[32] with options: –M 4 (find best hits), –l 30 (seed length), –r 1 (random assignment of multiple hits), and –v 5 (maximum number of mismatches). Mapped reads were subsequently filtered using a 30 bp length and 95% identity cutoff and gene-length normalized base counts were calculated using the soap.coverage script (available at: http://soap.genomics.org.cn/down/soap.coverage.tar.gz). For samples where 11 M or more sequence reads were obtained (n = 393), 11 M sequence reads were drawn randomly (without replacement). These randomly drawn reads were mapped to the gene catalogue and the number of reads counted to form a downsized depth or abundance matrix. The 11M downsized depth matrix was used to estimate co-abundance gene group (CAG) abundances, gene and MGS richness. Similar downsizings were done for the reduced sampling depths as indicated in Supplementary Fig. 11.

Taxonomical annotation

Catalogue genes were assigned taxonomical annotation by sequence similarity to a database of 3,048 reference genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ and ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT/, July 2012), using

BLASTN[33], only accepting alignments with 100 bp or longer. Sequence similarity of 95%, 85% and 75% or better was used for species, genus and phylum level taxonomical annotation, respectively. MGS were assigned a species level annotation if more than 50% of the genes comprised in the CAG were assigned a given species level taxonomy (incl. genes with no match). MGS were described to have 'clear and unambiguous similarity to a known species' when 90% or more of the genes were annotated to the same species. Selected CAGs that appear in figures and could not be assigned genus or species level taxonomy by DNA similarity (MGS:11, MGS:17, MGS:124 and MGS:225) were in addition taxonomically annotated by similarity to the UniProt database (BLASTP, best hit, $E < 0.001$) to get an approximate taxonomical annotation.

Phage definition and taxonomy annotation

A CAG was called phage-like if it passed one of two criteria. a) If a CAG contained a minimum of 10 *phage-taxonomy annotated genes* and 80% of these were consistent at species, genus or family level. Here *phage-taxonomy annotated genes* were defined as genes with a top-3 blastp[33] hit ($E < 0.001$, against the combined NCBI nr Sept. 2013 and ACLAME[24] 0.4 database) to a viral organism listed in the International Committee on Taxonomy of Viruses (ICTV) master species list (release 2012)[19]. b) If a CAG encoded five or more distinct *characteristic phage functions* and ≥ 40% of the CAG genes were *most similar to known phage genes*. *Phage-functional classes* were defined: as proteins with a best-hit (hmmscan[34], domE < 0.001, against Pfam-A[35] 27.0) to one of 16 phage specific Pfam functions defined by Minot et al.[10], or as proteins matching the corresponding set of functions identified among phage orthologous groups (blastp, $E < 0.001$, against POC VQ[36]). A *characteristic phage function* was only counted once per GAC. Furthermore, a gene *most similar to known phage genes* was defined as a gene with a best-hit (blastp, $E < 0.001$, against the combined NCBI nr and the ACLAME 0.4 database) to a viral organism. All phage-like CAGs were taxonomically annotated to species, genus or family level using a 50% consistency criteria across ICTV annotated genes (top-3 blastp hits, $E < 0.001$, against the combined NCBI nr and ACLAME[24] database). Interestingly, the functions "tail", "portal" "terminase" and "capsid" were each found in ≥ 70% of all phage-like CAGs and on average in only 5% of other small CAGs.

Gene annotations and enrichment analysis

Functional annotation (incl. CRISPR associated genes) of the gene catalogue was obtained by aligning predicted proteins to the UniProt database using BLASTP (best hit with $e < 0.001$) and proteins from the eggNOG (v3) database[27] using BLASTP (WU-BLAST 2.0, default parameters except $E = 1x10^{-5}$ B = 10000) and were assigned to an orthologous group as described elsewhere[37].

Genes of MGS:11, CAG:4957, MGS:17 and MGS:124 (appearing in Supplementary Fig. 16C) was aligned to proteins listed by Roessner et al.[38] as experimentally verified and strictly anaerobe corrin ring biosynthesis proteins (60 coverage, 40% identity). CRISPR repeat-spacer segments were identified with CRT (ver. 1.2)[28] in selected CAG assemblies. Genes were annotated as virulence or antibiotic resistance genes when BLASTP alignments exceeded 80% identity over 80% of the length of protein in the VFDB[26] (February 2012 version) or ResFinder[39] (version 1.2) database, respectively.

From 271 essential genes from the genome of *Bacillus subtilis* strain 168[40], 252 COGs were deduced (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Bacillus_subtilis_168_uid57675/NC_000964.ptt manually curated, see Supplementary Data 9). Genes aligning to these COGs were termed essential genes.

CAGs significantly enriched for a specific annotation were identified using Fisher's exact test ($P < 0.001$ for Fig. 1B). Significant biases in eggNOG[27] annotation, as a function of the MGS observation frequency across the samples, were identified using Wilcoxon rank sum test ($P < 1x10^{-15}$, Supplementary Data 8).

Co-abundance clustering

The canopy-based clustering of the gene catalogue was performed by iteratively picking a seed gene among the not yet clustered genes and aggregate genes with abundance profiles within a fixed distance from the seed gene abundance profile (Pearson correlation coefficient > 0.9 and Spearman's rank correlation coefficient > 0.6) into the seed canopy. Canopies with median abundance profiles within a distance of 0.97 PCC from one another were merged. Canopies with 2 or less genes (1.7 M genes), or for which the canopy abundance signal from any three samples constituted 90% or more of the total signal across all samples, for which the median profile was detected in less than 4 samples, or for which one sample made up 90% of the total signal (1.1M genes), were discarded for having insufficient supporting evidence (based on Monte Carlo simulation, see Supplementary Fig. 17). Canopies that passed these criteria were called CAGs. CAGs with more than 700 genes are also referred to as MGS or just species. Note, that the number of clusters was not pre-defined for the canopy-based clustering. CAG abundance profiles were calculated as the sample-wise median gene depth signal (downsized). A CAG was considered observed in a sample when its abundance profile exceeded zero in that sample.

MGS augmented assembly

For each of the 741 MGS we performed a *de novo* MGS augmented assembly, using the subset of sequence reads that mapped to the contigs from where the MGS genes originated. For each MGS we perform independent and sample-specific augmented assemblies on the two samples from where most sequence reads mapped to the MGS and the sample from which most of the MGS gene containing *de novo* contigs were derived. For a given sample the reads were aligned using Burrows-Wheeler Aligner[41] (bwa-0.5.9) to the MGS specific scaffolds and the mapped reads, including unmapped mates, were extracted. These reads were then corrected by Quake[42] using $k = 15$. The reads were then *de novo* assembled with Velvet (1.2.01) using *k*-mers from 21 to 45 and the parameters '-cov_cutoff auto' and '-exp_cov auto'. As several samples were used for assembly of each MGS, the best assembly was selected based on ranking of contig N50 and the number of contigs in the assemblies[22]. Contigs with read depth of less than half the average depth of all contigs were removed from the assemblies[22,43]. The contigs and scaffolds were then filtered to 100 and 500 bp minimum lengths, respectively, and gaps in scaffolds were filled using SOAPdenovo GapCloser (1.10).

Assembly statistics

General assembly statistics were calculated using assemblathon_stats.pl[44] and coverage was calculated by aligning reads to the contigs using bwa (0.5.9) and BEDtools[41,45]. To assess the quality of the assemblies, we adopted the six high quality

draft assembly criteria from the Human Microbiome Project (HMP)[46]. Five of these criteria address the contiguity of the assembly and one criterion genome completeness, by counting core genes contained in the assembly. The criteria are *i*) 90% of the genome assembly must be included in contigs > 500 bp, *ii*) 90% of the assembled bases must be at > 5 X read coverage, *iii*) The contig N50 must be > 5 kb, *iv*) scaffold N50 must be > 20 kb, *v*) average contig length must be > 5 kb and *vi*) > 90% of the core genes must be present in the assembly. The core gene ratios were determined using HMP standard operating procedure for both bacteria and archaea. In short blastx was used to identify core genes from the scaffolds and proteins with at least 30% identity and 30% coverage for Bacteria and 50% identity and 70% coverage for Archaea were considered a core gene hit[33]. The ratio of core genes identified was then calculated using get_coregroups_coverage.pl (HMP tools and protocols). In total 360 sample-specific MGS augmented assemblies, from 247 unique MGS passed all six criteria (Supplementary Data 3). In addition, 139 unique assemblies passed five criteria.

We determined the number of novel species by aligning all proteins to Uniprot[25] using blastp and converted taxids from strain to species level using NCBI-taxonomy. An assembly was considered previously un-sequenced if less than 10% of the genes could be aligned with a minimum of 95% identity over 33aa to genes from a species. 181 of the 238 HQ assembled draft genomes plus 83 assemblies passing 5 criteria were identified as novel species.

Screening for chimeric assemblies

Because the HMP criteria were created for single genome assembly, we applied three additional metrics to account for putative chimeric assemblies arising from metagenomic data, i) uniformity of the contig read depth distribution, ii) identification of multiple copies of conserved 40 Clusters of Orthologous Genes (COGs)[27] and iii) inter-assembly tetra nucleotide frequency (TNF) consistency.

Because assemblies consisting of genomic regions from different organisms are likely to have multi-modal coverage distributions we performed peak detection on the contig read coverage distributions for all assemblies passing 4-6 HMP criteria and assemblies with more than 1 peak were manually inspected. From the presence of multiple copies of COGs we were able to identify three assemblies as chimeric. Of the 247 unique high-quality draft assemblies 9 (3.6%) were identified as potentially chimeric and for the additional 139 assemblies that passed 5 criteria, we identified 3 potential chimeric assemblies (2.3%) and one without any core genes (MGS:3246). The remaining assemblies have been deposited at the European Nucleotide Archive (ENA).

Furthermore, tetranucleotide frequencies z-scores were calculated for all assemblies and HMP reference assemblies as described by Teeling *et al.*[47] For each assembly the frequencies were calculated in windows of 5 kb to avoid biases introduced by different scaffold lengths. If a scaffold was shorter than the window size it was still included in the calculations. Within each assembly a median tetranucleotide frequency z-profile were created and the tetranucleotide frequency z-scores of each 5 kb window were correlated to this median profile using PCC. The resulting high-quality draft genomes showed comparable TNF correlations to the single organism HMP reference genomes indicating low rate of chimeric assemblies (Supplementary Fig. 9).

Comparison of MGS augmented assemblies and reference genomes

To estimate the completion level of the MGS augmented assemblies, 299 draft reference genomes from the human intestinal tract HMP DACC database and the NCBI collection of complete reference genomes (both version updated from 2012/04) were used as reference set for a blast comparison procedure. 44 of the assemblies that passed 5 or more of the 6 HMP criteria (including the bacteria/archaea core ratio criteria) were similar to a reference genome. The contigs and scaffolds of these assemblies were projected on their closest reference genomes using the GAGE pipeline for assembly quality evaluation[48]. First *nucmer* (default parameters) was used to align the contigs/scaffolds to the reference genome. Then *delta-filter* was used to remove low identity match (parameters: -I 95, -o 80). Finally *dnadiff* was used to compare the assemblies and the closest reference genome and estimate the mean identity and coverage of each contigs and scaffolds (Supplementary Data 4). Additionally, the MGS:337 assembly, which did not meet the six criteria, was 99.9% identical to *Bifidobacterium animalis subsp. lactis CNCM I-2494*[18] and covered 95% of this reference genome (Supplementary Fig. 10).

To search for potential contaminants, unaligned scaffold fragments were blasted to the complete reference genome set, and the best hit (with identity and coverage threshold of ≥ 95% and ≥ 80%, respectively) were extracted. Scaffold that matched to a different genus were considered potential contaminants. Of the 44 MGS augmented assemblies, only 16 contained any scaffolds with similarity to an alternative genus. In general these scaffolds were small with an average size of only 2721 bp. If we consider unaligned scaffold with similarity to an alternative genus as potential contaminant, the mean contamination rate was estimated to 1.00 scaffold per HQ assembly.

MGS augmented assembly gap closure using Sanger sequence data

To further experimentally validate the coherence of the sample specific MGS augmented assemblies we used Sanger sequence data from eight samples[16]. Faecal microbial DNA from those individuals was used to construct plasmid-based (pCNS) clone libraries of 3 kb long inserts, containing 250,000 clones each. Clone insert ends were sequenced using the Sanger technology (ABI3730XL). Sequences were subsequently subjected to vector cleaning and quality trimming, generating the on average 230,468 (+/- 5,145) reads per sample. The same DNA was used for pyrosequencing (454GSFlx-Titanium) resulting in on average 2,362,978 reads per sample (+/- 3,245,603). For each reference subject, Sanger and 454 reads as well as Velvet contigs generated from Illumina sequencing of the same DNA were combined for assembly using the 454-Newbler software (v2.3). CAGs detected in a given reference subject was compared with Sanger reads from that individual using blastn. High-scoring segment pair (HSPs) covering at least 90 % of the length of the smallest read or velvet contig with at least 90 % identity were selected, and corresponding reads extracted. Scaffolding of the CAG contigs with paired Sanger reads was then achieved using the bambus software[49]. Only assemblies with > 1 X coverage were kept, and used to assess rate of gap-closure (Supplementary Data 3). On average 64% of the assembly gaps were closed and in particular the MGS:710 assembly was closed to only 3 scaffolds from initial 32 scaffolds.

Phylogeny of the MGS assemblies

We used all non-chimeric assemblies passing 5 and 6 HMP criteria (139 and 247 assemblies, respectively) and 296 HMP gut microbiome reference genomes

(HMPDACC) and 1506 reference genomes to construct a phylogeny based on 40 phylogenetic marker proteins (COGs)[50]. For each assembly, proteins were predicted using Prodigal[51] and aligned using blastp to the individual COG proteins and the best hits were selected requiring at least 50% id over 50% of the COG sequence. For each COG the MGS assembly and reference proteins were aligned using *muscle* [52] and here joined to a single alignment file for each COG using *muscle -profile*. The 40 individual protein alignments were concatenated to a single alignment for each reference genome/MGS assembly and alignments with containing less than 35 COGs were removed from further analysis resulting in 337 MGS assemblies for the final tree. The phylogenetic tree was constructed with FastTree using the JTT substitution matrix with the parameters "-gamma -pseudo -spr 4 -mlacc 3 -slownni"[53] and visualized using ITOL[54].

Co-assembly of E. coli and dependency-associated CAGs

A pool of the main *E. coli* (MGS:4) and its nine dependency-associated CAGs (CAG:427, CAG:1345, CAG:2136, CAG:2318, CAG:2530, CAG:2610, CAG:3070, CAG:3196 and CAG:5108) were used for recruiting 1708 contigs for a pooled assembly, across 247 selected samples. Subsequently, *de novo* assembly (as described above) from 13 of these samples passed five or more HMP criteria (Supplementary Data 7). A consensus assembly was generated from the contigs of these assemblies using minimus2 where each assembly were joined to the consensus in separate steps[55]. The consensus assembly contained 4.3 Mb sequences in 45 contigs and with a contig N50 of 129 kb. Subsequently all the individual assemblies were aligned with blastn to the consensus assembly and contigs without a significant hit were pooled and clustered using cd-hit-est with the parameters "-c 0.8 -n 7"[56]. To further reduce redundancy of the extra contigs they were cut into 500 bp 'reads' with 250 bp overlap and re-assembled using Newbler 2.6. The resulting 157 contigs were then added to the consensus assembly obtained from minimus2 to a final assembly of 4.91 Mb in 202 contigs.

Dependency-associations

A CAG was considered dependency-associated on another CAG when the sample-wise overlapping detections of the CAG pair was statistically significantly overrepresented (Fisher's exact test, upper tail, Bonferroni corrected $P < 1x10^{-10}$) and the dependency-associated CAG was not detected independently.

Smaller CAGs (<700 genes) were considered 'co-existence associated' when their detections were significantly enriched (Fisher's exact test, Bonferroni corrected $P < 0.05$) in samples where *an MGS pair* was co-observed, and never occurred independently of one of the two MGS (the host). Here an MGS pair consisted of a host MGS and a companion MGS. An MGS was considered a potential companion if it co-existed with a potential host species in samples from ten or more individuals and was found independently of the host species in samples from ten or more human individuals. For the co-existence associated relationships where the small CAG was not observed independently of any of the two MGS the host species were determined, as the MGS with the strongest abundance correlation to the small CAG, across samples where both were detected, and by the sample specific co-assembly. No inconsistency between these measures was found.

Dependency-associated small CAGs were considered significantly absent in samples where a specific companion species were found when: it among host

containing samples, were significantly enriched in samples where the companion species was absent (Fisher's exact test, Bonferroni corrected $P < 0.05$). Furthermore, the small CAG could never be observed independent of the hosting species. Again, an MGS was considered as a potential companion species if it co-existed with a host species in samples from 10 or more individuals.

For all types of dependency-associations a CAG was considered detected in a sample if the CAG abundance profile exceeded zero. Furthermore, only CAGs detected in ≥ 10 and ≤ 308 samples were considered. To ensure independence between the observations only one sample per individual was used (n = 318). Dependency-associated, 'co-existence associated' and co-existence absent' CAGs showed correlation to the species richness comparable to that of all CAGs.

Estimation of CAG persistence probabilities

Data from 73 human individuals, which were sampled twice, were used to estimate the annual persistence probabilities of MGS with or without dependency-associated CAGs (Supplementary Data 6, Supplementary Fig. 18). All of the 2x73 stool-samples included in this analysis resulted in at least 11M sequence reads, and samples yielding more than this were downsized to 11M reads. Furthermore, all included sample pairs exhibited strong stability between the samplings, in that they were more similar to each other than to 99% of the other samples in the cohort (using the Spearman correlation coefficient of the MGS abundances as similarity measure).

The main idea in this analysis was the following: for a fraction of the 73 sample pairs, a given MGS is present in the sample obtained at time point 1. For a subset of these sample pairs, the same MGS was also present at time point 2. Based on these, data logistic regression can be used to estimate an annual persistence probability for the MGS. The predictor variable (time between two consecutive samples) is continuous, while the outcome variable (presence or absence of an MGS) is binary. Logistic regression is used to estimate how the probability of an MGS still being present depends on the amount of time passed.

This computation is based on the assumption that an MGS has a typical probability *per time unit* of persisting in the gut of an individual. Thus the likelihood of observing an MGS at time point 2 is expected to be smaller the more time has passed between the two samplings. Specifically, this decline is assumed to be exponential, thus if the probability that a given MGS will persist for a year is P(1) = 0.7, then the probability that it will persist for two years is P(2) = 0.7^2 = 0.49, etc. This assumption seems to fit well with survival curves constructed from these data (see Fig. 3A). Of course, the persistence of a given MGS in any individual is likely to depend on the specific conditions in that individual. We simply assume that there is a typical overall annual persistence probability associated with the MGS (on average, a given MGS has a typical tendency to persist in the gut of any individual), and real data will be scattered around this average according to unidentified covariates and stochastic effects.

Annual persistence probabilities were estimated in a probabilistic (Bayesian) model-based framework that explicitly accounts for time-dependence. In this approach we assume that the annual persistence probability for an MGS is determined by the inherent resilience of the MGS itself, in combination with possible additional effects (positive or negative) caused by a set of dependency-associated CAGs. Specifically, we assume that the annual persistence probability, P, for a given MGS, depends on the effects of a set of dependency-associated CAGs in the form of a

logistic regression model: $\ln(P/[1-P]) = \text{logit}(P) = b_0 + \text{Sum}[b_i X_i]$ or, equivalently: $P = \text{expit}(b_0 + \text{Sum}[b_i X_i])$. Here, the regression coefficient b_0 corresponds to the inherent persistence tendency of the MGS itself, b_i corresponds to the effect of dependency-associated CAG number i, and X_i is a binary variable indicating whether dependency-associated CAG number i is present or absent for a given sample. "Expit" is the sigmoidal, logistic function (the inverse of the logit function). The index, i, runs over all the dependency-associated CAGs for a given MGS.

The probability that a CAG will survive for t days, P(t), can be found from its annual persistence probability, P, in the following way: $P(t) = P^{[t/365]}$. The likelihood for a data point where the MGS survives (i.e., where it is still present at the second sample, after t days have elapsed) is therefore given by the following expression: $L = P^{[t/365]} = [\text{expit}(b_0 + \text{Sum}[b_i X_i])]^{[t/365]}$. For data points where a CAG does not survive, the likelihood is simply: $L = 1 - P^{[t/365]} = 1 - [\text{expit}(b_0 + \text{Sum}[b_i X_i])]^{[t/365]}$. As recommended in Gelman *et al.*[57] the priors for all b_0 regression coefficients (which correspond to the inherent persistence of all MGS) are t-distributions with $\mu=0$, df=1, and rate=0.1 (corresponding to scale=10). The priors for all b_i regression coefficients (corresponding to the effects on persistence of the dependency-associated CAGs) are t-distributions with $\mu=0$, df=1, and rate=0.4 (corresponding to scale=2.5). These are conservative priors that help keep the correlation coefficients close to zero. Given these expressions for priors and likelihoods, it is possible to perform a Bayesian analysis of the model, resulting in estimates of the above-mentioned regression coefficients. However, since the regression coefficients themselves can be difficult to interpret, we instead report the following derived measures: i) the annual persistence probability for each MGS. This can be computed as: $P = \text{expit}(b_0)$. ii) The annual persistence probability for a specific MGS when together with a given dependency-associated CAG. This can be computed as: $P = \text{expit}(b_0 + b_j)$, where j refers to the specific dependency-associated CAG. iii) The *effect* of the dependency-associated CAG. We have chosen to simply expres this as the *absolute difference* between the above two measures. (For instance: If the annual persistence probability of an MGS, together with a specific dependency-associated CAG is 0.75, and the annual persistence probability of the MGS alone is 0.5, then the effect of the dependency-associated CAG is reported as $0.75 - 0.5 = 0.25$).

For the analysis of coexistence between a pair of MGS and an associated CAG, there was insufficient data to obtain estimates for each individual CAG. We therefore pooled all data points for CAGs having a positive effect on the persistence of their MGS host, and estimated an overall effect for these.

Note that, in the Bayesian framework, estimates are expressed as probability distributions over the possible values for parameters of interest[21]. We therefore obtain both an estimate of a parameter, and quantification of how certain we are of the estimate. To declare an effect to be substantially different from zero, we require that its 95% highest posterior density interval (the "95% HDI") should be located entirely outside of a "region of practical equivalence" to 0 (a "ROPE"). In this analysis the ROPE was defined to be [-0.02, 0.02]. The 95% HDI is the narrowest interval that includes 95% of the probability. By design, all parameter values inside a 95% HDI will be more likely than all values outside. In this work we have identified 26 dependency-associated CAGs where we are more than 95% certain that they have a non-zero effect on the persistence probability of an MGS (Supplementary Data 6).

The model was implemented and analysed in a Bayesian framework by Markov chain Monte Carlo (MCMC) using the JAGS package[58]. Convergence of MCMC was checked by running two independent chains and verifying that they arrived at similar posterior distributions. In particular it was checked that the potential scale reduction factor ("R-hat") for each estimated parameter was < 1.1[59].

REFERENCES (FOR ONLINE METHODS)

1. Qin, J. *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
2. Fodor, A. A. *et al*. The "most wanted" taxa from the human microbiome for whole genome sequencing. *PLoS One* **7,** e41294 (2012).
3. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–20 (2005).
4. Turnbaugh, P. J. *et al*. A core gut microbiome in obese and lean twins. *Nature* **457,** 480–4 (2009).
5. Rajilić-Stojanović, M., Heilig, H. G. H. J., Tims, S., Zoetendal, E. G. & de Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* (2012). doi:10.1111/1462-2920.12023
6. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–36 (2011).
7. Schloissnig, S. *et al*. Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).
8. Fitzsimons, M. S. *et al*. Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. *Genome Res.* gr.142208.112– (2013). doi:10.1101/gr.142208.112
9. Reyes, A. *et al*. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466,** 334–8 (2010).
10. Minot, S. *et al*. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21,** 1616–25 (2011).
11. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22,** 1985–94 (2012).
12. Zhang, Q., Rho, M., Tang, H., Doak, T. G. & Ye, Y. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biol.* **14,** R40 (2013).
13. Wang, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* **28,** i356–i362 (2012).
14. Albertsen, M. *et al*. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31,** 533–538 (2013).
15. Karlsson, F. H. *et al*. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498,** 99–103 (2013).
16. Qin, J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
17. Le Chatelier, E. *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–6 (2013).
18. Chervaux, C. *et al*. Genome sequence of the probiotic strain Bifidobacterium animalis subsp. lactis CNCM I-2494. *J. Bacteriol.* **193,** 5560–1 (2011).
19. *Virus Taxonomy: Classification and Nomenclature of Viruses : Ninth Report of the International Committee on Taxonomy of Viruses*. 1327 (Elsevier, 2012). at <http://books.google.com/books?hl=en&lr=&id=KXRCYay3pH4C&pgis=1>
20. Terns, M. P. & Terns, R. M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* **14,** 321–7 (2011).
21. Kruschke, J. K. Bayesian data analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* **1,** 658–676 (2010).
22. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–9 (2008).
23. Li, R. *et al*. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).
24. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.* **38,** D57–61 (2010).
25. Bairoch, A. *et al*. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33,** D154–9 (2005).
26. Chen, L., Xiong, Z., Sun, L., Yang, J. & Jin, Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* **40,** D641–5 (2012).

27. Powell, S. *et al*. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res*. **40,** D284–9 (2012).
28. Bland, C. *et al*. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).
29. Kultima, JR, Sunagawa, S, et al. MOCAT: a metagenomics assembly and gene prediction toolkit.
30. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res*. **38,** e132 (2010).
31. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res*. **12,** 656–64 (2002).
32. Li, R. *et al*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–7 (2009).
33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol*. **215,** 403–10 (1990).
34. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. **39,** W29–37 (2011).
35. Punta, M. *et al*. The Pfam protein families database. *Nucleic Acids Res*. **40,** D290–301 (2012).
36. Kristensen, D. M., Cai, X. & Mushegian, A. Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J. Bacteriol*. **193,** 1806–14 (2011).
37. Tringe, S. G. *et al*. Comparative metagenomics of microbial communities. *Science* **308,** 554–7 (2005).
38. Roessner, C. A. & Scott, A. I. Fine-tuning our knowledge of the anaerobic route to cobalamin (vitamin B12). *J. Bacteriol*. **188,** 7331–4 (2006).
39. Zankari, E. *et al*. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother*. **67,** 2640–4 (2012).
40. Kobayashi, K. *et al*. Essential Bacillus subtilis genes. *Proc. Natl. Acad. Sci. U. S. A*. **100,** 4678–83 (2003).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).
42. Kelley, D. R., Schatz, M. C. & Salzberg, S. L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol*. **11,** R116 (2010).
43. Mavromatis, K. *et al*. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* **4,** 495–500 (2007).
44. Earl, D. *et al*. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res*. **21,** 2224–41 (2011).
45. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).
46. Chain, P. S. G. *et al*. Genomics. Genome project standards in a new era of sequencing. *Science* **326,** 236–7 (2009).
47. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol*. **6,** 938–47 (2004).
48. Salzberg, S. L. *et al*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. **22,** 557–67 (2012).
49. Koren, S., Treangen, T. J. & Pop, M. Bambus 2: scaffolding metagenomes. *Bioinformatics* **27,** 2964–71 (2011).
50. Ciccarelli, F. D. *et al*. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311,** 1283–7 (2006).
51. Hyatt, D. *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32,** 1792–7 (2004).
53. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).
54. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*. **39,** W475–8 (2011).
55. Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next generation sequence assembly with AMOS. *Curr. Protoc. Bioinformatics* **Chapter 11,** Unit 11.8 (2011).
56. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–9 (2006).

57. Gelman A., Jakulin A., Pittau M. G., Su, Y. A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models. *Ann. Appl. Stat.* **2(4),** 1360–1383 (2009).
58. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. in *Proc. 3rd Int. Work. Distrib. Stat. Comput. March 2022* (Hornik, K., Leisch, F. & Zeileis, A.) 0 (Citeseer, 2003). at <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/>
59. Gelman, A. & Rubin, D. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7,** 457–511 (1992).

Appendix B: Supplementary Information - Variable gene modules predict persistence of microbes in the human gut

SUPPLEMENTARY INFORMATION

MGS completion compared to reference genomes

On average the MGS with closely related reference genomes include 69% of the genes annotated to that species (for the MGS augmented assemblies 78% of the sequence is covered with an average identity of 98.4%). This discrepancy may reflect variation in the genetic makeup of the species across samples, limitations of the co-abundance segregation or dissimilarity between the observed species and the reference genome. As a benchmark, 19 Spanish individuals consumed a defined fermented milk product containing *Bifidobacterium animalis subsp. lactis CNCM I-2494*[18] (Supplementary Data 1). In this case 95% of the *B. animalis* reference genome genes were captured in MGS:337. Still the co-abundance segregation is likely to separate the genomic core from clone specific and mobile genetic elements; a notion that is supported by the very significant enrichment of essential genes (encoding *Bacillus subtilis* essential COGs[40], Fishers exact test: $P << 1 \times 10^{-100}$) in the MGS.

MGS vs. small CAGs

The distinction, between the 741 MGS with more than 700 genes and the smaller CAGs, should roughly identify the CAGs that represent cellular microbial species. This division is not clear-cut and some core genomes may fall below and some clone specific variants and mobile elements may exceed the 700-gene threshold. The threshold was chosen based on a combination of expectations and observations. While some bacterial genomes have been reported to contain very low number of genes[60] most known bacterial genomes encode more than some 1,000 genes, whereas most phages and plasmids have less that 500 genes (Supplementary Fig. 4). Prior knowledge therefore suggests a threshold somewhere between 500 and 1000 genes[61]. In addition, three independent observations in our data suggest a threshold around 700 genes. First, the observed bimodal size distribution of the CAGs shown in Fig. 1A narrows around 700 genes and therefore suggest a natural distinction around 700 genes. Second, a significant enrichment for genes essential to bacterial life, detected by homology to the *Bacillus subtilis* essential gene set[40] was found primarily in CAGs with more than 700 genes. Finally, if small CAGs represent genetic heterogeneity of biological organisms or bacteriophages they should depend on an organism to proliferate. The number of small CAGs with statistically determined dependency-associations to MGS is highest at the 700-gene threshold (Supplementary Fig. 19) and drops for higher thresholds. With this threshold the odds ration for small CAGs to MGS dependency-associations is 12.7.

The majority of 'reference species gene sets' demonstrates incoherent gene abundance profiles

32% of the 3.9 M gene catalogue could be assigned taxonomy at phylum level by similarity to known microbial organisms (best hit with over 75% identity over 100 bp). The majority of these resemble Firmicutes and Bacteroidetes genes (57% and 28%, respectively). However, only 10% of the catalogue genes could be assigned taxonomy at species level (best hit with over 95% identity over 100 bp). Using this criterion, 161 'reference species gene sets', with more than 700 genes assigned to the same species, were defined. Here, these gene sets serve as representatives of a reference genome based structuring of the metagenomics data. Correlation analysis of the abundance profiles of the genes within the 161 'reference species gene sets', show that many of these gene sets do not behave as a coherent entity (Supplementary Fig.

15). Hence, 88 of the 'reference species gene sets' have significant sub-populations of genes that do not correlate with the bulk of the genes, i.e. 25% or more of the genes have a Pearson correlation coefficient (PCC) < 0.5. The genes within 56 of these sets could be identified as members of multiple distinct MGS, which in turn include additional genes without similarity to these reference genomes. For 12 'reference species gene sets' almost no internal gene correlations could be found (less than 25% of the genes have PCC of 0.5 or more). The highly inconsistent 'reference species gene sets' are found across the major phyla: Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria, but the Bacteroidetes gene sets stand out as particularly incoherent with an average within gene set PPC of 0.5. At genus level Bacteroides, Faecalibacterium and Prevotella are the most incoherent groups, all with an internal PCC average below 0.5. This level of inconsistency is problematic both because the annotation does not reflect the biological organization of the system and because association between inconsistent 'reference species gene sets' and clinical data, may be misleading or fail to identify an underlying organism. Further work will be required to clarify the reasons for the inconsistency, but we suggest that CAGs may be more suitable than the homology-based gene sets both for understanding the gut microbial communities and their association to health and disease.

MGS profiles are coherent in an independent sample series

To demonstrate that the MGS behave as general biological entities the coherence of abundance profiles for the MGS genes was investigated in 115 independent human stool samples that were not used for the clustering[17]. In this independent sample set the median gene-wise Pearson correlation coefficient for intra MGS gene profiles was as high as 0.98. Hence, the MGS indeed appear to be general descriptors of coherent genetic entities across similar microbial systems.

The MGS profiles are distinct and robust

Individual gene abundance profiles of a typical MGS are highly coherent and distinct from the profiles of the genes not included in that particular MGS. Consequently, relaxing the gene inclusion criterion from PCC > 0.9 to PCC > 0.8 relative to the median profile of the MGS only extends the MGS gene set on average by 5%. Similarly, raising the inclusion criterion to PCC > 0.95 reduces the number of genes included by 17 % on average. Hence, the co-abundance based clustering is robust to changes in parameters that determine the cluster boundary and importantly the MGS are separated from other genes. This feature is so strong that more than 30 MGS stand out as distinct and dense gene clouds in a two-dimensional Pearson scatter-plot (Supplementary Fig. 3), even when the plan/dimensions of the plot is not targeting the separation of these specific MGS. In addition, even the dependency-associated CAGs are clearly distinct from their hosts, thus on average the PCC between the host and the dependency-associated CAGs is 0.46 (+/- 0.2).

21% of the abundance-uncorrelated genes can be linked to the MGS

1.2M catalogue genes, with abundance profiles exceeding the filtering criteria (more than 3 samples must constitute 90% of the total abundance signal) did not segregate with any CAG. The detection rate of these abundance uncorrelated genes was however comparable to that of the correlated genes, as their sequence coverage and re-detection rate in paired samples from the same individual, were similar to that of the correlated genes. Although, the abundance-uncorrelated genes on average were detected in significantly fewer samples than the correlated genes (mean: 50 and 93

samples, respectively) the bulk of these were detected in a sufficient number of samples to allow these to be segregated, if they were correlated to other genes.

These abundance-uncorrelated genes are in contrast to the clustered genes significantly underrepresented in essential genes. Interestingly, we found that, genes involved in antibiotic resistance, with the exception of vancomycin resistance, had distinct single gene abundance profiles. This is in line with the fact that most antibiotic genes, except vancomycin resistance genes, are known to single-handedly provide antibiotic resistance and suggests that some genes may be highly dynamic and perhaps are best understood non-contextually, at the single gene level.

21 % of these abundance-uncorrelated genes, however, can be linked to an MGS by shared sequence contigs in at least one sample, indicating that some of these genes may be clone or strain specific genes of the species. In support of this, these contig-extended genes are likewise significantly underrepresented in essential genes (encoding *Bacillus subtilis* COGs[40], Fisher's exact test $P = 0.002$).

For instance the very common *Bacteroides vulgatus* (MGS:6) comprises 2,271 genes but can be contig extended to include additional 326 genes, across all samples. The average sample however, only comprises 161 of these genes, and the abundance profiles of these genes show little correlation to the *B. vulgatus* (MGS:6) profile (mean PCC = 0.3). This abundance profile inconsistency of the contig extended genes may to some extend resemble the inconsistencies observed for 'reference species gene sets' and as such illustrate the difference between example based gene sets and CAGs.

Common species comprise genes for protection against ROS and Vitamin B_{12} metabolism

Comparisons of MGS between any pair of individuals show an average overlap of 50% (±12%). A substantial part of the shared MGS belongs to a core set of 31 MGS that were detected in at least 90% of the samples (Supplementary Data 11) and together they account for 25% of the abundance signal. Of these, 18 have clear similarity to taxonomically known species, and, hence, there is an overrepresentation of taxonomically known species among the core MGS (odds ratio 3.0). The most common MGS across the sample series are the *Blautia wexlerae* (MGS:9) and *Bacteroides vulgatus* (MGS:6) which were found in 395 and 392 of the 396 samples, respectively. *B. vulgatus* (MGS:6) is in addition the most abundant species across the samples, matched by on average 6% of the mapped reads and is the dominating species in 58 samples.

A number of orthologous groups (eggNOG)[27] are found significantly more frequently among common species than in less common species (Wilcoxon rank sum test, $P < 1\text{x}10^{-15}$, Supplementary Data 8) and suggests a set of functions important for bacterial existence in the human gut. These include enzymes for protection against reactive oxygen and vitamin B_{12} biosynthesis.

Only very few contigs may be the result of chimeric assemblies

For the assembly of the data we used the MOCAT pipeline[29] which performs a revision of the initial assembly that specifically tries to identify and break chimeric contigs (see Methods). However, to assess the rate of potential chimeric contigs we re-mapped all reads to the assemblies using bwa (bwa-0.7.5)[41] and calculated the number of bases per contig that was not bridged by properly paired read pairs. The properly paired read-pairs are read-pairs that map in the expected orientations and with the expected insert length. Absence of these in regions of an assembly indicates a

point of miss-assembly and a potential chimera in metagenomic assemblies. Across all contigs we found that only a small fraction (0.0058) had one or more bases that were not covered by properly paired reads, pointing to a very low contig chimera rate. Among the cross dependency-associated CAG contigs we only found 31 of 7,966 (0.0039) without proper paired read coverage.

The MGS richness is in concordance with gene richness and indicative of Crohn's disease

The number of different species, commonly known as the species richness, is an important measure of an ecological system, partially because it is believed to reflect the general health and stability of a system[17]. Obviously, this measure depends on proper detection of species in the ecological system, and we propose the MGS count as an estimate. Across our cohort of stool samples this number ranges from 33 to 307 with a mean of 155. The number of MGS with species level annotation (shown as coloured bars in Supplementary Fig. 20) is more constant across the cohort than the number of unknown species (shown with black bars). Importantly, sample-wise gene-richness correlates significantly better to the MGS richness estimate (PCC = 0.97), than to the taxonomically known species richness (PCC = 0.55) or to richness estimates based on reference genome detection (see Methods, PCC = 0.52). In addition, MGS richness significantly associates the occurrence of Crohn's disease (t-test, $P = 4x10^{-15}$), much stronger than does species richness derived by sequence similarity to known reference genomes (Crohn's disease: $P = 0.09$). While association between species richness and Crohn's disease has been reported before[62], the MGS richness measure clearly enhances the correlation.

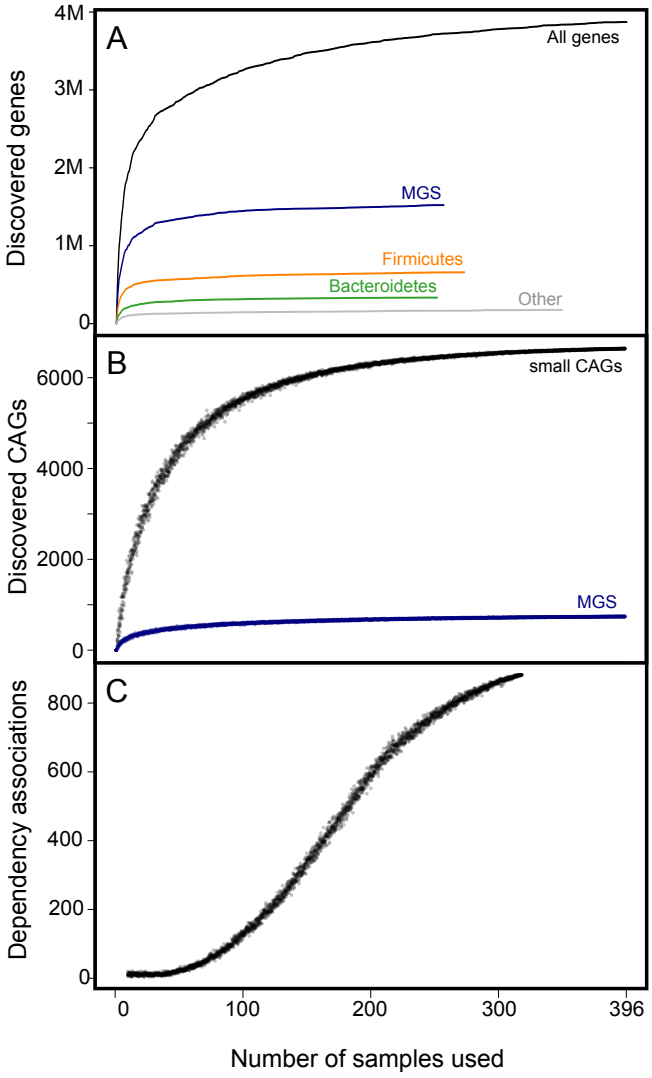
Co-existence associated CAGs

For the microbial species the presence of other companion species in the community may be a major factor to which they may adapt. Such adaptations may be indicated by significantly increased occurrence of specific dependency-associated CAGs in samples where a companion species is also found. A subset of 66 dependency-associated CAGs does exactly that (Supplementary Fig. 16A shows an example) and these are therefore candidates for adaptations to co-existence. In 18 of these relationships the dependency-associated CAGs coincide with significantly enhanced persistence probabilities of the hosting MGS when found jointly with a companion species (Supplementary Fig. 16B, Supplementary Data 12). The companion species on the other hand appear to be only marginally affected by the presence of the dependency-associated CAG, with only a slight but insignificant increased persistence.

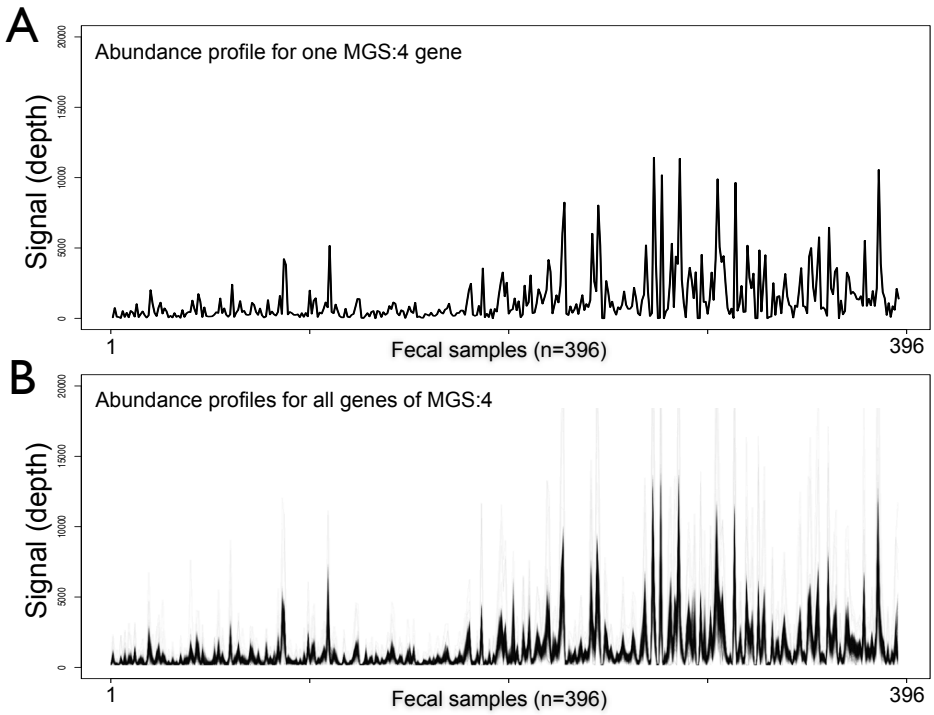
This set of co-existence associated CAGs is very significantly enriched in genes encoding parts of the TonB complex that is important for extracellular sensing and that in *Pseudomonas aeruginosa* has been associated with biofilm formation and quorum sensing[63]. As an example, the *Odoribacter splanchnicus* (MGS:225) dependency-associated CAG:3500 contains genes that encode a 'TonB-dependent receptor plug protein', a 'two-component sensor histidine kinase' and a 'transcriptional response regulator rprY', proteins that have been reported in signal transduction, chemotaxis and quorum sensing[63,64]. Furthermore, the set of co-existence associated CAGs are enriched in the broad-spectra acriflavin resistance proteins and conjugation-coupling factor proteins.

In contrast, to the co-existence associated CAGs, another set of dependency-associated CAGs were significantly absent in samples where specific companion species were found (Supplementary Fig. 16C and Supplementary Data 12). The MGS:11 (Oscillibacter) has several such associated CAGs. In particular, CAG:4957 encodes 16 proteins that are orthologous to proteins in two companion species (MGS:17, Ruminococcus like and MGS:124, Pseudoflavonifractor like), but not to any proteins in the hosting MGS:11. Seven of these proteins are in the anaerobic corrin ring biosynthesis part of the Vitamin B_{12} pathway (Supplementary Fig. 16C lower panel). Hence a possible role for CAG:4957 is to compensate for the biosynthetic potential of the companion species when they are absent.

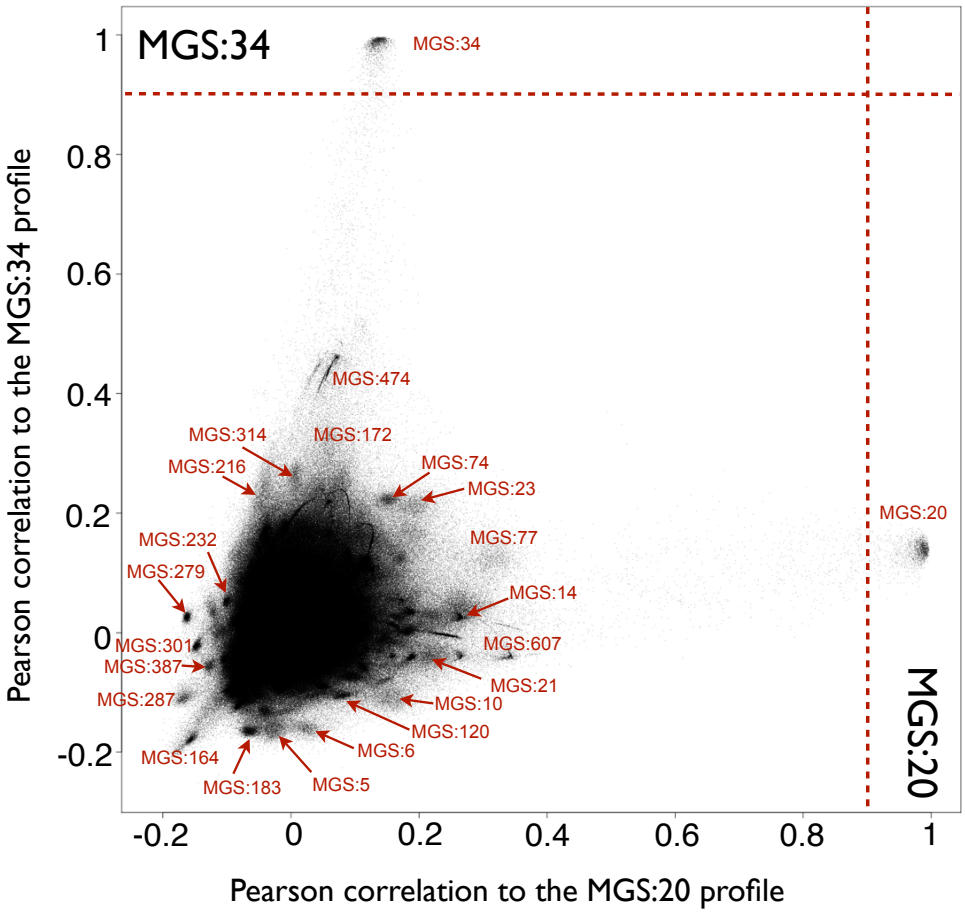
SUPPLEMENTARY FIGURES



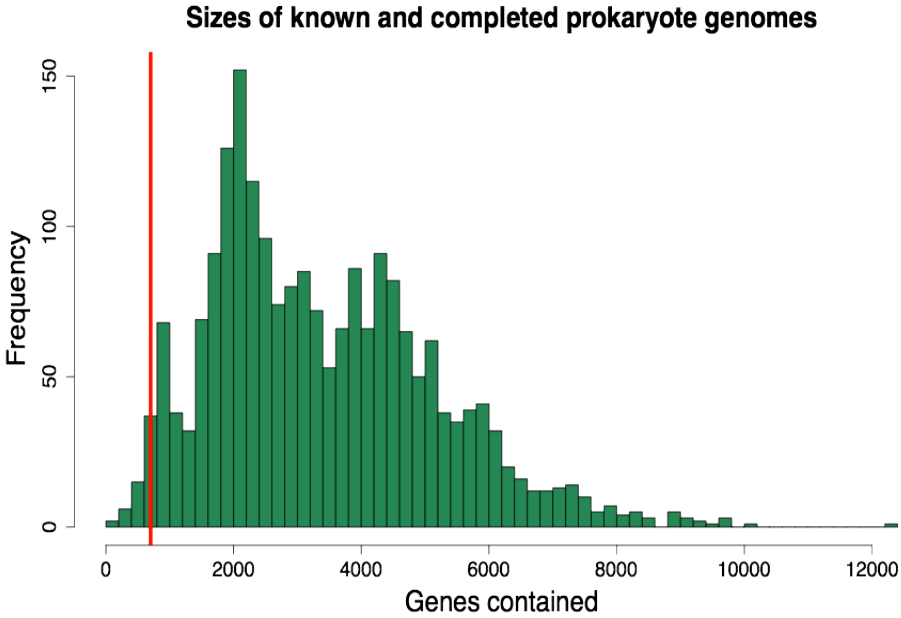
Supplementary Figure 1. Accumulation curves. A) The five curves show the count of genes for the different types. The last sample, on average, discovers only 584 new genes or 0.02%. Reordering the samples resulted in almost identical cumulative curves (not shown). B) The number of small CAGs (semitransparent black) or MGS (semitransparent blue) found three or more times in random subset of samples of the indicated sample sizes (x-axis). C) The number of significantly dependency associations identified in a random set of samples (from independent individuals) of the indicated sample sizes (1 – 318). In B) and C) 10 independent random drawings were made for each sample size.



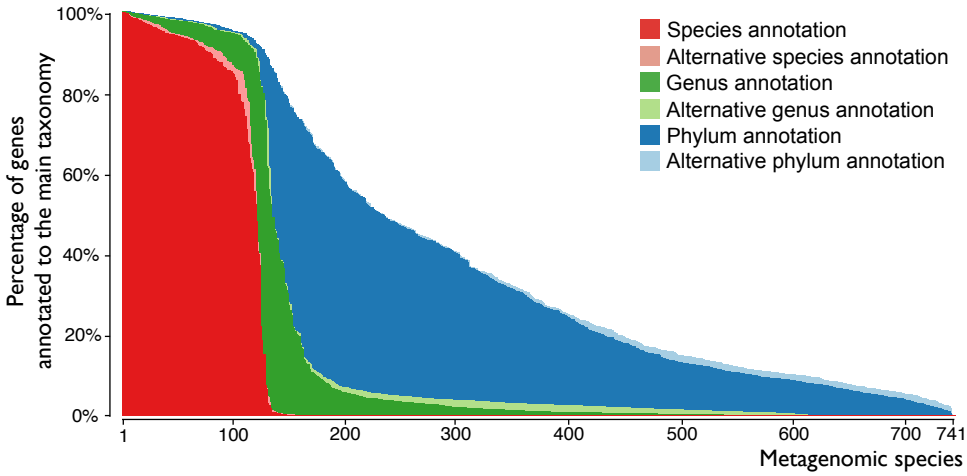
Supplementary Figure 2. Gene profiles of MGS:4. A) The abundance profile of a single gene from MGS:4 (*E. coli*) across 396 samples. B) The abundance profile for all of the 3,523 genes of MGS:4 (shown as 3,523 semi-transparent lines). The median Pearson correlation coefficient between the abundance profiles of the MGS:4 genes was 0.98.



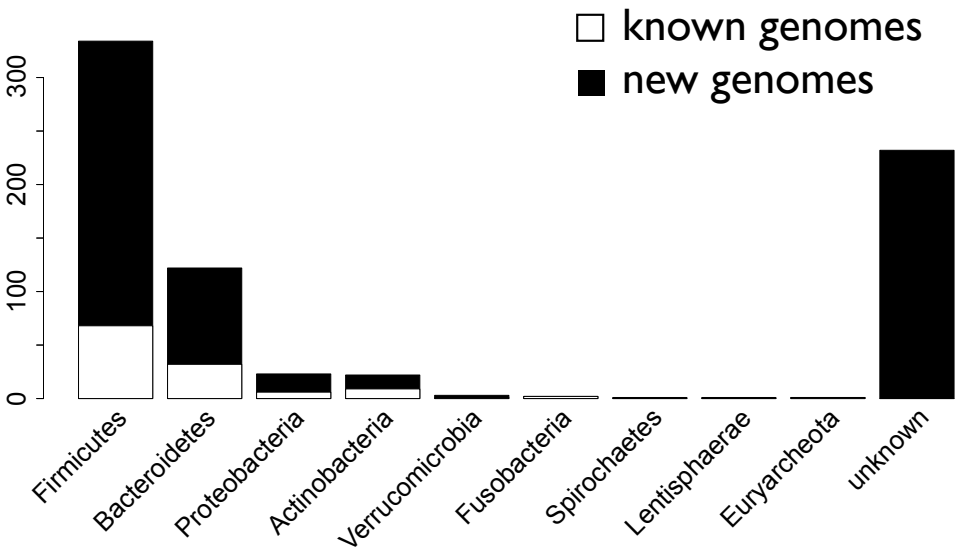
Supplementary Figure 3. Scatter plot showing the Pearson correlation coefficients between the abundance profile of the 3.9M catalogue genes (points) and the abundance profile of MGS:20 (x-axis) and MGS:34 (y-axis). Genes belonging to MGS:20 (n = 2119) and MGS:34 (n = 1799) are defined as genes with correlation coefficients exceeding 0.9 (Pearson) on the x and y axis, respectively (see Methods). In addition to the axis defining MGS, several other MGS are visible as distinct gene clouds at the periphery of the main gene cloud. The IDs of the most visible MGS are indicated.



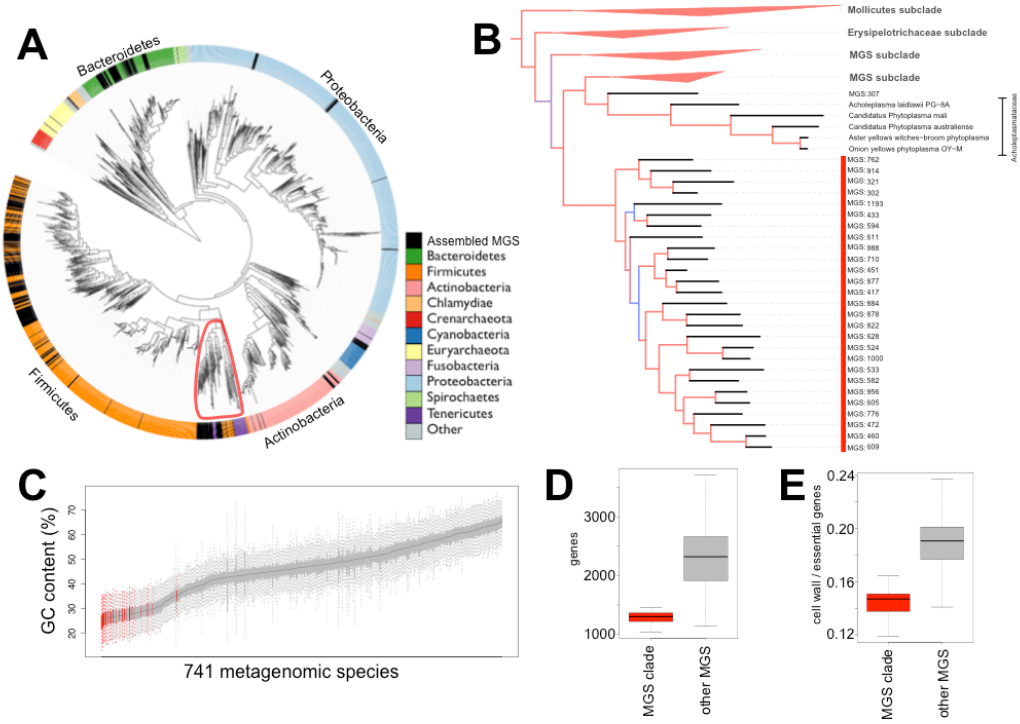
Supplementary Figure 4. Number of genes encoded by complete prokaryote genomes. The vertical red line indicates the 700 genes threshold between MGS and small CAGs. Gene numbers from all complete prokaryotes in the NCBI genome browser (http://www.ncbi.nlm.nih.gov/genome/browse/) are shown.



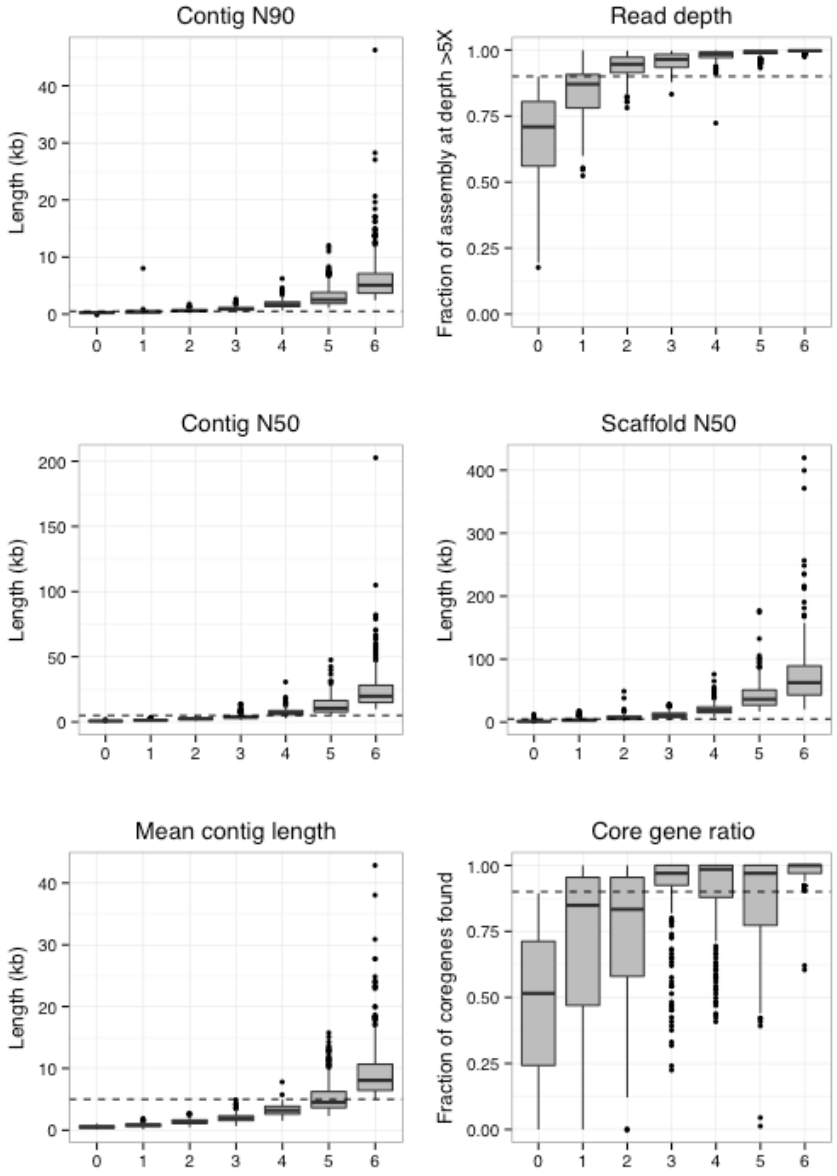
Supplementary Figure 5. Bar plot showing the taxonomical consistency of the MGS. The percentage of genes with the most common gene-wise taxonomical annotation for the given MGS is indicated in red, green and blue for species, genus and phylum level annotation, respectively. The percentage of genes annotated to an alternative species, genus or phylum is indicated in light-red, light-green and light-blue, respectively. The area above the bars indicates the percentage of genes without taxonomy annotation. On average, only 1.8% of the genes in an MGS are more similar to alternative species and 518 MGS have no species level similarity to any previously sequenced genome. For the remaining genes, no taxonomical assignment at the indicated level was found. Species, genus and phylum level taxonomical annotation was defined as best sequence match with 95%, 85% and 75% identity over ≥ 100 bp (for details see Methods and Supplementary Data 2).



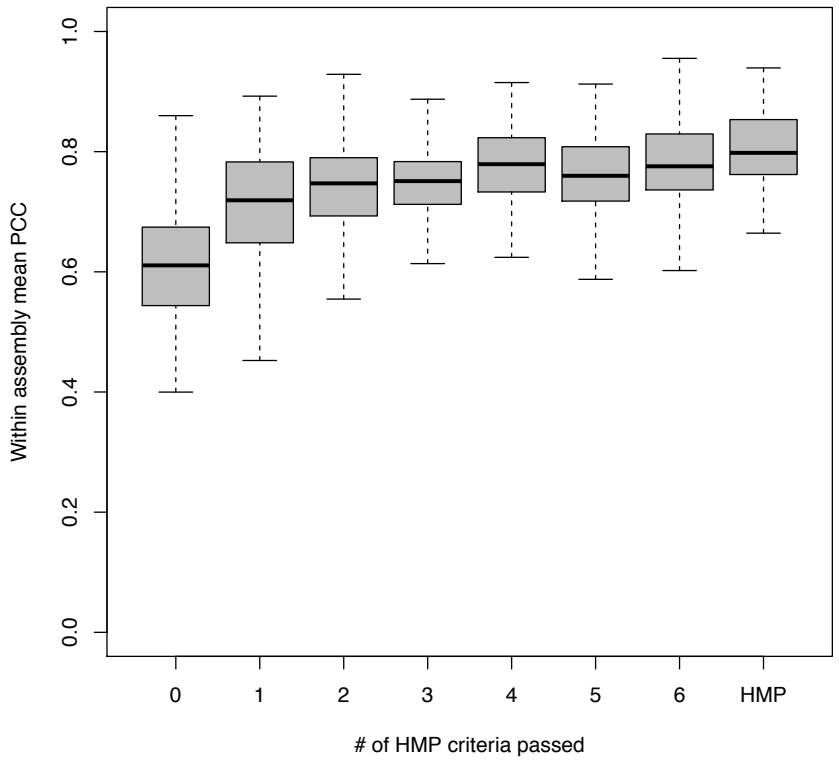
Supplementary Figure 6. Bar plot showing the number of known and previously unsequenced MGS for the indicated phylum. The MGS were assigned to a phylum if 90% of the genes annotated (best hit >75% identity over ≥ 100 bp) indicated the same phylum and more than 100 genes were annotated (for details see Supplementary Data 2).



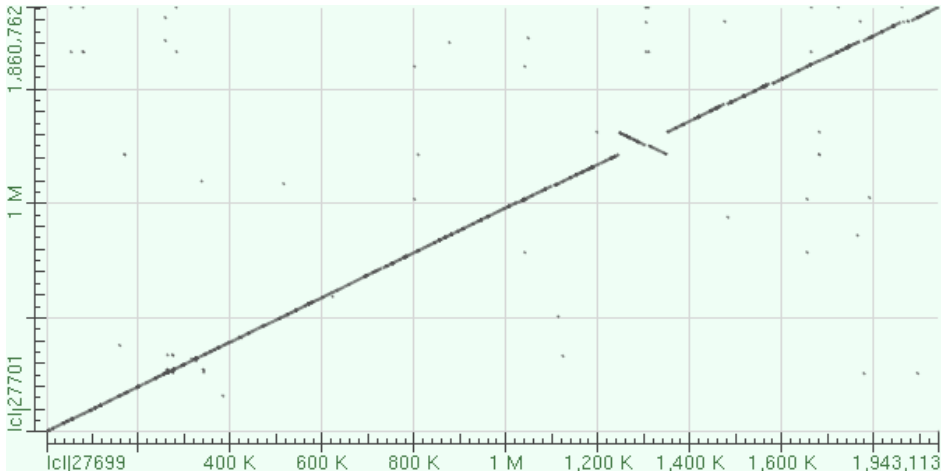
Supplementary Figure 7. Phylogenetic analysis of MGS augmented assemblies. A) Phylogeny of 337 assemblies (passing 5 or more HMP criteria) plus 1,637 reference genomes including 296 HMP microbiome gastrointestinal tract reference genomes[65]. The coloured ring shows the taxonomy of the reference genomes: green: Bacteroidetes, orange: Firmicutes, light pink: Actinobacteria, light orange: Chlamydia, red: Crenarchaeota, dark blue: Cyanobacteria, yellow: Euryarchaeota, light purple: Fusobacteria, light blue: Proteobacteria, light green: Spirochaetes, purple: Tenericutes and in black: CAG assemblies. The phylogenetic tree was created using the approximate maximum likelihood method implemented in FastTree on an alignment of 40 marker proteins, and visualized using ITOL[54]. The clade marked by the red ellipsoid is shown in B. B) Sub-tree of A, containing the *Tenericutes*, some *Firmicutes* and a clade of 27 CAGs (indicated with a red bar). The branches are coloured by bootstrap support values, where red shows values of 0.95 or higher and blue below 0.95. The MGS clade only consists of MGS augmented assemblies and forms a sister group to the family Acholeplasmataceae (class: Mollicutes). The assembly quality of the species in this clade is comparable to other MGS augmented assemblies (Supplementary Data 3). C) Box-plot of the distribution of gene-wise GC content of MGS. CAGs belonging to the clade shown in B are indicated in red and demonstrate low CG content. D) Box-plot of the gene content of CAG assemblies for the indicated groups. E) Box-plot showing the ratio between the number of cell wall and other essential genes[40] for the assemblies as indicated (Wilcoxon rank sum test, $P = 6 \times 10^{-16}$).



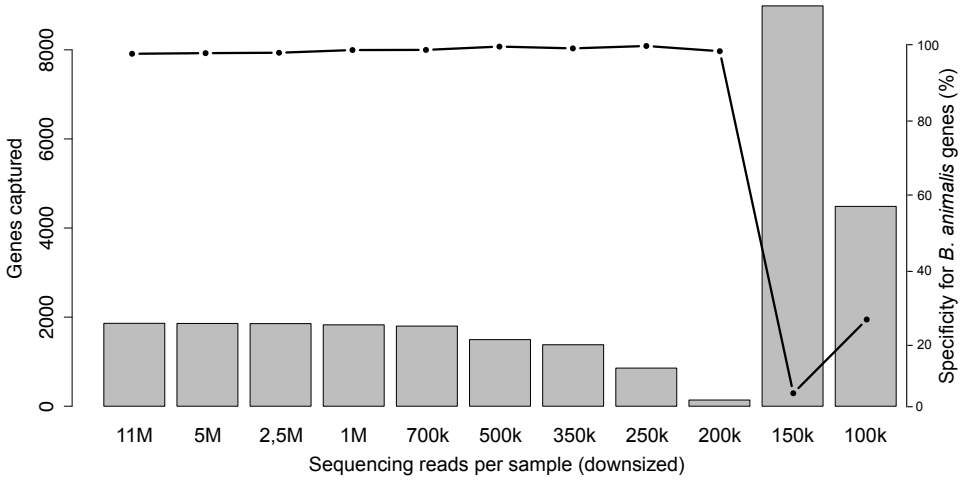
Supplementary Figure 8. Statistics of the 741 MGS augmented assemblies and visualization of HMP high quality draft genome criteria. The assemblies are divided by how many HMP criteria they pass (x-axis) where passing six criteria equals a high quality assembly. The horizontal dashed lines represent the HMP thresholds for the particular criteria. The lower and upper hinges correspond to the 25th and 75th percentiles, the whiskers represents the 1.5 * Inter-Quartile Range (IQR) extending from the hinges and the dots represents outliers from these. The two assemblies in the "Core gene ratio" panel that pass six criteria but only identified 60% of the core genes are archaeal organisms and they pass the archaea core gene ratios criteria.



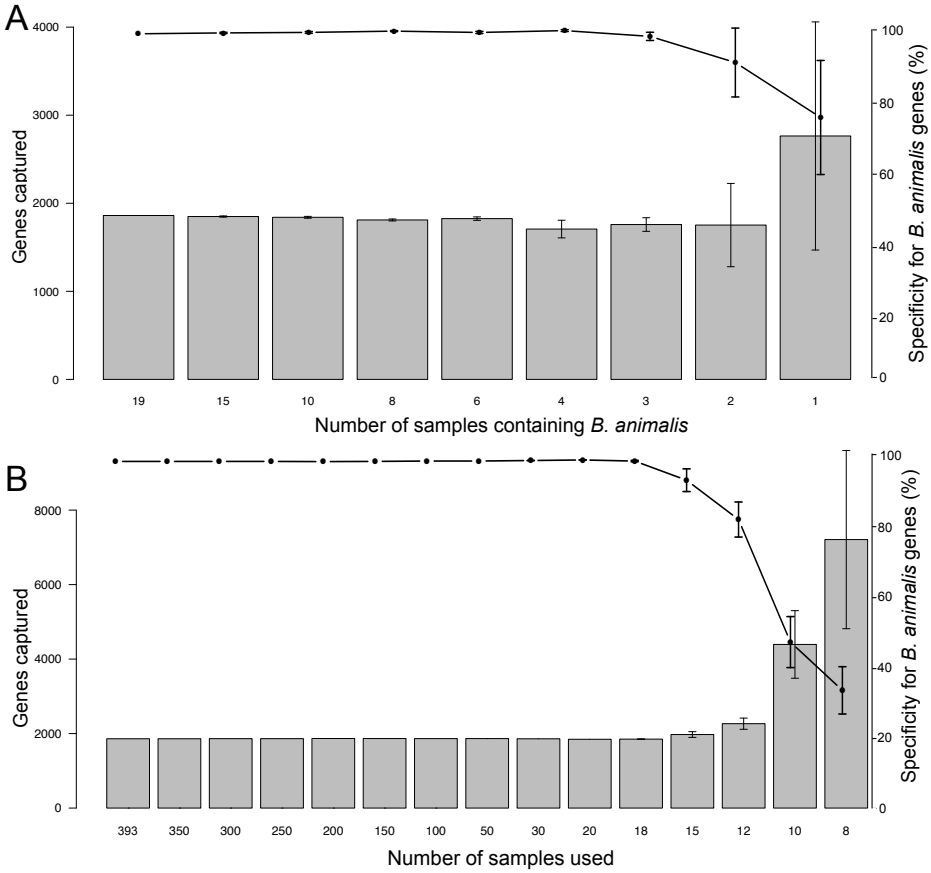
Supplementary Figure 9. Intra-assembly correlation of Tetra Nucleotide Frequency z-scores (TNF-z) to assembly-specific TNF-z median profiles. The TNF-z profile of all scaffolds within an assembly was correlated using the Pearson Correlation Coefficient (PCC) to the median TNF-z profile of the particular assembly. The figure shows the distribution of assembly mean PCC binned by the number of HMP criteria that the particular assemblies passed (0-6, 6 equals high quality draft) and the 296 HMP gastrointestinal tract reference genomes. The high quality assemblies show similar average PCCs to their median profile as the HMP reference genomes, indicating a similar coherency of the MGS high quality assemblies as the HMP reference genomes created using standard growth, sequencing and assembly techniques.



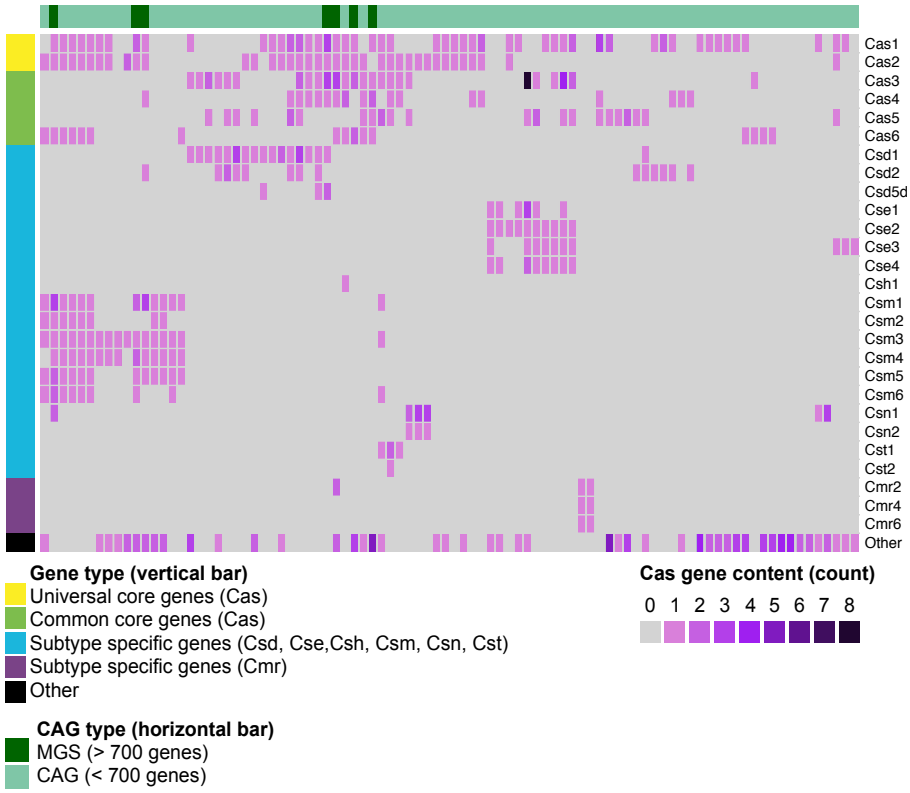
Supplementary Figure 10. Blast dot-plot of MGS:337 assembly (y-axis) vs. the *Bifidobacterium animalis subsp. lactis CNCM I-2494* reference genome (x-axis), showing very high consistency between the MGS assembly and the reference genome. 19 human subjects consumed a defined fermented milk product containing *Bifidobacterium animalis subsp. lactis CNCM I-2494*; MGS:337 was assembled from one of them (sample O2.UC47-2).



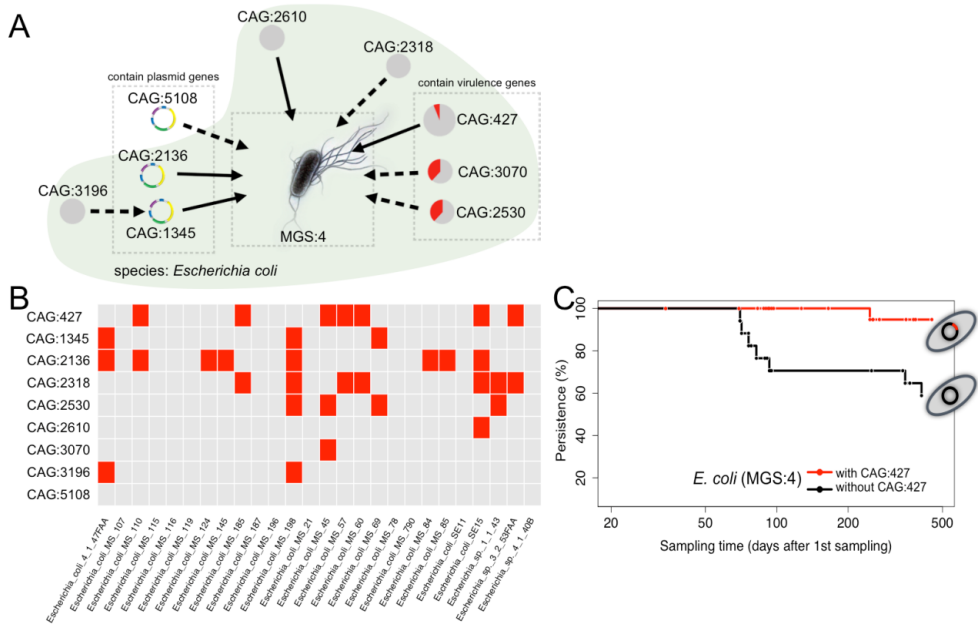
Supplementary Figure 11. Sensitivity and specificity of the co-abundance clustering at reduced sequencing depths. A series of co-abundance clusterings based on data downsized to the sequencing depths indicated at the x-axis was performed. From these the size (number of genes captured) and specificity in terms of *Bifidobacterium animalis subsp. lactis CNCM I-2494* matching genes (> 95% identity) of the MGS:337, are shown as bars and line, respectively. At a sequence depth of 700K reads 97% of the *B. animalis subsp. lactis CNCM I-2494* genes were captured and at a depth of 200K read 98.6% of the genes matched the reference genome. Here 393 stool samples (with 11M or more reads, before downsizing), including 19 samples from individuals who had consumed a defined fermented milk product contain *B. animalis subsp. lactis CNCM I-2494*, were used. Across these 19 samples on average 0.3% of the reads mapped unambiguously to MGS:334 genes and at a sequencing depth of 700k this corresponded to on average 1,962 reads per sample.



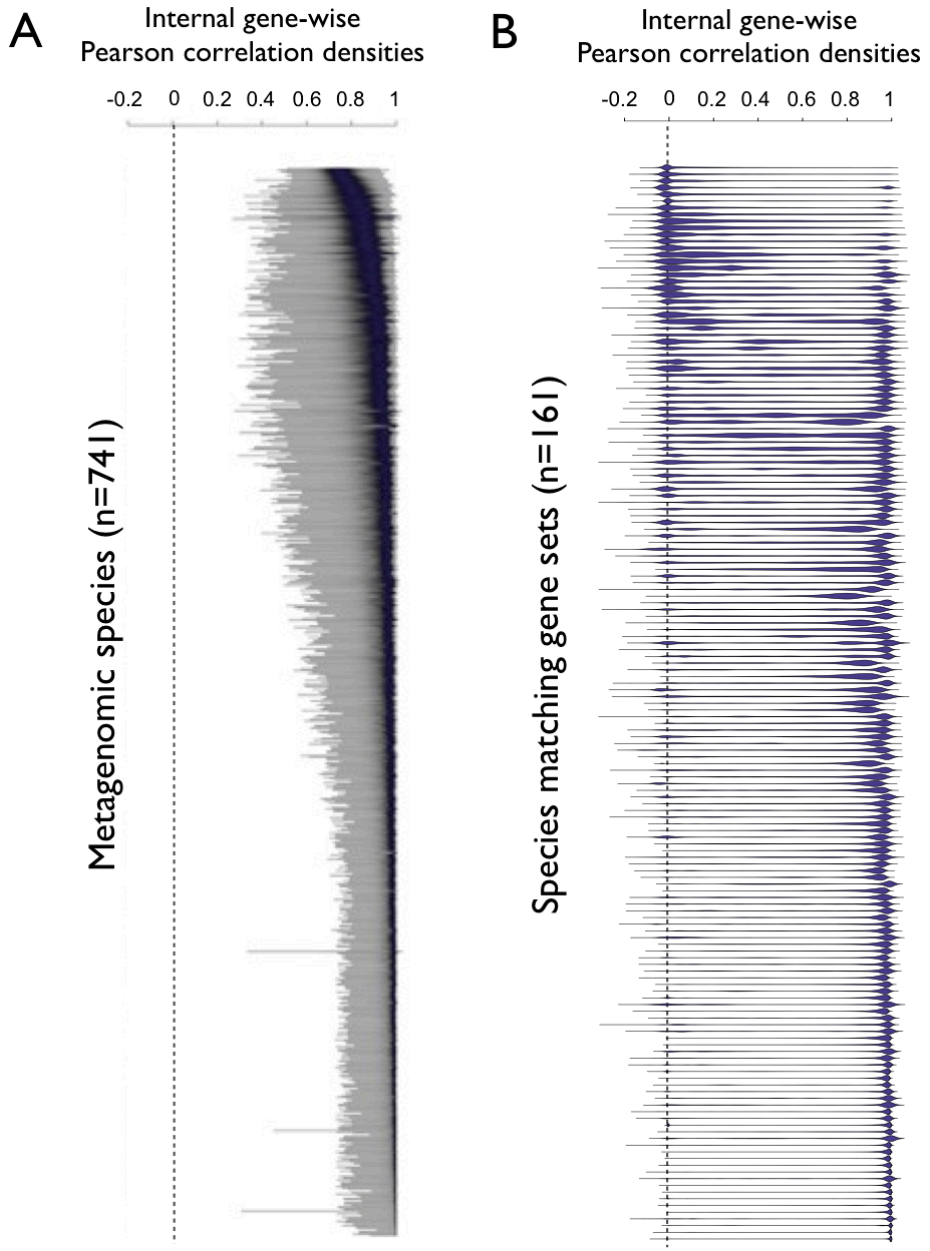
Supplementary Figure 12. Co-abundance clustering of the *Bifidobacterium animalis subsp. lactis CNCM I-2494* (MGS:337) in subsets of samples. The size (left axis) and specificity (right axis) of the *B. animalis* MGS from a series of co-abundance clusterings on random sample subsets are shown. All samples were downsized to 11M sequence reads. The bars indicate the mean number of genes captured in the MGS and the line indicates the mean percentage of the captured genes with strong similarity to *B. animalis subsp. lactis CNCM I-2494* (95% identity over 100 bp or better). Whiskers indicate +/- one standard deviation from the mean (n = 5). In **A**) the results of co-abundance clusterings of 375 stool samples, including the indicated number of samples (x-axis) from individuals, who had consumed a defined fermented milk product (DFMP), containing *B. animalis subsp. lactis CNCM I-2494*, are shown. **B**) Shows the results of co-abundance clusterings on a series of random sample subsets of the indicated size (x-axis). These subsets contained the maximal number of individuals possible who had consumed the DFMP (19-8 DFMP consuming individuals).



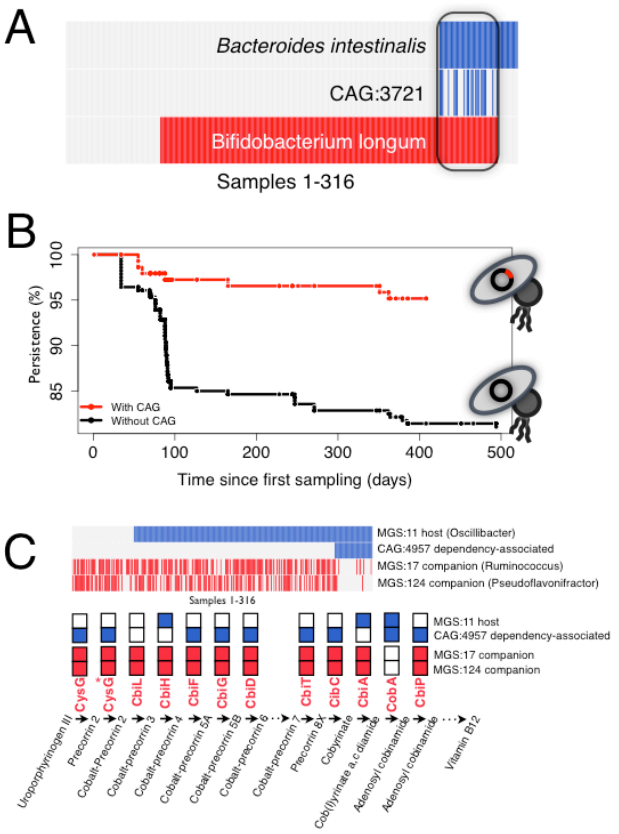
Supplementary Figure 13. Heatmap showing CRISPR-associated (Cas) genes annotated for MGS and CAGs that were found enriched for CRISPR related genes. The rows show the occurrence of Cas genes in 83 Cas enriched MGS and CAGs (columns). The colour coding at the left corresponds to the subtypes of Cas genes and the colours on top indicate the CAG type. The CAGs cluster according to subtypes of Cas genes they contain.



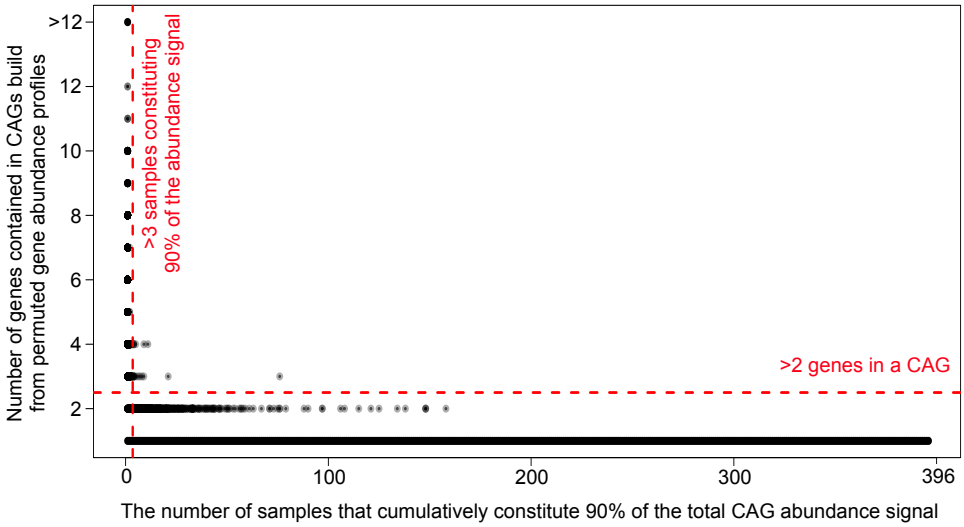
Supplementary Figure 14. *E. coli* dependency-associated CAGs. **A**) The dependency-association sub-network centred around *E. coli* (MGS:4). Arrows show directional dependency-associations. The green background colouring indicates CAGs dominated by genes with species level similarity to *E. coli*. The proportion of virulence genes is indicated by pie charts for CAG:427, CAG:3070 and CAG:2530. Together, these three CAGs contain 15% of all the virulence genes found in the entire gene catalogue. CAG:5108, CAG:2136 and CAG:1345 are significantly enriched for plasmid genes. **B**) Identification of the nine MGS:4 dependency-associated CAGs across 25 HMP reference genomes of *E .coli* isolated from the human intestinal tract[65]. Red rectangles indicate a sequence match with >90% identity over >90% of the gene length between the indicated CAG and HMP reference genome. **C**) The cumulative persistence of (MGS:4) with or without the dependency-associated CAG:427 as observed between longitudinal samplings of 34 individuals. Points indicate sampling time of the second sample (in days) relative to the first sampling from the same individual.



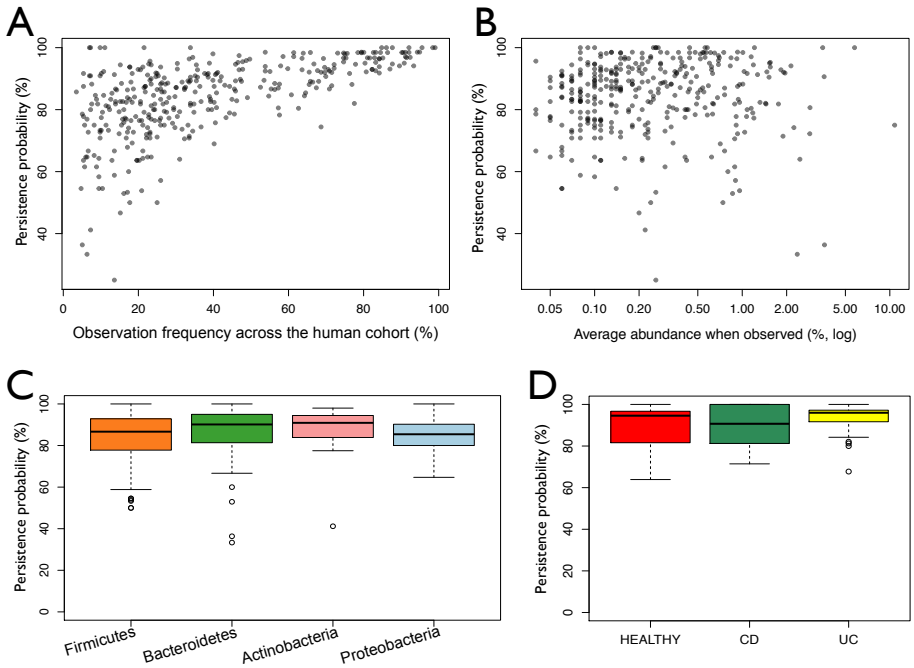
Supplementary Figure 15. Violin plots showing the densities of the internal gene abundance Pearson correlation coefficients for gene sets defined by A) MGS and B) 'reference species gene sets'. The thickness of the horizontal blue 'violins' (lines) indicates the densities of the distribution of Pearson correlation coefficients between the genes within a given gene set. The 'reference species gene sets are defined as sets, that share species level taxonomical assignment by sequence match to a reference genome (best hit, 95% identity over 100 bp or better). The horizontal scale is the same in the two plots. The 'reference species gene sets' and MGS are ordered vertically by the median PCC.



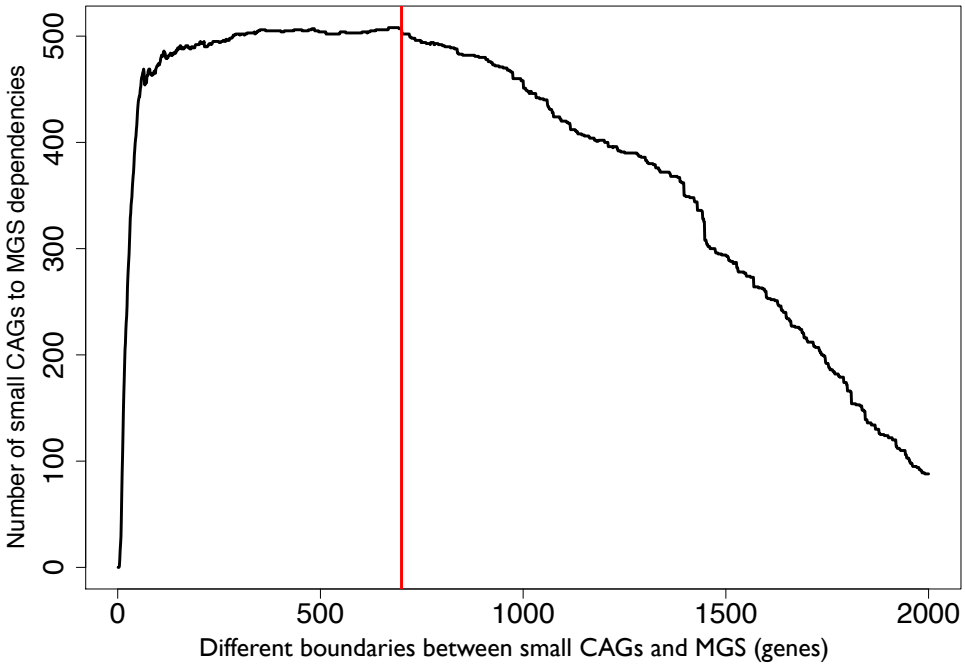
Supplementary Figure 16. Adaptations to the co-existence or absence of a companion species. **A**) Sample-wise detections of *Bacteroides intestinalis* (MGS:315) and CAG:3721 shown in blue bars and the occasional companion species *Bifidobacterium longum* (MGS:69) in red bars. CAG:3721 is significantly associated to the co-existence of the two species ($P = 2 \times 10^{-9}$). **B**) The cumulative persistence of species that live in co-existence with a companion species is shown as a function of dependency-associated CAGs. Points indicate sampling time (in days) of the second of two longitudinal samples relative to the first sampling. The curves show the joint observation of 18 inter-species relationships across 73 individuals, where a CAG coincide with substantially increased persistence of the hosting MGS. The average annual effect of carrying a positive CAG was 29% as estimated by Bayesian modelling (95% credible interval: 17 to 41 percentage points). **C**) Top panel shows the sample-wise detections of the host (MGS:11, *Oscillibacter*), the dependency-associated CAG:4957 and the two companion species (MGS:17, Ruminococcus like and MGS:124, Pseudoflavonifractor like) as indicated. The detection of CAG:4957 is significantly anti-correlated to the detections of the companion species (MGS:17: p = 0.0007 and CAG124: p = 0.002). Lower panel show the genetic potential for the anaerobic corrin ring part of the Vitamin B12 biosynthesis pathway[38] for the indicated CAGs as filled boxes. The precorrin 2 to cobalt-precorrin 2 step (marked with *) may be catalysed by both CysG and CbiK. Enzymes catalyzing steps between cobalt-precorrin 6 and 7 lagging experimentally verification are not shown[38]. A possible role for CAG:4957 is to compensate for the biosynthetic potential of the companion species in their absence.



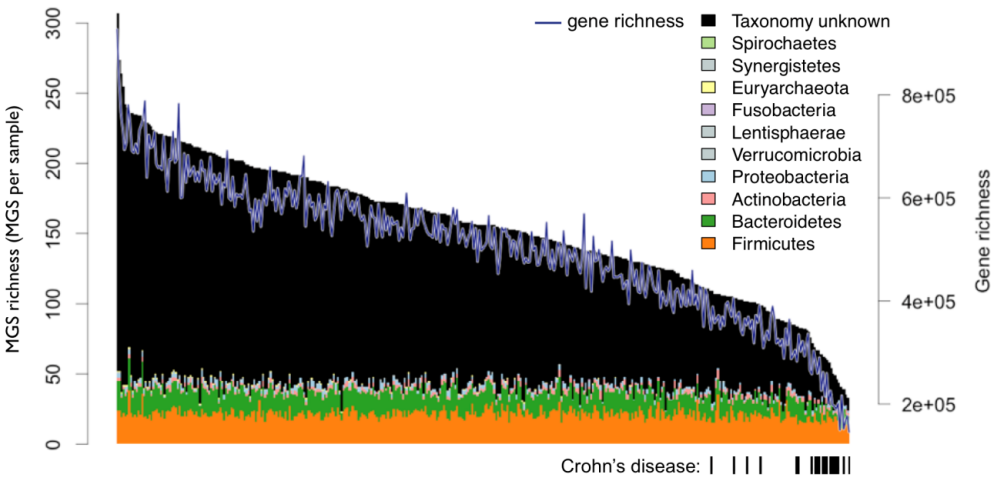
Supplementary Figure 17. Canopy clustering on permuted abundance profiles. The result of an exhaustive co-abundance binning of a gene-wise shuffled abundance matrix is shown. The size (number of genes) and minimal number of samples that constitute 90% of the total abundance signal from the resulting 1,840,781 random CAGs are shown. Only 18 CAGs escape the QC filter indicated with red dashed lines. All of these contained 3 or 4 genes and were observed in a few samples. 1,539,760 of the random CAGs contained 1 gene and 799 contained more than 12 genes. For all of the latter 90% or more of the abundance signal originated from only one sample. The estimated number of randomly occurring CAGs in the non-permuted canopy clustering (*i.e.* the real data) was very low and only expected among the rare and very small CAGs (FDR ~10% for CAGs with 3 or 4 genes).



Supplementary Figure 18. Persistence probabilities of MGS. A) Persistence probabilities (estimated from the 2 × 73 re-sampled individuals) as a function of the observation frequency across 318 independent human samples. B) Persistence probabilities as a function of the average abundance across the samples where the MGS was detected. C) Box-plot showing the persistence probabilities for MGS assigned to the four main phyla. D) The MGS persistence probabilities across the patient groups: Healthy, Crohn's disease (CD) and ulcerative colitis (UC). Persistence probabilities were estimated for MGS observed in 5 or more individuals.



Supplementary Figure 19. The number of statistically significant small CAG to MGS dependency-associations as a function of the gene-number boundary between these CAG classes. The vertical red line indicates the 700-gene definition. The odds ratio for small CAGs to MGS is 12.7 with a boundary at 700 genes.



Supplementary Figure 20. MGS richness across the 396 samples. The height of the bars indicates the number of MGS found in each sample (left axis) and the bar colour shows the phylum level taxonomy of the MGS that represents known species (range: 25 to 81, mean: 50). Black indicates the number of MGS without species level taxonomical annotation. The blue line indicates the sample-wise gene richness (right axis). The PCC between the MGS richness and the gene richness is 0.96 and only 0.55 for the taxonomically known species. Rectangles below the bar plot indicate samples from individuals with Crohn's disease.

SUPPLEMENTARY DATA

Supplementary Data 1, sample description
Supplementary Data 2, MGS taxonomical statistics
Supplementary Data 3, MGS augmented assembly statistics
Supplementary Data 4, MGS augmented assemblies comparison to reference genomes
Supplementary Data 5, summary information on the 6640 small CAGs
Supplementary Data 6, dependency-association network
Supplementary Data 7, MGS:4 + dependency-associated CAG assembly statistics
Supplementary Data 8, eggNOG prevalent in frequently observed MGS
Supplementary Data 9, *Bacillus subtilis* essential COG list
Supplementary Data 10, gene catalogue comparison
Supplementary Data 11, MGS frequencies and abundance statistics
Supplementary Data 12, dependency-associations with or without companion species

REFERENCES (FOR SUPPLEMENTARY INFORMATION)

60. Juhas, M., Eberl, L. & Glass, J. I. Essence of life: essential genes of minimal genomes. *Trends Cell Biol*. **21,** 562–8 (2011).
61. Toussaint, A. & Chandler, M. Prokaryote genome fluidity: toward a system approach of the mobilome. *Methods Mol. Biol*. **804,** 57–80 (2012).
62. Manichanh, C. *et al*. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–11 (2006).
63. Abbas, A., Adams, C., Scully, N., Glennon, J. & O'Gara, F. A role for TonB1 in biofilm formation and quorum sensing in Pseudomonas aeruginosa. *FEMS Microbiol. Lett*. **274,** 269–78 (2007).
64. Wolanin, P., Thomason, P. & Stock, J. Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome Biol*. **3,** reviews3013.1–reviews3013.8 (2002).
65. Nelson, K. E. *et al*. A catalog of reference genomes from the human microbiome. *Science* **328,** 994–9 (2010).